Visualisation and target prioritisation using compuational and experimental omics datasets in human

Jonathan Lees[1], Jean-Karim Heriche[2], Juan-Ranea[3], James Perkins[1], Christine Orengo[1]
[1]University College London, [2] European Molecular Biology Laboratory, [3] University of Malaga

Correspondence: lees@biochem.ucl.ac.uk, +44 (0)20 7679 2000, Darwin Building, Gower Street, London WC1E 6BT

With the ever increasing number of large scale biological datasets (e.g. Y2H, microarray, Next-Gen, RNAi) it is becoming difficult for experimentalists to keep track of all publications associated with their proteins or pathways of interest. We have developed a set of tools that combine many high-throughput datasets with genome wide computational based predictions (e.g. text mining of literature) to allow experimentalists to view systems of interest integrated with experimental data to allow novel hypotheses to be developed. Where necessary orthology assignments are used to inherit relevant datasets from model organisms for human-centric visualization. As part of this we have developed powerful tools for target prioritization for predicting novel members of biological pathway / system . We implement this using graph kernels (Ref 1 for review) providing the ability to query the integrated high throughput datasets and retrieve related proteins visually in a similar manner to a search engine such as Google but where the search terms are gene names belonging to a pathway of interest (see Figure 1 for overview of the pipeline). We have computationally validated our ranking method and show it to be state of the art in comparison to the existing methods. Furthermore we have experimentally validated our method using the chromosome condensation biological process as our query pathway. We tested the top 100 genes predicted to be part of this pathway. Of these we 40 showed positive phenotypes for this pathway in a medium throughput screen. The predictor used to generate the results is publicly available as a separate website at http://FunL.org. We have combined the ranking algorithm as part of a wider data integration strategy to discover novel mediators of pain (http://PainNetworks.org). This site contains the data-types mentioned above along with adaptations of existing clustering algorithms (e.g. Ref 2) that leverage the kernel and expression data to allow for amongst other things tailoring of the clustering. The work has been carried out in close collaboration with experimental groups to ensure the tools are highly accessible and contains the most relevant information.

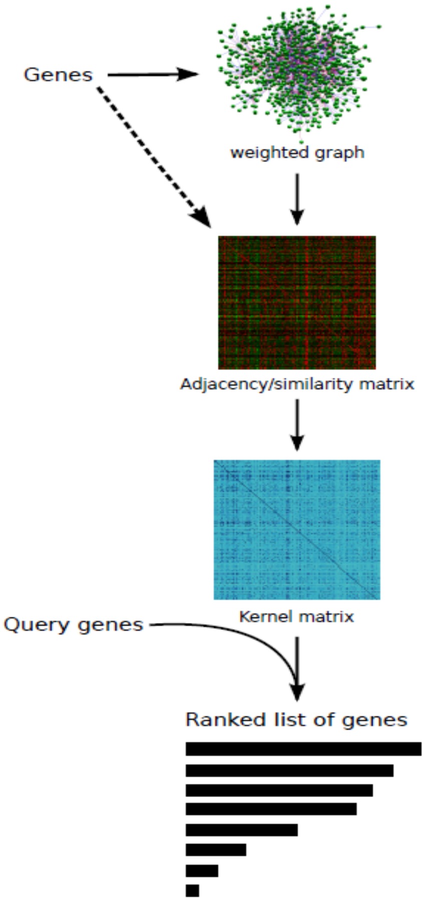

Figure 1: Schematic of ranking Pipeline. From Top to bottom: networks are converted to kernel matrices which are integrated by summation. The integrated kernel is then used to derive a ranking of the genome by the Query genes. Gene association networks integrated include experimental protein interactions, text mining associations, inherited interactions (HIPPO) and Gene expression similarity.

References

1 . Lees JG, Heriche JK, Morilla I, Ranea JA, Orengo CA. Systematic computational prediction of protein interaction networks. Phys Biol. 2011 Jun;8(3):035008.

2. Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature. 2010 Aug 5;466(7307):761-4.