Mining Protein Interactions of Phosphorylated Proteins from the Literature using eFIP

Catalina O Tudor[1], Cecilia N Arighi[1,2], Qinghua Wang[1,2], Cathy H Wu[1,2], K Vijay-Shanker[1]
[1] Department of Computer and Information Sciences
[2] Center for Bioinformatics and Computational Biology
University of Delaware, Newark DE, USA

Abstract

There has been a general shift in paradigm from dedicating a lifetime's work to analyzing of a single protein to the analysis of cellular and biochemical processes and networks. Although bioinformatics tools have greatly assisted in data analysis, both protein identification and functional interpretation are still major bottlenecks. In this regard, public knowledge bases constitute a valuable source of such information, but the manual curation of experimentally determined biological events is slow compared to the rapid increase in the body of knowledge represented in the literature. Hence, literature still continues to be a primary source of biological data. Nevertheless, manually finding relevant articles is not a trivial task, with issues ranging from the ambiguity of some names to the identification of those articles that contain the specific information of interest. One important aspect of proteins is their phosphorylated states and their implication in protein interacting networks. We have developed eFIP, a web-based tool, which aids scientists to find quickly abstracts mentioning phosphorylation of a given protein (including site and kinase), coupled with mentions of interactions, and evidence for impact of phosphorylation on the interaction. eFIP combines information provided by applications such as eGRAB, RLIMS-P, and an in-house PPI module, and displays the results in a highlighted and tabular format for a quick inspection.

1. Introduction

One important aspect of proteins is their phosphorylated states and their implication in protein function and protein interacting networks. Phosphorylation of specific intracellular proteins/ enzymes by protein kinases and dephosphorylation by phosphatases provide information of both activation and deactivation of critical cellular pathways, including regulatory mechanisms of metabolism, cell division, cell growth and differentiation. Often, protein phosphorylation has some functional impact. Proteins can be phosphorylated on different residues, leading to activation or down-regulation of their activity, alternative subcellular location, and binding partners. One such example is protein Smad2, whose phosphorylation state determines its interaction partners, its subcellular location, and its cofactor activity.

However, protein-protein interaction (PPI) data involving phosphorylated proteins is not yet well represented in the public databases. Extracting this information is critical to the interpretation of PPI and prediction of the functional outcomes, and this is the main motivation for capturing this type of information in Protein Ontology (PRO).

To help curators find quickly abstracts mentioning phosphorylation of a given protein (including site and kinase), coupled with mentions of interactions and possible impact of phosphorylation, we have developed eFIP (Extracting Functional Impact of Phosphorylation) [1]. This tool currently focuses on identifying modified forms of proteins and the participation of these phosphorylated forms in protein-protein interaction. The aim of eFIP is to be used directly in the curation process for Pro Ontology.

2. The System's Modules

Figure 1 illustrates the pipeline of the system. This involves a document retrieval module, called eGRAB [4], an information extraction tool for identifying mentions of phosphorylation, called RLIMS-P [2,3], and a protein-protein interaction module developed in house. The goal, delimited by the last step in the pipeline, is to identify the participation of phosphorylated forms of a protein in protein-protein interaction.

Possible inputs are a gene name, a gene identifier, or a list of PMIDs. Although the gene identifiers are species-specific, we do not limit the abstracts to the ones mentioning the gene with the corresponding species. Instead, we use the identifier to retrieve all possible names of the gene to be used in the document retrieval module. We concentrate on computing results for genes coming from vertebrates. Currently, eFIP considers only the abstract of a paper, since

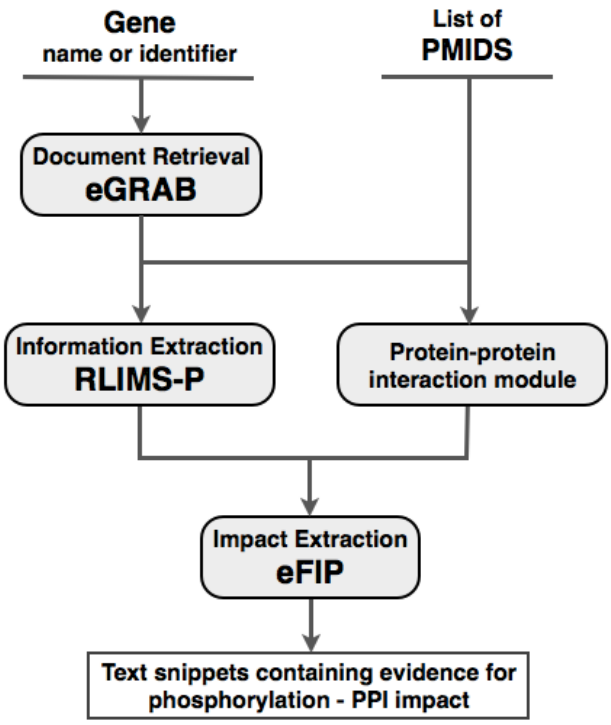

Figure 1. General pipeline of the system, including the Document Retrieval, Information Extraction, Protein-Protein Interaction and Impact Extraction Modules.

one of the components, RLIMS-P, works only with abstracts. We plan to extend our approach to the Results section in the near future.

2.1 Extractor of Gene-Related ABstracts (eGRAB)

eGRAB is used to gather the literature for a given gene/protein. eGRAB starts by gathering all possible names and synonyms of a gene/protein from knowledge bases of genes and proteins (such as EntrezGene and UniProtKB), searches PubMed using these names, and returns a set of disambiguated Medline abstracts to serve as the gene's literature. This technique filters potentially irrelevant documents that mention the gene names in some other context, by creating language models for all the senses and assigning the closest sense to an ambiguous name. eGRAB is currently being used in other systems. The approach and its evaluation are provided in [4].

2.2 Rule-based LIterature Mining System for Protein Phosphorylation (RLIMS-P)

RLIMS-P is a system designed for extracting protein phosphorylation information from MEDLINE abstracts. It extracts the three objects involved in this process -- the protein kinase, the phosphorylated protein (substrate), and the phosphorylation site (residue/position being phosphorylated). RLIMS-P utilizes extraction rules that cover a wide range of patterns, including some specialized terms used only with phosphorylation. Additionally, RLIMS-P employs techniques to combine information found in different sentences, because rarely are the three objects (kinase, substrate, and site) found in the same sentence. RLIMS-P has been benchmarked and the results are presented in [2]. A more detailed description of the system can be found in [3].

2.3 The Protein-Protein Interaction Module

Many PPI tools have been described in the literature. However, we could not find a PPI tool available for download that we could use in eFIP's pipeline. Additionally, the PPI tool would have to be easily adaptable for our needs. One example of additional features we wanted to be able to incorporate is the ability to detect interactions involving only one partner, in the cases in which the other partner is implicit, or the ability to detect anaphora resolution when one of the partners or both are described by pronouns "it" or "they".

Hence, the PPI module is an in-house implementation designed to detect mentions of PPI in text. This tool extracts text fragments, or text evidence, that explicitly describe a type of PPI (such as binding and dissociation), as well as the interacting partners. The primary engine of this tool is an extensive set of rules specialized to detect patterns of PPI mentions. The interacting partners identified are further sent to a gene mention tool to confirm whether they are genuine protein mentions. Consider the sample phrase "several proapoptotic proteins commonly become associated with 14-3-3." "14-3-3" is a protein, whereas "several proapoptotic protein" prompts the need to further identify the actual proteins (Bad and FOXO3a) that interact with 14-3-3.

2.4 The extraction of phosphorylation impact on PPI

Our main goal is to find interacting information about a particular protein when it is in its phosphorylated state. For this, we select abstracts that contain both phosphorylation and PPI mentions involving the same protein. We define impact as the influence or dependency of the phosphorylation on the protein-protein interaction. Therefore, the impact can fall into any of these categories: none (it cannot be established confidently or the interaction does not depend on protein phosphorylation), enables interaction (the phosphorylation creates a binding site for a given protein binding partner), prevents interaction (the phosphorylation abrogates a binding site), increases interaction (the phosphorylation increases the affinity for the binding partner), and decreases interaction (the phosphorylation decreases the affinity for the binding partner)

For example, consider the following sentence: “Phosphorylated Bad binds to the cytosolic 14-3-3”. In this example, we can tell that the phosphorylation happens before the binding, as one of the interactants is reported to be “phosphorylated Bad”. However, we cannot tell if the phosphorylation has any impact on the binding itself, i.e., if 14-3-3 binds to Bad regardless of its form, phosphorylated or non-phosphorylated. In contrast, the next sentence shown here not only mentions the phosphorylation happening before the interaction, but also describes how the interaction is dependent on the phosphorylation: “Bad phosphorylation induced by survival factors leads to its preferential binding to 14-3-3 and suppression of the death-inducing function of Bad.”

3. Results and user interaction

The input in eFIP is a gene name (or identifier), or a list of PMIDs. The output is a ranked list of PMIDs, each accompanied by a summary of the information found within.

If a gene name or identifier is provided, eFIP outputs all relevant articles where the corresponding protein is phosphorylated and implicated in a protein-protein interaction. eFIP ranks these papers, taking into consideration the confidence assigned to each of the steps involved: the detection of the phosphorylation mention, the detection of the partners of the interaction, and the detection of the impact of phosphorylation on the interaction. If a list of PMIDs is provided as input, then multiple phospho-proteins might be involved in protein-protein interactions. eFIP lists relevant PMIDs for one phospho-protein at a time, first considering the phospho-protein that has more mentions of phosphorylation and PPI in the documents provided.

As an example, consider protein BAD as input to eFIP. There are 1,331 documents linked to protein BAD as determined by eGRAB. Alternatively, we can provide the following list of PMIDs:

8929531, 10949026, 11526496, 11583580, 12351720, 12438947, 15896972, 16139821, 16403219, 19221220.

An example output is shown in Figure 2.

1. Survival-factor-induced phosphorylation of Bad results in its dissociation from Bcl-x(L) but not Bcl-2
Hirai I, Wang HG

Relevant Irrelevant

PMID 11583580 see in PubMed | read abstract here Validate the results
Info: prevent BAD - Bcl-x(L); prevent BAD (Ser-112) - Bcl-x(L); enable BAD (Ser-112) - 14-3-3 proteins

2. Serine phosphorylation of death agonist BAD in response to survival factor results in binding to 14-3-3 not Bcl-X(L).
Zha J, Harada H, Yang E, Jockel J, Korsmeyer SJ

Relevant Irrelevant

PMID 8929531 see in PubMed | read abstract here Validate the results
Info: enable BAD - 14-3-3; enable BAD - Bcl-X(L)

3. 14-3-3 proteins and survival kinases cooperate to inactivate BAD by BH3 domain phosphorylation
Datta SR, Katsov A, Hu L, Petros A, Fesik SW, Yaffe MB, Greenberg ME

Relevant Irrelevant

PMID 10949026 see in PubMed | read abstract here Validate the results
Info: prevent BAD (Ser-155) - prosurvival Bcl-2 proteins

Figure 2. Example output of ranked PMIDs for protein BAD.

The user can click on “read abstract here”, and this will display the abstract with the relevant information (phospho-protein, phospho-site, interactant, impact words) being underlined. Most of the times the information can be found in one sentence taken from the abstract, but the information could also span multiple sentences if the different mentions involved in the impact are scattered in text.

The user will interact with the system both at the input and output stages. In the case in which the input is a gene name, the system interacts with the user by providing all the genes that match the specified name, and it then asks the user to select the correct gene from the list. Once the results are displayed, the user will be able to tell the system which instances are correctly identified with an impact and which instances are wrong. Corrections can also be provided both for the phosphorylation mention and for the PPI mention. Moreover, the user will be able to link the phosphorylated protein to an actual knowledge base entry, as well as provide a way to normalize the protein names when different textual variants are used (e.g., “CPS1”, “CPSI”, and “CPS 1” all refer to the same protein).

4. Proposed Task for the Curator

To assess the system in terms of precision and recall, we will ask one curator to build a gold standard based on a list of 50 PMIDs that mention phosphorylation of various proteins. Given this set of abstracts, the curator will manually read these abstracts and record in a spreadsheet the phosphorylation, interaction, and impact information. We will then use this set of abstracts to run eFIP and compare the results with the gold standard. To assess the usefulness of the system, we will ask a group of curators to annotate the same set of PMIDs using eFIP. The curators will go over the list of articles and validate the ranking by clicking on the relevant/non-relevant buttons, as shown in Figure 2. The curators will also assess the information extracted, by clicking on the "Validate the results" link for any given PMID, as well as provide corrections for each of the phospho-protein, phospho-site, interactant, and impact, as shown in Figure 3.

Survival-factor-induced phosphorylation of Bad results in its dissociation from Bcl-x(L) but not Bcl-2.

The pro-apoptotic Bcl-2-family protein Bad heterodimerizes with Bcl-2 and Bcl-x(L) in the outer mitochondrial membranes, nullifying their anti-apoptotic activities and promoting cell death. We report that interleukin-3 (IL-3) stimulation induces Bad phosphorylation and triggers its translocation from mitochondria to cytoplasm in cells expressing Bcl-x(L) but not Bcl-2. Overexpression of Bad sensitized Bcl-x(L)-expressing FL5.12 cells to apoptosis induced by IL-3 deprivation, but had no effect on the viability of cells expressing Bcl-2. IL-3 stimulation induced Bad phosphorylation at Ser-112, impairing its binding to Bcl-x(L) and resulting in its association with 14-3-3 proteins in the cytosol. However, Ser-112 phosphorylation could not trigger Bad dissociation from mitochondria in FL5.12 cells expressing Bcl-2...

Phospho-protein	Phospho-site	Interactant	Impact		
BAD	-	Bcl-x(L)	prevent	Accept	Remove
BAD	Ser-112	Bcl-x(L)	prevent	Accept	Remove
BAD	Ser-112	14-3-3 proteins	enable	Accept	Remove

Figure 3. Example output of information extracted from PMID 11583580.

References

1. Arighi, C.N., Siu, A.Y., Tudor, C.O., Nchoutmboube, J.A., Wu, C.H., and Shanker, V.K. (2011) eFIP: A Tool for Mining Functional Impact of Phosphorylation from Literature. *Bioinformatics for Comparative Proteomics,* Methods in Molecular Biology, vol. 694, 63--75.
2. Hu, Z.Z., Narayanaswamy, M., Ravikumar, K.E., Vijay-Shanker, K., and Wu, C.H. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics* 21, 2759--2765.
3. Narayanaswamy, M., Ravikumar, K.E., and Vijay-Shanker, K. (2005) Beyond the clause: extraction of phosphorylation information from Medline abstracts. *Bioinformatics* 21 Suppl 1, i319--i327.
4. Tudor, C.O., Schmidt, C.J., Vijay-Shanker, K. (2010) eGIFT: Mining Gene Information from the Literature. *BMC Bioinfomatics*. vol. 11, 418.