VISUAL ANALYTICS OF CHROMOSOME 3D STRUCTURE
Kenneth Sabir
Dr Sean O’Donoghue (Garvan Institute of Medical Research, CSIRO), Prof. Seokhee Hong
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Understanding Function from Structure

Background
For the past 50 years, the study of genetics has mainly focused on analysing DNA in a linear fashion. However, increasing data points to the importance of the three-dimensional structure of chromosomes, for example groups of co-expressed genes are believed to occur in close proximity due to the physical winding of the chromosomes, even though they may appear distant within the one-dimensional string of nucleotide residues, or may even occur on separate chromosomes.

Hypothesis
“Applying visual analytics to understand genomic proximity relationships will help gain knowledge into the spatial properties of the genome and gain insight about its underlying biological function.”

Contribution
My research will apply visual analytics to a newly emerging type of experimental data that is helping to shed light on the 3D structure of chromosomes, and how this structure impacts the coordination of gene expression and other key genomic functions.

I have developed an algorithm for finding all the chromosomal regions (including genes) close in proximity to a physical location. This algorithm can be systematically run across the entire genome to produce numerous gene sets to each be analysed for any relationships of functional similarity or experimental interaction.

I have created a database of terms from all full text research articles from PubMedCentral. Map-Reduce techniques are used to find all papers for a gene, and then all terms for each of those papers. The frequency of terms for a given gene list can then be compared to a given background to find over-represented terms.

Source dictionaries indexed:
- Wikipedia,
- Proteins,
- Chemicals,
- Diseases,
- Pathways

Unpublished raw data provided by Lars Jensen (Uni of Copenhagen, Denmark)

Map-Reduce - new database of full-text PubMed terms

MONGO DB
GENE -> [PMID]
PMID -> 6 BILLION TERMS

MAP REDUCE