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

Abstract

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.

Main Text

In biomedicine, molecular pathways are used to infer the mechanisms underlying disease conditions and infer 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[1]. Proteins are known to exist in various forms, referred to as proteoforms[2], 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 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 (github.com/PathwayAnalysisPlatform/PathwayMatcher), an open-source standalone application for mapping multiple types of omics data to pathways without loss of granularity. PathwayMatcher is readily available for workflow integration *via* Bioconda[3] and Galaxy[4]. We demonstrate how this high-specificity mapping provides refined pathway search results and enables the building of proteoform-centric biological networks. Finally, we provide the first comparison between gene- and proteoform-centric networks.

PathwayMatcher relies on Reactome[5], a free, open-source, curated knowledgebase containing documented biochemical reactions categorized in hierarchical pathways. Proteins in Reactome are defined according to UniProt[6] following a gene-centric paradigm. 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 1A**).

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 1B**, 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.

Compared to protein accession matching, this proteoform-level matching reduces the number of pathways mapped per entity, with a median number of pathways per entity of four for proteoforms *versus* eleven for accession numbers (**Figure 1C**). Proteoform pathway mapping is therefore substantially more specific than the current gene- or protein-centric summarization approach. Practical examples of omics data mapping for each category are available in the online documentation (github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki).

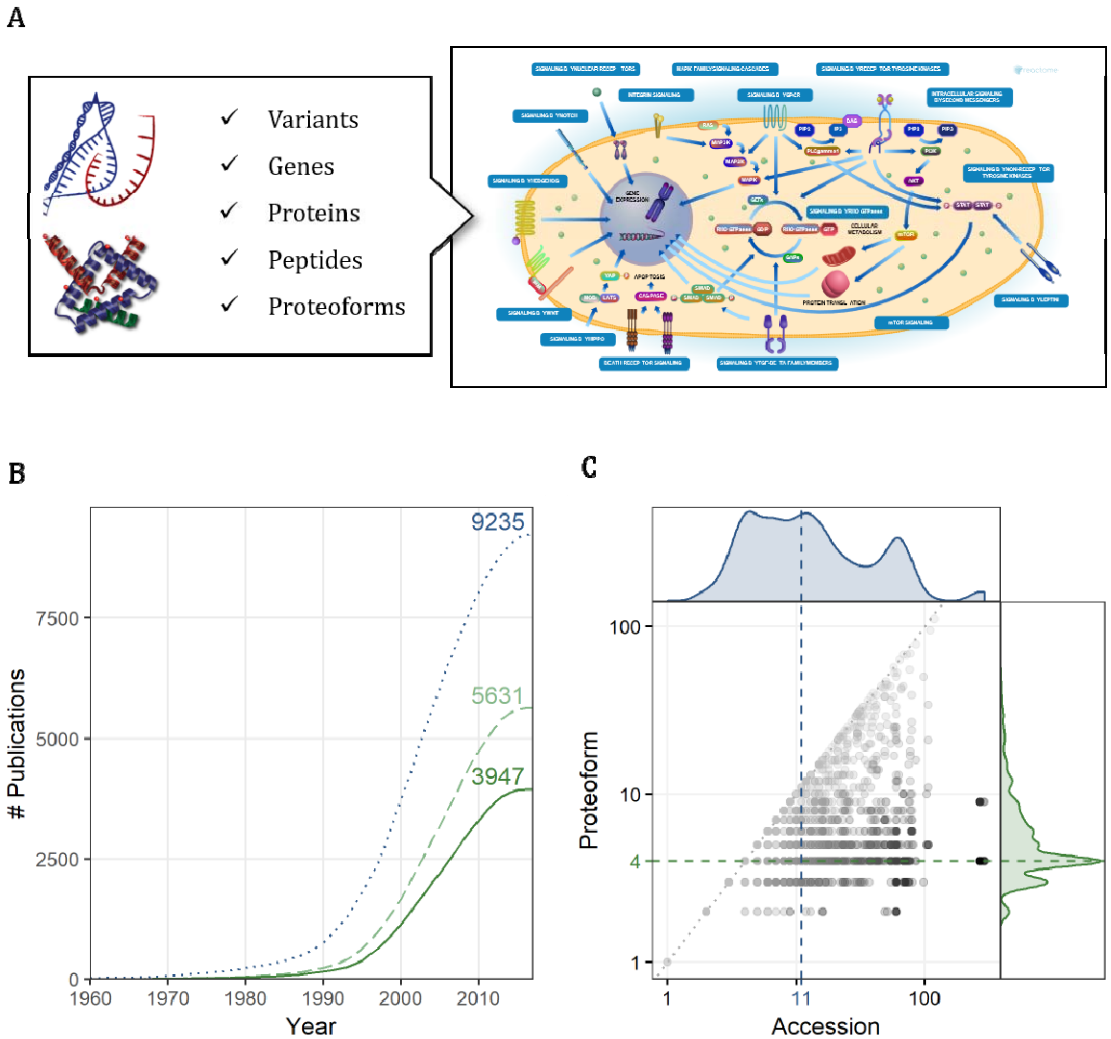


Figure 1: PathwayMatcher enables the mapping of omics data to pathways without loss of granularity. (A) PathwayMatcher supports the mapping of multiple omics data types (left) to the Reactome database (right). (B) The number of publications indexed in Reactome documenting at least one reaction between 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. (C) For all modified protein annotated in Reactome, the number of pathways mapped using the proteoform versus its protein accession is plotted in black. The density of the number of pathways mapped are indicated at the top (blue) and right (green) for accession and proteoform matching, respectively. The median number of pathways mapped are indicated with dashed lines.

Pathways can be represented as a network where each reaction is indicated by a connection between its participants. 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[7]. Consequently, connections between different proteoforms, otherwise aggregated per gene, become distinguishable. One can therefore anticipate that proteoform-centric networks provide a more fine-grained model to study biological systems. However, while gene networks have proven their ability to identify genes associated with diseases[8], networks of finer granularity remain largely unexplored.

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, it is then possible to construct proteoform networks. To illustrate the difference between proteoform networks and the more common gene-centric networks, we mapped a public dataset of phosphorylation changes associated with cAMP/PKA-dependent signaling in platelets[9]. **Figure 2A** shows the main component of the network obtained by accession matching with PathwayMatcher. According to Reactome, of the 378 connections, only 29 (7.7%, solid green lines) are documented with the actual proteoforms provided, 85 (22%, solid blue lines) are reported with other proteoforms, and 264 (70%, dotted red lines) do not involve phosphorylated versions of these proteins. Consequently, analyzing the data as a gene-centric network relies essentially on interactions between the non-phosphorylated forms of the proteins, or other proteoforms, which may have a completely different biological function, while adding information at the proteoform level allows the identification of clusters of interacting proteoforms.

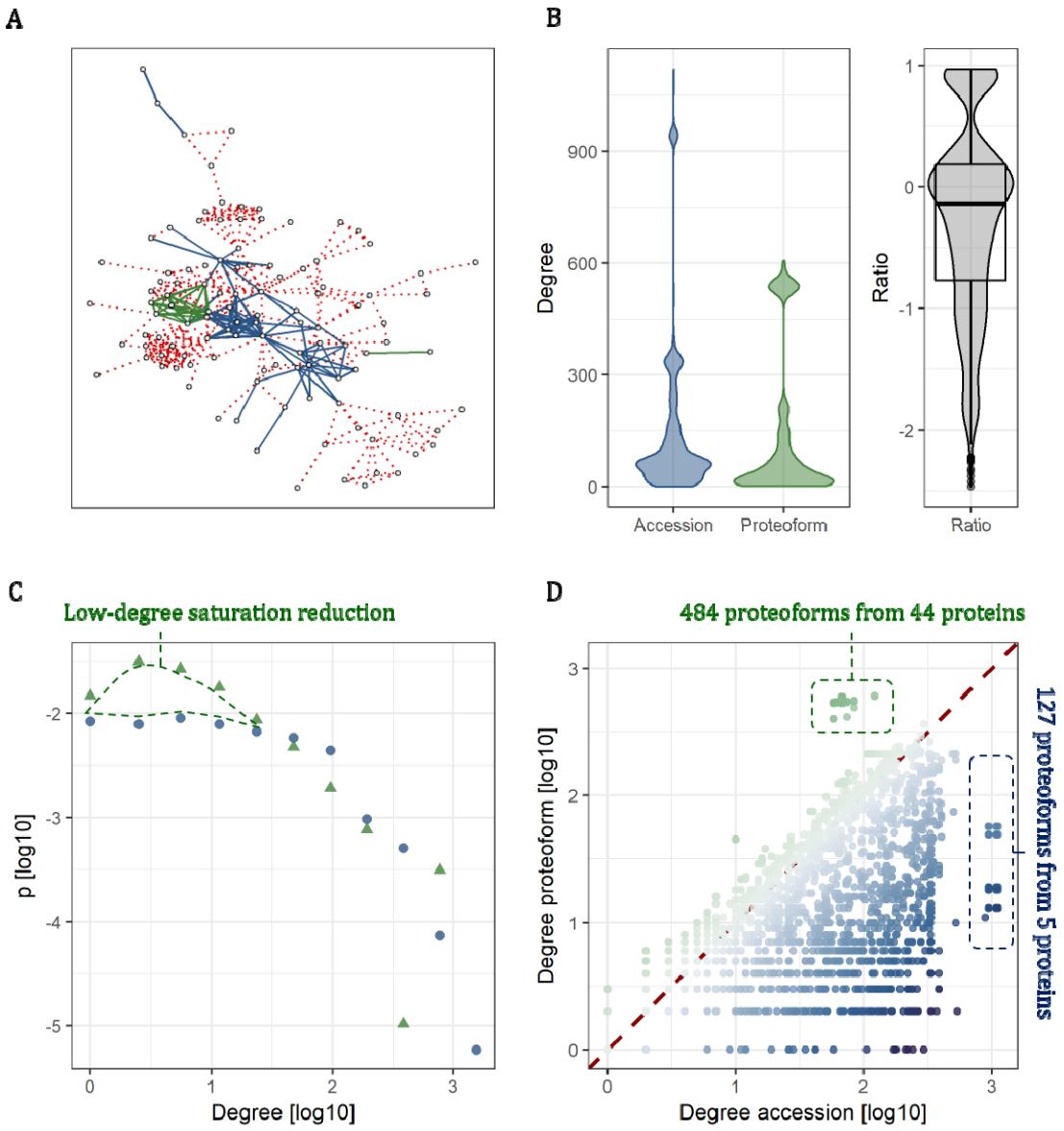


Figure 2: Comparison of accession-centric, i.e. gene-centric, versus proteoform-centric mapping of omics data to pathways. (A) The main component of a network generated by accession matching from a public phosphoproteomics dataset; connections not documented using proteoforms as dashed red lines, connections documented using proteoforms but with proteoform mismatch as solid blue lines, and connections documented with the exact same proteoforms as solid green lines. (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). (C) The probability of finding a node of a given degree in the network built in B when mapping by accession (blue dots) or by proteoform (green triangles); the reduction of low-degree saturation in the network is highlighted with a green area. (D) The degree of each node in the network built in B with proteoform versus accession mapping, colored with a blue-grey-green gradient corresponding to the ratio in B.

On a global scale, 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[7]. 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.

Figure 2C displays the degree probability for gene- and proteoform-centric networks built using all reactions involving proteoforms only. The gene-centric network shows a classical degree probability for gene networks[10]: starting with a plateau, the low-degree saturation, followed by a linear decrease characteristic of a power law, and finally a drop at the highest degrees, the high degree cut-off. In comparison, as highlighted on the left-side of the plot, the proteoform-centric network presents a reduced low-degree saturation, indicative of a higher prevalence of specialized entities. Moreover, as can be seen on the right-side of the plot, the high-degree cut-off is much more pronounced, showing a dramatic decrease in the prevalence of large hubs, with the notable exception of nodes with close to 1,000 connections.

As illustrated in **Figure 2D**, this exception is due to approximately 484 proteoforms originating from 44 proteins. Interestingly, these proteoforms are almost exclusively participating in extracellular matrix organization and are not outliers when considering the gene-centric network. Conversely, the five outliers in the gene-centric degree distribution with around 1,000 connections, break down into 127 proteoforms in the proteoform-centric network, each with a degree between 10 and 100. We hypothesize that the structural difference between gene- and proteoform-centric networks can shed light on how biological networks, built from a limited set

of genes, specialize to achieve a vast number of very specific functions. PathwayMatcher therefore makes it possible for researchers to study post-translational level mechanisms that would otherwise be lost in data summarization.

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. Our findings further indicate that gene- and proteoform-centric networks present structural differences that may carry important biological meaning. At the moment, 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

1 Burger, B., Hernandez Sanchez, L. F., Lereim, R. R., Barsnes, H. & Vaudel, M. Analysing the structure of pathways and its influence on the interpretation of biomedical datasets. *bioRxiv*, doi:10.1101/333492 (2018).

2 Smith, L. M., Kelleher, N. L. & The Consortium for Top Down, P. Proteoform: a single term describing protein complexity. *Nature methods* **10**, 186-187, doi:10.1038/nmeth.2369 (2013).

3 Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods* **15**, 475-476, doi:10.1038/s41592-018-0046-7 (2018).

4 Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* **46**, W537-W544, doi:10.1093/nar/gky379 (2018).

5 Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res*, doi:10.1093/nar/gkx1132 (2017).

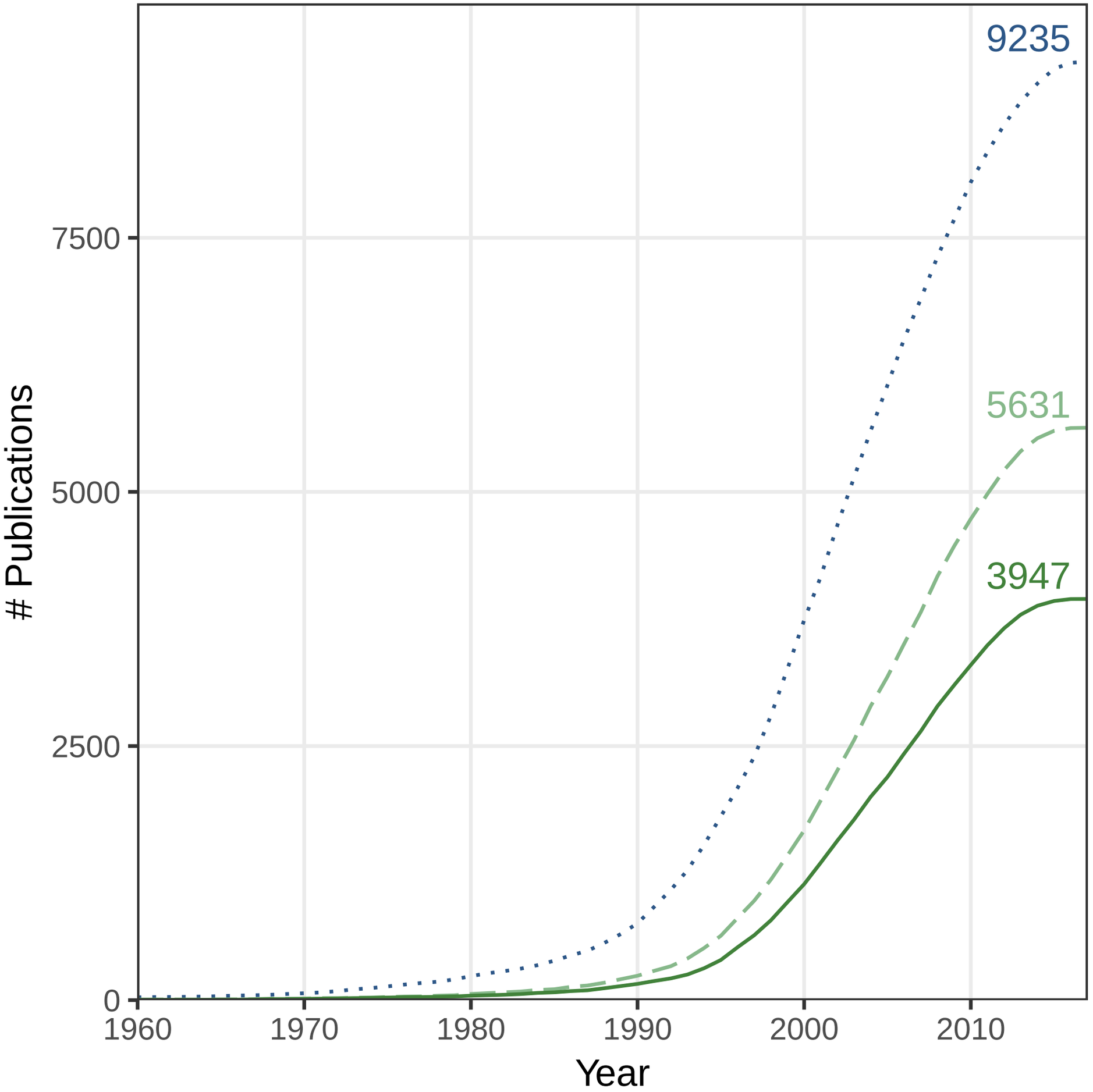
6 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

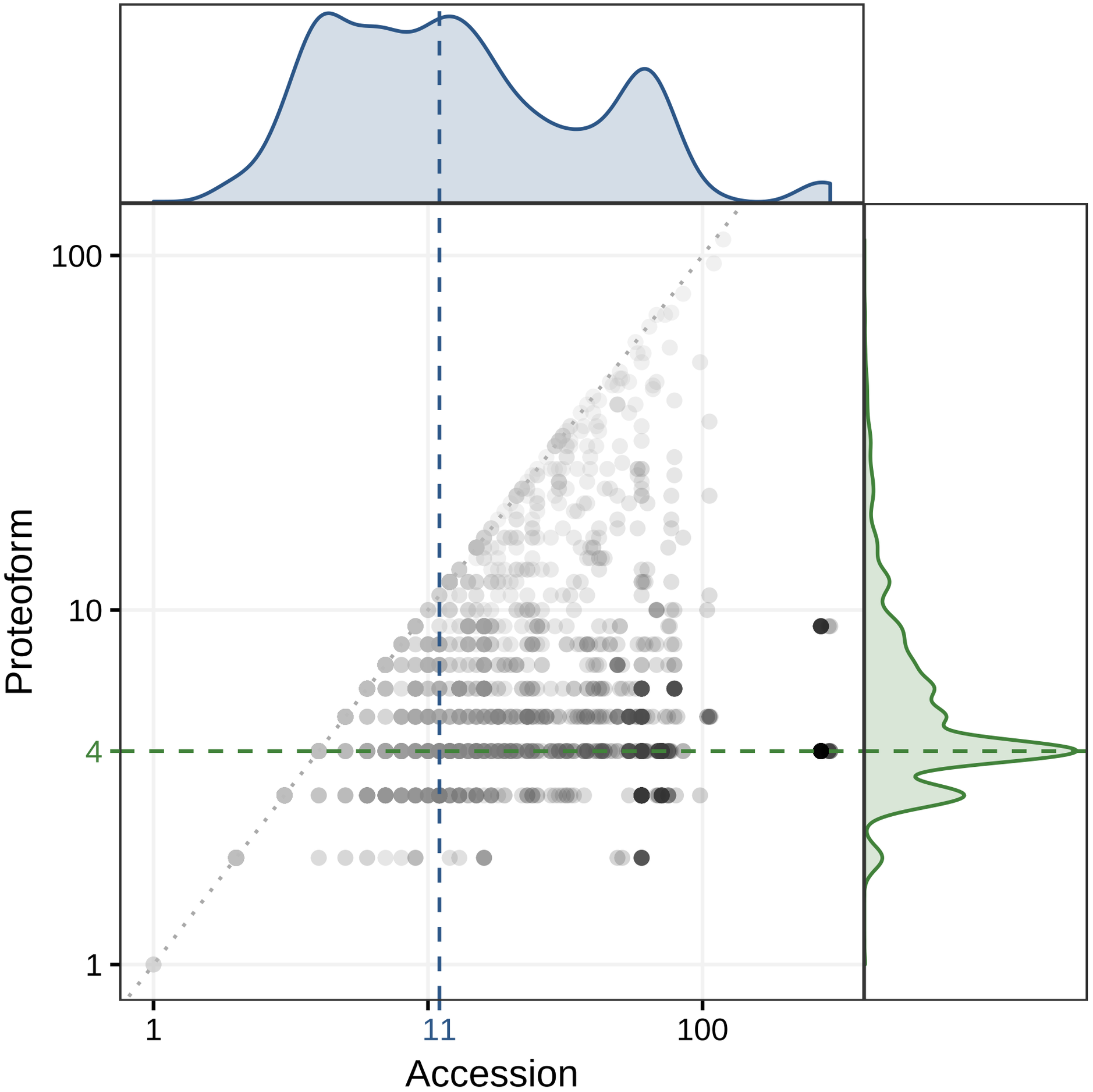
7 Aebersold, R. *et al.* How many human proteoforms are there? *Nature Chemical Biology* **14**, 206, doi:10.1038/nchembio.2576 (2018).

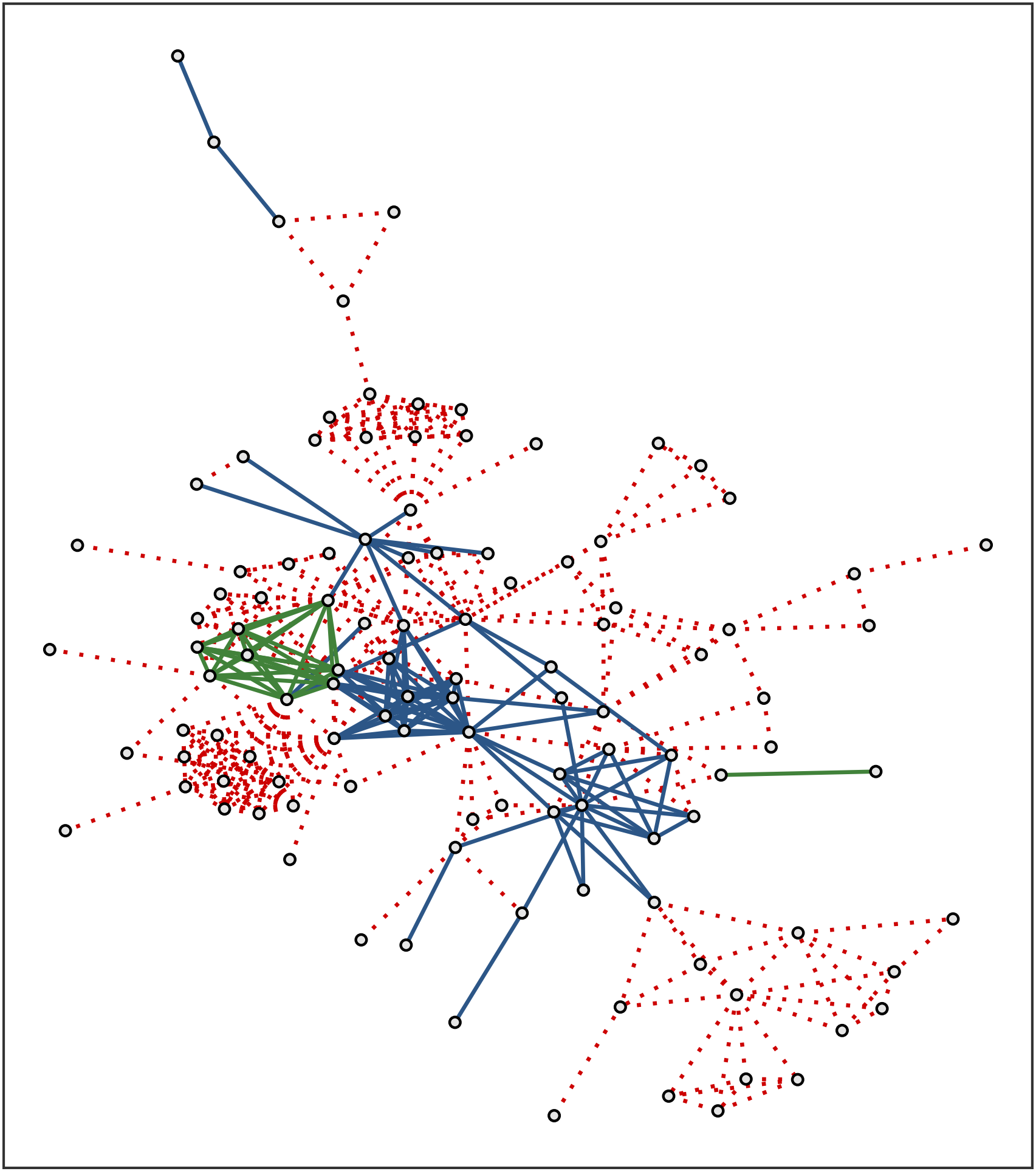
8 Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, doi:10.1126/science.1257601 (2015).

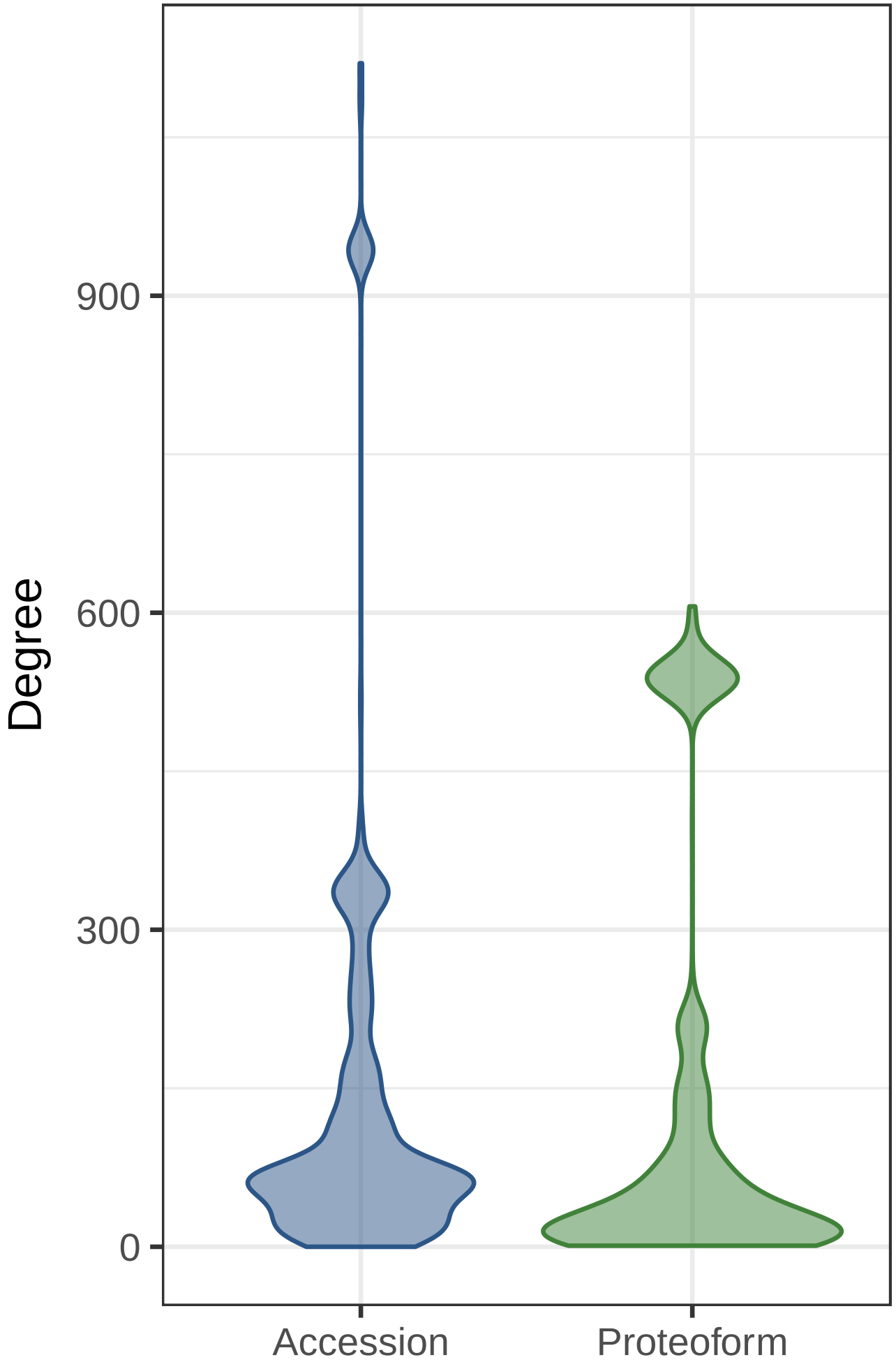
9 Beck, F. *et al.* Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. *Blood* **123**, e1-e10, doi:10.1182/blood-2013-07-512384 (2014).

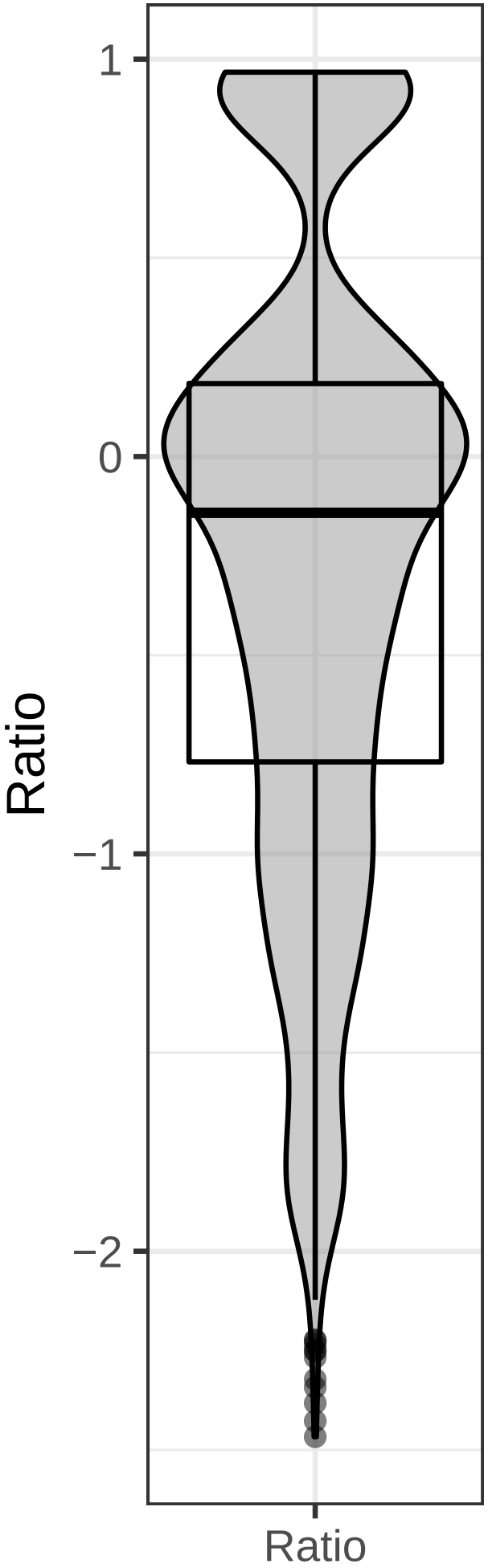
10 Barabási, A.-L. & Pósfai, M. *Network science*. (Cambridge university press, 2016).

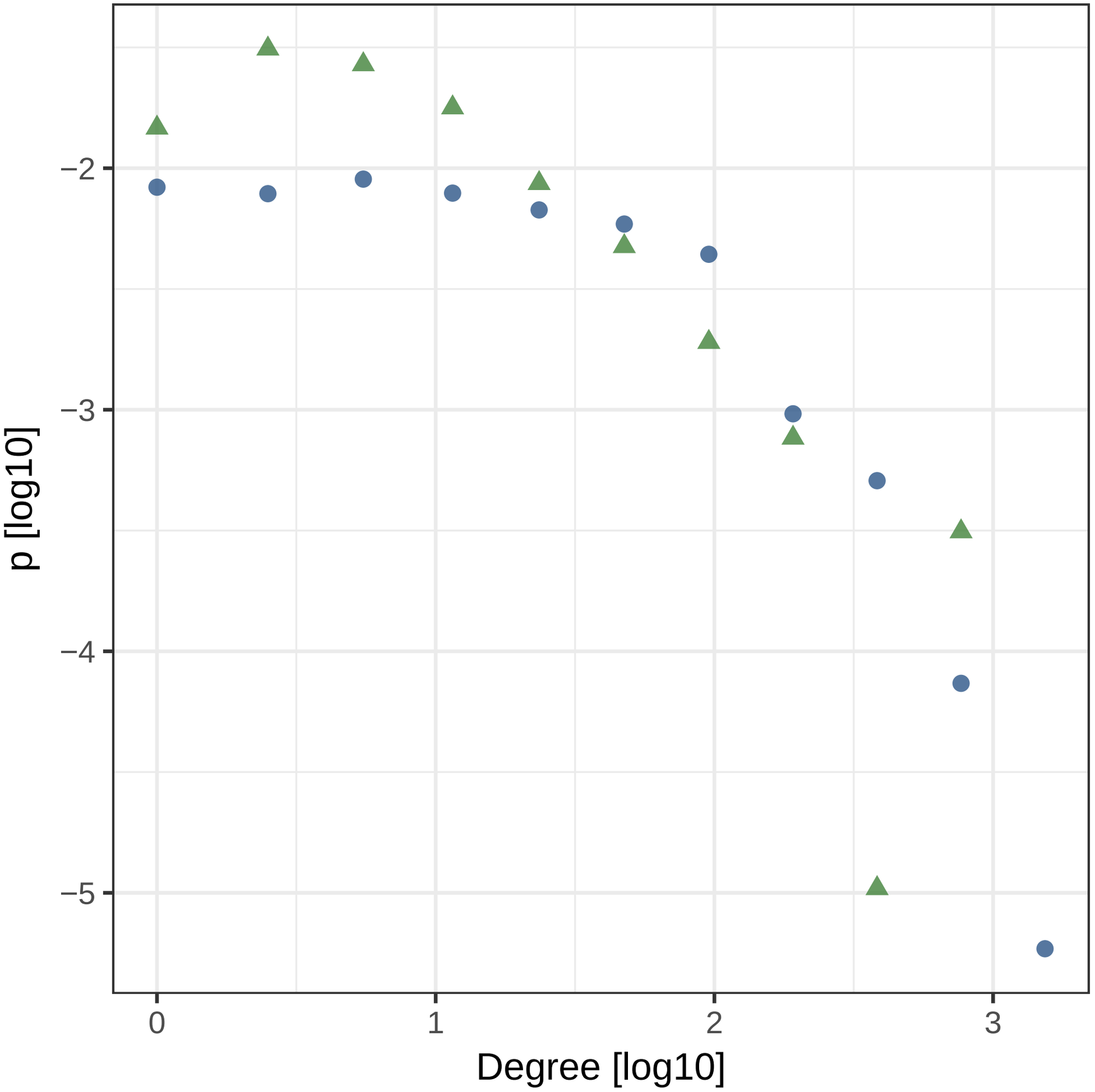

9235
5631
3947
7500
5000
2500
0
Publications
1960
1970
1980
1990
2000
2010
Year

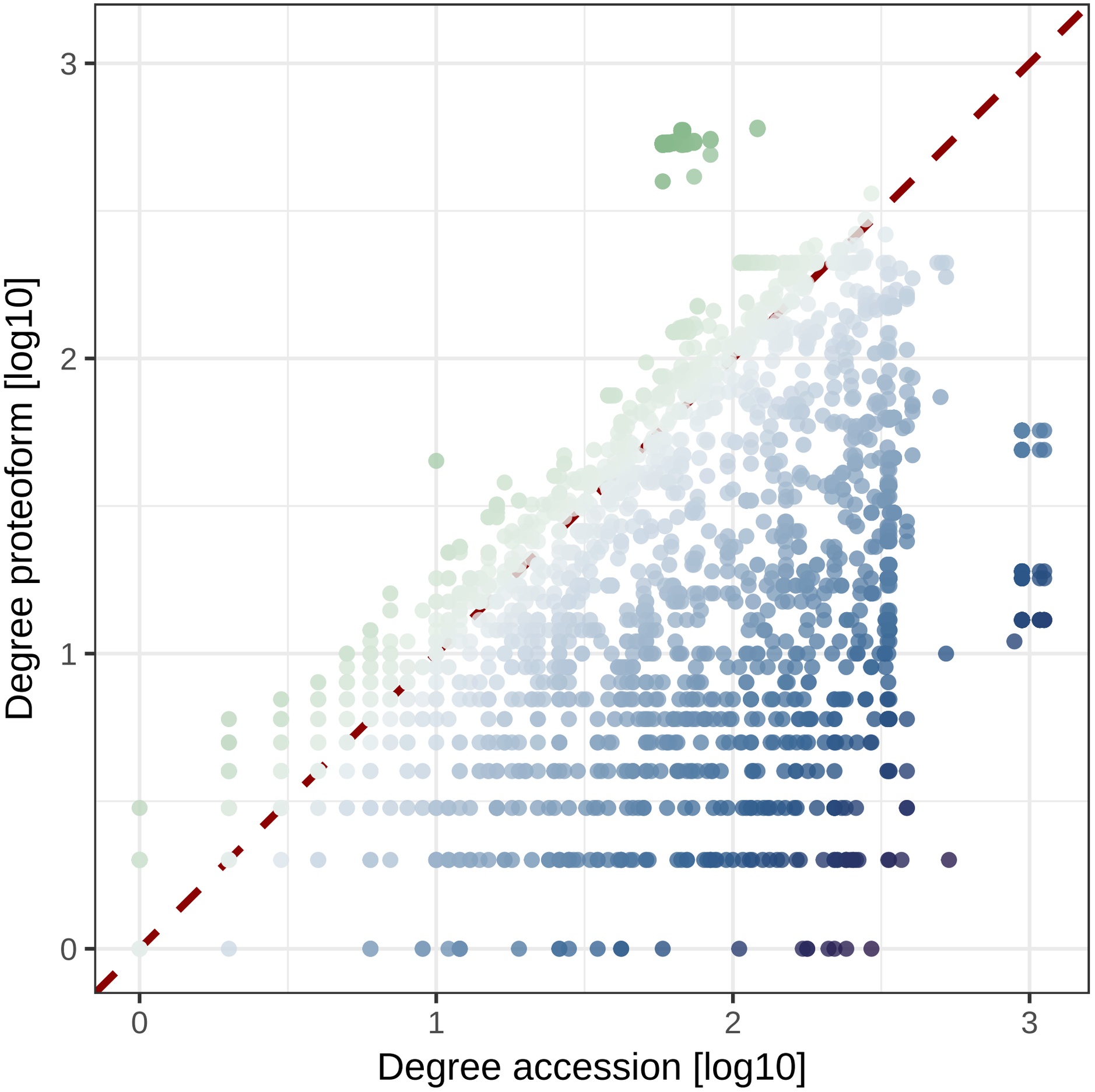

100
10
4
1
Proteoform
1
11
100
Accession




Degree
900
600
300
0
Accession
Proteoform


1
0
−1
−2
Ratio
Ratio


−2
−3
−4
−5
p [log10]
0
1
2
3
Degree [log10]


Degree proteoform [log10]
Degree accession [log10]
0
1
2
3
0
1
2
3