**Supplemental Table 2.** A list of the 19 protein annotation tools used to generate 122 annotation features for antigens and non-antigens in the VaxiJen and BPA training data sets1.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Program** | **Version** | **Loca-tion** | **Annotation** | **AF No.** | **AF Name** | **Description** | **URL** | **Ref.** |
| NA | NA | NA | NA | 1 | Length | Protein length | NA | NA |
| DictyOGlyc | 1.1 | Web | Glycosylation | 2 | MaxScore | Maximum score of the predicted glycosylation sites | http://www.cbs.dtu.dk/services/DictyOGlyc/ | [1] |
|  |  |  |  | 3 | AvgScore | Average score of all predicted glycosylation sites |  |  |
|  |  |  |  | 4 | MaxDiff | Maximum difference between the site-score and threshold |  |  |
|  |  |  |  | 5 | AvgDiff | Average difference between the site-score and threshold |  |  |
|  |  |  |  | 6 | Count | Number of predicted glycosylation sites |  |  |
|  |  |  |  | 7 | CorrCount | Number of predicted glycosylation sites, normalized for protein length |  |  |
| NetAcet | 1.0 | Web | Acetylation | 8 | A-Flag | 1 if the acetylated residue is Alanine, otherwise 0 | http://www.cbs.dtu.dk/services/NetAcet/ | [2] |
|  |  |  |  | 9 | G-Flag | 1 if the acetylated residue is Glycine, otherwise 0 |  |  |
|  |  |  |  | 10 | S-Flag | 1 if the acetylated residue is Serine, otherwise 0 |  |  |
|  |  |  |  | 11 | T-Flag | 1 if the acetylated residue is Threonine, otherwise 0 |  |  |
|  |  |  |  | 12 | Score | Score for the potential acetylation site |  |  |
|  |  |  |  | 13 | SiteFlag | 1 if there is a potentially acetylated residue, otherwise 0 |  |  |
|  |  |  |  | 14 | AcetFlag | 1 if the residue is predicted to be acetylated, otherwise 0 |  |  |
| NetGlycate | 1.0 | Web | Glycation | 15 | MaxScore | Maximum score of predicted glycation sites | http://www.cbs.dtu.dk/services/NetGlycate/ | [3] |
|  |  |  |  | 16 | AvgScore | Average score of all predicted glycation sites |  |  |
|  |  |  |  | 17 | Count | Number of predicted glycation sites |  |  |
|  |  |  |  | 18 | CorrCount | Number of predicted glycation sites, normalized for protein length |  |  |
| NetPhosBac | 1.0 | Web | Phosphorylation | 19 | MaxScore | Maximum score of predicted phosphorylation sites | http://www.cbs.dtu.dk/services/NetPhosBac-1.0/ | [4] |
|  |  |  |  | 20 | AvgScore | Average score of all predicted phosphorylation sites |  |  |
|  |  |  |  | 21 | Count | Number of predicted phosphorylation sites |  |  |
|  |  |  |  | 22 | CorrCount | Number of predicted phosphorylation sites, normalized for protein length |  |  |
| NetPhosYeast | 1.0 | Web | Phosphorylation | 23 | MaxScore | Maximum score of predicted phosphorylation sites | http://www.cbs.dtu.dk/services/NetPhosYeast/ | [5] |
|  |  |  |  | 24 | AvgScore | Average score of all predicted phosphorylation sites |  |  |
|  |  |  |  | 25 | Count | Number of predicted phosphorylation sites |  |  |
|  |  |  |  | 26 | CorrCount | Number of predicted phosphorylation sites, normalized for protein length |  |  |
| ProtParam | Biopython 1.53 | Local | Basic Protein Stats | 27 | Isoelectric | Predicted isoelectric point | http://ca.expasy.org/tools/protparam.html | [6] |
|  |  |  |  | 28 | Instability | Predicted instability index |  |  |
|  |  |  |  | 29 | MolecWeight | Predicted Molecular Weight |  |  |
|  |  |  |  | 30 | Aromaticity | Predicted protein aromaticity |  |  |
|  |  |  |  | 31 | GRAVY | Predicted grand average of hydropathy |  |  |
|  |  |  |  | 32 | PercTurn | Percent of protein predicted to be a part of a turn |  |  |
|  |  |  |  | 33 | PercHelix | Percent of protein predicted to be a part of an alpha-helix |  |  |
|  |  |  |  | 34 | PercSheet | Percent of protein predicted to be a part of a beta-sheet |  |  |
|  |  |  |  | 35 | PercAlanine | Percentage of residues that are alanine |  |  |
|  |  |  |  | 36 | PercCysteine | Percentage of residues that are cysteine |  |  |
|  |  |  |  | 37 | PercAsparticAcid | Percentage of residues that are aspartic acid |  |  |
|  |  |  |  | 38 | PercGlutamicAcid | Percentage of residues that are glutamic acid |  |  |
|  |  |  |  | 39 | PercPhenylalanine | Percentage of residues that are phenylalanine |  |  |
|  |  |  |  | 40 | PercGlycine | Percentage of residues that are glycine |  |  |
|  |  |  |  | 41 | PercHistidine | Percentage of residues that are histidine |  |  |
|  |  |  |  | 42 | PercIsoleucine | Percentage of residues that are isoleucine |  |  |
|  |  |  |  | 43 | PercLysine | Percentage of residues that are lysine |  |  |
|  |  |  |  | 44 | PercLeucine | Percentage of residues that are leucine |  |  |
|  |  |  |  | 45 | PercMethionine | Percentage of residues that are methionine |  |  |
|  |  |  |  | 46 | PercAsparagine | Percentage of residues that are asparagine |  |  |
|  |  |  |  | 47 | PercProline | Percentage of residues that are proline |  |  |
|  |  |  |  | 48 | PercGlutamine | Percentage of residues that are glutamine |  |  |
|  |  |  |  | 49 | PercArginine | Percentage of residues that are arginine |  |  |
|  |  |  |  | 50 | PercSerine | Percentage of residues that are serine |  |  |
|  |  |  |  | 51 | PercThreonine | Percentage of residues that are threonine |  |  |
|  |  |  |  | 52 | PercValine | Percentage of residues that are valine |  |  |
|  |  |  |  | 53 | PercTryptophan | Percentage of residues that are tryptophan |  |  |
|  |  |  |  | 54 | PercTyrosine | Percentage of residues that are tyrosine |  |  |
| NetPhosK | 3.1 | Local | Phosphorylation | 55 | MaxScore | Maximum score of predicted phosphorylation sites | http://www.cbs.dtu.dk/services/NetPhosK/ | [7] |
|  |  |  |  | 56 | AvgScore | Average score of all predicted phosphorylation sites |  |  |
|  |  |  |  | 57 | Count | Number of predicted phosphorylation sites |  |  |
|  |  |  |  | 58 | CorrCount | Number of predicted phosphorylation sites, normalized for protein length |  |  |
| YinOYang | 1.2 | Local | O-Linked Beta-N-acetylglucosamine | 59 | MaxRank | Maximum rank of predicted glc-n-ac sites | http://www.cbs.dtu.dk/services/YinOYang/ | [8] |
|  |  |  |  | 60 | AvgRank | Average rank of predicted glc-n-ac sites |  |  |
|  |  |  |  | 61 | MaxScore | Maximum score of predicted glc-n-ac sites |  |  |
|  |  |  |  | 62 | AvgScore | Average score of predicted glc-n-ac sites |  |  |
|  |  |  |  | 63 | MaxDiff1 | Maximum difference between the prediction scores and the lower threshold |  |  |
|  |  |  |  | 64 | AvgDiff1 | Average difference between the prediction scores and the lower threshold |  |  |
|  |  |  |  | 65 | MaxDiff2 | Maximum difference between the prediction scores and the higher threshold |  |  |
|  |  |  |  | 66 | AvgDiff2 | Average difference between the prediction scores and the higher threshold |  |  |
|  |  |  |  | 67 | Count | Number of predicted glc-n-ac sites |  |  |
|  |  |  |  | 68 | CorrCount | Number of predicted glc-n-ac sites, normalized for protein length |  |  |
| LipoP | 1.0a | Local | Lipoproteins & Signal Peptides | 69 | PeptidaseI | Predicted non-lipoprotein signal peptide cleavage site | http://www.cbs.dtu.dk/services/LipoP/ | [9] |
|  |  |  |  | 70 | PeptidaseII | Predicted lipoprotein signal peptide cleavage site |  |  |
|  |  |  |  | 71 | Transmembrane | Predicted n-terminal transmembrane helix |  |  |
|  |  |  |  | 72 | Cytoplasmic | No predicted signal protein - predicted localization to the cytoplasm |  |  |
|  |  |  |  | 73 | Score | Score associated with the above prediction |  |  |
| TargetP | 1.1 | Local | Subcellular Localization | 74 | SecretScore | Score for secretory pathway signal peptide | http://www.cbs.dtu.dk/services/TargetP/ | [10] |
|  |  |  |  | 75 | MitoScore | Score for the mitochondrial targeting peptide |  |  |
|  |  |  |  | 76 | OtherScore | Score for non-secretory, non-mitochondrial localization |  |  |
|  |  |  |  | 77 | SecretFlag | 1 if the program predicts that the protein is secreted, otherwise 0 |  |  |
|  |  |  |  | 78 | MitoFlag | 1 if the program predicts that the protein is localized to the mitochondria, otherwise 0 |  |  |
|  |  |  |  | 79 | OtherFlag | 1 if the program predicts that the protein is neither secreted nor mitochondrial, otherwise 0 |  |  |
| NetNGlyc | 1.0a | Local | N-Glycosylation | 80 | MaxScore | Maximum score of predicted N-glycosylation sites | http://www.cbs.dtu.dk/services/NetNGlyc/ | [11] |
|  |  |  |  | 81 | AvgScore | Average score of predicted N-glycosylation sites |  |  |
|  |  |  |  | 82 | MaxRank | Maximum rank of predicted N-glycosylation sites |  |  |
|  |  |  |  | 83 | AvgRank | Average rank of predicted N-glycosylation sites |  |  |
|  |  |  |  | 84 | Count | Number of predicted N-glycosylation sites |  |  |
| NetOGlyc | 3.1d | Local | O-Glycosylation | 85 | MaxGforT | Maximum generalized score for all predicted threonine O-glycosylation sites | http://www.cbs.dtu.dk/services/NetOGlyc/ | [12] |
|  |  |  |  | 86 | MaxIforT | Maximum isolated score for all predicted threonine O-glycosylation sites |  |  |
|  |  |  |  | 87 | AvgGforT | Average generalized score for all predicted threonine O-glycosylation sites |  |  |
|  |  |  |  | 88 | AvgIforT | Averages isolated score for all predicted threonine O-glycosylation sites |  |  |
|  |  |  |  | 89 | HitsForT | Number of predicted threonine O-glycosylation sites |  |  |
|  |  |  |  | 90 | CountForT | Number of possible threonine O-glycosylation sites |  |  |
|  |  |  |  | 91 | MaxGforS | Maximum generalized score for all predicted serine O-glycosylation sites |  |  |
|  |  |  |  | 92 | MaxIforS | Maximum isolated score for all predicted serine O-glycosylation sites |  |  |
|  |  |  |  | 93 | AvgGforS | Average generalized score for all predicted serine O-glycosylation sites |  |  |
|  |  |  |  | 94 | AvgIforS | Averages isolated score for all predicted serine O-glycosylation sites |  |  |
|  |  |  |  | 95 | HitsForS | Number of predicted serine O-glycosylation sites |  |  |
|  |  |  |  | 96 | CountForS | Number of possible serine O-glycosylation sites |  |  |
| ProP | 1.0c | Local | Cleavage Sites | 97 | Furin-MaxScore | Maximum score of predicted cleavage sites (furin mode) | http://www.cbs.dtu.dk/services/ProP/ | [13] |
|  |  |  |  | 98 | Furin-AvgScore | Average score of predicted cleavage sites (furin mode) |  |  |
|  |  |  |  | 99 | Furin-Count | Number of predicted cleavage sites (furin mode) |  |  |
|  |  |  |  | 100 | General-MaxScore | Maximum score of predicted cleavage sites (general mode) |  |  |
|  |  |  |  | 101 | General-AvgScore | Average score of predicted cleavage sites (general mode) |  |  |
|  |  |  |  | 102 | General-Count | Number of predicted cleavage sites (general mode) |  |  |
| BepiPred | 1.0b | Local | Linear B-cell epitopes | 103 | MaxScore | Maximum score of predicted linear epitopes | http://www.cbs.dtu.dk/services/BepiPred/ | [14] |
|  |  |  |  | 104 | AvgScore | Average score of predicted linear epitopes |  |  |
|  |  |  |  | 105 | PercEpitope | Percent of the protein predicted to be a part of a linear epitope |  |  |
| TMHMM | 2.0c | Local | Transmembrane Helices | 106 | TransAAcount | Number of amino acids predicted to be a part of trans-membrane helices | http://www.cbs.dtu.dk/services/TMHMM/ | [15] |
|  |  |  |  | 107 | StartAAcount | Number of amino acids, from the first 60, predicted to be a part of trans-membrane helices |  |  |
|  |  |  |  | 108 | Count | Number of predicted trans-membrane helices |  |  |
| HMMTOP | 2.1 | Local | Transmembrane Helices | 109 | Count | Number of predicted trans-membrane helices | http://www.enzim.hu/hmmtop/ | [16] |
| PSORTb | 3.0 | Local | Subcellular Localization | 110 | ProbCytoMem | Predicted probability that the protein is localized to the cytoplasmic membrane | http://www.psort.org/psortb/ | [17] |
|  |  |  |  | 111 | ProbCytoplasm | Predicted probability that the protein is localized to the cytoplasm |  |  |
|  |  |  |  | 112 | ProbPeriplasm | Predicted probability that the protein is localized to the periplasm |  |  |
|  |  |  |  | 113 | ProbExtracell | Predicted probability that the protein is exported from the cell |  |  |
|  |  |  |  | 114 | ProbOuterMem | Predicted probability that the protein is localized to the outer membrane |  |  |
|  |  |  |  | 115 | ProbCellWall | Predicted probability that the protein is localized to the cell wall |  |  |
| SignalP | 3.0 | Local | Signal Peptides | 116 | NN-MaxC | Neural network predicted maximum cleavage site score | http://www.cbs.dtu.dk/services/SignalP/ | [18] |
|  |  |  |  | 117 | NN-MaxY | Combined neural network predicted scores for cleavage and signal peptide probability |  |  |
|  |  |  |  | 118 | NN-MaxS | Neural network predicted maximum signal peptide site score |  |  |
|  |  |  |  | 119 | NN-AvgS | Average of all signal peptide scores n-terminal to the predicted cleavage site |  |  |
|  |  |  |  | 120 | D-score | Average value of AvgS and MaxY, known to provide better discrimination than MaxY alone |  |  |
|  |  |  |  | 121 | HMM-MaxC | Alternate HMM-based prediction of the likelihood of protein cleavage |  |  |
|  |  |  |  | 122 | HMM-ProbS | HMM-based predicted probability of whether the protein has a signal peptide or not |  |  |

1In many cases several different annotation features could be derived from a single protein annotation tool and these are indicated in the table. Abbreviations: *AF No*., refers to a number assigned to each of the 122 annotation features derived from the 19 protein annotation tools; *AF name*, refers to the label given to the annotation feature; *Ref.*, refers to the literature reference for the protein annotation tool listed below; *NA*, not applicable. Shading is for display purposes only.

**References for Supplemental Table 2:**

[1] Gupta R, Jung E, Gooley AA, Williams KL, Brunak S, Hansen J. Scanning the available Dictyostelium discoideum proteome for O-linked GlcNAc glycosylation sites using neural networks. Glycobiology 1999;9:1009-22.

[2] Kiemer L, Bendtsen JD, Blom N. NetAcet: prediction of N-terminal acetylation sites. Bioinformatics 2005;21:1269 -70.

[3] Johansen MB, Kiemer L, Brunak S. Analysis and prediction of mammalian protein glycation. Glycobiology 2006;16:844 -53.

[4] Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I. NetPhosBac – A predictor for Ser/Thr phosphorylation sites in bacterial proteins. PROTEOMICS 2009;9:116-25.

[5] Ingrell CR, Miller ML, Jensen ON, Blom N. NetPhosYeast: prediction of protein phosphorylation sites in yeast. Bioinformatics 2007;23:895 -7.

[6] Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 1999;112:531-52.

[7] Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. PROTEOMICS 2004;4:1633-49.

[8] Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput, 2002.

[9] Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 2003;12:1652-62.

[10] Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005-16.

[11] Gupta R, Jung E, Brunak S. Prediction of N-glycosylation sites in human proteins. 2004.

[12] Julenius K, Molgaard A, Gupta R, Brunak S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 2005;15:153-64.

[13] Duckert P, Brunak S, Blom N. Prediction of proprotein convertase cleavage sites. Protein Engineering Design and Selection 2004;17:107 -12.

[14] Larsen J, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Research 2006;2:2.

[15] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567-80.

[16] Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849-50.

[17] Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 2010;26:1608 -15.

[18] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved Prediction of Signal Peptides: SignalP 3.0. Journal of Molecular Biology 2004;340:783-95.