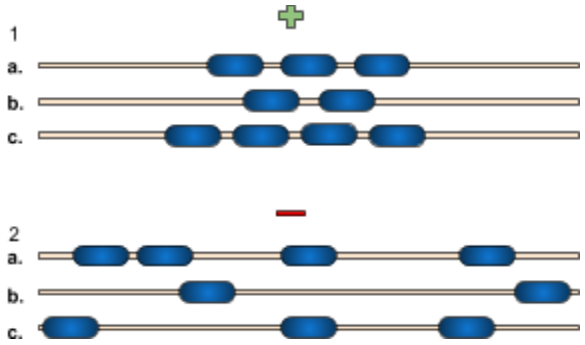
# ASSIGNMENT ON GENOMIC SEQUENCE

- Use machine learning for classifying the sequences with motif clusters (see pic below). This is a binary classification task on motif grammar.

1. 5 text files with a total of 10,000 sequences (ATGCTGA....) with labels (except for test set) split into 3 sets are provided. The positive class refers to the sequences that contain multiple instances of motifs in a specific region. The negative class corresponds to the sequences that contain the motifs anywhere.
   1.1. For simplicity, only one motif is present within and across the sequence(s).




Some examples of the positive & negative class. [not to scale]
1(a-c): Instances of the multiple motif occurring in a specific region 2: Instances of motifs found anywhere on the sequence.

2. Deliverables
   2.1. You are required to submit a report. Your report should cover the following:
      2.1.1. Methodology, implementation details - Number of parameters, hyperparameters, model summary
      2.1.2. Performance in terms of accuracy, F1 & AUPRC.
      2.1.3. Describe in brief your experiments, the rationale for design decisions and observations.
      2.1.4. The report must contain the training curve (loss vs epochs)
      2.1.5. Information about motif, distance between the instances and specific region of clustering (interpretation/visualization)
   2.2. Predictions for the test set. The predictions should be submitted as a text file in the same format as the labels.
   2.3. Submission of the report is due by 4th March.

3. How to submit:
   Your code, text file for prediction and report (rollNum.pdf) must be submitted in a zipped folder (rollNum.zip).

4. Notes:
   4.1. The best place to begin would be using a 1D CNN as discussed in the class. You are encouraged to read and try different approaches. For example alternate encoding of the sequence data, different architectures as discussed in the class.
   4.2. You may use set 2 for validation or merge it & use k fold cross-validation.
   4.3. The objective of this assignment is **not to get the best performance** but explore & experiment with the sequence data.
   4.4. **Do NOT plagiarise. Plagiarism of code or report will be rewarded with 0. Renaming the variables or changes in ordering will not help.**
   4.5. Please contact Ms Ruchi Chauhan (email ID- ruchi.chauhan@research.iiit.ac.in) for your queries regarding the assingment.