Modelling threshold phenomena in OWL:
Metabolite concentrations as evidence for disorders

Janna Hastings[1,2⋆], Ludger Jansen[3,4], Christoph Steinbeck[1], and Stefan Schulz[5]

[1] Chemoinformatics and Metabolism, European Bioinformatics Institute, UK
[2] Swiss Centre for Affective Science, University of Geneva, Switzerland
[3] Department of Philosophy, University of Rostock, Germany
[4] Department of Philosophy, RWTH Aachen University, Germany
[5] Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

Abstract. While genomic and proteomic information describe the overall cellular machinery available to an organism, the metabolic profile of an individual at a given time provides a canvas as to the current physiological state. Concentration levels of relevant metabolites vary under different conditions, in particular, in the presence or absence of different disorders. Metabolite concentrations thus mediate an important link between chemistry and biology, contributing to a systems-wide understanding of biological processes and pathways. However, there are a number of challenges in the ontological representation of such information.
Firstly, concentration information is numeric and ranges over continuous values, while ontologies consist of discrete classes. Secondly, ontologies usually model only what is certain, and their logical formalisms are adapted to reasoning from certain axioms to logical deductions, however, the link between chemicals and diseases via concentration levels, like many threshold phenomena, is both uncertain and vague.
In this paper we evaluate the representation of this knowledge using a combination of concrete domains and probabilistic reasoning. We parse concentration values from HMDB and create an ontology able to distinguish normal from abnormal concentrations and able to evaluate a probabilistic risk category for the presence of an associated disorder.

Introduction

Metabolomics is the study of the small molecule products of metabolic processes present in living organisms, called metabolites. Concentration levels of different metabolites in the fluids of the body provide evidence for which processes have taken place, and thereby can reliably indicate disorders [17], as well as providing additional support for functional genomics expression studies [7].

The ChEBI ontology is an ontology of chemical entities and their roles in biological contexts, presently containing around 25,000 classes. 'Metabolite' is

⋆ To whom correspondence should be addressed: hastings@ebi.ac.uk

included in ChEBI as a *role* which chemical entities take in biological contexts. ChEBI does not currently provide information on the differing concentration levels of metabolites, nor on their association with disorders. This information is provided by metabolome databases such as the Human Metabolome Database [18], but these resources are not organised into an ontology, with the disadvantage that they do not allow for automated reasoning and semantic computational processing. It is therefore crucial to provide an ontological *view* on this metabolomics data, especially in the context of the ChEBI project.

Concentration information has historically been difficult to represent ontologically, for three reasons:

1. Until recently, OWL did not provide support for defining classes based on data value ranges (in Description Logics (DLs) [1], these are known as *concrete domains*). This functionality is included in OWL 2 DL.
2. The link between concentrations and disorders is not *certain*, that is, it is information about what *may* be associated with a given disorder rather than what is *always* an indicator for the disorder. Logical reasoning cannot directly draw inferences from such associations.
3. The threshold between normal concentration levels and the levels associated with disorders is *vague*, that is, there is no hard numeric cutoff between normal concentration levels and disordered concentration levels.

In this paper, we present an approach to representing and reasoning with metabolite concentration levels associated with disorders, using OWL 2 data ranges and *probabilistic* DL reasoning [13] as implemented in Pronto [12], a probabilistic extension of Pellet [15]. We draw the metabolite data from the Human Metabolome Database [18].

Our implementation is guided by the following questions:

1. Can we differentiate normal from abnormal metabolite concentrations?
2. What is the *likelihood* that a patient has a given disorder, considering specified values for his/her concentrations of different metabolites in biofluids?
3. Can we *accumulate* the evidence (i.e. increase the likelihood) for the presence of a given disorder if there are multiple metabolite concentration values pointing towards it?

1 Background

1.1 ChEBI, metabolomics data and the HMDB

ChEBI is an OBO Foundry [16] ontology for the structural features and biological roles of biologically interesting chemicals [5]. Many of the biologically interesting chemicals are metabolites, which are are found in ChEBI together with their structural chemical classification and their biological roles including 'metabolite'. Roles are associated with chemicals using the *has role* relationship. However, there is currently no information formally captured in the ontology as to the *context* in which a chemical has a particular role.

The identification and annotation of the metabolites found in the human organism together with associated contextual information such as the disorders linked to different metabolic profiles, is being undertaken by the Human Metabolome Project [17], from which has arisen the Human Metabolome Database (HMDB) [18]. HMDB contains physicochemical, spectral, clinical, biochemical and genomic information for all known human metabolites. Each metabolite contains an extensive collection of information in text fields and images including measured concentration values taken from human samples of different biofluids (such as blood, urine, cerebrospinal fluid), from persons of different ages and with different underlying conditions.

In this paper, we focus in particular on metabolites for which HMDB contains both a normal and an abnormal (associated with some disease) concentration level for an adult subject. The difference between the normal and abnormal concentration values indicates a *threshold* between these scenarios, such that we would be able to infer the likelihood of a sample concentration being from a disordered organism by virtue of the numeric value being closer to the known disordered concentration than to the known normal concentration.

1.2 Reasoning with data ranges in OWL 2

OWL properties are separated between those that range over objects (descendents of `owl:Thing`) and those that range over data values for different types of data, such as integers or strings.

In OWL 2 [8], data restrictions can be used to define classes by referring to an operator and a range of values of a data property, such as strings or integers. For example (in Manchester syntax [11]),

$$Adult \texttt{ subClassOf } Human \texttt{ and } hasAgeInYears \texttt{ some int}[>= 18]$$

specifies that *Adult*s are those *Human*s that have ages greater than or equal to 18 years.

In the Description Logics underlying the OWL language, such data ranges are called *concrete domains* [2]. Concrete domains are defined with respect to a domain over which values can range, and a set of allowed predicates that operate on that domain. In our example above, the domain over which the *hasAgeInYears* data property ranges is the domain of *non-negative integers*, $\mathbb{N}0$, and the predicates which operates on that domain (in OWL, allowed predicates correspond to XSD facets) include '$\leq$', '=', '$>$'.

For the concentration values being represented in our metabolite concentration ontology, the domain is non-negative real numbers ($\mathbb{R}$), which we represent for sake of the necessary precision as XSD doubles (i.e. 64 bit floating point numbers), and the predicates we use are $\leq$ and $\geq$.

1.3 Probabilistic Description Logics

While standard Description Logics are designed to represent information that is *certain*, as in chemistry it is certain that all members of the class *carboxylic acids*

contain at least one *carboxy group*, a recent DL extension allows the association of probabilistic *uncertainty* with DL axioms [13].

Probabilistic DL-based ontologies extend classical DLs with probabilistic knowledge about classes and properties (known as *terminological* probabilistic knowledge) as well as about individuals (known as *assertional* probabilistic knowledge). Terminological probabilistic knowledge expresses knowledge about *randomly chosen* individuals belonging to classes, that is, *generic* members of the class, while assertional probabilistic knowledge is about specific named individuals in the knowledge base [13].

Probabilistic DLs extend traditional DLs with the ability to quantitatively model and reason with partially overlapping classes (specifying the degree to which two classes overlap), and to associate with each axiom in the ontology a probability value which represents the degree of reliability or *certainty* of the axiom. It is the latter capability that we will make use of. Probabilistic knowledge consists of *conditional constraints* [13].

Definition 1. *A* conditional constraint *is an expression of the form* $(\psi \mid \phi)[l, u]$*, where* ϕ *and* ψ *are classes in the ontology, and* l *and* u *are real numbers in the range [0, 1]. Informally,* $(\psi \mid \phi)[l, u]$ *encodes that* ϕ *is a subclass of* ψ *with probability between* l *and* u*.*

For example, we may wish to express the knowledge that *if* a certain patient has a measured metabolite concentration within a certain range (ϕ), *then* the probability of them having a certain disorder (ψ) is in the range [0.75, 0.85].

2 Creating the ontology

2.1 Data extraction and threshold calculation

The HMDB database was programmatically parsed from the downloadable metabocards export. Metabolites for which there was both a normal and an abnormal concentration in the same biofluid, were extracted. The normal and the abnormal concentrations were then used to generate a threshold condition which was half-way between the normal and the abnormal, and which was directed in the direction of the abnormal (either greater than or less than the threshold depending on which side the abnormal concentration fell).

For example, a pair of sample values for metabolite *D-glucose* in blood were 4440 uM for a normal adult and 7000 uM for an adult with the disorder *Diabetes Mellitus Type 2*. In this case we create a threshold at 5700 uM, having abnormal concentrations greater than the threshold.

Note that the threshold being set half-way between normal and abnormal is an artificially introduced constraint for the purpose of this paper. Identifying true thresholds between normal and abnormal concentration levels is of course a much more complex procedure requiring large numbers of samples and sophisticated techniques for eliminating noise in the underlying data [6, 3]. However, for our purposes in evaluating the representation of such information in OWL, we can safely ignore this additional complexity.

2.2 Populating the OWL ontology with data

The OWL ontology was created using the OWL API [10] and reasoned over with a slightly modified form of Pronto[6] [12]. The full generated ontology, illustrated in Figure 1, includes data for 48 metabolites associated with 39 different disorders[7].

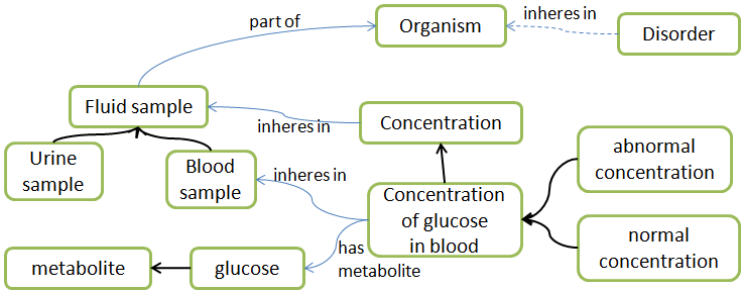

Fig. 1. Metabolite ontology: Fluid samples, of which blood and urine are two examples, are considered part of organisms. Concentrations of various different metabolites inhere in these fluid samples. Concentrations may be normal or abnormal, and if abnormal are associated with a disorder, which inheres in the organism from which the sample was extracted. Unlabelled arrows represent *is a* (`subClassOf`) relationships.

Note that the ontology shows fluid samples as part of organisms, although this is a simplification since fluid samples are in actual fact typically no longer part of an organism, and concentration values may depend on how the sample was extracted and processed.

Calculated threshold values were added to the ontology as classes defined with data ranges. For example, we fully define the class *concentration of D-glucose in Blood associated with Diabetes mellitus type 2* as:

```
'concentration of D-glucose in Blood associated with Diabetes mellitus type 2'
  equivalentTo ( 'concentration in blood'
    and (hasMetabolite some 'portion of D-glucose')
    and (hasConcentrationValue some double[>= 5700.0]) )
```

In addition to the simplification involved in setting the threshold half way between the normal and abnormal concentrations, there is a deeper underlying problem with this threshold model. Even if we included an accurate threshold between normal and abnormal, this threshold represents at the class level what is *generally* true across many individuals, but obscures the underlying *individual variance* in phenotype and metabolism which might affect the actual threshold for each individual. Furthermore, it represents normal and abnormal as a *binary* phenomenon whereas in reality there is a continuum between the normal and the abnormal [14]. Thus, we cannot create a straightforward DL relationship

[6] Version 0.2, upgraded to the latest version of the OWL API and Pellet, since data ranges were not available in the implemented OWL 1.1. version.

[7] The ontology (META.owl) and software (META.zip) are available for download from `http://www.ebi.ac.uk/~hastings/concentrations/`.

between a given metabolite concentration and a disorder, since, according to the current model and the underlying DL semantics, each concentration instance would then be associated with at least one disorder instance. It is to address this gap that we propose the use of probabilistic DL.

2.3 Adding probabilistic constraints

The challenge is to be able to infer, based on measured metabolite concentration values, the *likelihood* of presence of a disorder. We will call this the *risk* of having the disorder, given the concentration value of the metabolite. We create classes for the categories of *low*, *medium* and *high* risk of having the given disorder. Note that the variation of risk with concentration value can be thought of, as a simplifying assumption, as a continuously valued function ranging over all possible concentration values[8]. However, as Pronto constraints take the form of *intervals* associated with *classes* (or instances), to create a finite number of OWL classes and associate probability intervals to them, it is necessary to discretize the probability function into fixed ranges. We will do this as illustrated in Figure 2.

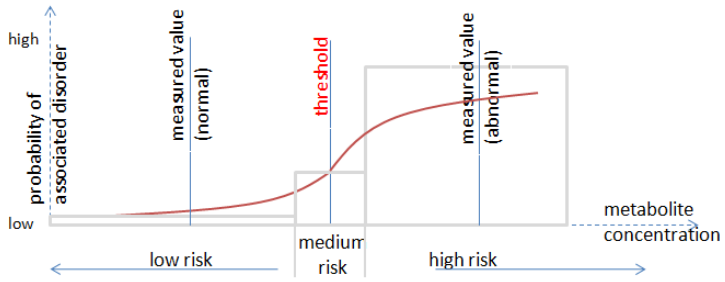

Fig. 2. Discrete approximation: We assume a continuous probability function for the relationship between metabolite concentration and the risk of having the associated disorder. We assign three risk ranges, with *medium* risk ranging around the threshold value. The diagram represents a scenario where the abnormal concentration of the metabolite is larger than the normal concentration.

For example, we fully define the class *person with low risk of having diabetes based on their blood glucose level* as:

```
'person with high risk of having Diabetes mellitus type 2 based on Blood sample of D-glucose'
equivalentTo (organism
   and hasPart some (bloodSample
     and  bearerOf some (concentration
       and hasMetabolite some 'portion of D-glucose'
       and hasConcentrationValue some double[>=6840.0] ) ) )
```

[8] In general, associative relationships between symptoms/signs and disorders are more complex, involving two parameters: (1) the probability of the disorder given the sign/symptom; and (2) the probability of the sign/symptom given the disorder [9]. We include only the first.

We create the relevant low, medium and high risk categories for each fluid type and disorder for the metabolites *D-glucose* and *acetoacetic acid*. Although it is possible to use our software to create such classes for every metabolite in the ontology, we have selected this subset to reduce overhead for reasoning.

Finally we create the conditional constraints that associate the given risk categories for the associated disorder, with a certain probability. We have arbitrarily selected the following probability ranges for the given risk categories: Low risk: [0.00;0.24]; Medium risk: [0.25;0.54]; High risk: [0.55;1.00]. As required by Pronto, conditional constraints are added to the ontology as an *annotation* on a `subClassOf` axiom; these axioms are then *removed* from the main ontology and added to the probabilistic knowledge base by the Pronto pre-processor.

For example, we add the constraint:

```
'person with high risk of having Diabetes mellitus type 2 based on Blood sample of D-glucose'
        subClassOf 'person with Diabetes mellitus type 2'
          pronto:certainty "0.55;1.00"
```

3 Results of reasoning

We test the reasoning capability of the generated ontology corresponding to the three questions listed in the introduction. As a simple probe, we create three individuals with different metabolite concentration measurements. Table 1 describes the individuals and their blood concentration values.

Individual	Metabolite	Concentration	Expected Risk (Diabetes)
Harry	Glucose	4000.0	Low
Sally	Glucose	10000.0	High
Barry	Glucose	6000.0	Med
Barry	Acetoacetic acid	2000.0	High

Table 1. Individuals with sample metabolite concentrations

3.1 Reasoning with data ranges

The first question tests reasoning with numeric thresholds for inferences about conditional properties. Here, the conditional property is the suspected presence of a *disorder* in the organism whose fluid sample has been measured for the particular metabolite concentration. Testing this is straightforward: the individual `Sally` in the ontology is associated with a blood glucose concentration value of 10000 uM. After classifying using Pellet, Sally's concentration is correctly classified as abnormal.

3.2 Reasoning with probability

Answering the second question involves the use of the probabilistic constraints. To allow interplay between the results of reasoning with the data ranges expressed in the ontology and the probabilistic reasoning, we executed a two-step

process: firstly, the classical reasoning was performed, and then the inferred class memberships were asserted back into the probabilistic ontology before performing the probabilistic reasoning. This allows us to ask Pronto to answer the question: *entail* that an individual (e.g. `Harry`) has the disease *Diabetes mellitus Type 2*. In response, Pronto provides a probability range and an explanation, which refers to the probabilistic constraints used in generating the conclusion. The results are illustrated in Table 2.

Individual	Risk of Diabetes [l;u]
Harry	[0.0;0.24]
Sally	[0.55;1.0]
Barry	[0.25;1.0]

Table 2. Individuals with inferred probability results for diabetes

The results for `Harry` and `Sally` are a straightforward result of the risk categories associated with the classes for which their membership is inferred. However, that of `Barry` is more complex since he has multiple concentration values implicating the disease.

3.3 Reasoning with multiple probabilistic constraints in combination

Pronto uses linear resolution to determine the probability range entailed by a set of constraints [12]. There are two scenarios: when multiple constraints can be resolved (into a probabilistic interval entailment), and when they conflict. Since `Barry` has a blood *D-glucose* concentration in the medium risk range and a blood *acetoacetic acid* concentration in the high risk range, and the two ranges do not conflict, the above result for `Barry` indicates Pronto's strategy in the absence of a conflict, resembling a *union* of the two underlying data ranges.

When multiple constraints conflict, Pronto prefers *more specific* statements to less specific. We evaluated this behaviour by changing the medium risk constraint to *overlap* with the high risk constraint, setting the upper bound for medium to 0.55 instead of 0.54. In this case, Pronto concludes that the probability for `Barry` having diabetes is [0.55;0.55] – the most specific (narrowest) resolution. If the medium risk ranges to 0.6, Pronto entails `Barry` the range [0.55;0.6]. Thus, it seems that the behaviour on conflict (at least for the two-axiom scenario we test here) resembles an *intersection* of the two underlying data ranges.

While it remains a task for future work to examine the reasoning behaviour under more complex scenarios, neither of these results is an optimal representation of the intuitive requirement driven by the use case: it would be better if the probabilistic combination of different types of evidence for the same conclusion *increased* the certainty of the conclusion. However, Pronto does allow for *overriding* inherited constraints in more specific subclasses. Thus, we can specify a new risk subclass for `Barry`'s combined risk categories, and associate this with the disease with a new probability range (e.g. [0.54;0.85]). However, this approach is in general somewhat cumbersome as it would require adding

many more classes and constraints to the knowledge base – for all interesting combinations of risk factors.

4 Conclusion

Metabolomics is the field which bridges between chemical data and biological data by investigating the chemical markers for biological processes, and therefore for their underlying disorders [18]. Accurately modelling the associations between metabolites and disorders goes beyond traditional OWL modelling constructs. We have evaluated a probabilistic representation strategy using Pronto. While probabilistic ontologies have been used to model, e.g. breast cancer risk factors [12], they have to our knowledge not previously been applied to chemical–disease associations, nor used in combination with concrete domains. Our prototype has illustrated the general applicability of the approach, but a more intuitive and flexible solution for reasoning with combined probability constraints would be mandatory for a real application based on this scenario.

Future work will involve the investigation of alternative probabilistic DL approaches, such as those which use an underlying Bayesian model [4], and ultimately address the extension of this prototype towards a full implementation linking ChEBI metabolites to diseases.

Acknowledgements

This work was partly supported by the Deutsche Forschungsgemeinschaft (DFG) grant JA 1904/2-1, SCHU 2515/1-1 GoodOD (Good Ontology Design) and by the BBSRC, grant agreement number BB/G022747/1.

References

1. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook: Theory, Implementation, and Applications, 2nd Edition. Cambridge University Press, 2 edn. (Sep 2007)
2. Baader, F., Sattler, U.: Description logics with aggregates and concrete domains. Information Systems 28(8), 979–1004 (Dec 2003), http://www.sciencedirect.com/science/article/B6V0G-481FTVC-1/2/06400b60da99c41bc6e07596ff8950c1
3. van den Berg, R., Hoefsloot, H., Westerhuis, J., Smilde, A., van der Werf, M.: Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7(1), 142 (2006)
4. da Costa, P.C.G., Laskey, K.B.: PR-OWL: A framework for probabilistic ontologies. In: International Conference on Formal Ontology in Information Systems. pp. 237–249 (2006)
5. de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C.: Chemical Entities of Biological Interest: an update. Nucl. Acids Res. 38, D249–D254 (2010)

6. Flöter, A., Nicolas, J., Schaub, T., Selbig, J.: Threshold extraction in metabolite concentration data. Bioinformatics 20(10), 1491–1494 (2004), http://bioinformatics.oxfordjournals.org/content/20/10/1491.abstract
7. Gieger, C., Geistlinger, L., Altmaier, E., Hrab de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J., Illig, T., Suhre, K.: Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. PLoS Genet 4(11), e1000282 (11 2008)
8. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: OWL 2: The next step for OWL. Web Semant. 6, 309–322 (November 2008), http://portal.acm.org/citation.cfm?id=1464505.1464604
9. Hall, G.H.: The clinical application of Bayes' theorem. The Lancet 290, 555–557 (1967)
10. Horridge, M., Bechhofer, S.: The OWL API: A Java API for working with OWL 2 ontologies. In: Hoekstra, R., Patel-Schneider, P.F. (eds.) Proc. of OWL Experiences and Directions 2009 (OWLED 2009) (2009)
11. Horridge, M., Patel-Schneider, P.F.: OWL 2 web ontology language manchester syntax (Oct 2009), http://www.w3.org/TR/2009/NOTE-owl2-manchester-syntax-20091027/
12. Klinov, P.: Pronto: A Non-monotonic Probabilistic Description Logic Reasoner. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) The Semantic Web: Research and Applications, Lecture Notes in Computer Science, vol. 5021, chap. 66, pp. 822–826. Springer Berlin Heidelberg, Berlin, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-68234-9_66
13. Lukasiewicz, T.: Probabilistic description logics for the semantic web. TU Vienna infsys research report (2007)
14. Schulz, S., Johansson, I.: Continua in biological systems. The Monist 4, 499–522 (2007)
15. Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Journal of Web Semantics 5, 51–53 (2007)
16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11), 1251–1255 (Nov 2007), http://dx.doi.org/10.1038/nbt1346
17. Wishart, D.S.: Current progress in computational metabolomics. Briefings in Bioinformatics 8(5), 279–293 (2007), http://bib.oxfordjournals.org/content/8/5/279.abstract
18. Wishart, D.S., Knox, C., Guo, A.C.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J.A., Lim, E., Sobsey, C.A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H.J., Forsythe, I.: HMDB: a knowledgebase for the human metabolome. Nucleic acids research 37(Database issue), D603–610 (Jan 2009), http://dx.doi.org/10.1093/nar/gkn810