BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 2 2015, pages 295–296
doi:10.1093/bioinformatics/btu637

Databases and ontologies Advance Access publication September 23, 2014

PIP-DB: the Protein Isoelectric Point database

Egle Bunkute[1], Christopher Cummins[2], Fraser J. Crofts[1], Gareth Bunce[1], Ian T. Nabney[2] and Darren R. Flower[1,*]

[1]School of Life and Health Sciences and [2]School of Engineering and Applied Science, University of Aston, Aston Triangle, Birmingham B4 7ET, UK

Associate Editor: Jonathan Wren

ABSTRACT

Summary: A protein's isoelectric point or pI corresponds to the solution pH at which its net surface charge is zero. Since the early days of solution biochemistry, the pI has been recorded and reported, and thus literature reports of pI abound. The Protein Isoelectric Point database (PIP-DB) has collected and collated these data to provide an increasingly comprehensive database for comparison and benchmarking purposes. A web application has been developed to warehouse this database and provide public access to this unique resource. PIP-DB is a web-enabled SQL database with an HTML GUI front-end. PIP-DB is fully searchable across a range of properties.

Availability and implementation: The PIP-DB database and documentation are available at http://www.pip-db.org.

Contact: d.r.flower@aston.ac.uk

Received on July 1, 2014; revised on September 8, 2014; accepted on September 18, 2014

1 INTRODUCTION

For a macromolecular polyprotic system—such as protein, DNA or RNA—the isoelectric or isoionic point—commonly referred to as the isoelectric point (pI)—can be defined by the point of singularity in a titration curve, corresponding to the solution pH value at which the net overall surface charge, and thus, the mobility, of the ampholyte sums to zero (Maldonado *et al.*, 2010). Protein pIs can be determined in several ways, but are generally determined using either polyacrylamide gel-based isoelectric focusing (or IEF) or capillary IEF (or cIEF) (Silvertand *et al.*, 2008; Righetti *et al.*, 2013). Separation by pI is a key component of 2D gel electrophoresis, a key precursor of proteomics, where discrete spots can be digested in-gel, and proteins subsequently identified by analytical mass spectrometry (Mauri and Scigelova, 2009). Analysis of whole proteomes indicate that at the system level, pIs exhibit a multimodal distribution indicative of significant phylogenetic constraints on surface charge (Wu *et al.*, 2006).

Theoretical calculation of pIs can aid proteome analysis: assuming the protein to be denatured, calculation is rapid, requiring only the sequence to be known (Cargile *et al.* 2004). Most techniques exploit tabulated values for pKa values for ionizable amino acid residues (Sillero and Ribeiro, 1989), which are assumed constant irrespective of structural context (Maldonado *et al.*, 2010).

A protein's pI is one of the most comprehensively determined and widely reported characteristic quantities in biochemistry and proteomics. However, such reports are typically almost incidental within the wider characterization of a protein or proteins. Thus far, no dedicated, web-accessible database of protein pI values has been made available. Thus, we describe here the Protein pI Database (PIP-DB), our ongoing attempt to catalogue comprehensively the pIs of proteins, as reported in the literature.

2 METHODS AND USAGE

2.1 Data acquisition

Protein pI data were collected through scrutiny of the primary scientific literature. Data acquisition progressed in two stages. In the first rapid stage, two early reviews (Righetti and Caravaggio, 1976; Righetti *et al.*, 1981) were interrogated supplying a core of information (pI, protein identity, *etc.*), which we supplemented by parsing online resources, primarily Brenda (Schomburg *et al.*, 2013). Using this initial limited dataset, we constructed a prototype PIP-DB. In the second, slow stage, we undertook quasi-exhaustive literature searches, using a variety of search terms within PUBMED, ISI Web of Knowledge, Scopus and Quertle, to identify papers containing pI data, together with prospective and retrospective searching using citation lookup. In all instances, where possible, we referred back to the original article to extract all archived data items, which we then supplemented as described below.

2.2 Data content

PIP-DB contains 5773 protein entries, each associated with either a single pI value or pI range. Each protein entry is linked to additional associated data: experimental data (stored within PIP-DB) and cross references to external data sources. Experimental data include, where available, temperature, method of analysis (IEF, cIEF, *etc.*), total measured molecular weight (MW), number of subunits, subunit MW, Enzyme Commission (EC) number, source organism and cellular and/or tissue location. Cross references include links to the NCBI Taxonomy browser (Federhen, 2012) and literature citations to PUBMED abstracts, publisher abstracts in lieu of PUBMED entries, and, where available, to full texts at publisher's websites. For just under one half of entries (2822 of 5773), PIP-DB also records the protein sequence, as abstracted from UniProt (UniProt Consortium, 2014) or NCBI (NCBI Resource Coordinators, 2014).

*To whom correspondence should be addressed.

2.3 Web implementation

PIP-DB has been implemented as a polylingual web-enabled database system, including source code written in Clojure LISP, JavaScript, Less CSS, M4sh, Make, Python and sh programming languages. Additional documentation is formatted in LaTeX, HTML and Markdown.

PIP-DB can be searched by keywords such as protein name, source organism, experimental method, and cellular and/or tissue location and by numerical properties such as pI, EC number, temperature and MW. A search engine and domain-specific language has been designed, which enables searching of PIP-DB by representing compound queries using tree structures in LISP.

Importantly, PIP-DB can also be searched using global sequence similarity, as implemented in BLAST (Altschul *et al.*, 1990). As much of the data we archive have a provenance, which is legacy in nature, often not linked unambiguously to explicit sequences, we have generally been conservative in our assignment of corresponding sequence data, with 2822 sequences of the 5773 entries.

Ease of use was a priority in constructing PIP-DB, with a particular emphasis on creating an intuitive search engine. Navigation through the website is facilitated by the tiered display of information: encompassing titles, then summary, to full entry, with links to further information.

3 DISCUSSION AND CONCLUSION

In line with our previous database generation exercises (Blythe *et al.* 2002; McSparron *et al.* 2004; Toseland *et al.*, 2005a,b; Toseland *et al.*, 2006; Ansari *et al.* 2010), we extract and record data as described in the original report, without making arbitrary changes to it. As it is not possible for logistic reasons to retest each pI value, we must trust these values are accurate. Context in the form of associated experimental data allows users to draw their own conclusions regarding data veracity.

Researchers from many disciplines can potentially benefit from PIP-DB: such as virtual gel methodology and theoretical IEF, as well as functioning as a tool in its own right to facilitate study of the physical chemistry of proteins. In particular, predictive pI calculation needs benchmarking, as the veracity of pI predictors is assumed and not yet proven. What explicit comparisons there have been, and they are few in number, have used small datasets (Patrickios and Yamasaki, 1995), datasets of peptides rather than proteins (Lengqvist *et al.*, 2011) or report poor accuracy (Henriksson *et al.*, 1995; Patrickios and Yamasaki, 1995). Thus, the sequence search feature of PIP-DB will provide an alternative of true utility to extant predictors and a gold-standard benchmarking resource for such work.

In summary, PIP-DB is a dedicated, manually-curated fully searchable database presently containing >5500 measured pI values, and associated data of many types. As new studies appear, and legacy data are continually polled, additional information will be integrated into PIP-DB. Moreover, we will extend the scope of the PIP-DB to include peptides, nucleotides and viruses (Subirats *et al.*, 2011).

ACKNOWLEDGEMENTS

The authors thank Sukhvir Johal, Anees Ul-haq, Saleh Ahmed and Mohammad Abdullah for their help.

Funding: This work was supported by Aston University.

Conflict of interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ansari,H.R. *et al.* (2010) AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res.*, **38**, D847–D853.

Blythe,M.J. *et al.* (2002) JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics*, **18**, 434–439.

Cargile,B.J. *et al.* (2004) Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J. Proteome Res.*, **3**, 112–119.

Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.

Henriksson,G. *et al.* (1995) Calculation of the isoelectric points of native proteins with spreading of pKa values. *Electrophoresis*, **16**, 1377–1380.

Lengqvist,J. *et al.* (2011) Observed peptide pI and retention time shifts as a result of post-translational modifications in multidimensional separations using narrow-range IPG-IEF. *Amino Acids*, **40**, 697–711.

Maldonado,A.A. *et al.* (2010) Isoelectric point, electric charge, and nomenclature of the acid-base residues of proteins. *Biochem. Mol. Biol. Educ.*, **38**, 230–237.

Mauri,P. and Scigelova,M. (2009) Multidimensional protein identification technology for clinical proteomic analysis. *Clin. Chem. Lab. Med.*, **47**, 636–646.

McSparron,H. *et al.* (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci.*, **43**, 1276–1287.

NCBI Resource Coordinators. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.

Patrickios,C.S. and Yamasaki,E.N. (1995) Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory. *Anal. Biochem.*, **23**, 82–91.

Righetti,P.G. and Caravaggio,T. (1976) Isoelectric points and molecular weights of proteins. *J. Chromatogr.*, **127**, 1–28.

Righetti,P.G. *et al.* (1981) Isoelectric points and molecular weights of proteins. A new table. *J. Chromatogr.*, **220**, 115–194.

Righetti,P.G. *et al.* (2013) Capillary electrophoresis and isoelectric focusing in peptide and protein analysis. *Proteomics*, **13**, 325–340.

Schomburg,I. *et al.* (2013) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.*, **41**, D764–D772.

Sillero,A. and Ribeiro,J.M. (1989) Isoelectric points of proteins: theoretical determination. *Anal. Biochem.*, **179**, 319–325.

Silvertand,L.H. *et al.* (2008) Recent developments in capillary isoelectric focusing. *J. Chromatogr. A*, **1204**, 157–170.

Subirats,X. *et al.* (2011) Recent developments in capillary and chip electrophoresis of bioparticles: viruses, organelles, and cells. *Electrophoresis*, **32**, 1579–1590.

Toseland,C.P. *et al.* (2005a) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.*, **1**, 4.

Toseland,C.P. *et al.* (2005b) DSD—an integrated, web-accessible database of Dehydrogenase Enzyme Stereospecificities. *BMC Bioinformatics*, **6**, 283.

Toseland,C.P. *et al.* (2006) PPD v1.0—an integrated, web-accessible database of experimentally determined protein pKa values. *Nucleic Acids Res.*, **34**, D199–D203.

UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.

Wu,S. *et al.* (2006) Multi-modality of pI distribution in whole proteome. *Proteomics*, **6**, 449–455.