	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ PRODUCTON
Primary Sequence Analysis of PTMs **Module Hours: 3**	**Effective Date: 04/26/2021**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

BACKGROUND

The central dogma of biology, emphasizing the transition from DNA to RNA to protein, is fundamentally important when establishing an understanding of protein functionality within cells. However, the genome alone cannot account for the exponentially larger number of proteins that exist compared to the number of genes in the human body. (ThermoFisher Scientific) This observation has led to increased investigation into the importance of numerous posttranslational modifications that allow for the diversification of proteins.

There is an extensive number of posttranslational modifications that can be applied to a given protein, including covalent addition of functional groups and proteolytic cleavage. Many different programs have been developed to analyze primary protein sequences and anticipate where such post translational modifications may occur.

- **PROSITE:** PROSITE is a conglomerate database that contains information about protein families and domains. (G-Preciado et al., 2009) It functions by grouping together proteins with similar domains that indicate similar functions. This site can be used to identify potential post translational modification sites of known proteins or to potentially determine the function of an unknown protein based on similarities to certain families or domains.
- **ProP:** ProP is a server that predicts arginine and lysine propeptide cleavage sites in eukaryotic protein sequences. (Duckert et al., 2004) ProP utilizes artificial neural networks to predict the cleavage sites. (Duckert et al., 2004)
- **NetPhos:** NetPhos is a server that predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using artificial neural networks. (Blom et al., 1999).
- **NetOGlyc:** NetOGlyc is a server that predicts GalNAc O-glycosylation sites in mammalian proteins using artificial neural networks. (Steentoft et al, 2013)
- **NetNGlyc:** NetNGlyc is a server that predicts N-glycosylation sites in human proteins using artificial neural networks. (Gupta & Brunak, 2002)

1. PURPOSE

The purpose of this procedure is to understand the basics of post translational modification and to become comfortable utilizing various databases that identify post translational modification sites in the primary protein sequence.

2. SCOPE

This procedure applies to qualified skills center users.

3. RESPONSIBILITY
3.1. It is the responsibility of the user to understand and perform the procedure described in this document.
3.2. It is the responsibility of the user performing the procedure to fully document any deviations from the written procedure.
3.3. It is the responsibility of the user to become trained on the procedure.

4. DEFINITIONS
4.1. Glycosylation – The enzymatic, post-translational alteration of a protein involving the covalent addition of a sugar molecule.
4.2. Phosphorylation – The post-translational addition of a phosphate group, often by a kinase, to specific sites on a protein.
4.3. Proteolytic cleavage – The hydrolysis of the peptide bonds in a protein that causes the protein to be cut in a specific region. This mechanism is catalyzed by peptidase enzymes.
4.4. PROSITE profile – A characterization of protein domains over the entire length. A profile is often used to predict the structural properties of a protein and is more sensitive than a PROSITE pattern. (G-Preciado et al., 2009)
4.5. PROSITE pattern – A short sequence motif confined to a small region with high sequence similarity. A pattern is often used to predict a protein's function. (G-Preciado et al., 2009)
4.6. FASTA format – A format for representing either nucleotide or protein sequences in which the base pairs or amino acids are represented using the specified single letter codes. (Zhang)

5. MATERIALS/EQUIPMENT
5.1. PROSITE application: https://prosite.expasy.org/
5.2. ProP application: http://www.cbs.dtu.dk/services/ProP/
5.3. NetPhos application: http://www.cbs.dtu.dk/services/NetPhos/
5.4. NetOGlyc application: http://www.cbs.dtu.dk/services/NetOGlyc/
5.5. NetNGlyc application: http://www.cbs.dtu.dk/services/NetNGlyc/
5.6. NCBI Protein database: https://www.ncbi.nlm.nih.gov/protein/

	SKILLS CENTER **STANDARD OPERATING PROCEDURE**	**A BIOFIZZ** ![brain logo] **PRODUCTON**
Primary Sequence Analysis of PTMs **Module Hours: 3**	**Effective Date: 04/26/2021**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

6. PROCEDURE

6.1. PROSITE

6.1.1. Go to the PROSITE application website: https://prosite.expasy.org/

6.1.2. On the home screen, there are many different options for searching the database. (Figure 2)

- The Search option is beneficial when there are a few terms, identifiers, etc. that will be used to guide the search.
- The Browse option is beneficial when there is not as specific of a protein that you are looking for, enabling you to peruse existing entries.
- The ScanProsite mode is used when searching for information about a protein with a known FASTA sequence, PDB identifier, or UnitProKB accession numbers/identifiers. After pasting the desired input, clicking Scan will bring up a new window with the desired information.
- For this SOP, the ScanProsite option will be used. The input is the NCBI protein database FASTA sequence for the first hit when searching "human kinase".
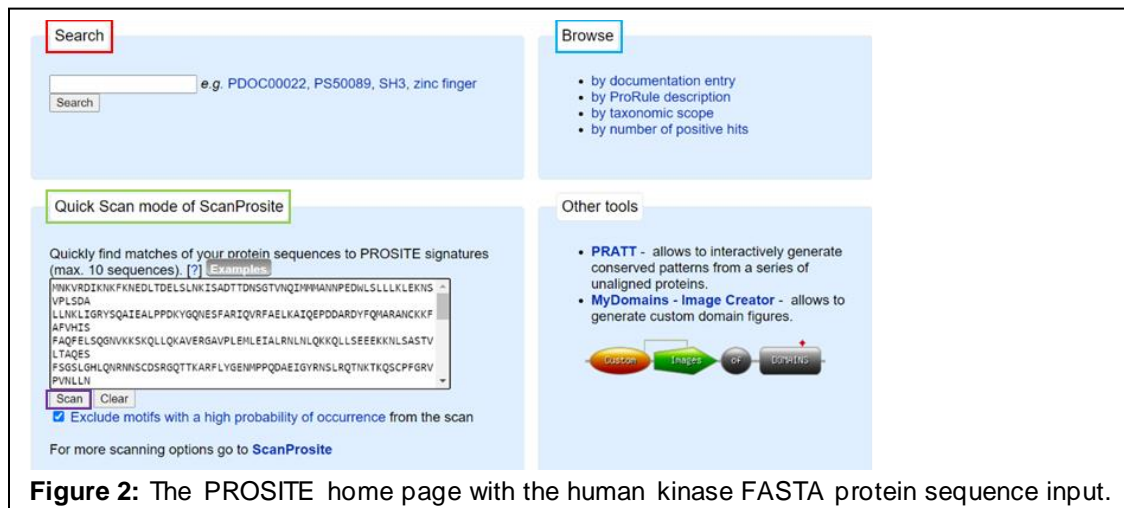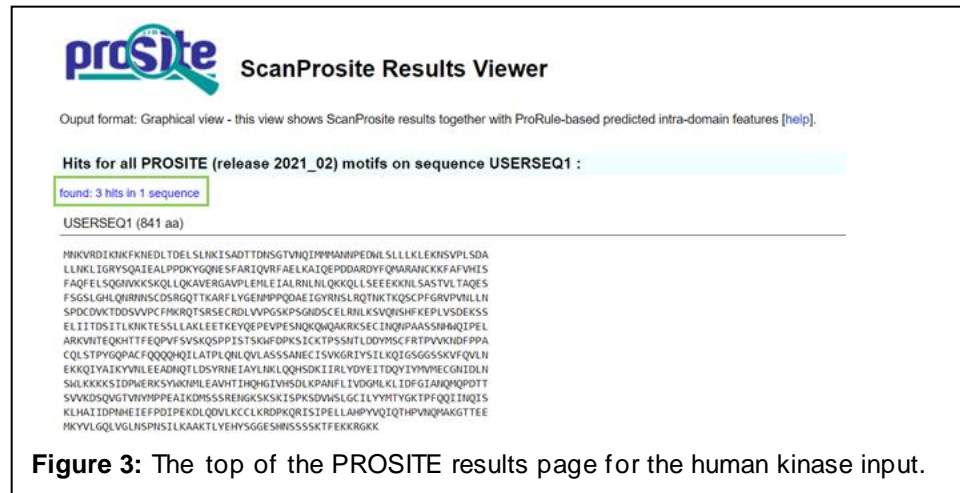

Figure 2: The PROSITE home page with the human kinase FASTA protein sequence input.

- Clicking the "exclude motifs with a high probability of occurrence" button allows for elimination of common repetitive motifs as hits that can throw off your search. This is most helpful with larger proteins.

6.1.3. The ScanPROSITE application will look for hits that are potential matches for the input. The resulting screen first delineates the amino acid sequence of

the given protein, along with the amount of hits found in X number of sequences. (Figure 3)

Figure 3: The top of the PROSITE results page for the human kinase input.

6.1.4. Scrolling down the results page will give a more detailed description of the hits found in the database.
   - A legend is given explaining result annotations. (Figure 4)
   - Hit options will be listed by both profile and pattern correlations (see definitions section for more details on the differentiation between these 2 categories).

Figure 4: The PROSITE legend and the first profile hit for the human kinase input.
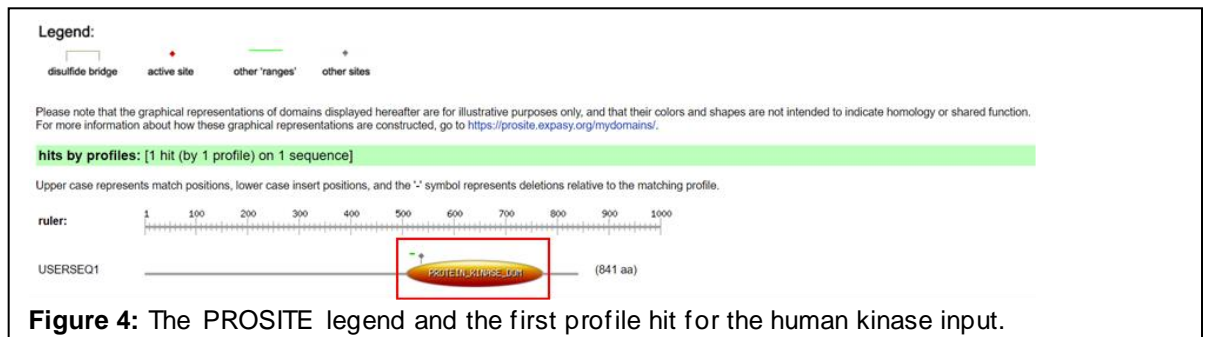
   - Rolling over the protein kinase domain annotation (Figure 4) in the given example will highlight the region in the amino acid sequence given in Figure 3.
   - Clicking on the protein kinase annotation will bring up a new window giving a detailed, generalized description about protein kinase signatures

and profiles. This includes descriptions about the protein family, conserved regions and specific patterns and profiles.

   6.1.5. Further down the results page there will be information about the resulting domain(s). (Figure 5)
- There is a description of the amino acid range spanned by the given domain.
- The score value gives an estimate of how well the given profile matches the profile for the domain in question. (Michael, 2016) A score equal to or above the cut off value indicates a motif occurrence. (Sigrist, 2002)
- The predicted features section describes features with fulfilled conditions. (de Castro et al., 2006)
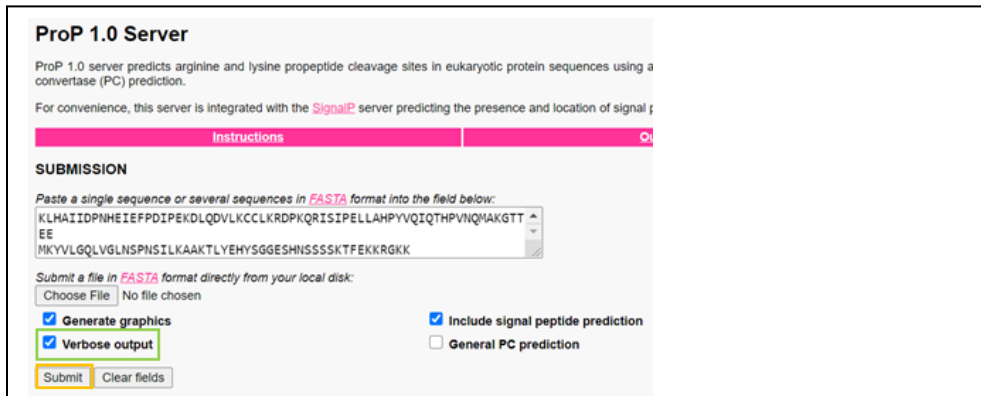- The absent features section describes features with unfulfilled conditions. (de Castro et al., 2006)

Figure 5: Detailed description of the protein kinase domain.
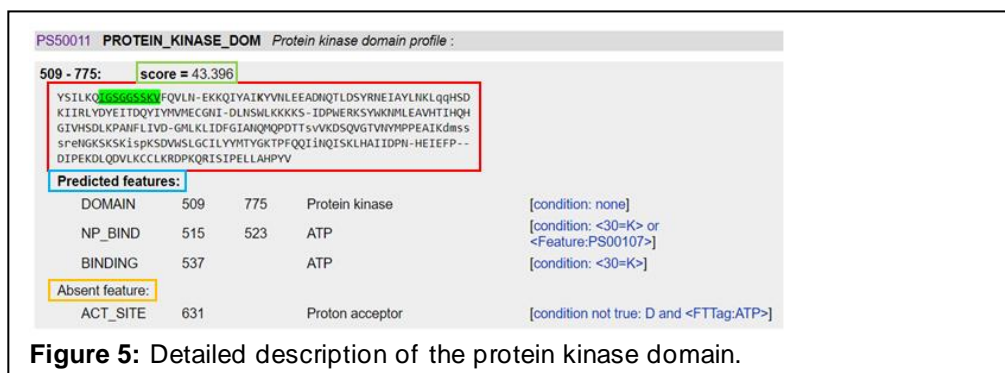
6.2. ProP
   6.2.1. Go to the ProP website: http://www.cbs.dtu.dk/services/ProP/
   6.2.2. Input the sequence of the protein of interest in the FASTA format, which can be found on the NCBI protein database: https://www.ncbi.nlm.nih.gov/protein/. (Figure 6)
- Checking the Verbose Output box will show each individual score for all 4 of the artificial neural networks used for computation in addition to the average score. Unchecking the box will just show the average score. (Blom, 2017) (Figure 6)

6.2.3. Click the Submit button to run the query. (Figure 6)
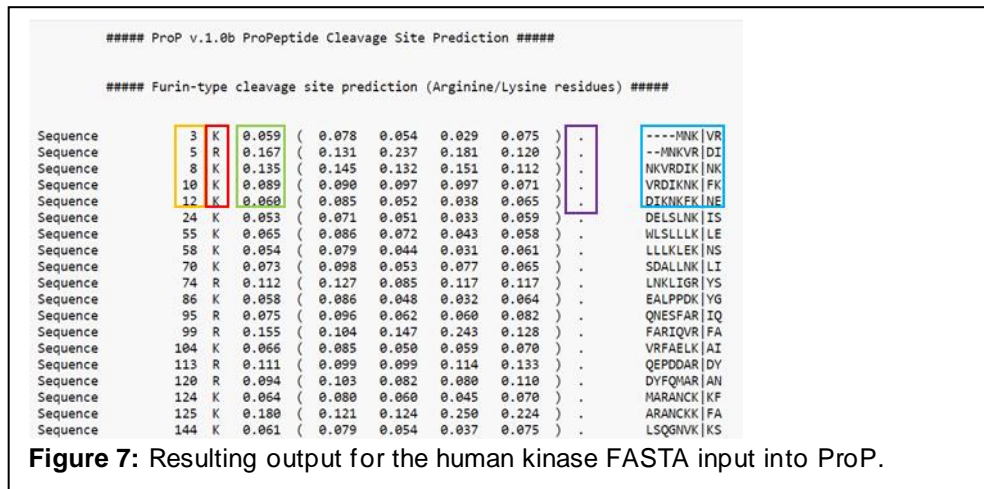

Figure 6: Home page of the ProP server with the human kinase FASTA input.

6.2.4. A new window will pop up with the results for the search, with each column giving different important information. (Figure 7)

- Position = shows what position in the propeptide that the R or K residue falls.
- Residue = shows whether the residue is an R or K.
- Score = a value between 0.000 to 1.000. The closer the score is to 1.000, the more likely the residue will be followed by a cleavage site. 0.500 is the threshold value; a number above threshold indicates that the R or K is likely followed by a cleavage site. (Blom, 2017)
- Answer = If the prediction of the site meets or exceeds the threshold score, then the output in this column will be "ProP"
- Context = shows the R or K in question and the predicted cleavage site, along with the surrounding amino acids.

	SKILLS CENTER **STANDARD OPERATING PROCEDURE**	**A BIOFIZZ** **PRODUCTON**
Primary Sequence Analysis of PTMs **Module Hours: 3**	**Effective Date: 04/26/2021**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

Figure 7: Resulting output for the human kinase FASTA input into ProP.

6.2.5. A graph summarizing the data might also be available, depending on the protein. (Figure 8)

- It is important to note that the horizontal grey line indicates the threshold score. Any predictions that rise past this bar are likely candidates for peptide cleavage.

Figure 8: Resulting summary graph for the human kinase FASTA input into ProP.
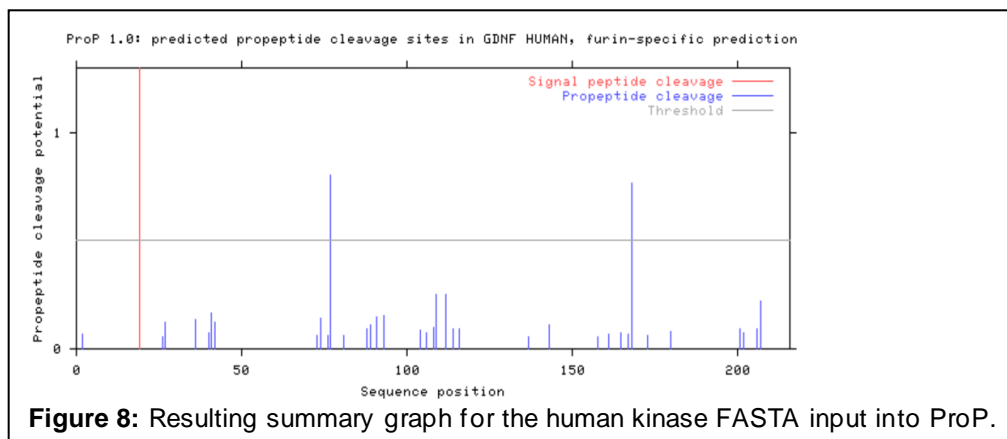
6.3. NetPhos

6.3.1. Go to the NetPhos website: http://www.cbs.dtu.dk/services/NetPhos/
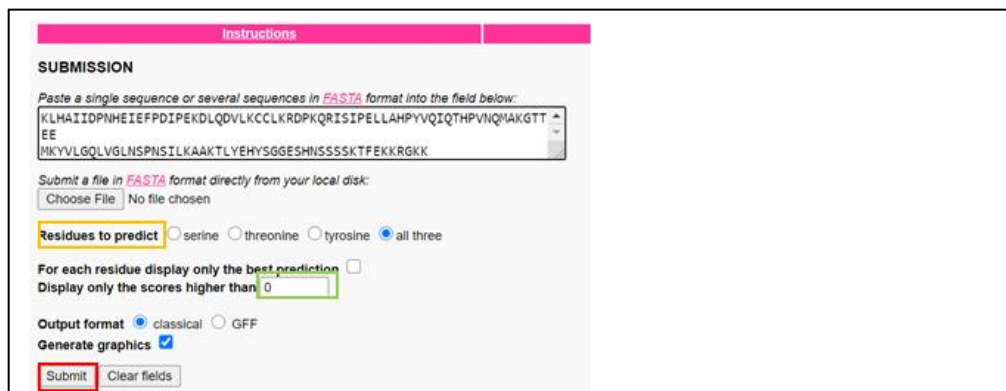
6.3.2. Input the sequence of the protein of interest in the FASTA format. (Figure

6.3.3. The desired residues can be filtered to serine, threonine, tyrosine or a combination of all three. (Figure 9)

6.3.4. The search can be filtered based on score. Changing the value to 0.5 will ensure that all hits shown will be positive predictions. (Figure 9) (Blom, 2017)

6.3.5. Click submit when the settings are as desired. (Figure 9)
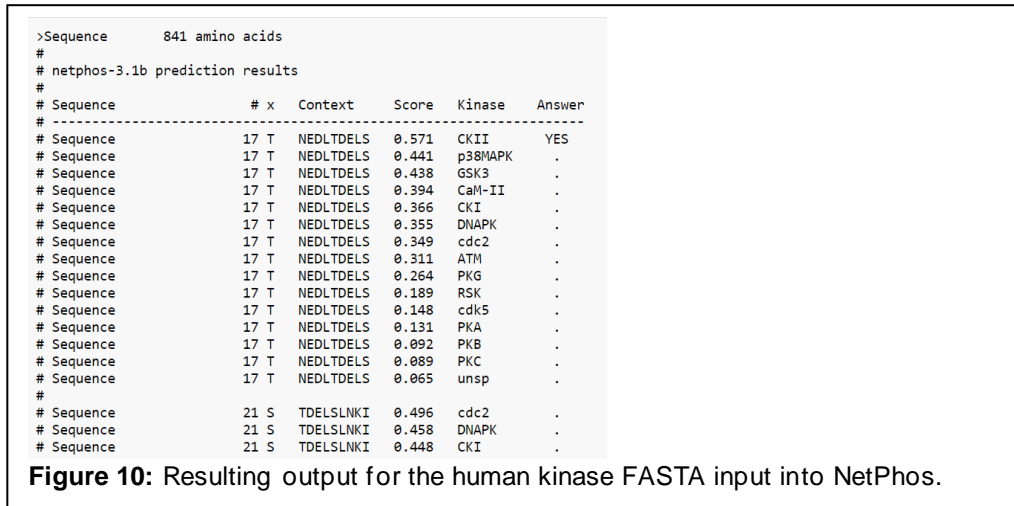

Figure 9: Home page of the NetPhos server with the human kinase FASTA input.

6.3.6. A new window will pop up with the results for the search, with each column giving different important information. (Blom, 2017) (Figure 10)

- # = shows the position of the potential phosphorylation residue
- x = shows the amino acid one letter code for the potential phosphorylation residue
- Context = shows the amino acid one letter codes for the sequence surrounding the potential phosphorylation residue, with the residue centered.
- Score = a value between 0.000 to 1.000. The closer the score is to 1.000, the more likely the residue is to be a true phosphorylation site. 0.500 is the threshold value, and indicates that the prediction is a strong potential candidate.
- Kinase = prediction of the specific kinase that would phosphorylate the potential phosphorylation residue
- Answer = a value of YES will be given if the prediction has a score above threshold and meets the criteria of a strong prediction.
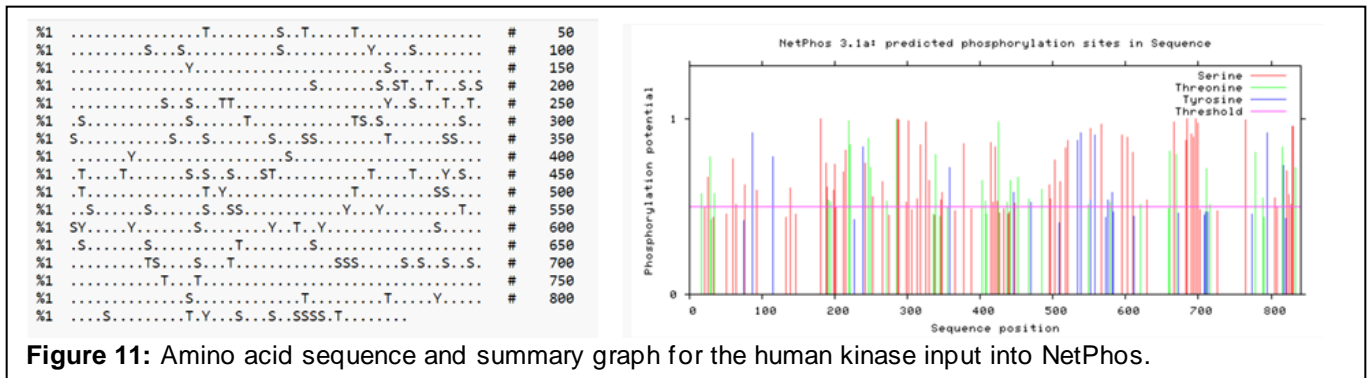
Figure 10: Resulting output for the human kinase FASTA input into NetPhos.

6.3.7. Scrolling to the bottom of the output page will show 2 new viewing formats. (Figure 11)

- First, there will be an amino acid sequence, with residue position labels along the right side. All amino acids except for the potential phosphorylation sites will be ommitted.
- Second, there will be a summary graph that allows for visualization of the different potential phosphorylation sites. Each type of phosphorylation site (serine, threonine or tyrosine) has its own color. It is important to note that the pink line indicates the threshold score.

Figure 11: Amino acid sequence and summary graph for the human kinase input into NetPhos.

	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ PRODUCTON
Primary Sequence Analysis of PTMs **Module Hours: 3**	**Effective Date: 04/26/2021**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**
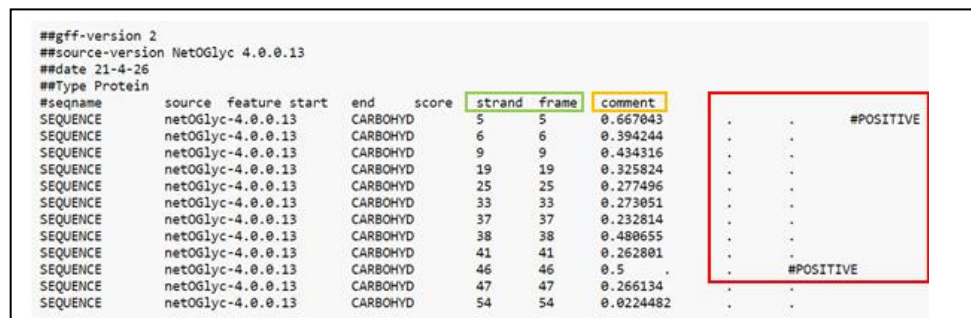
6.4. NetOGlyc

6.4.1. Go to the NetOGlyc website: http://www.cbs.dtu.dk/services/NetOGlyc/

6.4.2. Input the sequence of the protein of interest in the FASTA format and then click submit to begin the query.

6.4.3. A new window will pop up with the results for the search, with each column giving different important information. (Figure 12)

- Strand/frame columns = give the position of the glycosylated residue.
- Comment = gives the score for the residue, a value between 0.000 to 1.000. The closer the score is to 1.000, the more likely the residue is to be a true glycosulation site. 0.500 is the threshold value, and indicates that the prediction is a strong potential candidate.
- Comment cont. = a value of POSITIVE will be given if the prediction has a score above threshold and meets the criteria of a good potential prediction.

Figure 12: Resulting output for the human kinase FASTA input into NetOGlyc.

6.5. NetNGlyc

6.5.1. Go to the NetNGlyc website: http://www.cbs.dtu.dk/services/NetNGlyc/
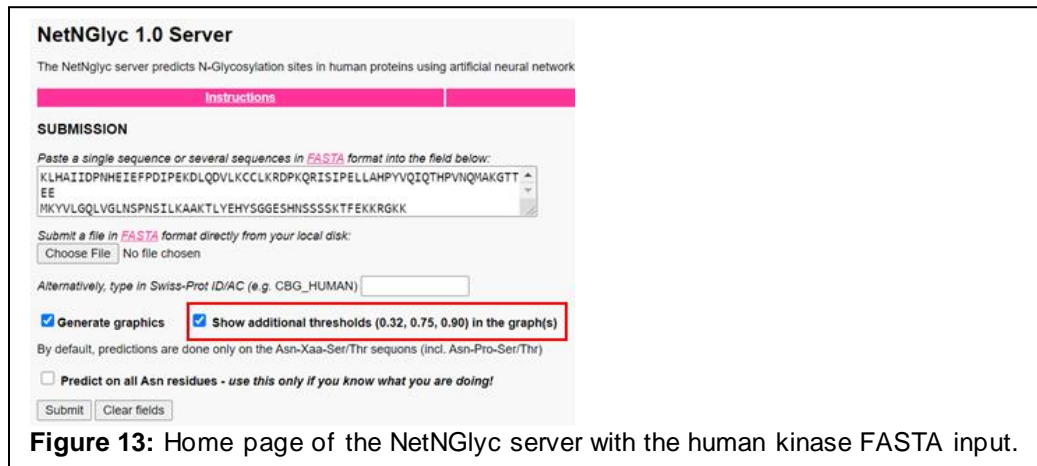
6.5.2. Input the sequence of the protein of interest in the FASTA format.

- By clicking the Show Additional Thresholds box, there is opportunity to create higher or lower confidence levels of the resulting sites. (Figure 13) (Gupta, 2017)

6.5.3. Click submit to begin the query.

Figure 13: Home page of the NetNGlyc server with the human kinase FASTA input.

6.5.4. A new window will pop up with the results for the search, with each column giving different important informatio n. (Figure 14)
- First, the amino acid sequence is given. Blue highlighted regions indicate Asn-Xaa-Ser/Thr sequons. Red highlighted regions indicate Asn residues that are predicted to be N-glycosylated. (Gupta, 2017)

Figure 14: Resulting output for the human kinase FASTA input into NetNGlyc.

- In the table format, like before, values like position and score are given.
- N-Glyc result column annotations for glycosylation sites: (Gupta, 2017)
    1. - = negative site
    2. + = threshold score above 0.5
    3. ++ = threshold score above 0.5 and jury agreement 9/9, or threshold score above 0.75
    4. +++ = threshold score above 0.75 and jury agreement 9/9
    5. ++++ = threshold score above 0.9 and jury agreement 9/9

6.5.5. There will be a summary graph that allows for visualization of the different potential glycosylation sites. It is important to note that the red line indicates the usual 0.5 threshold score. (Figure 15)
   - Additional dotted grey lines indicate the other potential threshold scores, if the show additional thresholds option was chosen in step 6.5.2.

Figure 14: Summary graph for the human kinase input into NetNGlyc.
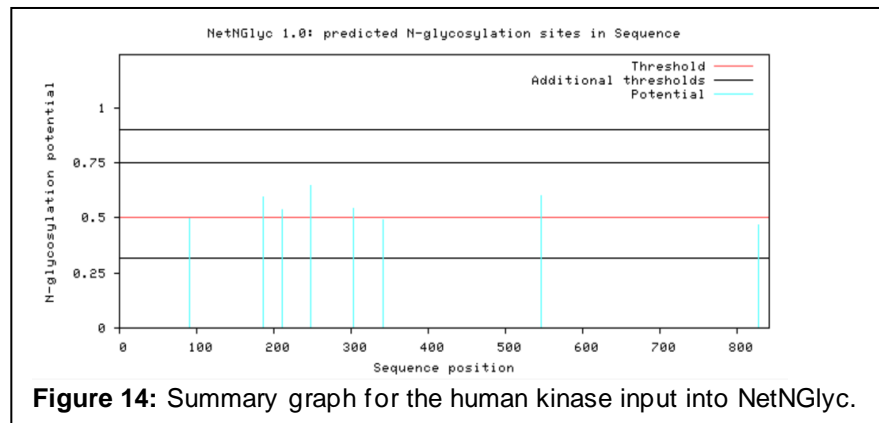
7. TROUBLE SHOOTING

8. REFERENCES

Blom, N. (2017, Jan 2). *Output Format.* DTU Bioinformatics.
   http://www.cbs.dtu.dk/services/ProP/output.php
Blom, N. (2017, Jan 2). *Usage instructions.* DTU Bioinformatics.
   http://www.cbs.dtu.dk/services/ProP/instructions.php
Blom, N. (2017, Jan 5). *NetPhos 3.1 usage instructions.* DTU Bioinformatics.
   http://www.cbs.dtu.dk/services/NetPhos/instructions.php
Blom, N. (2017, Jan 5). *Output Format.* DTU Bioinformatics.
   http://www.cbs.dtu.dk/services/NetPhos-2.0/output.php
Blom N., Gammeltoft S. & Brunak S. (1999) Sequence and structure based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology, 294*(5), 1351-1362. http://www.cbs.dtu.dk/services/NetPhos/
de Castro E., Sigrist C.J.A., Gattiker A., Bulliard V., Petra S. Langendijk-Genevaux P.S., Gasteiger E. Bairoch A. & Hulo N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research, 34*, W362-W365.

	SKILLS CENTER STANDARD OPERATING PROCEDURE	**A BIOFIZZ** **PRODUCTON**
Primary Sequence Analysis of PTMs **Module Hours: 3**	**Effective Date: 04/26/2021**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

https://doi.org/10.1093/nar/gkl124

Duckert P., Brunak S. & Blom, N. (2004) Prediction of proprotein convertase cleavage sites. *Protein Engineering, Design and Selection, 17,* 107-112. http://www.cbs.dtu.dk/services/ProP/

G-Preciado, A., Peimbert, M. & Marino, E. (2009). Genome Sequence Databases: Types of Data and Bioinformatic Tools. *Encyclopedia of Microbiology (Third Edition),* 211-236. https://doi.org/10.1016/B978-012373944-5.00027-4

Gupta, R. (2017, Jan 5). *Instructions.* DTU Bioinformatics. http://www.cbs.dtu.dk/services/NetNGlyc/instructions.php

Gupta, R. (2017, Jan 5) *Output Format.* DTU Bioinformatics. http://www.cbs.dtu.dk/services/NetNGlyc/output.php

Gupta, R. & Brunak, S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput.* 310-322. https://pubmed.ncbi.nlm.nih.gov/11928486/

Hulo N., Bairoch A., Bulliard V., Cerutti L., De Castro E., Langendijk-Genevaux P.S., Pagni M. & Sigrist C.J.A. (2006). The PROSITE database. *Nucleic Acid Research, 34,* D227-D230. https://pubmed.ncbi.nlm.nih.gov/16381852/

Joshi, H. (2017, Jan 3). *Output Format.* DTU Bioinformatics. http://www.cbs.dtu.dk/services/NetOGlyc-4.0/output.php

ThermoFisher Scientific. (n.d.) *Overview of Post-Translational Modifications (PTMs).* ThermoFisher Scientific: Protein Biology Resource Library. https://www.thermofisher.com/us/en/home/life-science/protein-biology/ protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-post-translational-modification.html

Sigrist C.J.A., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bairoch A. & Bucher P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform, 3*(3), 265-74. doi: 10.1093/bib/3.3.265.

Steentoft C., Vakhrushev S.Y., Joshi H.J., Kong Y., Vester-Christensen M.B., Schjoldager K.T., Lavrsen K., Dabelsteen S., Pedersen N.B., Marcos-Silva L., Gupta R., Bennett E.P., Mandel U., Brunak S., Wandall H.H., Levery S.B. & Clausen H. (2013) Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J, 32*(10),1478-88. 10.1038/emboj.2013.79.

Steward, K. (2019, Sep 26). *Amino Acids – the Building Blocks of Proteins.* Applied Sciences from Technology Networks. https://www.technologynetworks.com /applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357

Zhang, Y. (n.d.) *What is FASTA format?* Zhang Lab, University of Michigan.
https://zhanglab.dcmb.med.umich.edu/FASTA/

9. MODULE MASTERY TASK

This task will test your ability to analyze the post-translational modifications of a protein of interest.

Finding a Protein of Interest (do not use the human kinase of the SOP)
1. What is your protein of interest and why did you choose this protein?
2. In one sentence, describe your protein of interest and its function.
3. Go to the NCBI database and obtain the FASTA sequence of your protein of interest. Paste that FASTA sequence here.

PROSITE
1. How many hits were found in how many sequences?
2. How many profile vs. pattern hits does your protein have?
3. Is there an annotated domain on your protein?
4. Give a brief description of the domain.
5. What range of amino acids does the domain from question 3 cover?
6. What is the score value for your domain?
7. Were there any absent features in the hit for this domain?

ProP
Pick a resulting hit that to answer the following questions.
1. What is the position of the residue/cleavage site?
2. Does the cleavage site follow an R or a K residue?
3. Document the context of the cleavage site.
4. What is the score of this hit? What does this indicate?
5. What would you conclude about the predictive strength of this hit?
6. Submit the summary graph of your protein's hits.

NetPhos
Pick a resulting hit that to answer the following questions.
1. What is the position of the potential phosphorylation site?
2. Which residue occurs at this phosphorylation site?
3. Document the context of the phosphorylation site.

4. What is the score of this hit? What does this indicate?
5. What would you conclude about the predictive strength of this hit?
6. Submit the summary graph of your protein's hits.

NetOGlyc
Pick a resulting hit that to answer the following questions.
1. What is the position of the potential glycosylation site?
2. What is the score of this hit? What does this indicate?
3. What would you conclude about the predictive strength of this hit?

NetNGlyc
Pick a resulting hit that to answer the following questions.
1. How many Asn-Xaa-Ser/Thr sequons are there in your protein? How many Asn residues are predicted to be N-glycosylated?
2. How many "+" hits were there for your protein? How many negative hits?
3. Submit the summary graph of your protein's hits.