PROTEOME AND GENE EXPRESSION ANALYSIS:

Gene expression begins when genes are transcribed into messenger RNAs (mRNAs), which are then translated to produce proteins. Total gene expression in cultured cells or a tissue sample can be detected in two main ways. One is the detection and quantification of total messenger RNA—the tran- scriptome—by DNA microarray technology. The other detects the total protein composition of the sample—the proteome—by separating the protein products of the genes by two-dimensional gel electrophoresis or chromatography, followed by identification of their constituent peptides by mass spectrometry. In both these cases a single experiment produces enormous amounts of raw data, and new tech- niques had to be devised for data collection, storage, and analysis. The transcrip- tome and proteome, unlike the genome, are changeable in response to conditions, and depend on the state of development, the environment, and the type of tissue.

Monitoring the simultaneous expression of multiple genes provides information that cannot be obtained by monitoring the expression of one or a few genes at a time. By revealing which genes are expressed together, or co-expressed, for example, these techniques can identify genes that may be functionally related, such as the various members of a multiprotein complex or a metabolic or signaling pathway. This infor- mation can be used to help assign possible functions to unidentified genes with the same expression patterns. Co-expression can also help indicate which genes are under the control of the same regulatory system.

Large scale gene expression:

High-throughput, whole-genome DNA microarrays have become a very useful tool in biological research. However, the interpretation of the large amount of data produced by a microarray experiment or a series of experiments can be time consuming. It is also difficult because different methods can yield alternative conclu- sions. The aim of these experiments is usually to extract some biological or functional meaning from the lists of genes, either by identifying critical genes that might be responsible for a biological effect, or by finding patterns within the genes that point to an underlying biological process and annotating each one of the genes.

Serial analysis of gene expression (SAGE) is also used to study global patterns of gene expression:

serial analysis of gene expression (SAGE). It has both an exper- imental and a bioinformatics component. It is based on the following observations:

Diagram

Description automatically generated

first, that a short sequence (a tag) contains enough information to uniquely identify a gene (provided that the tag is obtained from a unique position within each gene); second, that the sequence tags from the total cellular RNA (converted into cDNA) can be linked together to form long DNA molecules, called concatemers. This DNA sequence is read and counted. The total number of times a particular tag is observed in the concatemers approximates the expression level of the corresponding gene (see Figure 15.3). The data produced by SAGE include a list of the tags with their corresponding counts, providing a digital output of cellular gene expression that can easily be analyzed further. The SAGE analysis programs SAGEmap and xProfiler are available on the NCBI website and allow the user to specify which organ is to be investigated. Libraries consisting of gene lists organized by the various types of tissues or cell lines are provided for further choice. The expression associated with these gene lists can be divided into two groups and compared with each other. The output from SAGE provides the SAGE tag, the UniGene ID (Identification number), the gene description, and color- and letter-coded differences in expression levels.

Digital differential display uses bioinformatics and statistics to detect differential gene expression in different tissues:

Another alternative to microarrays for looking at differential gene expression in some circumstances is purely computational. Digital differential display (DDD) is a method for comparing EST-based expression profiles in different tissues or condi- tions from various libraries or between pools of EST libraries. An EST library contains short sequences cloned from the total cellular mRNA (converted to cDNA) of a particular tissue or particular condition. The theory is that genes expressed at a high level will be represented by more ESTs than those expressed at a lower level. Genes whose expression levels differ significantly from one set of EST libraries to the next are identified using a statistical test.

The NCBI’s UniGene database forms the core of the DDD method. In UniGene, all the human EST sequences in the databases have been put into distinct clusters, where each cluster represents a single gene. The DDD methods then compare the number of sequences from each EST library assigned to a particular UniGene cluster, and identifies those differences between the clusters that are likely to be biologically significant. The user can select EST libraries from a list on the DDD Web page and may combine selected libraries into specific pools. Figure 15.4 shows a DDD analysis of two selected pools. Each of the three columns on the left repre- sents a particular pool, and the rows represent UniGene clusters. On the right is the gene description, which gives the name of the cluster, and the UniGene ID number for that cluster. Clicking on this ID provides a summary report for that cluster.

The simplest method of analyzing gene expression microarray data is hierarchical cluster analysis:

The main aim of gene expression analysis is the identification of common patterns of gene expression; for example, which genes are being co-expressed, and which genes have been downregulated or upregulated in one sample compared to the other. Hierarchical clustering is the most widely used method for analyzing patterns of gene expression in microarray data.

The results of a hierarchical cluster analysis of gene expression microarray data are typically displayed as a dendrogram with a color-coded grid. The grid of colored squares contains rows that represent a gene for which expression was detected in one or other of the samples, and the columns represent the different samples or conditions (see Figure 15.5). The precise shade of each square represents the fold increase or decrease of the expression level of the gene in that sample in relation to some reference level, often the median level of expression of the gene over all the samples. Typically, higher expression is shown by red, lower expression by green or blue, and expression similar to the reference as black. In the display, it is easy to see how the rows of genes are arranged into distinct blocks composed of genes with similar expression patterns in the same tissue. In the experiment depicted in.

Other more sophisticated clustering methods have been applied to the analysis of gene-expression data, of which k-means clustering and techniques such as self- organizing maps (SOMs) and self-organizing tree algorithms (SOTA) are just a few.

Facilitating the integration of data from different places and experiments:

In general it is difficult to compare and integrate data from different laboratories and experiments. Therefore a group of scientists have set up a consortium (the MGED society) to standardize the output and annotation of microarray data, which will facilitate sharing the data and creating a consolidated database. The set of stan- dardization rules is called MIAME, which stands for Minimum Information About a Microarray Experiment. This includes information that is essential for someone else to interpret the results of the experiment and even to reproduce the experi- ment. The journal Nature and other Nature research journals will in general only accept articles dealing with microarray data that comply with MIAME. The public repositories ArrayExpress at the EBI and GEO at NCBI have been set up to store and distribute MIAME-compliant microarray data. In addition a MicroArray Quality Control (MAQC) project is under way to assess the quality of DNA microarray data. This project has recently concluded that with careful experimental design, data transformation, and analysis, microarray data can be compared between different formats and laboratories.

Clustered gene expression data can be used as a tool for further research:

Clustered data on gene expression patterns obtained from either gene expression microarrays or genome bioinformatics can be used as a predictive tool to identify new transcription factors or other cell-regulatory proteins. Regulatory elements are identified using both gene expression patterns and the clustering of genes according to function, based on functional annotations obtained experimentally or from sequence homology. The clustered genes (or proteins) can be analyzed with respect to protein–protein interaction data to see if the genes can form a function- ally related pathway. For example, Figure 15.10 illustrates how a cluster identified by SOM clustering (node 0, see list in Figure 15.8B) has been subjected to the pSTIING database to obtain an interaction map. We can see that many of the genes submitted (in red) are connected either directly or through another interaction partner. Therefore it is likely these genes are part of a specific functional pathway. Some of the genes form individual clusters. They may be connected via a gene product (protein) that has not been identified as significantly changed, or may not be on the actual chip. The map can be extended to see if other connections can be found (see Figure 15.10B). In such ways a more complete pathway or interaction map can be generated.

A vast collection of data from many gene expression and protein expression exper- iments is now freely available on the Web and can be mined for biological reanalysis or used as reference data for the development of new bioinformatics tools. For example, the L2L tool is a repository of microarray data that users can search with their own up- or downregulated gene list to see if anyone else has got a similar gene expression pattern.

15.2 Analysis of Large-scale Protein Expression:

The proteome refers to all the proteins that make up an organism or, on a smaller scale, the total number of proteins found in a particular cell type at a specific point in time and under specific conditions. An organism will have different protein expression in various parts of its body. The protein expression will also differ between the separate stages of an organism’s life cycle and under different environmental conditions. To understand how an organism or a cell functions both under normal and abnormal (such as disease) conditions it is important to know how protein expression is affected.

Two-dimensional gel electrophoresis is a method for separating the individual proteins in a cell:

Proteomics experiments that aim to characterize the proteome of a cell type or tissue at a particular point in time commonly use the technique of 2D gel elec- trophoresis, either on its own or followed by mass spectrometry (MS) of the sepa- rated proteins in order to identify them. Two-dimensional gel electrophoresis separates proteins in two dimensions according to two independent properties. In the first step, isoelectric focusing (IEF) of the protein mixture separates proteins on the basis of their isoelectric points (or pI/pH). Electrophoresis [in the presence of the detergent sodium dodecyl sulfate (SDS)] of the separated proteins in a direction along the gel at right angles to the IEF then resolves each spot into its constituent proteins on the basis of their molecular weights (see Figure 15.12). A 2D gel thus provides a kind of map (with pH on one axis and molecular weight on the other) on which, in theory, any given protein will always occur as a spot at the same location, whatever the sample being analyzed. Once the gel has been run and the proteins separated, it is stained and scanned into an image-analysis program to detect and identify the spots.

The problem with this technique is that on even the best gel only about 3000 spots are visible. There are many more different proteins than that in a cell: some esti- mates are as high as 100,000. The limitations of 2D gels are due to a number of factors, one being that the gel is not large enough to separate proteins that have very similar molecular weights or pH range. In addition, membrane proteins often become insoluble during the first-dimension run and low-abundance proteins are often not detectable.

Clustering methods are used to identify protein spots with similar expression patterns:

As with gene expression data, clustering is a useful method of grouping similar gels or spots together and extracting protein expression patterns that can indicate biological differences or similarities between samples. Many of the clustering methods used for microarray data can be applied to 2D gel data. Figure 15.16 shows an unrooted hierarchical clustering tree constructed from 2D gel data from Swiss 3T3 cells that were stimulated with different growth factors at different times. Protein synthesis after the various stimulation regimes, or with no stimula- tion, was measured and compared using 2D gels and then analyzed by hierarchical clustering. In this instance, the gels clearly fall into two groups (see Figure 15.16). One group clusters on the same main branch as the unstimulated sample, suggesting that these treatments had little effect on protein expression. The other branch clusters together those samples that were stimulated for a longer time and those that were stimulated by more than one growth factor and platelet-derived growth factor (PDGF).