Deliverable

D1.10 Adaptation of BOQA algorithm to its use in the ontology of unsolved RD	
Version | Status	V1 | final
Work package	WP1
Lead beneficiary	Charité (Sebastian Köhler)
Due date	31.12.2018 (M12)
Date of preparation	15.12.2018
Target Dissemination Level	Public[1]
Author(s)	Sebastian Köhler (Charité), Svitlana Havrylenko (INSERM-Orphanet)
Reviewed by	Peter Robinson (JAX)
Approved by	Ana Rath (INSERM-Orphanet)

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257.

[1] In November 2021, the authors of this deliverable report decided to make it publicly available via the Solve-RD website.

Explanation according to GA Annex I:

The degree of similarity of unsolved cases to each other and known diseases will be calculated based on a Bayesianalgorithm called BOQA (Bauer 2012 Bioinformatics). The resulting ontology will be scalable, in order to include also cases coming from math-making initiatives in the future, as well as other ontologies of interest such as those based on animal model phenotypes.

Abstract:

WP1 not only aims to collect standardized phenotypic information from unsolved rare disease (RD) cases, but also aims to transform their phenotypic descriptions into diagnostic hypotheses. One of the proposed means to produce these diagnostic hypotheses is calculating a numerical similarity value that reflects how well the phenotype information of an unsolved case aligns with other solved or unsolved cases as well as how well it overlaps with known rare diseases.

HPO provides a standardised phenotype ontology that is used for data collection in this WP. Orphanet will build the ontology of unsolved RD cases (RDCO) (D1.9). Here we present the work, which is located at the interface between the data collection efforts (e.g. D1.1) and the work on the ontological model for unsolved diseases (D1.9). Specifically, we have developed a software tool that adapts several algorithms to calculate the degree of similarity between unsolved cases and known diseases using solved cases from the RD-Connect project (http://rd-connect.eu/). The results of this tool will be imported directly into the ontology of rare unsolved cases (RDCO).

We have tested the tool and the therein implemented algorithms on 107 PhenoPackets obtained from solved cases. We identified two algorithms that show superior performance on this relatively small test set and were able to run this tool on Orphanet computers which is required for D1.9.

Introduction:

Following data collection, one of the central components for the success of Solve-RD will be the ability to align data of unsolved cases with data about solved cases as well as known rare disease entities. This is an essential step to make use of existing knowledge in an automated computational approach with the ultimate goal of solving as many unsolved cases as possible.

One type of data that is being collected within the Solve-RD project is phenotype data. This is done using the Connect Genome-Phenome Analysis Platform (GPAP), which makes use of the Human Phenotype Ontology (www.human-phenotype-ontology.org, HPO) to represent clinical features noted for the submitted cases. HPO-encoded phenotype data has already been used in multiple projects and software tools to improve the differential diagnosis procedure (e.g. Phenomizer, http://compbio.charite.de/phenomizer/) or identify likely pathogenic DNA aberrations in a phenotype-driven fashion (e.g. Exomiser, https://github.com/exomiser/Exomiser).

However, the algorithms implemented in those tools are not easily accessible and thus hard to apply to Solve-RD data. Herein, we present the adaptation and implementation of several algorithms to determine the similarity between two PhenoPackets and between PhenoPackets and ORDO entries. The developed software uses HPO and the Orphanet disease HPO annotation data.

Report:

There are several different algorithms based on different principles to determine the degree of similarity between two sets of HPO classes. A priori it is not possible to know which algorithm is the best choice in the setting of Solve-RD. To our knowledge, there have been no large-scale comparative studies, especially for evaluating different algorithms when comparing the phenotypic profiles of patients with each other. Thus, it is unclear which algorithm will have the best performance or which one fits best the needs of the users of the software. For example, in some algorithms it is easier to track the computation and to explain the inner workings to a user.

For the purpose of this project we decided to work with eight different algorithms (Table 1). Note that we have not used the original implementation of BOQA and have instead implemented Algorithm 1 from the publication by Bauer et al. (2012). We have also added two algorithms that can make use of the frequency information for each associated HPO class in ORDO.

Algorithm	Uses Frequency Information	Based on	Reference(s)
Resnik (symmetric)	No	Information content (IC)	Resnik et al. (1995), Pesquita et al. (2009), Köhler et al. (2009)
Resnik (asymmetric)	No		
Bernoulli with Grid (BOQA-like)	No	Bayesian approach & Inherent IC	Bauer et al. (2012)
PhenoDigm	No	Information content (IC)	Smedley et al. (2013)
Jaccard	No	HPO subclass hierarchy	Pesquita et al. (2009), Köhler (2018)
Jaccard (weighted)	Yes		
Cosine	No	HPO subclass hierarchy	Pesquita et al. (2009), Köhler (2018)
Cosine (weighted)	Yes		

Table 1. The algorithms tested in this project.

In order to evaluate the performance of these eight algorithms (Table 1) it was decided to test them on as many solved cases as we could currently identify. Initially, the PhenoPackets of 107 solved cases were sent by Orphanet to Charité. They correspond to cases marked as “solved” in the RD-connect GPAP PhenoTips instance and having a clinical diagnosis and/or causal gene variant information available (batch 1 and 2 in Deliverable 1.1). Not all 107 PhenoPackets were useful for the purpose of our analysis (see Figure 1). We finally included 59 cases in the analysis.

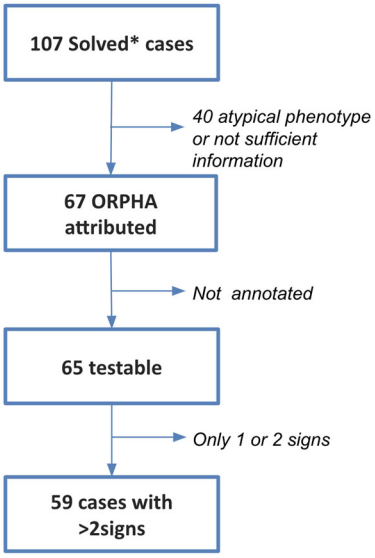

Figure 1. Solved cases obtained from GPAP. The Figure shows which criteria we used to exclude cases that were not useful for the purpose of algorithm testing.

For the analysis we took all the phenotype data as HPO classes and used the different methods to rank all ORDO diseases by similarity to the data in the PhenoPacket. The rank of the correct diagnosis was then recorded. The lower the rank, the better the performance of the algorithm, as a potential user would have look through less entries before arriving at the correct diagnosis. For this study we recorded only the rank of the correct diagnosis if it was under the first 50 disease in the ranked list. In Figure 2, it becomes evident that four algorithms (Resnik symmetric, Resnik asymmetric, Bernoulli with grid (variation of BOQA), and PhenoDigm) show the best performance.

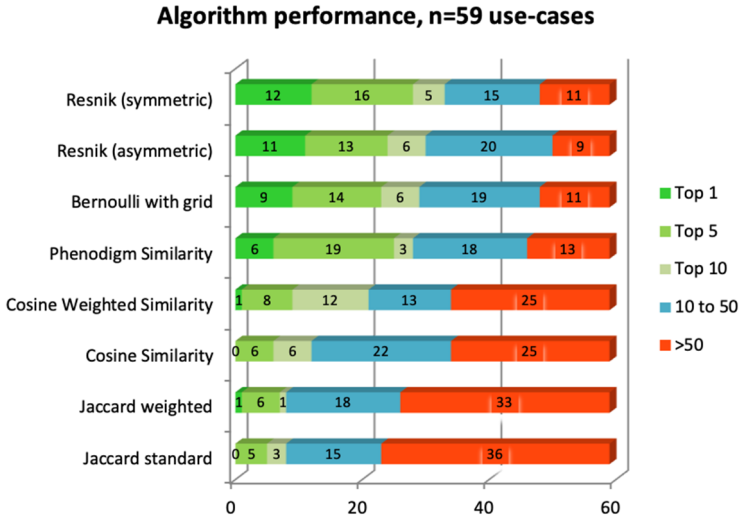

Figure 2. Performance of the algorithms on placing sought diagnoses in the first 50 ranks.

Given these results it was decided to continue for now with the Resnik-based similarity measure, but to design the downstream software and infrastructure (e.g. RDCO) in a way that will allow switching to another algorithm at a later stage if needed. This may be needed as the number of solved cases that we could test in this phase of the project is only a very small excerpt of the reality of unsolved cases. It may be the case that another algorithm will show a better performance with more data available.

Figure 3 shows a screenshot of the yet private GitLab repository, which has been shared with developers from Orphanet (esp. developers for RDCO). The code has been optimized to enable users to run the complete analysis on a laptop in an adequate time.

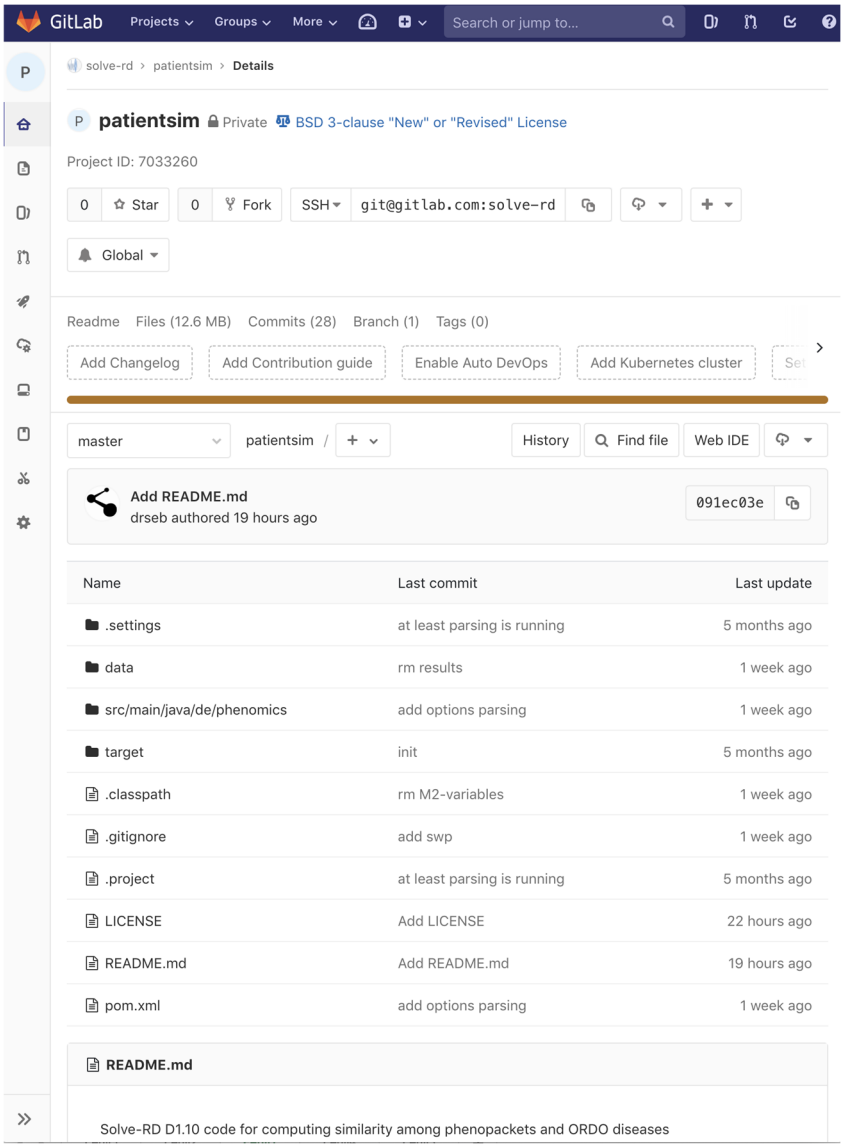

Figure 3. Screenshot of the GitLab repository for the software.

From the GitLab repository shown in Figure 3 it is possible to generate an executable jar-file that can be run on any machine, that has Java installed. In Figure 4, it is shown which parameters are required (or optional) to run this software. Required are the data, such as HPO, Orphanet-disease-annotations, and the folder with the PhenoPackets to process. Optional is the removal of duplicates or the number of top ranked entries to report in the results.

```
Sebastian-Kohlers-iMac-2:jars sebastiankohler$ java -jar runSolveRdAnalysis.jar
Missing required options: o, x, p
usage: java -jar <yourjarfile>.jar
 -d,--remove-duplicates                  option to turn on the removal of
                                         'duplicate' annotations
 -m,--numberResultsPhenopackets <arg>    the number of top scoring ordo
                                         entries to report in results
 -n,--numberResultsDiseases <arg>        the number of top scoring ordo
                                         entries to report in results
 -o,--hpo <arg>                          the HPO obo file
 -p,--phenopackets <arg>                 the folder containing the
                                         phenopackets (.yaml)
 -r,--create-report                      option to turn on the generation
                                         of debug-reports for Resnik
 -x,--orphaxml <arg>                     the en_product4_HPO.xml file
```

Figure 4. Screenshot of calling the provided executable file without any parameters, to induce the help screen to be shown. The help thus shows the different options that are currently provided.

For each PhenoPacket, the software then computes the most similar ORDO entries and the most similar other PhenoPackets. The results are stored in three folders (Figure 5).

Figure 5. Results produced by one run of the software. Three folders are produced which organise the results by reports, disease-related results and case-related results.

The first folder is called “reports” and currently contains a report for invalid PhenoPackets, such as empty PhenoPackets. It also computes reports on “duplicate” HPO classes inside a PhenoPacket. A duplicate HPO class is for example “Muscle weakness”, when “Distal muscle weakness” is already assigned to a PhenoPacket, because these two classes are in subclass-relationship. It is also possible to remove such “duplicates” before the analysis using the parameter “-d” (see Figure 4).

The second folder contains the most similar ORDO entries (Orphanet diseases) for each PhenoPacket. We have decided to create multiple formats for the same results, as e.g. JSON is easier to process with downstream software and the txt-version (Figure 5) is easier to understand for a human reader.

The third folder is very similar to the previously described folder, but instead of ORDO entries, other PhenoPackets are being ranked by their phenotypic similarity to the current case.

To give an example of a result file, Figure 6 shows how we provide the results in a human readable fashion. The pound sign indicates a comment-line – here we provide information about the used HPO version, Orphanet disease annotation version, and parameters that were used when calling the software. In each line (starting from line 8) we show the results for one

of the currently implemented algorithms. The lines start with the name of the algorithm followed by a ranked list of items. Each item is represented with four elements:

- The item id (e.g. ORPHA:98915)
- The label of item (e.g. Synaptic congenital myasthenic syndromes)
- The rank (e.g. 1)
- The score (e.g. 0.534)

These results are also produced in JSON format in a separate file, which will be used by Orphanet to populate RDCO.

```
# ontology version: releases/2018-07-25
# annotation data version: 2018-07-24 13:31:12
# parameters:
#  - number most similar disease to report: 100
#  - number most phenopackets to report: 100
#  - do create debug-report ? : true
#  - do remove duplicates ? : false
Resnik (symmetric)      ORPHA:98915 | Synaptic congenital myasthenic syndromes | 1.0 | 0.534    ORPHA:590 | Congenital myasthenic syndrome
Resnik (asymmetric)     ORPHA:590 | Congenital myasthenic syndrome | 1.5 | 0.730        ORPHA:98914 | Presynaptic congenital myasthenic syn
Bernoulli with grid     ORPHA:98915 | Synaptic congenital myasthenic syndromes | 1.0 | 0.000    ORPHA:590 | Congenital myasthenic syndrome
Cosine Similarity       ORPHA:230800 | Toxin-mediated infectious botulism | 1.0 | 0.482 ORPHA:363623 | Autosomal recessive limb-girdle musc
Cosine Weighted Similarity      ORPHA:98915 | Synaptic congenital myasthenic syndromes | 1.0 | 0.196    ORPHA:590 | Congenital myasthenic s
Jaccard standard        ORPHA:230800 | Toxin-mediated infectious botulism | 1.0 | 0.313 ORPHA:178481 | Intestinal botulism | 2.0 | 0.273
Jaccard weighted        ORPHA:590 | Congenital myasthenic syndrome | 1.5 | 0.069        ORPHA:98914 | Presynaptic congenital myasthenic syn
Phenodigm Similarity    ORPHA:363623 | Autosomal recessive limb-girdle muscular dystrophy type 2T | 1.0 | 0.701 ORPHA:486815 | Congenital m
```

Figure 6. An example result file, showing the most similar ORDO entries for the given PhenoPacket.

Conclusion:

We have successfully implemented a system to compute the degree of similarity of unsolved cases to each other and to known diseases. A modified version of a Bayesian algorithm (BOQA) is one of several implemented algorithms. We will continue to add more algorithms to make use of negated phenotypes, but these are less established at the moment and need more thorough testing and evaluation. We will also include opposite-of information as has been shown in Köhler et al (2018, bioRxiv). We will re-run all analyses to evaluate the different algorithms as more solved and unsolved cases will become available throughout the next months.