With the accumulation of a significant amount of data about prokaryotic glycosylation, it is tempting to organize this information in a streamlined way so that the concerned researchers would be able to take some benefit from this. For this reason, we set out to create the first database of prokaryotic glycoproteins. ProGlycProt (http://www.proglycprot.org/) is an open access, manually curated, comprehensive repository of bacterial and archaeal glycoproteins with at least one experimentally validated glycosite (glycosylated residue). To avoid confusion between characterized and uncharacterized glycoproteins, the database has been accordingly divided into two sections: (i) ProCGP—the main data section consisting of 95 entries with experimentally characterized glycosites and (ii) ProUGP—a supplementary data section containing 245 entries with experimentally defined glycosylation but uncharacterized glycosites. Manifold information acquired from published sources is provided for each entry, which is also fully cross-referenced. The fields constituting each entry include source organism, coding gene, protein, glycosites, glycosylation type, attached glycan, associated oligosaccharyl/glycosyl transferases (OSTs/GTs), supporting references, and applicable additional information. About 174 entries are there in ProGlycProt for which characterization (including that of glycosites) is unavailable in Swiss-Prot release 2011_07. For characterized glycoproteins, a dedicated gallery of homology models and crystal structures is provided in addition to two new tools developed in view of emerging information about prokaryotic sequons that are absent or rarely seen in eukaryotic glycoproteins. ProGlycProt provides an extensive compilation of experimentally identified glycosites (334) and glycoproteins (340) of prokaryotes that could serve as an information resource for research and technology applications in glycobiology.

Prokaryotic glycoproteins have caught an intent attention of many researchers of late due to their potential of being used in technology based applications. One of the initial studies that are routinely carried out vis-à-vis glycoprotein characterization entails the analysis of

glycosites (glycosylated-residues), and associated sequence and structural features. In spite of the availability of many glycosite prediction tools, none is reliably fit for predicting prokaryotic glycosites. Therefore, in this study, new Support Vector Machine (SVM) based algorithms (models) were developed for predicting glycosites with high accuracy in prokaryotic protein sequences. For these models, binary profile of patterns (BPP), composition profile of patterns (CPP), and position-specific scoring matrix (PSSM) profile of patterns (PPP) were used as training features. The prediction tool GlycoPP was developed in collaboration with Dr. GPS Raghava's research group at IMTECH. In this study, 59 experimentally characterized glycoproteins of prokaryotes were used to extract an extensive dataset of 107 N-linked and 116 O-linked glycosites that includes validated N-glycosites from phyla *Crenarchaeota*, *Euryarchaeota* (domain Archaea), *Proteobacteria* (domain Bacteria) and validated O-glycosites from phyla *Actinobacteria*, *Bacteroidetes*, *Firmicutes* and *Proteobacteria* (domain Bacteria). Based on the fact that prokaryotic glycosylation predominantly occurs on residues present in loops/accessible areas in folded proteins, hybrid models were developed using predicted secondary structure and surface accessibility features in various combinations with training features. With these models, prediction accuracies of 82.71% (MCC 0.65) and 73.71% (MCC 0.48) were obtained for N-glycosites and O-glycosites, respectively. Suitability of these models for reliably predicting N- and O-glycosites in potential glycoproteins was ascertained by evaluating the best performing models with 28 independent prokaryotic glycoproteins. The web server GlycoPP employing these models can be accessed under Tools given at http://www.proglycprot.org/.