Introduction to Bioinformatics

Initial Analysis of Reads from Viral Metagenome

A. Overview of the project

You are now beginning the task that all of your prior training has prepared you for, and you are now ready to assume your proper role within the Global Viral Metagenome Project. The sequences you consider, taken from either the Bear Paw Hot Springs metagenome or the Octopus Hot Springs metagenome [Schoenfeld T et al. (2008) Appl. Environ. Microbiol. 74:4164-4174] have been seen by no one else besides yourself. Tom Schoenfeld and his colleagues have analyzed the sequences in the aggregate, but no one has looked at individual reads, until now. It is yours to determine for each of the sequences under your care what is interesting about it. How do you know if you get the right answer? The right answer isn't known. You will be the first to know it. The first task, then, is discovery, but the equally important second task is presenting your insights in such a way that others may understand what you have understood. Otherwise your efforts will come to nothing.

But what is "interesting"? Certainly each of you will want to determine as best you can whether your sequences contain genes. You'll want to predict to the extent possible the function of any encoded protein. What kind of virus might your sequence have come from? As you seek what is interesting, ask yourself, "*Have I seen something like this before*?" The answer will sometimes be yes. Find genes? Think about the exercise using `READING-FRAMES-OF` and GeneMark. Protein function? Think about your mystery sequence. Viral source? Think about Karlin's article and your efforts with codon and dinucleotide biases. But if that's all there was, you wouldn't be necessary. These tasks can be taught to a computer and automated. You are essential because sometimes you should be struck by something in your sequence you haven't seen before. Maybe nobody has seen it before -- we can't teach a computer to look for something we have never considered. It is critical that you go beyond automated processes and remain alive to the story your sequence has to tell. To do this, you will need to work in a way that is foreign to most of you, putting aside the comfortable general principles and simplifications found in text books and grapple directly with unvarnished biological information.

But on with the show!

B. Obtaining a sequence from a metagenome

Almost all of you have already obtained a sequence, but I'll review the steps here.

Log onto ViroBIKE (perhaps getting there through the BioBIKE Portal linked from the course Resources and Links page). Functions specific to this project are not part of BioBIKE but rather are found in a module called `viral-metagenome`. To make these functions available to your session, `ENTER viral-metagenome`, just as you previously obtained special functions from the `Alien-world` module. After executing the command, a FUNCTION button should appear on your palette, containing at least two functions: `CLAIM-READ-FOR` and `READS-OF`. The first enables you to claim a short read sequence from one of the two metagenomes, identifying it as uniquely yours. The second reminds you about what reads you have claimed.

If you haven't already done so, claim a read for yourself now by choosing your name from the options and executing the `CLAIM-READ-FOR` function. There is also a `FROM` option, in case you should ever want to claim a read from a specific metagenome (either Octopus or Bear-paw), but initially you have no basis for a preference, so just take pot luck by not using the FROM option.

SQ1. (If you haven't done so already) Use CLAIM-READ-FOR to get a read. You probably got *two reads. Examine the names of the two reads carefully. Why two?*

The first characters before the dot identify the metagenome (either BPHS for Bear Paw Hot Springs or OctHS for Octopus Hot Springs). The next eight characters before the hyphen identify the clone. The two characters after the hyphen identify the read. Notice that the metagenome IDs and the clone IDs are the same. The only difference is the read IDs. When you claim a read, you get every read associated with the clone, since the two reads should come from the same viral genome and be physically close to each other.

These reads you claimed are yours permanently. You can check at any time what reads you have claimed using the READS-OF function (specifying your name). Don't try to use to retrieve your reads with the CLAIM-READ-FOR function. You'll just get one or more ***additional*** reads to add to your collection. It is not easy to unclaim what you have claimed!

SQ2. Use READS-OF to display your set of reads. What is the *form* of the result (e.g., is it a table of values?)

C. Basic characterization of reads

What are these reads you now have? One thing you can do quickly is to get a description.

SQ3. Use a function used to describe genes and proteins to describe your reads. Copy/paste into its argument the result from READS-OF.

I wouldn't be surprised if the result of executing the function told you that your list of reads is a "Simple list of 2 elements". You knew that! You wanted to know what ***each*** read is. Specify the EACH prefix before the argument of DESCRIPTION-OF and try again.

SQ4. What type of entity does BioBIKE consider a read? (Look at the poorly named "Organism-entity-type"). What metagenome do they come from? How long is each read? Are they circular (like plasmids)?

Besides this, there's not much to learn from the descriptions. Perhaps the sequences will be informative. Even if it isn't now, you'll certainly need to have the sequence on hand as you analyze it.

SQ5. Display the two sequences. Do you see anything unusual?

Most reads have one or more N's at the beginning or at the end of the sequences. N indicates that the automated nucleotide caller saw a nucleotide at the given position but it couldn't decide exactly what it is, so it inserted N as a place holder. You very seldom see N's in published sequences or sequences in databases because these uncertain parts are cut out. Seeing them is a cause for concern. Evidently we have unedited sequences and especially in the early and late regions, we should not have a great deal of confidence in the sequence.

D. Initial analytical steps: Search for overlaps

You would definitely like to identify genes in your sequences, if they exist.

SQ6. What is the likelihood of finding a full-length gene in your first sequence? Presume that a gene is 1000 nt long (a typical average). Presume also that viruses contain only genes, one after another (this is unrealistic since you'd expect about 10% to be noncoding, but let's keep it simple).

Not good. You would have better luck if you were able to extend your sequence by combining them with other reads from the metagenome that overlap with yours.

SQ7. Have you done something like this before? How did you find overlapping fragments?

SQ8. How many reads are there in your metagenome? Have you done something like this before? You've counted how many genes there are in genomes, but reads are considered by BioBIKE to be what? Search in the GENOME menu a function that will enable you to list the reads and hence count them.

Doing an alignment of all those reads would take hours. And looking for end sequences in Word would be a nightmare. We need a more powerful way to look for sequence similarities.

SQ9. Do you know a powerful way of looking for sequence similarities?

Go to the STRINGS-SEQUENCES menu and bring down `SEQUENCE-SIMILAR-TO`. This is the BioBIKE function that accesses Blast. Use it to search for reads within your metagenome that have nucleotide sequence similarity to one of your reads. If you use as the query one of your reads and use as the target your metagenome, BioBIKE will take care of labeling. If you instead use something like `(SEQUENCES-OF Octopus)` you'll get the comparison but lose the names of the reads.

SQ10. How might the results of a Blast help you in finding overlapping reads? Draw a picture of the situation you're looking for, with (made up) coordinates. Give two different scenarios, drawing a different picture (and different coordinates) for each.

Now perform the blast, executing a completed `SEQUENCE-SIMILAR-TO` function. By the way, in copying and pasting your reads, you may encounter something like this: #$OctHS.APNO1000-b2. The #$ are internal symbols indicating that what follows is a frame (i.e. a collection of information). Feel free to delete the symbols if they bother you.

SQ11. Examine the first (best) match. How good is it? How much of the query matches? How long is the query? (try clicking on the query to find its length) How much of the target matches? Why is this match so good?

SQ12. If the match does not begin at nucleotide 1 and go all the way to the end, why not?

The first hit is no doubt spectacular but, so long as you're blasting the metagenome of your read, not very interesting. What about the other matches? Hold on to this result. You'll need it. But for purposes of discussion, I'm going to direct you to a result we can all talk about.

SQ13. Redo the Blast, this time using the read AOIX2864-b2 as the query and bearpaw as the target. Describe in general terms the matches you find. Describe the two best matches (apart from the match of the read with itself). There's a good deal to describe.

SQ14. Note that the T-START (start of the target match) has a larger coordinate than T-END (end of the target match). How can that be?

The two best matches (shown on lines 2 and 3 of the Blast result table) are very interesting. Notice what part of the query is matched and what part of the target. Notice also ***which*** target is matched.

SQ15. Draw a picture of showing the relationship between the original read and the two matches listed on lines 2 and 3 of the Blast result table, labeled with coordinates.

SQ16. Have BioBIKE align the three sequences (the original read plus the two partial matches) using the `ALIGN-BLAST-RESULT` function (you can get this function off of the STRING-SEQUENCES menu, SEARCH/COMPARE submenu). Copy and paste the Blast result table (in the <u>Results window</u>, not the <u>printout</u> of the table!) into the argument, and use the FROM and TO options to specify lines 1 through 3 of the table. Look carefully at the alignment. Does it correlate well with your drawing? In what way? In what way not?

SQ17: What does the "R" at the end of two of the labels of the alignment signify? (one of the labels is BPHS.AOIX3283-g2:723-885R)

SQ18. Repeat the execution of the `ALIGN-BLAST-RESULT` function, this time specifying lines 1 through 2. How does the alignment \differ from the previous? How can you explain the difference?

SQ19. Final alignment: Try using `ALIGNMENT-OF` to align the complete sequences of the original read AOIX2864-b2 and the read AOIX3283-g2 containing the matches it found. Since AOIX3283-g2 matches the original read only when it is inverted, then you need to supply the function with the `SEQUENCE-OF` the read, using the `INVERT` option. How does this alignment compare with the others?

The `ALIGN-BLAST-RESULT` function and the similar `ALIGNMENT-OF` function uses a program called Clustal, which is undoubtedly the most widely used program to align multiple sequences. However, another way of aligning sequences is using your own eyes and a good word processor. Judging Clustal's results against the evidence of your eyes, what do you conclude? The sad fact is, Clustal isn't perfect, and often it is positively awful. It can produce alignments far faster than you can by eye, but you cannot trust the results without checking them yourself. It is particularly bad at aligning sequences where there are large gaps, as in the case above.

E. Solving the first mystery: Massive overlap

Now consider the remaining matches in the Blast result table.

SQ20. How would you describe the matches from line 4 onwards (with one exception)?

Hundreds of matches! And this is no isolated incident. Almost all of your own reads have the same sort of match! There is something special about the first few dozen nucleotides. Why are they repeated in so many other reads? This is a major mystery, crying out to be solved. One important question to answer is how common this special characteristic amongst the reads. We'll take as a semi-random sample the first 100 reads of the Octopus metagenome

SQ21. Define a variable containing the first 100 reads of the Octopus metagenome. Use the `FIRST` function (you can find it on the LIST-TABLES menu, LIST-EXTRACTION

submenu). Click the Number prefix and enter 100. For the entity, put in the `CONTIGS-OF` octopus.

SQ22. Align the first 100 nucleotides of these 100 sequences. (If you don't know how to get them, think of *What is a Gene?*). You can make your life easier later on if you set the GROUP-LENGTH option to a very large number (e.g. 1000). This eliminates the spaces between groups of nucleotides. Save the window containing the alignment. You'll need it later. What do you conclude about the first 100 nucleotides of the reads of Octopus?

Incredible isn't it? What do we make of such massive similarity?

SQ23. Copy the part of the alignment that is most similar (~35 nucleotides). Delete hyphens if necessary. Peer at this sequence. Look for strangeness. Peer more. Carry it with you wherever you go and take it out on occasion to refresh your memory. Write it on the ceiling of your room so you see it when you wake up. Find anything?

SQ24. Maybe someone else has seen this sequence before. We can ask GenBank. You know how to do that by going to the NCBI web site, but try a different way, from the comfort of your own ViroBIKE. Bring down SEQUENCE-SIMILAR-TO, paste the sequence you found in SQ23 as the query. For the target, go to the DATA menu and click on *GENBANK*. What kind of Blast do you want to perform? BioBIKE tries to guess by looking at the sequences you give it, but it generally pays to specify. Included in the options are DNA-vs-DNA (i.e. BlastN), Translated-DNA-vs-Protein (i.e. BlastX) and others. Choose one and go. Be prepared to wait up to a minute for an answer. Don't expect an orchestral fanfare for an answer. Just look at the Result window.

Depressing. But wait! Blast returns only those matches that are better than a certain threshold E-value. BioBIKE sets the threshold at 0.001. Maybe we can find something a bit similar but still significant if we lower the bar.

SQ25. Repeat the Blast of SQ24 but first setting the THRESHOLD option to 10. What does an E-value of 10 mean? What E-values do the resulting matches have. Are you surprised to find such matches?

Oh well. The most amazing repeated sequence you'll ever find, and GenBank can tell us nothing. So we're on our own. Now what? Well, a good maxim is "*When you have no idea what to do, read the directions*". The directions in this case is the procedure Schoenfeld et al followed to obtain these reads. Go back to his paper and read how they constructed the library of clones used in the sequencing.

SQ26. Describe as best you can the events that produced the clones. It is very likely that some of the steps will be totally mysterious to you. Be sure to identify these holes in your knowledge so that you can ask for enlightenment. One source of enlightenment will be www.genycell.com/images/productos/protocolos/41015-1__39.pdf which is a company brochure on pSMART (see Appendix D in particular).

SQ27. Given your knowledge of how the library was made, reconsider the alignment you made in SQ22. I suggest that you copy it into a word processor and color in the sequences that are of special significance.