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).
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 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