Pilot Work for Task 1

Models:

Establishment of a computational platform of LogicTRN. We have established a comprehensive computational platform, LogicTRN, to determine the TF-TF interactions in gene regulation and to reconstruct TRNs. By combining cis-regulatory logics and transcriptional kinetics into one model framework, LogicTRN can naturally integrate gene expression and TF-DNA binding as indivisible aspects of gene transcription in order to identify TF regulatory logics. Using LogicTRN, we constructed the dynamic TRNs in breast cancer cell development. We first decoded how ESR1 interact with TFs (FOXA1, GATA3, FOXM1, CEBPB, JUN, FOS, JUND) or cofactors (EP300 and CTCF) during the E2-induced tumor progression. The resultant networks indicate that some key logics control dynamic transcriptions a majority of important genes. Meanwhile, using LogicTRN, we have reconstructed a transcriptional regulatory network during the human pluripotent stem cells (hiPSC)-derived cardiomyocytes (CM) differentiation. LogicTRN identified 12 regulatory logics targeting the eight key cardiac TFs. The derived networks predicted MESP1 is a main regulator to control transcription of other TF genes, supporting MESP1 function as a master player in determining cardiac cell lineage commitment by triggering the cardiac TFs. The manuscript is under reviewed by Nature Communications.

Figure 1. Regulatory logics and the networks for the E2-induced breast cancer development. A, models of logics in regulating TF genes FOXA1 and FOXM1. B, top ranked TF logics, number of differentially expressed genes (DEGs) at T1-T2 and T2-T3 stages, and their targeted pathways or biological processes. C-D, TRNs showing that GATA3 and ESR1 logics are dominant regulators in facilitating cell cycle and proliferation at T1-T2 (C) and T2-T3 (D) stages. In C-D, blue triangle nodes represent TFs, cyan circle nodes represent regulatory logics, oval nodes
Establishment of LDC network reconstruction methodology. We developed a linear-dynamic cascaded (LDC) method to reconstruct dynamic gene networks from the sample-based transcriptional data. LDC was developed based on the intra-stage steady-rate assumption and the continuity assumption, which can properly characterize the dynamical and continuous nature of gene transcription in a biological process. Simulation study shows that the LDC method can significantly improve the performance of network inference comparing to static approaches. LDC was further applied to reconstruct dynamic gene networks of hepatocellular carcinoma (HCC) progression. The derived HCC networks were verified by the functional analysis and network enrichment analysis. Furthermore, it was shown that the modularity and network rewiring in the HCC networks can clearly characterize the dynamical patterns of HCC progression. (Hailong Zhu, et al, NAR, 2012).

Figure 2. Gene regulatory networks during hepatocellular carcinoma (HCC) progression (A) in normal stage, (B) cirrhotic stage, (C) dysplastic stage, (D) early HCC stage and (E) advanced HCC stage.
Construct evolution- and structure-based computational strategy for revealing the impact of deleterious missense mutations on MODY 2

Heterozygous mutations in the central glycolytic enzyme glucokinase (GCK) can result in an autosomal dominant inherited disease, namely maturity-onset diabetes of the young, type 2 (MODY 2). MODY 2 is characterised by early onset: it usually appears before 25 years of age and presents as a mild form of hyperglycaemia. In recent years, the number of known GCK mutations has markedly increased. As a result, interpreting which mutations cause a disease or confer susceptibility to a disease and characterising these deleterious mutations can be a difficult task in large-scale analyses and may be impossible when using a structural perspective. The laborious and time-consuming nature of the experimental analysis led us to attempt to develop a cost-effective computational pipeline for diabetic research that is based on the fundamentals of protein biophysics and that facilitates our understanding of the relationship between phenotypic effects and evolutionary processes. In this study, we investigate missense mutations in the GCK gene by using a wide array of evolution- and structure-based computational methods, such as SIFT, PolyPhen2, PhD-SNP, SNAP, SNPs&GO, fathmm, and Align GVGD. Based on the computational prediction scores obtained using these methods, three mutations, namely E70K, A188T, and W257R, were identified as highly deleterious on the basis of their effects on protein structure and function. Using the evolutionary conservation predictors Consurf and Scorecons, we further demonstrated that most of the predicted deleterious mutations, including E70K, A188T, and W257R, occur in highly conserved regions of GCK. The effects of the mutations on protein stability were computed using PoPMusic 2.1, I-mutant 3.0, and Dmutant. We also conducted molecular dynamics (MD) simulation analysis through in silico modelling to investigate the conformational differences between the native and the mutant proteins and found that the identified deleterious mutations alter the stability, flexibility, and solvent-accessible surface area of the protein. Furthermore, the functional role of each SNP in GCK was identified and characterised using SNPeffect 4.0, F-SNP, and FASTSNP. We hope that the observed results aid in the identification of disease-associated mutations that affect protein structure and function. Our in silico findings provide a new perspective on the role of GCK mutations in MODY2 from an evolution-based structure-centric point of view. The computational architecture described in this paper can be used to predict the most appropriate disease phenotypes for large-genome sequencing projects and to provide individualised drug therapy for complex diseases such as diabetes. (Hailong Zhu, et al, Theranostics, 2014).
Figure 3. Conservation analysis of the GCK protein sequence from 1 to 465 aa using ConSurf. The amino acids are coloured based on their conservation grades and conservation levels. A grade of 1 indicates rapidly evolving (variable) sites, which are colour-coded in turquoise; 5 indicates sites that are evolving at an average rate, which are coloured white; and 9 indicates slowly evolving (evolutionarily conserved) sites, which are colour-coded in maroon.
Reference:
