# Chapter IX
# Information Extraction of Protein Phosphorylation from Biomedical Literature

**M. Narayanaswamy**
*Anna University, India*

**K. E. Ravikumar**
*Anna University, India*

**Z. Z. Hu**
*Georgetown University Medical Center, USA*

**K. Vijay-Shanker**
*University of Delaware, USA*

**C. H. Wu**
*Georgetown University Medical Center, USA*

## ABSTRACT

*Protein posttranslational modification (PTM) is a fundamental biological process, and currently few text mining systems focus on PTM information extraction. A rule-based text mining system, RLIMS-P (Rule-based LIterature Mining System for Protein Phosphorylation), was recently developed by our group to extract protein substrate, kinase and phosphorylated residue/sites from MEDLINE abstracts. This chapter covers the evaluation and benchmarking of RLIMS-P and highlights some novel and unique features of the system. The extraction patterns of RLIMS-P capture a range of lexical, syntactic and semantic constraints found in sentences expressing phosphorylation information. RLIMS-P also has a second phase that puts together information extracted from different sentences. This is an important feature since it is not common to find the kinase, substrate and site of phosphorylation to be mentioned*

*in the same sentence. Small modifications to the rules for extraction of phosphorylation information have also allowed us to develop systems for extraction of two other PTMs, acetylation and methylation. A thorough evaluation of these two systems needs to be completed. Finally, an online version of RLIMS-P with enhanced functionalities, namely, phosphorylation annotation ranking, evidence tagging, and protein entity mapping, has been developed and is publicly accessible.*

## INTRODUCTION

Protein post translational modification (PTM), a molecular event in which a protein is chemically modified during or after its being translated, is essential to many biological processes. Protein phosphorylation is one of the most common PTMs, which involves the addition of a phosphate group to serine, threonine or tyrosine residues of a protein, and is fundamental to cell metabolism, growth and development. Many cellular signal transduction pathways are activated through phosphorylation of specific proteins that initiate a cascade of protein-protein interactions, leading to specific gene regulation and cellular response. It is estimated that one third of the mammalian genome coding sequences code for phosphoproteins. The phosphorylation state of cellular proteins is also highly dynamic, detection, quantification and functional analysis of the dynamic phosphorylation status of proteins, and the kinases involved are essential for understanding the regulatory networks of biological pathways and processes, which are under extensive investigation by researchers of many areas of biological research.

While PTMs are fundamental to our understanding of cellular processes, the experimental PTM data are largely buried in free-text literature. For example, a recent PubMed query for protein phosphorylation returned 103,478 papers. Although PTMs, especially phosphorylation, are among the most important protein features annotated in protein databases, currently only limited amount of data are annotated in a few resources, such as UniProt Knowledgebase (UniProtKB) (Wu et al., 2006), and specialized databases including Phospho.ELM and PhosphoSite, which can not keep up with the fast-growing literature. With the increasing volume of scientific literature now available electronically, efficient text mining tools will greatly facilitate the extraction of information buried in free text. Information extraction of PTM information on specific proteins, sites/residues being modified, and enzymes involved in the modification are particularly useful not only to assist database curation for protein site features and related pathway or disease information, but also to allow users to quickly browse and analyze the literature, and help other bioinformatics software to integrate text mining component into pathway and network analysis.

There are many BioNLP relation extraction systems that have been developed in the past few years. Some of these employ special rule/pattern based approaches (e.g., Blaschke et al., 1999; Pustejovsky et al., 2002). Other approaches for extracting protein-protein interactions include detecting co-occurring proteins (Proux et al., 2000; Stapley and Benoit, 2000; Stephens et al., 2001), or using a text parser tailored for the specialized language typically found in the biology literature (e.g., Friedman et al., 2001; Daraselia et al., 2004). The rule-based approach involves designing patterns to extract specific types of information, while the parser approach requires development of grammars, methods for disambiguation and further effort to provide methods that map parse information to objects involved in the relation. More modern approaches employ machine learning for relation extraction (e.g., Bunescu and Mooney, Gioliana et al). Such methods require an annotated corpus, where the sentences