Raw Data Preparation

Transcriptome Data:

The processing of the raw microarray data was performed in an R-Bioconductor environment. The Affymetrix gene chip .cel files were read by using *affy* R-package. The raw expression data was normalized by Robust Multi-array Average (RMA) pre-processing method. The log2 gene expression values are provided as the input score of each gene. However users can also provide any scaled, ranked gene expression value or significance analysis result as input data file. Then the following steps will automatically done by the signal transduction score flow algorithm Cytoscape plug-in.

The rank score of each gene was computed based on fold change ratios between experiments; the expression change of a gene was called the *score* of that gene. If *r(x)* indicates the order of the score *x* when all the scores are in sorted ascending order, then the rank of *x, R(x)* is given by

$$R(x) = \frac{r(x)}{TS}, \qquad (1)$$

where *TS* is the total number of scores. Note that *R(x)* score ranges from 0 to 1. *R(x)* was scaled to provide better interpretation during the scoring of pathways as follows.

$$I(x) = (1 - R(x)) * 100. \qquad (2)$$

I(x) score might be interpreted as the *individual rank of x* which was obtained from either microarray or ChIP-seq data sources. On the other hand, the expression value of gene *x* was exactly provided as the *I(x)* score of *x*, if the microarray experiment is based on Affymetrix gene chip which does not have negative values unlike double channel expression arrays.

ChIP-seq Data

ChIP-seq Analysis:

The CisGenome framework performs an essential analysis of ChIP-seq data, and thus we used its peak detection method for our raw data. The first step was identification of peak regions from short DNA-reads. Thus the peak detection method of CisGenome tool was used (http://www.biostat.jhsph.edu/~hji/cisgenome).

The peak detection method essentially searches the entire genome with a sliding window (width=100, slide=25) and determines regions with read counts greater than 10. The next step of the analysis was the mapping of experiment-related (TF site neighboring) genes to the peak regions. We mapped the transcription start site (TSS) of each gene in the genome within a distance of ±10000 base pair.

When a peak region in that range was found within the set region, DNA-read counts were taken for the rank value of that gene. $r(x)$ is set to 1 for the gene x, which is located in the neighbouring region of the most significant (having the highest read count) peak region. Hence, the $I(x)$ value of that gene x is very close to 100.