PathwayMatcher: multi-omics pathway mapping and proteoform network generation

Luis Francisco Hernández Sánchez[1,2,3], Bram Burger[4,5], Carlos Horro[4,5], Antonio Fabregat[3], Stefan Johansson[1,2], Pål Rasmus Njølstad[1,6], Harald Barsnes[4,5], Henning Hermjakob[3,7], and Marc Vaudel[1,2,]*

[1] K.G. Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Norway

[2] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

[4] Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

[5] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

[6] Department of Pediatrics, Haukeland University Hospital, Bergen, Norway

[7] Beijing Proteome Research Center, National Center for Protein Sciences Beijing, Beijing, China

* To whom correspondence should be addressed

Summary

Mapping biomedical data to functional knowledge is an essential task in biomedicine and can be achieved by querying gene or protein identifiers in pathway knowledgebases. Here, we demonstrate that including fine-granularity information such as post-translational modifications greatly increases the specificity of the analysis. We present PathwayMatcher (github.com/PathwayAnalysisPlatform/PathwayMatcher), a bioinformatic application for mapping multi-omics data to pathways and show how this enables the building of biological networks at the proteoform level.

Keywords: Pathway, post-translational modification, network, proteoform

Main Text

In biomedicine, molecular pathways are used to infer the mechanisms underlying disease conditions and potential drug targets. Pathways are composed of series of biochemical reactions, of which the main participants are proteins, that together form a complex biological network (Burger et al., 2018). Proteins are known to exist in various forms, referred to as proteoforms (Smith et al., 2013), depending on the genetic profile of patients, sequence cleavage and folding, and post-translational modification (PTM) state. Proteoforms can carry PTMs at specific sites, conferring each proteoform unique structure and properties. Notably, many pathway reactions can only occur if all or some of the proteins involved are in specific post-translational states.

However, when analyzing omics data, both input and pathways are summarized in a gene- or protein-centric manner, meaning that the different proteoforms and their reactions are grouped by gene name or protein accession number, and the fine-grained structure of the pathways is thus lost. Here, we present PathwayMatcher, an open-source standalone application for mapping multiple types of omics data to pathways. PathwayMatcher is programmed in java, open source, freely available at github.com/PathwayAnalysisPlatform/PathwayMatcher under the permissive Apache 2.0 license. It is readily available for workflow integration *via* Bioconda (Grüning et al., 2018) and Galaxy (Afgan et al., 2018). Further information on the implementation, availability, performance, as well as test data sets can be found in supplementary information.

By taking advantage of the extensive annotation of Reactome (Fabregat et al., 2017), a free, open-source, curated knowledgebase containing documented biochemical reactions categorized in hierarchical pathways, PathwayMatcher enables the mapping of omics data to pathways at several biological levels: (i) genetic variants, (ii) genes, (iii) proteins, (iv) peptides, and (v) proteoforms (**Figure 1**).

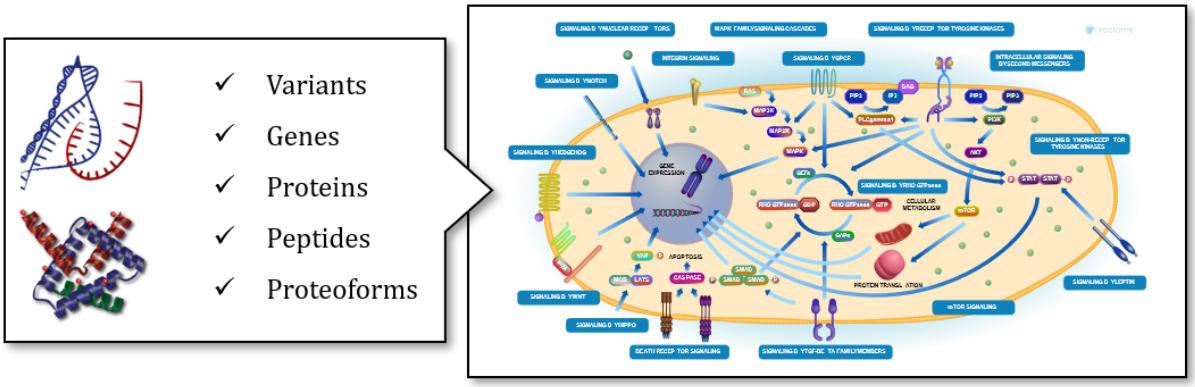

Figure 1: PathwayMatcher enables the mapping of omics data to pathways without loss of granularity. PathwayMatcher supports the mapping of multiple omics data types (left) to the Reactome database (right).

The ability to map proteoforms is particularly interesting, as it provides a one-to-one mapping to the elementary building blocks of pathways, and is achieved by using information about the post-translational state required for a protein to participate in a given reaction. As displayed in **Figure 2A**, the availability of information regarding proteoform interactions has dramatically increased during the past two decades, with 3,947 and 5,631 publications indexed in Reactome (Version 64 at time of writing) describing at least one reaction between modified proteins or between a modified and an unmodified protein, respectively. By harnessing this vast amount of knowledge, PathwayMatcher allows pinpointing precisely the pathways in which a proteoform participates. Proteoform pathway mapping is therefore more specific than gene- or protein-centric summarization approaches. Practical examples of omics data mapping for each category are available in the online documentation (github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki).

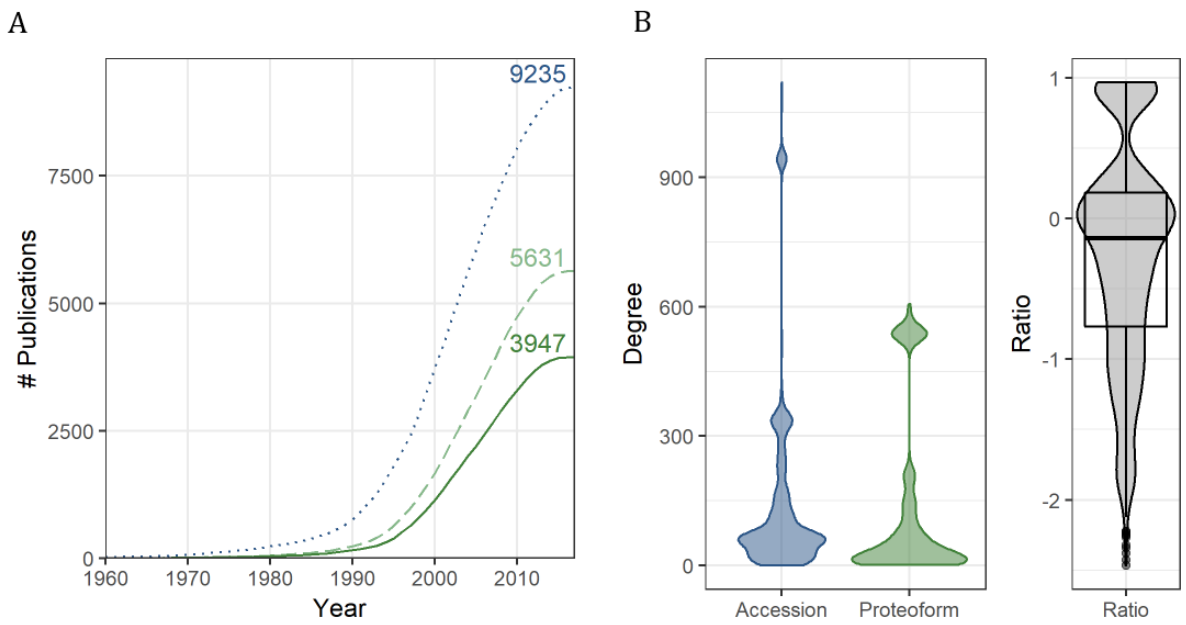

Figure 2: Building proteoform-centric network from the Reactome database. (A) The number of publications indexed in Reactome documenting at least one reaction between two proteins with PTMs (solid dark green line), between one protein with PTMs and one without (dashed light green line), and two proteins without PTMs (dotted blue line), counting all publications with a year earlier than or equal to the x-axis value. The number of publications in each category at time of writing (Reactome Version 64) is indicated to the right. (B) Degree distribution of the network of all proteoforms in Reactome mapped by accession (blue). proteoform (green), and the accession over proteoform degree ratio (black).

Pathways can be represented as a network where each reaction is indicated by a connection between its participants. In addition to standard over-representation analysis and mapping of input to pathways, PathwayMatcher allows building molecular networks based on the reactions between entities in the input (inner network) and reactions with other entities in Reactome (outer network). By conducting proteoform matching, PathwayMatcher then makes it possible to construct proteoform networks. In a gene-centric paradigm, the network comprises a limited number of nodes, approximately 20,000 for humans, whereas in a proteoform-centric paradigm, the human network is expected to have several million nodes (Aebersold et al., 2018). In the current version of Reactome, building the entire gene- and proteoform-centric networks produces 9,759, and 12,775 nodes with 443,229 and 672,047 connections, respectively. Within the proteoform network 9,694 nodes with 432,169 connections involve two accessions, 7,734 nodes with 95,539 connections involve one accession and one proteoform, and 2,806 nodes with 144,339 connections involve two proteoforms.

These results indicate that rather few proteoforms are currently annotated in pathways, less than 1% of the several millions expected (Aebersold et al., 2018). As displayed in **Figure 2B**, the degree, *i.e.* the number of connections, between proteoforms is globally reduced when matching by proteoform instead of accession. The degree ratio ranges from over 300-fold decrease to 10-fold increase, with a median ratio of 72%, corresponding to 28% less connections per node. Hence, in a proteoform network, connections otherwise aggregated per gene become distinguishable. One can therefore anticipate that proteoform-centric networks provide a rich paradigm to study biological systems. However, while gene networks have proven their ability to identify genes associated with diseases (Menche et al., 2015), networks of finer granularity remain largely unexplored.

In conclusion, PathwayMatcher is a versatile application enabling the mapping of several types of omics data to pathways. Thanks to the fine-grained information available in Reactome, PathwayMatcher supports refining the mapping to the level of proteoforms. To date, only limited information is available on the interactions between proteoforms, but as our understanding of protein interactions continues to increase, and as our ability to characterize them in samples progresses, proteoform-centric networks will surely become of prime importance in biomedical studies. Notably, the effect of genetic variation on genes, proteins, and proteoforms is currently only resolved for a fraction of the genome. The rapid development of this field suggests that it will be soon possible to specifically identify biological functions affected by variants within the proteoform network.

Acknowledgements

LFHS, SJ, PRN, and MV are supported by the European Research Council and the Research Council of Norway. BB, CH, and HB are supported by the Bergen Research Foundation. HB is also supported by the Research Council of Norway. LFHS, SJ, PRN, and MV are supported by the European Research Council and by the Research Council of Norway. This work has been supported by National Institutes of Health BD2K grant (U54 GM114833) and National Human Genome Research Institute at the National Institutes of Health Reactome grant (U41 HG003751).

Author Contributions

LFHS did all of the programming, testing, documentation, and wrote the manuscript. BB and CH contributed with programming, testing, documentation, ideas, and manuscript writing. AF, SJ, PRN, and HB contributed with ideas and manuscript writing. HH contributed with project design, manuscript writing, and supervised the work. MV contributed with project design, programming, documentation, testing, supervised the work, and wrote the manuscript. All authors participated in the preparation of the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

References

Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A.*, et al.* (2018). How many human proteoforms are there? Nature Chemical Biology *14*, 206.
Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A.*, et al.* (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research *46*, W537-W544.
Barabási, A.-L., and Pósfai, M. (2016). Network science (Cambridge university press).
Beck, F., Geiger, J., Gambaryan, S., Veit, J., Vaudel, M., Nollau, P., Kohlbacher, O., Martens, L., Walter, U., Sickmann, A.*, et al.* (2014). Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. Blood *123*, e1-e10.
Burger, B., Hernandez Sanchez, L.F., Lereim, R.R., Barsnes, H., and Vaudel, M. (2018). Analysing the structure of pathways and its influence on the interpretation of biomedical datasets. bioRxiv.
Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B.*, et al.* (2017). The Reactome Pathway Knowledgebase. Nucleic Acids Res.
Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., and Köster, J. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods *15*, 475-476.
Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. Science *347*.
Smith, L.M., Kelleher, N.L., and The Consortium for Top Down, P. (2013). Proteoform: a single term describing protein complexity. Nature methods *10*, 186-187.
The UniProt, C. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res *45*, D158-D169.