PROJECT SUMMARY

The past decade has seen a burst of large-scale sequencing projects offering novel and diverse resources to probe molecular function. While thousands of previously unseen genes have been identified, their roles are yet to be determined. Thus a fundamental goal in bioinformatics is to find the biological function of uncharacterized genes on a genomic wide scale.

In this project we aim to build a general system that, given a gene (protein coding or non-coding RNA), can suggest a hypothesis for its function that can then be validated experimentally. This multi-scale prediction will be carried out exploiting the structure and information recorded in biological networks. More specifically the first aim is dedicated to the development of biological network analysis tools using logic circuit models. For this we are going to integrate network and functional genomics data to characterize gene interactions in an objective manner using logical operations. The second aim is focused on integrating phenotype annotations extracted from phenotype ontologies with biological network information. For this we propose to construct a relational network and predict phenotypes of uncharacterized genes using the guilt by association principle. Finally, our third aim is to combine the information obtained from AIMs 1 and 2 towards optimizing the phenotypic function predictions. We will also develop a number of stand-alone and web-based applications for phenotype function prediction using biological networks and logic circuit models.

The project will be developed as a bilateral collaboration between the groups of Dr Mark Gerstein (US NSF PI) at Yale University and Dr Alberto Paccanaro (UK BBSRC PI) at Royal Holloway University of London. The two PIs have a long history of successful collaborations on many network based approaches for biological problems.

INTELLECTUAL MERIT

Our previous work on analysing biological system using networks, has already proven to have a substantial impact in the genome analysis community. Also publications stemming from the network construction, analysis and predictions tools that we developed have generate a large number of citations. In relations to its core aim, key aspects of the current proposal include the development of robust computational resources to help understand and predict gene function that will be freely provided to the scientific community. Much of the proposals intellectual merit relies in its new approach for the study of gene function, integrating objective classification methods using logic circuit models, with biological networks and semi-supervised machine learning techniques to infer phenotypic functions.

BROADER IMPACTS

The proposed work will lead to impacts beyond the functional genomics community – in particular, in the fields of personal genomics and network science. In recent years, there has been an avalanche of personal genomics sequencing data, providing new needs for the large-scale interpretation and identification of gene function.

The proposal to integrate function prediction with networks will have direct relevance to the network scientific community. It will help transform biological networks from static node-and-edge representations into dynamic entities that better reflect reality. Insights gained from such network analyses may reveal relationships between higher-level network characteristics and the individual attributes/functions of each of the nodes (genes).

The results will be integrated into a series of web tools that will serve as a bridge for investigators coming from networks and functional genomics/protein function communities. Finally, we outline a comprehensive set of plans to use the developed tools as devices for education in bioinformatics, as well as designing, organizing and running various workshop and webinars on

the topic.