Tommy Angelini
Biological Information
Prof. Nigel Goldenfeld
Fall 2001

Homework 3-2

In problem 3-1, we compared unknown nucleotide and amino acid sequences to documented ones. More broadly, in this course our studies of bioinformatics have been entirely based upon the methods of comparison, relying on the assumption that somewhere there is indeed real information. After all, comparisons are not made for their own sake, and meaningful comparisons must make reference to some known facts. Where does the information come from in bioinformatics? In this essay I will present one of the methods for collecting such information. To keep this essay readable I will define certain terms as they come up. I ask that readers very advanced in molecular biology be patient.

In order to test the functions of proteins, we must first be able to produce large amounts of a given protein. Most often, scientists use phage vectors (viruses) to infect a host and genetically trick it into expressing a certain protein (by inserting some viral DNA). This is done by preparing random cDNA libraries in phage vectors. A cDNA library is a large collection of randomly mutated pieces of cDNA. Complimentary DNA (cDNA) is a piece of DNA made using isolated mRNA as a template. When, for example, a single bacterium is infected by a single virus, the bacterium dies and releases new viruses, along with the protein it was tricked into expressing. The new viruses infect nearby bacteria and so on. If one does this between layers of agarose gel, eventually a big pile of dead bacteria and protein, called a plaque, builds up which can be sucked up with a pipette and purified.

This method works well for many applications, but it has many limitations. Most of the cDNA library encodes proteins using the wrong reading frame, and most of the proteins expressed are not the complete proteins. Also, bacterial expression of eukaryotic genes often fails to yield correctly folded proteins and overproduces products from abundant sequences. Finally, this plaque extracting procedure cannot be miniaturized, which prevents us from performing high-throughput studies.[1,2]

Instead of using cDNA libraries, we can use a large collection of a *single* protein encoding region of DNA at a time. In ref. 3, the authors created a collection of 5800 yeast open reading frames cloned into yeast expression vectors. This has two levels of significance. First, an open reading frame is the most general way to translate a piece of DNA into a protein: it includes all reading frames and even leaves out stop codons. 5800 is 93.5 % of all possible reading frames for translating the yeast genome into proteins. Secondly, since they used yeast as their host, the proteins produced will be folded correctly.

The method for characterizing these proteins is as follows:

1. Clone each of the 5800 reading frames separately, using expression vectors in yeast.
2. Purify each sample in a parallel fashion.

3. Use a robot to pipette a few microliters of each protein solution onto a “chip”
4. Carry out bio-chemical tests on the chips, seeing how each of the “pixels” of protein responds (e.g. binds to something).

A “chip” is nothing more than a chemically treated microscope slide. In ref. 3 the authors used an aldehyde treated slide, to which the primary amines at the NH_2 termini or other residues attach. Also, during the cloning process, a promoter was induced which fused the proteins to glutaione-S-transferase-polyhistidine (GST-His6X) at their NH_2 termini. This allowed them to use nickel-coated slides, to which the proteins attach at their His6X tags. A robot is used to squirt a few nanoliters of a specific protein solution onto a small section of the chip (~150-200 microns in diameter). Done in this automated way, the robot/computer knows which protein is on which “pixel”.

To test the proteome for protein-protein interactions, all one has to do is expose the whole chip to some fluorescently labeled proteins and wash away any that didn’t bind. In the study, the yeast proteome was probed with biotinylated calmodulin and detected with Cy-3 labeled streptavidin (streptavidin is a protein and biotin is a ligand). It is common to use biotin on one side and streptavidin on the other because they have an incredibly large free energy of association. This test allowed the authors to detect six of the known calmodulin binding proteins, and to identify 33 previously unknown calmodulin binding candidates. Doing sequence comparisons, they found that 14 of the 39 camodulin-binding proteins contained a common motif that was closely related to a calmodulin-binding motif in myosins. They also detected a protein on their chip that binds directly to Cy-3 labeled streptavidin.

One can perform the same type of test for other kinds of molecules. For example, protein-lipid binding is important to understand since phospholipids are the major constituents of biomembranes. Protein-drug interactions have an obvious importance. Again, biotinylated lipids were used and Cy-3 labeled streptavidin was used to detect the lipid binding proteins. Six different liposomes were used and 150 lipid binding proteins were identified. Ninety-eight of these were known proteins, and 13 of the unknown proteins were predicted to be associated with lipid membranes. Further, comparisons were made based on the relative strength of the binding, as well as the specificity of the binding (will *any* lipid do, or just certain kinds?).

The authors claim that their protein chips can be used to screen for protein-drug interactions, but presented no such experimental data. It seems correct that one can screen for protein-*almost anything* interactions. The biggest drawback I see is that making and purifying the protein still seems like quite a job. If one wanted to characterize the entire human proteome, a protein chip doesn’t make the earlier steps any easier than other methods. On the other hand, it apparently gives more complete, more reliable, and easier to analyse results. Further, since the process is automated (after the protein is made), the possible amount of information one can extract is really phenomenal.

References:

1. G. MacBeath, S.L. Schreiber, *Science* **289** 1760 (2000)
2. J Sambrook, E.F. Fritsch, T Maniatis, *Molecular Cloning: A Laboratory Manual*, Sold Spring Harbor Laboratory Press, 1989.
3. H. Zhu, *et. al.*, *Science* **293** 2101 (2001)