LinkedPPI: Enabling Intuitive, Integrative Protein-Protein Interaction Discovery

Laleh Kazemzadeh[1], Maulik R. Kamdar[1], Oya D. Beyan[1], Stefan Decker[1], and Frank Barry[2]

[1] Insight Center for Data Analytics, National University of Ireland, Galway
{laleh.kazemzadeh,maulik.kamdar,oya.beyan,stefan.decker}@deri.org
[2] Regenerative Medicine Institute, National University of Ireland, Galway
frank.barry@nuigalway.ie

Abstract. Understanding the dynamics of protein-protein interactions (PPIs) is a cardinal step for studying human diseases at the molecular level. Advances in "sequencing" technologies have resulted in a deluge of biological data related to gene and protein expression, yet our knowledge of PPI networks is far from complete. The lack of an integrated vocabulary makes querying this data difficult for domain users, whereas the large volume makes it difficult for intuitive exploration. In this paper we employ Linked Data technologies to develop a framework 'LinkedPPI' to facilitate domain researchers in integrative PPI discovery. We demonstrate the semantic integration of various data sources pertaining to biological interactions, expression and functions using a domain-specific model. We deploy a platform which enables search and aggregative visualization in real-time. We finally showcase three user scenarios to depict how our approach can help identify potential interactions between proteins, domains and genomic segments.

Keywords: Protein-Protein Interaction Network, Linked Data, Domain-specific Model, Visualisation

1 Introduction

1.1 Background

The study of biological networks forms the integral core of biomedical research related to human diseases and drug development. The ultimate goal of such studies is to understand the connections between different genes and proteins, how the cell signals propagate across these networks and regulate their functionality. Hence understanding the Protein-Protein Interaction (PPI) networks underlying each such cellular mechanism is important to specifically target the dysfunctional proteins, leading towards the discovery of potential drugs and treatments for diseases. Studying PPI networks helps understand the interconnectedness between different cellular mechanisms and pathways. Biological pathways are not independent of each other, but their interactions are harmonious, which makes them part of a bigger network. Thus it is important to investigate the dynamics of the cell system as a whole.

Human genome contains more than 20000 protein-coding genes which interact tightly in order to regulate various cellular pathways and mechanisms. The major challenge in developing a thorough understanding of these cellular mechanisms and pathways is to complete the PPI network for each mechanism. However, experimental validation of the binary interactions between total number of proteins is an inconceivable task thus computational models can be used to aid the researchers. These models help identify the sequential, structural and physicochemical properties of known interacting protein pairs and highlight the underlying patterns. Researchers then apply these patterns to narrow down the potential interacting partners for any protein(s) under investigation. Therefore wet-lab validation of the hypothesis formed around the predicted links and protein partners is realistic and achievable.

In computational models, experimentally validated PPIs form the backbone of PPI networks, however data pertaining to gene-expression, domain-domain interactions and genomic locations have proved their valuable contribution in inference and prediction of new links between protein pairs [6,19]. Each of these data sources has been published to address specific, albeit very different research problems. Therefore the data representation, data model and formats may vary from one data source to the other. Challenges stemming from the heterogeneity of the data emphasis the need for a framework which can bridge these biologically different concepts in order to highlight and extract the ubiquitous patterns, inconspicuous in the *bigger picture.*

1.2 Motivation

Due to advances in sequencing technologies, enormous amount of experimental data has been generated and stored as independent databases. Databases such as BioGRID [28], HPRD [9], MINT [17] contain the experimentally validated binary interactions, while UniProt [30], Ensembl [8], Entrez-Gene [21] and Gene Ontology [2] offer sequence information, genome localisation and cellular functionality of individual genes and proteins. On the other hand, knowledge bases like Pfam [27] contain information regarding the functional and structural protein subunits (domains).

The main motivation of this work is to provide researchers with a framework which enables them to retrieve the answers to their research questions from these disparate data sources. A researcher interested in the list of protein domains in a specific protein can look up the UniProt website[3] which is extensively rich in protein information. Genomic locations of protein-coding genes are publicly available from several websites such as CellBase [5]. However questions like, *'List of all the proteins which contain the exact or partial set of protein domains?'* or *'What is the relation of a set of interacting proteins and the genomic location of their underlying genes?'* cannot be answered through these websites.

The challenges in the aggregation and exploration of the aforementioned massive biological data sources have sparked the interests of several domain researchers and led them towards the adoption of a new generation of integrative

[3] `http://www.uniprot.org`

technologies, based on Semantic Web Technologies and Linked Data concepts, thus giving birth to Integrative Bioinformatics [25,7].

2 Methods

The final goal of this research is the identification and extraction of potential PPI networks from various publicly available data sources. The core structure of the PPI network consists of proteins and their experimentally proven interactions. Fig. 1 depicts an overview of the LinkedPPI architecture. Following subsections will describe data selection, RDFization and integration methodologies used.

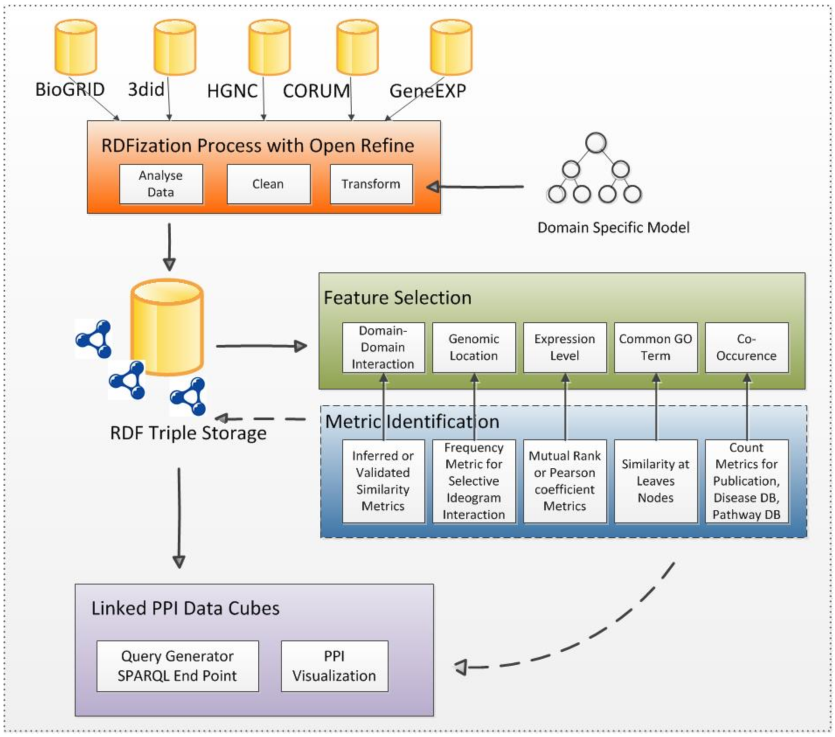

BioGRID
3did
HGNC
CORUM
GeneEXP
RDFization Process with Open Refine
Analyse Data
Clean
Transform
Domain Specific Model
RDF Triple Storage
Feature Selection
Domain-Domain Interaction
Genomic Location
Expression Level
Common GO Term
Co-Occurence
Metric Identification
Inferred or Validated Similarity Metrics
Frequency Metric for Selective Ideogram Interaction
Mutual Rank or Pearson coefficient Metrics
Similarity at Leaves Nodes
Count Metrics for Publication, Disease DB, Pathway DB
Linked PPI Data Cubes
Query Generator SPARQL End Point
PPI Visualization

Fig. 1. LinkedPPI Architecture

2.1 Selection of Relevant Data Sources

Validated Interactions: Experimentally validated interactions were retrieved from BioGRID (Biological General Repository for Interaction Datasets), one of the most comprehensive PPI databases [28]. Only physical interactions were included in our work, regardless of their classifications as raw or non-redundant.

Protein Complexes: In most cellular processes proteins act as a complex, instead of binary interactions between a pair of single proteins [1], during the same time and within the same cellular compartment. Such proteins are tightly interacting and play key roles in PPI networks. Elucidation of the dynamics of

PPI networks and functionality of individual proteins can benefit from identification of essential protein complexes, since different subunits contribute to drive a cellular function. In this work, we have used the latest release of CORUM (Comprehensive Resource of Mammalian protein complexes) [24].

Gene Expression: Understanding the correlation between gene-expression networks and protein interaction networks is an ongoing challenge in PPI studies. Proteins coded from co-expressed genes are more likely to interact with each other [10] and there is higher probability that an interacting pair of proteins share cellular functions [3]. We have used the COXPRESdb database [23] which publishes recent gene expression microarray datasets for Human.

Genomic Locations: Neighboring genes show similar expression pattern and are often involved in similar biological functions [22] which suggest that they might share same activation and translation mechanisms. These interactions may not be limited to the adjacent genes but can be long-range interactions to fulfil the cellular functionality [18]. Such evidences encouraged us to introduce a layer for the genomic locations of the protein-coding genes in our framework. We do not define 'genomic location' as the exact start/stop position of genes on a chromosome, but as the Ideogram band in which the genes reside. Ideograms are schematic representations which depict fixed staining patterns on a tightly coiled chromosome in Karyotype experiments. Karyotype describes number of chromosomes, their shape and length and banding patterns of chromosomes in the nucleus. Ideogram data was downloaded from the Mapping and Sequencing Tracks in the Human Genome Assembly (GRCh37/hg19, Feb 2009) at the UCSC Genome Browser[4]. The start/stop coordinates of the genes were retrieved from CellBase [5] and used to determine the genes within each ideogram. HGNC (HUGO Gene Nomenclature Committee) was used to map common genes referenced by different identifiers (Entrez-Gene, Ensembl and UniProt) [26].

Protein Domains: Proteins functionality and their structures are defined by their domain specification. Each protein consists of single or multiple domains, mutual sharing of which may lead to interaction with other proteins. However, identification of domain interactions through experimental validation for all possible protein pairs is an insurmountable task. Therefore domain knowledge bases can shed light on PPIs as well as help identify novel domain-domain interactions. We used 3did (Database of three-dimensional interacting domains) [29] which contains high resolution three-dimensional structurally interacting domains.

Gene Co-occurrence: Studying co-occurrence networks of genes can lead to the prediction of novel PPIs and discovery of hidden biological relations [13]. Previously Kamdar et al. generated co-occurrence scores as a weighted combination of the total number of diseases, pathways or publications in which any two genes occur simultaneously [14].

2.2 LinkedPPI Data Integration

One of the crucial challenges in integrative bioinformatics is the heterogeneous nature of biological data sources. Even though several attempts have been made

[4] `http://genome.ucsc.edu/`

in the standardization of the data through controlled vocabularies and guidelines, various hurdles still need to be surpassed. The proteomic standards initiative-molecular interaction (PSI-MI) is widely accepted by the community for the modelling of biological networks [12]. Even though some of the data sources are represented using the PSI-MI format there is no decipherable interconnectedness between them. To introduce the desired interconnectedness or 'bridges' between these data sources, we decided to use Linked Data technologies.

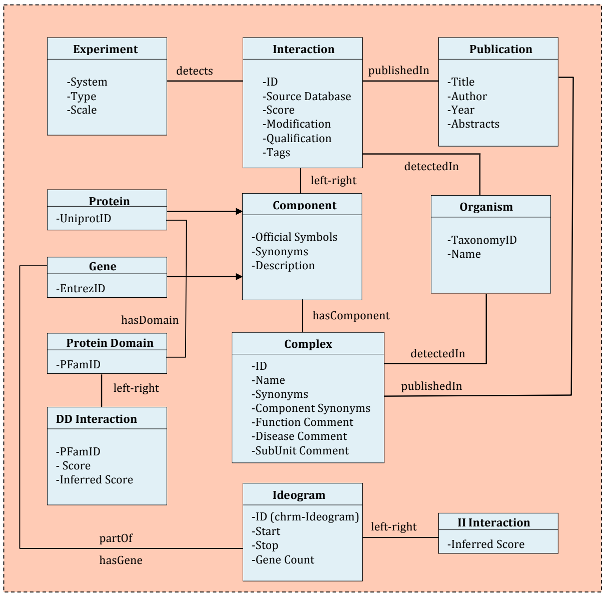


Fig. 2. Class Diagram of LinkedPPI Domain-specific Model

We proposed a simple concise domain model, for the modelling of the PPIs retrieved from BioGRID, complexes from CORUM, protein domains and genomic location. A domain-specific model is beneficial over an extensive well-construed ontology due to the absence of non-domain-specific concepts (`Thing`, `Continuant`, etc.) and is much smaller and self-contained to address a specific problem. Being native to a particular domain (e.g. *Protein-protein interactions*), it serves as an intermediate layer between the user and the underlying data, and enables intuitive knowledge exploration and discovery [15]. Our model comprises 12 concepts, which are termed relevant in this domain, and is shown in Fig. 2. The core concept in this model is a *Component*. A *Component* can either be a *Gene* or a *Protein*. A *Component* can be part of a *Complex* or can interact with another *Component* through an *Interaction*. The *Organism*, in which both the *Interaction* and the *Complex* are detected, is also available as a distinct concept. The *Experiment* concept embodies the attributes related to the experimental system which was used to detect the interaction (e.g. Y2H, AP-MS), the scale of the experiment (high or low-throughput), and whether it is a physical interaction or genetic. *Publication* documents the experiments, and links to resources described in the PubMed repository. A *Gene* is contained within an *Ideogram*, and *IIinteraction* represents inferred interactions between two ideograms from experimentally validated PPIs. *Domains* associated with protein-coding genes are represented using Pfam IDs and *DDinteraction* models interaction scores between two domains retrieved from 3did and inferred from BioGRID.

Open Refine provides a workbench to clean and transform data and eventually export it in required format. We used its RDF Extension [20] to model and convert the tab-delimited files downloaded from CORUM, BioGRID, 3did and CellBase to RDF graphs and stored them in a local Virtuoso Triple Store[5]. Data from COXPRESdb was already published on the web as RDF, and we re-used their data model and URIs. Similarly, for other data sources we re-used the URIs for the genes, proteins and publications, from those provided by Entrez-gene, UniProt and PubMed. To determine which gene is responsible for the encoding of which protein (mapping between Entrez-Gene ID and UniProt ID), we used the ID mapping table[6] provided by UniProt. One of the major advantages of using this approach was that the mapping also linked the relevant Gene Ontology (GO) [2] terms to the Entrez-Gene ID, thus providing additional information regarding the localisation and function of the specific genes.

3 Results

After RDFization, the BioGRID data source contains around 11 million triples (11357231), which establish 634996 number of distinct interactions between 14135 Human proteins. The data source also links out to 38952 unique PubMed publications documenting these PPIs. The CORUM data source consists of 156364 triples, with 2867 distinct complexes. The 3did data source consists of 320690 triples with 6818 distinct protein domains and 61582 validated and inferred domain-domain interactions. We inferred a total of 13493 interactions between 405 ideograms, referenced through 80092 triples. 60676 mappings were instantiated between the genes and 7750 extracted GO child leaves.

3.1 Search and Visualization

As such, relevant information could be retrieved from the SPARQL Endpoint through the formulation of appropriate queries. However as SPARQL requires a steep learning curve, the non-technical domain user needs intuitive, interactive visualization tools, which aggregate this information from the multiple data sources and summarize it. We devised a PPI Visualization Dashboard[7] based on ReVeaLD (Real-time Visual Explorer and Aggregator of Linked Data) [15] to accommodate our requirements for the search and visual exploration of the LinkedPPI networks. As the user starts typing the official symbol of the desired protein, a list of possible alternatives will be retrieved from the indexed entities. On selection, the entity URI is passed as a parameter through a set of

[5] `http://srvgal78.deri.ie/sparql`

[6] `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/README`

[7] `http://srvgal78.deri.ie/linkedppi`

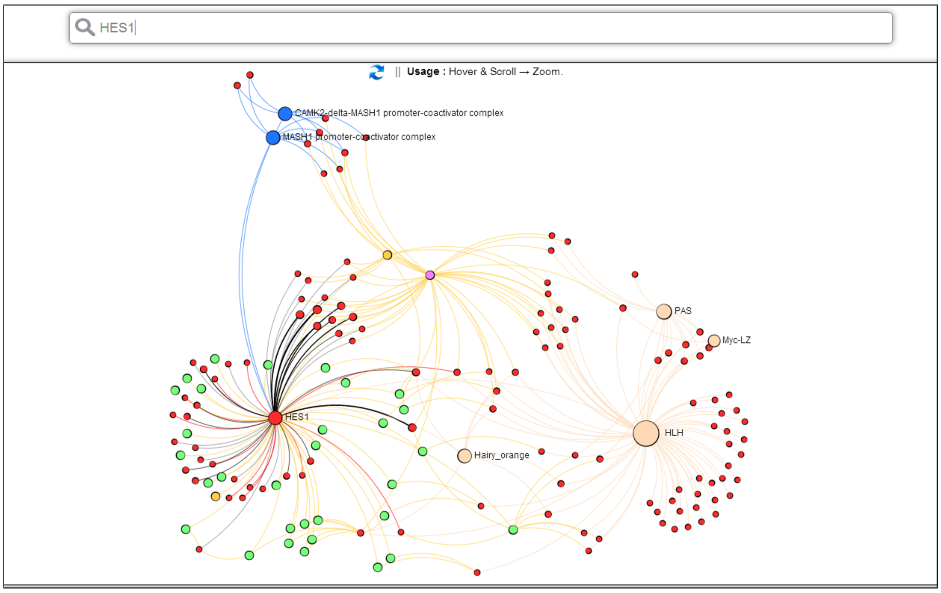
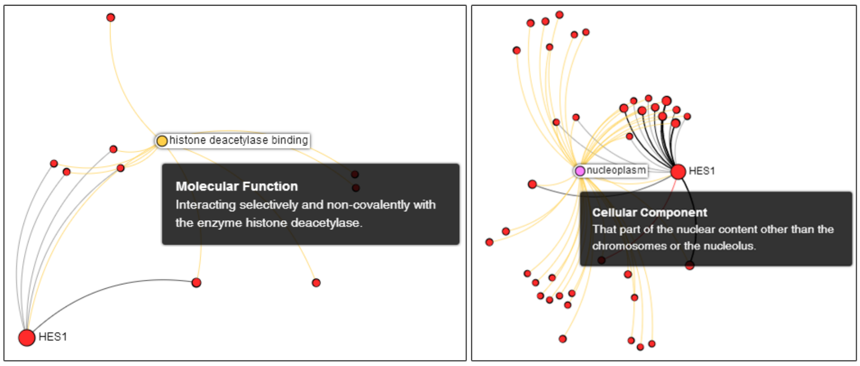

HES1
|| Usage : Hover & Scroll → Zoom.
CAMK2-delta-MASH1 promoter-coactivator complex
MASH1 promoter-coactivator complex
PAS
Myc-LZ
HES1
HLH
Hairy_orange

Fig. 3. Searching HES1 protein using the PPI Visualization Dashboard

pre-formulated SPARQL SELECT queries[8] targeting the various data sources. As shown in Fig. 3, the PPI network associated with the searched protein (e.g., HES1 *entrezgene:3280*) is rendered in a force-directed layout. The list of entities retrieved from the data sources are represented as circular nodes, with the size of each node directly proportional to the number of associated nodes. The nodes are rendered using different colors for the sake of visual differentiation - *Red* for *Components* (*Proteins* of BioGRID or *Genes* of COXPRESdb), *Blue* for CORUM *Complexes*, *Light Brown* for 3did *Protein Domains*. The three categories of GO Child Nodes - *Biological Processes*, *Molecular Functions* and *Cellular Components* are displayed using *Green*, *Yellow* and *Purple* colors.

The interactions between different proteins are represented as edges - the color of the edges is directly dependent on whether the associations have been retrieved from BioGRID, COXPRESdb or Co-occurrence Data (*Black*, *Red* and *Purple* respectively). The thickness of *Black* edges depends on the total number of publications which have experimentally validated the underlying interactions. The thickness of the *Red* and *Purple* edges depends on the PCC (Pearson Correlation Coefficient for Gene Expression) and Co-occurrence scores respectively. The *Protein* nodes, which are present in the same complex, possess interacting domains or have underlying coding genes associated to the same GO terms, are not connected directly to each other by edges. They are all connected using similar colored edges to the respective node (complex, domain or GO term), however there may be instances with experimental interactions or co-expression between the connected entities. The resulting network is hence densely clustered, rather

[8] `https://gist.github.com/maulikkamdar/a47fbecddecc6ba4b373`



Molecular Function
Interacting selectively and non-covalently with the enzyme histone deacetylase.

Cellular Component
That part of the nuclear content other than the chromosomes or the nucleolus.

Fig. 4. Subgraph of HES1 PPI network based on GO terms

than a simplistic radial layout of nodes. Hovering over any node highlights subgraph of the network which only displays the first-level connected nodes and their relations (Fig. 4), hence allowing any domain user to intuitively deduce answers to simple questions like, *'Which protein-encoding genes in the network share the same molecular function and have experimental co-expression?'* An information box is also displayed beside the hovered node to show additional information like GO term descriptions, Pfam or PubMed IDs, and PCC scores. Zooming and panning across the visualization is possible using a mouse.

3.2 Use Cases

The following subsections describe three different scenarios depicted in Fig. 5 that our framework could be employed to facilitate extraction of implicit information which can be used as predictors of novel protein-protein interactions. The relevant SPARQL Queries are documented at `http://goo.gl/xesMjR`.

Use Case 1: Extraction of Potential Protein-Protein Interactions Based on the Domain-Domain Interactions. Proteins carry on their functions through their protein domain(s), is a well-known fact. In this scenario we aim to extract possible PPIs based on the known domain-domain interactions. For the sake of simplicity in this use case we assume we are interested in proteins which contain single domains. A researcher has a *protein* in mind for which the sequence specification and domain composition are known. An interesting question might be, list of potential protein partners for this *protein*. Using our framework researcher can retrieve the list of protein pairs in which at least one of the proteins contains the same protein domain as the protein under question. Possible outcomes are: a) the protein under investigation itself shows up in the result set which forms the list of its experimentally validated protein partners. However this could be queried from the BioGRID web site directly. b) List of proteins with one single domain. In this case with a naive and straightforward conclusion, the researcher may accept the list. However in most cases further application of GO enrichment or advanced statistical analysis offer a more concise list but these analyses are beyond the scope of this work. c) The query results to a set of proteins consisting of several domains which requires further statistical or domain expert knowledge refinement. Despite the need of further investigation in such cases, the shortlisted hypothetical interaction partners are expected to be brief to save a tremendous amount of time and effort. The SPARQL Query uses the example of the HES1 (*entrezgene: 3280*) protein. We obtain the list of domains present in HES1 - Hairy_Orange (*pfam:PF07527*) and HLH(*pfam:PF00010*), and the list of proteins (e.g. HEY2) which share these domains, or have domain-domain interactions. We then retrieve validated PPIs in which the protein participates (e.g. HEY2-SIRT1). We can also obtain the PubMed publication documenting each PPI.

Use Case 2: Identification of Potential Domain-Domain Interactions. Protein-protein interactions can be identified experimentally through various types of experiments (e.g: Yeast Two-Hybrid). However it is not possible to identify the interacting domains between two proteins from same experiments and it requires a set of different experiments and protocols. Often protein domains act as signature elements and repeatedly interact with each other within the same organism. Therefore these frequent observations assist in identification of novel domain-domain interactions which is enlightening in identification of latent PPIs. Nevertheless in this work such observations are inferred implicitly from the validated PPI dataset (BioGRID) and require further statistical significance analysis. In our SPARQL Query example, we retrieve the validated and the inferred scores for domain interactions with the HLH domain (*pfam:PF00010*).

Use Case 3: Identification of Selective Interactions between Segments of Human Genome. Human chromosomes are compact in 3D space with each chromosome folding into its own territory [18]. Even though the exact relation of spatial conformation of genes and their functionality is not fully understood yet, studies have shown that the structure of the human genome follows its functionality [16]. It is widely believed that chromosomal folding bring functional elements in close proximity regardless of their inter- or intra-chromosomal distance in base pair unit. In other words the concept of close and far in relation to the spatial map of genome is represented differently. Also, it has been shown that the contacts between small and gene-rich chromosomes are more frequent [18]. These evidences suggest linkages between chromosomal conformation, gene activity and their products (proteins) functionality. Identifying the significance of association of genomic location of genes and their products partner selection will aid in completion of the proximity pattern followed by genes and lead by their arrangement. The prospective pattern can be employed to the prediction models in order to infer potential protein interactions.

In this work we have selected the boundary of ideogram bands on each chromosome as genomic location of each gene. Several genes may reside on one ideogram as well as genes that may fall between two consecutive ideograms. The reason for this selection of distance unit is to take into account the effect of the co-expression of neighboring genes and the possible shared mechanism. Protein pairs from PPI dataset are mapped to their genomic location and simple frequency calculation from retrieved data can identify the significance of interactions between two genomic segments. Based on these findings researchers can propose the genomic location pattern in which proteins preferentially select their interacting partners. These regions may contain genes involved in the same pathways or share the same functionality which are yet to be identified by further gene enrichment analysis. As shown in the provided SPARQL Query, we could retrieve the pre-determined inferred score between two ideograms, e.g. *Chromosome 3 - q29* and *Chromosome 10 - q24.32* containing the protein-coding genes for HES1 and SIRT1 (*entrezgene:23411*) proteins respectively.

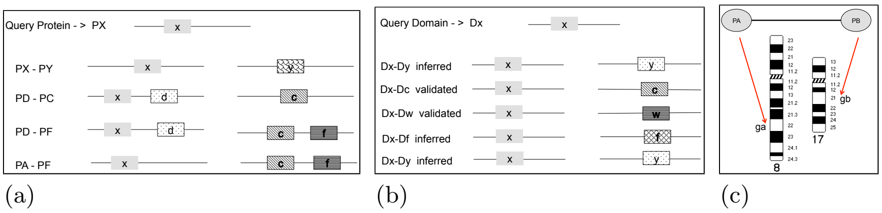


Fig. 5. Illustration of the three use cases. **a) Use Case 1:** *Px* is the protein under query which consist of domain *x*. **b) Use Case 2:** *Dx* is the domain of interest independent of the containing protein. The returning result is a list of binary interactions between domain pairs which labels the interaction either as *validated* or *inferred*, the former is retrieved from 3did database while the latter is deduced from PPI data. **c) Use Case 3:** The genomic location of *PA* and *PB*, two interacting proteins, are mapped to their ideogrammatic location on two different chromosomes.

4 Related Work

Jiang et al. developed a semantic web base framework which predicts targets of drug adverse effect based on the PPIs and gene functional classification algorithm [11]. Chem2Bio2RDF [4] integrates data sources from Bio2RDF[9] in order to study polypharmocology and multiple pathway inhibitors which also requires thorough understanding of underlying PPI network.

5 Conclusions

The incorporation of complementary datasets for the expansion of PPI networks is a useful approach to gain insight into biological processes and to discover novel PPIs which have not been documented in the current PPI databases. However, there is an inherent high level of heterogeneity at the schema and instance level of these data sources, due to lack of a common representation schema and format. Hence, we decided to apply Linked Data concepts in the integration, retrieval and visualisation of concealed information. The enormous amount of publicly available data and its dynamicity, in terms of regular updates, is currently a

[9] `http://bio2rdf.org`

rate-limiting step to our data-warehousing approach for centralised analysis. We have proposed a domain-specific model which can accommodate the needs in the field of PPI modelling. The use of a domain-specific model and an interactive graph-based exploration platform for search and aggregative visualisation makes our integration approach more intuitive for the actual users who deal with PPI predictions. We have also proposed a set of three user scenarios depicting how LinkedPPI framework could be used for the prediction of potential interactions between proteins, domains and genomic regions.

6 Future Work

The approach which has been presented in this work is used in extraction of valuable information with regard to PPI network, domain-domain interactions and selective genomic interactions. However the observations reported in the outcome of such data retrieval is raw and could be a valuable asset for simulations and prediction methods if further analysis is done. As part of the future work we intend to apply statistical analysis on significance of such observations in order to be able to develop a classifier algorithm which is able to predict interacting and non-interacting protein pairs.

Acknowledgements This work has been done under the Simulation Science program at the National University of Ireland, Galway. SimSci is funded by the Higher Education Authority under the program for Research in Third-level Institutions and co-funded under the European Regional Development fund.

References

1. Alberts, B.: The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. Cell 92(3), 291–294 (Feb 1998)
2. Ashburner, M., Ball, C.A., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25(1), 25–29 (May 2000)
3. Bhardwaj, N., Lu, H.: Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics 21(11), 2730–2738
4. Bin Chen, Xiao Dong, D.J.H.W.Q.Z.Y.D., Wild, D.J.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 11 (2010)
5. Bleda, M., Tarraga, J., de Maria, A., Salavert, F., et al.: CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic acids research 40(W1), W609–W614 (2012)
6. Chatterjee P, Basu S, K.M.N.M.P.D.: PPI$_S$VM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables
7. Chen, H., Yu, T., Chen, J.Y.: Semantic Web meets Integrative Biology: a survey. Briefings in Bioinformatics 14(1), 109–125 (Jan 2013)
8. Flicek, P., et al.: Ensembl 2012. Nucleic acids research p. gkr991 (2011)
9. Goel, R., Harsha, H.C., Pandey, A., Prasad, K.S.: Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. Molecular bioSystems 8(2), 453–463 (Feb 2012)

10. Grigoriev, A.: On the number of protein-protein interactions in the yeast proteome. Nucleic Acids Res 31, 4157–4161 (2003)
11. Guoqian Jiang, Chen Wang, Q.Z., Chute, C.G.: A Framework of Knowledge Integration and Discovery for Supporting Pharmacogenomics Target Predication of Adverse Drug Events: A Case Study of Drug-Induced Long QT Syndrome. AMIA Summits Transl Sci Proc p. 8892 (2013)
12. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., et al.: The HUPO PSI's molecular interaction formata community standard for the representation of protein interaction data. Nature biotechnology 22(2), 177–183 (2004)
13. Jelier, R., et al.: Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics 21(9), 2049–2058 (2005)
14. Kamdar, M.R., Iqbal, A., Saleem, M., Deus, H.F., Decker, S.: GenomeSnip: Fragmenting the Genomic Wheel to augment discovery in cancer research. In: Conference on Semantics in Healthcare and Life Sciences (CSHALS). ISCB (2014)
15. Kamdar, M.R., Zeginis, D., Hasnain, A., Decker, S., Deus, H.F.: ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of biomedical informatics 47, 112–130 (2014)
16. Kosak, S.T., Groudine, M.: Form follows function: the genomic organization of cellular differentiation. Genes Dev 18, 1371–1384 (2004)
17. Licata, L., Briganti, L., et al.: MINT, the molecular interaction database: 2012 update. Nucleic acids research 40(Database issue), D857–D861 (Jan 2012)
18. Lieberman-Aiden, E., van Berkum, N., Williams, L., Imakaev, M., et al.: Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326(5950), 289–293 (2009)
19. Liu, Z.P., et al.: Inferring a protein interaction map of mycobacterium tuberculosis based on sequences and interologs. BMC Bioinformatics 13(Suppl 7), S6 (2012)
20. Maali, F., Cyganiak, R., Peristeras, V.: Re-using cool uris: Entity reconciliation against lod hubs. In: Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M. (eds.) LDOW. CEUR Workshop Proceedings, vol. 813. CEUR-WS.org (2011)
21. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucleic acids research 39(Database issue), D52–7 (Jan 2011)
22. Michalak, P.: Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomic 91, 243248 (2007)
23. Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Motoike, I.N., Kinoshita, K.: COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. Nucleic Acids Research 41(D1), D1014–D1020 (Jan 2013)
24. Ruepp, A., et al.: CORUM: the comprehensive resource of mammalian protein complexes - 2009. Nucleic Acids Research 38(Database-Issue), 497–501 (2010)
25. Ruttenberg, A., Clark, T., Bug, W., et al.: Advancing translational research with the Semantic Web. BMC bioinformatics 8 Suppl 3(Suppl 3), S2+ (2007)
26. Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W., Bruford, E.A.: genenames.org: the HGNC resources in 2011. Nucl. Acids Res. 39(Suppl 1), D514–D519 (2011)
27. Sonnhammer, E.L., Eddy, S.R., et al.: Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28(3), 405–420 (Jul 1997)
28. Stark, C., Breitkreutz, B.J., Reguly, T., et al.: BioGRID: a general repository for interaction datasets. Nucleic acids research 34(suppl 1), D535–D539 (2006)
29. Stein, A., Russell, R.B., Aloy, P.: 3did: interacting protein domains of known three-dimensional structure. Nucleic acids research 33(suppl 1), D413–D417 (2005)
30. Uniprot-Consortium: The Universal Protein Resource (UniProt) 2009. Nucleic acids research 37(Database issue), D169–174 (Jan 2009)