ABSTRACT
Summary: A protein’s isoelectric point or pI corresponds to the solution pH at which its net surface charge is zero. Since the earliest 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 this legacy 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. PIPD is a web-enabled SQL database with an html GUI front-end. PIPD is fully searchable across a range of characteristics.
Availability and documentation are available at: http://www.pip-db.org.
Contact: d.r.flower@aston.ac.uk
Supplementary information: none

1 INTRODUCTION
For a macromolecular polyprotic system - such as a protein, DNA, or RNA - the isoelectric or isoionic point – commonly referred to as the 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 pI can be determined in several ways, but are generally determined using either polyacrylamide gel-based isoelectric focusing (or IEF) or capillary isoelectric focusing (or cIEF) (Silvertand et al., 2008; Righetti et al., 2013).
Separation by pI is a key component of 2-D gel electrophoresis, a key precursor of proteomics, where discrete spots can be digested in-gel, and proteins subsequently identified by analytical mass spectrometry (Mauri et al., 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). Proteomic analysis is aided by the theoretical calculation of pIs: assuming the protein is denatured, calculation is rapid, requiring only the sequence to be known. Most techniques exploit tabulated values for pKa values for ionisable amino acid residues, which are assumed constant irrespective of context.

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. Here, we describe the Protein pI Database (PIP-DB), our attempt to comprehensively catalogue the isoelectric points of proteins, as reported in the literature.

2 METHODS AND USAGE
2.1 Data Acquisition.
Protein data was collected through scrutiny of the primary scientific literature. Two early reviews (Righetti and Caravaggio, 1976; Righetti et al., 1981) identified a core of information upon which we subsequently built PIP-DB, by cross-referencing certain on-line resources, primarily Brenda (Schomburg et al., 2013), and by undertaking quasi-exhaustive retrospective literature searches using a variety of keywords and search terms, together with more limited prospective and retrospective citation searching.

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 with PIP-DB) and cross-references to external data sources. Experimental data includes, 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 website. For approximately a fifth of entries, PIP_DB also records the protein sequence, as abstracted from UniProt (UniProt Consortium, 2014) or NCBI (NCBI Resource Coordinators, 2014).

2.3 Web Implementation.
PIP-DB has been implemented as a web-enabled database system. Development of PIP-DB was a polylingual project including source code written in Clojure LISP, JavaScript, Less CSS, M4sh, Make, Python, and sh programming languages. The documentation is formatted in LATEX, HTML and Markdown.
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. Since it is not possible for logistic reasons to retest each pI value, we must take it on trust that these values are correct. PIP-DB is a dedicated, manually-curated, fully-searchable database currently containing over 5500 measured pI values, and associated data of several types. As new studies appear, and legacy data is continually polled, additional information will be integrated into PIP-DB. Moreover, we will also extend the scope PIP-DB to include peptides, nucleotides, and viruses (Subirats et al., 2011). Researchers from many disciplines can potentially benefit from this data resource. For example, for benchmarking predictive pI calculation, virtual gel methodology, and isoelectric focusing. As well as a tool in its own right to facilitate study of the physical chemistry of proteins, particularly as an aid to analyzing proteomes.

ACKNOWLEDGEMENTS

We 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