Metadata of the chapter that will be visualized in SpringerLink

Book Title	Bioinformatics and Biomedical Engineering
Series Title	
Chapter Title	Constructing a Quantitative Fusion Layer over the Semantic Level for Scalable Inference
Copyright Year	2018
Copyright HolderName	Springer International Publishing AG, part of Springer Nature

Author	Family Name	**Gezsi**
	Particle	
	Given Name	**Andras**
	Prefix	
	Suffix	
	Role	
	Division	Department of Measurement and Information Systems
	Organization	Budapest University of Technology and Economics
	Address	Budapest, Hungary
	Division	
	Organization	Abiomics Europe Ltd.
	Address	Budapest, Hungary
	Email	gezsi@mit.bme.hu
	URL	http://bioinfo.mit.bme.hu/
Corresponding Author	Family Name	**Bruncsics**
	Particle	
	Given Name	**Bence**
	Prefix	
	Suffix	
	Role	
	Division	Department of Measurement and Information Systems
	Organization	Budapest University of Technology and Economics
	Address	Budapest, Hungary
	Division	
	Organization	Abiomics Europe Ltd.
	Address	Budapest, Hungary
	Email	bruncsics@mit.bme.hu
Author	Family Name	**Guta**
	Particle	
	Given Name	**Gabor**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Abiomics Europe Ltd.

	Address	Budapest, Hungary
	Email	guta@mit.bme.hu
Author	Family Name	**Antal**
	Particle	
	Given Name	**Peter**
	Prefix	
	Suffix	
	Role	
	Division	Department of Measurement and Information Systems
	Organization	Budapest University of Technology and Economics
	Address	Budapest, Hungary
	Division	
	Organization	Abiomics Europe Ltd.
	Address	Budapest, Hungary
	Email	antal@mit.bme.hu

Abstract

We present a methodology and a corresponding system to bridge the gap between prioritization tools with fixed target and unrestricted semantic queries. We describe the advantages of an intermediate level of networks of similarities and relevances: (1) it is derived from raw, linked data (2) it ensures efficient inference over partial, inconsistent and noisy cross-domain, cross-species linked open data, (3) preserved transparency and decomposability of the inference allows semantic filters and preferences to control and focus of the inference, (4) high-dimensional, weakly significant evidences, such as overall summary statistics could also be used in the inference, (5) quantitative and rank based inference primitives can be defined, and (6) queries are unrestricted, e.g. prioritized variables, and (7) it allows wider access for non-technical experts. We provide a step-by-step guide for the methodology using a macular degeneration model, including drug, target and disease domains. The system and the model presented in the paper are available at bioinformatics.mit.bme.hu/QSF.

Keywords
(separated by '-')

Semantic web - Graph databases - Linked open data - Data and knowledge fusion - Recommender systems - Explanation generation

Constructing a Quantitative Fusion Layer over the Semantic Level for Scalable Inference

Andras Gezsi[1,2], Bence Bruncsics[1,2](✉), Gabor Guta[2], and Peter Antal[1,2]

[1] Department of Measurement and Information Systems,
Budapest University of Technology and Economics, Budapest, Hungary
{gezsi,bruncsics,antal}@mit.bme.hu
[2] Abiomics Europe Ltd., Budapest, Hungary
guta@mit.bme.hu
http://bioinfo.mit.bme.hu/

Abstract. We present a methodology and a corresponding system to bridge the gap between prioritization tools with fixed target and unrestricted semantic queries. We describe the advantages of an intermediate level of networks of similarities and relevances: (1) it is derived from raw, linked data (2) it ensures efficient inference over partial, inconsistent and noisy cross-domain, cross-species linked open data, (3) preserved transparency and decomposability of the inference allows semantic filters and preferences to control and focus of the inference, (4) high-dimensional, weakly significant evidences, such as overall summary statistics could also be used in the inference, (5) quantitative and rank based inference primitives can be defined, and (6) queries are unrestricted, e.g. prioritized variables, and (7) it allows wider access for non-technical experts. We provide a step-by-step guide for the methodology using a macular degeneration model, including drug, target and disease domains. The system and the model presented in the paper are available at bioinformatics.mit.bme.hu/QSF.

Keywords: Semantic web · Graph databases · Linked open data
Data and knowledge fusion · Recommender systems
Explanation generation

1 Introduction

Integration of cross-domain information has been targeted at different levels: at the level of data, such as in the joint statistical analysis of cross-domain omic datasets [1], at the level of knowledge, such as in the pharmaceutical integration approaches using semantic web technologies [2–4], and even at the level of computational services, such as in the scientific workflows [5,6]. However, significant part of scientific knowledge is uncertain, weakly significant, poorly represented and remains inaccessible for cross-domain integration, although the importance

I. Rojas and F. Ortuño (Eds.): IWBBIO 2018, LNBI 10813, pp. 1–13, 2018.
https://doi.org/10.1007/978-3-319-78723-7_4

of the analysis and interpretation of such weak signs have already been recognized in many standalone high-dimensional omic domains. This is illustrated by data fusion in molecular similarity [7], kernel-based data and knowledge fusion [8], cross-species gene prioritization [9], Bayesian fusion [10] and network boosted analysis of genome-wide polymorphism data [11].

Semantic technologies, relying heavily on the Resource Description Framework (RDF), provide an unprecedented basis for cross-domain data and knowledge fusion, as demonstrated by the emergence of large-scale, unified knowledge space in life sciences (the Life Sciences Linked Open Data Space, LSLODS, see e.g. BIO2RDF [12], CHEM2BIO2RDF [13], Open PHACTS [3], integrated WikiPathways [14], biochem4j [15], DisGeNET-RDF [16,17]). However, there are serious limitations concerning its computational complexity of inference [18] and practical IT accessibility [19], its inaccessibility for non-technical users [3,20,21]. Furthermore, most importantly, its ability to cope with uncertain facts, evidences, and inference is still an open challenge (for representing uncertain scientific knowledge, see e.g. HELO [22]; for combination of uncertain evidences, see e.g. [10,23–25]).

To tackle these challenges, we propose the construction of an intermediate, quantitative knowledge level of structured similarities and we created a corresponding system to demonstrate its advantages, the Quantitative Semantic Fusion (QSF) system (Fig. 1). This approach is related to multiple earlier approaches in fusion, such as (1) Linked Open Data (LOD) cubes to support computationally efficient SPARQL queries [26], (2) knowledge graphs [27], (3) probabilistic logic, Markov logic for semantic web integration inference and approximation of inference in large-scale probabilistic graphical models [28], and (4) relational generalization of kernel-based fusion [8,29].

We demonstrate the properties of this approach and the corresponding QSF system using a specific model for macular degeneration.

2 The Quantitative Semantic Fusion Framework

The Quantitative Semantic Fusion (QSF) System is an extensible framework that incorporates distinct annotated semantic types (also called: entities) and links between them by integrating different data sources from the Linked Open Data world. The QSF System then enables the users to quantitatively prioritize a freely chosen entity based on evidences propagated from any other, possibly multiple entities through the connecting links.

Currently, the system contains genes, taxa, diseases, phenotypes, disease categories (UMLS semantic types and MeSH disease classes), pathways, substances, assays, cell lines and the targets of the compounds. Besides, associations between genes and diseases are further described by related single nucleotide polymorphisms and the source of the association information. Links define associations between entities. For example, genes and pathways are connected with a link which represents gene-pathway associations. Certain links have additional annotations which can be used for (1) weighting associations during similarity computations and/or for (2) filtering links based on the annotation values. In order to enable cross-species information fusion, we also added gene ortholog links.

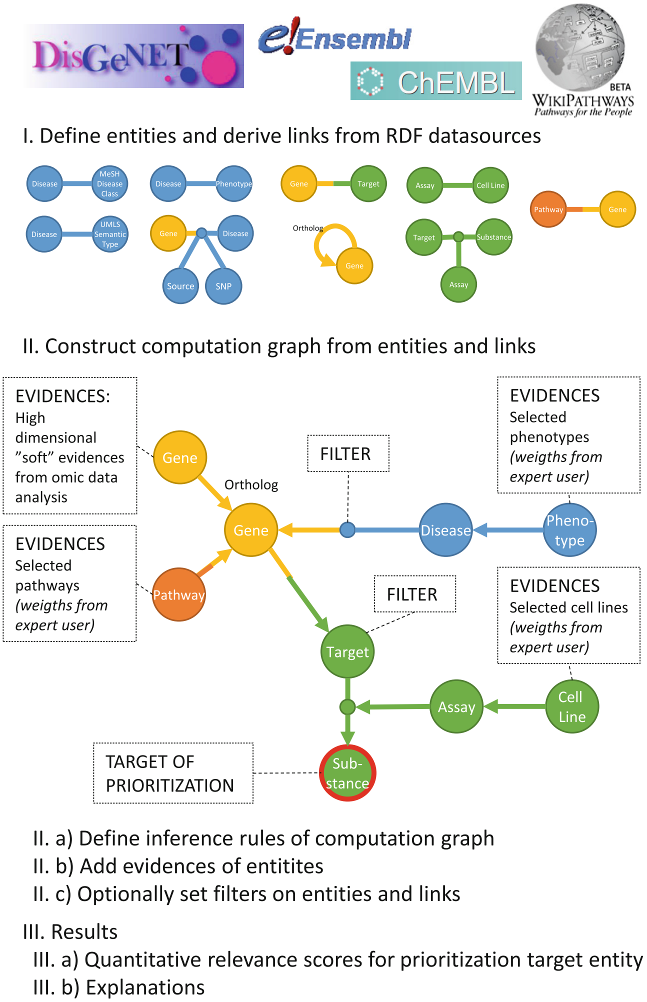

DisGeNET
e!Ensembl
ChEMBL
BETA
WikiPathways
Pathways for the People
I. Define entities and derive links from RDF datasources
Disease
MeSH Disease Class
Disease
Phenotype
Gene
Target
Assay
Cell Line
Pathway
Gene
Disease
UMLS Semantic Type
Gene
Disease
Source
SNP
Ortholog
Gene
Target
Substance
Assay
II. Construct computation graph from entities and links
EVIDENCES: High dimensional ”soft” evidences from omic data analysis
Gene
Ortholog
FILTER
EVIDENCES Selected phenotypes (weigths from expert user)
Gene
Disease
Pheno-type
EVIDENCES Selected pathways (weigths from expert user)
Pathway
FILTER
EVIDENCES Selected cell lines (weigths from expert user)
Target
Assay
Cell Line
TARGET OF PRIORITIZATION
Sub-stance
II. a) Define inference rules of computation graph
II. b) Add evidences of entitites
II. c) Optionally set filters on entities and links
III. Results
III. a) Quantitative relevance scores for prioritization target entity
III. b) Explanations

Fig. 1. Quantitative Semantic Fusion (QSF) System (I) The QSF System incorporates distinct annotated semantic types (i.e. entities) and their quantitative pairwise relations (i.e. links) by integrating different data sources from the Linked Open Data world. Predefined entities and links from DisGeNET [16], Ensembl [30], ChEMBL [31] and WikiPathways [14] are shown in the top. Together entities and links form the *structure* and *parameters* of the QSF System. (II) The user can freely construct so-called computation graphs using the available entities and links and can select any entity as the target of the prioritization. An example computation graph is shown in the middle. Then, the user defines the (II.a) *inference rules*, sets (II.b) *evidences* of possibly multiple entities and (II.c) optionally sets filters on specific entities and links. The main results of the prioritization are (III.a) the quantitative relevance scores for the target entity and (III.b) the most dominant *explanations* of the prioritization results.

Furthermore, to be able to expand the evidences related to gene or substance entities we enriched the system by adding gene-gene similarities based on Gene Ontology semantic similarity using GOssTo [32] and substance-substance similarities based on MACCS fingerprints computed by Tanimoto similarity.

The user can freely construct so-called *computation graphs* from the available entities (used as nodes) and links (as edges). Entities can be reused in the graph, i.e. multiple nodes in the graph can have the same semantic type. The user then arbitrarily selects an entity to prioritize and gives evidences in other entities.

For example, given a selection of phenotypes related to a disease, together with relevant drugs and substances in relevant clinical trials, and certain related genes in model organisms, the user may want to prioritize human genes based on all these evidences. The evidences propagate through the edges as similarity calculations between the entity vector of the source node and the row entity vectors of the linker matrix between the source and the target node. Seven different similarity calculation methods were implemented which are Cosine similarity, Dice and Overlap coefficients, Tanimoto similarity, and three kernel based similarities: linear, polynomial and radial basis function kernels. The vectors can be weighted by numeric annotations of the links, and information retrieval based corrections can also be used. Default similarity calculations are suggested for each link type based on internal tests and cross-validation, but the default calculation mode can be overridden by the user. In case of a node that has more than one incoming edges in the computation graph, the calculation of the scores of the node can be given with a mathematical formula over the incoming edges.

3 A Simple Model for Age-Related Macular Degeneration

To illustrate the methodology and the QSF system, a simple model was set up using age-related macular degeneration (AMD) as an example. In this case, the disease database contains 21 AMD subtypes or related diseases, but only one of them (URI: http://linkedlifedata.com/resource/umls/id/C0242383) contains relevant genetic information (including 391 genes). To expand the genetic information, the GWAS catalog [33] with 131 hits for AMD were used for human genetic source, and 72 rat genes from RGD (Rat Genome Database) and 42 mouse genes from MGD (Mouse Genome Database) [34] were used for ortholog genes, representing the two most common animal modes for AMD. For chemical information, a currently used AMD drug and over 30 drug candidates from clinical AMD trials were used, taken from DrugBank [35]. For pathways three complement and angiogenesis-related pathways were identified in the literature as underlying mechanisms.

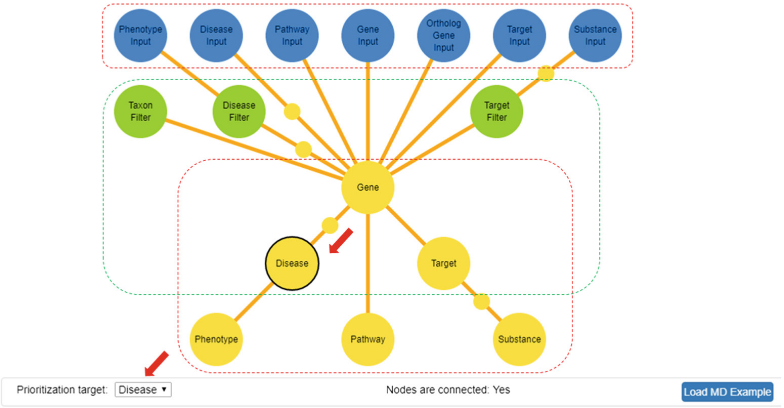
4 Phases of the Methodology

The main phases of the methodology starting from deriving relations from RDF resources to visualization of the most relevant proofs for an inference are as follows:

1. Resource and model overview: Overview the modeled phenomena and the available relevant resources and their connections.

2. Model structure: Design entities and their relations.
3. Model parameters: Derive parameters for the planned relations using SPARQL queries or direct RDF conversion.
4. Inference rules: Specify the inference rules for the propagation and combination of evidences, especially in multiply connected structures (with loops).
5. Evidences: Construct hard (logical) and soft (weighted) evidences.
6. Dynamic knowledge base: Define the active parts of the knowledge base, e.g. by selecting relevant model organisms and resources, and analogously disable certain parts of the knowledge base by semantic filtering.
7. Inference: Perform off-line inference using a computational cluster.
8. Results: Export prioritization and scoring results for targets: e.g. for external enrichment analysis.
9. Explanations: Export the most relevant explanations visualized as graphs.
10. Sensitivity analysis: Check the sensitivity of the results for settings.

The graphical user interface (GUI) of the QSF framework can be used for answering a large number of various questions using the predefined computation graphs. Furthermore, the development of the system allows further integration of new databases and the computation graph of the evidence propagation is easily customizable. To support non-technical users, the GUI contains prepared computation graphs, which are capable of handling typical questions and demonstrating functionalities. The presented computation graph is a simple tree-based fusion model over genes, diseases, phenotypes, pathways, targets (proteins) and substances. We use this model to demonstrate and explain the QSF phases (Fig. 2).



Prioritization target: Disease ▾

Nodes are connected: Yes

Fig. 2. A simple computation graph for macular degeneration in the QSF system. Blue, green and yellow denote the inputs, the filters and the outputs, respectively. (Color figure online)

4.1 Resource and Model Overview

The first step is the overview of the relevant information sources for the modeled phenomena collecting information for the involved phenotypes, genes, drugs and pathways (resources for macular degeneration are presented in Sect. 3).

4.2 Model Structure, Parameters, and Inference Rules

The second step starts with the construction of the computation graph, which contains the input nodes and the paths with possible filtering nodes. For the AMD model, the inputs and the target determine the computation graph (Fig. 2), but the framework can handle arbitrary graphs for complex models.

Remark: The building blocks for the computation graph are the nodes and the edges connecting them (Fig. 1).

Example: If a set of genes and substances are the available inputs and the question is about the pathways involved, then by clicking on the Gene and Substance Input Nodes and by selecting the Pathway Node as 'Prioritization target' (see later in Subsect. 4.5) the relevant parts (i.e. paths connecting the corresponding nodes) of the predefined computation graph will be automatically selected and the Gene and Target Filter nodes will be automatically added to the final computation graph.

4.3 Adding Input Evidences

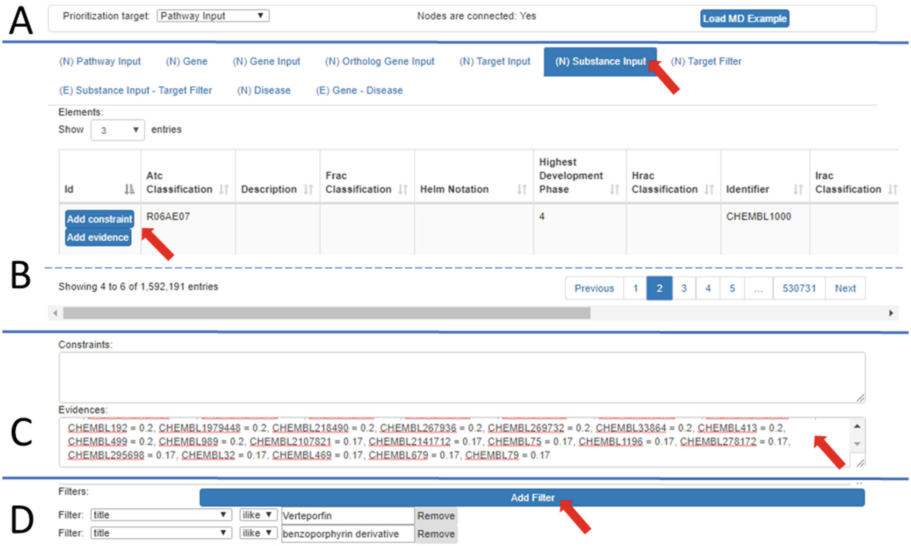
The framework can incorporate three types of inputs: (1) constraint information or list of entities without any weight, (2) evidence information or a list of entities with corresponding weights or evidences and (3) conditional input or filter parameters on a node choosing all the entities where the condition applies.

Remark: The QSF approximates Bayesian information propagation therefore for quantitative results the inputs are required to represent probabilities, although using any other kind of weights are allowed and will result in meaningful prioritization values, but the quantitative interpretation is more problematic.

Example: If the inputs are the drugs of running trials for a given disease, then the inputs can be added manually by clicking on 'Add constraint' or 'Add evidence', and the IDs will be shown in a list for each node (see Fig. 3B).

Converting IDs: The GUI allows to choose entities one by one for any node by name or ID, but for a larger number of values using lists and list of IDs is suggested. For genes Ensemble IDs, for diseases UMLS IDs, for phenotypes HPO IDs, for pathways WikiPathways IDs and for protein targets and substances ChEMBL IDs are used. Using these IDs, a large amount of input can be entered into the model, therefore converting data from diverse origin to the presented IDs is highly recommended, in order to utilize a maximum amount of data.

Defining soft evidences: Quantitative evidences are values of weights for each input entity representing relevance. It can be any numeric value, but optimally they are values between 0 and 1 representing the probability of the input.



Prioritization target: Pathway Input

Nodes are connected: Yes

Elements:

Show 3 entries

Id	Atc Classification	Description	Frac Classification	Helm Notation	Highest Development Phase	Hrac Classification	Identifier	Irac Classification
Add constraint Add evidence	R06AE07				4		CHEMBL1000	

Showing 4 to 6 of 1,592,191 entries

Constraints:

Evidences:

CHEMBL192 = 0.2, CHEMBL1979448 = 0.2, CHEMBL218490 = 0.2, CHEMBL267936 = 0.2, CHEMBL269732 = 0.2, CHEMBL33864 = 0.2, CHEMBL413 = 0.2, CHEMBL499 = 0.2, CHEMBL989 = 0.2, CHEMBL2107821 = 0.17, CHEMBL2141712 = 0.17, CHEMBL75 = 0.17, CHEMBL1196 = 0.17, CHEMBL278172 = 0.17, CHEMBL295698 = 0.17, CHEMBL32 = 0.17, CHEMBL469 = 0.17, CHEMBL679 = 0.17, CHEMBL79 = 0.17

Filters:

Filter: title ilike Verteporfin

Filter: title ilike benzoporphyrin derivative

Fig. 3. GUI interface for inputs and filters: (A) Choosing prioritization target (B) Choosing a node and providing manual constraints and evidences (C) Giving constraints and evidences using lists (D) Adding filters and specifying filter conditions

Using lists: For a larger number of inputs, usage of an input list is suggested. For example, if the drug trials for macular degeneration disease are considered as input, then by converting these drug names into ChEMBL, IDs can be added by separating them by commas. Quantitative (soft) evidences can also be used; the format is similar, but for each drug a certainty or relevance weight can be specified by an equality sign, where the number that follows is preferably a probability (Fig. 3C). In this case, the trial phase (0, I–IV) is known for the drugs and the probabilities could be approximated by the acceptance rate of ophthalmology trials, which are 0.17 for phases 0 and I, 0.2 for phase II, 0.45 for phase III [36].

Conditional inputs: Inputs can also be specified by using statements for any parameter of a given input node. Example statements are the following: for a disease node: the title contains the term “macular degeneration”; for a gene node: the chromosome number is 5; for a substance node: the title (or chemical name) contains a name (Fig. 3D) or a specific structure (like “Cyclopropyl-6-fluoro” and “carboxylic acid”).

4.4 Adding Filters

The semantic control over the inference, e.g. filtering out gene-diseases interactions purely based on keywords, is a novel function, which is completely missing from currently prevailing monolithic gene prioritization systems. Further improvements could be achieved by filtering out the less reliable links, e.g. the weak substance-target interactions, although the selection of threshold values for filtering is an open issue in our fusion methodology as well.

Remark: The filtering method is the same for input and filter nodes (Fig. 3D), except that if there is a filter statement in an input node (without parents), then it includes all the entities matching the statement and in case of any intermediate or filter node, it excludes all the entities from further propagation.

Example: In case of macular degeneration wide range of sources contain data about low vision in general, therefore filtering out common factors causing low vision like cataract can improve the quality of the inputs. Additionally, filtering on the Target-Substance edge also allows excluding chemicals with low affinity to the target by sorting out the weak associations where the pChembl (−log(IC50 or Ki)) is below a certain number.

4.5 Determining Outputs and Visualization

The next step is to define a target for the prioritization. It determines the type of the output and the path(s) of the propagation.

Remark: The 'Prioritization target' determines the path(s) of the information propagation in the graph; therefore it is an interpretation or an aspect of the model.

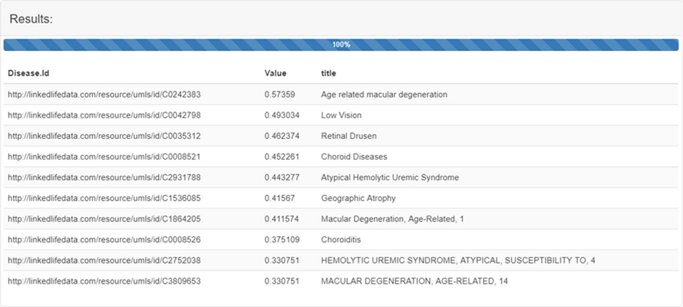
Example: For example, if the question is which diseases are involved in a biological setup, by clicking on the disease, that node is chosen for the prioritization target. It can be changed later by choosing the target from the list of the involved nodes (Figs. 2 and 3A).

The GUI supports interpretation using a simple tabular result prioritization and a graphical visualization.

Example: The macular degeneration model uses the known macular degeneration-related pathways, human and model animal genes, drugs and their known targets. Choosing a disease node as the prioritization target, the results (Fig. 4) and the contribution of the individual inputs (Fig. 5) are informative for evaluating the model.

Prioritization: The results contain entity identifiers, a numeric value representing the relevance of each entity and further descriptive parameters (Fig. 4).

Tabular view of prioritization: The matrix view plots parallel results in columns corresponding to all the inputs and for each individual input node. This technique supports the understanding of the contributions of the inputs and their redundancy, complementarity. The color scheme helps the visual tracking of the entities ranked differently by various inputs (Fig. 5).

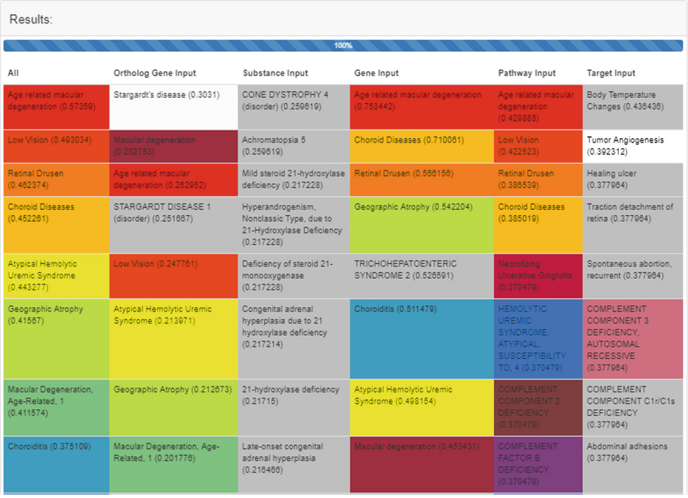


Results:

100%

Disease.Id	Value	title
http://linkedlifedata.com/resource/umls/id/C0242383	0.57359	Age related macular degeneration
http://linkedlifedata.com/resource/umls/id/C0042798	0.493034	Low Vision
http://linkedlifedata.com/resource/umls/id/C0035312	0.462374	Retinal Drusen
http://linkedlifedata.com/resource/umls/id/C0008521	0.452261	Choroid Diseases
http://linkedlifedata.com/resource/umls/id/C2931788	0.443277	Atypical Hemolytic Uremic Syndrome
http://linkedlifedata.com/resource/umls/id/C1536085	0.41567	Geographic Atrophy
http://linkedlifedata.com/resource/umls/id/C1864205	0.411574	Macular Degeneration, Age-Related, 1
http://linkedlifedata.com/resource/umls/id/C0008526	0.375109	Choroiditis
http://linkedlifedata.com/resource/umls/id/C2752038	0.330751	HEMOLYTIC UREMIC SYNDROME, ATYPICAL, SUSCEPTIBILITY TO, 4
http://linkedlifedata.com/resource/umls/id/C3809653	0.330751	MACULAR DEGENERATION, AGE-RELATED, 14

Fig. 4. A result of disease prioritization in the macular degeneration model.



Results:

100%

All	Ortholog Gene Input	Substance Input	Gene Input	Pathway Input	Target Input
Age related macular degeneration (0.57359)	Stargardt's disease (0.3031)	CONE DYSTROPHY 4 (disorder) (0.259619)	Age related macular degeneration (0.753442)	Age related macular degeneration (0.429885)	Body Temperature Changes (0.436436)
Low Vision (0.493034)	Macular degeneration (0.282753)	Achromatopsia 5 (0.259619)	Choroid Diseases (0.710061)	Low Vision (0.422523)	Tumor Angiogenesis (0.392312)
Retinal Drusen (0.462374)	Age related macular degeneration (0.262952)	Mild steroid 21-hydroxylase deficiency (0.217228)	Retinal Drusen (0.566156)	Retinal Drusen (0.386539)	Healing ulcer (0.377964)
Choroid Diseases (0.452261)	STARGARDT DISEASE 1 (disorder) (0.251667)	Hyperandrogenism, Nonclassic Type, due to 21-Hydroxylase Deficiency (0.217228)	Geographic Atrophy (0.542204)	Choroid Diseases (0.385019)	Traction detachment of retina (0.377964)
Atypical Hemolytic Uremic Syndrome (0.443277)	Low Vision (0.247761)	Deficiency of steroid 21-monooxygenase (0.217228)	TRICHOHEPATOENTERIC SYNDROME 2 (0.526591)	Necrotizing Ulcerative Gingivitis (0.370479)	Spontaneous abortion, recurrent (0.377964)
Geographic Atrophy (0.41567)	Atypical Hemolytic Uremic Syndrome (0.213971)	Congenital adrenal hyperplasia due to 21 hydroxylase deficiency (0.217214)	Choroiditis (0.511479)	HEMOLYTIC UREMIC SYNDROME, ATYPICAL, SUSCEPTIBILITY TO, 4 (0.370479)	COMPLEMENT COMPONENT 3 DEFICIENCY, AUTOSOMAL RECESSIVE (0.377964)
Macular Degeneration, Age-Related, 1 (0.411574)	Geographic Atrophy (0.212673)	21-hydroxylase deficiency (0.21715)	Atypical Hemolytic Uremic Syndrome (0.498154)	COMPLEMENT COMPONENT 2 DEFICIENCY (0.370479)	COMPLEMENT COMPONENT C1r/C1s DEFICIENCY (0.377964)
Choroiditis (0.375109)	Macular Degeneration, Age-Related, 1 (0.201776)	Late-onset congenital adrenal hyperplasia (0.216466)	Macular degeneration (0.453431)	COMPLEMENT FACTOR B DEFICIENCY (0.370479)	Abdominal adhesions (0.377964)

Fig. 5. The result of the disease prioritization using all macular degeneration-related inputs (leftmost column) and the contribution of the individual inputs (other columns).

Explanation visualization: To visualize the most relevant paths (i.e. the explanations) between the input nodes and the target node an explanation graph is exported into Cytoscape. The graph can be processed further using the add-ons and resources developed by the broad community of Cytoscape (Fig. 6).

4.6 Checking Robustness of the Results

Currently, we are implementing methods to support the comparison of results under different settings, e.g. using various inference rules, evidence weighting or semantic filtering. For example, our preliminary evaluation for the disease axis using the AMD model suggests the use of Tanimoto similarity for narrow queries and cosine similarity for broader queries with heterogeneous, soft evidences, e.g. for data analytic evidences.

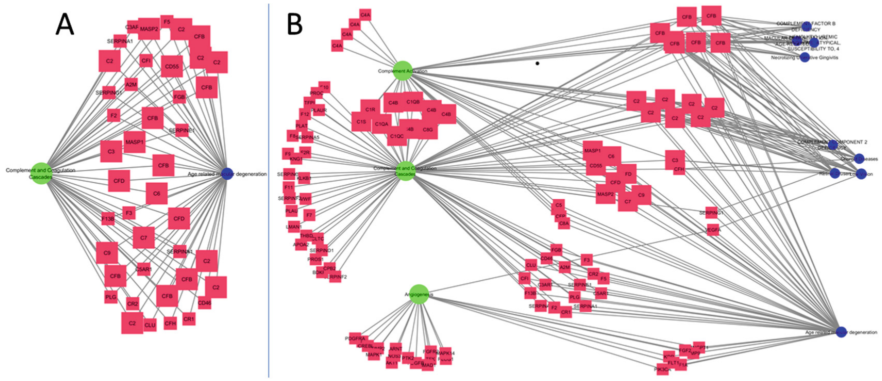


Fig. 6. Explanation graphs: (A) The association between a disease and a pathway can be determined by numerous genes (B) Graph representation of the most relevant explanation between pathways (green) and diseases (blue) trough genes (magenta) (Color figure online)

5 Conclusion

The availability of voluminous and heterogeneous semantically linked open data and knowledge provides an unprecedented opportunity for cross-domain fusion. However, uncertainty over the measurements and knowledge fragments, and also over the evidences poses a fundamental challenge for the practical use of these resources in research and development. We proposed an intermediate level of data and knowledge to cope with high-dimensional uncertainty, at which level quantitative relevances can be propagated through similarities and the inference process can also be semantically controlled and focused. Currently, we are evaluating the quantitative performance of the QSF system in prioritization tasks.

Acknowledgments. The research has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.2-16-2017-00013) and by OTKA 112915. This work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No 633589. This publication reflects only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

References

1. Zhu, Z., et al.: Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. **48**(5), 481–487 (2016)
2. Chen, H., Ding, L., Wu, Z., Yu, T., Dhanapalan, L., Chen, J.Y.: Semantic web for integrated network analysis in biomedicine. Briefings Bioinform. **10**(2), 177–192 (2009)

3. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open PHACTS: semantic interoperability for drug discovery. Drug Discov. Today **17**(21–22), 1188–1198 (2012)
4. Chen, B., Wang, H., Ding, Y., Wild, D.: Semantic breakthrough in drug discovery. Synth. Lect. Semant. Web **4**(2), 1–142 (2014)
5. Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., Brass, A.: TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics **16**(2), 184–186 (2000)
6. Karim, M.R., Michel, A., Zappa, A., Baranov, P., Sahay, R., Rebholz-Schuhmann, D.: Improving data workflow systems with cloud services and use of open data for bioinformatics research. Briefings Bioinform. (2017). bbx039 AQ3
7. Ginn, C.M., Willett, P., Bradshaw, J.: Combination of molecular similarity measures using data fusion. Perspect. Drug Discov. Des. **20**, 1–16 (2000). Virtual Screening: An Alternative or Complement to High Throughput Screening? Springer AQ4
8. Lanckriet, G.R., De Bie, T., Cristianini, N., Jordan, M.I., Noble, W.S.: A statistical framework for genomic data fusion. Bioinformatics **20**(16), 2626–2635 (2004)
9. Tranchevent, L.C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., Moreau, Y.: Candidate gene prioritization with endeavour. Nucleic Acids Res. **44**(W1), W117–W121 (2016)
10. Province, M.A., Borecki, I.B.: Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans. Pac. Symp. Biocomput. **13**, 190–200 (2008)
11. Nakka, P., Raphael, B.J., Ramachandran, S.: Gene and network analysis of common variants reveals novel associations in multiple complex diseases. Genetics **204**(2), 783–798 (2016)
12. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 200–212. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_14
13. Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinform. **11**(1), 255 (2010)
14. Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willighagen, E.L., Evelo, C.T., Pico, A.R.: Using the semantic web for rapid integration of wikipathways with other biological online data resources. PLoS Comput. Biol. **12**(6), e1004989 (2016)
15. Swainston, N., Batista-Navarro, R., Carbonell, P., Dobson, P.D., Dunstan, M., Jervis, A.J., Vinaixa, M., Williams, A.R., Ananiadou, S., Faulon, J.L., et al.: biochem4j: Integrated and extensible biochemical knowledge through graph databases. PLoS ONE **12**(7), e0179130 (2017)
16. Queralt-Rosinach, N., Piñero, J., Bravo, À., Sanz, F., Furlong, L.I.: DisGeNET-RDF: harnessing the innovative power of the semantic web to explore the genetic basis of diseases. Bioinformatics **32**(14), 2236–2238 (2016)
17. Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., Furlong, L.I.: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. **45**(D1), D833–D839 (2017)

18. Gray, A.J., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C., Burger, K., Chichester, C., Evelo, C.T., Goble, C., Harland, L., et al.: Applying linked data approaches to pharmacology: architectural decisions and implementation. Semant. Web **5**(2), 101–113 (2014)
19. Beek, W., Rietveld, L., Schlobach, S., van Harmelen, F.: LOD Laundromat: why the semantic web needs centralization (even if we don't like it). IEEE Internet Comput. **20**(2), 78–81 (2016)
20. Dong, X., Ding, Y., Wang, H., Chen, B., Wild, D.: Chem2Bio2RDF dashboard: ranking semantic associations in systems chemical biology space. Future Web Collaboratice Sci. (FWCS) WWW (2010)
21. Kamdar, M.R., Musen, M.A.: PhLeGrA: graph analytics in pharmacology over the web of life sciences linked open data. In: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 321–329 (2017)
22. Soldatova, L.N., Rzhetsky, A., De Grave, K., King, R.D.: Representation of probabilistic scientific knowledge. J. Biomed. Semant. **4**(Suppl. 1), S7 (2013)
23. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. **7**(1), 496 (2011)
24. Callahan, A., Cifuentes, J.J., Dumontier, M.: An evidence-based approach to identify aging-related genes in caenorhabditis elegans. BMC Bioinform. **16**(1), 40 (2015)
25. Fu, G., Ding, Y., Seal, A., Chen, B., Sun, Y., Bolton, E.: Predicting drug target interactions using meta-path-based semantic network analysis. BMC Bioinform. **17**(1), 160 (2016)
26. Abelló, A., et al.: Fusion cubes: towards self-service business intelligence (2013) AQ5
27. Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. Semantic web **8**(3), 489–508 (2017)
28. Domingos, P., Lowd, D., Kok, S., Poon, H., Richardson, M., Singla, P.: Just add weights: Markov logic for the semantic web. In: da Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW 2005-2007. LNCS (LNAI), vol. 5327, pp. 1–25. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89765-1_1
29. De Bie, T., Tranchevent, L.C., Van Oeffelen, L.M., Moreau, Y.: Kernel-based data fusion for gene prioritization. Bioinformatics **23**(13), i125–i132 (2007)
30. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al.: Ensembl 2016. Nucleic Acids Res. **44**(D1), D710–D716 (2015)
31. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics **30**(9), 1338–1339 (2014)
32. Caniza, H., Romero, A.E., Heron, S., Yang, H., Devoto, A., Frasca, M., Mesiti, M., Valentini, G., Paccanaro, A.: GOssTO: a stand-alone application and a web tool for calculating semantic similarities on the gene ontology. Bioinformatics **30**(15), 2235–2236 (2014)
33. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al.: The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res. **45**(D1), D896–D901 (2017)

34. Twigger, S., Lu, J., Shimoyama, M., Chen, D., Pasko, D., Long, H., Ginster, J., Chen, C.F., Nigam, R., Kwitek, A., et al.: Rat genome database (RGD): mapping disease onto the genome. Nucleic Acids Res. **30**(1), 125–128 (2002)

35. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al.: DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. **42**(D1), D1091–D1097 (2013)

36. Thomas, D.W., Burns, J., Audette, J., Carrol, A., Dow-Hygelund, C., Hay, M.: Clinical Development Success Rates 2006–2015. Biomedtracker/BIO/Amplion, San Diego, Washington, DC, Bend (2016)

Author Queries

Chapter 4

Query Refs.	**Details Required**	**Author's response**
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city name in affiliation "Aff2". Please check and confirm if the inserted city name is correct. If not, please provide us with the correct city name.	
AQ3	Kindly provide volume and page range for Refs. [6, 20].	
AQ4	Kindly check and confirm if the updated Ref. [7] is correct.	
AQ5	Kindly provide complete details for Ref. [26].	