

InSoLiTo project: The research software graph-based network from OpenEBench

Sergi Aguiló-Castillo^{1,2}, José M. Fernández^{1,2}, Josep Ll. Gelpí^{1,2,3} and Salvador Capella-Gutiérrez^{1,2}

¹ Spanish National Bioinformatics Institute (INB/ELIXIR-ES)

² Barcelona Supercomputing Center (BSC)

³ Dept Biochemistry and Molecular Biomedicine, Universitat de Barcelona (UB)



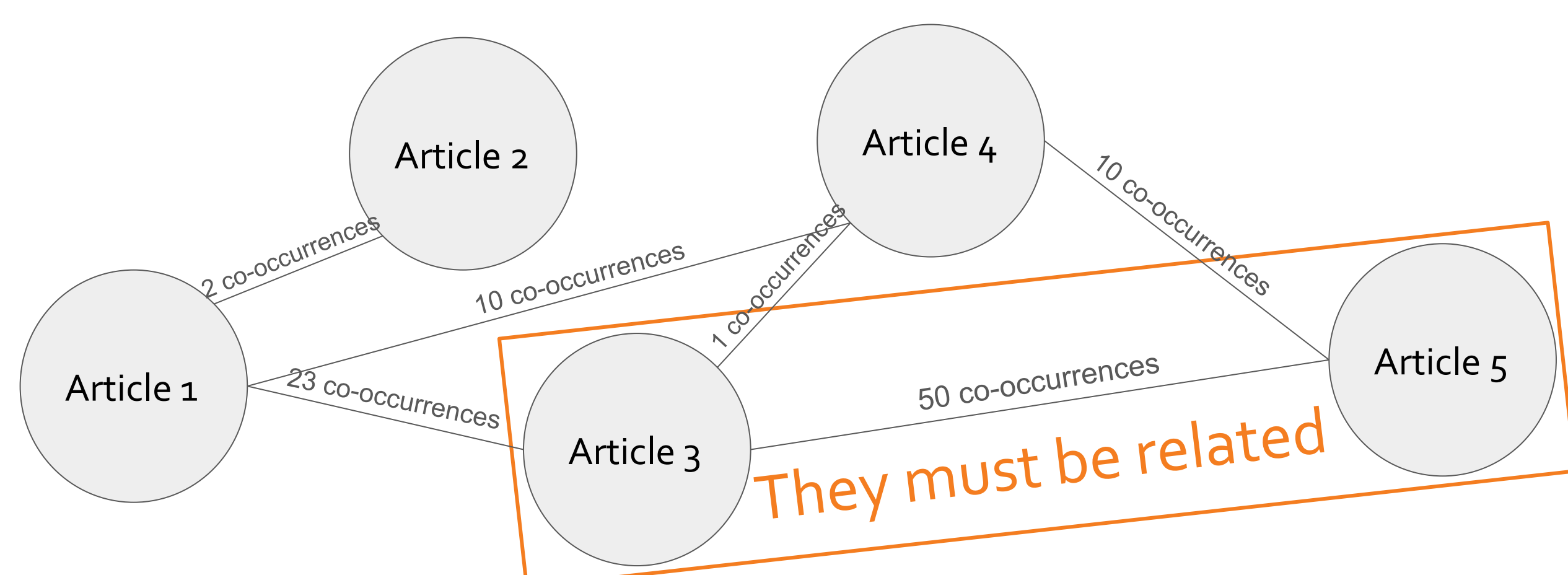
ELIXIR All Hands 2022, 7-10 June, Amsterdam, Netherlands



InSoLiTo is a graph-based network of the co-usage of research software from the tools collection hosted by OpenEBench. Co-usage is understood as being cited in the same scientific publications. The initial aim of the project is to identify **how bioinformatics software relate to each other** allowing to **potentially infer analytical workflows** commonly used from the literature.

Development

Calculate the Co-occurrences: How many scientific publications two articles have been referenced together.



The most direct way of knowing how the bioinformatics community uses their software is by looking to the **references** in their scientific articles.

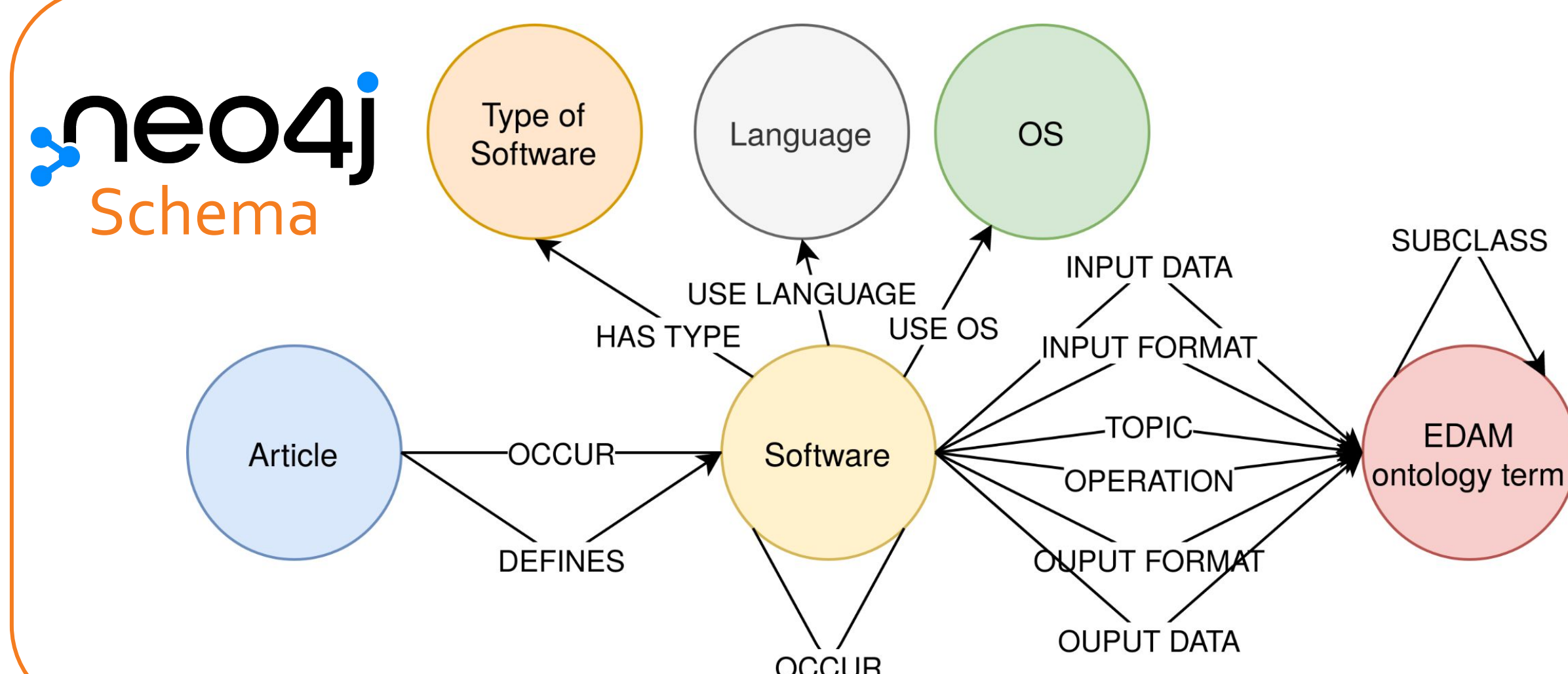
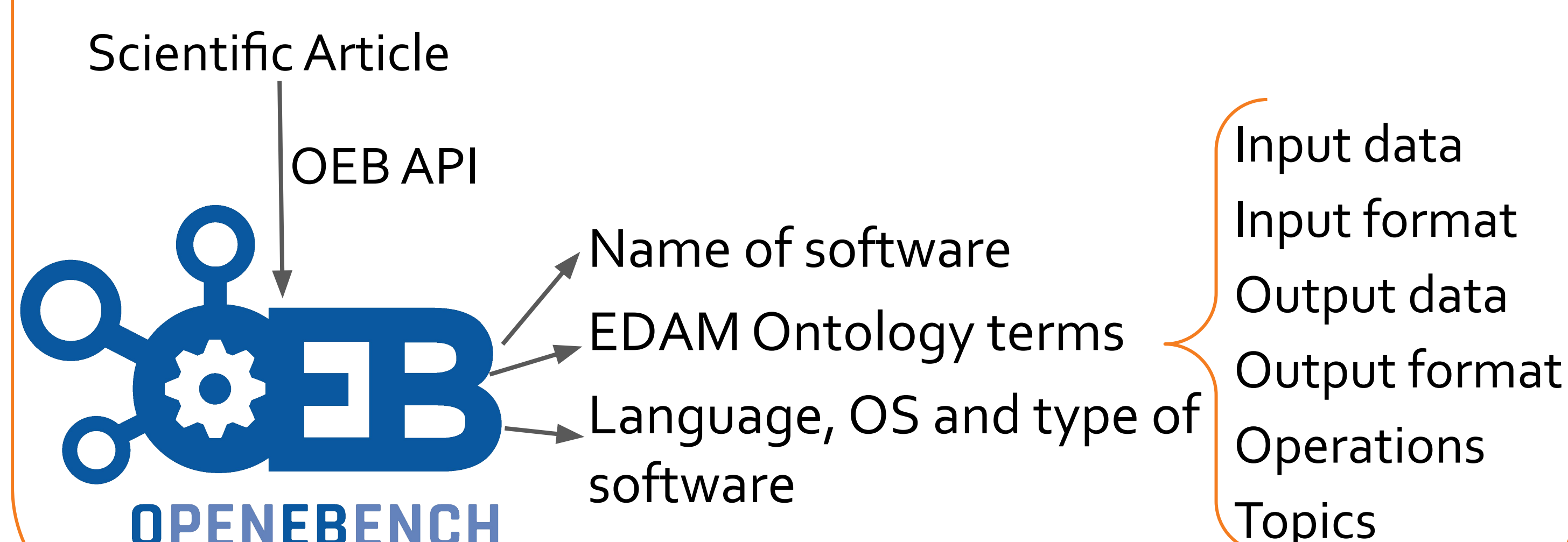
The theory is simple:

Two articles are closely related when there is high co-occurrence between them.

Transform Articles to Research Software

The tools, databases and services stored in the OpenEBench Technical Monitoring database, among other information, include references to the articles associated to the software.

The **OpenEBench API** is used to assign the articles to the cited software and collects the additional information stored in OEB.

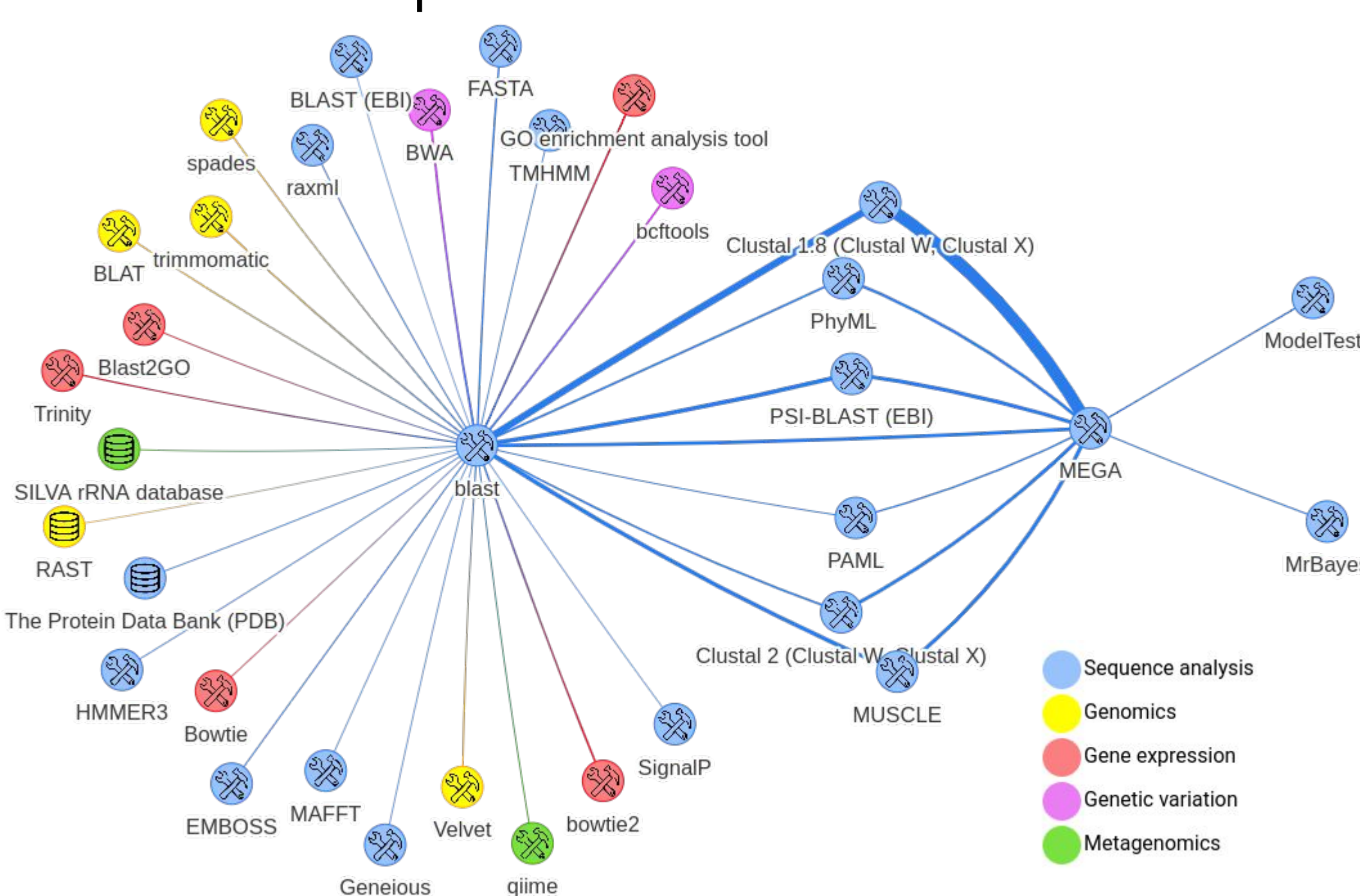


Graph database

Input data is **filtered by the minimum number of co-occurrences** (in this case >10 co-occurrences) to remove unrelated or random relationships between the data. Data is stored in a Neo4j **property graph database**, and clustered with the Louvain approach.

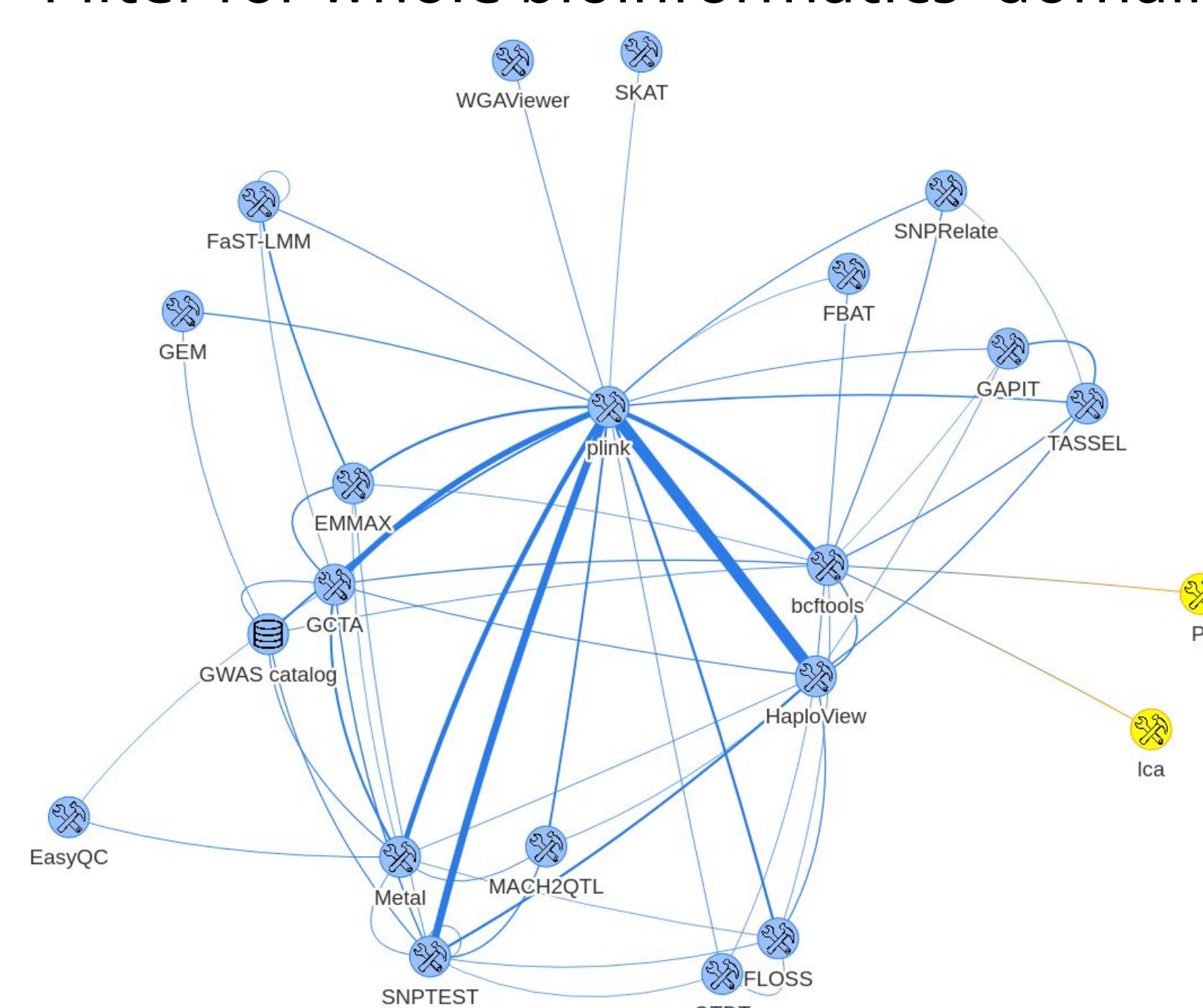
Interestingly, clusters show a high correlation with EDAM Topic (Ex: 71% of Proteomics software is in the same cluster).

Filter by software, see common relationships and the main topic of their cluster.



Filter by Blast and MEGA with >400 co-occurrences in the relationships

User Interaction
Filter for whole bioinformatics' domain



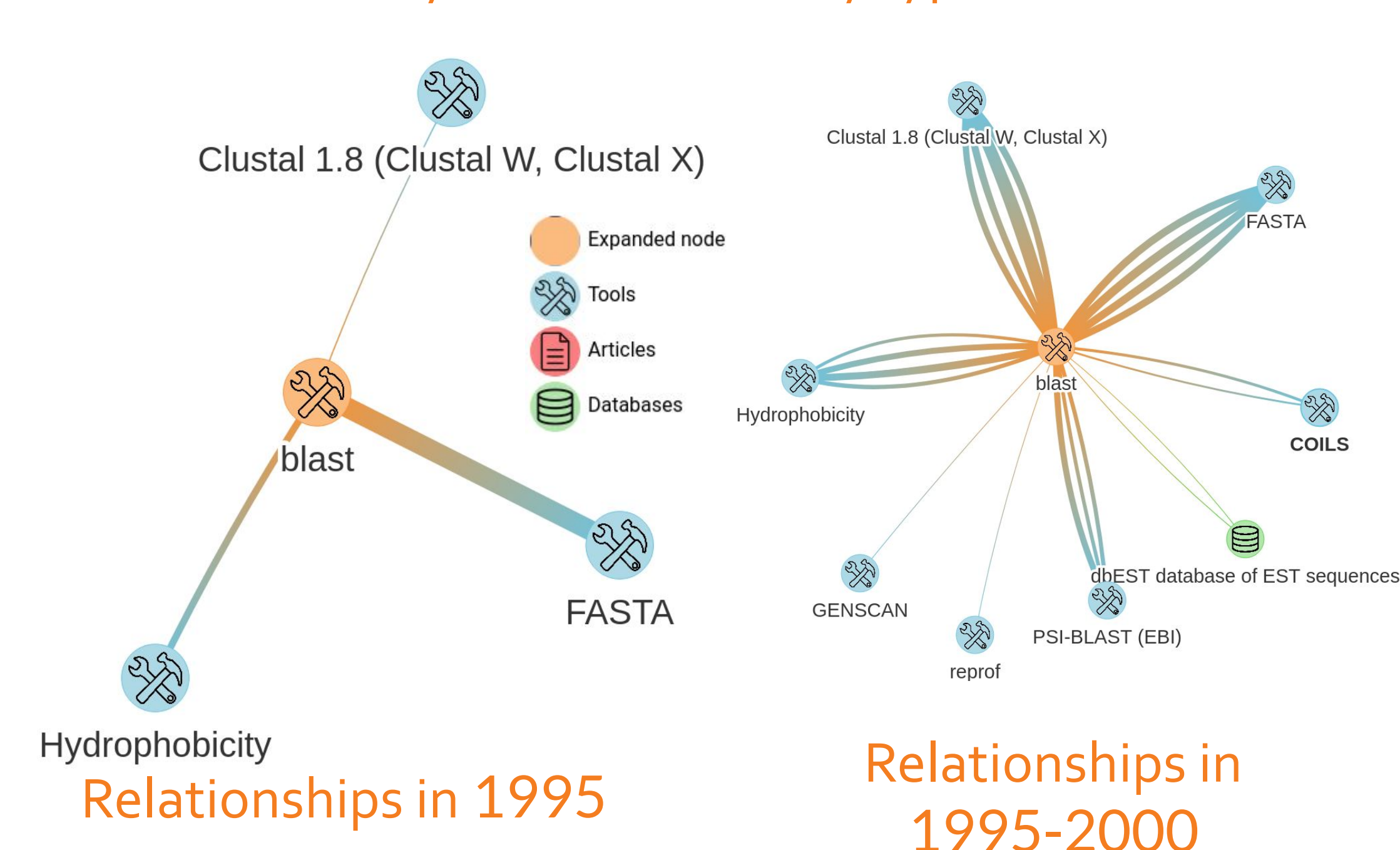
Filter by GWAS Study

Why it is necessary?

Results can be used by research communities **organizing scientific benchmarks** on popular pipelines in their domains. Include widely used databases, popular articles and other relevant resources as part of those workflows.

Filter per year to see the evolution of the usage of other software in relation to the selected tool

Filter by Blast. Colored by type of node



Contact

Sergi Aguiló-Castillo
Junior Research Engineer
INB/ELIXIR-ES
sergi.aguiló@bsc.es



insolito.openebench.bsc.es



These projects have received funding from the European Union's Horizon 2020 research and innovation programme

