Introduction to Bioinformatics
Repeats, Tandem Repeats, and Pattern Matching

I. An argument for the study of repeated sequences

Genome analysis generally centers on genes:

- Where are they?
  Where do they begin and end? How do they cluster into logical units?
- What are they?
  What is the predicted functions of the encoded proteins?
- How do they differ from each other?
  What regions are conserved from one organism to the next? What can differences tell us about evolutionary history?

This, perhaps, is as it should be. Proteins, and therefore the genes that encode them, to a great extent direct the workings of the cell, so we are well advised to scrutinize them.

However, there is more to DNA than genes -- much more. Most obviously, there are the DNA sequences that control the expression of genes, to ensure that liver cells have the proteins required for liver function while skin cells (containing exactly the same DNA) have the proteins required for skin function. We've come to understand how the information contained in genes relate to protein sequences through the genetic code. We're much further behind in deciphering the other messages in DNA.

Egyptian hieroglyphs were not deciphered until the Rosetta Stone was found, containing hieroglyphs side-by-side with relatively intelligible Greek text. In the case of DNA, we don't have the equivalent of intelligible text to guide us in deciphering the parts of DNA we don't yet understand, so how can we proceed?

Imagine a message sent to us from an alien civilization (Fig. 1A). We would very much like to receive and understand this message, but it is written in the alien's own language, somehow transliterated into Roman letters (Fig. 1B). Even though we don't know the language and don't know the message, it isn't difficult to appreciate that there ***is*** a message there to be understood – the repeated elements gives that away. We can still detect the repeated elements, hence the promise of meaning, even when the message (Fig. 1C) is embedded in noise (perhaps the result of passage through interstellar space).

GREETINGS PRIMITIVE EARTHLINGS!	VYNNWGZVT LYGXGWGDN NUYWHPGZVT!	RBCVYNNWGZVTOLYGXGWGDNXS MMOPFNENUYWHPGZVTEIMJOFR
WE COME IN PEACE.	QN 2SXN GZ LNU2N.	KHARQNG2SXNGZNLNU2NXL12V
WE INVITE YOU TO LIVE IN PEACE WITHIN A GALACTIC COMMUNITY…	QN GZDGWN RSF WS PGDN GZ LNU2N QGWHGZ U VUPU2WG2 2SXXFZGWR…	QNNGZDGWNKRSFKWSVPGDNTGZ KLNU2NFQGWHGZGUVUPU2WG2F A2XJFNJ2SXXFZGWRBXMUP2TS
TESTING, TESTING, 1 2 3	WNTWGZV, WNTWGZV K J O	HWNTWGZVNZWNTWGZV2KPJSOH
(hey, is this thing on?)	(gkq, xt rgxt rgxjb dj?)	22GKGMNXTWRGXTURGXJB3DJA

Fig. 1: Message from alien civilization. (**A**) Meaning of alien message. (**B**) Message transliterated into Roman alphabet. (**C**) Transliterated message embedded in noise.

The same is true with DNA. Repeated sequences within genomic DNA hold out the promise that they are there for a reason, either by virtue of a process that is inherently periodic (e.g. the noise that comes from a screen door that's banging in the wind) or by virtue of a positive function, preserving the repeats through selection.

Your research projects this semester will focus on repeated sequences in the genomes of bacteria and bacteriophages. Since they are so little studied, it you will have the real possibility of discovering something that is not already known. That's why the projects are described as "research" rather than "exercise".

II. Taxonomy of repeated sequences

Repeated sequences may be divided into two broad categories: tandem and dispersed. Tandem repeats (Fig. 2A) are characterized by consecutive copies of a repeated unit. These are well known in human DNA, where an excessive number of consecutive copies of the trinucleotide "CAG" is responsible for Huntington's Disease, and expansion of a "CGG" repeat causes Fragile X Syndrome.[1] The relatively high rate at which the number of tandem repeats change in the human genome is also the basis for an important method of forensic identification.[2]

Much less is known about tandem repeats in bacterial and phage genomes. You will do your part to rectify that situation.

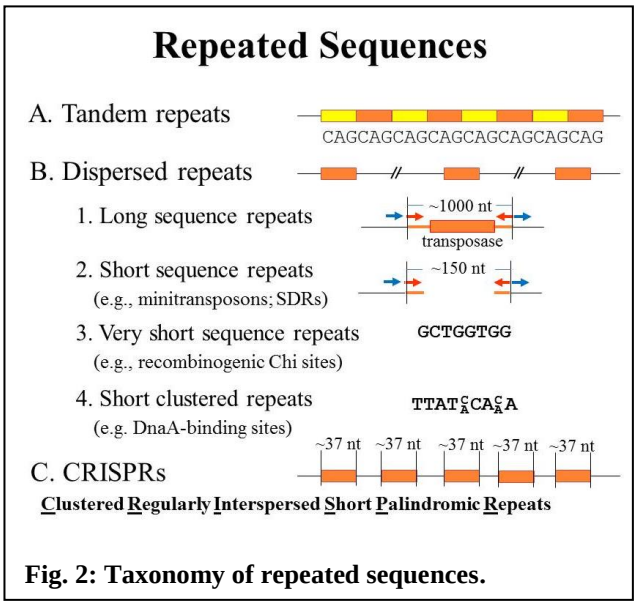

Fig. 2: Taxonomy of repeated sequences.

Dispersed repeats (Fig. 2B) are characterized by units that are positioned at positions distant from one another. The best known in this category are transposable elements, which can move from one position in a genome to another, catalyzed by an enzyme, transposase, encoded by a gene within the element itself. These are not repeats of merely academic interest -- about half of the human genome is taken up by transposable elements![3] Most transposable elements in bacteria are of the DNA type shown in Fig. 2B1, where the transposase gene(s) is often flanked by short inverted DNA repeats[4] (shown as orange arrows). The process of insertion in many cases is achieved by a cut and fill mechanism, which generates direct repeats (shown as blue arrows) that lie just outside the transposon. The inverted repeats are recognition sites for transposase, and seemingly parasitic DNA fragments have arisen multiple times in evolution, consisting of these repeats and little else. These fragments, often called minitransposons (Fig. 2B2), can also be acted on by transposase, even though they themselves possess no genes.

---

[1] Budworth H, McMurray CT (2013). A brief history of triplet repeat diseases. In: *Trinucleotide Repeat Protocols* (Kohwi Y, McMurray CT, eds), Methods in Molecular Biology 1010:3-17.
[2] Thompson R, Zoppis S, McCord B (2012). An overview of DNA typing methods for human identification: Past, present, and future. In: *DNA Electrophoresis Protocols for Forensic Genetics* (Alonso A, ed), Molec Biol 830:3-16.
[3] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409:860-921.
[4] Siguier P, Fileé J, Chandler M (2006). Insertion sequences in prokaryotic genomes. Curr Opin Micriol 9:526-531.

Dispersed repeats may be very small, for example the 9-nucleotide chi sites (Fig. 2B3) that occur about every 5000 nucleotides in E. coli and promote recombination between homologous DNA.[5] You've already seen one instance of such a dispersed repeat – the clustering of DnaA-binding sites at origins of replication[6,7] (Fig. 2B4). Unlike transposable elements, however, these are not capable of moving from one site to another.

Within the last 12 years a hybrid category of repeated sequences has been shown to be abundant in a wide range of bacterial and archaeal genomes.[8] These sequences are not tandem repeats because the repeating units are separated by small chunks of non-repeating DNA, but they are invariably found a nearly constant distance one from another. They have been shown to be important in imparting resistance to bacteriophage attack.

SQ1. In what category would you place lambda operators (as presented in Gene Regulation and Bacteriophage)?

SQ2. Consider the repeated sequences that are not able to transpose (i.e. everything but those shown in Fig. 2B1 and 2B2). What ways can you imagine by which they might arise in so many positions in the genome?

III. Discovery of extensive tandem repeats in bacteria

Most repeated sequences were discovered by accident, by people who weren't looking for them and had no particular interest in such things. To see a case in point look up:

> Mazel D, Houmard J, Castets AM, Tandeau de Marsac N (1990).
> Highly repetitive DNA sequences in cyanobacterial genomes.
> J Bacteriol 172:2755-2761.

Our goal is to understand what repeats they found, which will undoubtedly entail finding them ourselves. With that in mind, on with the article!

Introduction

What's this? Photosynthesis? Hot springs? Light-harvesting antennae? You can see clearly in the first two paragraphs what the authors do for a living. Their first interest is definitely not repeated sequences. The third paragraph is of more interest. This time, they talk about repeated sequences, but nothing about tandem repeats. The main reason is that, so far as prokaryotes are concerned, there was nothing to say. The necessary portion of the Introduction is the short last paragraph, which can be summed up: they found some tandem repeats by accident and now they want to find out how common they are.

Materials and Methods

(as usual, I suggest skipping this section unless there's some overwhelming need to read it that presents itself later)

---

[5] Stahl F (2005). Anecodotal, historical and critical commentaries on genetics Chi: A little sequence controls a big enzyme. Genetics 170:487-493.
[6] Class notes: Companion to Fuller RS et al (1984): The dnaA protein complex... Cell 38:889-900.
[7] Class notes: Computational detection of origins of replication
[8] Sorek R, Kunin V, Hugenholtz P (2008). CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Rev 6:181-186.

Results

The title of the first section sounds good (I like "Characterization", "repeated sequences", genome"). Maybe we can focus on this section to achieve our goal.

SQ3. What's your goal in reading this article?

Unfortunately, the text of this section itself is very dense. A good figure would have helped a lot, but none is present, so we're going to have to find one ourselves.

SQ4. What is the core technique used by Mazel et al?

You may not recognize it, but the key word is in the first sentence – "hybridization". What's that? With some reluctance, it's time to head to the Materials and Methods section.

SQ5. What does the Materials and Methods section say about the core technique?

Of course you blip over where they bought chemicals and enzymes, how they grew their strains, etc. Those things aren't going to help you understand the experiment. The last paragraph of the M&M section looks plausible, however... until you read it. Then you see that it's just the names of membranes, the temperature – details important for replicating the experiment but not for understanding it. The one thing you might gain from the section is the word attached to "hybridization". You learn that that word is shorthand for "Southern hybridization". Good. At least you have a handle that's Google-able. Do that, adding the word "animation". You'll find many. Choose one or more[9] and teach yourself what is Southern hybridization (also known as a Southern blot).

SQ6. What elements do you have to have in order to do a Southern blot?

SQ7. Look again at the first paragraph of the Results. Can you identify the needed elements for the blot?

SQ8. The first sentence talks about a 413-bp *Xho*I restriction fragment. If that's a typical size for what you get when you use *Xho*I to cut the *Calothrix* genome, then about how many total fragments would you expect to be produced? Bear in mind that *Calothrix* is a cyanobacterium (see Introduction). How many bands would you expect to be stained by ethidium bromide? (If you don't know what I mean by "ethidium bromide" then you may not have chosen a good animation)

SQ9. How many bands would you expect to see in an autoradiogram? Well, that depends. Depends on what?

SQ10. In this light, consider Fig. 1, the autoradiogram resulting from the Southern blot. What do you see?[10] In the first sentence of the Result section, the authors speak of a "smear". What are they referring to, and what does it signify?

OK, I'm convinced that there's something special about the 413-bp *Xba*I fragment they used as a probe, but what ***is*** that fragment? Reading further in the first paragraph, I'm directed to their Fig. 2, which gives a bit of sequence, but I want the whole thing! What is everything on that fragment?

[9] I examined only one, from McGraw Hill. It's pretty good for our purposes.

[10] I once showed a three-year old an autoradiogram of a blot and asked her what she saw. She pointed and said "Dirt!" That wasn't a bad analysis, but you should do a bit better.

At this point I would ordinarily head to BioBIKE to bring up the sequence of Calothrix PCC7601, but as it turns out, that bacterium has not been sequenced in the 24 years since Mazel et al wrote their article, and so it can't be in BioBIKE. I must be content with the DNA fragments that Mazel et al sequenced themselves. In particular, I want the sequence that contains the *Xba*I fragment from near the *cpeBA* operon. Where can I find it? The text tells me to see Table 1.

SQ11. Table 1 and its accompanying footnote gives two ways to find the sequence containing the *cpe*BA operon. What are they?

I'd much rather have the sequence in electronic form rather than printed in an article, so I prefer to get the sequence from EMBL (European Molecular Biology Laboratory) or GenBank from NCBI (the National Center for Biotechnological Information).[11] I choose NCBI. To get the sequence, go to NCBI (see the Resources & Links section of the course web site), and enter the GenBank accession number into the wide search box. Specify "Nucleotides" as the database (the menu next to the search box), and click Search. That should get you to the GenBank entry for the DNA sequence that includes the *cpeBA* operon. If you scroll down, you'll find it tells you where the various genes begin and end. Note the coordinates of *cpeB* and *cpeA*, then scroll down to the sequence itself. Copy it and paste it into your favorite word processor.

SQ12. Using the coordinates, find the beginning and end of *cpeB*. Select the letters of the gene and change the font to red. Do the same with *cpeA*. How does your annotated sequence compare with the map Mazel et al provide in their Fig. 2A?

Now we need to find the *Xba*I sites and the repeated sequences. For that, it's easier to let a computer do the searching, so bring the sequence into BioBIKE. To do this, go to any instance of BioBIKE and define a variable containing the GenBank sequence, using SEQUENCE-OF with the FROM-GENBANK and LABELED options specified. Put the accession number into the *entity* box (of course in quotes, otherwise BioBIKE will think you mean a variable), and execute DEFINE. You now have a variable containing the *cpeBA* region.

SQ13. Use the DIGESTION-OF function to digest the *cpeBA* region with *Xba*I (of course put *Xba*I in quotes, otherwise BioBIKE will think you mean a variable). From the coordinates that pop up after execution, locate the *Xba*I sites in your sequence and underline them. What sequence is recognized by *Xba*I? Find the letters of the 413-nt *Xba*I fragment and change their font to bold.

Now to find the tandem repeats. You can do this by eye, guided by Mazel et al's Fig. 2A and 2B, but again, why not give the computer a chance to help?

SQ14. Use MATCHES-OF-PATTERN to find in the *cpeBA* region (the *target*) the tandem repeats. What can you use as the *pattern* in order to find the individual instances of the repeating unit? You can figure out a pattern from Fig. 2B, or you can use the pattern provided by Mazel et al at the end of the first paragraph. They call it STRR1, What are the coordinates of the matches? Highlight STRR1 in your sequence.

SQ15. Did you find another instance of STRR1 besides the one in the 413-nt *Xba*I fragment? Highlight that as well.

[11] As it turns out, the reference they cite doesn't have the complete sequence anyway.

SQ16. Do the STRR1 instances reside inside of genes or between genes? What fraction of DNA of a typical bacterium lie between genes? How do you account for the position of STTR1?

This pretty much gives you what you want, but the pattern doesn't give only tandem repeats. It will find single instances of the repeating unit as well. There are ways of refining the pattern to filter out the single instances. Use the **Help** box to find a page that describes more about how to use patterns.

In the PDF document entitled *BioBIKE Pattern Matching*, look at the section on **Repetition Symbols**. You'll find there the "…" symbol, which asks for the maximal number of repeats. For example, the pattern "(AG)…" applied to the target "TTAGAGAGCC" will return "AGAGAG", the maximal number of repeats of "AG".

SQ17. Find in the cpeBA region the longest stretch of consecutive A's.

You probably noted a problem. The pattern "A…" returns long strings of consecutive A's, but it also returns short strings. There's a more flexible specification: *{n,m}*, *{n,}*, and *{,m}*. These give a range of desired repetitions. For example. The pattern "A{3.5}" asks for "AAA", "AAAA", or "AAAAA". The pattern "A{4,}" asks for at least 4 A's in a row.

SQ18. Use the *{n,}* specification to find the longest stretch of consecutive A's.

SQ19. Use the *{n,}* specification to find instances where the STRR1 unit is repeated at least 3 times.

This is good if you already know the pattern, but what if you don't? There's another trick. Go to the **Other special symbols** section and notice the remarkable *`n* specification (note that the first part is a downwards quote, not a straight quote. It's probably on the upper left of your keyboard). This specification means to repeat the string captured by the pattern within the first set of parentheses (`2 means to use the string captured by the second set, and so forth). You can use this to find repeated sequences even when you don't know the repeating unit. For example, the pattern "(T*)`1" applied to "TTAGCTATACTCTCAA" will find both "TATA" and "TCTC" (the star matches any letter).

SQ20. Use the `1 notation to find in "AABABCCC" all strings that are tandem repeats of a two-letter unit. It should return "ABAB" (overlapping repeats are not found).

SQ21. Use the `1 notation to find in the *cpeAB* region any tandem repeat where the repeating unit is 7 nucleotides and the length of the repeat is at least 3 units. How does this answer compare to what you got in SQ14? Why?

In short, MATCHES-OF-PATTERN can be used in at least two ways to find tandem repeats:

1. If you know the pattern, you can search for some number of tandem copies of it using the {n,} specification.
2. You can look for any tandem repeat using variations on the ( )`1 pattern.

Unfortunately, the first strategy requires that you already know the pattern, and the second strategy works only if the tandem repeat has identical units.

Is there a way of finding inexact tandem repeats when you don't know the pattern?

Indeed there is, as you'll discover in the next installment.