BioInformatics and OORexx

What is BioInformatics?

- Any coding related to biological sciences
- Statistical Functions
  - Counts, averages, percentages, chi squared
- DNA related functions
  - DNA complements
  - Identifying genes
  - Finding SNPs and other mutations
  - Measuring genetic drift
  - Protein mapping and transcription
  - Primer related functions (targets, meltpoints etc)

Molecular Biology (crash course)

- DNA provides the basic code used to assemble proteins. It's all about proteins.
- Double helix

Data is encoded as "base pairs" – the rungs of the ladder

Bases are the 4 amino acids –

Adenine (bonds only with Tyrosine)

Tyrosine (bonds only with Adenine)

Cytosine (bonds only with Guanine)

Guanine (bonds only with Cytosine)

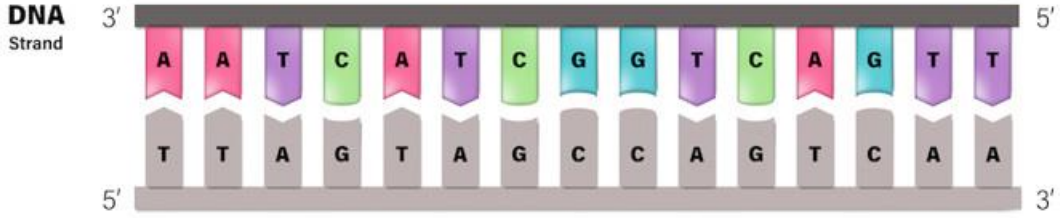
Molecular Biology - Enzymes

- To read the DNA, the helix must first be split so the bases are exposed.
- DNA is a large molecule. An enzyme is used to split it. The particular enzyme used is specific to this task (helicase).
- DNA is really, really long when split. The fastest method determined to read the data was to use more enzymes to cut the DNA into shorter lengths. This was called “shotgun” sequencing.
- These “short reads” then had to be logically reassembled using overlapping sections on each end. The problem then became computational.

Molecular Biology – Base Pairing

Once split, we can display a single strand:

It is displayed from 3’ (“3 prime”) to 5’ (“5 prime”). The cellular machinery (and the sequencers) read from 5’ to 3’. Because the DNA backbone is not symmetrical, the other side (known as the complement) will be 5’ to 3’



Molecular Biology – DNA structure

- Determining the complement of a DNA strand we must first work out the base's logical complement then reverse the resultant string to ensure the 3' – 5' orientation is preserved.

- Why does this matter? If we are looking at the output of a sequencer, it will have both the original strand and its complement. As we only note the original strand, we need to eliminate the "mirror image" complement. If searching for primer target sites, we must test for the complement also using "mirror" primers. With damaged DNA (eg ancient DNA), we might only recover the complement accurately.

Molecular Biology – DNA Data

BioInformatics includes taking the short reads and assembling them into a single DNA string. It is normally 80 bytes long per line and will have as many lines as required, with the last line only as long as required.

Most people start with blocks of DNA that have already been merged into a single block. When reading the files, each line will need to be concatenated into a single 3’ to 5’ string.

Rexx is very good with strings, both in size (storage) and functions

Molecular Biology – Sample DNA Data

Sample DNA data

>MT251972.1:25356-26183 ORF3a protein [organism=Severe acute respiratory syndrome coronavirus 2] [isolate=SARS-CoV-2/human/USA/WA-UW236/2020]

ATGGATTTGTTTATGAGAATCTTCACAATTGGAACTGTAACTTTGAAGCAAGGTGAAATCAAGGATGCTA

CTCCTTCAGATTTTGTTCGCGCTACTGCAACGATACCGATACAAGCCTCACTCCCTTTCGGATGGCTTAT

TGTTGGCGTTGCACTTCTTGCTGTTTTTCAGAGCGCTTCCAAAATCATAACCCTCAAAAAGAGATGGCAA

CTAGCACTCTCCAAGGGTGTTCACTTTGTTTGCAACTTGCTGTTGTTGTTTGTAACAGTTTACTCACACC

TTTTGCTCGTTGCTGCTGGCCTTGAAGCCCCTTTTCTCTATCTTTATGCTTTAGTCTACTTCTTGCAGAG

TATAAACTTTGTAAGAATAATAATGAGGCTTTGGCTTTGCTGGAAATGCCGTTCCAAAAACCCATTACTT

TATGATGCCAACTATTTTCTTTGCTGGCATACTAATTGTTACGACTATTGTATACCTTACAATAGTGTAA

CTTCTTCAATTGTCATTACTTCAGGTGATGGCACAACAAGTCCTATTTCTGAACATGACTACCAGATTGG

TGGTTATACTGAAAAATGGGAATCTGGAGTAAAAGACTGTGTTGTATTACACAGTTACTTCACTTCAGAC

TATTACCAGCTGTACTCAACTCAATTGAGTACAGACACTGGTGTTGAACATGTTACCTTCTTCATCTACA

ATAAAATTGTTGATGAGCCTGAAGAACATGTCCAAATTCACACAATCGACGGTTCATCCGGAGTTGTTAA

TCCAGTAATGGAACCAATTTATGATGAACCGACGACGACTACTAGCGTGCCTTTGTAA

Molecular Biology – Codons and Amino Acids

- Amino acids are the building blocks of proteins
- There are about 500 known amino acids
- Only 20 are normally used by humans (actually 22 in total)
- 3 bases can be used to map to each amino acid. These 3 bases are called a ***codon***. The codons are "aligned" like word alignment on zOS. The word length in this case is 3 bases. This is called its ***frame***
- From the next table you will see that whilst 3 amino acids define a codon, we cannot uniquely tell the amino acids in the codon.

Molecular Biology - Transcription

Note:

AUG

UAA

UAG

UGA

FIRST LETTER	SECOND LETTER: U	SECOND LETTER: C	SECOND LETTER: A	SECOND LETTER: G	THIRD LETTER
U	UUU Phenylalanine (F) UUC UUA Leucine (L) UUG	UCU UCC Serine (S) UCA UCG	UAU Tyrosine (Y) UAC **UAA** stop codon **UAG** stop codon	UGU Cysteine (C) UGC **UGA** stop codon UGG Tryptophan (W)	U C A G
C	CUU CUC Leucine (L) CUA CUG	CCU CCC Proline (P) CCA CCG	CAU Histidine (H) CAC CAA Glutamine CAG (Q)	CGU CGC Arginine (R) CGA CGG	U C A G
A	AUU AUC Isoleucine (I) AUA ***AUG*** start codon* (M)	ACU ACC Threonine ACA (T) ACG	AAU Asparagine AAC (N) AAA Lysine (K) AAG	AGU Serine (S) AGC AGA Arginine (R) AGG	U C A G
G	GUU GUC Valine (V) GUA GUG	GCU GCC Alanine (A) GCA GCG	GAU Aspartic acid GAC (D) GAA Glutamic acid GAG (E)	GGU GGC Glycine (G) GGA GGG	U C A G

* The start codon encodes the amino acid methionine

RNA is used for transcription and uses U instead of T

Molecular Biology – Protein Assembly

DNA is restricted to the cell nucleus. For a protein to be assembled enzymes copy the relevant code from the DNA into an RNA template, which is carried out of the nucleus and then assembles the protein in the cell.

RNA can reuse portions of the same DNA code. Part of the protein can be assembled from codons using the standard alignment, then restart reading from a different point, shifting the codon frame by one or two base pairs and completing the protein using the codons described using the changed frame alignment

Molecular Biology – Finding Genes

C-G bonds are stronger than A-T bonds and DNA has evolved to put blocks of C-G bonds at the end of each DNA string. These blocks are called telomeres and are at the ends of each DNA strand to prevent it from fraying and unravelling. It is like the caps on shoelaces. Some DNA is circular and avoids the problem by having no ends.

When looking at DNA, we normally ignore these blocks. Somewhere after the C-G block we should be able to find a ***start*** codon (AUG). This indicates the start of a gene. The gene goes until a stop codon is encountered (within the current frame). There will be a stretch of non-coding bases, until another gene start codon is found

Molecular Biology – Mutations and Errors

- SNPs (pronounced “snips”) are Single Nucleotide Polymorphisms. This just means a nucleotide (a single base) has an unexpected value.
- Deletions. Sometimes a gene will be missing a base or even whole sections. Deletion or mutation of a start codon deletes the gene.
- Insertions. Sometimes a single base, or even whole portions of DNA can be inserted. The effect on the organism is unpredictable.
- Damaged or partial DNA. Some DNA may provide ambiguous or very poor quality reads. This can sometimes be “logically” corrected.
- Patterns of mutations can indicate various evolutionary trends.

Molecular Biology - PCR

The Polymerase Chain Reaction (PCR)

DNA samples are typically small. In addition, the sample molecules will be unique and complex. Until PCR was invented, it was nearly impossible to determine the DNA code.

With the correct polymerase and a soup of ATCG (as raw materials), it is possible to “amplify” the DNA. Many copies will be made and a bigger sample can be sequenced. Sequencing requires duplicates to be removed, but the duplicates are first used to statistically pick the base.

BioInformatics – A Rexx Programmer's Dream

From what you have seen so far, you will have realised that classic Rexx has a number of functions ready-made for handling genetic data.

Pos()	- for finding sequences of bases
Translate()	- creating complements, isolating bases
Length()	- counting some or all bases
Reverse()	- creating complements

Rexx is also comfortable using very large strings.

BioInformatics – OORexx Advantages

With Rexx, a whole genome can be treated as a single string. Under OORexx the genome can be treated as an Object and the methods applied to the genome.

A standard toolset of classes and methods can be developed

Collection classes. Stems are particularly useful.

Performance. Can be compiled for even better performance.

BioInformatics – OORexx Advantages

Working with large DNA datasets can be heavy on I/O, cpu and memory. Many applications would benefit from distributed processing.

OORexx allows multi-threading

Potential for distributing processing over multiple machines. Features of OORexx would facilitate this.

BioInformatics – File Formats

FASTA is an encoding method that can store either the actual base or the codon, all using alphabetic characters (upper case). From a programmer's point of view it could probably be better, but it is widely used. You will need code to translate between DNA and FASTA (Walter highlighted one during his *Rosetta Pearls* talk).

Fna type files have a header line and then the raw DNA in 80 byte lines. Multiple DNA sequences may be in the one file.

Others are used, both common (eg txt, xml) and custom. Net search.

BioInformatics - Samples

Verify the DNA string only contains valid bases:

```
::method ValidDNA
use arg dna
bases = Space(Translate(dna, '    ','ATCG'), 0)
If Length(bases) > 0 Then Return -1
Return 0
```

BioInformatics - Samples

Complement the DNA:

```
::method Complement
use arg dna
Return Reverse(Translate(dna, ‘TAGC‘,’ATCG’))
```

BioInformatics - Samples

Count “G” bases in the DNA:

```
::method CountG
use arg dna
Return Length(Space(Translate(dna, ‘   ‘,’ATC’), 0))
```

BioInformatics - Samples

Count *any* bases in the DNA:

```
::method CountBases
use arg dna, bases
mask = Space(Translate('ATCG','    ',bases),0)
Return Length(Space(Translate(dna, '   ',mask), 0))
```

BioInformatics - Applications

1. Collect a sample of water (from pond, lake, aquarium, pool)
2. Sequence the water (biome sequencing) – all organisms sampled
3. Identify organism of first fragment (manual search of database)
4. Download that organism’s genome
5. Eliminate all known fragments for that organism
6. Check next unknown fragment and repeat steps 3 to 5
7. Build a list of organisms
8. Bill client

BioInformatics - Applications

Check all isolates for primer sites

- Check for any new isolates
- Download new isolate
- Search isolate for desired target site
- Report success/fail

BioInformatics - Applications

Primer design and selection. Check for number and types of dimers for a given primer

Single stranded DNA (eg a primer) is like a wet noodle with Velcro patches spaced along one side. It can adhere to itself (very messy)

Primers that dimer with themselves waste primer and reduce effect

Brute force determine every permutation of primer-primer binding

Report number and type of dimers possible.

BioInformatics - Samples

Calculating Melt Points

A polymerase allows the base pair bonds to be broken at certain temperatures. Measurements have shown the exact energy required to break specific bonds. It depends upon both the base pair *and* the pair next to it.

Using the primer DNA and its target site it is possible to determine the Celsius temperate required to “melt” the DNA.

BioInformatics – Coding Exercises

Introduction to BioInformatics

https://rosalind.info/

This site explains the reason for much of the code required and sets a number of exercises for you to test your algorithms.

BioInformatics – Downloading Genomes

Unless you are accessing data directly from a sequencer, you will want to download genomes etc.

National Center for BioTechnology Information (NCBI) holds over 1,000 genomes and descriptions. These also contain notes and generally a gene map.

https://www.ncbi.nlm.nih.gov/genome

Note: This is really difficult to navigate, but probably easy by algorithm