Nucleic Acids Research Advance Access published October 26, 2016

doi: 10.1093/nar/gkw978

Proteome-*pI*: proteome isoelectric point database

Lukasz P. Kozlowski*

Quantitative and Computational Biology Group, Max Planck Institute for Biophysical Chemistry, Göttingen, Lower Saxony, 37077, Germany

Received August 19, 2016; Revised October 07, 2016; Editorial Decision October 10, 2016; Accepted October 13, 2016

ABSTRACT

Proteome-*pI* is an online database containing information about predicted isoelectric points for 5029 proteomes calculated using 18 methods. 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, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing, liquid chromatography–mass spectrometry and X-ray protein crystallography. The database, available at http://isoelectricpointdb.org allows the retrieval of virtual 2D-PAGE plots and the development of customised fractions of proteome based on isoelectric point and molecular weight. Moreover, Proteome-*pI* facilitates statistical comparisons of the various prediction methods as well as biological investigation of protein isoelectric point space in all kingdoms of life. For instance, using Proteome-*pI* data, it is clear that Eukaryotes, which evolved tight control of homeostasis, encode proteins with *pI* values near the cell pH. In contrast, Archaea living frequently in extreme environments can possess proteins with a wide range of isoelectric points. The database includes various statistics and tools for interactive browsing, searching and sorting. Apart from data for individual proteomes, datasets corresponding to major protein databases such as UniProtKB/TrEMBL and the NCBI non-redundant (*nr*) database have also been precalculated and made available in CSV format.

INTRODUCTION

Isoelectric point (*pI*) is the pH at which a particular molecule carries no net electrical charge. For polypeptide chains, *pI* depends primarily on the dissociation constants (pK_a) for the ionisable groups of the seven charged amino acids: glutamate, aspartate, cysteine, tyrosine, histidine, lysine and arginine. Moreover, the charge of the terminal groups (NH_2 and COOH) can affect the *pI* of short peptides. It is also important to consider posttranslational modifications, the exposure of charged residues to solvent, the Born effect (dehydration), hydrogen bonds (charge-dipole interactions), and charge-charge interactions (1). *pI* has broad usage in currently used biochemical and proteomic techniques. For example, during electrophoresis, the direction of protein migration on the gel depends on the charge. Thus, it is possible to separate proteins in a gel based on their *pI*. Given the sequence, we can try to computationally predict *pI* using the Henderson–Hasselbalch equation (2), by calculating the charge of the molecule at a certain pH using pK_a values of charged residues. More than 600 different pK_a values have so far been reported for the ionisable groups of amino acids (3). The final result, predicted *pI,* will most likely be different than the real one, given that many proteins are chemically modified (e.g., amino acids can be phosphorylated, methylated, or acetylated), and this can influence their charge. Nevertheless, even an approximate isoelectric point is a highly valuable and frequently used parameter.

In the past, much work has gone into creating databases storing experimentally verified *pI* values for proteins, yet none of these databases contains more than five thousand proteins (4,5), which is very few compared to the protein sequence data currently available. Thus, Proteome-*pI* database is an attempt to decrease this gap; hopefully, it will expand the body of knowledge regarding isoelectric points in a more genome-wide fashion.

MATERIALS AND METHODS

Sequences

Protein sequences of model organisms were obtained from UniProt as of April 2016, release 2016_04 (6). This includes 5029 complete proteomes (with splicing isoforms for Eukaryote) from the entire tree of life. In total, protein isoelectric point, molecular weights and other statistics were calculated for >21 million protein sequences (Table 1).

Predictions

To predict isoelectric points, Proteome-*pI* currently uses 18 different algorithms and programs, which can be divided into three categories. The first category consists of methods

*To whom correspondence should be addressed. Tel: +49 551 201 2895; Fax: +49 551 201 2803; Email: lukasz.kozlowski.lpk@gmail.com

Table 1. General statistics of the Proteome-*pI* database

	Number of proteomes	Total number of proteins	Mean number of proteins (±SD)	Mean size of proteins (±SD)	Mean mw of proteins (±SD)
Viruses	504	20 920	42 ± 89	297 ± 375	33 ± 42
Archaea	135	318 388	2358 ± 920	283 ± 212	31 ± 23
Bacteria	3776	12 082 903	3200 ± 2510	311 ± 240	34 ± 26
Eukaryote	614	9 299 039	15 145 ± 11 830	438 ± 429	49 ± 48
Eukaryote (major)	614	8 629 591	14 055 ± 9899	434 ± 416	48 ± 46
Eukaryote (minor)	448	669 448	1494 ± 5130	495 ± 564	55 ± 63

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

that predict the isoelectric point based on the Henderson–Hasselbalch equation with different pK_a values corresponding to different charged groups (2). Those methods usually use nine different pK_a values established empirically in separate experiments (seven pK_a values for charged amino acids and two for polypeptide chain termini). For example, pK_a values obtained by Thurlkill *et al.* were measured in 0.1 M KCl at 25°C using alanine pentapeptides with a charged residue in the centre and with blocked terminal groups (7). Further, nine-parameter models are used for calculation of isoelectric points in methods named after the lead author of the study or the source of the pK_a values: EMBOSS (8), DTASelect (9), Solomons (10), Sillero (11), Rodwell (12), Wikipedia, Lehninger (13), Grimsley (3), Toseland (14), Thurlkill (7), Nozaki (15) and Dawson (16). Additionally some algorithms use different numbers of pK_a values (Patrickios (17) uses only six, Bjellqvist (18) uses 17, and ProMoST (19) uses 72 pK_a values depending on the location of amino acid with respect to the protein termini). In the next category, we have IPC_protein and IPC_peptide models, which use computationally optimised nine-parameter pK_a sets (20). Finally, the consensus from all methods apart from Patrickios (highly simplified model with only six parameters) is also reported.

RESULTS

Database use

The Proteome-*pI* database incorporates multiple browsing and searching tools. First, it can be searched and browsed by organism name, average isoelectric point, molecular weight or amino acid frequencies (see also Table 2). Proteins with extreme *pI* values are also available. For individual proteomes, users can retrieve proteins of interest given the method, isoelectric point and molecular weight ranges (this particular feature can be highly useful to limit potential targets in analysis of 2D-PAGE gels or before conducting mass spectrometry). Additionally, precalculated fractions of proteins according to isoelectric point are also available. Finally, some general statistics (total number of proteins, amino acids, average sequence length, amino acid frequency) and links to other databases (UniProt, NCBI) can be found (see Figure 1 for an example).

Moreover, apart from the data for individual proteomes, one can also obtain precalculated isoelectric points from all major protein databases, including *nr* (21), UniProt, PDB (22) and SwissProt (23) (more details in Supplementary Data).

DISCUSSION

The main content of the Proteome-*pI* database is the comprehensive isoelectric point prediction using numerous methods. 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, such as 2D-PAGE gel electrophoresis (24,25), capillary isoelectric focusing (26), liquid chromatography–mass spectrometry (LC–MS) (27) and X-ray protein crystallography (28,29). Additional goals of the database include facilitating biological investigation of protein isoelectric point space. For instance, it is well known that distribution of protein isoelectric points of proteomes is bimodal, with a low fraction of proteins having *pI* values close to the cell physiological pH (Supplementary Figure S1) (30). Interestingly, if we divide proteomes into the kingdoms of life, one can notice that Eukaryota have the largest proteins restricted to narrow isoelectric point range. On the other side, Archaea possess usually small proteins, but the isoelectric points of their proteins can vary significantly (Figure 2). This is most likely due to the adaptation to the extreme conditions in which many Archaea live (31). Finally, viruses form a completely separate group. Their proteins have isoelectric point which is strongly correlated with the *pI* of its host proteins and therefore can vary significantly. Simultaneously, the molecular weight of viral proteins is significantly lower than that of host proteins due to the compactness of virions (significant evolutionary pressure to minimise the overall size) (32).

It should be noted that there is at least one other similar database storing isoelectric points for some proteomes. The JVirGel website (33) contains*pI* data for 227 relatively small, prokaryotic proteomes, precalculated using only one method. In contrast, the Proteome-*pI* database aggregates predictions of isoelectric points calculated by 18 different methods and algorithms across >5000 proteomes from all kingdoms of life (over 21 million proteins).

Future work

The principal future goal is to include more isoelectric point algorithms and proteomes for further investigation. The next future goal is to provide more tools for online analysis, e.g., tools for Gene Ontology searching (34). Another possible extension could be to add putative digestion products of trypsin and their respective isoelectric points (35). We will be grateful for any contribution to the database from the community.

Figure 1. Proteome-*pI* example report for *Salmonella enterica*. At the top, the average isoelectric point, precalculated fractions of proteins according to isoelectric point and virtual 2D-PAGE plot for the proteome are shown. In the next section, the user can retrieve a subset of proteins within specified isoelectric point and molecular weight ranges calculated using a particular method. Next, proteins with minimal and maximal isoelectric points are presented along with some general statistics.

Table 2. Amino acid frequency for the kingdoms of life in the Proteome-*pI* 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	6.61	1.76	5.81	6.04	4.25	5.79	2.15	6.53	6.35	8.84	2.46	5.41	4.62	3.39	5.24	7.06	6.06	6.50	1.19	3.94	6 150 189
Archaea	8.20	0.98	6.21	7.69	3.86	7.58	1.77	7.03	5.27	9.31	2.35	3.68	4.26	2.38	5.51	6.17	5.44	7.80	1.03	3.45	89 488 664
Bacteria	10.06	0.94	5.59	6.15	3.89	7.76	2.06	5.89	4.68	10.09	2.38	3.58	4.61	3.58	5.88	5.85	5.52	7.27	1.27	2.94	3 716 982 916
Eukaryota	7.63	1.76	5.40	6.42	3.87	6.33	2.44	5.10	5.64	9.29	2.25	4.28	5.41	4.21	5.71	8.34	5.56	6.20	1.24	2.87	3 743 221 293
All	8.76	1.38	5.49	6.32	3.87	7.03	2.26	5.49	5.19	9.68	2.32	3.93	5.02	3.90	5.78	7.14	5.53	6.73	1.25	2.91	7 555 843 062

*Similar statistics for all 5029 proteomes included in Proteome-*pI* are available online on individual subpages. For di-amino acid frequencies see Supplementary Table S2.

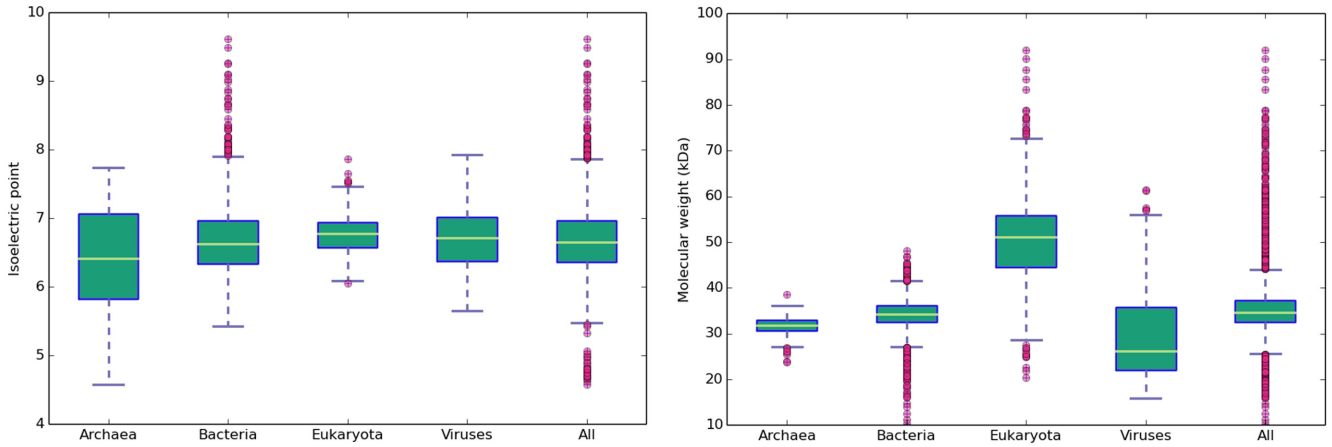

Figure 2. Isoelectric points and molecular weights across kingdoms of life. Data for the proteomes of 135 Archaea, 127 viruses (>50 proteins), 3775 bacteria and 614 eukaryotes.

AVAILABILITY

All data in the Proteome-*pI* database are available for download free of charge. Proteome-*pI* can be accessed at http://isoelectricpointdb.org The database will be available at given web address for at least ten years.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Proteome-*pI* author acknowledges all authors of previous works related to different pK_a sets and databases, especially developers of UniProt database.

FUNDING

Funding for open access charge: The open access publication charge for this paper has been waived by Oxford University Press—*NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.
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. 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.
4. Hoogland,C., Mostaguir,K., Sanchez,J.C., Hochstrasser,D.F. and Appel,R.D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, **4**, 2352–2356.
5. Bunkute,E., Cummins,C., Crofts,F.J., Bunce,G., Nabney,I.T. and Flower,D.R. (2015) PIP-DB: the protein isoelectric point database. *Bioinformatics*, **31**, 295–296.
6. The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212.
7. 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.
8. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277.
9. 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.
10. Solomons,T.G. (1992) *Organic Chemistry*. John Wiley&Sons.
11. Sillero,A. and Ribeiro,J.M. (1989) Isoelectric points of proteins: theoretical determination. *Anal. Biochem.*, **179**, 319–325.
12. Rodwell,J.D. (1982) Heterogeneity of component bands in isoelectric focusing patterns. *Anal. Biochem.*, **119**, 440–449.
13. Nelson,D.L., Lehninger,A.L. and Cox,M.M. (2008) *Lehninger Principles of Biochemistry*. Macmillan.
14. 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 pK(a) values. *Nucleic Acids Res.*, **34**, D199–D203.
15. Nozaki,Y. and Tanford,C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions: estabilishment of a hydrophobicity scale. *J. Biol. Chem.*, **246**, 2211–2217.
16. Dawson,R.M.C. (1986) *Data for Biochemical Research*. Clarendon Press, Oxford.
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. 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.

19. 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.
20. Kozlowski,L.P. (2016) IPC - Isoelectric Point Calculator. *Biol. Direct*, **11**, 55.
21. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
22. Rose,P.W., Prlić,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D. and Woo,J. (2015) The RCSB protein data bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
23. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Plant Bioinformatics: Methods Protocols*, 23–54.
24. O'Farrell,P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.*, **250**, 4007–4021.
25. Klose,J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. *Humangenetik*, **26**, 231–243.
26. Righetti,P.G., Castagna,A., Herbert,B., Reymond,F. and Rossier,J.S. (2003) Prefractionation techniques in proteome analysis. *Proteomics*, **3**, 1397–1407.
27. Heller,M., Ye,M., Michel,P.E., Morier,P., Stalder,D., Jünger,M.A., Aebersold,R., Reymond,F. and Rossier,J.S. (2005) Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.*, **4**, 2273–2282.
28. 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.
29. Kantardjieff,K.A. and Rupp,B. (2004) Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics*, **20**, 2162–2168.
30. Kiraga,J., Mackiewicz,P., Mackiewicz,D., Kowalczuk,M., Biecek,P., Polak,N., Smolarczyk,K., Dudek,M.R. and Cebrat,S. (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics*, **8**, 163.
31. Oren,A. (2008) Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Syst.*, **4**, 13.
32. Grenfell,B.T., Pybus,O.G., Gog,J.R., Wood,J.L., Daly,J.M., Mumford,J.A. and Holmes,E.C. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**, 327–332.
33. Hiller,K., Grote,A., Maneck,M., Münch,R. and Jahn,D. (2006) JVirGel 2.0: computational prediction of proteomes separated via two-dimensional gel electrophoresis under consideration of membrane and secreted proteins. *Bioinformatics*, **22**, 2441–2443.
34. The Gene Ontology Consortium (2015) Gene ontology consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
35. Shevchenko,A., Tomas,H., Havli,J., Olsen,J.V. and Mann,M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.*, **1**, 2856–2860.