Topic 4: Bioinformatic tools for evaluation of genomic variability of AnGR
Lecture

The topic of this lecture is Bioinformatics tools for assessing genome variability of animal genetic resources. The lecture is part of module 2, Conservation and sustainable utilization of animal genetic resources. The creation of this presentation was supported by the ERASMUS+ KA2 grant as part of the project ISAGREED, Innovation of content and structure of study programs in the management of animal genetic and food resources using digitization.

What is bioinformatics and what is its importance? Bioinformatics is an interdisciplinary field that combines biology, computer science, statistics, and mathematics to analyze and interpret biological data using computational tools and algorithms. In animal genetics, it plays a crucial role in obtaining and analyzing extensive genomic data, such as DNA sequences, to gain insights into the genetic makeup and biological processes of animals. In animal genetics, bioinformatics helps identify and characterize genes responsible for specific traits, diseases, or abnormalities in animals. For example, it can help identify genes associated with fur color, milk production, growth rate, or susceptibility to diseases.

It allows researchers to compare and analyze genomes of different animal species, understand their evolutionary relationships, and identify common genetic elements. This can provide insights into the diversity and evolution of animal species. Bioinformatics helps in developing new breeding strategies and improving breeding practices. By analyzing genetic data, it helps identify animals with desirable traits for breeding programs, improve traits such as productivity, disease resistance, or adaptability. Bioinformatics supports the protection and conservation of endangered animal species by studying their genomes and identifying genetic markers for monitoring populations, assessing genetic diversity, and assisting in captive breeding programs.

How are the data used? For bioinformatic analyses, internet connection, computer, programs, and own data are needed. Data such as DNA, RNA, or protein sequences are used. Also genomic, transcriptomic, proteomic, metabolomic, phylogenetic, or structural data. Data for bioinformatic analyses are initially uploaded to online databases. There were about 2000 databases available online in January 2024. The most significant sequence databases are GenBank, ENA, Uniprot, and the genomic database Ensembl.

On servers where the databases are located, there are tools for searching, aligning, and analyzing bioinformatic data. Pairwise sequence alignment is used to identify regions of similarity that may indicate functional, structural, and/or evolutionary relationships between two biological sequences (proteins or nucleic acids).

Multiple sequence alignment (MSA) is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between sequences can be studied.

The basic tool for aligning two sequences is BLAST, Basic Local Alignment Search Tool - on NCBI server. BLAST searches for areas of similarity between biological sequences. The program compares nucleotide or protein sequences with sequence databases and calculates statistical significance.

It is possible to compare your own sequence with database sequences (in GenBank). It is possible to compare specific two sequences. The primary focus is on local alignment (also available for global alignment)

Another tool for pairwise sequence alignment on the European EMBL-EBI server is EMBOS.

When to use local or global alignment? Local (using Smith-Waterman algorithm) is used for more different, evolutionarily distant sequences; it is limited to assigning unique segments and stops where the sequences diverge significantly.

Global alignment (using Needleman-Wunsch algorithm) is the most suitable for sequences that are similar and approximately the same length; attempt to align sequences over their entire length even at the cost of introducing gaps into one or both sequences.

We will demonstrate a model alignment process. We want to determine which two sequences A and B or C and D are more similar to each other? Align the sequences over their entire length, i.e. write them in two rows placed below each other so that identical positions (bases or amino acids) are aligned. Each pair and non-pair will be assigned a value, for example, 1 for a match and 0 for mismatch. Both alignments show that the first pair of sequences A and B have 8 match and 2 mismatch, and the second pair of sequences C and D have 17 match and 3 mismatch. However, which pair of sequences is more similar?

It is necessary to calculate the normalized similarity values (score). We can compare the similarity of pairs of sequences of different lengths. Multiply the number of matches by their value (1) and add to it the number of mismatches multiplied by their value (0). The normalized score is determined by dividing the calculated value by the length of the alignment. In our case, the alignment of sequences C and D has a higher score, so they are more similar.

In another example, we align two sequences of different lengths. If we determine the score for unaligned sequences, it would have a value of 6. After alignment, the score increased to 9. The score increases by inserting gaps. Gaps increase the number of aligned identical residues.

There are many online tools for multiple sequence alignment (MSA). Among the oldest is Clustal, but others have been gradually developed such as MAFFT, T-Coffee, MUSCLE, Kalign, or COBALT. Each was developed for different types of sequences and their lengths.

The principle of MSA, aligning 3 or more sequences, is similar to that of BLAST, based on pairwise alignments. However, the calculations are more complex. This can reveal mutations, substitutions, or insertions/deletions. These comparisons are used to derive evolutionary relationships through phylogenetic analysis and can highlight homologous features between sequences. The result may be a phylogenetic tree expressing evolutionary distances between sequences.

Thank you for your attention.