Phenotype Inference from Text and Genomic Data

Maria Brbić[1], Matija Piškorec[1], Vedrana Vidulin[1], Anita Kriško[2], Tomislav Šmuc[1], and Fran Supek[1,3](✉)

[1] Ruđer Bošković Institute, Zagreb, Croatia
[2] Mediterranean Institute of Life Sciences, Split, Croatia
[3] Centre for Genomic Regulation, Barcelona, Spain
fran.supek@irb.hr

Abstract. We describe ProTraits, a machine learning pipeline that systematically annotates microbes with phenotypes using a large amount of textual data from scientific literature and other online resources, as well as genome sequencing data. Moreover, by relying on a multi-view non-negative matrix factorization approach, ProTraits pipeline is also able to discover novel phenotypic concepts from unstructured text. We present the main components of the developed pipeline and outline challenges for the application to other fields.

Keywords: Phenotypic trait · Microbes · Comparative genomics
Late fusion · Text mining · Non-negative matrix factorization

1 Introduction

With the development of next-generation DNA sequencing techniques, the number of available microbial genomes has rapidly increased. However, this explosive growth of genomics data is not followed by the phenotypic annotations of organisms, such as growth at extreme temperatures, resistance to radiation, or the ability to cause disease in plants, animals or humans. The systematic annotation of organisms with phenotypic traits is of importance for discovering the associations between genes to phenotypes that would suggest a biological basis for various traits. Existing databases [7,11] rely on manual annotation of organisms, which results in limited coverage. On the other hand, there is a vast amount of unstructured data with phenotype descriptions available in scientific articles and other textual resources. Motivated by this abundance of genomic and of textual data, we developed ProTraits [2] - a machine learning-based pipeline that systematically assigns predictions across large number of organisms and phenotypes. Along with predicting existing phenotypic labels, ProTraits pipeline is also able to define novel phenotypic concepts from unstructured text using a multi-view approach based on non-negative matrix factorization followed by clustering and manual curation. Here, we briefly describe main components of

Y. Altun et al. (Eds.): ECML PKDD 2017, Part III, LNAI 10536, pp. 373–377, 2017.
https://doi.org/10.1007/978-3-319-71273-4_34

our pipeline and present an overview of results. The proposed approach can easily be extended to other fields with the abundant unstructured textual data. The ProTraits database of microbial phenomes is available at http://protraits.irb.hr/.

2 Methodology

In this section, we describe the main components of the ProTraits pipeline (Fig. 1): (i) unsupervised phenotype discovery based on multi-view non-negative matrix factorization; (ii) a supervised machine learning framework for phenotype inference from textual and genomic data; (iii) a late-fusion based component for the combination of predictions coming from 11 independent models, and (iv) a user-friendly web interface providing searchable predictions.

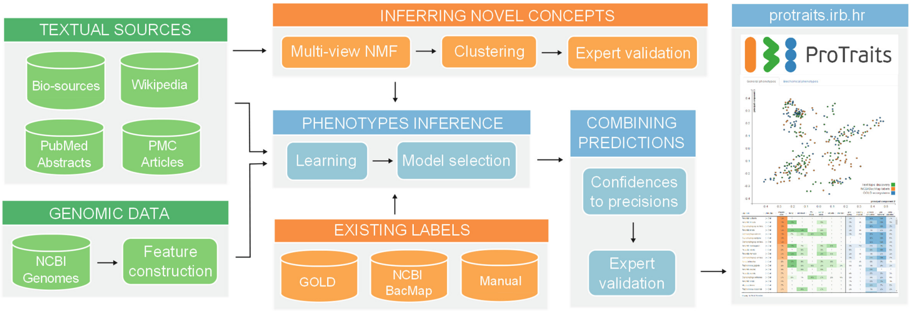

Fig. 1. System architecture of the ProTraits pipeline

2.1 Initial Data

Text documents describing bacterial and archaeal species were downloaded from six textual resources including Wikipedia, the MicrobeWiki student-edited resource, PubMed abstracts of scientific publications, PubMedCentral full-texts, and an additional set of assorted microbiology resources. The initial set of phenotype assignments was collected from NCBI, BacMap [11] and GOLD databases [7]. The set of biochemical phenotypes was collected manually from individual publications where various microbial species were initially characterized.

2.2 Inferring Phenotypic Concepts

We applied non-negative matrix factorization (NMF), commonly used for topic discovery tasks, to each text resource separately to discover novel phenotypic concepts. We then clustered the NMF factors, while requiring that a concept has to be consistently discoverable in at least three text resources. Since the NMF algorithm has a stochastic component, we ran the algorithm multiple times with different random seeds while also varying the number of factors parameter, in order to maximize the diversity of discovered concepts. These groups were then examined by an expert and those describing new phenotypes were retained and used in the same way as labels collected from the existing databases. In total, we discovered 113 non-redundant novel phenotypic concepts.

2.3 Phenotype Prediction

In the phenotype prediction task, the learning examples were species and the class label was the presence/absence of a phenotype in that species. A separate model was trained for each of the 424 phenotypes and 10-fold cross-validation used to estimate the accuracy. Once a model was learned, it was applied to the species with unknown phenotypic annotations. To make the functioning of our models more interpretable to biologists, we also provide sets of most important features of all models.

Predictions from textual data. We used bag-of-words representation with tf-idf weighting of word frequencies across documents assigned to species in a given text corpus. A Support vector machine (SVM) classifier with a linear kernel was trained on all combinations of text resources and phenotypes.

Predictions from genome data. We constructed five different genomic representations for each microbial species: (i) the proteome composition [1,9]; (ii) the gene repertoire encoded as presence/absence of Clusters of Orthologous Groups (COG) gene families [4,6]; (iii) co-occurrence of species across environmental sequencing data sets [3]; (iv) gene neighborhoods [8] encoded as pairwise chromosomal distances between gene family members; and (v) genomic signatures of translation efficiency in gene families [5,10]. Again, we trained models on all combinations of representations and phenotypes. We used the Random Forest (RF) classifier which we found to outperform other tested algorithms.

Combining predictions. To combine predictions from different models and provide an interpretable estimate of confidence in each prediction, the confidence scores of each prediction were converted to precisions, based on cross-validation precision-recall curves. Precision scores for organisms in the initially unlabeled set of organisms were calculated via linear interpolation between the neighboring confidence points and then assigned to both positive and negative class for each prediction and further adjusted to account for difference in class sizes, ensuring that the minimum precision of each class is 0, regardless of the number of positive/negative examples. The systematic validation performed by two experts on a random sample of $2,500$ predictions showed that the precisions combined using late fusion schemes agree well with human judgment, particularly when requiring agreement of two independent models (either text or genomics-derived).

Web interface and results. In summary, ProTraits covers $3,046$ microbial organisms and 424 microbial phenotypes. It provides predictions across six textual resources and five independent genomic representations. At the precision threshold higher than 0.9, ProTraits assigns ≈545,000 novel annotations, out of which ≈308,000 are supported in two or more independent predictions. A web interface at http://protraits.irb.hr/ provides precision scores across 11 individual predictors and an integrated score calculated using the two-votes late fusion scheme.

3 Challenges and Conclusions

Training separate classifiers for each of the phenotypes does not scale well in terms of computation time required, especially for high-dimensional genomic datasets. However, using existing multi-label classifiers was not straightforward for our datasets since most of the target values were missing. Another challenge was collecting initial labels, as this requires tedious manual curation. While the two existing microbial phenotype databases alleviated this problem in our work, for other important problems in the life sciences, similar databases may not be available. Crucially, the input of field experts has allowed us to validate predictions and inferred concepts, demonstrating that our models are trustworthy.

Acknowledgments. This work has been funded by the by the European Union FP7 grants ICT-2013-612944 (MAESTRA) and Croatian Science Foundation grants HRZZ-9623.

References

1. Brbić, M., Warnecke, T., Kriško, A., Supek, F.: Global shifts in genome and proteome composition are very tightly coupled. Genome Biol. Evol. **7**, 1519–1532 (2015)
2. Brbić, M., Piškorec, M., Vidulin, V., Kriško, A., Šmuc, T., Supek, F.: The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. **44**, 10074–10090 (2016)
3. Chaffron, S., Rehrauer, H., Pernthaler, J., von Mering, C.: A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. **20**, 947–959 (2010)
4. Feldbauer, R., Schulz, F., Horn, M., Rattei, T.: Prediction of microbial phenotypes based on comparative genomics. BMC Bioinform. **16**, 1–8 (2015)
5. Kriško, A., Copić, T., Gabaldón, T., Lehner, B., Supek, F.: Inferring gene function from evolutionary change in signatures of translation efficiency. Genome Biol. **15**, R44 (2014)
6. MacDonald, N.J., Beiko, R.G.: Efficient learning of microbial genotype-phenotype association rules. Bioinformatics **26**, 1834–1840 (2010)
7. Reddy, T.B.K., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A., Kyrpides, N.C.: The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. **43**, D1099–1106 (2015)
8. Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., Koonin, E.V.: Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. **30**, 2212–2223 (2002)

9. Smole, Z., Nikolic, N., Supek, F., Šmuc, T., Sbalzarini, I.F., Kriško, A.: Proteome sequence features carry signatures of the environmental niche of prokaryotes. BMC Evol. Biol. 11–26 (2011)
10. Supek, F., Škunca, N., Repar, J., Vlahoviček, K., Šmuc, T.: Translational selection is ubiquitous in prokaryotes. PLoS Genet. **6**, e1001004 (2010)
11. Stothard, P., Van Domselaar, G., Shrivastava, S., Guo, A., O'Neill, B., Cruz, J., Ellison, M., Wishart, D.S.: BacMap: an interactive picture atlas of annotated bacterial genomes. Nucleic Acids Res. **33**, D317–D320 (2005)