Published online 28 October 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D1535–D1540***
https://doi.org/10.1093/nar/gkab944

Proteome-*pI* 2.0: proteome isoelectric point database update

Lukasz Pawel Kozlowski *

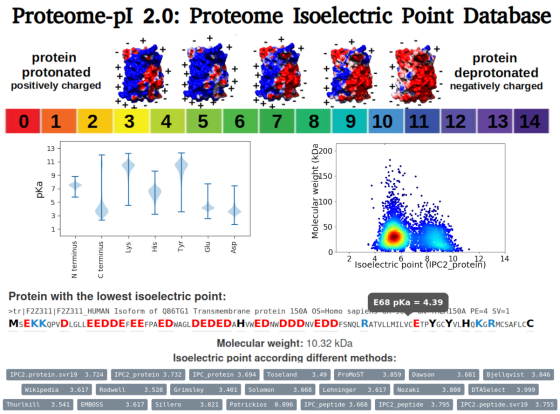
Institute of Informatics, Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Mazovian Voivodeship 02-097, Poland

Received September 19, 2021; Revised September 28, 2021; Editorial Decision September 28, 2021; Accepted October 04, 2021

ABSTRACT

Proteome-*pI* 2.0 is an update of an online database containing predicted isoelectric points and pK_a dissociation constants of proteins and peptides. The isoelectric point—the pH at which a particular molecule carries no net electrical charge—is an important parameter for many analytical biochemistry and proteomics techniques. Additionally, it can be obtained directly from the pK_a values of individual charged residues of the protein. The Proteome-*pI* 2.0 database includes data for over 61 million protein sequences from 20 115 proteomes (three to four times more than the previous release). The isoelectric point for proteins is predicted by 21 methods, whereas pKa values are inferred by one method. To facilitate bottom-up proteomics analysis, individual proteomes were digested *in silico* with the five most commonly used proteases (trypsin, chymotrypsin, trypsin + LysC, LysN, ArgC), and the peptides' isoelectric point and molecular weights were calculated. The database enables the retrieval of virtual 2D-PAGE plots and customized fractions of a proteome based on the isoelectric point and molecular weight. In addition, isoelectric points for proteins in NCBI non-redundant (nr), UniProt, SwissProt, and Protein Data Bank are available in both CSV and FASTA formats. The database can be accessed at http://isoelectricpointdb2.org.

GRAPHICAL ABSTRACT



INTRODUCTION

The charge of a protein is one of its key physicochemical characteristics and is related to the pK_a dissociation constant (pK_a is a quantitative measure of the strength of an acid in solution). For proteins and peptides, the ionizable groups of seven charged amino acids should be considered: glutamate (γ-carboxyl group), cysteine (thiol group), aspartate (β-carboxyl group), tyrosine (phenol group), lysine (ϵ-ammonium group), histidine (imidazole side chains), and arginine (guanidinium group) (1). Taken together, the pK_a values of all charged groups can be used to calculate the overall charge of the molecule in any pH or to estimate the isoelectric point (*pI*, IEP), that is, the pH at which there is an equilibrium of positive and negative charges and therefore the total net charge of the molecule is equal to zero (2). Both pK_a and isoelectric point estimates have been used in numerous techniques, such as two-dimensional gel electrophoresis (2D-PAGE) (3,4), crystallization (5), capillary isoelectric focussing (6), and mass spectrometry (MS) (7,8). It should be stressed that experimental measurements of pK_a values [PKAD database (9)] and isoelectric point [SWISS-2DPAGE (10)] are very limited (a few thousand records at most), but there are many computational methods that can be used to predict these features. In this work, I present a

*To whom correspondence should be addressed. Tel: +48 22 55 44 454; Fax: +48 22 55 44 400; Email: lukaszkozlowski.lpk@gmail.com

major update of the original Proteome-*pI* database (Figure 1) (11). The following changes have been introduced:

- the number of proteomes included has been increased four-fold (from 5029 to 20 115);
- new algorithms for isoelectric point prediction have been added (21 algorithms in total);
- the prediction of pK_a dissociation constants for over 61 million proteins have been included;
- the prediction of isoelectric point for *in silico* digests of proteomes with the five most commonly used proteases (trypsin, chymotrypsin, trypsin + LysC, LysN, ArgC) have been added.

MATERIALS AND METHODS

Datasets

Proteome-*pI* 2.0 is based on UniProt (12) reference proteomes (2021_03 release) and contains over 61 million protein sequences coming from 20 115 model organisms (Table 1 and Supplementary Table S1). The data are divided according to the major kingdoms of the tree of life and include splicing variants for eukaryotic organisms. Additionally, the isoelectric point is predicted for the most commonly used protein sequence databases, such as the entire UniProt TrEMBL with 219 million sequences (12), SwissProt with 561 000 proteins (13,14), NCBI nr (non-redundant) with 409 million sequences (15), and Protein Data Bank with 601 000 protein chains (16).

Predictions for proteins

Each proteome is analysed by various methods. The prediction of the isoelectric point is currently performed using 21 methods (including four new ones), which can be grouped into two categories. The simplest methods of isoelectric point prediction are based on experimentally derived pK_a sets and the Henderson–Hasselbach equation: Patrickios (17), Solomons (18), Lehninger (19), EMBOSS (20), Dawson (21), Wikipedia (pK_a values as presented in Wikipedia page in 2005), Toseland (22), Sillero (23), Thurlkill (24), Rodwell (25), DTASelect (26), Nozaki (27), Grimsley (28), Bjellqvist (29) [whose method was implemented as ExPASy 'Compute pI/Mw Tool' (30)] and ProMoST (31). The second group includes methods that are based on machine learning [IPC_protein, IPC_peptide, IPC2_protein, IPC2_peptide, IPC2.peptide.svr19, and IPC2.protein.svr (32,33)]. Moreover, in Proteome-*pI* 2.0, a completely new category of predictions has been introduced, namely the prediction of pK_a dissociation constants. In this case, only one algorithm is used [IPC2.pKa (33)], as other methods for pK_a prediction are prohibitively slow and additionally require structural data (not available in Proteome-*pI*) (34–37).

Predictions for peptides

To facilitate bottom-up mass spectrometry analysis, *in silico* proteolytic digestion of proteins by the five most commonly used proteases (trypsin, chymotrypsin, trypsin + LysC, LysN, ArgC) has been introduced (38). The proteolytic products (i.e. peptides) are treated as the surrogates of the parent proteins for further qualitative or quantitative analysis. The proteases generally cleave proteins at specific amino acid residue sites, but digestion is frequently incomplete (missed cleavage sites are widespread). To predict proteolysis, the Rapid Peptides Generator (RPG) program was used (with a 1.4% miscleavage rate) (39). The resulting five datasets are further categorized according to the molecular mass of the peptides (Figure 1 and Supplementary Table S2): ESI Ion Trap (600–3500 Da), LTQ Orbitrap (600–4000 Da), MALDI TOF/TOF (750–5500 Da), MS low (narrow range of mass, 800–3500 Da), and MS high (wide range of mass, 600–5500 Da) (35). Finally, for the resulting peptides, the isoelectric point is predicted.

RESULTS

A single results page for Proteome-*pI* displays a comprehensive overview of the complete proteome (one from 20 115 model organisms). The isoelectric point predictions for all proteins (including splicing isoforms or alternative sequences) are available, together with a virtual 2D-PAGE plot. The user can retrieve customized datasets according to specified isoelectric point and molecular mass ranges. Extreme examples (proteins with minimal and maximal isoelectric point predictions) are then presented. The information is complemented with plots depicting global isoelectric point and pK_a predictions according different methods (Supplementary Figure S1). In the next panel, the user can find *in silico* digests of the whole proteome with trypsin, chymotrypsin, trypsin + LysC, LysN and ArgC proteases suitable for different mass spectrometry machines, such as the ESI Ion Trap (600–3500 Da), LTQ Orbitrap (600–4000 Da), MALDI TOF/TOF (750–5500 Da), MS low (narrow range of mass, 800–3500 Da), and MS high (wide range of mass, 600–5500 Da). This can result in a huge number of potential peptides (e.g. for human proteins, trypsin digests can exceed two million peptides; Supplementary Table S2). At the bottom, general statistics such as amino acid and di-amino acid frequencies can be found. Additionally, each page is interconnected to external databases, such as UniProt and NCBI Taxonomy.

Furthermore, Proteome-*pI* 2.0 provides global analyses related to the distribution of molecular weight and isoelectric points across kingdoms, or amino and di-amino acid statistics (Table 2 and Supplementary Table S3). Such data can be useful for high-throughput analysis of specific taxons, such as plants (40), fungi (41) or groups of interacting proteins (42).

DISCUSSION

The Proteome-*pI* 2.0 database update is a significant improvement upon the previous version, both quantitatively (covering more proteomes and using more algorithms) and qualitatively (including peptide digests and pK_a predictions). Nevertheless, apart from the technical extension of the database (analysing more organisms), it is always worth checking how the addition of new data may have affected some global conclusions drawn from the data available at the time of evaluation.

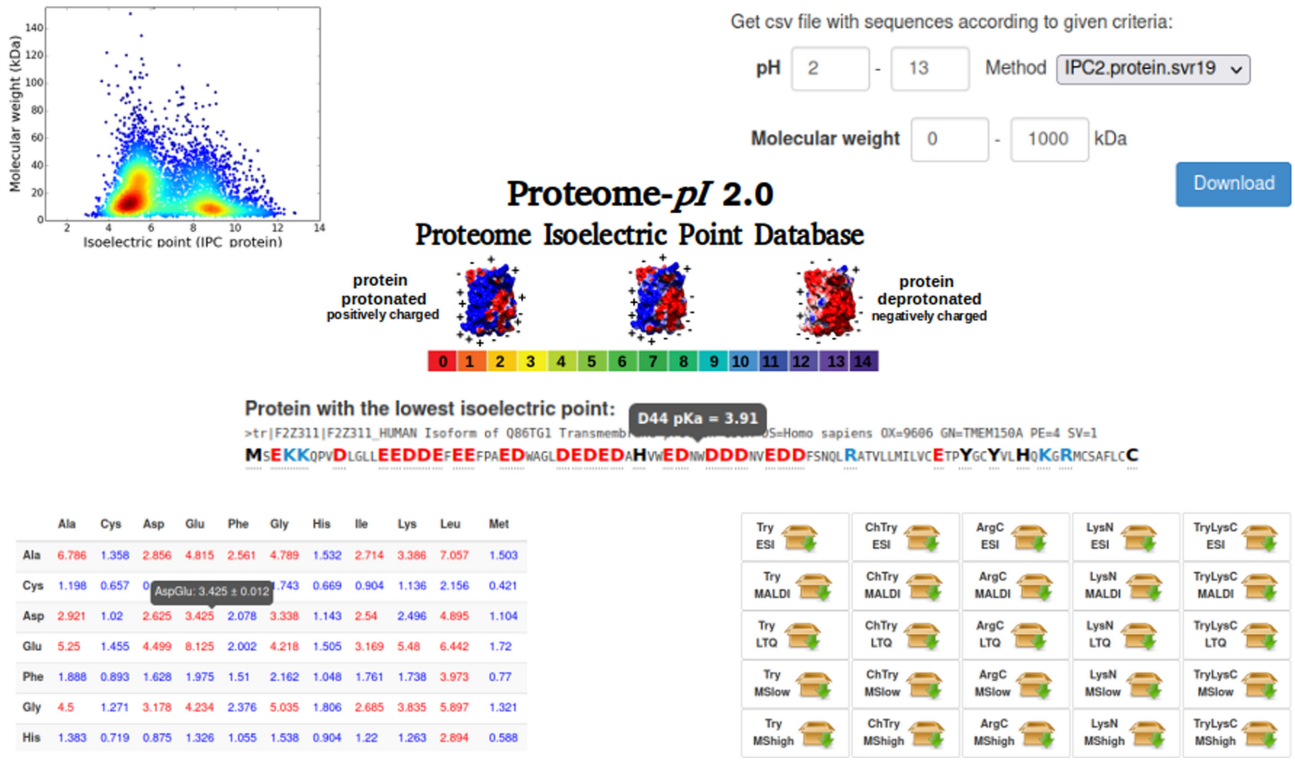


Figure 1. An overview of the Proteome-*pI* 2.0 database. Isoelectric points and molecular weights for individual proteins from 20 115 proteomes are visualized on virtual 2D PAGE plots (top left) and can be retrieved according to the predictions from one of 21 algorithms (top right). The data for individual proteins are accompanied by dissociation constant (pK_a) predictions (middle). The proteomes are digested *in silico* by one of the five most commonly used proteases (trypsin, chymotrypsin, trypsin + LysC, LysN, ArgC) (bottom right). Additionally, auxiliary statistics are provided (e.g. di-amino acid frequencies) (bottom left).

Table 1. General statistics of the Proteome-*pI* 2.0 database (20 115 proteomes with 61 329 034 proteins in total)

	Number of proteomes	Total number of proteins	Mean number of proteins (±SD)	Mean size of proteins (±SD)	Mean mw of proteins (±SD)
Viruses	10 064	518 140	51 ± 85	237 ± 300	26.6 ± 33.2
Archaea	331	767 951	2320 ± 1263	278 ± 211	30.6 ± 23.1
Bacteria	8108	30 290 647	3736 ± 1785	320 ± 246	35.1 ± 26.8
Eukaryote	1612	29 752 296	18457 ± 16804	467 ± 471	52.1 ± 52.4
Eukaryote (major)	1612	25 437 198	15780 ± 11138	438 ± 420	48.8 ± 46.7
Eukaryote (minor)	637	4 315 098	6774 ± 14244	638 ± 676	71.2 ± 75.4

mw, molecular weight in kDa; mean size in amino acids. For more statistics, see Supplementary Table S1. 'Major' and 'minor' refer to splicing isoforms of proteins used for calculation of the statistics.

For instance, one of the scientifically important by-products of creating Proteome-*pI* was the observation that the isoelectric points and molecular weights of proteins in different kingdoms vary considerably. For example, Archaea have the smallest proteins (except for viruses), but the isoelectric point of the proteome can differ greatly among individual species. This may be because Archaea are known for living in extreme environments (e.g. low or high pH), which affects the range of isoelectric point in their proteomes. In 2016, when the first version of the database was created, only 135 Archaeal organisms were included, whereas in the current version we have 331 such proteomes. Careful comparison of Figure 2 from Kozlowski (11) with Supplementary Figure S2 shows that indeed the trend is following an analysis of more Archaea, highlighting how unique and diverse these organisms can be in terms of their proteins' charge (see also Supplementary Figure S1).

Similarly, many statistics calculated previously have been repeated on the larger dataset, using a new version of a proteome or extending the calculation from the statistical perspective. For instance, two auxiliary statistics that Proteome-*pI* provides are amino and di-amino acid frequencies for whole proteomes. In the current version, we added error estimates (with × 100 bootstrapping at the protein level) to assess the possible variability of the calculations. This is not a purely technical aspect, as our knowledge about what constitutes the proteome of a given organism changes over time, and consequently we can draw conclusions different to those based on the data from the past. This is a highly dynamic situation, even for intensively studied organisms. For example, the human proteome in

Table 2. Amino acid frequency for the kingdoms of life in the Proteome-*pI* 2.0 database

Kingdom	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	Total amino acids
Viruses	7.81	1.29	6.20	6.46	3.91	6.72	1.96	6.05	6.24	8.28	2.51	4.99	4.25	3.62	5.31	6.47	6.14	6.66	1.42	3.71	122 870 810
Archaea	8.95	0.90	w7.00	7.94	3.65	7.84	1.86	6.03	4.18	9.11	2.14	3.36	4.36	2.48	5.83	6.12	5.84	8.16	1.06	3.18	213 285 886
Bacteria	10.64	0.90	5.67	6.06	3.76	8.01	2.08	5.52	4.22	10.12	2.31	3.35	4.82	3.49	6.18	5.75	5.58	7.42	1.31	2.81	9 693 905 784
Eukaryota	7.38	1.85	5.34	6.55	3.79	6.35	2.50	4.94	5.64	9.38	2.27	4.13	5.56	4.27	5.71	8.45	5.56	6.24	1.24	2.81	13 901 635 566
All	8.72	1.46	5.49	6.36	3.78	7.04	2.32	5.19	5.05	9.67	2.29	3.81	5.24	3.94	5.90	7.33	5.57	6.74	1.27	2.81	23 931 698 046

Similar statistics for the 20 115 individual proteomes included in Proteome-*pI* are available online on separate subpages. Additionally, the online version of the table http://isoelectricpointdb2.org/statistics.html is accompanied by an error estimated with 1000 bootstraps. For di-amino acid frequencies, see Supplementary Table S3.

2016 constituted 21 006 proteins with 71 173 splicing isoforms (92 179 in total). Now, we have 20 600 protein annotations with 79 500 splicing isoforms (100 100 in total), and this does not take into account the recent T2T-CHM13 reference genome update (43). The situation may be even more dramatic for proteomes that may have been only recently studied intensively in terms of proteomics. For example, X*enopus tropicalis* in 2016 had 18,252 annotated proteins, with an average isoelectric point of 6.70 and an average molecular mass of 60.1 kDa, accompanied by 5346 splicing isoforms (23 598 in total). Now, it has 22 514 proteins (average isoelectric point of 6.64 and average mass of 71.9 kDa), and 23 799 splicing isoforms have been identified. Accordingly, we decided to maintain the previous version of Proteome-*pI* (http://isoelectricpointdb.org) and present the new release as a completely new resource (http://isoelectricpointdb2.org).

Future prospects

The number of reference proteomes has increased 4-fold during the last five years (5029 in Proteome-*pI* 1.0 versus 20 115 in the current release); therefore, constant addition of new proteomes is of great interest. Furthermore, users frequently request respective data for proteomes of interest to them, such as a particular strain of bacteria or virus not included in the official release but relevant to their ongoing studies (44). In parallel, the addition of new algorithms for isoelectric point and pK_a prediction is foreseen. The latter is especially worth consideration, as the database currently includes the prediction of pK_a values by only one method. This limitation will not be easy to overcome, as most of the pK_a predictors [e.g. Rosetta pKa (45), H++ (35), MCCE (36)] rely on protein structure information. However, the advance of the SWISS-MODEL Repository (46) and recently the AlphaFold Protein Structure Database (47) gives hope that Proteome-*pI* could be also extended by 3D-based protein predictions. It is worth mentioning here that there are already some efforts for making predictions of isoelectric points and pK_a values based on available protein structures [pKPDB database (48)]. Finally, one of the most important additions to the Proteome-*pI* database was introducing *in silico* proteome digests derived from the five most commonly used proteases. Furthermore, the resulting datasets were categorized by molecular mass to facilitate analysis with specific mass spectrometry techniques. Such an approach could be seen as highly simplistic, and further grinding of *in silico* digests is possible. Future plans in this respect include adding the prediction of peptides' hydrophobicity, retention time (49), electrophoretic mobility (50), and the use of more sophisticated methods than can be utilized for the prediction of *in silico* digests [e.g. DeepDigest (51)]. Finally, adding information about the uniqueness of peptides versus coverage after digestion would be also valuable. We would be grateful for any contribution or ideas from the community with respect to future improvements to the database.

DATA AVAILABILITY

All data in the Proteome-*pI* 2.0 database are available for download free of charge. For more information see Supple-

mentary Data. The database will be maintained for at least 10 years and can be accessed at http://isoelectricpointdb2.org or http://isoelectricpointdb2.mimuw.edu.pl (mirror).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

I would like to thank all authors of the previous works related to isoelectric point and pK_a set measurements and computational methods. Special acknowledgement is extended to the developers of the UniProt database, upon which Proteome-*pI* depends heavily.

FUNDING

National Science Centre, Poland [2018/29/B/NZ2/01403]. Funding for open access charge: National Science Centre, Poland [2018/29/B/NZ2/01403].
Conflict of interest statement. None declared.

REFERENCES

1. Pace,C.N., Grimsley,G.R. and Scholtz,J.M. (2009) Protein ionizable groups: pK values and their contribution to protein stability and solubility. *J. Biol. Chem.*, **284**, 13285–13289.
2. Po,H.N. and Senozan,N.M. (2001) The Henderson-Hasselbalch equation: its history and limitations. *J. Chem. Educ.*, **78**, 1499.
3. Klose,J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. *Humangenetik*, **26**, 231–243.
4. O'Farrell,P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.*, **250**, 4007–4021.
5. Kirkwood,J., Hargreaves,D., O'Keefe,S. and Wilson,J. (2015) Using isoelectric point to determine the pH for initial protein crystallization trials. *Bioinformatics*, **31**, 1444–1451.
6. Zhu,M., Rodriguez,R. and Wehr,T. (1991) Optimizing separation parameters in capillary isoelectric focusing. *J. Chromatogr. A*, **559**, 479–488.
7. Branca,R.M., Orre,L.M., Johansson,H.J., Granholm,V., Huss,M., Pérez-Bercoff,Å., Forshed,J., Käll,L. and Lehtiö,J. (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods*, **11**, 59.
8. Cologna,S.M., Russell,W.K., Lim,P.J., Vigh,G. and Russell,D.H. (2010) Combining isoelectric point-based fractionation, liquid chromatography and mass spectrometry to improve peptide detection and protein identification. *J. Am. Soc. Mass Spectrom.*, **21**, 1612–1619.
9. Pahari,S., Sun,L. and Alexov,E. (2019) PKAD: a database of experimentally measured pKa values of ionizable groups in proteins. *Database (Oxford)*, **2019**, baz024.
10. Hoogland,C., Mostaguir,K., Sanchez,J.-C., Hochstrasser,D.F. and Appel,R.D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, **4**, 2352–2356.
11. Kozlowski,L.P. (2017) Proteome-pI: proteome isoelectric point database. *Nucleic Acids Res.*, **45**, D1112–D1116.
12. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
13. SIB Swiss Institute of Bioinformatics Members (2016) The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Res.*, **44**, D27–D37.
14. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
15. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
16. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M. *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
17. Patrickios,C.S. and Yamasaki,E.N. (1995) Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory. *Anal. Biochem.*, **231**, 82–91.
18. Graham Solomons,T.W., Fryhle,C.B. and Snyder,S.A. (2017) *Solomons' Organic Chemistry*. 12th edn, global edition, Wiley Wiley.com.
19. Nelson,D.L. and Cox,M.M. (2017) *Lehninger Principles of Biochemistry*. 7th edn, Macmillan Learning for Instructors.
20. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
21. Dawson,R.M.C., Elliott,D.C., Elliott,W.H. and Jones,K.M. (1987) *Data for Biochemical Research*. 3rd edn., Wiley, p. 97.
22. Toseland,C.P., McSparron,H., Davies,M.N. and Flower,D.R. (2006) PPD v1.0–an integrated, web-accessible database of experimentally determined protein pKa values. *Nucleic. Acids. Res.*, **34**, D199–D203.
23. Sillero,A. and Ribeiro,J.M. (1989) Isoelectric points of proteins: theoretical determination. *Anal. Biochem.*, **179**, 319–325.
24. Thurlkill,R.L., Grimsley,G.R., Scholtz,J.M. and Pace,C.N. (2006) pK values of the ionizable groups of proteins. *Protein Sci.*, **15**, 1214–1218.
25. Rodwell,J.D. (1982) Heterogeneity of component bands in isoelectric focusing patterns. *Anal. Biochem.*, **119**, 440–449.
26. Tabb,D.L., McDonald,W.H. and Yates,J.R. (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.*, **1**, 21–26.
27. Nozaki,Y. and Tanford,C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.*, **246**, 2211–2217.
28. Grimsley,G.R., Scholtz,J.M. and Pace,C.N. (2009) A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci.*, **18**, 247–251.
29. Bjellqvist,B., Basse,B., Olsen,E. and Celis,J.E. (1994) Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis*, **15**, 529–539.
30. Wilkins,M.R., Gasteiger,E., Bairoch,A., Sanchez,J.C., Williams,K.L., Appel,R.D. and Hochstrasser,D.F. (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.*, **112**, 531–552.
31. Halligan,B.D., Ruotti,V., Jin,W., Laffoon,S., Twigger,S.N. and Dratz,E.A. (2004) ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels. *Nucleic Acids Res.*, **32**, W638–W644.
32. Kozlowski,L.P. (2016) IPC - Isoelectric Point Calculator. *Biol. Direct*, **11**, 55.
33. Kozlowski,L.P. (2021) IPC 2.0: prediction of isoelectric point and pKa dissociation constants. *Nucleic Acids Res.*, **49**, W285–W292.
34. Pahari,S., Sun,L., Basu,S. and Alexov,E. (2018) DelPhiPKa: Including salt in the calculations and enabling polar residues to titrate. *Proteins Struct. Funct. Bioinf.*, **86**, 1277–1283.
35. Anandakrishnan,R., Aguilar,B. and Onufriev,A.V. (2012) H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.*, **40**, W537–W541.
36. Song,Y., Mao,J. and Gunner,M.R. (2009) MCCE2: improving protein pKa calculations with extensive side chain rotamer sampling. *J. Comput. Chem.*, **30**, 2231–2247.
37. Reis,P.B.P.S., Vila-Viçosa,D., Rocchia,W. and Machuqueiro,M. (2020) PypKa: a flexible Python module for Poisson–Boltzmann-based pKa calculations. *J. Chem. Inf. Model.*, **60**, 4442–4448.
38. Giansanti,P., Tsiatsiani,L., Low,T.Y. and Heck,A.J.R. (2016) Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nat. Protoc.*, **11**, 993–1006.
39. Maillet,N. (2020) Rapid Peptides Generator: fast and efficient in silico protein digestion. *NAR Genomics Bioinformatics*, **2**, lqz004.

40. Mohanta,T.K., Khan,A., Hashem,A., Abd_Allah,E.F. and Al-Harrasi,A. (2019) The molecular mass and isoelectric point of plant proteomes. *BMC Genomics*, **20**, 631.
41. Mohanta,T.K., Mishra,A.K., Khan,A., Hashem,A., Abd-Allah,E.F. and Al-Harrasi,A. (2021) Virtual 2-D map of the fungal proteome. *Sci. Rep.*, **11**, 6676.
42. Chasapis,C.T. and Konstantinoudis,G. (2020) Protein isoelectric point distribution in the interactomes across the domains of life. *Biophys. Chem.*, **256**, 106269.
43. Nurk,S., Koren,S., Rhie,A., Rautiainen,M., Bzikadze,A.V., Mikheenko,A., Vollger,M.R., Altemose,N., Uralsky,L., Gershman,A. *et al.* (2021) The complete sequence of a human genome. bioRxiv doi: https://doi.org/10.1101/2021.05.26.445798, 27 May 2021, preprint: not peer reviewed.
44. Scheller,C., Krebs,F., Minkner,R., Astner,I., Gil-Moles,M. and Wätzig,H. (2020) Physicochemical properties of SARS-CoV-2 for drug targeting, virus inactivation and attenuation, vaccine formulation and quality control. *Electrophoresis*, **41**, 1137–1151.
45. Kilambi,K.P. and Gray,J.J. (2012) Rapid calculation of protein pKa values using Rosetta. *Biophys. J.*, **103**, 587–595.
46. Bienert,S., Waterhouse,A., de Beer,T.A.P., Tauriello,G., Studer,G., Bordoli,L. and Schwede,T. (2017) The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res.*, **45**, D313–D319.
47. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
48. Reis,P.B.P.S., Clevert,D.-A. and Machuqueiro,M. (2021) pKPDB: a protein data bank extension database of pKa and pI theoretical values. *Bioinformatics*, btab518.
49. Spicer,V., Yamchuk,A., Cortens,J., Sousa,S., Ens,W., Standing,K.G., Wilkins,J.A. and Krokhin,O.V. (2007) Sequence-specific retention calculator. a family of peptide retention time prediction algorithms in reversed-phase HPLC: applicability to various chromatographic conditions and columns. *Anal. Chem.*, **79**, 8762–8768.
50. Chen,D., Lubeckyj,R.A., Yang,Z., McCool,E.N., Shen,X., Wang,Q., Xu,T. and Sun,L. (2020) Predicting electrophoretic mobility of proteoforms for large-scale top-down proteomics. *Anal. Chem.*, **92**, 3503–3507.
51. Yang,J., Gao,Z., Ren,X., Sheng,J., Xu,P., Chang,C. and Fu,Y. (2021) DeepDigest: prediction of protein proteolytic digestion with deep learning. *Anal. Chem.*, **93**, 6094–6103.