(30 March 2014)

Bioinformatic Analysis of Small Dispersed Repeats

I. How to find ill-defined tandem repeats?

You have just embarked on a research project to analyze the occurrences of some type of repeated sequence. It would be helpful to have some model of how your project might proceed. No doubt you are scouring the literature looking for such models pertinent to your chosen topic, but in parallel, let's look back on the articles we have considered collectively in this course.

Early in the semester, you read part of a chapter by Compeau and Pevzner (2013) (and a tour of same) that led you through figuring out how to find origins of DNA replication. You may have gotten some idea of how to develop an algorithm and some tools of either general use (e.g. COUNTS-OF-K-MERS) or more specific use to origins (we didn't get that far in the chapter), but you didn't see the end product, because the chapter was not a research article. Its purpose was not to describe the methods and results of a research project. It may be useful for inspiration, but it isn't a good model for your own research project, particularly if your project focuses on clustered dispersed repeats.

Later, you read Shine and Dalgarno (1975) (and a tour of same), a research article that described an inquiry into sequence motifs that precede protein-encoding genes. You used the article and 40 years of progress to look more broadly at these sequence motifs and how they may help you detect the boundaries of genes. However, the article itself does not provide a good model for your project, since in 1975 the amount of sequence available was miniscule, a very different situation from your own.

Last week, you read Mazel et al (1990) (and a tour of same) and learned about tandemly repeated sequences. The authors were unable to perform a global search for tandem repeats in a bacterial genome, because the first bacterial genome sequence became available only in 1995. Mazel et al were forced to use biochemical methods to estimate the frequency of the tandem repeats they observed.

Today, for the first time, we will consider an article that was written in the age of genomic sequences and has its express purpose the analysis of a type of repeated sequence. Although its primary focus is short dispersed repeats (see Repeats, Tandem Repeats, and Pattern Matching for what I mean by that and related terms), the article also brings into the analysis tandem repeats and very short dispersed repeats, as well as a brief excursion into the subject of regulatory sequences. There's something for everyone in this article, and the methods employed may be useful to some people in all of the research groups.

So it's time to get the article:

> Elhai J, Kato M, Cousins S, Lindblad P, Costa JL (2008).
> Very small mobile repeated elements in cyanobacterial genomes.
> Genome Res 18:1484-1499.

You might want to skim the article with the following questions in mind (all of the time thinking as well about your own project):

- What questions did the authors find interesting concerning short dispersed repeats?
- What kind of analyses were performed? What kind of tools would be necessary to do these kind of analyses?
- What was the motivation behind the article? Why the emphasis on mobility?

II. Motivation behind the study

You might believe from generalities presented in text books that introns are common in genes. This is true only in certain eukaryotic genomes. Alternatively, you might believe that introns are found only in eukaryotes. This also is not true. Introns are found in prokaryotic genomes and phage genomes but only in a small fraction of genes. One such gene determines tRNA-leu (Fig. 1A).

Costa et al (2002)[*] studied the sequence of the intron in this gene and found it to be highly variable in *Nostoc* strains, far more so than the variability of sequences in other genes, which are pretty constant amongst related strains. The reason turned out to be the presence of 7-mer repeats, of the same sort seen by Mazel et al (1990). However, a fraction (8 of 54) introns have variability that went beyond the number of 7-mer units. Costa et al found that in these cases the sequence appeared to have taken up an additional short sequence (Fig. 1B). The variability of tandem repeats was expected. Such repeats are known to add or subtract units rapidly in evolutionary time. But the appearance of the interpolations was very puzzling.

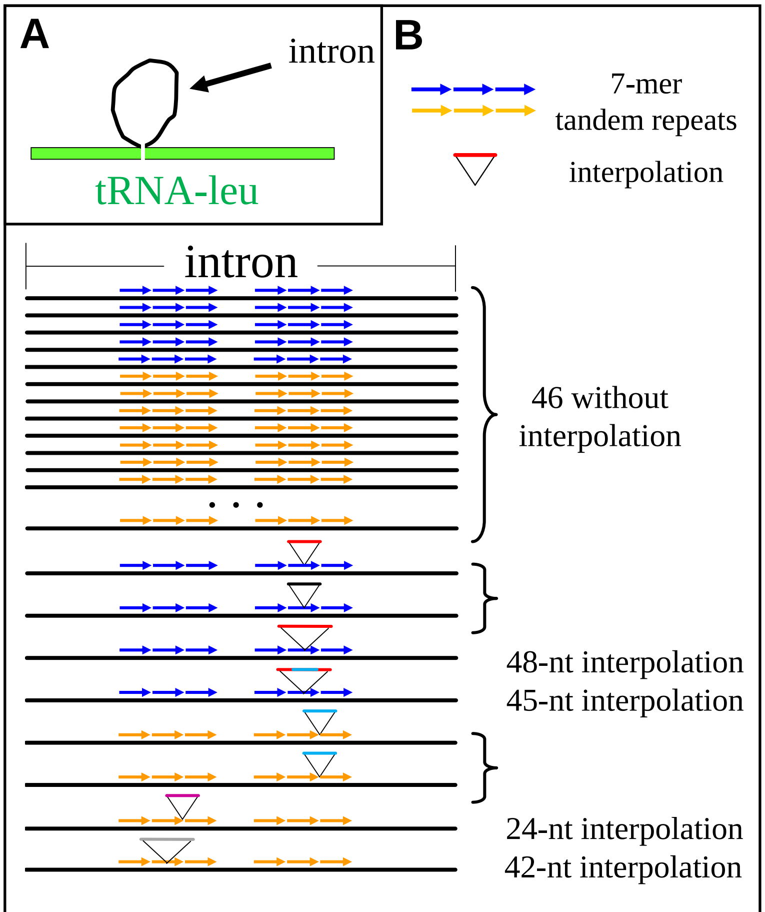

Fig. 1: Structure of introns from tRNA-Leu gene sequenced from 54 isolates of *Nostoc*. (**A**) cyanobacterial tRNA-leu genes contain introns. (**B**) The introns of these genes in *Nostoc* contain two sets of 7-mer tandem repeats of variable lengths. In some cases, the introns have short sequences interpolated into one of the sets of tandem repeats (Costa et al (2002) Mol Biol Evol 19:850-857).

SQ1. Why was this result puzzling? What class of DNA sequences are able to suddenly appear in new positions? Why wouldn't this class explain what was observed as interpolations into the introns?

SQ2. Considering the results of this article, are the interpolations specific to the introns of tRNA genes? Do they appear elsewhere in the genome?

SQ3. Are the interpolations always associated with tandem repeats?

SQ4. Why are the authors so interested in the mobility of these sequences?

[*] Costa JL, Paulsrud P, Lindblad P (2002). Mol Biol Biol 19:850-857.

III. **Characterization of short dispersed repeats**

Consider SDR1 as an example of a short dispersed repeat considered by this article, to get an idea as to what kinds of questions were asked and how they were answered.

SQ5. What are two ways that instances of SDR1 were found in the genome of Nostoc punctiforme?

SQ6. Implement those two ways yourself, using either CyanoBIKE (where the organism has the nickname Npun) or Phantome (where it has the nickname Npun-73102). Do you get similar results to those shown in Fig. 1?

Fig. 1 shows context of the SDR1 instances. If you are using a pattern to find the instances, you can use the following trick to display the context:

```
MATCHES-OF-PATTERN "*{14}pattern*{14}" IN Npun
```

where *pattern* is whatever pattern you use to find SDR1. The overall pattern causes BioBIKE to return 14 nucleotides on each side of a match.

SQ7. The authors make a big deal about a 4-nt x 2 gapped palindromic sequence. Is it really surprising to see such a palindrome? What is the probability of finding such a palindrome in a 24-nt sequence?

SQ8. If it isn't probability, then what convinced the authors that the palindrome is of biological significance?

Later on in the article, there's a section that considers where SDR elements are found in the genome.

SQ9. Why might one expect to find the elements within genes? Why might one NOT expect to find them in genes?

SQ10. In fact, where do SDRs reside relative to genes? Why such a difference from one SDR to another?

SQ11. Why might certain regulatory sequences be found preferentially in the C context as compared to the D context (where contexts are defined as in Table 4 of the article)? There are two SDRs that appear in the C context more frequently than the others. Why do you suppose that is?

SQ12. Confirm the results presented in the article by finding the genetic contexts of a few of the repeats you identified in SQ6, using CONTEXT-OF.

Still later in the article, there is a discussion of results of where in the genome the SDRs insert -- their ***target*** preferences.

SQ13. How is the question of the target of an SDR related to the question of its length?

SQ14. What strategy was used to address these questions?

Try it out, attempting to recreate the result with SDR5.1. The first step is to identify the gene into which SDR5 was inserted. Fig.6 says it's npf3461. Go to Phantome/BioBIKE and find that gene.

SQ15. Go to Phantome/BioBIKE and find the gene. How would you go about doing that?

Well, it would be too easy (and too uncommon) if the name of the gene in the article matched the name of the gene in BioBIKE, so you need a more reliable strategy. The surest bet is to use the ***sequence***, if you're fortunate enough to have a bit of it, as you do in this case.

SQ15. In Phantome/BioBIKE use SEQUENCES-SIMILAR-TO and the sequence fragment shown for npf3461 in Figure 6 G to find the gene. Here's how:

- **Type at least 20 nucleotides into the *query* box (of course surrounded by quotation marks – otherwise BioBIKE would think that the letters spell out the name of a variable).**
- **Use the options DNA-vs-DNA and IN.**
- **Bring down the function GENES-OF into the *value* box of IN**
- **Type Npun-73102 in the *entity* box of GENES-OF (or get the genome from ORGANISM's Bacteria menu)**
- **Execute the function**

What's the name of the gene that contains this instance of SDR5.1?

SQ16. Find other genes that are similar to that gene. Here's how:

- **Put p- before the name of the gene to change it into the corresponding protein and put it in the *query* box of SEQUENCES-SIMILAR-TO. (Why do you want to use the protein rather than the gene to find similar genes?)**
- **Use the options PROTEIN-vs-PROTEIN, BYPASS-LOOKUP, and IN. BYPASS-LOOKUP avoids the fast pre-computed Blast table (which you don't need if your target is only several organisms) and ensures that even organisms that are new to the database are included in the search.**
- **Bring down the function ORGANISM/S-NAMED into the *value* box of IN and type "Nostoc" into its *name* box. This function will return organisms with "Nostoc" in their names. Try it out by executing just ORGANISM/S-NAMED. Happy with the result?**
- **Execute the function**

How would you describe the results of the Blast (which is what SEQUENCE-SIMILAR-TO does)?

You can draw a line somewhere, keeping the best matches and tossing the feeble ones using the FIRST function and choosing the **number** pre-option. For example, FIRST 5 in *entity* will give you the top 5 matches, if you drag the results of the previous function into the *entity* box.

SQ17. How closely do the proteins you found in the previous step align with each other? Here's how you can find out:

- **Bring down the ALIGNMENT function and drag the result containing all the good protein into the *sequence-list* box.**
- **Execute the function**

Is there any region of the protein alignment that stands out as unusual? How long is that region? How long would be the DNA that encodes that region? What is that region?

SQ18. How closely do the genes encoding the proteins align with each other? Here's how you can find out:

- **Mouse over the green action icon of the list of proteins in the *sequence-list* box and click Surround with**
- **Bring down the GENES-OF function. It should surround the list of proteins.**
- **Execute the function**

Is there any region of the DNA alignment that stands out as unusual? Note its coordinates. How do the coordinates relate to the coordinates of the unusual region in the protein alignment? How does the DNA alignment in that region compare with Fig. 6G?

SQ19. If the SDR5.1 sequence weren't there (or, if you imagine going back in time before it somehow got inserted) what would be the sequence of the DNA at the insertion site? From the text of the article, what is special about that sequence?

SQ20. The article claims that the sequence is highly overrepresented... is it? Compare the number of times it occurs in npun-73102 with an estimate of the number of times you would expect it to occur.

If you can follow these steps to recreate the results of the published article, you can use the same steps for similar purposes of your own, obtaining results that are not already known.

IV. **Unbiased search for SDR elements in genomes**

This article resulted from the accidental discovery of short inserted sequences in tRNA genes. What if that accident hadn't taken place? Suppose you wanted to find out what SDRs occur in a genome knowing nothing about what sequences to look for?

This question is addressed in the section of the article entitled "*Exhaustive search for SDR elements...*", and the method is briefly described in the **Methods** section, in the third paragraph of **Database search**.

SQ21. Do you recognize the method? Hint: Take a look (or another look) at Problem Set 7, Question 4.

PS7.4 describes an algorithm that is essentially the same as what was used in the article, however, it is implemented in the problem set in a way that is too inefficient to work with a genome much larger than lambda. COUNTS-OF-K-MERS is an efficient implementation of the algorithm.

SQ22. Bring COUNTS-OF-K-MERS into your workspace as previously described (e.g. in the notes *Computational Detection of Origins of Replication*). Execute it, using Npun-73102 as the *organism-or-seq* and 24 as the *window-size* (both as used to construct Table 3 of the article). In addition, use the BOTH-STRANDS and THRESHOLD option, setting the threshold to 10 to save some time – you're not interested in repeats with lower copy number. Then execute the function. You won't get the same results as in Table 3, because COUNTS-OF-K-MERS isn't clever enough to know what a CRISPR is or what SDRs are, but you'd think the numbers at least would be right. Are they? Make any sense of the sequences?

COUNTS-OF-K-MERS may be useful to certain members of each of the research groups.