OGP: A Repository of Experimentally Characterized O-Glycoproteins to Facilitate Studies on O-Glycosylation

Jiang-Ming Huang[1,2,#,a], Meng-Xi Wu[1,2,#,b], Yang Zhang[1,2,c], Si-Yuan Kong[1,d], Ming-Qi Liu[1,e], Bi-Yun Jiang[1,2,f], Peng-Yuan Yang[1,2,3,*,g], Wei-Qian Cao[1,2,3,*,h]

[1] *Department of Chemistry and Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China*

[2] *The Fifth People's Hospital of Shanghai, Fudan University, Shanghai 200240, China*

[3] *NHC Key Laboratory of Glycoconjugates Research (Fudan University), Shanghai, 200032, China*

[#] Equal contribution.

[*] Corresponding author(s).

E-mail: pyyang@fudan.edu.cn (Yang P), wqcao@fudan.edu.cn (Cao W)

[§]Current address:

Department of Chemistry and Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China

Running title: *Huang J et al / Association Repository of O-glycoproteins*

[a] ORCID 0000-0002-5316-4266

[b] ORCID 0000-0002-4780-8109

[c] ORCID 0000-0002-8856-266X

[d] ORCID 0000-0002-2829-2557

[e] ORCID 0000-0002-7971-9418

[f] ORCID 0000-0002-8579-0229

[g] ORCID 0000-0001-5779-1008

[h] ORCID 0000-0002-1007-4607

3125 words, 4 figures, 3 supplementary figures, and 3 supplementary notes; 21 References, 8 References in supplementary information and 12 References from 2014

Abstract

Numerous studies on cancer, biopharmaceuticals, and clinical trials have necessitated comprehensive and precise analysis of protein O-glycosylation. However, the lack of updated and convenient databases deter the storage of and reference to emerging O-glycoprotein data. To resolve this issue, an O-glycoprotein repository named OGP was constructed on the basis of a collection of O-glycoprotein data from different sources. OGP, the largest O-glycoprotein repository thus far, includes 9354 O-glycosylation sites and 11633 site-specific O-glycans mapping to 2133 O-glycoproteins. Based on data regarding sites, an O-glycosylation site prediction tool was developed. Moreover, an OGP-based website is already available (*http://www.oglyp.org/*). This website comprises four functional, specially designed and user-friendly modules: Statistic Analysis, Database Search, Site Prediction and Data Submit. The first version of OGP and the OGP-based website allows users to obtain various data regarding O-glycoproteins, such as protein accession numbers, sequence, site-specific glycan structures, experimental methods, and potential glycosylation sites. Information regarding O-glycosylation mining can be obtained efficiently on this website efficiently, thus greatly facilitating studies on O-glycosylation.

KEYWORDS: O-glycoproteins; Association repository; Site prediction;

Introduction

Comprehensive and precise analysis of O-glycoproteins would potentially further the current understanding of the role of O-glycosylation in many physiological and pathological phenomena, such as intercellular communication [1], hereditary disorders, immune deficiencies, and cancer [2−4]. Numerous studies have attempted to understand the complexity of O-glycosylation. Recent technological advancements in many fields, especially mass spectrometry, have yield increasing data on O-glycoproteins [5−7]. However, the lack of up-to-date and comprehensive databases largely deters referring to, archiving, and utilizing emerging data on O-glycoproteins.

Studies have attempted to develop glycosylation-related databases [8−21]. However, most of these databases are focused on N-glycoproteins. Only a few databases contain data on O-glycoproteins. Furthermore, the most extensively used repository, UniCarbKB [9], which primarily provides data regarding N-glycoproteins, provides limited data in O-glycoproteins without data on site-specific glycans. dbPTM [11, 12] is an integrated resource for post-translational modifications (PTMs) containing over 130 types of PTMs; however, it does not provide information regarding site-specific O-glycosylation. O-GLYCBASE [8] provides information regarding both glycans and glycosylation sites and is the most widely used database in studies on O-glycosylation; however, it has not been updated since 2002. Moreover, it contains merely 189 O-glycoproteins and 2142 O-glycosylation sites, lagging in data regarding the current O-glycoproteome. Current O-glycoprotein databases are less satisfactory with notable issues including insufficient records, unknown reliability, a non-user unfriendly interface, and outdated or unrelated data.

The dearth of O-glycoprotein databases has greatly impeded studies on O-glycosylation and development. Recent large-scale analysis of O-glycosylation sites and intact O-glycopeptides have gradually become possible. For example, Steentofh et al. developed a glycoengineering-based method called “SimpleCell” for large-scale identification of O-glycosylated sites [5]. Yang et al developed a method called “EXoO” for large-scale analysis of intact O-glycopeptides [7]. However, functional analyses of O-glycoproteins are yet limited. In addition to the complexity of O-glycosylation, another primary factor limiting studies on O-glycosylation is the

difficulty in filtering large data to select candidate O-glycoproteins. Thus, an updated O-glycosylation database providing extensive information regarding protein O-glycosylation status, site-specific O-glycans, analytical methods, and other related information is required and would accelerate studies on O-glycosylation.

In this study, an O-glycoprotein repository named OGP was constructed, containing 9354 O-glycosylation sites and 11633 site-specific O-glycans mapping to 2133 O-glycoproteins. To our knowledge, OGP is the most comprehensive experimentally characterized O-glycoprotein repository thus far. An O-glycosylation site prediction tool was also developed on the basis of the recorded site data. An OGP-based website is already available (*http://www.oglyp.org/*) to facilitate database access, containing four modules: Statistic Analysis, Database Search, Site Prediction, and Data Submit. All the aforementioned O-glycoprotein data can be easily retrieved from the website. Such a comprehensive, reliable, user-friendly, and open-access O-glycoprotein repository would greatly benefit studies on O-glycosylation, including the development of O-glycoprotein-targeting drugs, and related basic and clinical studies.

Construction of the OGP repository

The OGP knowledgebase was constructed by integrating experimentally verified O-glycoproteins between 1998 and 2018 and other existing O-glycoprotein databases [8] (Fig. 1A). All proteins were manually curated, aligned with UniProt entries, and merged. Detailed extraction methods from the literature are described in Supplementary Note I. In total, 9354 glycosylation sites and 11633 site-specific glycans mapping to 2133 glycoproteins of different species have been recorded in the database (Fig. 1B, C). The scale of the OGP repository is more than 20-fold that of the primary existing database O-GlycBase 6.0, which is the largest O-glycoprotein repository thus far (Fig. 1D, E). This database is intended to be further updated to maintain its scale and content by expanding it with more published data.

The database includes data on proteins, peptide sequences, glycosylation sites, and site-specific glycans. For each site and site-specific glycan, detailed analytical strategies, such as sample sources, digestion enzymes, enrichment methods, analytical

methods, and related studies are integrated. All glycoproteins recorded in the database have been aligned with their UniProt entries. Thus, additional data including protein sequence annotation, subcellular locations, and other post-translational modifications can be conveniently obtained. To better obtain topological information regarding glycans, a linear coding method (Supplementary Note II) has been used in this database to record site-specific glycan structures. Furthermore, experimental strategies for every glycopeptide, such as immunoprecipitation, gel filtration, and mass spectrometry, were manually extracted, verified, and recorded in the database. These data are easily retrievable from the OGP website constructed on the basis of the database.

Development of an O-glycosylation site prediction model

Since O-glycosylation is highly complex but important, it is important to better understand its modification pattern. As a meaningful trial and for better utilization of OGP, an O-glycosylation site prediction model was developed using meticulously selected O-glycosylation sites with high confidence, recorded in OGP. This prediction model was generated through three primary steps (Fig. 2A, Supplementary Note III): 1) construction of a dedicated training set; 2) optimization of parameters; 3) evaluation of site prediction performance. Through systematic optimization, a dedicated training dataset was established with a 1:1 positive: negative instances ratio for random sampling (1754 positive site-central sequences and 1754 negative site sequences) (Fig. 2B, Supplementary Note III-1.1). Sequences with 11 amino residues were considered preferable (Fig. 2C, Supplementary Note III-1.2). Thereafter, we compared the performance of different algorithms to predict O-glycosylation sites using Weka 3.8 as the data mining tool, revealing that the Random Forest algorithm displayed the best performance (Fig. 2D, E, Supplementary Note III-1.3). Finally, the Random Forest algorithm was optimized to construct the prediction model. Ten-fold cross- validation indicated that the prediction model has high accuracy and sensitivity (ROC value=0.982, Recall value=0.902) (Fig. 2D, E, Supplementary Note III-1.3).

Construction of the OGP-based website

Based on the present OGP knowledge base, a dedicated website was constructed using hypertext markup language (HTML), cascading style sheet (CSS), and hypertext preprocessor (PHP); its design is shown in Figure 3A. The website generally contains three repositories in the underlying database layer: OGP, Prediction model, and Data submission. OGP is the core database that stores O-glycosylated protein sequences, sites, site-specific glycans, corresponding experimental data, and bibliography. The prediction model was optimized and contained an inherent training set. Data submission preserves user-uploaded information. By performing a set of actions including protein query and prediction model training and application in the operation layer, the website outputs four modules: Statistic analysis, Database search, Site prediction, and Data submit. The website is supported by common web browsers such as Internet Explorer, Mozilla Firefox, Google Chrome, Safari, and Opera.

Utility and the interface of the OGP website

The OGP-based website, equipped with a user-friendly graphical interface, is already available (*http://www.oglyp.org/*) and comprises four main modules: Statistic analysis, Database search, Site Prediction, and Data Submit. Furthermore, other functions including studies focused on O-glycosylation and links to databases (UniProt, UniCarbKB, and O-GLCBASE) are provided herein and have been updated to provide more information regarding O-glycosylation. The homepage of this website is shown in Figure 3B. Furthermore, the website provides detailed instructions and frequently asked questions to facilitate its convenient.

The Statistic Analysis module provides a general overview of the OGP database status, including the scale of total proteins, glycosylation sites, and site-specific glycans (Supplementary figure 1A), frequency distribution of O-glycoproteins in major taxonomic categories (Supplementary figure 1B), database scale comparison between the OGP and the other primary databases (Supplementary figure 1C), data regarding glycoprotein-involved pathways, molecular and cellular functions of proteins, and related diseases determined through Ingenuity Pathway Analysis (Supplementary figure 1D, E, F). Furthermore, this module provides more essential information. For example, more than 95% of the reported O-glycoproteins are present in mammalians, 55% of which are present in *Homo sapiens*, indicating that

O-glycosylation in other species warrants further analysis. All information would be updated in real time with the expansion of the OGP database.

In the Database Search module, an O-glycosylation keyword search can be carried out using the protein name and UniProt accession number (Supplementary figure 2). Figure 4 shows a webpage returning results from a query for fibrinogen gamma chain as an example. These results comprise well-structured data on protein O-glycosylation including basic information (*i.e.*, protein name, UniProt accession number, and species, [Figure 4A]), data on global mapping of all recorded O-glycosylation sites of the queried O-glycosylated protein (Figure 4B), experimentally verified O-glycopeptides and site-specific glycans (Figure 4C), and corresponding experimental methods, identifiers and source references (Figure 4D, E).

The site prediction model developed herein has been incorporated in the OGP website as one of the core modules to predict O-glycosylation sites in proteins. As is shown in Supplementary figure 3A, users can either complete the template file with site-centered sequences as instructed or simply upload a typical protein FASTA-type file and click on "predict" (Supplementary figure 3A). The prediction results for each site can be then displayed directly on the right side of the web page. Prediction scores range between 0 and 1; scores higher than 0.5 indicate positive sites, while those less than or equal to 0.5 indicate a highly probably non-O-glycosylated site. The higher the score, the greater the probability of a site being O-glycosylated and *vice versa.* The results can also be downloaded as shown in Supplementary figure 3B.

The Data Submit module enables users to upload new data into the OGP database and insert comments. All the new submitted data and feedback are carefully recorded in a backend database and revised manually by experts at regular intervals. Both a template form and an online form are accepted during a submission.

Conclusion and Discussion

The OGP repository, containing 9354 O-glycosylation sites and 11633 site-specific O-glycans mapping to 2133 O-glycoproteins, is the most comprehensive, experimentally verified O-glycoprotein repository thus far. All data contained in the OGP repository have been manually curated, and protein sequences have been aligned

with UniProt entries and merged. Based on recorded site data, an O-glycosylation site prediction tool has been developed to facilitate the prediction of O-glycosylation sites. The OGP-based website is already available (*http://www.oglyp.org/*) and contains four specially designed, user-friendly, functional modules: Statistic Analysis, Database Search, Site Prediction, and Data Submit. The initial version of the OGP repository and OGP-based website provides various data on O-glycoproteins, such as protein access number, sequence, site-specific glycan structures, experimental methods, and potential glycosylation sites. Mining of data on O-glycosylation can be carried out efficiently using this website. The OGP repository would greatly facilitate studies on O-glycosylation. The scale and content of this database is intended to be continuously expanded in subsequent versions of the OGP repository.

Authors' contributions

HJ developed the site prediction tool, participated in the database construction and helped to draft the manuscript. WM performed the database uniform, participated in the database and website construction and helped to draft the manuscript. ZY participated in website construction. KS participated in data collection. LM helped to revise the manuscript. JB participated in data collection. YP conceived the study, helped to draft the manuscript. CW designed the database, conceived the study, and draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgements

This work was supported by grants from the National Key Research and Development Program (2018YFC0910300, 2016YFA0501303 and 2016YFB0201702), the National Natural Science Foundation of China Project (91853102, 31600665 and 31570825). We would like to thank Mr. Jian-Qiang Wu to help with website construction, and Elsevier Premium Language Editing Services to help with the language editing.

References

[1] Hudson HF, Gerald WH, Ronald LS. Glycosylation precursors. In: Varki A, Richard DC, et al. editors. Essentials of Glycobiology, 3rd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA; 2017, p. 293– 304.

[2] Boskovski MT, Yuan S, Pedersen NB, Goth CK, Makova S, Clausen H, et al. The heterotaxy gene GALNT11 glycosylates Notch to orchestrate cilia type and laterality. Nature 2013; 504: 456–9.

[3] Wolfert MA, Boons GJ. Adaptive immune activation: glycosylation does matter. Nat Chem Biol 2013; 9: 776–84.

[4] Campos D, Freitas D, Gomes J, Magalhães A, Steentoft C, Gomes C, et al. Probing the O-glycoproteome of gastric cancer cell lines for biomarker discovery. Mol Cell Proteomics 2015; 14: 1616–29.

[5] Steentoft C, Vakhrushev SY, Vester-Christensen MB, Schjoldager KTBG, Kong Y, Bennett EP, et al. Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered SimpleCell lines. Nature Methods 2011; 8: 977–82.

[6] Ali L, Flowers SA, Jin C, Bennet EP, Ekwall AKH, Karlsson NG. The O-glycomap of lubricin, a novel mucin responsible for joint lubrication, identified by site-specific glycopeptide analysis. Mol Cell Proteomics 2014; 13: 3396–409.

[7] Yang W, Ao M, Hu Y, Li QK, Zhang H. Mapping the O-glycoproteome using site-specific extraction of O-linked glycopeptides (EXoO). Mol Syst Biol 2018; 14: e8486.

[8] Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. Nucleic Acids Res 1999; 27: 370–2.

[9] Campbell MP, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita KF, et al. UniCarbKB: building a knowledge platform for glycoproteomics. Nucleic Acids Res 2014; 42: D215–21.

[10] Malik A, Firoz A, Jha V, Ahmad S. PROCARB: A Database of Known and Modelled Carbohydrate-Binding Protein Structures with Sequence-Based Prediction Tools. Adv Bioinformatics 2010; 2010: 436036–9.

[11] Huang KY, Su MG, Kao HJ, Hsieh YC, Jhong JH, Cheng KH, et al. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. Nucleic Acids Res 2016; 44: D435–46.

[12] Huang KY, Lee TY, Kao HJ, Ma CT, Lee CC, Lin TH, et al. dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. Nucleic Acids Res 2019; 47: D298–D308.

[13] Bhat AH, Mondal H, Chauhan JS, Raghava GPS, Methi A, Rao A. ProGlycProt: a repository of experimentally characterized prokaryotic glycoproteins. Nucleic Acids Res 2012; 40: D388–93.

[14] Cooper CA, Harrison MJ, Wilkins MR, Packer NH. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. Nucleic Acids Res 2001; 29: 332–5.

[15] Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. Nucleic Acids Res 2003; 31: 511–3.

[16] Kao HJ, Huang CH, Bretaña NA, Lu CT, Huang KY, Weng SL, et al. A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. BMC Bioinformatics 2015; 16: S10.

[17] Kawasaki T, Nakao H, Takahashi E, Tominaga T. GlycoEpitope: the integrated database of carbohydrate antigens and antibodies. Trends Glycosci Glycotechnol 2006; 18: 267–72

[18] Baycin-Hizal D, Tian Y, Akan I, Jacobson E, Clark D, Chu J, et al. GlycoFly: a database of Drosophila N-linked glycoproteins identified using SPEG--MS techniques. J Proteome Res 2011; 10: 2777–84.

[19] Zhang H, Loriaux P, Eng J, Campbell D, Keller A, Moss P, et al. UniPep--a database for human N-linked glycosites: a resource for biomarker discovery. Genome Biol 2006; 7: R73.

[20] Kaji H, Saito H, Yamauchi Y, Shinkawa T, Taoka M, Hirabayashi J, et al. Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. Nat Biotechnol 2003; 21: 667–72.

[21] Kaji H, Yamauchi Y, Takahashi N, Isobe T. Mass spectrometric identification of N-linked glycopeptides using lectin-mediated affinity capture and glycosylation site-specific stable isotope tagging. Nature Protocols 2006; 1: 3019–27.

Figure

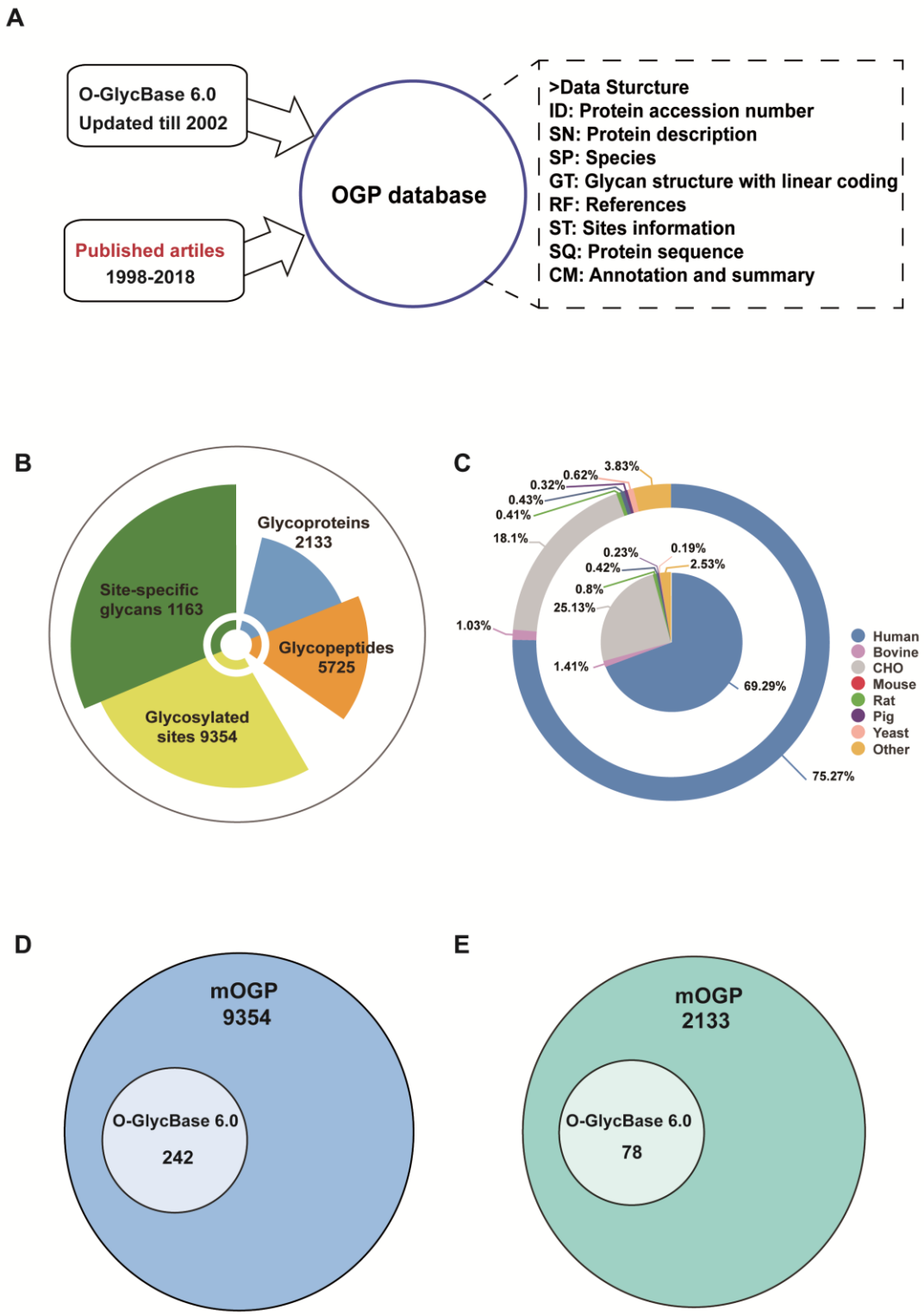

Figure 1 Overview of OGP repository.

A. OGP Data collection. **B.** The scale of the OGP repository. **C.** Species distribution of O-glycoproteins in OGP. **D.** Comparison of OGP with O-GlycBase 6.0 in

glycosylation site level. **E.** Comparison of OGP with O-GlycBase 6.0 in glycoprotein level.

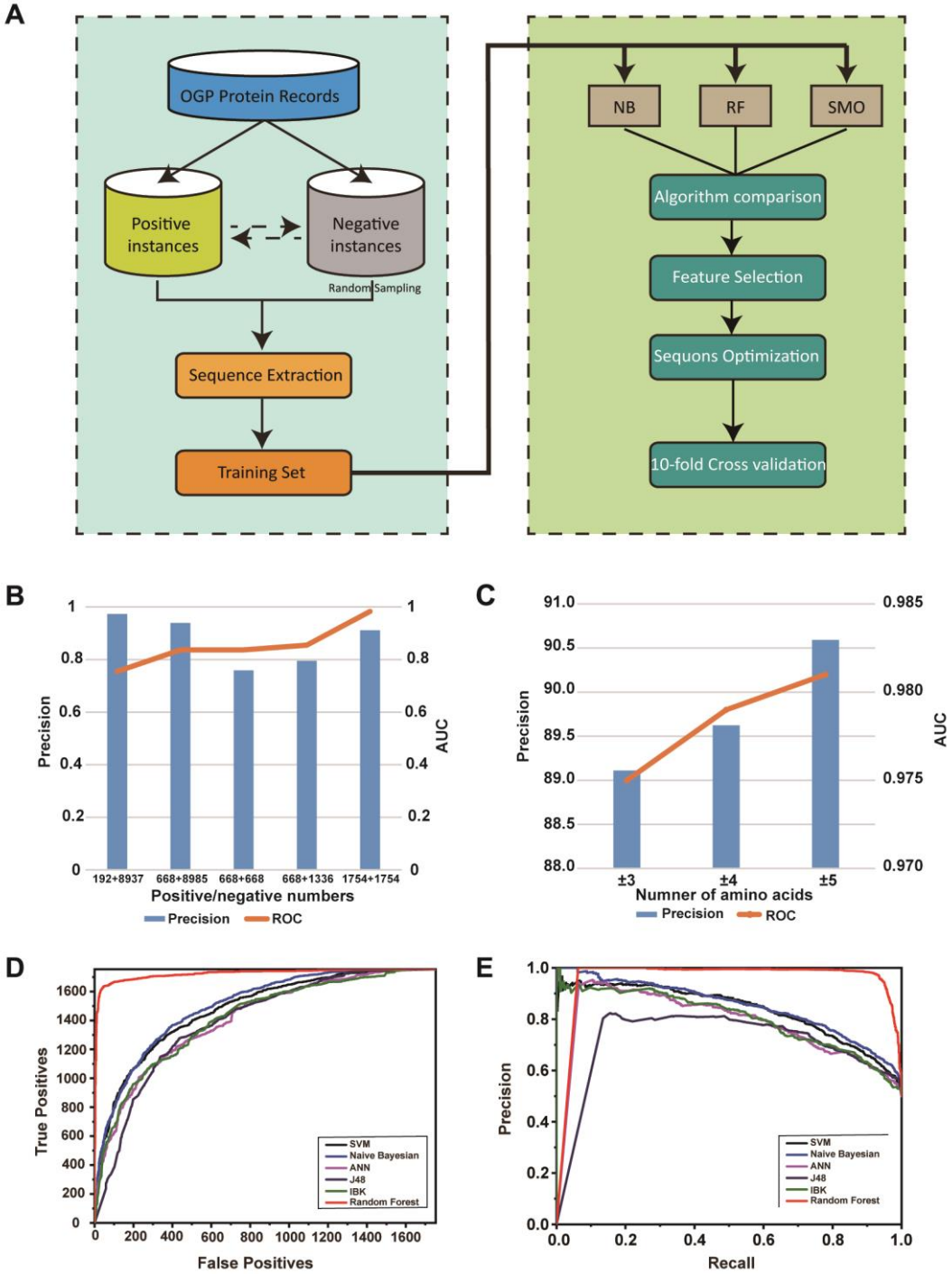

Figure 2 Development of O-glycosylation site prediction model

A. Workflow for building OGP based O-glycosite prediction model. **B.** Effect of positives and negatives scales and ratios on model prediction performance (AUC

represented for area under curve). **C.** Influence of amino acid residuals on the performance of site prediction model. **D.** ROC curves of each classification algorithm. **E.** PR curves of each classification algorithm.

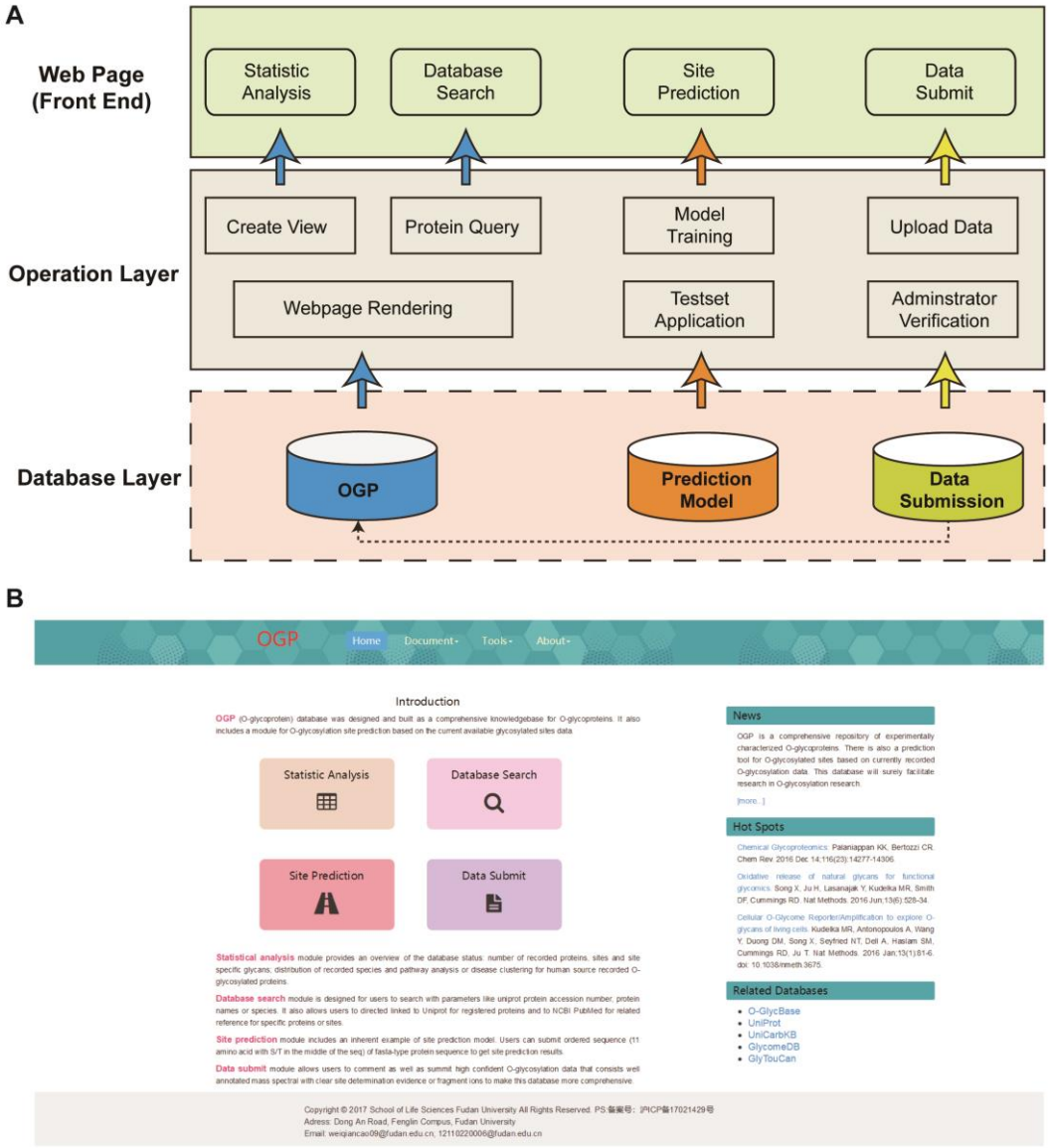

Figure 3 Construction of OGP-based website

A. The website structural design. **B.** Homepage of the website.

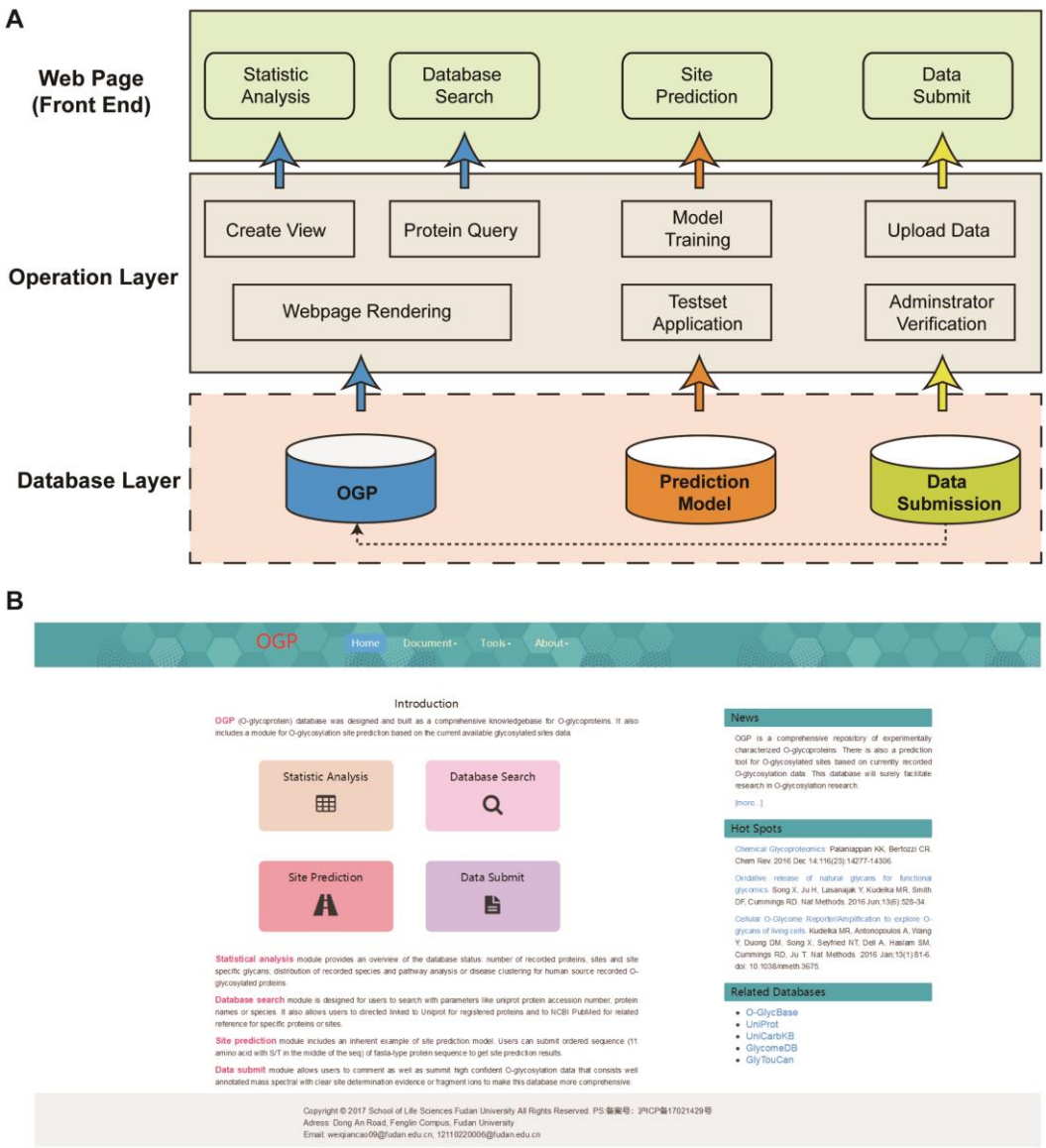

Figure 4 A webpage returning results from a query for Fibrinogen gamma chain

A. O-glycoprotein basic information. **B.** Protein Sequences and all recorded O-glycosylation sites with pink highlight. **C**. Experimentally verified O-glycopeptides and site-specific glycans. **D.** Corresponding experimental methods. D. Related source references.