CIRFESS: An interactive resource for querying the set of theoretically detectable peptides for cell surface and extracellular enrichment proteomic studies

Matthew Waas (https://orcid.org/0000-0003-4537-1502)[1], Jack Littrell (https://orcid.org/0000-0003-1264-894X)[1], Rebekah L. Gundry (https://orcid.org/0000-0002-9263-833X)[1]

[1]CardiOmics Program, Center for Heart and Vascular Research; Division of Cardiovascular Medicine; and Department of Cellular and Integrative Physiology, University of Nebraska Medical Center, Omaha, NE, 68198, USA

Running Title: CIRFESS informs cell surface proteomics

Address reprint requests to: Rebekah L. Gundry, PhD, University of Nebraska Medical Center, Department of Cellular and Integrative Physiology, 985850 Nebraska Medical Center, Omaha, NE, 68198-5850, USA, Telephone: 402-559-4426, Fax: 402-559-4438, Email: rebekah.gundry@unmc.edu

Abstract

Cell surface transmembrane, extracellular, and secreted proteins are high value targets for immunophenotyping, drug development, and studies related to intercellular communication in health and disease. As the number of specific and validated affinity reagents that target this subproteome are limited, mass spectrometry (MS)-based approaches will continue to play a critical role in enabling discovery and quantitation of these molecules. Given the technical considerations that make MS-based cell surface proteome studies uniquely challenging, it can be difficult to select an appropriate experimental approach. To this end, we have integrated multiple prediction strategies and annotations into a single online resource, Compiled Interactive Resource for Extracellular and Surface Studies (CIRFESS). CIRFESS enables rapid interrogation of the human proteome to reveal the cell surface proteome theoretically detectable by current approaches and highlights where current prediction strategies provide concordant and discordant information. We applied CIRFESS to identify the percentage of various subsets of the proteome which are expected to be captured by targeted enrichment strategies, including two established methods and one that is possible but not yet demonstrated. These results will inform the selection of available proteomic strategies and development of new strategies to enhance coverage of the cell surface and extracellular proteome. CIRFESS is available at www.cellsurfer.net/cirfess.

Introduction

The emergence of proteomics as a major discipline within the life science has been in no small part due to the development and eager adoption of bioinformatic strategies to enable the rapid analysis of mass spectrometry (MS) data files and inferred biological results. Since 1994, when the Yates laboratory introduced SEQUEST[1], the first computational tool for fully automated database searching, continued developments in database construction, algorithm design, and software development have propelled the evolution of MS-based proteomics[2–9]. All aspects of MS-based proteomics, including interpretation of raw spectra and database searching, visualization of results, and subsequent biological inferences benefit from advances in bioinformatics. Beyond the analysis of experimental data, bioinformatics tools that integrate machine-learning or ontological resources have become increasingly popular for prediction and classification of protein-level information[10]. Such approaches rely on experimental data to train or inform predictions and annotations, and in turn, the prediction strategies can benefit experimental design.

To scientists at the bench, perhaps the most exciting and impactful bioinformatic tools are those that can inform the next experiment. To this end, web-based formats have become increasingly popular resources as they often require less setup (*e.g.* installation), avoid operating system compatibility issues, and can be used in a familiar framework. Hundreds of web-based bioinformatics tools are now available for proteomics (*e.g.* www.expasy.org/proteomics). Current tools span a broad range of utility, including systems-level distribution of proteins based on experimental observations, visualization of experimental results, and prediction and cataloging of specific post-translational modifications, interactomes, and subcellular proteomes. Despite the increase in availability of web-based proteomics tools, there are currently relatively few resources designed to specifically assist in experimental design and analysis of the cell surface proteome.

The cell surface and extracellular space contain proteins which play key roles in a wide range of biological processes and can be utilized as valuable markers for immunophenotyping and drug targets. Despite their importance, the cell surface proteome remains relatively poorly characterized compared to the depth that most intracellular proteomes have been described. Given the relative low abundance, presence of hydrophobic transmembrane spanning regions, and dynamic nature of the cell surface proteome owing to continuous cycling of proteins due to internalization, secretion, and stimulus-triggered recruitment to the plasma membrane, specialized techniques are typically required to enhance the detection of cell surface proteins by MS. Such proteomic approaches include those that specifically label and enrich cell surface proteins using affinity-based approaches[11–13]. Application of these approaches have supported efforts to catalog the cell surface and secretome and have led to large scale efforts in experimentation[14] and collation[15]. As with most proteomic methods, currently available strategies to probe the cell surface and secreted proteome are biased towards proteins that contain specific features (*e.g.* presence of an *N*-glycosylation site or lysine within an extracellular region of the protein that will generate a detectable peptide after trypsin digestion). Also, the implementation of these approaches can be inconsistent among users, resulting in variability in the specificity (*i.e.* cell surface versus non-cell surface) of enrichment observed. Hence, the integration of bioinformatic predictions with experimental data provides an orthogonal means to interpret and filter results. Relevant predictions for surface and extracellular proteins include the presence of signal peptides[16–22] and transmembrane domains[16,17,23–25]. Other bioinformatic approaches have applied manual curation, ontological annotations, or machine learning approaches to predict the subset of proteins that are localized to the cell surface and extracellular regions[26–29]. However, not all cell surface proteins contain canonical signal peptides[30]. Also, GPI-anchored and extracellular matrix or secreted proteins do not contain transmembrane domains, and gene ontology annotations may be insufficiently specific *(e.g.* cell surface versus membrane). Thus,

filtering a proteomic dataset by these constraints often does not provide the complete picture of the cell surface proteome for a specific cell type.

Based on these limitations, in theory, it remains necessary to rely on experimental data to precisely define the proteins localized to the cell surface in a specific cell type. This leads to the question: *Which experimental approach is the best to use?* No doubt, the answer will be context dependent. If a specific monoclonal antibody is available, flow cytometry can be an effective approach for determining the surface localization of a protein. If antibodies are not available, the MS-based proteomic method of choice will depend on whether a cell surface proteome-wide screen or detection of a particular protein or subclass of proteins is desired. It will depend on the type and availability of the source material (*e.g.* metabolic labeling approaches cannot be used routinely for the analysis of primary human cells). Given the numerous technical considerations that make cell surface proteome studies uniquely challenging, it can be difficult to decide which approach to use. Currently, there is no single bioinformatic tool that can assist the investigator in determining which MS-based method is likely to be the most suitable approach for surface proteome studies.

To address this, we constructed a resource that integrates multiple prediction strategies and annotations relevant for the analysis of cell surface and extracellular proteins by MS and applied it to interrogate the human proteome. The results from these resources were compiled into a single interface and are accessible via a web-application termed Compiled Interactive Resource for Extracellular and Surface Studies (CIRFESS), accessible at www.cellsurfer.net/cirfess. By bringing together key resources used to interrogate the surface and extracellular space, CIRFESS helps to prevent duplication of efforts and continued pinging of separate prediction servers for the same set of proteins. We expect CIRFESS will be informative for a broad range of future applications and will inform the selection or development of proteomic strategies to enhance coverage of the cell surface and extracellular proteome.

Methods

Database and prediction server access

The human reference proteome was downloaded from UniProt (canonical only, 20416 entries, accessed Sept 13, 2019). The proteome was filtered and split to meet the requirements of the individual prediction servers (*e.g.* length of proteins, number of entries). Default scoring settings were applied, and the outputs were collected as specified: TMHMM– 'one line per protein', Phobius – 'Short', PrediSi, – 'Text', Signal P – 'Short output'.

Integrating prediction outputs

The outputs from the independent prediction servers were parsed and integrated using Python 3.7. The source code is made available as a Jupyter Notebook to enable implementation on batch predictions from TMHMM, Phobius, PrediSi and SignalP[16,22,25,31] for other species with minimal alteration (https://github.com/GundryLab/cirfess). A schematic of the inputs and the generated data structure is shown in Figure 1.

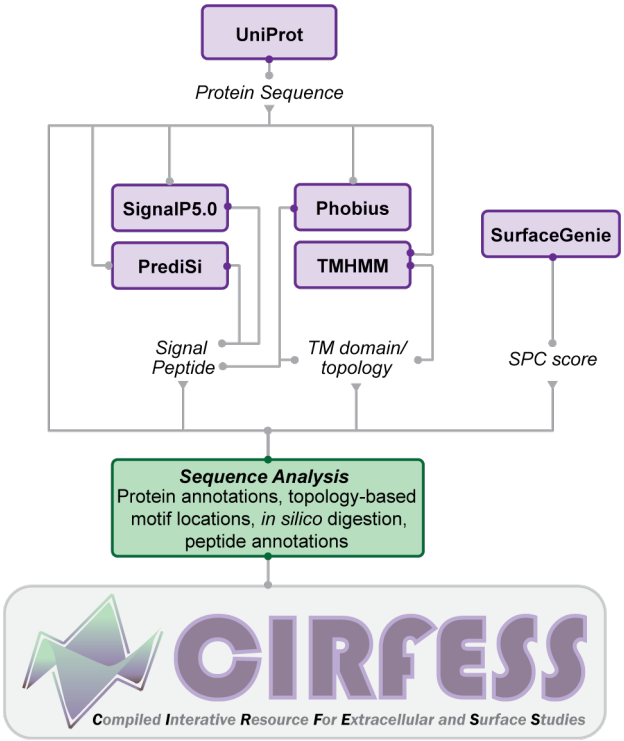

Figure 1. Schematic of the resources used in the construction of CIRFESS.

Evaluating peptides, motifs, and topology

Protein sequences were digested *in silico* to generate a list of potential peptides using the canonical tryptic cleavage site, X[R/K] where X is not P, allowing for up to two missed cleavages. This list of peptides is subsequently annotated with the following information: (1) presence of motifs for relevant proteomic capture strategies; N[!P][S/T/C/V] for *N*-glycan based capture, C for cysteine-based capture, K for lysine-based capture, (2) topological information – which residues and motifs are predicted to be intracellular and extracellular, and (3) suitability for a standard bottom-up proteomic experiment – length > 5, *m/z* of 2+ or 3+ charge state peptide < 2000.

CIRFESS Web application

A web application (CIRFESS) for accessing the data structure containing the parsed prediction outputs was developed in R using the Shiny package and is available at www.cellsurfer.net/cirfess. Source code is available at https://github.com/GundryLab/cirfess.

Results and Discussion

Proteome-wide comparison of prediction strategies

There are multiple sources of information to consider for evidence of surface or extracellular localization. Here we utilized transmembrane (TM) predictions[16,22], signal peptide (SP) predictions[17,25], and the Surface Prediction Consensus (SPC)[32] score. To highlight the complementarity of these measures and justification for inclusion in this resource, we performed set analysis of the human proteome using the UpSetR[33] web application. As shown in Figure 2A, the combination of these measures reveals distinct subsets of proteins. For instance, proteins which have a SP, but no TM domain (subsets b,f) are often considered to be the set of secreted proteins. For the set of proteins with predicted TM domains and SP, the SPC score is helpful for distinguishing between organelle membrane proteins (subset d) and cell surface proteins (subset

a). Finally, proteins with an SPC score but not TM or SP may contain GPI-anchored proteins or those without a canonical SP (subset g).

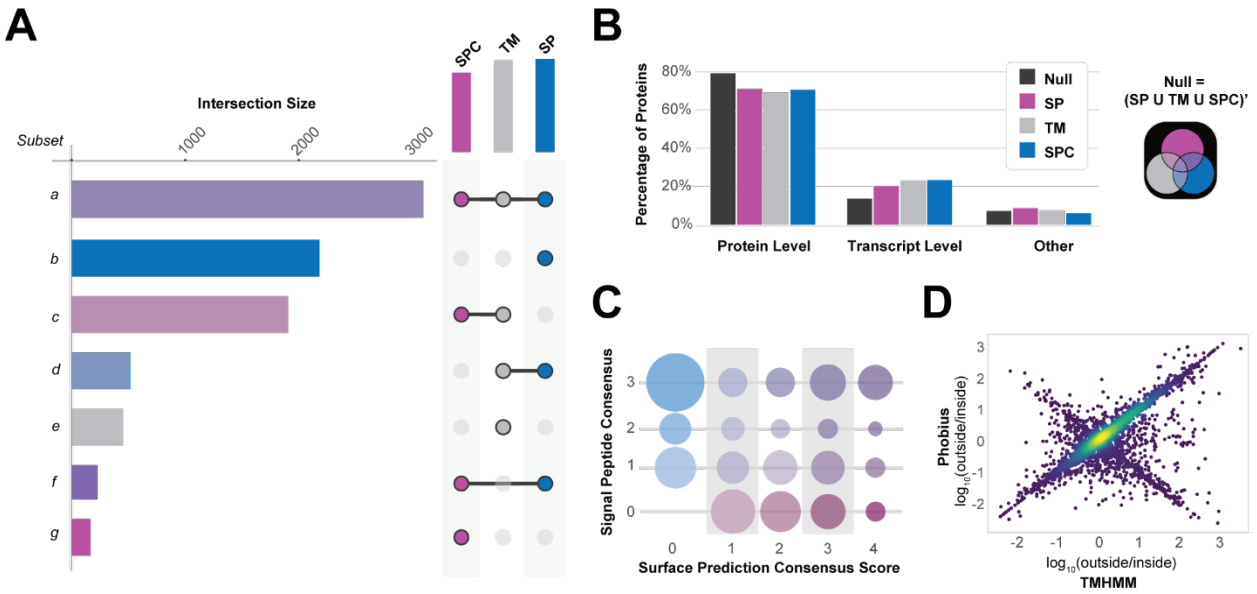

Figure 2. Contrasting views of the human proteome based on prediction strategies relevant for the cell surface proteome. (A) UpSet plot illustrating overlap in human proteins that are classified as containing a SP, TM domain, or SPC > 0. (B) Bar graph depicting the percentage of the human proteome with different levels of evidence in UniProt that are classified as containing a SP, TM domain, or SPC >0, or none of these features (Null). (C) Relationship between different levels of consensus for SP prediction and SPC score. (D) Plot depicting the $\log_{10}$ ratio of extracellular to intracellular residues predicted by TMHMM and Phobius highlighting that the opposite orientation is predicted for a subset of proteins.

To assess the level of experimental data that currently exists for cell surface proteins, we considered the “protein existence” annotation within UniProt. For these and further analyses, we compared proteins with positive SP predictions, positive TM predictions, or those with SPC scores > 0, with proteins that were negative for all three analysis – which we term the ‘Null set’ of proteins (shown visually in Figure 2B). Compared to the Null set, all three classes of proteins have a lower percentage of members with protein-level evidence, 79% for Null proteins compared to 71%, 69%, and 71% for SP, TM, and SPC proteins, respectively (Figure 2B). This analysis highlights the need for further experimental investigation of these subproteomes and the value of integrating these disparate bioinformatics predictions into a single analysis. For example, stratifying proteins with positive SP predictions (6026 proteins) by the number of algorithms for which it was positive reveals that the majority (3192, 53%) are predicted by all 3 algorithms. Plotting the number of positive SP predictions against SPC score reveals a positive relationship between SPC score and number of positive SP predictions for proteins with SPC score >0 (Figure 2C). Moreover, it also reveals that the majority of proteins with three positive SP predictions has an SPC score of 0 (1648 of 3192, 51.6%). This suggests that secreted proteins may have more recognizable signal peptide sequences than proteins translocated through the membrane.

Focusing on TM proteins, TMHMM and Phobius predict 5353 and 5471 proteins with TM domains, respectively, with 4846 proteins in common and 1132 proteins unique to a single prediction strategy (507 and 625 in TMHMM and Phobius, respectively). While overall there is strong consensus between the two algorithms for predicting which proteins contain TM domains, the number of TM domains predicted differs for 1306 out of the 4846 commonly predicted proteins. Furthermore, the opposite membrane orientation was predicted for a subset of proteins, visualized by plotting the $\log_{10}$ ratio of the predicted extracellular to intracellular residues (Figure 2D). Altogether, these analyses demonstrate the value of integrating data from multiple sources and

reveal that no single feature is sufficient to comprehensively predict the set of cell surface and extracellular proteins.

Motif coverage of extracellular and surface predicted proteins

As prediction strategies alone are insufficient to define the set of proteins localized to the cell surface and extracellular space, experimentation is required. To aid in the selection of proteomic strategies that are likely to produce the desired coverage of the cell surface proteome, the SP, TM, and SPC analyses described above were integrated with *in silico* analyses designed to predict which proteins would generate tryptic peptides likely to be detectable by electrospray MS, and of those, which are expected to be captured by application of commonly used biorthogonal enrichment strategies targeting *N*-glycans and lysines[11–13]. We also considered cysteines as they are enriched in surface proteins compared to nonsurface proteins[26] and numerous affinity reagents are available for targeting cysteines, although this is not yet a widely used approach for cell surface proteins. Important for the *N*-glycan approach, although strategies that specifically enrich peptides from the extracellular space (glycan biotinylation is performed on cells with intact plasma membranes) provide an additional level of experimental evidence for surface localization, it is possible to capture *N*-glycopeptides from whole cell lysate. In this case, the MS-based evidence for a glycan modifying an asparagine within the consensus motif for *N*-glycosylation is proposed to serve as standalone evidence for surface localization. Canonically, the consensus motif has been described as NXS/T where X is any amino acid except proline. However, more recently, evidence for *N*-glycosylation has been put forth at NXC[34,35] and NXV[36]. Here, we investigated the frequency of the various consensus motifs occurring in SP, TM, SPC and Null (meaning the protein contains no SP, TM, or SPC) sets of proteins. First, the probability of each motif occurring within the subset of proteins was calculated with respect to the amino acid frequencies. The expected frequencies based on amino acid compositions was consistent among the sets of proteins for each motif (0.26 ± 0.003 %, 0.18 ± 0.009 %, 0.09 ± 0.009 %, and 0.21 ±

0.019 % for NXS, NXT, NXC, and NXV respectively). Next, the observed frequency of each motif was calculated for each subset of proteins. The natural log of the odds ratio of observed to expected for each subset of proteins was calculated and plotted for each motif as well as the canonical (NXS/T) and complete consensus motifs (NXS/T/C/V) (Figure 3A). The results reveal that the NXS and NXT occur more frequently than expected and NXC and NXV occur less frequently than expected for SP, TM, and SPC proteins. Whereas NXS and NXT occur at about the expected rate for the Null set of proteins, NXC and NXV occur slightly above the expected frequency. While the complete consensus motif occurs more frequently than expected for SP, TM, and SPC proteins, the canonical motif demonstrates a much higher odds ratio, especially relative to the odds-ratio for the Null set. This analysis suggests that while the mere presence of the consensus motif provides some evidentiary weight to the localization of a surface protein – (1) it should not be considered conclusive, and (2) the canonical motif provides more meaningful information than the complete consensus motif. Based on these analyses, we elected to only consider the canonical motif as potential targets for *N*-glycan capture for subsequent analyses.

By integrating the topology information provided by the TM predictions with the locations of motifs within proteins, we estimated the coverage that each capture strategy would provide for each subset of the proteome (Null, SP, TM and SPC). For this analysis, proteins were categorized based on whether they contained the relevant motif and whether the motif was in a region determined to be extracellular by one or both TM prediction strategies. The percentage of proteins for which a predicted extracellular motif was located within an MS-suitable peptide was recorded (Figure 3B). This analysis revealed that while 72% of Null proteins contain a consensus motif for *N*-glycosylation, none of the glycopeptides are predicted to be in the extracellular domain. In contrast, of the SPC proteins which contain the consensus motif, 86% of those proteins contain at least one peptide contains the consensus motif within the predicted extracellular domain. These results were further summarized by calculating the percentage of each subproteome that is

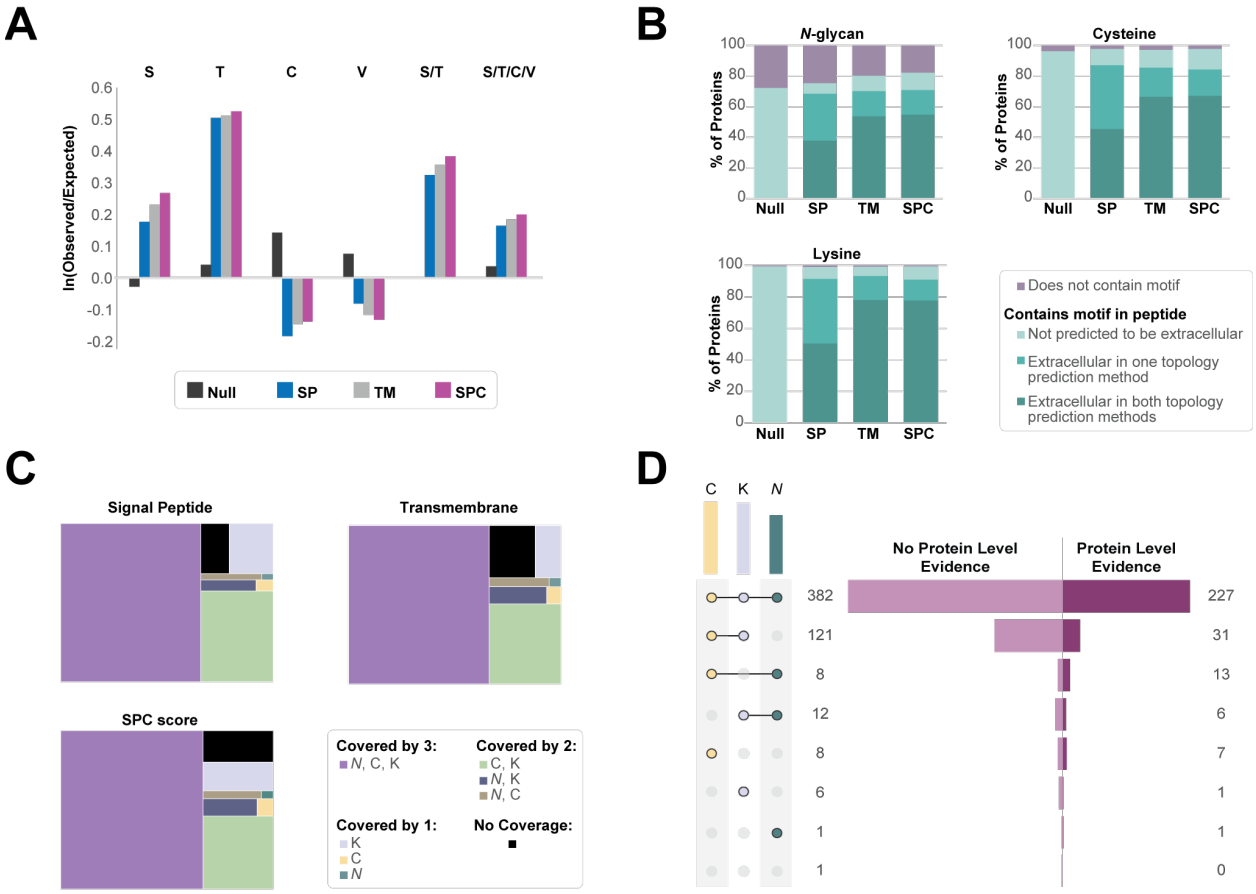

Figure 3. Results of CIRFESS analysis of the human proteome to assess predicted coverage provided by three common cell surface proteomic enrichment strategies. (A) The natural log of the odds ratio for observed-to-expected frequency of each permutation of the *N*-glycan consensus motif along with the canonical (S/T) and complete (S/T/C/V) consensus motif. (B) The expected coverage of the different subsets of proteins for each enrichment strategy broken down by which proteins have peptides with predicted extracellular motifs by one or both TM prediction methods. (C) The makeup of SP, TM, and SPC score proteins based on the overlapping coverage of the three individual enrichment strategies. (D) The set of human GPCRs based on expected coverage for enrichment strategy and level of evidence in UniProt.

predicted to be covered by each or multiple capture strategies (Figure 3C). Overall, querying the results from this analysis provides a strategy for investigators to rapidly interrogate the human proteome to determine which experimental strategy is most likely to be useful to address their biological question. In summary, 66.4 ± 0.4% of SP, TM and SPC proteins are likely to be captured by any of the three strategies, 17.3 ± 1.8% are detectable by cysteine or lysine capture, but not detectable by *N*-glycan strategies, 7.4 ± 1.4% are detectable by a single strategy, and 5.7 ± 1.2% are not detectable by any of the three strategies considered here. The identity of the proteins within each classification are provided in Supplemental Table 1A-C and these results provide actionable data related to high interest targets. For example, of the 825 human G-protein coupled receptors (GPCR), a striking 65.3% lack protein-level evidence within UniProt. Of these, all but one are predicted to be captured by at least one enrichment strategy and 70.9% of them are predicted to be captured by all three strategies. Supplementary Table 1D contains the identity of the GPCR proteins and which enrichment strategies are predicted to capture them.

Development of a web application

The web application CIRFESS was generated using R Shiny and can be accessed at www.cellsurfer.net/cirfess. All supporting resource files are available for download at https://github.com/GundryLab/cirfess. Within the CIRFESSS interface, users are able to perform single and batch querying of protein accession numbers to extract protein-level and peptide-level annotations as well as information about numbers of motifs and motif-containing peptides.

Critical Considerations

Results from the current implementation of CIRFESS are limited to human proteins digested with trypsin and the resulting peptides are detectable in the 2+ or 3+ charge state. These criteria were selected based on common implementation of bottom-up proteomic methods. However, all source files and code are publicly available in the Github repository and a user-specific version of

CIRFESS could be generated, requiring minimum alteration to change the *in silico* digestion strategies or criteria for MS-compatible peptide filtering. Implementation on other species would require submission of proteins to the individual prediction servers, but the source code includes scripts to parse and integrate the generated output files. Another critical assumption is related to the *N*-gycan capture strategy where detection depends on the glycosite being occupied by a glycan which is sensitive to the oxidation strategy applied (*e.g.* cis diols for meta-periodate[37]). Currently, as it is not possible to predict which sites will be occupied with specific glycan structures, the peptides predicted to be observable by this strategy should be considered with this caveat in mind. It is possible that post-translational modifications may interfere with the digestion, capture, ionization, and identification of peptides in any of the strategies, and therefore experimental observations may not be fully predictable by this bioinformatics approach. Moreover, for enrichment strategies which use cleavable linkers, residual portions of the linker that remain after cleavage will increase the mass of the resulting peptide. However, the 2000 *m/z* range used here for predicting detectable peptides should accommodate most commonly used reagents. Finally, it may be beneficial to combine the results from CIRFESS analysis with predictions for peptide detectability[38–40] or proteotypicity[41] to better inform the set of peptides which are most likely to be observable or informative.

Conclusion

CIRFESS is a web-based tool designed to accelerate cell surface proteome studies by eliminating the need to query each bioinformatics source separately and integrating disparate features into a single streamlined resource and output. Results may be queried for proteins or protein classes of interest to inform the design of the next experiment. We anticipate that CIRFESS will be broadly applicable for multiple applications across a broad range of biology and disease studies. Overall, these analyses suggest that acquiring protein-level evidence for the majority of predicted cell surface proteins is a matter of applying the right technology to a relevant biological system.

Overall, we expect CIRFESS will promote the rational selection of the most apt cell surface proteomic methods and will inspire new method development (*e.g.* cysteine-targeting) to enable detection of the human proteome not predicted to be accessible by established surface protein enrichment methods.

Author Contributions

R.L.G. and M.W. conceived the study; R.L.G. supervised the study; M.W. and J. L. developed the algorithms and developed the web application; M.W. and R.L.G. analyzed data; M.W. generated figures; M.W. and R.L.G. co-wrote the manuscript; All authors approved the final manuscript.

Acknowledgements

This work was supported by the National Institutes of Health [R01-HL126785 and R01-HL134010 to R.L.G.; F31-HL140914 to M.W.]; and JDRF [2-SRA-2019-829-S-B to R.L.G.]; Special thanks to Dr. Christopher Ashwood and Linda Berg Luecke for critical review of the manuscript and insightful discussions and Dr. David Tabb for helpful consultation. Funding sources were not involved in study design, data collection, interpretation, analysis or publication.

Supplementary Information

Supplemental Table 1. A. Coverage of SP proteins. B. Coverage of TM proteins. C. Coverage of SPC Proteins. D. Coverage and evidence of G-protein coupled receptors.

References

1. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994;5(11):976–89.

2. Eng JK, Jahan TA, Hoopmann MR. Comet: An open-source MS/MS sequence database search tool. *Proteomics*. 2013 Jan;13(1):22–4.

3. Gluck F, Hoogland C, Antinori P, Robin X, Nikitin F, Zufferey A, Pasquarello C, Fétaud V, Dayon L, Müller M, Lisacek F, Geiser L, Hochstrasser D, Sanchez JC, Scherl A. EasyProt - An easy-to-use graphical platform for proteomics data analysis. *J Proteomics*. 2013 Feb 1;79:146–60.

4. Wenger CD, Coon JJ. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *J Proteome Res*. 2013;12(3):1377–86.

5. Dorfer V, Pichler P, Stranzl T, Stadlmann J, Taus T, Winkler S, Mechtler K. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res*. 2014 Aug 1;13(8):3679–84.

6. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun*. 2014 Oct 31;5.

7. Risk BA, Spitzer WJ, Giddings MC. Peppy: Proteogenomic search software. *J Proteome Res*. 2013;12(6):3019–25.

8. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008 Dec;26(12):1367–72.

9. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification

by searching sequence databases using mass spectrometry data. In: Electrophoresis. Wiley-VCH Verlag; 1999. p. 3551–67.

10. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141).

11. Kalxdorf M, Gade S, Eberl HC, Bantscheff M. Monitoring Cell-surface N-Glycoproteome Dynamics by Quantitative Proteomics Reveals Mechanistic Insights into Macrophage Differentiation. *Mol Cell Proteomics*. 2017 May;16(5):770–85.

12. Turtoi A, Dumont B, Greffe Y, Blomme A, Mazzucchelli G, Delvenne P, Mutijima EN, Lifrange E, De Pauw E, Castronovo V. Novel Comprehensive Approach for Accessible Biomarker Identification and Absolute Quantification from Precious Human Tissues. *J Proteome Res*. 2011 Jul;10(7):3160–82.

13. Wollscheid B, Bausch-Fluck D, Henderson C, O'Brien R, Bibel M, Schiess R, Aebersold R, Watts JD. Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol*. 2009 Apr 6;27(4):378–86.

14. Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, et al. A Mass Spectrometric-Derived Cell Surface Protein Atlas. Wanjin H, editor. *PLoS One*. 2015 Apr 20;10(4):e0121314.

15. Uhlén M, Karlsson MJ, Hober A, Svensson A-S, Scheffel J, Kotol D, et al. The human secretome. *Sci Signal*. 2019 Nov 26;12(609).

16. Hiller K, Grote A, Scheer M, Münch R, Jahn D. PrediSi: Prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*. 2004 Jul 1;32(WEB SERVER ISS.).

17. Käll L, Krogh A, Sonnhammer EL. A Combined Transmembrane Topology and Signal

Peptide Prediction Method. *J Mol Biol*. 2004 May 14;338(5):1027–36.

18. Frank K, Sippl MJ. High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics*. 2008 Oct;24(19):2172–6.

19. Signal-3L 3.0: Improving signal peptide prediction through combining attention deep learning with domain rules [Internet].

20. Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. *PLoS Comput Biol*. 2008 Nov;4(11).

21. Savojardo C, Martelli PL, Fariselli P, Casadio R. DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics*. 2018 May 15;34(10):1690–6.

22. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 2019 Apr 1;37(4):420–3.

23. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*. 1998;14(4):378–9.

24. Bernhofer M, Kloppmann E, Reeb J, Rost B. TMSEG: Novel prediction of transmembrane helices. *Proteins Struct Funct Bioinforma*. 2016 Nov 1;84(11):1706–16.

25. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*. 1998;6:175–82.

26. Bausch-Fluck D, Goldmann U, Müller S, van Oostrum M, Müller M, Schubert OT, Wollscheid B. The in silico human surfaceome. *Proc Natl Acad Sci U S A*. 2018;115(46):E10988–97.

27. da Cunha JPC, Galante PAF, de Souza JE, de Souza RF, Carvalho PM, Ohara DT, et al. Bioinformatics construction of the human cell surfaceome. *Proc Natl Acad Sci U S A*. 2009 Sep 29;106(39):16752–7.

28. Town J, Pais H, Harrison S, Stead LF, Bataille C, Bunjobpol W, Zhang J, Rabbitts TH. Exploring the surfaceome of Ewing sarcoma identifies a new and unique therapeutic target. *Proc Natl Acad Sci U S A*. 2016 Mar 29;113(13):3603–8.

29. Díaz-Ramos MC, Engel P, Bastos R. Towards a comprehensive human cell-surface immunome database. *Immunol Lett*. 2011 Jan 30;134(2):183–7.

30. Kuchler K, Thorner J. Membrane translocation of proteins without hydrophobic signal peptides. *Curr Opin Cell Biol*. 1990;2(4):617–24.

31. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*. 2004 May 14;338(5):1027–36.

32. Waas M, Snarrenberg ST, Littrell J, Lipinski RAJ, Hansen PA, Corbett JA, Gundry RL. SurfaceGenie: A web-based application for prioritizing cell-type specific marker candidates. *https://doi.org/101101/575969*. 2020;Revis.

33. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph*. 2014 Dec 31;20(12):1983–92.

34. Wang J, Yang P, Tang B, Sun X, Zhang R, Guo C, Gong G, Liu Y, Li R, Zhang L, Dai Y, Li N. Expression and characterization of bioactive recombinant human α-lactalbumin in the milk of transgenic cloned cows. *J Dairy Sci*. 2008;91(12):4466–76.

35. Sato C, Kim J-H, Abe Y, Saito K, Yokoyama S, Kohda D. Characterization of the iV-Oligosaccharides Attached to the Atypical Asn-X-Cys Sequence of Recombinant Human Epidermal Growth Factor Receptor [Internet]. Vol. 127, J. Biochem. 2000.

36. Zielinska DF, Gnad F, Wiśniewski JR, Mann M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*. 2010 May;141(5):897–907.

37. BOBBITT JM. Periodate oxidation of carbohydrates. *Adv Carbohydr Chem*. 1956;48(11):1–41.

38. Muntel J, Boswell SA, Tang S, Ahmed S, Wapinski I, Foley G, Steen H, Springer M. Abundance-based classifier for the prediction of mass spectrometric peptide detectability upon enrichment (PPA). *Mol Cell Proteomics*. 2015 Feb;14(2):430–40.

39. Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotechnol*. 2009 Feb;27(2):190–8.

40. Eyers CE, Lawless C, Wedge DC, Lau KW, Gaskell SJ, Hubbard SJ. CONSeQuence: Prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol Cell Proteomics*. 2011 Nov;10(11).

41. Searle BC, Egertson JD, Bollinger JG, Stergachis AB, MacCoss MJ. Using data independent acquisition (DIA) to model high-responding peptides for targeted proteomics experiments. *Mol Cell Proteomics*. 2015 Sep 1;14(9):2331–40.