Bioinformatics, 32(17), 2016, 2719–2721
doi: 10.1093/bioinformatics/btw170
Advance Access Publication Date: 3 April 2016
Applications Note

Databases and ontologies

ProbOnto: ontology and knowledge base of probability distributions

Maciej J. Swat[1,*], Pierre Grenon[2] and Sarala Wimalaratne[1]

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK and [2]CHIME, University College London, London, UK

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 29, 2016; revised on March 23, 2016; accepted on March 24, 2016

Abstract

Motivation: Probability distributions play a central role in mathematical and statistical modelling. The encoding, annotation and exchange of such models could be greatly simplified by a resource providing a common reference for the definition of probability distributions. Although some resources exist, no suitably detailed and complex ontology exists nor any database allowing programmatic access.

Results: ProbOnto, is an ontology-based knowledge base of probability distributions, featuring more than 80 uni- and multivariate distributions with their defining functions, characteristics, relationships and re-parameterization formulas. It can be used for model annotation and facilitates the encoding of distribution-based models, related functions and quantities.

Availability and Implementation: http://probonto.org

Contact: mjswat@ebi.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

When encoding probabilistic uncertainties using a given parametric distribution, its name and parameters are usually sufficient to specify the intended distribution unambiguously, as in many cases such parameter set is unique. However, in multiple cases, two or more parameterizations exist and it becomes essential to specify correctly the parameters for the distribution in order to obtain the correct model. A well-structured, independent, standard reference would greatly facilitate the specification of a distribution and its declaration in a programming language or exchange format as is shown by interoperability issues between tools with distributions differing in their parameterizations (LeBauer *et al.*, 2013).

Many resources are available online (Dinov *et al.*, 2015; Williams *et al.*, 2008), as tools (Marichev and Trott, 2013) and in printed format (Forbes *et al.*, 2011; Johnson *et al.*, 2005; Leemis and Mcqueston, 2008) but a comprehensive ontology which formalizes the theory of kinds and relations of probability distributions does not yet exist.

Some Bioportal (Noy *et al.*, 2009) ontologies provide a simple classification and little information; in most cases parameters, defining functions, or quantities are not defined and information about the relationships between distributions is not available. Such ontologies, because of their simplicity, are insufficient for our purposes and it is the same with existing databases of distributions.

2 Methods

We compiled an initial ‘common denominator’ collection of parametric distributions from UncertML (Williams *et al.*, 2008) and Matlab Statistical Toolbox (MathWorks, 2015) and then extended this set to cover models used in various statistical modelling areas, in particular, pharmacometrics. A number of important and more exotic discrete data models were treated with alternative versions of the Negative Binomial and the Generalized Poisson, or the Conway–Maxwell–Poisson distribution.

We further included all distributions and relevant parameterizations used in the following tools: Monolix (Lixoft Team, 2014), NONMEM (Beal *et al.*, 2009) and winBUGS (Lunn *et al.*, 2009), and a few found in MCSim (Bois, 2010). We completed this collection with many relationships/re-parameterizations between distributions (Leemis and Mcqueston, 2008).

We used the following sources to populate the database with probability functions, their relationships and quantities: Forbes *et al.* (2011), Johnson *et al.* (2005) and the probability distribution pages of the Wikipedia.

The following list summarizes the features of ProbOnto 1.0

- Over 80 uni- and multivariate distributions and alternative parameterizations.
- Probability density or mass functions and available cumulative distribution, hazard and survival functions.
- Supports encoding of univariate mixture distributions.
- Over 130 relationships and re-parameterization formulas.
- Related quantities such as mean, median, mode and variance.
- Parameter and support/range definitions and distribution type.
- Latex and R code for mathematical functions.

LeBauer *et al.* (2013) showed that re-parameterizations between related distributions are essential for the interoperability between existing tools using only two tools and five distributions. ProbOnto extended coverage in this respect beyond proof of concept with a wider scope and number of tools and distributions. ProbOnto contains, for example, six alternative parameterizations of the log-normal or five of the negative binomial distribution, all in use in different contexts. Scientists and tool developers can look all these up in ProbOnto (Supplementary Material).

3 Ontological model

ProbOnto is a knowledge base built from a simple ontological model. At its core, a probability distribution is an instance of the class thereof, a specialization of the class of mathematical objects. A distribution relates to a number of other individuals, which are instances of various categories in the ontology. For example, these are parameters and related functions associated with a given probability distribution. This strategy allows for the rich representation of attributes and relationships between domain objects. The ontology can be seen as a conceptual schema in the domain of mathematics and has been implemented as a PowerLoom (MacGregor *et al.*, 1997) knowledge base. An OWL version is generated programmatically using the Jena API (McBride, 2001). Output for ProbOnto are provided as supplementary materials and published on or linked from the ProbOnto website. The OWL version of ProbOnto is available via Ontology Lookup Service (OLS) to facilitate simple searching and visualization of the content (Jupp *et al.*, 2015). In addition the OLS API provides methods to programmatically access ProbOnto and to integrate it into applications.

4 Use case

ProbOnto was first designed to facilitate the encoding of nonlinear mixed effect models and their annotation in PharmML, Pharmacometrics Markup Language, (Swat *et al.*, 2015) developed by DDMoRe (Harnisch *et al.*, 2013). The scope and features of the language made ProbOnto invaluable in encoding of diverse models applicable to discrete (e.g. count, categorical and time-to-event) and continuous data (Supplementary Material).

Despite its PharmML original context, ProbOnto is purpose independent and does not put implementation constraints on tool designers, thus allowing an open ended number of usage scenarios.

When using ProbOnto with PharmML, a small generic XML schema was enough to allow for flexible encoding of distributions relevant in pharmacometric modelling, their parameters and functions (see Supplementary Material, Appendix). The following example shows how the negative binomial distribution is encoded by using its codename and declaring that of its parameters (‘rate’ and ‘overdispersion’).

```
<Distribution>
  <po:ProbOnto name=“NegativeBinomial2”>
    <po:Parameter name=“rate”>
      <ct:Assign>
        <ct:SymbRef symbIdRef=“lambda”/>
      </ct:Assign>
    </po:Parameter>
    <po:Parameter name=“overdispersion”>
      <ct:Assign>
        <ct:SymbRef symbIdRef=“tau”/>
      </ct:Assign>
    </po:Parameter>
  </po:ProbOnto>
</Distriution>
```

To specify any given distribution unambiguously using ProbOnto, it is sufficient to declare its code name and the code names of its parameters.

5 Future plans and conclusions

ProbOnto provides a means for the encoding and annotation of statistical models, thus facilitating their exchange between software tools. Due to its generic construction which does not enforce a specific implementation in target software, ProbOnto can be applied across various modelling platforms and databases.

Although it already incorporates a high number of distributions, this collection is still growing with new distributions added on regular basis; ProbOnto will ultimately include all distributions supported by tools such as STAN (STAN Development Team, 2015) and R (R Core Team, 2015). In the future, additional features will extend the knowledge base such as applications of a probability distribution, its mathematical properties, such as linear combination, convolution and scaling (Leemis and Mcqueston, 2008) and related data type.

Acknowledgements

We would like to thank R. Bizzotto, F. Bois, V. Chelliah, I.Dinov, H. Hermjakob, J.M. Hilbe, S.Jupp, N. Juty, M. Lavielle, L.M. Leemis for helpful comments and suggestions.

Funding

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115156, resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ in kind contribution. The DDMoRe project is also supported by financial contribution from Academic and SME partners.

Conflict of Interest: none declared.

References

Beal,S.L. *et al.* (2009) NONMEM User’s Guides. (1989–2009). Technical report, Icon Development Solutions, Ellicott City, MD, USA.

Bois,F. (2010) Physiologically based modelling and prediction of drug interactions. *Basic Clin. Pharmacol. Toxicol.*, **106**, 154–161.

Dinov,I.D. *et al*. (2015) Probability Distributome: a web computational infrastructure for exploring the properties, interrelations, and applications of probability distributions. *Comput. Stat*., **15**, 1–19.

Forbes,C. *et al*. (2011) *Statistical Distributions*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Johnson,N.L. *et al*. (2005) *Univariate Discrete Distributions*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Jupp,S. *et al*. (2015) A new Ontology Lookup Service at EMBL-EBI. In: Malone, J. et al. (eds.) *Proceedings of SWAT4LS International Conference 2015*.

Harnisch,L. *et al*. (2013) Drug and disease model resources: a consortium to create standards and tools to enhance model-based drug development. *CPT Pharmacometrics Syst. Pharmacol*., **2**, e34.

LeBauer,D.S. *et al*. (2013) Translating probability density functions: From R to BUGS and back again. *R Journal*, **5**, 207–209.

Leemis,L.M. and Mcqueston,J.T. (2008) Univariate distribution relationships. *Am. Stat*., **62**, 45–53.

Lixoft Team. (2014) Lixoft, Monolix 4.3.

Lunn,D. *et al*. (2009) The BUGS project: Evolution, critique and future directions. *Stat. Med*., **28**, 3049–3067.

MacGregor,R. *et al*. (1997) *Powerloom Manual*. ISI, University of South California, Marina del Rey.

Marichev,O. and Trott,M. (2013) The Ultimate Univariate Probability Distribution Explorer.

MathWorks. (2015) *MATLAB and Statistics Toolbox*. Natick, MA: The MathWorks Inc.

McBride,B. (2001) Jena: Implementing the RDF model and syntax specification. In: *SemWeb*.

Noy,N.F. *et al*. (2009) Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*., **440**, W170–W173.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

STAN Development Team. (2015) Stan Modeling Language Users Guide and Reference Manual, Version 2.9.0.

Swat,M.J. *et al*. (2015) Pharmacometrics Markup Language (PharmML): opening new perspectives for model exchange in drug development. *CPT Pharmacometrics Syst. Pharmacol*., **4**, 316–319.

Williams,M. *et al*. (2008) UncertML: An XML Schema for Exchanging Uncertainty. In: *Proceedings of the GIS Research UK 16th Annual Conference GISRUK* 2008, pp. 275–279.