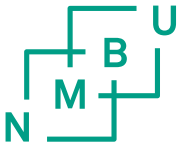

U
B
M
N

Filtering the Co-expression Networks of Populus Trichocarpa

A Thesis Submitted
To Department of Chemistry,
Biotechnology and Food Science,
Norwegian University of Life Sciences，
Ås, Norway.

By
Xiao Nie
Supervisor: Torgeir R. Hvidsten

May, 2014

Acknowledgement

I would like to thank Professor Torgeir R. Hvidsten and all those who have given me their generous helps, commitment and enthusiasm, which have been the major driving force to complete the current paper.

Ås Norway

May. 2014
Xiao Nie

Abstract

Before further study of gene-Co-expression network, it is prior to filter the network, the analysis aim to check whether co-expression network can be improved by filtering .Through different tools and algorithm to explore the optimal network, can cover enrichment analysis, protein prediction, efficient clustering etc.

Key words: Filtering Co-expression network

Table of Contents

Chapter One Introduction

1.1 Research Background

Gene network is a group of lined data sets and tools used to study complex networks of genes, molecules and gene function. The development of high-throughput genome sequencing and bioinformatics technologies have achieved the depth understanding and researching of molecular biology. For the section of genes co-expression, even we conquer the difficulty of handling huge amounts of biological data and successfully extract the information by organizing a network cover the transcription process. Still face the difficulty of explore internal connection among countless pairs of genes. One guideline for alleviation of the task is to filter the less meaningful elements, or remove genes considered irrelevant, generally genes whose variation across samples is less than an arbitrary threshold value will be deleted. For put less work under the limited resource. Because during the procedure exists the situation of wrong way of filtering or unnecessary filtering, the worse is remove the useful information. We believe it is necessary to construct a meticulous and authentic expression network which is the basement to start a filtering. After that we attempt to figure it out whether filtering will strength and prove the network in many ways through multiple tools. When facing a large-scale network, take the whole network to compare with the filtered network. We can understand whether the filtering is feasible and necessary. Does filtering really reduce the amount of calculation? Is the principle of focusing on most representative content the most crucial question? Meanwhile, modules are the small ones but can see big things in the network, our observation for checking the difference of filtered and unfiltered network is carried out on concerned module.

Populus trichocarpa is an economically important source of timber, its rapid growth and compact genome size has made it to a woody model organism in plant biology, entire genome has been sequenced in 2006 as well. Here comes several advantages why it is used as a model species, first the genome size reaches 485 MB which is larger than most of the other model plant, then the high

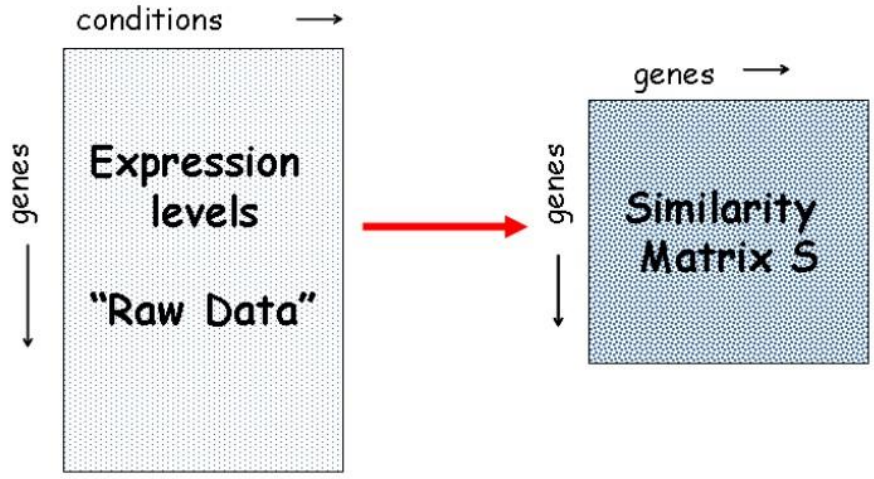
speed of growth result in reproductive maturity 4 to 6 years. Last and most important is that Populus trichocarpa represents a phenotypically diverse genus, Yong's research (1) reveals significant genetic differences between the roots, leaves and branches of the same tree, this kind of situation named somatic mosaicism, occurs when the somatic cells of the body are more than one genotype. Arousing the people's interests on evolutionary biology.

1.2 Build genes co-expression network

We downloaded Affymetrix experiments for Populus (462 experiments from multiple populus species). Raw data were normalized using the RMA normalisation method as implemented in the Bioconductor package 'affy' for the R statistical language using default settings. In the experiments, tissue samples were collected to create transcript population, Hybridizing labelled mRNA that we are interested, and the expression levels of genes are then derived from the hybridization intensities, the experiment resulted in an expression matrix which records the expression levels of different genes. In the procedure of wood formation, the developing secondary xylem of poplar is highly organized and easily recognized with distinct boundaries between the different developmental stages. For getting stable mRNA, sampled through the wood development region and subsequently analysed the sample by using cDNA-microarray, in different kinds of cell and sampled in different zone, diverse expression level results were collected, compare the similarity, normalize to a total general frame, we also get that lots of poplar genes analysed by transcript profiling are homologous to exist identified genes in other plant species, such as Arabidopsis. (2)

Collection of diverse expression data is the fundament of gene network, massive information about presentation and comparisons of proteins are supposed to be treated by one easily calculate and visual pattern, Such as Gene co-expression networks constructed from gene expression microarray data，capture the relationships between transcripts. The goal of building co-expression network is summarize the relationship between individuals from each pair of genes inside, fulfil the excellent pattern which can multiple and interpret the network thoroughly. For the purpose of easily module finding and network construction, the raw gene expression matrix is commonly transformed into a similarity matrix. The aim of this similarity matrix is to identify pairs of individuals which pair are closely related, for many kinds of measures, the similarity value is depend on their own expression composition. The matrix is symmetric, because the value type of similarity here is scalar instead of vector, in other words, the order of gene in the expression comparison does not affect the

value, such like compare 2 different personal names. Concrete Affymetrix values can be calculated by special algorithm. Here an unsupervised network inference method was used. Context likelihood of relatedness (CLR), uses transcriptional profiles of an organism across set of conditions to systematically determine transcriptional regulatory interactions. Through mutual information to score the similarity between the pairs' expression levels in a set of microarrays. Regulatory result comes from the biological chemicals (mRNA and proteins arise with gene expression) interact with other all kinds of proteins and molecules in different environment, the various of cellular biological function between transcription factor comprise a regulatory networks， the definition which is merging recently years, Owing the cellular component have to continually adapt the changing conditions by alter their gene expression patterns, The network describes gene expression as a function of regulatory inputs specified by interactions between proteins and DNA. But it's not practical to focus in great detail and mess small proteins, instead of considering the autonomous, taking the module as unit of network to account.(3) These "links" among proteins (or DNA sequences, also the mRNAs transcribed from DNA) can be used to analyse and calculate to create a new kind similarity matrix. A threshold and criteria was set to detect the mutual information whether their potential regulator is higher or lower, CLR increases the contrast between the physical interactions and the indirect relationships by taking the network context of each relationship into account.(4)



The Affymetrix data we got was a big txt file which only contain the similarity value, inputted them to language R platform: R studio, programmed and edited them into matrix, annotated with

gene names, transformed the triple matrix to symmetric matrix. The genes Co-expression network is composed of almost 30 thousand genes (31995 genes name). The form of gene name is with prefix POPTR and number accession, such as POPTR_0001s00240, actually the protein accession number just add dot and 1 at the end, like POPTR_0001s00240.1 in protein dataset. The genes are arranged in order of accession number.Every gene represents a code in the network. Like we have mentioned, in the matrix S, Sij and Sji are equivalent, so the Affymetrix matrix is symmetric. Inside contains lots of 0.000000 but no NA values, means the two genes have no similarity at all. With higher value more common expressed they are. The number of pairs is the genes' numbers squared. So the amount of information is very sizeable.

1.3 Using Cytoscape to achieve visualization of network

Cytoscape is an open source software for integrating bimolecular interaction networks with high-throughput expression data and related framework. It provides basic functionality to layout and analyse the network and to integrate the network with visually existed data. It's powerful to transform interaction into visualization of nodes and edges as a two-dimensional network. It provides user several effective and visual features to highlight aspects of the network. The strong point of Cytoscape that it is able to ignores what is the network made up of, no matter the element is gene, protein, or module, all can be symbolized as a node and interactions represented as a link (edge) between nodes, In different situations the parameters themselves are same in a mathematically sense, but the biological meaning is totally different once the object is changed, Anyhow we manage to attain a lot information from micro to macro models through the software.

Make sure the visualization is more directly perceived, here I just use one section of whole gene network which contain 3000 genes, then programmed them into three columns, two of them are gene name, another is interaction value or similarity value in the network, even so they are quite large, then we remove the interaction value is lower than 4, at last the filtered dataset is extracted from a sub network of 3000 genes, it includes 792 nodes and 3410 edges. We import the previous handled data, then use the function NetworkAnalyzer, it will compute the degree of both input and output, also a variety of other parameters. The system will store computed values as attributes of the corresponding nodes and edges. It means the layout of each node and the length of every edge is elaborate. Let's one by one check the Network parameters: Number of connected components is an important topological invariant of a graph. Connected components mean the isolated group of

vertices, when two vertices are connected to each other, and has no connection to additional vertices. This parameter indicates the connectivity of a network, lower value of connected components suggests a stronger connectivity, If the number is 1 represent all the nodes are connected. The network diameter is the largest distance of between two nodes. The average shortest path length means the expected distance between two connected nodes. Clustering coefficient is the proportion of connections among its neighbours which are really reached then compared with the number of all possible connections. So coefficient=1 there is a total full connected network, every node has edges to other nodes, coefficient=0 the other nodes are isolated, or all the nodes with less than two neighbours. Network clustering coefficient is the average of the clustering coefficients for all nodes in the network. The definition has been used to interpret network architecture. Neighbourhood connectivity is the node's number of neighbours, or defined as the probability that two randomly selected neighbours are connected each other. Closeness centrality is a measure of how fast information spreads from a given node to other reachable nodes in the network (5).

An exhaustive topological analysis of huge network is cumbersome and time consuming, in detail, computation of local parameters are dramatic faster than computation of global parameters. For reduce the complexity of a large interaction network, it's selectively display subsets of nodes and edges. But the result can explain and exhibit various aspects of the whole network thoroughly in both visual graph and statistical parameter. And the motivation for proposed filtering is self-evident, too many genes contribute little information, the huge data will cause pressure on computing, it's proved that cluster analysis and principal components are strongly affected by filtering. (6) After import and define the type of each column from the imported data. It is available to choose different kinds of layout, here we pick the default graph. Then use the function NetworkAnalyzer which emphasize on the connectivity of each node, combine with the distance, shared neighbours and shortest path to highlight the extraordinary nodes and the most penetrating path (take notice that not only the nodes are coloured but also the edges.)

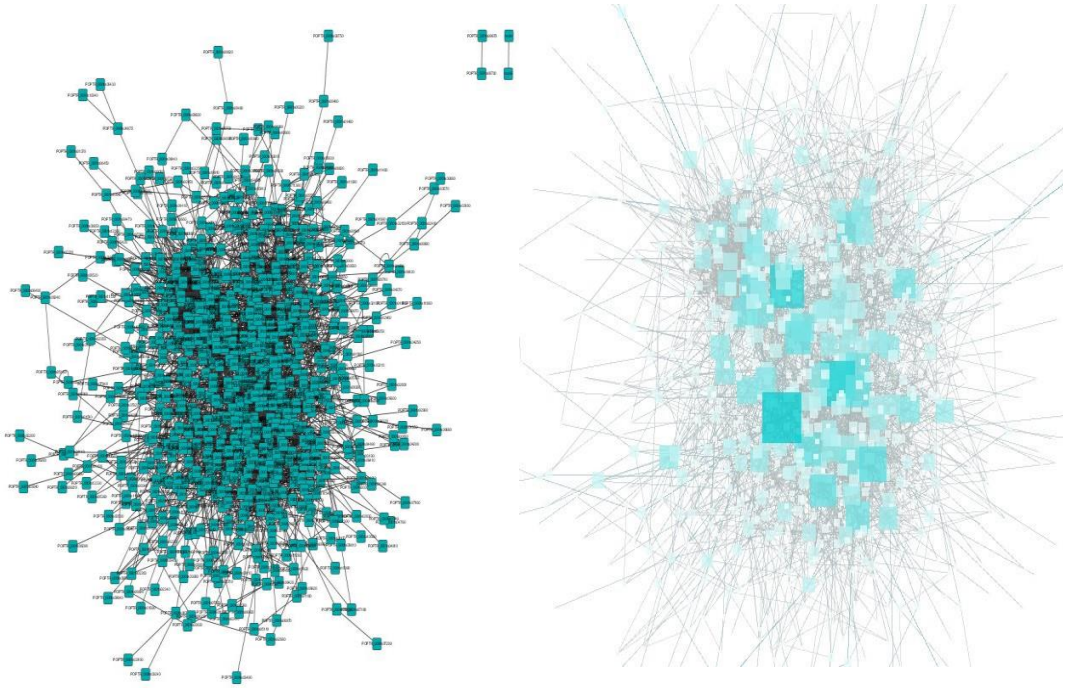


Figure1:the left is layout of filtered network. The right is layout of the same network after using the function NetworkAnalyzer

From the graph we can feature the network thoroughly, each edge means a linkage between two nodes longer distance means higher connection degree, in the right side of graph the larger size of square means more neighbours it has, colour depth is related to connection degree and in direct proportionate, the layout of network is 2-demensional, the node with less connection will be put on the border of network covered area, on the contrary, highly connected gene group is located in the central part.

Clustering coefficient :	**0.020**	Number of nodes :	**792**
Connected components :	**5**	Network density :	**0.011**
Network diameter :	**9**	Network heterogeneity :	**0.971**
Network radius :	**1**	Isolated nodes :	**2**
Network centralization :	**0.056**	Number of self-loops :	**16**
Shortest paths :	**617014 (98%)**	Multi-edge node pairs :	**0**
Characteristic path length :	**3.777**	Analysis time (sec) :	**37.937**
Avg. number of neighbors :	**8.571**		

Figure 2: the table of simple parameters of the Analyse Network result

The Network Analysis shows charts of the distribution of node degrees, neighbourhood connectivity, and average clustering coefficients. Here the average value is 0.02, the sub network does not has lots of shared nodes, but the clustering coefficient of central part is 0.025, we can interpret that the core part of the network is more connected and interactive. According the PCA

statistical analysis function named Extract connected components, only 6 edges regarded as irrelevant. So the whole gene network is highly connected. But the central of network with higher density of nodes and edges. Then branch out with very less destinations. In other word the system default put significant nodes at the central part. The core node with most neighbours are accountable and mostly supposed to analysis. That's the key for later series research. And the other nodes are quite equally distributed. We would extract the central for sub network analysis, in terms of Shortest Path Length Distribution, closeness centrality ,etc. are more stable and increased, means the core of the network(the central part of network) is more active and information interoperable, we can find the bar chart shows the Frequency is similar to normal distribution, means the similarity between nodes after filtering is stable, most of length of edges in some aspects explain the gene expression is controlled and affected by lots of genes. Not an individual and independent event of single RNA but a continuous efforts which means regulated by multiple genes. After all, even the characters we mentioned just cover a small part of whole network, but it is possible to extend and predict larger network.

Chapter Two Weighted Gene Co-expression Network Analysis

2.1 Introduction of R package: "WGCNA"

In weighted gene co-expression, network analysis is a systems biological method for describing the correlation patterns among genes across microarray samples. WGCNA can be used for finding clusters of highly correlated genes and co-expression modules, for calculating module membership measure and highly connected intramodular hub genes. Network analysis methods are increasingly used to represent the interactions of genes and genes' transcripts. Maybe these genes and gene production linked have a similar biological function or part of the same biological pathway. (7) According the definition of weighted network which in a network exist ties among nodes have weights assigned to them. A gene network is a system whose elements are somehow connected. Gene networks are increasing being used in bioinformatics applications .With the explosion of biological information coming up, the ability to handle huge data become a key point. Using Weighted Gene Co-expression Network Analysis (WGCNA) a network analysis method which has

been widely used to identify biologically meaningful gene modules in a great deal of organism. WCGNA has been implemented in R, a free and open source statistical programming language which is widely used. The efforts are not limited in correlation between individual pairs of genes, and also the extent to gathering similar genes module. WGCNA is an advanced application of hierarchical clustering. The basic principle is start by assigning each item to its own cluster, so that if you have N items, now you have N clusters. Each containing just one item. Through the algorithm make the distances between clusters represent the similarities between the items what they contain. Try to find the closest or most similar pair of clusters and merge them into a single cluster. Compute the distances between the new cluster and each of the old cluster. Repeat the former steps until all items are clustered into a single cluster of size N.

We transform the Affymetrix based matrix to adjacency matrix by using adjacency function which will standardized original scale to adjacency matrix scale from 0 to 1.

a_{ij}= 1 (there is an edge from node j to node i) and a_{ij}=0(otherwise)

After construction of the similarity (adjacency) matrix, a threshold must be imported to separate significant, biologically meaningful modules. Automatic block-wise network construction and module detection, constructing a weighted gene network require the choice of the soft thresholding power to which co-expression similarity is raised to calculate adjacency, aim to pick an appropriate soft-thresholding power for network construction. The power we get will influence the whole clustering, the key point to choose the soft-thresholding is relied on the criterion of approximate scale-free topology. We know in the scale-free network, the clustering coefficient distribution will decrease as the node degree increase. So even high degree can find the hub gene which has highest degree, in graph theory, the degree of a vertex is the number of edges incident to the other vertex, in the biology network, vertex is node here, and hub gene is the key to define and bridge the different expressed module. On the contrary the high degree will remove too much nodes, make lots of methods and algorithms impotent, it is a dilemma. We tend to find an appropriate soft-thresholding power. (8)

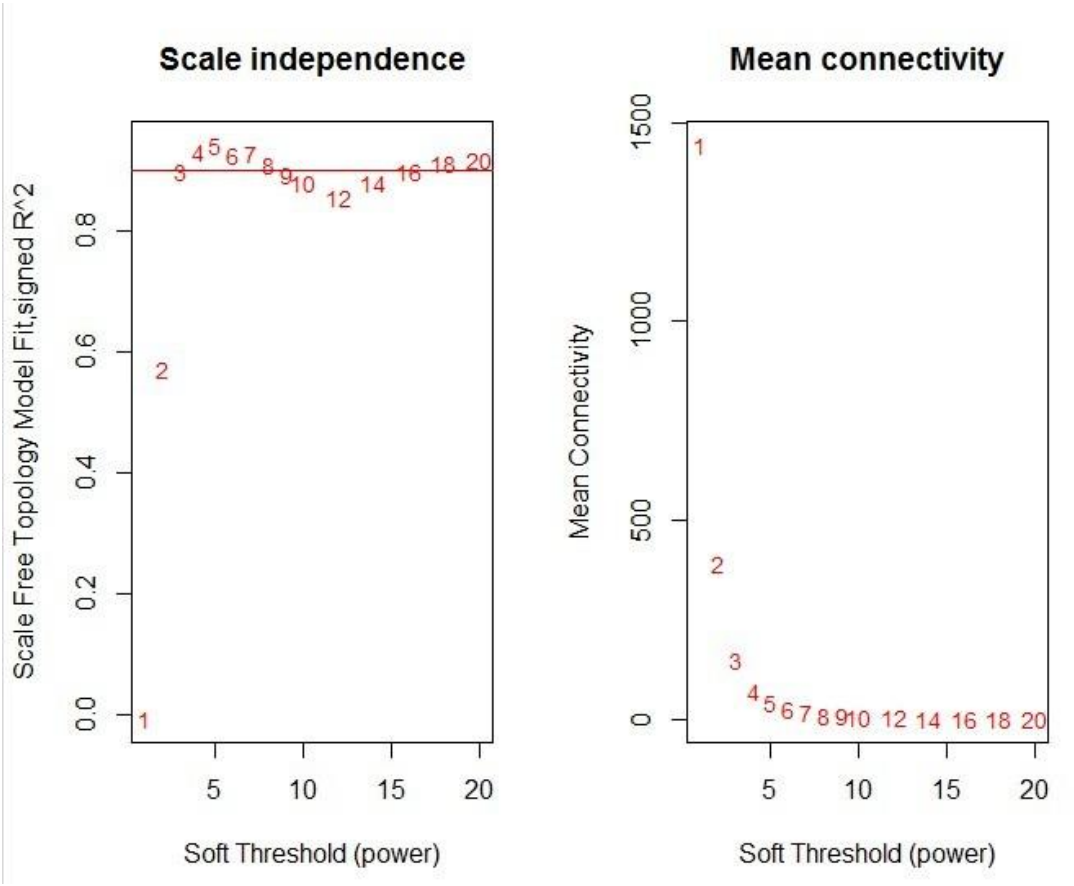


Figure3: Analysis of network topology for various soft-thresholding powers. The left panel shows the scale-free fit index (y-axis, denoted as scale.law.R.2 or power law) as a function of the soft-thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis).

Power 4 is lowest power for the fit index reaching a high value (over 90%) and the highest mean connectivity, we choose that as reasonable criterion. So in further plotting of clustering, the branch will end at height 0.4, because the default minimum height was set here.

2.2 Topological Overlap Matrix (TOM) based network construction

WGCNA methods depend on topological properties. Several studies have shown that two proteins having higher topological overlap are more likely to belong to the same functional class than those less topological overlap. Topological properties are invariant under homeomorphisms, in other word, it's a property of the space that can be expressed using open sets. The common problem in topology is decided whether two topological spaces from a pair are hemimorphic or not, because proteins' construct are complicated and changed always, similar to protein continuous deformation, the topological comparisons are essential. Also easily reach new modelling that relies on understood

statistical methods and improves on complex networks. The resulting topological overlap matrix converted to a dissimilarity measure and submitted to hierarchical clustering. The dendrogram showed below demonstrates the similarity expressed genes divided into different branches according the principle of hierarchical clustering. The topological overlap matrix can be used for module definition.

We transform the former adjacency matrix to topological overlap matrix.

a_{ij}= 1 (node j to node i share the same neighbors) and a_{ij}=0(they do not have common neighbors)

The topological overlap of two nodes reflects their similarity in terms of commonality of the nodes they connect to. Two nodes have high topological overlap if they are connected to roughly the same group of nodes in the network, such as they share the same neighbourhood. Because the topological overlap matrix is symmetric and the value is limited between 0 to 1, the assumptions is similar to adjacency matrix, roughly speaking, the topological overlap matrix can be considered as smoothed out version of the adjacency matrix. The definition of topological overlap provide other measure of connection strength based on shared neighbour's(nodes), Hierarchical clustering is a widely used method for detecting clusters in genomic data. In the experiment, we transform the similarity matrix to adjacency. Here adjacency is constructed from similarity degree. High numbers means high similarity. Then use hierarchical clustering to produce a hierarchical clustering tree of genes to identify modules. The graph is drew in the form of dendrogram. As we know hierarchical clustering has several advantages over other procedures. First it is a fully unsupervised method, in networks it is allowed to cluster all unites without having to specify a priori number of clusters present. Secondly the generation of a hierarchical tree provide both partitions of the network and visualization of clusters are combined into higher level groups.

Before further analysis, we need check the principle of defining a cluster criteria, to provide flexibility in clustering, we would better notice that a cluster is required to have a certain minimum number of member objects, in this point, our data is huge big, so we can pass it. Then if object is too far from a cluster are exclude to from the cluster even they have lots of common neighbours reached to same module. So in the Figure 4, you will see same coloured cluster are placed dispersedly. Also each cluster should be separated from its surrounding by a gap for easily distinguish those module. Last the core of each cluster should be tightly connected, make sure the core part of modules can fully cover the character of the module, it is representative in terms of

prediction and detection of module,

Since a particular module network may encode a pathway or a protein complex, here we use a clustering procedure to identify modules of nodes with high topological overlap. Module detection are clusters that result from using pairwise node from dissimilarity matrix as input of average linkage hierarchical clustering, actually lead into this dissimilarity is for easily detection of each node, it was defined as Dissimilarity of Topoverlap=1 – Topoverlap

Branches in the dendrogram are referred as modules. For a weighted network the topological overlapping measure interconnectedness. Height value 1 means topological overlap dissimilarity equal 0 when all of its neighbours are also neighbours of the other node or it is linked to the other node. By contrast, the lower value indicates the pairwise nodes are less unlinked or have less common neighbours. (9)

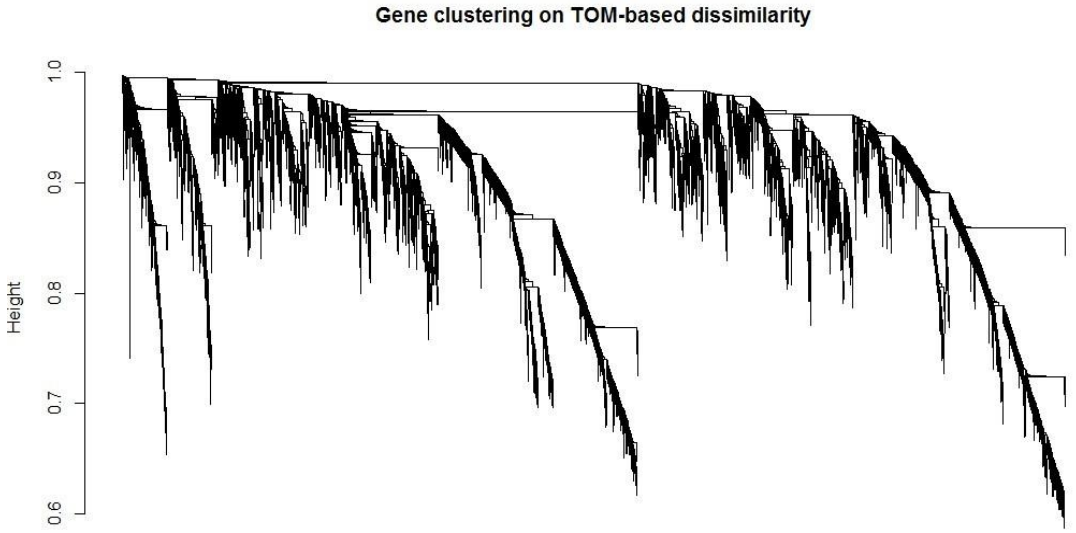


Figure4：Gene clustering on TOM-based dissimilarity, each vertical black line represent one gene.

2.3 Modules detection and research

There exist evidence that genes the protein originated from are organized into functional modules rely on cellular processes and pathways. (10) Gene co-expression networks is used to describe the relationships between gene transcripts. Therefore a major goal of network analysis is to identify module of densely interconnected genes. This kind of groups are often identified by searching for genes with similar patterns of connection strengths to other genes. And module is a proper object to manipulate for further analysis, after all, module is the set of our interesting genes, for a whole genome of plant, devote ourselves to check the function of filter on individual gene is impractical.

A bigger object has to be come up with.

In hierarchical clustering, clusters are defined as branches of cluster tree. In the dendrogram clustering tree, each leaf or black shot vertical line, corresponds to a gene. Branches of the dendrogram group together densely interconnected highly co-expressed genes. Module identification amounts to the identification of individual branches. We used the Dynamic Tree Cut from package dynamicTreeCut, which is a top-down algorithm that relies solely on the dendrogram, the algorithm implements an adaptive branch pruning of hierarchical clustering dendrograms. When we choose a fixed height on the dendrogram, and each contiguous branch of objects below that height is considered a separate cluster. The horizontal line means the cut height, from this line cause the vertical distance to the core is the gap. The algorithm can be understood as overlap height cut in to large clusters that will be decomposed into smaller ones by subsequence processing, then extract the cluster-based dendrograms, identify significant break points based on forward run length, update the current list of clusters until no new clusters are produced.(11)

We use average linkage hierarchical clustering here to handle with gene dissimilarity measure to define a cluster tree of the network, when the branches were settled down, a height cut-off was introduced to divide a clustering one by one. Coloured modules correspond to branches of the dendrogram. As an alternative dissimilarity measure. The interpretation of the minimum cluster size, when we decide a number of the size, the resulting cluster size will never be smaller than that.

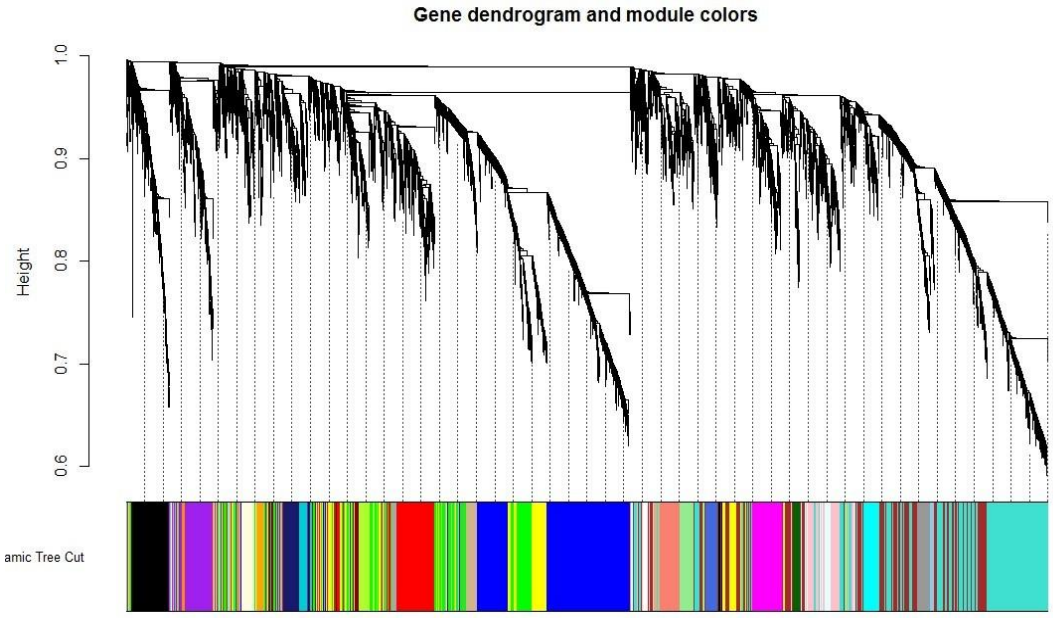


Figure5: Clustering dendrograms of genes, with dissimilarity based on topological overlap, together with assigned module colours.

Genes with highly correlated expression levels are biologically interesting for us, since they imply common regulatory mechanisms and participation in similar biological processes. An important aim of metabolic network analysis is to detect subsets (modules) of nodes that are tightly connected to each other. Generally the dendrogram exhibits distinct branches corresponding to the desired modules. But no single fixed cut height can identify them correctly. To automatically detect module, the tree cut method has to identify branches based on their shape instead absolute height. In terms of function, the fundamental concept of identifying modules is that a pair of RNA transcripts interacting with each other which has higher probability of share the same function. It's similar to detecting clusters in a network according the topological information. (12)

From the graph above, it's not hard to see too many modules will affect the representativeness of module itself and core gene which play a role of bridge and tie. To those expression profiles are very similar. We need to merge modules and narrow down the number of modules, but the breakthrough point should be the representative genes, those can achieve the integration of functional similar modules instead of slapdash geometrical merging. Furthermore the procedure of merging is exhibiting functional enrichment in the same categories. Specifically when merge the modules whose expression profiles are very similar. The DTC (Dynamic Tree Cut) may identify modules whose expression profiles are very similar.it will merge such modules since their genes are highly co-expressed. Here we skipped the Clustering dendrogram of genes, with dissimilarity based on topological overlap, together with assigned merged module colours and the original module colours. To quantify co-expression similarity of entire modules. Many module detection focus on these expression profiles are highly correlated. For such modules, if one representative gene can summarize the sub network or module expression profile. That gene is our interested module eigengene. We recalculate Eigengenes and cluster on their correlation. The module eigengene corresponds to the first principal component of a given module. It can be considered the most representative gene expression in a module.

We focus on their eigengenes which symbolise each module to cluster them on their correlations. In the Figure 5, the red line crossed whole colour-signed module to merge those highly related module reach the goal of simplification. After all too many modules will cause the difficulty in computation. And some different modules will contain similar genes would affect further analysis on module function and protein related research.

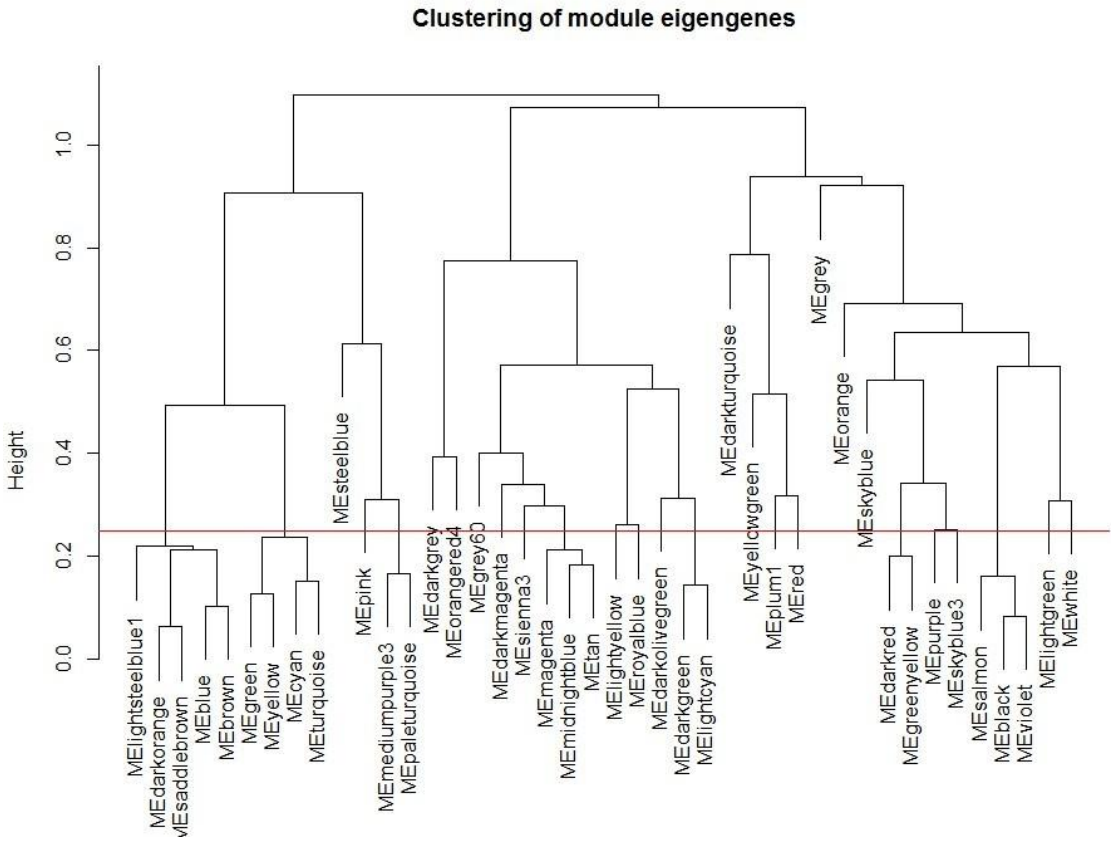


Figure 6: Clustering of genes, together with assigned merged module colours and the original module colours. The network consists of 42 modules which are shown in the dendrogram. Adjacent modules are more similar in expression than those more located distant instead nearby modules. Modules here are labelled with a prefix of "ME" plus specific colour. The branch was constructed using the eigenvectors derived from expression profiles of nodes in each module.

Module detection depends on several parameters choice, first and foremost is how to cut off branches of a hierarchical cluster tree, that's we have decided. But for our expected result, some adjust need to be import to merge from mess module. Here merged these modules' height lower than 0.25, and module eigengenes (defined as its first principal component) clustering the principal components, if 2 module eigengenes are highly correlated then the modules should be merge. A general rule is if the distance between the two is smaller than 0.15, they should be merged.

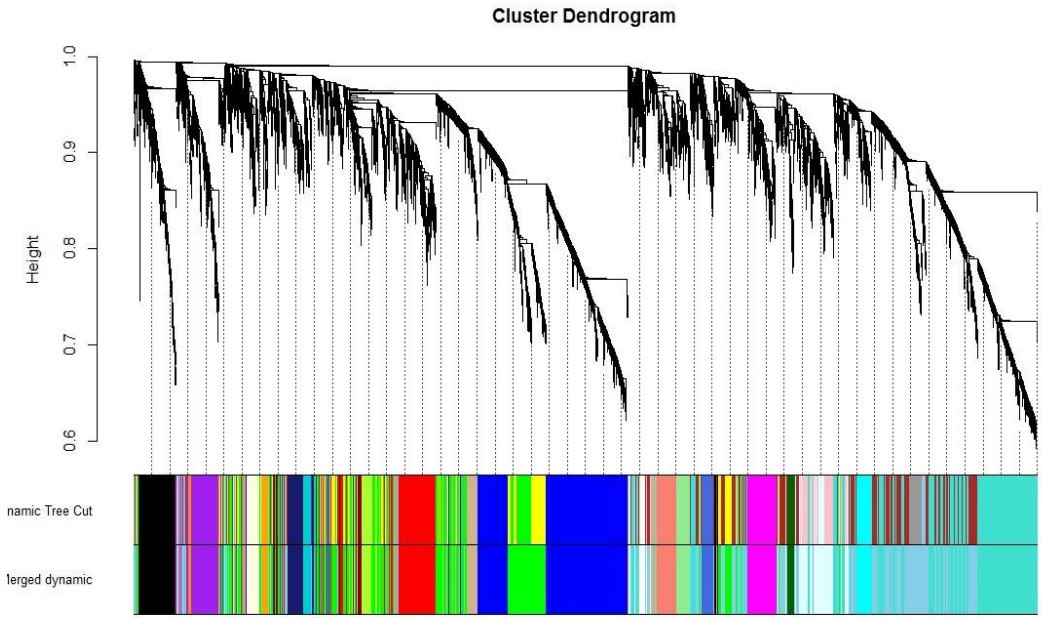


Figure 7: Clustering dendrogram of genes together with merged modules. Genes not assigned to any of the modules are coloured grey.

It is natural to think of what are these genes comprise each module. First we divert attention from module clustering to quantization of eigengene, we know that the hub gene that is most highly connected gene, should be signed up as a crucial element, and are thought to play an important role in organizing the behaviour of biological networks. But not every module has hub gene, or hub gene cannot reflect the whole module, as the most representative gene, ideally, quantization of eigengene is supposed to be fast way to detect and represent the whole module. The conception of gene significance is how we quantitate eigengene, in the data frame with genes and samples, a measure of gene significance is defined by forming the absolute value of the spearman correlation between trait and gene expression values. In the networks of gene expression similarity, gene significance is the average value of correlation inside the module. The mean gene significance for a particular module can be considered as a measure of module significance. We check Intramodular connectivity which measures how a given gene connected, or co-expressed, in terms of how important of gene, we pick up the average correlation of codes involved in the co-expression network. As we mentioned before, in the highly related network, stronger connection strength means more evolved and interesting. After the first stage of calculating the intramodular connectivity for each gene. We plot the gene significance against intramodular connectivity, as demonstration we choose up 2 of the module below.

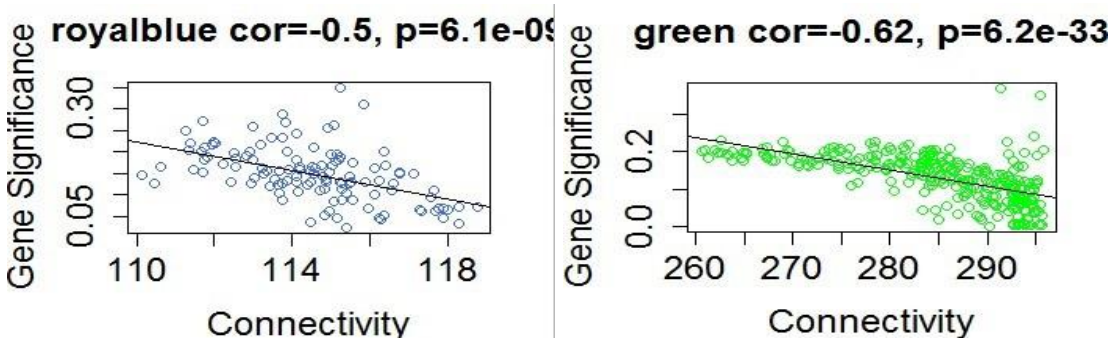


Figure 8: Gene significance (y-axis) verse intramodular connectivity (x-axis) showed up separately for 2 of whole module. There are some elements we need to look into carefully. The number of points in the graph, more plotted out means larger module. Then the trend of line, decreasing means with increasing connection the gene has less importance and influence to rest of genes in the module.

Our previous analysis has decide the interesting module which has higher significance. Gene significance was expressed as biologically significant in the specially appointed gene, the value of gene significance can take as either positive or negative. Next we locate the exact gene with high gene significance and high intramodular connectivity. In the hierarchical structure, hub gene plays a central role in directing the cellular response to a former given stimulus, actually most nodes make a small number of connections render a biological network. We can flexible adjust the criteria to limit the number of interesting gene to satisfy various analysis and online research. The combination of different gene from different module can make analysis more multivariate and accurate. So we can conclude that WGCNA is a highly robust, systems approach for integrating high-dimensional, multi-scale data, also identify modules and key driver genes that related to our concern outcomes. In terms of comparison of unfiltered genes network, we get more modules then the filtered one, Even the merge function also make great effort to simplify, but still have more we have now, in our viewpoint of handling huge network which we mentioned in the introduction part, too many modules undoubted make the job harsher, so in this point, it is apparent need not to spend too much effort on the advantages of module forming after filtering.

K-means clustering for genes network is another possible alternative method. As we know a popular and diverse set of clustering approaches that have been readily available. Except the hierarchical clustering we just used, K-means clustering actually is still one kind of traditional method. One of the simplest unsupervised learning algorithms Aim to assign the observations into k clusters in which each observation belongs to the cluster with the closest centroid, when all objects

have been assigned, recalculate the positions of the pre-set centroids. Repeat the process until the centroids no longer move.

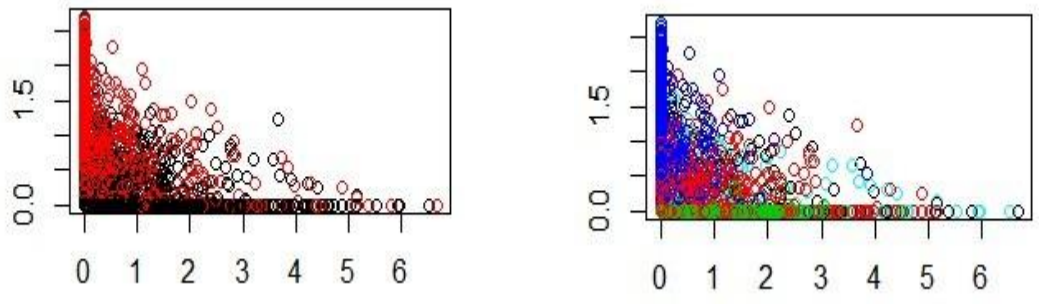


Figure 9: This plot is a tow-dimensional projection from a k dimensional space, where k is the number of variables. Left is 2 clusters and right is 5 clusters, the sample is part of whole genes network. The x-axis represents the distance of each node, and the y-axis is the orthogonal projection of the distance between clusters.

Actually k-means clustering is not common used in co-expression network, because the symmetric matrix is not suitable to be treated as several classes' limited data. K-means is more appropriate to implement on the data was built by several factor contribute. Even here the result is obvious. The other our big concerns is the filtering. Unlike hierarchical clustering, the limitation on the height related to the similarity score help us to filter less connected genes. Here one point in the flat or 3-dimension space is a whole, get the process become very hard to filter or remove less expressed genes. That is why majority research focus on hierarchical clustering. But from the plot, it is clear to find the modules, when the number of clusters increase, the points are distinguished. In the right graph, blue and green represent 2 low level of genes' transcript, but belong to 2 different clusters, the distance between the clusters is quite far. (The kmeans function tend to handle file with matrix format), but red and sky-blue represented modules with higher values and placed closer than the previous 2 clusters, anyway the modules with higher level of expression are our interested. Compare to K-means clustering, hierarchical is easier to filter, but when classify, sometimes overlap similarity is too prior instead of similarity score, that is why some module with large range of height, very long and narrow branch. K-means clustering is very directly perceived to divide the data.

There is a systematic comparison of clustering algorithms conclude that: (13) the effectiveness of clustering on the genes produced upon the positive match to known pathway. Methods were also tested to determine whether they were able to identify clusters by other clustering methods. Graph-

based techniques perform better. WGCNA includes functions for network construction, module detection, and gene selection, calculations of topological properties, data stimulation and visualization. Especially interfacing with external software make it more application-oriented.

Chapter Three Methods of predicting cellular component from sequence

3.1 Gene Ontology (GO)

Protein function is a broad conception with different meanings depending on context. In the computational pattern setting, protein function is described via terms from one of several controlled vocabularies. Because exist the degree of specificity to describe the function. The Gene Ontology (GO) is most prevalent of such controlled vocabulary systems. GO is a set of associations from biological phrased to specific genes that are either chosen from trained curators or generated automatically also the core of tools for the unification of biological access to all aspects of various bioinformatics, it is used to unify the representation of gene and related product attributes across all species. GO classifies proteins function into three separate categories, each of them consists of a set of terms that related to each other. Biological process means a biological objective to which the gene or gene product contributes. A process is finished via one or more ordered assemblies of molecular function. Process consist of chemical or physical transformation, seem like the stuff from input and output are something different. Molecular function we mentioned numerous times is defined as the biochemical activity of gene production, or elemental activities of a gene product at the molecular level. At last Cellular component refers to the place in the cell where a gene production is active. There terms reflect integrating of cell structure. (14)

The structure of GO show that the GO terms are organized hierarchically , higher level terms are more general and are assigned to more genes, caused that very common to see a large amount of genes annotated similar GO terms. Also more specific decedent terms are related to parents by either whole one or part of relationships. GO annotations can be understood as a common words in common language to describe aspects of a gene product's biology. Like the viewpoint we referred before, a coordinated change and bilateral efforts in the similarity matrix among many gene products

can produce potent biological effects. For Gene Ontology, how to cover and detect the subtle difference meanwhile address basic ontological principles. Another problem is some genes labelled numerous GO access, in the network module, become tricky to identify exact function and biological process, in another word, we have to find way out how to make sure the level architecture and hierarchical structure.

Name	GO terms	GO source
POPTR_0001s07390	GO:0018024 histone-lysine N-methyltransferase activity GO:0005634 nucleus GO:0008270 zinc ion binding GO:0016568 chromatin modification	PF02182 PF05033 PF00856
POPTR_0001s00210	GO:0000159 protein phosphatase type 2A complex GO:0023060 signal transmission GO:0023046 signaling process GO:0023052 signaling GO:0007165 signal transduction GO:0008601 protein phosphatase type 2A regulator activity	PF01603
POPTR_0001s02150	GO:0044260 cellular macromolecule metabolic process GO:0006281 DNA repair GO:0005634 nucleus GO:0003684 damaged DNA binding GO:0006974 response to DNA damage stimulus GO:0033554 cellular response to stress GO:0008853 exodeoxyribonuclease III activity	PF02144
POPTR_0001s06680	GO:0006468 protein amino acid phosphorylation GO:0005524 ATP binding GO:0044260 cellular macromolecule metabolic process GO:0004672 protein kinase activity	PF00097 PF00069 PF00023
POPTR_0001s01590	GO:0003735 structural constituent of ribosome GO:0005840 ribosome GO:0034645 cellular macromolecule biosynthetic process GO:0006412 translation	PF01632

Figure 10: the presentation of GO in our research, in gene enrichment analysis tool agriGO, Take whole Populus trichocarpa genome transcript ID as suggested backgrounds or reference. Query list is composed mainly our interesting gene. Result in several key genes' Gene Ontology annotation for one special biological process.

Additionally, GO has a cellular component category that describe the places where the proteins is evolved, proteins with same molecular function can play a role in different pathways, also a pathway is composed of proteins of various molecular functions. This distinction affect which method are the most applicable for computational predictions of protein function of each type. Because molecular function always referred prediction based on sequence or structural similarity to proteins of known function. On the other hand, Biological are fundamentally collaborative, so it's natural and reasonable to predict them based on protein's interaction partners. Anyhow network analysis of interatomes can be a useful for prediction on protein's cellular component. (15)

3.2 **Application of prediction tools**

Proteins functions involved in various cellular processes also must be localized at their appropriate subcellular compartment to perform their desired function. Knowledge of subcellular localizations (SCLs) of plant proteins relates to their functions and aids in understanding the regulation of biological processes at the cellular level. Protein subcellular localization prediction involves the computational prediction of location where protein resides in a cell. In former clustering of proteins, we get an insight into how elementary biological unites interact to form complex cellular networks, also through the analysis of the detected functional modules enable to further research functional repertoire of proteins and biological processes in which they evolved. The module containing well-known proteins along with unknown function, expect the functional prediction for the unknown proteins and interesting proteins' unknown function. The research shows interactions between functional modules is highly connected, few biological processes are isolated unites.(16) The connection between functional modules can be used to examine the organization and coordination of multiple complex cellular processes and determine the way be organized in to pathways depend on the interaction data.

In recent years, various prediction methods have been developed, give us a huge flexibility in the application of prediction. According the different algorithm and practical situation, adopt and compare different result, the way people access to bioinformatics has been greatly enriched all of a sudden. These approaches can be classified into different categories based on the follow features. Whether be generated on sequence information, or generated by making use of Gene Ontology annotations, or generated by hybrid methods. Here we will introduce PlantLoc for predicting SCLs of plant proteins from sequences without any annotation information.

PlantLoc is an accurate web server for predicting plant protein subcellular localization by substantiality motif (LM). PlantLoc made the LMs by using training dataset, gradually generate a big library for lots kind of plant proteins. In the prediction procedure, according the hit numbers from query lists to all LM libraries. PlantLoc will let all query sequences have chance to hit each localization domains. LM was defined as a gapped or ungapped fragment of amino acids that were a conserved pattern in a subcellular domain and existed in the N-terminus peptides of sequences. In PlanLoc, the principle of running total prediction is composed on three steps. First the LM program which means query sequences was generated by assembling some LMs, be carried out to generate

candidates of LMs from training dataset. Then LMs were selected from millions of generated candidates. After selection, the LMs and their frequencies from the training sets for special SCL were constituted in an LM library. In the third step, a query sequence was identified as belonging to a subcellular domain according to the hit numbers of LMs in LM libraries. When a library had the highest hit number, the query was definitely identified as belonging to the domain.

In terms of clarify plant protein SCLs, PlantLoc with totally 80.8% precision. Compare to other webserver also based on SCL prediction, PlantLoc perform better on multi-localization. (17)

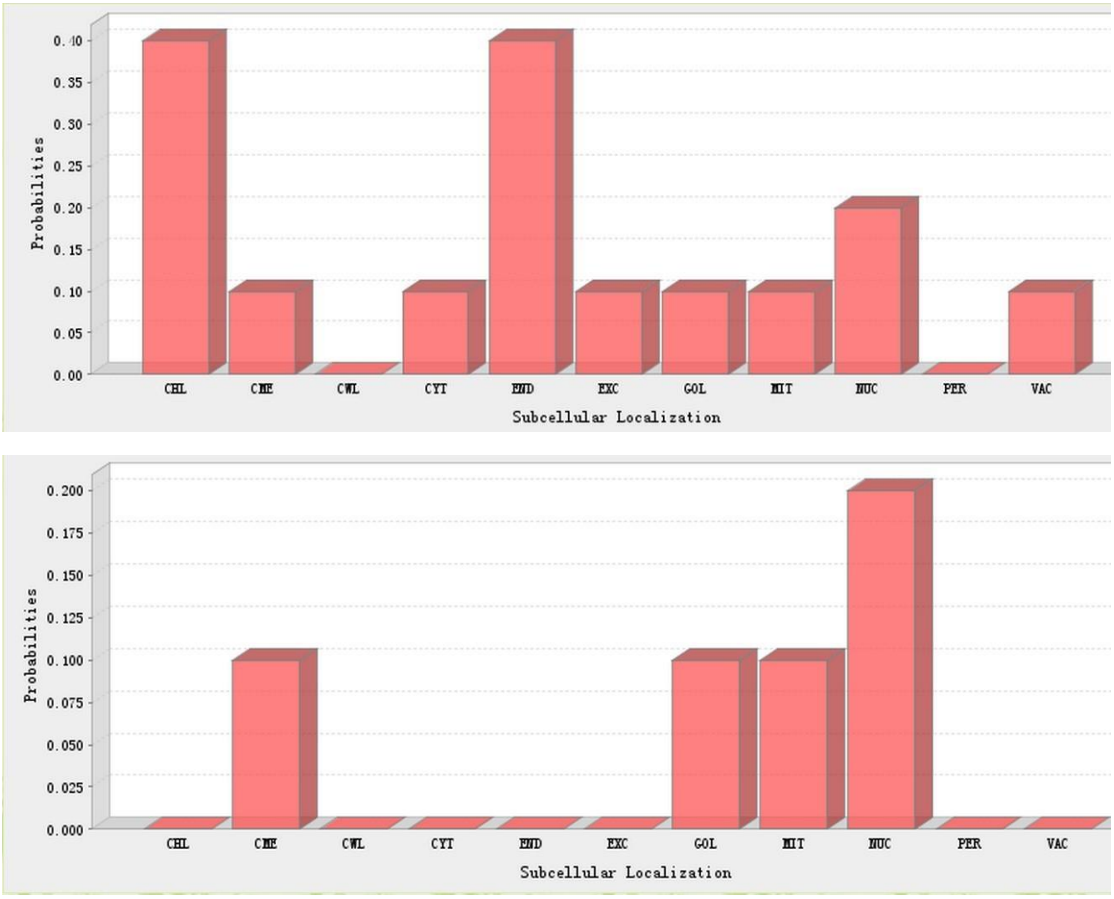


Figure 11: the result of subcellular localization prediction by gene product POPTR 0001s00210g and POPTR 0001s13930g, which CHL means chloroplast, END means endoplasmic reticulum, NUC means nucleus.

For individual gene result it's not good as we expected, because you are not sure for the exact position of the cell. Either the specification of the gene, owing to its function involved in almost every proceed of the biological activities of the process, you can find its track everywhere. Or there exist some degree of prediction deviation. That's common in complex algorithm. For second gene, it's obviously only 4 parts of cell need to be focused, and NUC has highest possibility. So it's

obviously signal gene prediction get hard to explain a whole network.

We try to extract the top eigengene from different modules twice, separately from filtered network and unfiltered network. Methods are as follows: for filtered network, we just pick dozens of highest gene significance from our interested module which has been merged. For unfiltered network, follow the same step as we did before just skip step of setting cut off value, then according the position of hub gene to locate the module, as we know the hub gene plays the function of cohesion, filtering will not affect that. At last pick the same number of eigengene to compare. These genes take over most component or highest similarity with other member gene. If the eigengene from the same module are same in two situation, we return back to the original module before merging, check the eigengene separately, input the list of this FASTA version of protein. Then extract the probability according to each cellular part one by one, import to R studio, using traditional statistical methods to summary the difference of prediction. Whether the filtered network perform better.

We conclude that in peroxisome and endoplasmic reticulum, filtered network get significant higher probability value (increase the 20% percentage, here 20% means in these item dominated result, doesn't include what they express in other cellular component dominated, here the criteria of domination is one component over 50%, 2 or more over 30%.). In nucleus, chloroplast, improve a little, make sure less noise included, such as when these two items has higher value the other like endoplasmic will become very low even 0. But in some specific module, it's biological function seem like be related to or limited to one or two molecular process, result can be explain as this protein only exist in this place because the function only happen here. Like hypothetical protein POPTR_0001s10850g, only can be found in extracellular space (because its' probability is 100%). Which means outside the cell, this space is usually taken to be outside the plasma membranes. For proteins, maybe one kind of final production of cell.

For accurate the result, avoid the too absolute situation, we intend to try another method named PSORT. Which is a computer program for the prediction of protein localization sites in cells. analyses the input sequence by applying the stored rules for various sequence features of known protein sorting signals on the basis of protein amino acid composition, It combines several prediction methods and algorithms for the amino-acid sequences which potentially represent localization signals in the cell. At last reports the possibility of the input protein to be localized at each candidate site with additional information. Here is the result of POPTR 0001s00210.

Cytoplasm --- Certainty= 0.450(Affirmative)

Microbody (peroxisome) --- Certainty= 0.196(Affirmative)

Mitochondrial matrix space --- Certainty= 0.100(Affirmative)

Chloroplast thylakoid membrane --- Certainty= 0.100(Affirmative)

Comparatively speaking, the second prediction seem like more reasonable, both main principal component are similar, after overall consideration, Cytoplasm is the most possible cellular component from the gene.

The protein which got the result of extracellular space rechecked in PSORT.

Plasma membrane --- Certainty= 0.460(Affirmative)

Microbody (peroxisome) --- Certainty= 0.150(Affirmative)

Endoplasmic reticulum (membrane) --- Certainty= 0.100(Affirmative)

Endoplasmic reticulum (lumen) --- Certainty= 0.100(Affirmative)

Here amplified the conception of extracellular, also attempt on accurate positioning, the plasma membrane is the exact place where protein involved in molecular function. An improved software not always improve the results, but can enrich the results, provides more details.

For POPTR 0001s13930 this kind of prediction, although have narrow down the area, but still exist several choices, a bit tricky to decide which is expected result, as we know plant cell organism has different cell structure and motion mechanism, when choosing prediction tool, should consider the character, pros and cons of the tool. The most important is the background analysis comes from previous multicomponent stimulated data. From the result of PlantLoc, the location is quite clear but still has space to improvement in terms of accuracy. We try another tool named Plant-mPloc which was evolved

From PlantLoc through a top-down approach improvement. Not only inherit the previous characters also with the ability of prediction is extend to cover from single location to multi location proteins. Predicting subcellular localization of plant proteins including those with multiple sites. It's developed by integrating the gene ontology information, functional domain information, and sequential evolutionary information through three different modes of pseudo amino acid composition. But the predictor is special for plant samples only covers 12 subcellular location: Cell wall, Chloroplast, Cytoplasm, Endoplasmic reticulum, Extracellular, Golgi

apparatus, Mitochondrion, Nucleus, Peroxisome, Plasma membrane, Plastid and Vacuole.(18) Plant-mPloc can be an alternative pool to deal with some problems PlantLoc are incapable.

Plant-mLoc has overcome some disadvantage of Plant-PLoc, which is when you input query sequence, need the FASTA format include accession number to utilize the advantage of GO, but lots of synthetic and hypothetical proteins or newly discovered are not being deposited yet. Even with the accession number, the many proteins can still not be meaningfully formulated in a GO database. At last, Plant-Ploc cannot deal with multiplex proteins that may simultaneously exist. The result for our interesting gene, Chloroplast is approbatory location. (19) But the result of Plant-mLoc is always individual, when we try to do the statistic research, maybe be short of data, caused the judge to the enrichment analysis result between filtered and unfiltered become a bit little hard. So the result is better took as a supplement of former work.

So we can do a small summary, nowadays too many prediction tools are accessible, according to tool's different developed stage and specialized spheres to take appropriate method. As far as possible to try more similar methods, when obtain the large-scale data, utilize adequate analysis method to get most optimal result.

Chapter 4 Gene Singular Enrichment Analysis

According to Huang et al. (20) enrichment tools can be classified into three categories: SEA (Singular Enrichment Analysis, GSEA (Gene Set Enrichment analysis) and MEA (Modular enrichment analysis). SEA is the tool with most traditional strategy, pick up the interesting genes those here means different expressed genes measured by package WGCNA, like higher significance in both pre-filter network module and filtered network module. The algorithm will test the enrichment of each annotation term one by one in a linear mode. For those enriched annotation terms over the enrichment p-value (normal set as 0.05) can be considered as significant enrichment probability. The advantages mentioned above makes SEA a simple and efficient way to extract major biological meaning behind large gene lists. The common of procedure of analysis is after be given a background gene set and a subset of interesting genes. These programs identify which GO terms are most commonly associated with the subset and test the claim that this enrichment is significantly different from what would be expected. Through the way of demonstration significant success in many genomic studies. For software AgriGO is classed as SEA which analysis computes

GO term enrichment in one set of genes by comparing it to another, AgriGO is an integrated gene ontology analysis toolkit for the agriculture community, web-based tool to perform GO-based functional enrichment analysis, supported organisms and gene identifiers were expanded over EasyGo tool, also involved several tools for predicting gene. In AgriGO, the number of supported organisms and identifiers is substantially increased compared with EasyGO. SEA seems more intend to accept a user-selected target list and uses the unbiased adjusted P-value as a single criterion to judge the GO term enrichment.(21) For statistical methods can be selected: Hyper geometric, Fisher's exact and Chi-square tests. The default statistical method is Fisher, if the query size is large and lack intersections with reference, Chi-square would be proper. In Chi-square test, term mapping counts in the query and background (reference) lists are used to form 2X2 contingency table, the difference between observation and expectation for each category is measure to derive a P-value from a Chi-square distribution. The hypergeometric test depend on hypergeometric distribution to calculate the probability of obtaining the contingency table as created above by chance. When the input list has few or no intersections with the reference list, the binomial and chi-square tests are more adequate. (22) Obey this rule we use the hypergeometric model to determine whether any terms annotate a specified list of genes at frequency greater than that would be expected by chance. The hypergeometric distribution defined as follow:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N is the total number of genes in the background distribution, M is the number of genes within that distribution that are annotated to the node of interest, n is the size of the list of genes of interest and k is the number of genes within that list which are annotated to the node. The background distribution by default is all the genes that have annotation. P-values were adjusted for multiple comparison

For comparing the difference of modules with filtered modules. Particularly what is the protein functions' change after network filtering? We adjust the threshold value to create an unfiltered network, but also merge those modules through Dynamic merge, due to too many modules will make experiment over complicated and inefficient. We pick up the most significant genes in our interesting module, those module cluster has higher value in figure 5. Input them to AgriGO start

the analysis.

We pick up 3 most important module (include the merged modules) from the previous research, also add the other modules' several Eigengene (the percentage of these "noise" just take account of 5% percentage of total sample) to make the sample more comprehensive. Because as we know more or less, the module will affect and contribute each other, If we exclude these "noise", perhaps in some of category, unfiltered modules have the advantage of amount, definitely score more in enrichment compare to filtered modules limited to special biological process and molecular functions. The reason we just detect the 3 modules because too many small and insignificant modules will weaken the test, but doesn't mean they are not meaningful, they could be practical and worth studying, just here for make the problem less complicated, just focus on the 3 modules.

The result is surprising, for unfiltered genes, even 31995 Annotated number in query list, only get we 4 significant GO terms (P: metabolic process, C: cell, cell part, intracellular), just can be drew one connection between two significant Go terms. On the contrary filtered genes with 2635 annotated number in query list, return 41 significant GO terms. Most concentrated upon biological process and cellular component. The p-value is dramatic small. Graphical results are posted as below.

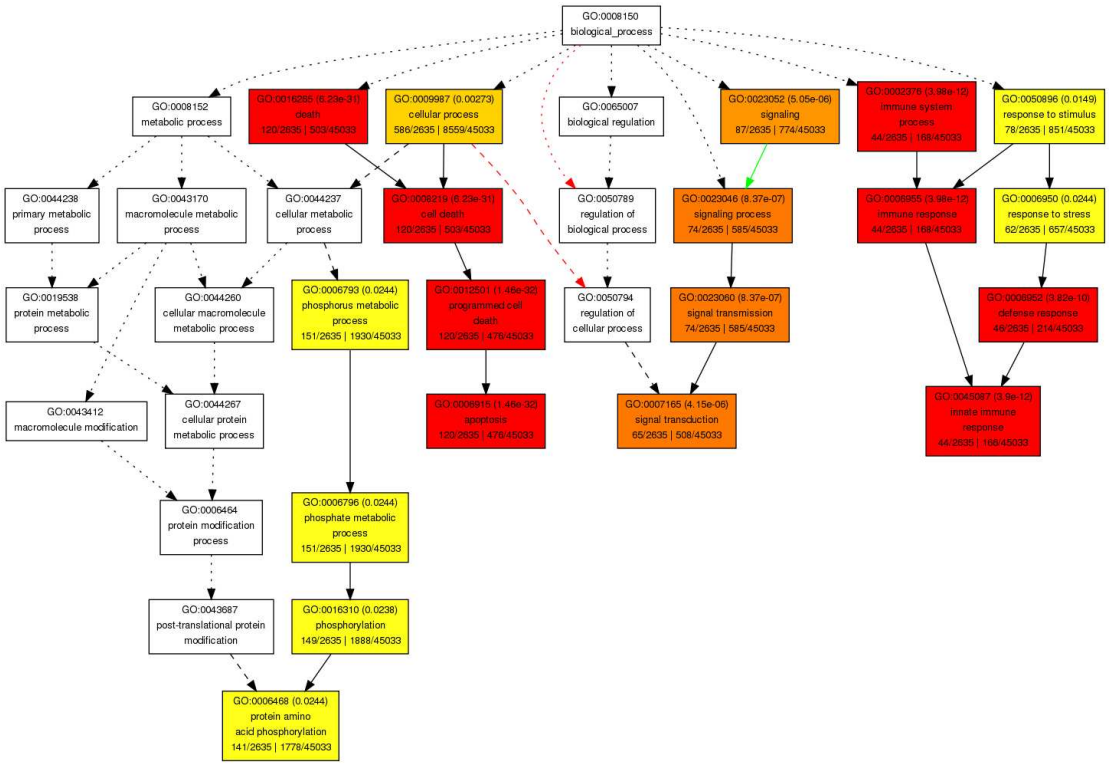


Figure 12. Hierarchical tree graph of overrepresented GO terms in biological process category

generated by SEA. Boxes in the graph represent GO terms labelled by their GO ID, term definition and statistical information. The significant term which adjusted P lower than 0.05 are marked with colour, while non-significant terms are shown as white boxes. The Top GO term is high level which summarize a total process, with the arrow down another block will gradually specialize a specific area or function. The diagram, the degree of colour saturation of a box is positively correlated to the enrichment level of the term, with darker the yellow the enrichment degree higher. Solid, dashed, and dotted lines represent two, one and zero enriched terms at both ends connected by the line, respectively. The rank direction of the graph is set to from top to bottom.

We can find the three clear and distinct line connected enriched GO term. They are signing and signing related, like transmission and transduction. Immune system process include immune response, defence response the last is around death, cell death and programmed cell death, end at apoptosis. That is the modules' main affects and works, and the p-value is different in three main streamline of Go terms, caused that they are separately based on corresponding module. It is undoubtable these modules' elements had been proved that they are enriched. According the place in the cell where these function and process happened, it is not hard to predict and locate the protein or proteome. That why we have mentioned that some tools' principle on predicting cellular components is GO. It's clearly GO terms indicate a process or function, through some kind of formulation, proteins are corresponding to biological process, then find the concrete cell organelles.

On the other hand, in the big section of metabolic process, there are not significant GO terms existed, except the unique influx of phosphorus metabolic process which end at protein amino acid phosphorylation. Maybe the query genes are same as reference genes in terms of expression and effects on basic biological process, means that the gene in the network engaged the basic process and general function, make the normal part of GO terms didn't show the significance. Or the query list just contain a little bit effective gene from other modules mapped to GO terms, but just 5% also means the eigengene fully represent the character of module.

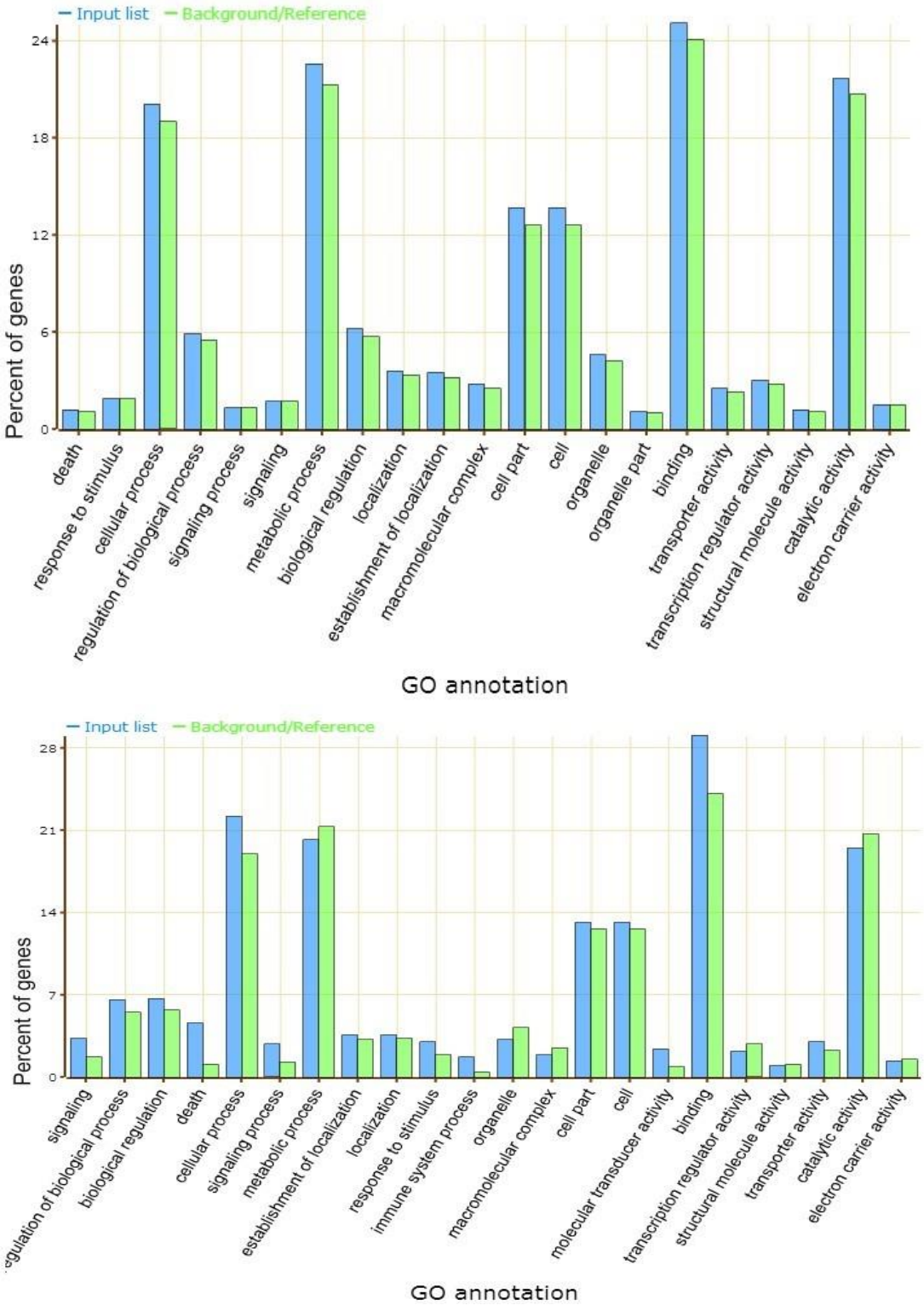


Figure 13: the comparison of Flash bar chart of overrepresented terms in all three categories. Obviously from to left to right, the name appeared in order of biological process, cellular component, and molecular function. And the scale of two graph in Y-axis are different. The Y-axis is the percentage of genes mapped by the term, means the abundance of the GO term. The X-axis is the

definition of GO terms. The percentage of the input list is calculated by the number of genes mapped to the GO term divided by the number of all genes in the input list. The same way of calculation was applied to the reference list to generate its percentage. The first is unfiltered genes from our interesting modules, and second is filtered genes from the same module. We can see in the biological process, filtered network has higher percentage at over 20, and module function of binding the number is over 28 and the other is almost 25, in the other three category, more or less filtered genes has higher number of percentage. In terms of number GO annotations, the filtered group has one more, do look down upon one annotation here, that is the representation of enrichment, and we can conclude the modules are enriched after filtering.

Honestly speaking, the percentage of genes mapped is relative low, special for the filtered genes. As we know the characters of Plant genome genes has much more in gene-families. Also the plant genomes contain large amounts of repetitive DNA, especially Populus trichocarpa's genome size is more than 500 Mbp. In former analysis some genes in module result in quite low genes mapped. Maybe these mapped genes are enriched in some special process, because they are highly similar, that is how we define significance before, means the filtering system will keep these high similarity pairs of genes, but discard the possible functional gene with less connection with main group of genes, result in some modules' genes are overrepresented, some genes are disposed too. We can consider to make an algorithm retrieve most reprehensive or deplete similar expression genes to make enrichment analysis more efficient and reliable.

Chapter Five Prediction on protein-protein interactions from sequence

Proteins are the most essential and participate in almost all process within the cell, execute nearly all of the functions. Meanwhile the Protein-protein interaction (PPI) is a core to predict protein function and ability of molecules, as far as we know, a protein never be isolated to finish a biological process, PPIs can be classified in numerous ways on the basis of their kinds of interaction, in terms of stability, there are obligate and nonobligate, in terms of persistence, they are transient and permanent. So the PPI world is super complicated. (23) When we try to figure out the multitude of protein-protein interaction to integrate the molecular machinery of the cell. A large amounts of

approaches for predicting either physical interactions or functional relationships between proteins have been developed. Protein-protein interaction prediction is a new aspect synthesize both bioinformatics and structural biology. It aims to identify and catalogue physical interaction between pairs or groups of proteins, meanwhile demonstrate full appreciation of cellular processes and networks at the protein level after analysing of all proteins' interactions and functions. The basic characteristics of proteins can suggest a complexity to understand the function of a protein which include critical aspects: Protein sequence and structure is used to discover motifs that predict protein function. Evolutionary history such as phylogenetic profiling, and covered sequences to identify key regulatory residues. Also the expression profile, similar to our data resource profiling of gene microarrays to reveal cell-type specificity and how expression is regulated. The post-translational modification, interaction with other proteins and intracellular localization under the area of consideration as well. (24) When it comes to the types of protein-protein interactions and meaningful result of prediction, the effects of biological effects of protein-protein interactions can be highlighted as follows: Change the kinetic properties of enzymes, perhaps the result of slight changes in binding or allosteric effects. And creating a new binding site, especially for small effector molecules, or change the specificity of a protein through the interaction with different biding partners. In case can be inactivate or destroy a protein. (25) The main idea of considering of prediction is interact of proteins are more likely to co-evolve, it is necessary to make inferences about interactions between pairs of proteins based on their phylogenetic distances. That why lots of method include the reference of various species background and use the phylogenetic tree.

ENTS (Elucidating Network Topology with Sequence) is a random-forest based software which can be used to make predictions of physical interactions between proteins. It has built various species fully covered database, most important past is the predictions has already calculated by ENTS, we need not to take energy on prediction but check the effect of filtering. ENTs uses pairwise combinations of conserved domains and predicted subcellular localization or proteins as input features. According the Organism and Cut-off value, different result can be derived or downloaded. According to the Minimum interaction confidence value, there are several options from 50% to 90%，the interval is 5%. (26)

ENTS is based on random forest approach which is a supervised learning approach, means all the input data will be compare to the existed proteins, depend on the foregoing algorithm to detect every

input protein, emerged from comparison among different species, through the procedure of determinant (such as cut-off value , kinds of organism, GO categories). Finally result in a reasonable prediction.

In the formulation of Calculation of domain pair odds, Aim to get Log-of-odds (LOD), Odds ratio is a measure of association between a hit and an outcome. The OR presents the odds that an outcome will occur given a particular hit. When OR=1 hit does not affect odds of outcome, OR over 1 hit associated with higher odds of outcome, OR lower than 1, hit associated with lower odds of outcome.

$$f(D_x, D_y) = \frac{n(D_x, D_y)}{\sum_{i=1}^{n_p} \sum_{j=1}^{i} n(D_i, D_j)}$$

$$f(D_x) = \frac{n(D_x)}{\sum_i n(D_i)}$$

$$LOD = \log \frac{f(D_x, D_y)}{f(D_x) f(D_y)}$$

Where D means the appearance of domain, larger value of n was accounted indicate more domain was observed. Bracket cover two domain means a pair of domain together also interaction named, because there exist the possibility absence of non-detection and non-existence. If LOD reaches to 0 represent no interacting protein pairs. So we can interpret that the protein involved most cellular process will has large value of LOD, and in unfiltered genes with more not significant interaction, but the observed number does not change. In filtered network, our interesting gene always show more significant and higher value of interaction. The LOD value will slightly higher, but it's just a hypothesis. We'll examine in next step.

We need to face the problem of false positives as well. They are interactions which are not biologically real and are only produced due to tools bias and error. Those false positive results can have dangerous effects. (27) In ENTS set a standard to create testing data which did not include any protein pairs used to calculate the domain pair odds, in the level of protein interactions did not contain overlap to random forest training data. Last calculate sensitivity and specificity values for the other predictor using the same set of testing data to provide single points on the ROC curve. Here sensitivity = TP / TP+FN, specificity= TN / TN+FP, TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of false positives. So according the geometric properties of ROC, it's possible to increase the sensitivity and

specificity to reduce the negative effect of false positives.

In terms of Functional similarity, every time start the measurement from a single present annotation, from that gradually sum up the intersection of two proteins' annotations divided according the amounts of pathways in the union. Like GO mentioned before, GO Semantic similarity also was calculated separately for Molecular function, biological process, and cellular component. The analysis was based on the information inherited from share parents. So the more times of node and any of its descendants occurred in the genome, the term of GO will be annotated more frequently, the similarity score definitely will be the maximum of all pairwise similarity scores. So in the list of protein prediction, the higher minimum interaction confidence value of protein share more similar or even equivalent GO terms, in another way with all confidence levels, we can get the expected result from a significant enrichment of co-expressed genes. It also give us a way to predict protein through GO term's attribution in different cellular component. Like the GO terms based transcripts of genes enrichment analysis, the p-value directly related to the protein annotated with Go terms, lower value represent more possibility this protein involve to either of the three GO categories. Consider the strength of protein interaction, when more proteins are highly annotated or enriched compare the reference background, not only the frequency of the occurrence, but also the togetherness of these protein, we can confirm that the proteins interaction according the GO annotation in some degree.

We have got the result from 3 modules by genes enrichment analysis, one side has only 4 significant terms, other side the filtered one has 41. The P-value for the network GO term is very low in latter compare to unfiltered network even in majority insignificant GO terms, anyhow, it's a viewpoint to predict protein interaction networks, although it can individually indicate how did the protein actually react to another, what is the percentage from one protein successfully predict another, at least we can prove that among the specific modules, the proteins are interacted very strongly, and has large possibility to reach other proteins. In terms of comparison between filtered and unfiltered network, it really works. But for whole genome protein interaction, there are some shortcoming we have discussed in last chapter, so the ENTS based on primary sequence data is more worthy of attempting.

Form the former research, we can get the filtered genes and visualization of network by Cytoscape. Firstly, download the existed prediction from http://ents.as.wvu.edu/index.php with cut-off 80%,

import to Cytoscape for creating a new protein interaction network to achieve the visualization then start to analyse. Through remove those pairs of genes with lower than similarity value 4 by the standard got from the first similarity network. Pay attention, this time the edge here represent confidence instead of similarity value mentioned in previous section. We also use the default layout, and NetworkAnalyzer, but this time need to treat the network as directed.

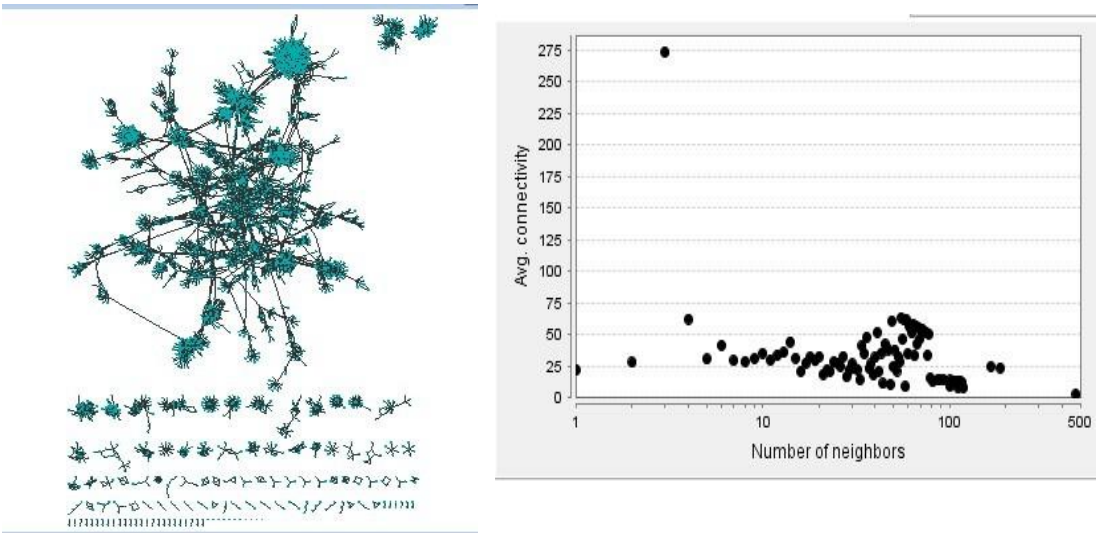


Figure 14: left is the layout of protein prediction network at cut-off value 0.8, every green point means a protein, one edge means prediction connection confidence between proteins. The length of path means the confidence value. Right is Neighbourhood Connectivity statistical data.

From the Average connectivity, the value is quiet similar, no matter how many neighbours the node has, the confidence value is almost same, in the appearance of LOD and Functional similarity, the occurrence of former various parameters like domain, GO terms from parents only have relation with the confidence value, no influence on the numbers of proteins predicted. In other words the formulation only determine the possibility of successful predict, not the possible amount. On the other hand, the lay out demonstrate very dramatic the module we got before, the gather of large amount proteins together, the high relative similarity to each other prove that the module share the same GO Semantic similarity, the interesting gene we got is the Eigengene, meanwhile some small group of proteins distance is quite small, means we made the dynamic merge achieve the integration of similar expressed module.

Clustering coefficient :	**0.090**	Number of nodes :	**3915**
Connected components :	**121**	Network density :	**0.0**
Network diameter :	**10**	Isolated nodes :	**11**
Network radius :	**1**	Number of self-loops :	**379**
Shortest paths :	**199824 (1%)**	Multi-edge node pairs :	**0**
Characteristic path length :	**2.597**	Analysis time (sec) :	**19.562**
Avg. number of neighbors :	**7.656**		

Figure 15: the table of simple parameters of the Analyse Network result

The average of 80% cut off clustering coefficient reached to 0.09 which is global clustering coefficient, comparably it's quite high, almost one tenth of nodes' neighbours are shared. We may reasonable believe these shared nodes are from same module, because in the principle of topological overlap we mentioned before, the structure and changing rules of proteins are important criteria to distinguish genes and modules. Meanwhile we can see the connected components reached 121, inside the network, there are so many independent groups, and we can speculate that because the high confidence value prevent some hub genes from connecting. In the 60 % cut-off situation, the value is lower. Then the average of confidence of unfiltered is 0.8345, the filtered network has just removed 11 genes and not over 100 interactions, but the average of confidence increased to 0.8383, because we pick the most reliable cut-off value to get the network, so even the filtered network didn't prove a lot, but when the cut-off value set to 0.6, before the average of confidence is 0.6775 with 17275 nodes, but almost half will be removed, the average of confidence climbed to 0.7322. Generally speaking with the confidence increasing, the percentage of filtered genes will descend, It means the filtered gene are mainly not significant on all kinds of biological process, when the confidence value gradually climb up, So Filter of network would improve the possibility of successful prediction in large extent. Meanwhile in the parameters of closeness centrality, Neighbourhood Connectivity and betweenness centrality, the Filtered network made a better performance. The effects are closely related to confidence value as well.

Chapter Six Discussion

Gene network analysis are a set of bioinformatics tools designed to manipulate gene sets that function in biological networks. With the mushroom development of biological technologies, a large amounts of toolkit and methods have been developed. It is important to pick the appropriate way to resolve different tasks. Actually during the procedure, how to interpret the principle of all sorts' methods, set up the related parameters and obtain the core information from mass dataset are

essential job.

Although we have concluded that the filtered network performed better in many respects, but there are some shortcomings in the process. First when coming to the huge data, because the limitation of the computer memory, in some R functions, we have to reduce the volume of input data. Result in data reduction, or data over simplified. Like the module detection, we try to combine the result from first part to whole second part data, it means in the "filtered" genes, some of them would have higher similarity exceed the criteria but was removed because they are unqualified in first part. Make the modules imperfect. Try to look on the bright side, the introduce of eigengene and hub gene is very convenient to swift the object from gene to module, because we utilize the character of network, in the network, there must exist some genes are able to highly connect a large group genes, as a representative sign. Secondly in the graph of merged modules, some of the same colour bar are separated, means the location of modules are a bit disordered, In some part of space, the density of different colour exist, represent it is not successful outcome of modules, the reason is related to the property of topological overlap matrix, consider the shared neighbours' score, sometimes the value of similarity and number of similar connection are not accordant, make it tricky to classify genes. On the other hand, new improved algorithm have come up to handle the dilemma, but regretful, the WGCNA package didn't update that. Meanwhile the numbers of module is a barrier for further compute, even we can focus on several high significant modules, but except the first merged module cover majority of one subset of genes, the other interesting modules are just minority of whole network, for specific biological function or process research, it's not a big deal, but for consideration of overall situation, still left some deficiency. Thirdly, in the subcellular localization, the predict tool can't treat the inquiry as multi-task, even you input a file contains a lot, the tool still calculate one by one, we are force to choose the most representative genes as input , then collect the outcome for further statistical. Actually the result did lack of convincing but whole procedure is inefficient and time-consuming. Sometimes one or two unexpected result need to improve or test through other tools, add the pressure on calculating. Then in Enrichment Analysis part, we run everything smoothly, except the too big difference between the comparisons, maybe owing to the coincidence, the mapped genes are too special and enriched in the filtered network, even the gap of percentage of mapped genes are not big. At last in the PPI prediction, I have to recognize use the visualization software is taking the easy way, did not consume too much time on

prediction itself, the breakthrough points are all we discussed before, like the definition of parameters: clustering coefficient. Exclude that many other kinds of prediction tool are available, but some of them can not fit the species, or too complicated system make us come back to what we have researched. After all PPI related research area is very new and original.

Return to the subject of filtering network, for the sake of accurate comparison, we try to do our best using number and graph to explain the result. Actually like I repeated mentioned, it is very difficult to compare as a whole network, we change the strategy to conquer the specific modules, more or less, and the result is not out of our expectation. GO annotation coverage is critical tool run through whole procedure, GO annotations are mainly generated by computational prediction. They flexible cover the huge network, offer a great space to assemble the similar gene from different modules, label various gene with three kinds of classification: biological processes which are made up of chemical reactions or other events that results in a transformation. Directly related to gene expression and protein modification. Cellular component refers to cells the structural and functional units of life and living organisms. The specific location of cell clarify the whole biological stream. Molecular function directly perceived through the senses that what the genes' function is. But rely on GO terms may lead to two issues: reduced quality of annotation means poor quality annotation will affect the GO distribution, can generate biased or misleading analysis results. Another is low annotation coverage, because the delayed development of GO system, cannot explain the new coming biological meaningful information in time. Result in leaving out some genes, make the result biased.

At last, depend on language R makes the statistical work well done, but when deal with mass data, the shortcomings were completely exposed. We'd better to divide the work into appropriate Programming language.

Reference

(1)Yong, Ed. (2012) "Tree's leaves genetically different from its roots". *Nature News*. Retrieved 2012-08-14.

(2) Hertzberg, M. (2001) “A transcriptional roadmap to wood formation”. Proc Natl Acad Sci U S A. 2001 Dec 4; 98(25):14732-7. Epub 2001 Nov 27.

(3) Alexandre Blais, Brian David Dynlacht. (2005) “Constructing transcriptional regulatory network”. Genes & Dev. 2005.19: 1499-1511.

(4) Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS.(2007) “Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles”. PLoS Biol. 2007 Jan; 5(1):e8.

(5) Shannon P, Markiel A, Ozier O. (2003) “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. Genome Res. 2003 Nov; 13(11): 2498-504

(6) David Tritchler, Elena Parkhomenko, Joseph Beyene. (2009) “Filtering Genes for Cluster and Network Analysis”. BMC Bioinformatics 2009, 10: 193 doi: 10

.1186/1471-2105-10-193

(7) Andy M Yip, Steve Horvath. (2007) “Gene network interconnectedness and the generalized topological overlap measure”. BMC Bioinformatics 2007, 8:22 doi: 10.1186/1471-2105-8-22

(8) Bin Zhang, Steve Horvath. (2005) “A General Framework for Weighted Gene co-expression Netwrok Analysis”. Statistical Applications in Genetics and Molecular Biology. Volume 4, Issue 1, ISSN (Online) 1544-6115, DOI: 10.2202/1544-6115.1128, August 2005

(9) Jun Dong, Steve Horvath. (2007) “Understanding network concepts in modules”. BMC systems biology 2007, 1:24

(10) Peter Langfelder, Steve Horvath. (2007) “Eigengene networks for studying the relationships between co-expression modules”. BMC Systems Biology. 2007; 1: 54. doi: 10.1186/1752-0509-1-54

(11) Peter Langfelder, Bin Zhang, Steve Horvath. (2007) “Dynamic Tree Cut: in-depth description, tests and applications”. November 22, 2007.

(12) Yu Keng Shih, Srinivasan Parthasarathy. (2012) “Identifying functional modules in interaction networks through overlapping Markov clustering”.Bioinformatics (2012) 28 (18):i473-i479.

(13) Jeremy J Jay, John D Eblen, Yun Zhang, ED. (2012) “A systematic comparison of genome-scale clustering algorithms”. *BMC Bioinformatics* 2012, **13**(Suppl 10):S7

(14) Zhou Du, Xin Zhou, Yi Ling, Zhenhai Zhang, Zhen Su. (2010) “agriGO: a GO analysis toolkit for the agricultural community”.Nucleic Acids Research Advance Access published on July 1, 2010, DOI 10.1093/nar/gkq310. Nucl. Acids Res. 38: W64-W70.

(15) Elena Nabieva, Mona Singh. (2009) “Prediction of Protein Structures, Functions, and Interactions”. John Wiley & Sons Ltd. 2009, (P 223)

(16) Jose B Pereira, Anto J Enright, Christos A Ouzounis. (2004) “Detection of Functional Modules from Protein Interaction Networks”. PROTEINS: Structure, Function, and Bioinformatics 54:49 – 57

(17) Shengnan Tang. Et al. (2013) “PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif”. Nucleic Acids Research, 2013, Vol. 41, web server issue W441-W447.

(18) Chou KC, Shen HB. (2010) “Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization”. Plos ONE; 2010, Vol 5 Issue 6, p1

(19) Chou KC, Shen HB (2007) “Large-scale plant protein subcellular location prediction”. Journal of Cellular Biochemistry 100: 665–678.

(20) Huang DW, et al. (2009) “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists”. Nucleic Acids Res. 2009;37:1-13.

(21) Michael Ashburner. Et al. (2000) “Gene Ontology: tool for the unification of biology”. Nat Genet 2000 May; 25(1):25-9

(22) Zhou X, Su Z. (2007) “EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species”. BMC Genomics 2007; 8:246.

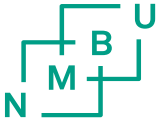
(23) I M A. Nooren, J M Thornton. (2003) “Diversity of protein-protein interactions,” The EMBO Journal, vol. 22, no. 14, pp. 3486–3492, 2003.

(24) Golemis E. (2002)" Protein-protein interactions: A molecular cloning manual". Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. Ix, 682 p. p.

(25) Phizicky E. M. and Fields S. (1995) "Protein-protein interactions: Methods for detection and analysis". Microbiol Rev. 59, 94-123.

(26) Rodgers-Melnick E, Culp M, DiFazio S (2013). "Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS".*BMC Genomics* 2013, 14:608

(27)Javad Zahin, Omid yaghoubi. (2013) "Protein–protein interaction prediction from PSSM based evolutionary information". Volume 102, Issue 4, October 2013, Pages 237-242.



Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no