Bioinformatics Resources - Introduction to Biological Databases

Francesco Penasa

March 12, 2020

`2020 03 12` Amino acid sequences (DNA/RNA), protein sequences and more complex data stored efficiently in databases to be accessible by who needs them to analyze data and generate new knowledge.

1. Primary databases: sequences of nucleotides and aminoacids
2. Derived and specialized databases (protein domains, structures, genes...)

Information gathered through literature, lab analyses and bioinformatics analyses. Each database is characterized by a central biological element. In primary data banks **each element is uniquely identified** by an accession number.

1. Sequences of nucleotides are represented by **4 characters** strings.
2. Sequences of amino acids are represented by **20 characters** strings.

$ACGT = DNA$ $ACGU = RNA$ amino acids are represented by a triplet of DNA alphabet, mapped into a new character.

Primary datatbases data format differs

1. GenBank: just store and arachive only nucleotide sequences. Every GenBank record is identified uniquely by the **ACCESSION and VERSION** codes.
2. EMBL datalibrary
3. DDBJ

Derived data banks

1. RefSeq: curated and not redundant collection of DNA, RNA and protein sequences.

All NCBI databases are accessible throught a nuque search engine called EWntrez `https://www.ncb`

Genome browser Allow us to browese data at various detail levels.

Reading the gene structure

1. Horizontal line with arrows : **INTRON**
2. Dark block: **EXON**
3. Thick and arrowless lines: **UTR**

1 Exercises

1. Display the RARA gene with the browser. How many alternative transcripts does it have? **2**
2. Considering the first transcript, how many introns and exons? **8 ; 9**
3. Now add to the displayed tracks the GC-percent track (it shows the percentage of GC bases along the sequence). Drag it and move it just under the sequence. **Mapping and Sequencing GC percent dense**
4. Now hide the alignment track which is shown by default. **boh**
5. Zoom out on the 5'UTR region of the transcript and check if there are any known SNPs in that region.

1. We want to retrieve the 3'UTR sequence for our gene, RARA
2. `http://genome.ucsc.edu/cgi-bin/hgTables`
3. identifiers paste list -¿ RARA
4. output format -¿ sequence, genomic
5. change some things and submit

1. Display the KRAS gene with the browser. How many alternative transcripts does it have ? Considering the first transcript, how many introns and exons? 4; 5; 6
2. Now add the 1000 Genomes – EUR common variants track to the display. Are there 5'UTR variants in KRAS?
3. Now zoom to the second exon of the first isoform of KRAS. How many missense variants are there ?

1. Get the Ensembl ID and MGI symbol for all mouse genes on chromosome 19. How many genes are there ?
2. Now limit this to protein coding genes. How many genes are there ?
3. Now save the results to a comma-separated values fil