BNFO301: Introduction to Bioinformatics

Advice on Pursuing a Phage Genome Research Project

I. Typical research project focus

There are hardly any limits on a question that can serve as the focus of your research project except one – ***find something interesting*** – and that doesn't exclude much. However, I can offer a strategy of finding something interesting that has proven to be broadly applicable to a variety of topics within research groups. Understand that this is just one strategy. You might well find something better for yourself.

I.A. Identify a specific protein that has an important role within the broad focus of your group

Many people, given the task of carving out something they can call their own within the broad focus of a group, tend to cast their eyes towards general functions. For example, a person in the DNA replication group might think of nucleotide synthesis. This is certainly required for DNA replication, but by itself, it doesn't give you a handle on what to look for in a phage genome. Genomes contain genes, and genes encode proteins. Much better to identify a specific sort of protein.

How to find such a protein? Review articles can provide an overview of proteins that are involved in a specific process. One place to find an article that may be useful is to visit the appropriate forum on the course Blackboard site. You're looking particularly for a protein that has conserved sequence motifs, something that you can search for in the phage proteins and thereby identify new instances of this type of protein. If you're lucky, the protein you adopt as your focus will have well conserved motifs that will enable you to identify new instances through their possession of similar motifs. But beware! You may well find that what sounds like a specific protein to you exists in nature as multiple types that may have little sequence similarity with one another. In order to make sense of the members of the classes of proteins and use their features for predictive purposes, you will need to separate one type from the other.

I.B. Become an expert on some small slice of the project

It takes years to understand the complexities of any interesting biological problem. But to know what there is to understand -- surprisingly, that takes hardly any time at all! You can collect almost everything that has ever been published on a suitably constrained topic just through a trip to PubMed plus a bit of noodling, and your collection will be no less exhaustive (and probably more so) than that of the world's leading expert on the topic.

The references you find through this exercise can be quite valuable, even if you have not read a single one of the articles you've found. The list tells you what questions have been asked and where to find the answers that have been obtained. It may give you a sense that you have at least drawn a frame around what might otherwise seem a formless topic.

I.C. Identify the features that will enable you to find instances of your protein

Most of a protein's sequence is subject to random mutation without fatal damage to the protein's function. However, there are often certain residues that cannot be changed without a decrease in functionality to the extent that selection will weed out mutants with these changes from a phage's population. If you can identify these residues, then you have a powerful tool to determine whether a protein that bears similarity to the class you're interested in truly exhibit the desired function.

Sometimes specific amino acids are invariant amongst all members of a class of proteins, but more often what you find are common sequence ***motifs***, a collection of nearby amino acids that are conserved, more or less, as a group. You would do well to find an article that identifies conserved motifs in the protein class on which you have chosen to focus.

I.D. Learn how to find the critical features yourself

It's one thing to find a pretty picture of a motif of your favorite class of proteins and quite another to find them yourself in protein sequences. Yet that is what you need to do if your goal is to identify proteins that have not previously been identified. Take the proven cases that you find in articles and collect their sequences within BioBIKE. Then with them in hand, use the function MOTIFS-IN to identify the motifs that are shown in the article.

I.E. Seek identifiable features in phage genes, particularly those without annotation

Once you've convinced yourself that you can find protein sequence motifs that have been described as important in the functioning of your favorite protein class, add proteins you have reason to believe may belong to that class and rerun MOTIFS-IN. You can get candidate proteins by a variety of means, e.g., by their provisional annotation or by sequence similarity. However you may find them, add them to the list of proteins of proven function and determine if the candidates have all of the sequence motifs typical of the class.

I.F. Talk with your colleagues

Science is a puzzle, and no one has all the pieces. Sharing your ideas with others and listening to their ideas in return may yield insights greater than either one could gain alone. It may be difficult to find time to meet with others in your group, a forum has been set up for each group so that you can post ideas, articles, and questions whenever the opportunity arises.