Workpackage 2

Link: http://dev.pubgene.com/complexdis/CM.html

Objectives

The goals of this workpackage are to extract and characterize gene networks associated to complex diseases using methods in text mining and database integration. The data sources employed to achieve these objectives are primarily from the literature (the 17+ million abstract in Medline) and experimentally validated interaction databases in the public domain.

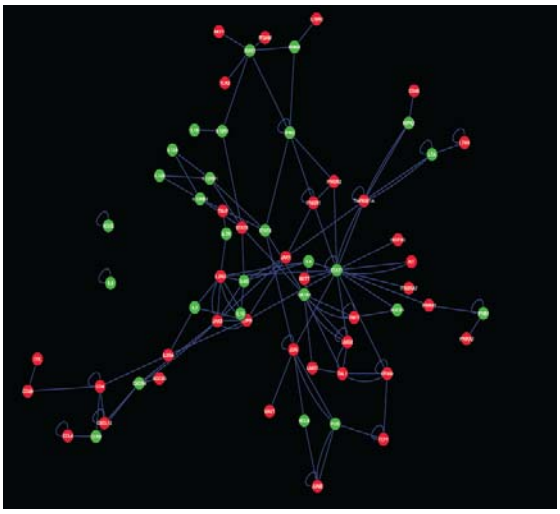
The network relationships of the entities mined from these sources are characterized by applying computational approaches in network theory that capture the semantic relationships that govern their network topology. Methods are to be applied and developed to extract knowledge from the literature that reflects the network organization of these cellular components and made available to our project partners in software and database solution developed and phased in two prototypes (D5 & D7). This software and database solution will be used to retrieve networks that will be used as templates to functionally annotate emergent modules identified using clique-based methods from our partners (D6) and agent based network models from our partners (D6) and also used in the context of analysis of high throughput data generated from our clinical partners (D8).

Progress towards objectives

Deliverable D6 (publication on Immunology gene networks)

Networks analytics and agent based modelling (in collaboration WP4 contractors) have been performed on the Immunology Gene networks from deliverable D5 & D7 and extended in collaboration with other network theory and modelling groups from the ComplexDis project and with biologist from the group to describe this network in a manuscript. The focus of this modular discovery is on T-cell differentiation and the Th1/Th2 balance. Deliverables D5 and D7 are used in a pipeline of data mining and filtering methods to retrieve sub-network modules in conjunction with protein-protein interaction databases (the consolidated set of experimentally validated protein interactions). The network module associated to the Th1/Th2 balance, derived from our methods is illustrated in Fig 2 below.

Fig 2: An example gene network module from our methods. Extracted using semantic based data mining approaches focused on the TH1/TH2 balance. This is a network module to be used in conjunction with WP4 in an agent based model. The modelling of this modular network will used in conjunction with the gene expression studies of allergic rhinitis in a publication



Deliverable D8: Publication on a study that validates NLP algorithms

Complex entangled networks of protein interactions are implicated in tumor immunity and the knowledge of such is latent in the vast biomedical literature. We have in this deliverable publication (in submission) extracted from the literature the knowledge attainable to the immune systems role in the signaling complexity for each gene in the human genome. This publication places, for the first time, a comprehensive immune analysis in the context of the complexity that exists in these signaling relationships. We in particular analyzed the immune systems role in the complex signaling interplays of the tumor microenvironment for highthrougput gene expression data.

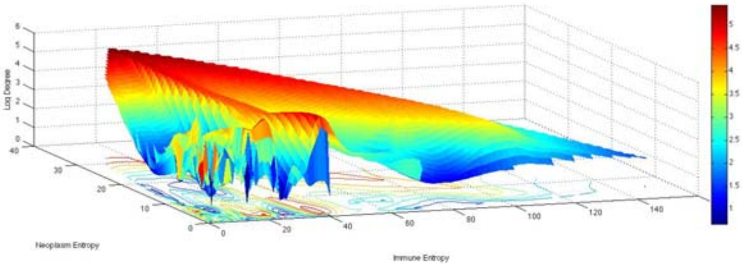


Figure 1: A three-dimensional heat surface plot representing the landscape of the interactome in the context of immune and cancer information using extracted information from the literature. Values above one on the log scale can be considered very high in terms of immune and cancer relevance. For degree centrality (C_{deg}) there are variable regions of scattered high and low connectivity in the interactome across the cancer immune landscape. The frequency of genes decreases with increasing information content in these plots. That which is apparent from this manner of representing the immune-interactome in the cancer context is the distinct areas of scattered high and then low connectivity value for genes in the cancer-immune landscape.

In this integrated analysis of previously manually curated and high-throughput efforts to catalogue immune genes we presented a strategy applying text-mining to glean the the entire immune information content of genes. This resulted in a genome-wide ranked immune relevance score that allowed the further characterization of tumor immunity against the interactome landscape (see Figure 1). In this information landscape, immune and cancer relevance correlations were described and also the relationship between immune relevance and the interactome.. We applied this consolidated immune information to a melanoma expression profile to find putative signatures of immunity interplay in a cancer's progression to metastasis.

Deliverables list

D6: Publication on the study of the topology robustness of literature networks and the analysis of such networks leading to the identification of putative critical hubs responsible for the topology of the disease gene network. This was produced on month 24 on its expected due date (in submission)

D8. Publication on a study that validates the text-mining and NLP algorithms to capture the modular organization of cellular components from the literature and the analysis of high-throughput biological data using the modular organization database. This has been delivered on month 36 on its expected due date and is in submission.

Milestones list and expected results

M4: Automatic systems for feature extraction of biological entities from the literature. We have reached a milestone (month 18) whereby all the elements of a software and database solution to achieve this objective are in place. For the benefit of our project partners this shall result in the classification of network hubs, extracted from the literature, and responsible for the topology of complex disease networks. A depiction of this milestone achievement is illustrated on Fig 4 where the sheer dominance of immunology genes implicated in complex disease from our database retrieval is classified in the context of the human protein interactome.

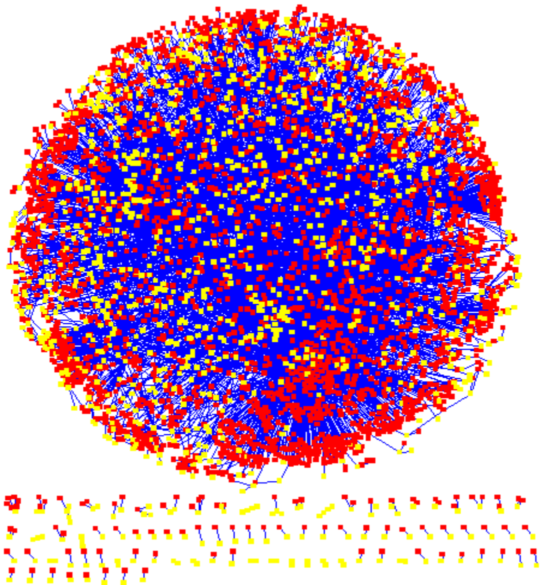


Fig 4: Complex disease genes (Yellow) with an immunology focus with their 1st degree human interacting partners (red), composing a giant sub-network covering >80% of the human protein interaction network database.

M5: We have completed considerable development that investigates the network robustness of gene networks derived from literature to identify critical hubs in experimental models of complex diseases in particular with respect to the immunological focus of complex disease (see Fig 5). The nature of this network topology and the modules that govern its behaviour will be investigated further in collaboration with our project partners, milestone that will be reached on month 24.

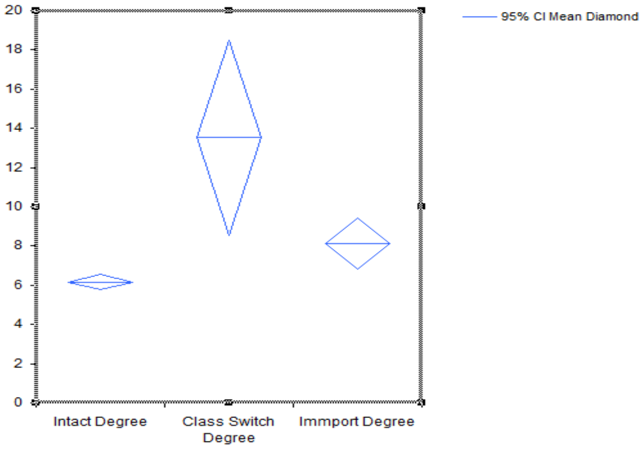


Fig 5: An illustration of how immunological genes possibly implicated in complex disease govern the robustness of the network module driving cellular processes.
On average immunology genes (Immport) are dominated by hubs more so than human protein ineractome (Intact). Furthermore, Ig Class switch modular genes have more hubs on overage than all classified immunology genes

M6: Our future plans for the next period is to concentrate on extraction and characterization of the modular organization of cellular functions in complex immunological diseases. The result will be to offer the capability of processing high-throughput biological data to form network modules by creating superimposition networks and their modular organization the. The First iteration reached month 6 (database), second iteration month 18 (database) and month 24 (publication in collaboration with WP4).