Gerstein Lab experience in biological networks.

In general, biological networks provide an efficient way of integrating biological datasets of various types. Those networks can represent different types of nodes (genes, cells, micro-RNAs etc.) and can represent different types of interactions between those nodes by allowing different types of edges. Network reconstruction requires integrating both large-scale yet noisy data from high-throughput experiments and reliable yet incomplete data from small-scale experiments. Gerstein lab has extensive experience and has a very strong track record in network reconstruction. In fact, Gerstein Lab is among the first to use statistical models to perform systematic network reconstruction of protein-protein interaction networks [PMID: 14564010]. Since then, Gerstein Lab has developed many computational methods for reconstructing a variety of biological networks. For example, Gerstein Lab developed the JEME method for reconstructing the enhancer-target gene interaction network by integrating epigenetic data from hundreds of human samples [PMID: 28869592]. Gerstein Lab developed the iTAR tool for identifying target genes of TFs for reconstructing TF binding networks [PMID: 27519564]. Gerstein Lab has used high-order neural networks and kernel methods to reconstruct peptide-major histocompatibility complex interactions [PMID: 26206306]. Gerstein Lab also developed the ProbRNA method for reconstructing protein-RNA interaction networks using structure-probing sequencing data [PMID: 24376038]. These are just some examples of the many methods Gerstein lab has developed.

Using the in-lab developed methods, Gerstein Lab has successfully reconstructed different types of networks in humans and model organisms, including large-scale efforts aimed at producing reference networks. As some notable examples, Gerstein Lab has reconstructed TF binding and miRNA binding networks in human cell lines and in *C. elegans*, respectively, based on data from ENCODE and modENCODE [PMID: 22955619, 24376038]. More recently, Gerstein Lab reconstructed gene regulatory networks in human neurons for the studying psychiatric conditions [PMID: 30545857]. Gerstein Lab analyzed the network change as those changes could result in phenotypic variations for the organisms eventually. Gerstein Lab has developed the DiNeR method for identifying such changes and analyzing their consequences for example to downstream gene expression programs [PMID: 32615918]. On a larger scale, some network perturbations may propagate to cause major network rewiring. Gerstein Lab has developed the TopicNet method to measure such rewiring in transcriptional regulatory networks [PMID: 32657410]. Gerstein Lab also applied this idea to study network rewiring in cancer genomics [PMID: 32728046].

Gerstein Lab will develop the TicTalk framework to integrate plasma non-coding RNA, gene (expression level or protein level) and pathogen abundance (shown in Fig 1). TicTalk uses the latent Dirichlet allocation procedure to create topics that group plasma non-coding RNA, gene and pathogen abundance with similar co-occurrence patterns across samples. The resulting topic matrix (in Fig 1 with topics) is similar to the derivation of gene expression signatures in non-negative matrix factorization. Within each patient, the weights for each topic are specified by a vector (in Fig 1). For a given gene and pathogen, we can determine the level of correlation of their representation across the various topics to obtain a raw linkage score. This score can be further normalized by comparison to a background distribution of all possible scores and then individualized to a particular patient by considering only the relevant topics active in that sample (the final score is indicated by in Fig 1 and represents a statistical significance value from the distribution).

 

From this, we can link a particular pathogen to a human gene or non-coding RNA. These linkages are further refined and related to known pathways using network propagation (also shown in Fig 1) to obtain a final set of linkages. Finally, TicTalk will construct an interactome between non-coding RNA, gene and pathogen for different samples, which can help to analyze the interactions between the tick and various components, with the potential to improve our strategies for treating and preventing Lyme disease and complications. TicTalk has four key advantages for integrating multiple data types. First, it takes advantage of the Dirichlet distribution of hyperparameters to handle sparse and noisy data. Second, it enforces a unitary topic distribution for each sample, allowing for easy comparison across samples and facilitating linkage identification between different data types. Third, it can be easily extended to multiple data types. Fourth, it can infer specific individual linkages.