Introduction to Bioinformatics

Introduction to Global Viral Metagenome Project

The time has come to begin to crank up the research project to determine what is the dominant (or at least most numerous) life form (or at least nucleic-acid-based entity) on this planet. A good introduction is provided by the following article, which I commend to your attention:

> Robert A. Edwards and Forest Rohwer (2005). Viral metagenomics.
> Nature Rev Microbiol (2005) 3:504-510.
>
> *Forest Rohwer and his colleagues have increased many fold the number of available environmental viral sequences, and he's a collaborator on the project.*

I'll accompany you as you read the article, offering questions that come to my mind. I hope that you will generate your own questions as well.

BOX 1: Cloning considerations and viral metagenomics

The Abstract and introductory paragraph raise interesting issues, but as they're all covered in more depth later in the article, I'll let them pass for now. The first order of business is to understand what metagenomic sequencing is and how it is done (hence what are its limitations). This is covered in BOX 1, on the second page of the article.

SQ1. How big, typically, are viral genomes? How does that size compare to the size of bacterial genomes (using the number they rather inaccurately give for "microbial" genomes)? How does the size of typical bacterial genomes compare with the sizes of cyanobacterial genomes?

SQ2. What is the critical difference between genome sequencing (that you're already familiar with) and metagenomic sequencing? What special challenges does metagenomic sequencing face?

SQ3. Metagenomic sequencing is considerably more difficult than genomic sequencing. Witness the fact that there are thousands of viral sequences known but only a handful of metagenomic sequences. Then why do it? What are the alternatives?

SQ4. About how many nucleotides of read sequence are available from a typical linker-amplified shotgun library? Suppose for the moment that there were only one virus in the environmental sample. What would be the coverage? Instead suppose, much more realistically, that there are ~5000 distinct viral types in the sample. Now what is the coverage?

SQ5. How is free DNA removed from the sample to be sequenced? How is cellular contamination removed from the sample to be sequenced?

Diversity of environmental viruses

The section starts out by citing four references to published studies on viral metagenomes. If you go back to the references and examine the list of authors, you'd find that all but one are from Forest Rohwer's group (The first author on three of them, Mya Breitbart, was at the time a graduate student in Forest's lab). This illustrates the magnitude of the problem. Precisely two research groups on earth have sequenced viral DNA from environmental samples. Is there any wonder why there are so great gaps in our knowledge?

SQ5. Figure 1 shows eight stacked columns. Describe the significance of the gray portions of the third and fourth columns (Mission Bay, 2002 and 2004). In light of your answer, the two faecal columns are very peculiar. Why?

SQ6. Draw a ninth column using information gleaned from the article, representing results from a typical microbial metagenome experiment.

SQ7. What evidence can you put forth regarding our present knowledge of viral genes as compared to our present knowledge of microbial genes? How do the results of Daubin and Ochman affect the interpretation of that evidence?

Phage phylogeny and taxonomy

SQ8. What's the difference between a phylogeny and a taxonomy? Why might virion characteristics (e.g. the shape of the virion particle) be more useful for a taxonomy than a phylogeny? Why might sequence-based systems be more useful for a phylogeny?

SQ9. From what information was the tree shown in Fig. 2 derived?

SQ10. From Fig. 2, discuss the proposition that siphophage comprise a phylogenetically coherent grouping (which would mean that all siphophage are more closely related to each other than any is to a nonsiphophage).

SQ11. At the bottom of p.505, the authors describe a simulation they performed. What was the motivation behind the experiment? What did they find? What are the implications for the analysis of available metagenomic sequences?

The proviral metagenome

SQ12. What's a provirus or prophage? Why would a virus do such a thing?

SQ13. At the bottom of p.506, the authors claim that "…about 1% of the microbial metagenomes encode phage proteins. How was this number derived?

Viral community structure and ecology

We want to know not only what viruses are out there but also how complicated is each sampled environment. An environment may contain a lot of viruses but (as in the case of faecal samples) not very many ***types*** of viruses. Or (as in the case of marine samples) the environment may contain the same number of viruses spread over many types. If we could sequence every virus in a sample, with a high enough coverage to assemble all the sequences, we'd know the answer directly. But to date, viral metagenomic projects have not obtained sufficient numbers of sequences to permit total assembly.

Nonetheless it is still possible to ***estimate*** the complexity of an environment. Consider this analogy. Suppose you're in front of a machine where you put in a quarter and it gives you a cheap toy. The kid next to you is screaming that he wants another toy. You keep putting in quarters, and he keeps screaming. After ten different toys (and $2.50), the eleventh comes out, and it's one that you've seen before. So you say to the kid, "Look, it's run out of toys. It's giving us the same ones as before." He may scream a lot, but he's no idiot. He says "No, there's LOTS of different toys left, AND I WANT THEM!" Who's right? And how many different toys ARE there likely to be in the machine?

The kid has a point. If the machine gives the toys in order – toy #1, toy #2, … all the way up to toy #10, and then it starts over, then you're right. But that would take a lot of work to set up. More likely, the toys are coming out at random, and if it takes 11 toys to get a duplicate, then there probably are many more distinct toys beyond the first 10.

That's the principle behind the Lander-Waterman algorithm mentioned in this section. If you get long contigs (many reads overlap), then you know that the number of possible sequences that can be read is being exhausted. On the other hand, if most reads have no overlap, then you know that the number of possible sequences that can be read must be much larger than what you have sampled. This qualitative assessment can be made quantitative and be used to predict the total number of possible reads and (assuming some average length of a virus) the total number of unique viral sequences.

SQ14. What does it mean when the authors say (on p.507, middle of the left column) that a sample contains ~10^{12} viral particles but only ~1000 viral genotypes? Make up two different scenarios that would be consistent with this claim.

[I'm going to jump now to the next section, skipping the description of community.]

Bioinformatics and viral metagenomics

When you assembled the miniplasmid sequence by hand a couple of weeks ago, you did essentially what assembly programs like Phred/Phrap does, with one important exception. In your case, there were no sequencing errors. You looked only for exact overlaps. But in fact sequencing errors are an unfortunate fact of life, and assembly programs must take them into account. It can allow overlaps between reads that have imperfect sequence similarity, so long as the number of mismatches is not so high (as judged by the known accuracy of the sequencing process).

SQ15. Why do the authors claim that assembling metagenomic sequences poses more of a difficulty than assembling sequences from a single organism?

SQ16. The authors describe a program, tblastx, that is very similar to BlastP and BlastN, which we have already used. What is the difference amongst these programs? Why does tblastx require much more computer time?

In BOX2, the authors propose analytical approaches that go beyond Blast. From what we have already done, you may be able to guess the nature of some of these approaches.

SQ16. What might it mean to analyze "codon usage"? What has to be true for this approach to be of any use?

SQ17. What might it mean to analyze GC/AT content? Note that BioBIKE has a function called GC-FRACTION-OF (under STRINGS-SEQUENCES/String Analysis). Play with it using sequences you have made yourself to figure out what it does. There is no function called AT-fraction. Why not?

SQ18. Are there differences amongst cyanobacterial genomes with respect to GC/AT content? Presuming the same is true with phage genomes, how might this help?