PSICQUIC and PSISCORE: accessing and scoring molecular interactions

Bruno Aranda[1,27], **Hagen Blankenburg**[2,27], **Samuel Kerrien**[1], **Fiona S L Brinkman**[3], **Arnaud Ceol**[4,5], **Emilie Chautard**[6,7], **Jose M Dana**[1], **Javier De Las Rivas**[8], **Marine Dumousseau**[1], **Eugenia Galeota**[5,9], **Anna Gaulton**[1], **Johannes Goll**[10], **Robert E W Hancock**[11], **Ruth Isserlin**[12], **Rafael C Jimenez**[1], **Jules Kerssemakers**[13], **Jyoti Khadake**[1], **David J Lynn**[14], **Magali Michaut**[12], **Gavin O'Kelly**[1], **Keiichiro Ono**[15], **Sandra Orchard**[1], **Carlos Prieto**[8,16], **Sabry Razick**[17,18], **Olga Rigina**[19], **Lukasz Salwinski**[20], **Milan Simonovic**[21], **Sameer Velankar**[1], **Andrew Winter**[22], **Guanming Wu**[7], **Gary D Bader**[12], **Gianni Cesareni**[5,9], **Ian M Donaldson**[17,23], **David Eisenberg**[20,24,25], **Gerard J Kleywegt**[1], **John Overington**[1], **Sylvie Ricard-Blum**[6], **Mike Tyers**[22,26], **Mario Albrecht**[2], and **Henning Hermjakob**[1]

Bruno Aranda: baranda@ebi.ac.uk

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK [2]Max Planck Institute for Informatics, Saarbrücken, Germany [3]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada [4]Institute for Research in Biomedicine, Barcelona, Spain [5]Department of Biology, University of Rome Tor Vergata, Rome, Italy [6]Institut de Biologie et Chimie des Protéines, Unité Mixte de Recherche 5086, Centre National de la Recherche Scientifique–Université Lyon 1, Lyon, France [7]Ontario Institute for Cancer Research, Toronto, Ontario, Canada [8]Cancer Research Center, Centro de Investigación de Cáncer–Instituto de Biología Molecular y Celular del Cáncer, Consejo Superior de Investigaciones Científicas, Universidad de Salamanca, Salamanca, Spain [9]Istituto di Ricovero e Cura a Carattere Scientifico, Fondazione S. Lucia, Rome, Italy [10]J. Craig Venter Institute, Rockville, Maryland, USA [11]Centre for Microbial Diseases and Immunity Research, University of British Columbia, Vancouver, British Columbia, Canada [12]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada [13]Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands [14]Animal and Bioscience Research Department, Animal and Grassland Research Innovation Centre, Teagasc, Ireland [15]University of California, Trey Ideker Lab, San Diego, School of Medicine, La Jolla, California, USA [16]Institute of Biotechnology of León, León, Spain [17]The Biotechnology Centre of Oslo, University of Oslo, Oslo, Norway [18]Biomedical Research Group, Department of Informatics, University of Oslo, Oslo, Norway [19]Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, Kongens Lyngby, Denmark [20]University of California, Los Angeles, Department of Energy Institute for Genomics and Proteomics, Los Angeles, California, USA [21]Faculty of Science, Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland [22]Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, Scotland [23]Department for Molecular Biosciences, University of Oslo, Oslo, Norway [24]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California, USA

[27]These authors contributed equally to this work.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

[25]Howard Hughes Medical Institute, University of California, Los Angeles, California, USA
[26]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

To the Editor

To study proteins in the context of a cellular system, it is essential that the molecules with which a protein interacts are identified and the functional consequence of each interaction is understood. A plethora of resources now exist to capture molecular interaction data from the many laboratories generating such information, but whereas such databases are rich in information, the sheer number and variability of such databases constitutes a substantial challenge in both data access and quality assessment to the researchers interested in a specific biological domain.

Integrating data from these disparate resources remained a challenge until 2004, when the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) released the PSI molecular interaction (MI) XML format, a community standard for the representation of molecular-interaction data. To concomitantly standardize annotation across the different databases, they also developed a controlled vocabulary enabling a detailed but consistent description of molecular interactions[1]. A simplified, standardized format for interaction data, the Molecular Interaction Tabular format (MITAB), is also available[2]. PSI-MI formats are now broadly accepted and widely implemented by over 30 databases and supported by key software tools.

The PSI-MI formats facilitate the integration of molecular interaction data from multiple sources, both by the user community and by dedicated software tools. However, users must still first collect data from each of the individual databases, which typically involves different queries at multiple websites or downloading data files from different web servers. Additionally, the retrieved data has then to be kept up to date with each release of the originating database. This challenge has led to the development of the PSI common query interface (PSICQUIC), a community standard for computational access to molecular-interaction data resources.

All data sources implementing PSICQUIC can be queried in the exact same way. Formulating the query once is sufficient to retrieve the relevant data from many interaction data sources. Independently published observations of an experimental system, curated by independent databases, are then integrated in response to a user query (Fig. 1). A PSICQUIC query can be a simple protein identifier or a complex construct using the syntax defined by the molecular interaction query language (MIQL) (Supplementary Note 1).

The existence of an open-source reference implementation for PSICQUIC allows the rapid setup of a local server for interaction data with limited effort. The PSICQUIC project site (http://psicquic.googlecode.com/) offers open-source client libraries and code examples, facilitating programmatic access to the PSICQUIC registry and services. Thus, PSICQUIC can be easily integrated with third-party applications. For instance, it is used by Cytoscape[3] to query multiple web services at the same time for rendering the resulting interaction networks. PSICQUIC is also used by the International Molecular Exchange consortium (IMEx) to facilitate high-quality, nonredundant data sharing (unpublished data).

As a result, more than 16 million interactions are already accessible from 16 PSICQUIC services (Supplementary Table 1), which includes servers hosted by most major molecular interaction providers. All these services are listed in the PSICQUIC registry. Each service is classified by tags from a controlled vocabulary, which help the user to select the services of

interest. The PSICQUIC architecture even allows seamless integration of commercial data sources with publicly available sources, based on access privileges of end users.

Another challenge in the field of molecular interactions is varying data quality. Owing to the diversity of techniques for experimental detection, computational prediction and curation of interaction data, adequate quality assessment methods have to account for the different evidence associated with each reported interaction. An interaction of two proteins can be supported, for example, by a single concurrent mention in a scientific publication or by multiple independent experimental observations, including details such as the protein-binding interface or assay parameters. Consequently, researchers require a system to retrieve confidence scores for user-defined sets of molecular interactions. This led to the development of the PSI confidence scoring system (PSISCORE) based on an earlier study[4] (Supplementary Note 2).

Confidence measures for molecular interactions can use different, potentially complementary, properties of biological systems. Evidence-based confidence scores are commonly derived from the applied experimental detection technique or based on standard reference sets, functional annotations, evolutionary conservation, structural knowledge, literature support or network topology. The diversity of confidence measures raises questions about their comparison and combination. To date, the community has not agreed on a generally accepted common scoring scheme for molecular interactions[5]. Therefore, PSISCORE is based on the concept of decentralization, where individual scoring servers can apply different scoring methods for assessing diverse biological and methodological aspects of interaction data (Fig. 1).

The start and end point of a PSISCORE use case is a user-defined PSI-MI file that describes a set of molecular interactions. The interaction data can be the result of a previous PSICQUIC query (Supplementary Note 3) or contain publicly available experimental interactions and unpublished or computationally predicted results. PSISCORE can also be integrated into existing workflows as a quality filter to add the computed confidence scores to the PSI-MI file. It is easy to programmatically access PSISCORE or to incorporate the user's own confidence scoring servers using the open-source libraries and the documentation at http://psiscore.googlecode.com/. All available scoring servers and their scoring methods are listed and described in the PSISCORE registry.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Hermjakob H, et al. Nat Biotechnol. 2004; 22:177–183. [PubMed: 14755292]

2. Kerrien S, et al. BMC Biol. 2007; 5:44. [PubMed: 17925023]

3. Shannon P, et al. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]

4. Blankenburg H, et al. Bioinformatics. 2009; 25:1321–1328. [PubMed: 19420069]

5. Orchard S, et al. Proteomics. 2007; 7:3436–3440. [PubMed: 17907277]
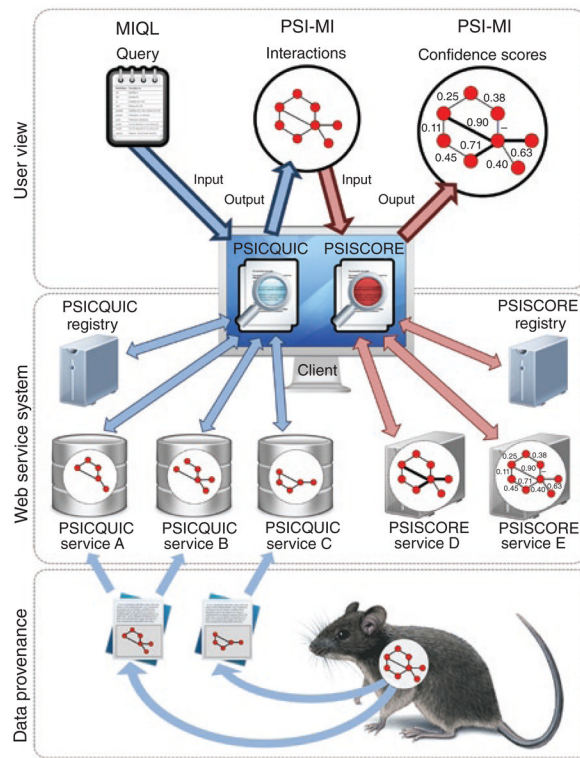
Figure 1.
PSICQUIC and PSISCORE architecture. A given biological system (sample) is observed by different experimental technologies, resulting in different publications reporting different, potentially partial, observations. A publication is potentially curated by more than one database. A PSICQUIC application sends a user query formulated in MIQL to all currently available PSICQUIC servers. Responses in unified PSI-MI format allow the PSICQUIC application to assemble a complete network view of the originally observed system. A given interaction network can be scored by multiple PSISCORE servers, each of them implementing one or more scoring methods (here symbolized by different line thickness (PSISCORE service D) and numbers (PSISCORE service E). The PSISCORE client application then presents the combined results to the user.