Laboratory of Systems and Synthetic Biology
www.wur.nl\ssb
www.systemsbiology.nl

Design and implementation of a FAIR framework for reproducible model usage

Supervisors: Robert Smith

Contacts: robert1.smith@wur.nl

Type of thesis: Computational

Required competences: Ability to code in Python and preferably having completed a modelling course such as Modelling Dynamic Systems (BCT20806), Introduction to Systems & Synthetic Biology (SSB50806) or Modelling in Systems Biology (SSB30806). It would be advantageous to have followed courses like iBiosystems (SSB20306), Bioinformation Technology (SSB20806) or Linked Data (INF35306) for a background in FAIR principles.

Acquired competences: FAIR data management (use of ontologies and querying), coding skills, understanding limitations and problems with mathematical model reusability.

Date: 16-05-23

Description

For the last 20 years, life scientists have become concerned by the lack of reproducibility and reusability of data. The lack of reproducibility means that researchers have to repeat experiments, slowing down exploration of new ideas to try in a bid to understand someone else's results. The lack of reusability will slow down technology development, as datasets that adhere to vastle different standards or frameworks cannot be automatically analysed or modelled. To try and solve these issues, many in the European research community advocate that all experimental data should follow FAIR principles (Findable, Accessible, Interoperable, and Reusable; www.go-fair.org). Following these principles, all experimental data is linked to metadata describing how an experiment was conducted, with which tools, and how data was collected. In this way, a researcher can understand and reproduce an experiment exactly. All this data is linked together with graph structures online (as you see, for example, on Wikipedia).

However, this issue is not suffered by experimental data alone. The simulation of mathematical models depends on use of specific solvers (e.g. ode45 for ODE models or GUROBI for constraint-based models) or certain platforms/packages (e.g. Python, R, Antimony, etc.) that have their own metaparameters. How a model is built is not always clear from publications either – such as which data was used for training and testing/validation, or how were model unknowns found (e.g. Which optimisation routine was used? What was the size of the search space tested?). Whilst all these details *should* be written down in publications, this is not always the case and means that even model results cannot be reproduced and models cannot be reused – in the same manner as life science experiments.

Current Systems Biologists worried about this situation has focussed on sharing models themselves using interoperable methods, such as the SBML coding language (Keating et al., 2020). This has led to model characteristics being described using the FAIR SBTab data format (Lubitz et al., 2016). More general ontologies are also available for mathematical terms (https://www.ebi.ac.uk/ols/ontologies/mamo). However, this does not solve the problem of storing information/metadata about model usage such that model simulations can be understood and reproduced. This project aims to solve this: the interested student will need to think about how to organise model metadata, check already available ontologies and develop their own ontology to connect data if needed, collect data from published sources, and validate the data network such that it can be queried with, e.g., GraphDB.

References

Lubitz T., et al. (2016) doi: 10.1093/bioinformatics/btw179

Keating S.M., et al. (2020) doi: 10.15252/msb.20199110