Experimental techniques for studying network nodes and edges

Introduction

Under the network paradigm of systems biology, the behavior of a biological system is defined as much by the molecules in the cell (i.e. the nodes of the system's graph) as by the physical and functional interactions between these molecules (i.e. the edges of the graph). This document contains a short description of some of the key experimental techniques for studying these nodes and interactions. One class of these techniques, namely those based on mass-spectroscopy, will be discussed in depth later on in the course.

Synthetic lethality screens

You have already heard about synthetic lethality screens in the self-study material for challenge 1. Important to note is that synthetic lethality is the only technique listed here, that is geared towards identifying functional interactions between network nodes that do not involve physical interactions. The other techniques you will encounter here are geared primarily towards determining the abundance of individual network nodes (e.g. protein or RNA molecules) or defining the physical interaction between these molecules.

Yeast-two-hybrid

The yeast-two-hybrid technique (Figure 1) is an experimental molecular-biology method that serves to detect the ability of two proteins to bind to one another. The technique is based on the activity of a naturally occurring activator protein from yeast (top row of figure 1). This activator protein (Gal4) consists of two protein domains, the DNA binding domain, which binds to a specific DNA sequence (UAS), and the activating domain, which activates the transcription of a nearby gene (the reporter gene). In the natural Gal4 protein a flexible linker links these two domains.

It is important to know that Gal4-mediated activation of the reporter gene is based purely on the fact that the binding domain of Gal4 brings the activating domain of Gal4 somewhere close to the reporter gene's promoter. The exact length and orientation of the linker does not matter.

In a yeast-two-hybrid experiment the gene coding for the natural Gal4 protein is replaced by two hybrid genes. In the hybrid gene the coding sequence for Gal4's DNA binding domain is fused to the gene of the protein, for which we would like to find binding partners. This protein is called the "bait". We use the "bait" to "fish" for potential binding partners.

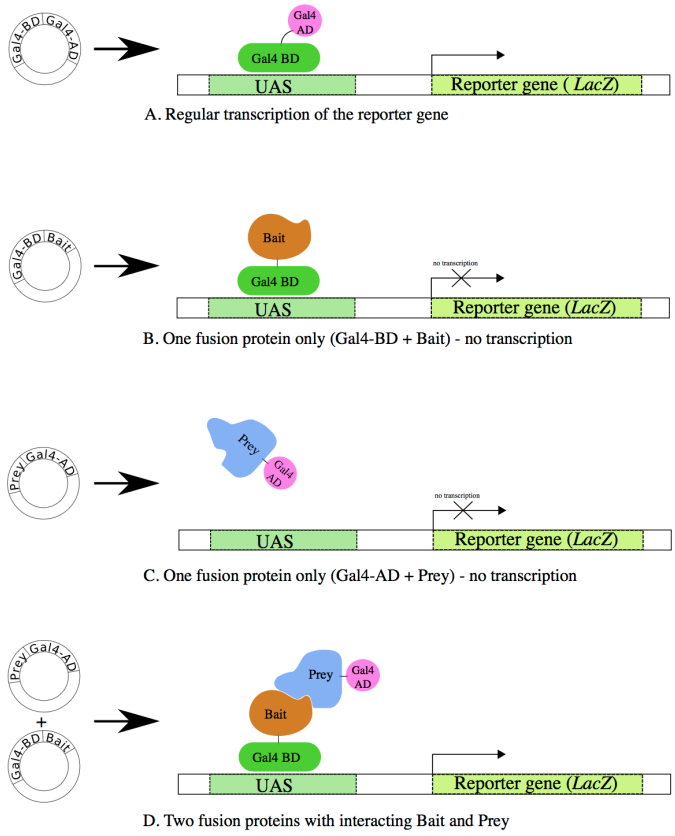

Figure 1 *Yeast two hybrid experiment. The first row shows the natural Gal4 gene (left) and the function of the resulting Gal4 protein (right). Binding of Glal4's DNA-binding domain (BD) to a specific DNA sequence (UAS) places its activator domain (AD) in the vicinity of the reporter gene's promoter. Thereby, the transcription of the reporter gene is stimulated. In a yeast-two-hybrid experiment the natural Gal4 protein gene is replaced by two genetically engineered fusion-genes. The first comprises the BD coding region artificially fused to the coding region of a protein, called the "bait". The second includes the AD coding region, fused to the coding region of another protein called "prey". Note, cells that only contain the bait or the prey fusion gene (middle two rows) show no activation of the reporter gene. Only when a pair of bait and prey fusion genes are expressed in the same cell and the bait and prey portions bind to one another will the reporter gene be activated.*

In the second hybrid gene the coding sequence for a potential binding partner (called the "prey") is fused to the coding sequence of Gal4's activating domain. These two hybrid genes are introduced into a yeast cell (e.g. by transfection with a plasmid) where they are expressed.

Subsequently the bait and prey portions of the two resulting hybrid proteins bind to one another. This binding event recruits the Gal4 activating domain to the vicinity of the reporter gene's promoter and thus stimulates transcription of the reporter gene.

Reporter genes typically employed in yeast-two-hybrid experiments, are genes that generate an easily detectable signal when activated. For example, the reporter gene could code for an enzyme that generates a colored metabolite. The colored cells then provide a readily detectable proof of the molecular binding event between the bait and the prey having occurred. Cells in which prey and bait portions of the fusion genes do not bind to one another will remain colorless.

The real power of the yeast-two-hybrid method lies in its ability to test a large number of potential interactions in a single experiment. For this a cell already containing a plasmid that carries the "bait" fusion protein is transformed with a large library of plasmids containing the genes for potential "prey"-proteins fused to a Gal4 activating domain. These cells are then grown into colonies of which those that are colored can easily be identified. By sequencing the prey-fusion gene plasmids of these cells, the "prey" protein that was able to bind to the "bait" can be identified.

The yeast-two-hybrid technique has two major limitations of which the first is conceptual, the second technical. The conceptual limitation is that the yeast-two-hybrid technique only investigates pair-wise interactions. Interactions that require three or more partners to be present at the same time cannot be detected. Moreover, indirect interactions (e.g. in large multi-protein complexes) have to be inferred through multiple pairwise interactions. The technical limitations arise from the fact that the binding events we would like to observe have to take place inside the complex environment of a yeast cell (or rather in the yeast cell's nucleus where transcription takes place). Certain fusion proteins may not be expressed or may not fold properly in yeast cells or they may not enter into the nucleus. Interaction between bait and prey protein may also require a posttranslational modification that is not done in yeast etc. etc.

Despite these drawbacks, yeast-two-hybrid experiments have been extremely successful and are very widely used. Their success has spawned a great variety of related experimental technique. For example, the use of reporter genes that allow the survival of the cell, when they have been activated, has become popular. This permits the isolation of cells expressing pairs of interacting prey and bait proteins via selection, which in turn facilitates the analysis of very large libraries of potential interaction partners. Further Variations of the basic yeast-two-hybrid method have been developed to identify protein-DNA and protein-small molecule interactions.

Affinity Purification Mass Spectrometry (AP-MS)

Proteins, which interact functionally, often bind to each other to form stable macromolecular complexes, too. Under appropriate conditions these complexes remain intact even after the cell containing them has been destroyed and the cell content released. In order to

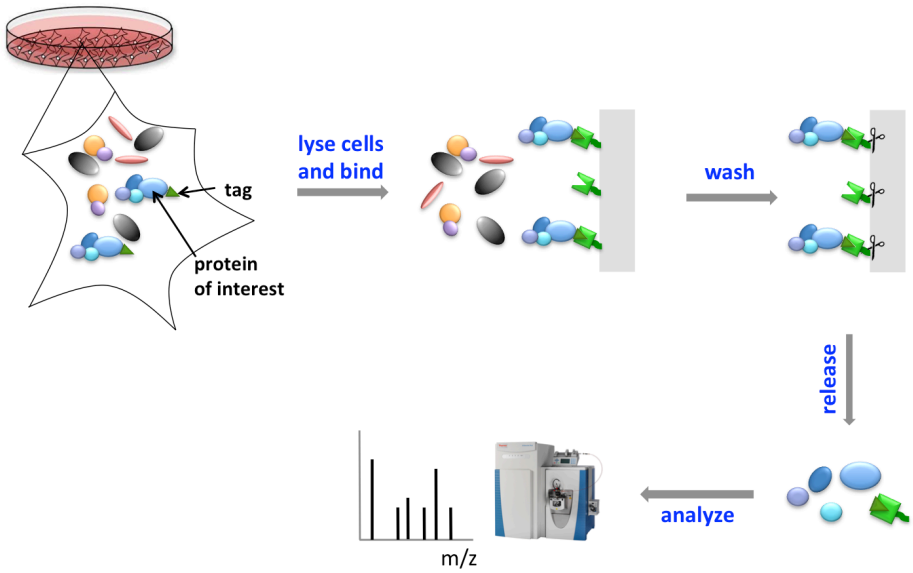

Figure 2 *Schematic representation of an affinity purification mass spectrometry (AP-MS) experiment. Cells are disrupted and the protein of interest, together with the other proteins bound to it, is immobilized on a solid support. Other cellular components are washed away, the purified complex is released and its components are identified by MS.*

identify the proteins, which form complexes with our protein of interest, we can use our protein as a "handle" to fish the entire complex out of the mixture of other cellular components. We therefore immobilize a highly selective binding partner (often an antibody) for our protein of interest by attaching it to a solid support. Binding between our protein of interest and the binding partner will immobilize the entire complex while other cellular components are washed away. Finally, the isolated complex is released and its components can be analyzed by Mass Spectrometry (**Error! Reference source not found.**).

While conceptually simple, this affinity purification mass spectrometry approach presents substantial technical challenges. First, antibodies that are suitable for trapping the protein of interest to a solid support are not available for all proteins. One therefore often has to work with a tagged version of the protein. Tagging involves fusing the gene for the protein of interest to the coding sequence for a short peptide or protein for which very specific, high-affinity antibodies are available. However, it is experimentally very difficult to introduce the gene for the tagged protein into the organism. It is easier to introduce the gene into a cell line that can be grown in tissue culture although this bears the risk, that the conditions in the chosen cell line are not sufficiently similar to those in the natural tissue and that the composition of the complexes may change.

Western blots

The traditional way in which the abundance of proteins in a tissue sample is measured, is via western blot (Figure 3): proteins are isolated from a sample, separated via gel electrophoresis and then transferred to and immobilized on a nitrocellulose membrane. The membrane is incubated with a solution containing an antibody, binding specifically to the protein of interest. Unbound antibodies are then washed away. Traditionally, these antibodies were radioactively labeled so that a photographic film placed onto the membrane would show a blackened band thus indicating the spot in which our protein is located on the membrane. By running two samples side-by-side on the same electrophoresis gel and by comparing the relative intensity of the resulting bands on the photographic film, the two samples can be compared in regard to the abundance of our protein of interest.

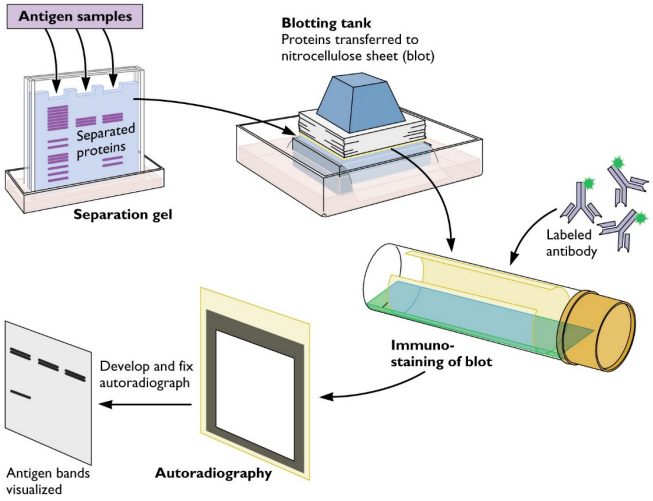

Figure 3 *Schematic representation of a western Blot experiment. Protein samples are separated by gel electrophoresis and transferred to a membrane. Labeled antibodies against the protein of interest are incubated with the membrane and thus reveal the presence and abundance of this specific protein on an autoradiograph. (image source www.virology.ws)*

Although the technique detecting antibodies has evolved to circumvent the use of radioactive labels, the basic principle of western blot has remained unchanged.

The main advantages of western blots are that they can be performed with very simple experimental equipment available in most biology labs and that data analysis is very straightforward. The principle disadvantage of western blots is that they are difficult to perform on more than a dozen samples or on two or three proteins per sample simultaneously. Western blots also require a specific antibody against each of the proteins that are to be analyzed. Generating these antibodies is a rather labor-intensive process and generating antibodies on a proteome-wide level is very difficult.

Gene expression profiling: Expression Microarrays & RNA Seq

Gene expression profiling is the experimental measurement of the relative abundance of RNA molecules generated by the transcription of an organism's genes. This technique will not figure prominently in the course's discussions around challenge 2, but this type of data plays an important role in a very large number of systems biology studies. In fact, for most studies aiming at the identification of functional networks, expression profiling is *the* central source of experimental data. Often it is even the *only* source. It is therefore important to give a brief outline of this technique.

In the lab, gene expression profiling is currently performed with the help of two experimental techniques: expression microarrays and RNA sequencing.

When the company Affymetrix launched their first gene expression microarrays in the late 1990 they caused a sensation. It suddenly became possible to measure the relative abundance of all the thousands of RNA's produced in an organism in a single experiment. Up to that day, such gene expression measurements had to be

performed one gene at a time employing a method called Northern Blot.

Gene expression microarrays were arguably the first true "omics" technology i.e. a technique that allows to measure the full complement of a certain class of molecules in a biological sample.

Gene-expression microarrays operate with short (25-50 bases long) single-stranded DNA molecules (often called oligos), which are synthesized on silicon surfaces using production techniques borrowed from semiconductor manufacturing. These surfaces are divided into thousands of small squares and of which each contains another type of oligo (0). Each of these oligos is complementary to a particular RNA sequence. When RNA samples, or more precisely cDNA copies of these RNAs, are applied to the chip, they will selectively bind to the complementary surface-bound oligos and will thus be very selectively enriched in the area of the chip where their complementary oligo is attached. Because we know which oligo sequence is attached to which square, the location of binding reveals the sequence of the bound cDNA.

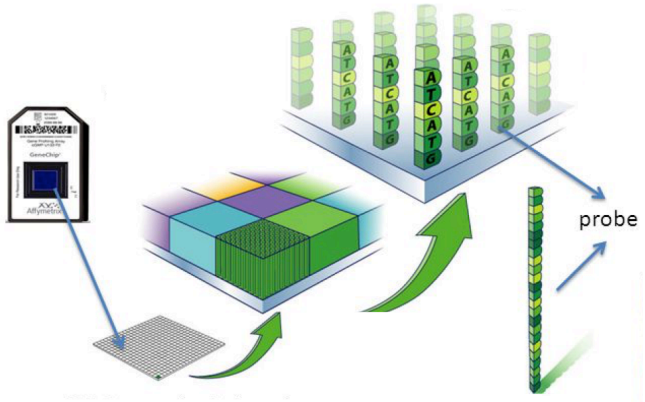

Figure 4 *Zoomed in views of an Affymetrix gene expression microarray. The chips' silicon surface is divided into many thousands of small squares. To each of the squares a different type (represented by different color in the middle panel) of single-stranded DNA oligo is attached, which is complementary to the RNA of a particular gene. The actual length of these probes is 25 bp. (image source Affymetrix website)*

A typical gene expression profiling experiment involves the following steps (Figure 5). Cells are collected for different two samples, of which the gene expression patterns are to be compared. These cells are disrupted in order to release their RNAs. Because RNAases are omnipresent in the cells and degrade RNA rapidly (RNAases are RNA degrading enzymes), the RNA samples are first reverse-translated into DNA molecules (so-called cDNAs). These cDNAs are isolated and then labeled with a fluorescent marker. For example, the cDNAs from sample A are labeled with a red fluorescent marker and those from sample B with a green fluorescent marker. The two cDNA samples are then combined and applied to the microarray.

The conditions chosen to reign on the microarray (salt concentration, temperature etc.) are such that the cDNAs will bind specifically to those surface-attached oligos that are complementary to the particular cDNA's sequence. This is to say, that they will bind to a specific square on the chip. All non-bound cDNA's are washed away. The intensity of the fluorescence for the two colors are measured in each of the squares. The relative intensity of the two fluorescence signals in a particular square then gives information about the relative abundance of the corresponding RNA in samples A and B.

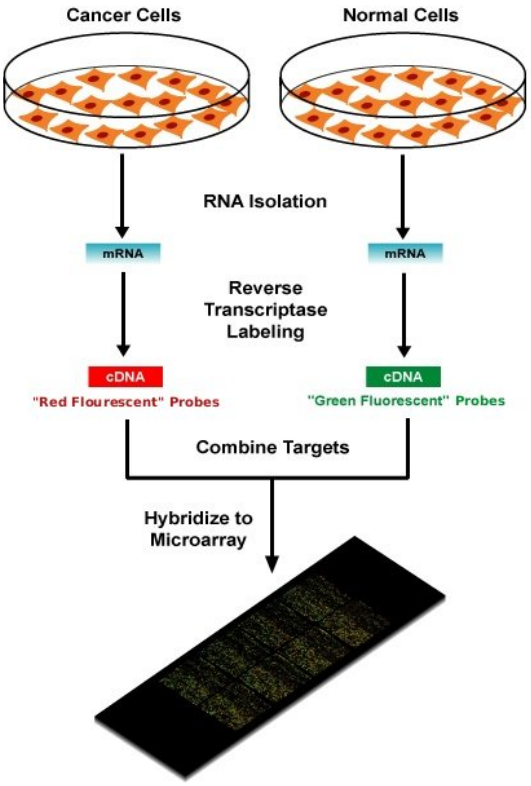

Figure 5 *Schematic workflow of a gene expression microarray experiment designed to determine the relative abundance of gene transcripts in two biological samples (cancer cells and healthy cells). RNA is extracted from the sample and converted to cDNA. Fluorescent labels are attached (red probes for the cancer and green probes for the normal sample). The labeled cDNAs are applied to a microarray surface that is covered with oligonucleotides. Each "dot" on the microarray contains oligonucleotides that are complementary to cDNAs from one specific gene. Non-specifically bound cDNAs are washed away. The relative intensity of the red and green fluorescent signal is measured and reveals for each gene the relative level of expression between cancer and normal cells. (image source: Wikipedia)*

This kind of data allows us to see if and for which genes translation has been up- or downregulated in tissue A vs. tissue B.

Though they were revolutionary just a little over 15 years ago, gene expression microarray experiments are now already about to be replaced by a new technology called RNA-Seq.

For RNA-Seq experiments the RNAs are also extracted from the cells, but before being transcribed into cDNAs they are subjected to a selection/depletion procedure in order to enrich RNAs of interest. Ribosomal RNAs for example, which are so abundant that they would swamp the subsequent analysis steps, are often depleted by binding them to beads containing complementary single-stranded DNA molecules.

Meanwhile, protein-coding RNAs can be enriched thanks to their characteristic poly-A tails.

The resulting RNA samples are then reverse transcribed into cDNA's and sheared into fragments a few hundred base pairs long. The resulting pools of cDNAs are then sequenced using random primers and next-generation sequencing techniques. These experiments generate millions and millions of short stretches of sequence, each approx. 100 bp long. These are known as reads and map onto the genome sequence of the organism being studied.

The simplest analysis of these data simply involves counting the number of reads that mapped to a particular gene. This number is corrected for the length of the corresponding cDNA and then serves as an indicator for the relative abundance of the corresponding RNA. However, this analysis only scratches the surface. The read data can also be used to identify new, previously unknown, RNAs. Furthermore, is it possible to detect genetic variations in the transcribed genes or alternative splice variants of genes.

For those who want to plunge deeper into the topic, a 2-hour lecture by David Gifford (MIT) on the biostatistical methods behind the analysis of RNA Seq data can be found under

https://www.youtube.com/watch?v=MniYgsZSp30.

ChIP-Seq

ChIP-Seq stands for chromatin immuno precipitation sequencing and is an experimental method for investigating protein DNA interactions, such as the interaction between transcription factors and the genomic sequences to which they bind.

The procedure is conceptually simple (figure 6). Cells are lysed and proteins are chemically cross-linked to the DNA, to which they are bound. The DNA is then sheared into small fragments. Antibodies are used to selectively isolate those protein-DNA complexes that contain the protein of interest. The DNA of those complexes is isolated and sequenced by next-generation sequencing.

Sequencing results in short stretches of sequences called reads that can be mapped onto the known genome sequence. They thus reveal the genomic locations to which this protein binds and the relative frequencies with which it binds to them. By comparing the results of ChIP-Seq experiments performed on samples collected from cells grown under different growth conditions, it is possible to directly detect the effect of different growth conditions on the protein-DNA interaction networks.

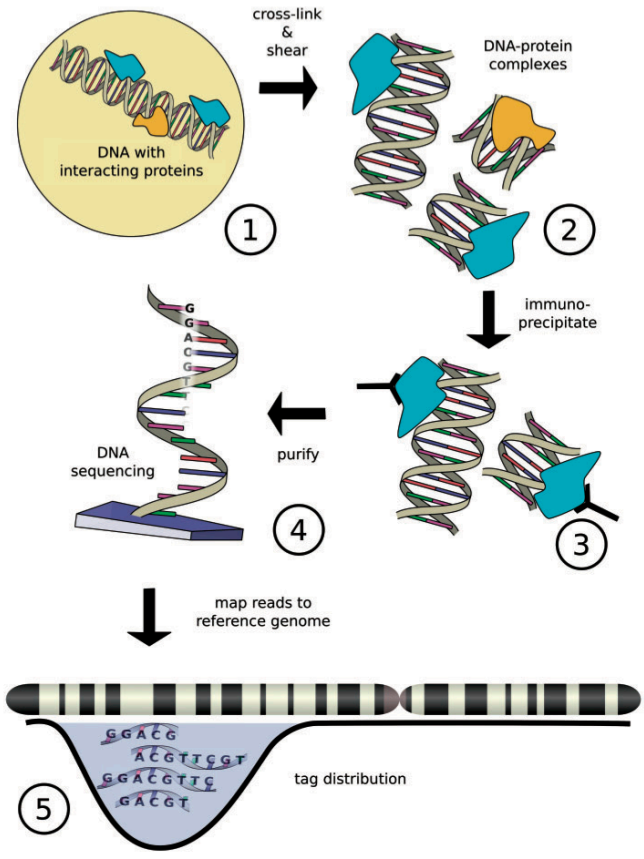

Figure 6 *Workflow of a ChIP-seq analysis: Chromatin in the nucleus (**1**) is cross-linked and sheared (**2**), followed by the enrichment of complexes containing the target protein using immunoprecipitation (**3**). Short reads obtained from massive parallel sequencing (**4**) are mapped to a reference genome (**5**) yielding a distribution of tags on the genome.*