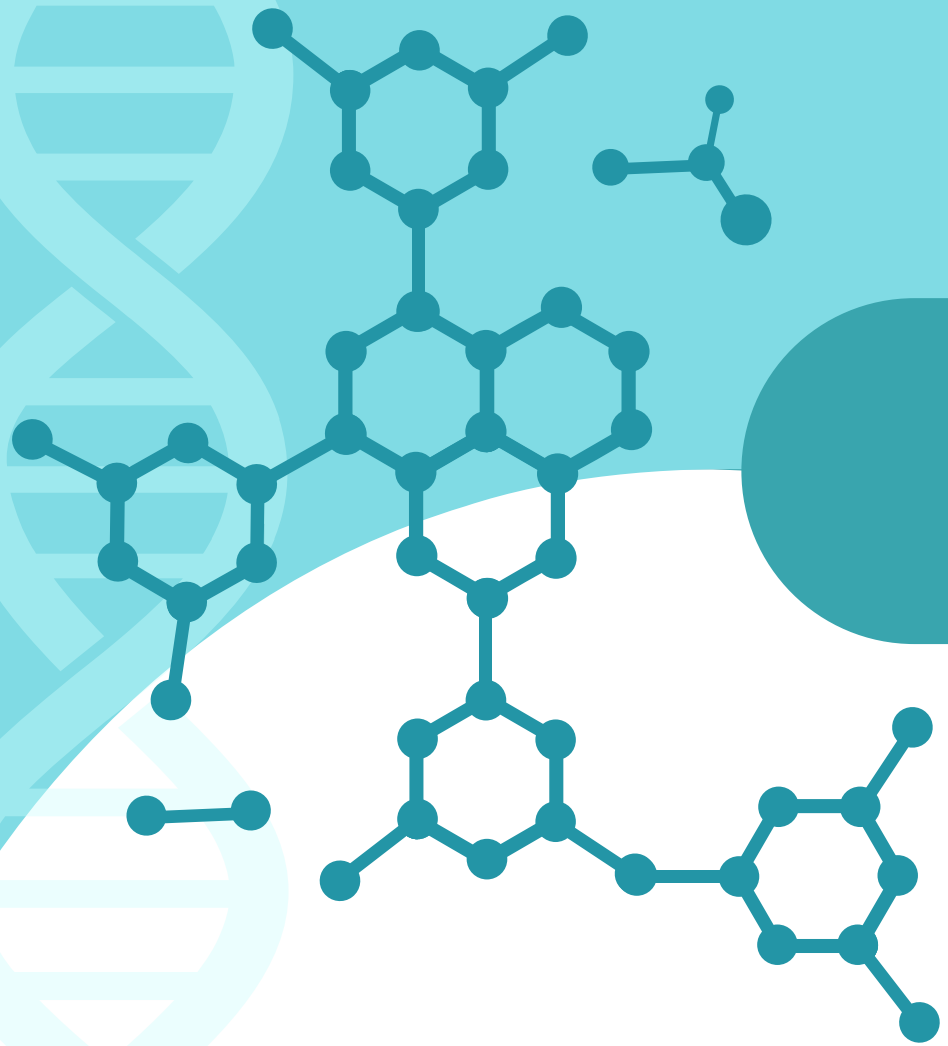
Bioinformatics LAB 4

Gene Fusions

Prof.ssa Elisa Ficarra
Prof.ssa Santa Di Cataldo
Eng. Marta Lovino
Eng. Alessio Mascolini

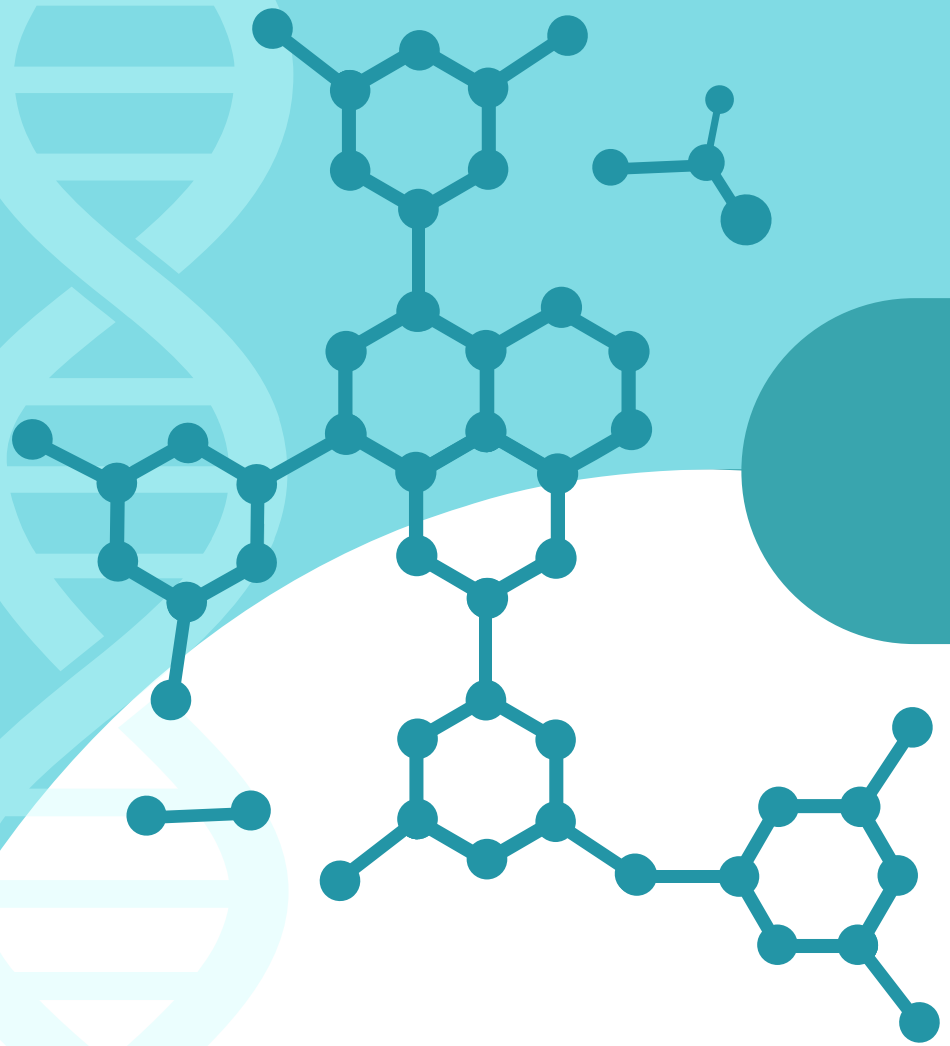
Politecnico di Torino
DAUIN
Dept. of Control and Computer Engineering



LAB 4 - Objectives

Objectives

- Understand how gene fusions work
- Build the gene fusion sequence starting from its breakpoints
- (Advanced, for the next labs): mono-dimensional CNN for gene fusions prioritization



LAB 3 - Assignments

Assignment 1: Build the gene fusion sequences

Build the gene fusion sequences starting from the `breakpoints.txt` file and the reference genome on chromosomes 10 and 18:

`breakpoints.txt`

Gene5p	chr5p	breakpoint5p	strand5p	Gene3p	chr3p	breakpoint3p	strand3p
ABL1	10	15566488	+	PACX3	18	8666148	+
PPCHS	18	747292	+	TTTCS	10	7393	-
CSSP	10	9999845	-	PPAJD	10	6628	-

Assignment 2: Perform gene fusion prioritization using CNN

Use the `training_set.csv` and `test_set_1.csv` files to train and test a 1-D CNN model to recognize oncogenic and not oncogenic gene fusions. The files are uploaded on the Teaching Portal. Please refer to the recording for all the details.

	FusionPair	Label	Version	Chr5p	Coord5p	5pStrand	5pCommonName	5pEnsg	5pGeneFunctionality	5pGeneDescription
	Chr3p	Coord3p	3pStrand	3pCommonName	3pEnsg	3pGeneFunctionality	3pGeneDescription	MainProteins	Proteins	
0	CSNK2B_NDUFA6	1	grch37	6	31637695	+	CSNK2B	ENSG00000204435	protein_coding	casein kinase 2, beta polypeptide [Source:HGNC Symbol;Acc:2460]
	22	42486683	-	NDUFA6	ENSG00000184983	protein_coding	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa [Source:HGNC Symbol;Acc:7690]			

['MEKCKGTSRMAGTTSADVKMSSSEEVSWISWFCGLRGNEFFCEVDEDYIQDKFNLTGLNEQVPHYRQALDMILDLEPDEELEDNPNQSDLIEQAAEMLYGLIHARYILTNRGIAQMLEKYQQGDFGYCPRVYCENQPMLPIGLSDIPGEAMVKLYCPKCMDVYTPKSSRHHHTDGAYFGTGFPHMLFMVHPEYRPKRPANQFVPRLYGFKIHPMAYQLQLQAASNFKSPVKTIGT']

['MEKCKGTSRMAGTTSADVKMSSSEEVSWISWFCGLRGNEFFCEVDEDYIQDKFNLTGLNEQVPHYRQALDMILDLEPDEELEDNPNQSDLIEQAAEMLYGLIHARYILTNRGIAQMLEKYQQGDFGYCPRVYCENQPMLPIGLSDIPGEAMVKLYCPKCMDVYTPKSSRHHHTDGAYFGTGFPHMLFMVHPEYRPKRPANQFVPRLYGFKIHPMAYQLQLQAASNFKSPVKTIGT',
'MEKCKGTSRMAGTTSADVKMSSSEEVSWISWFCGLRGNEFFCEVDEDYIQDKFNLTGLNEQVPHYRQALDMILDLEPDEELEDNPNQSDLIEQAAEMLYGLIHARYILTNRGIAQMLEKYQQGDFGYCPRVYCENQPMLPIGLSDIPGEAMVKLYCPKCMDVYTPKSSRHHHTDGAYFGTGFPHMLFMVHPEYRPKRPANQFVPRLYGFKIHPMAYQLQLQAASNFKSPVKTNGTG',
'MSSSEEVSWISWFCGLRGNEFFCEVDEDYIQDKFNLTGLNEQVPHYRQALDMILDLEPDEELEDNPNQSDLIEQAAEMLYGLIHARYILTNRGIAQMLEKYQQGDFGYCPRVYCENQPMLPIGLSDIPGEAMVKLYCPKCMDVYTPKSSRHHHTDGAYFGTGFPHMLFMVHPEYRPKRPANQFVPRLYGFKIHPMAYQLQLQAASNFKSPVKTIGT',
'MSSSEEVSWISWFCGLRGNEFFCEVDEDYIQDKFNLTGLNEQVPHYRQALDMILDLEPDEELEDNPNQSDLIEQAAEMLYGLIHARYILTNRGIAQMLEKYQQGDFGYCPRVYCENQPMLPIGLSDIPGEAMVKLYCPKCMDVYTPKSSRHHHTDGAYFGTGFPHMLFMVHPEYRPKRPANQFVPRLYGFKIHPMAYQLQLQAASNFKSPVKTNGTG']

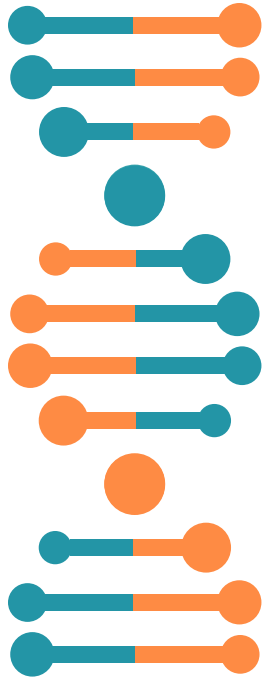
1	IFT57_CALM1	1	grch37	3	107884314	-	IFT57	ENSG00000114446	protein_coding	intraflagellar transport 57 homolog (Chlamydomonas) [Source:HGNC Symbol;Acc:17367]
	14	90871138	+	CALM1	ENSG00000198668	protein_coding	calmodulin 1 (phosphorylase kinase, delta) [Source:HGNC Symbol;Acc:1442]			

['MTAALAVVTTSGLEDGVPRSRGEGTGEVVLERGPGAAYHMFVVMEDLVEKLKLLRYEEEFLRKSNLKAPSRHYFALPTNPGEQFYMFCTLAAWLINKAGRPFEQPQEYDDPNATISNILSELRSFGRTADFPPSKLKSGYGEHVCYVLDCFAEEALKYIGFTWKRPIYPVEELEEESVAEDDAELTLNKVDEEFVEEETDNEENFIDLNVLKAQTYHLDMNETAKQEDILESTTDAAEWSLEVERVLPQLKVTIRTDNKDWRIHVDQMHQHRSGIESALKETKGFLDKLHNEITRTLEKISSREKYINNQLENLVQEYRAAQAQLSEAKERYQQGNL']

['MTAALAVVTTSGLEDGVPRSRGEGTGEVVLERGPGAAYHMFVVMEDLVEKLKLLRYEEEFLRKSNLKAPSRHYFALPTNPGEQFYMFCTLAAWLINKAGRPFEQPQEYDDPNATISNILSELRSFGRTADFPPSKLKSGYGEHVCYVLDCFAEEALKYIGFTWKRPIYPVEELEEESVAEDDAELTLNKVDEEFVEEETDNEENFIDLNVLKAQTYHLDMNETAKQEDILESTTDAAEWSLEVERVLPQLKVTIRTDNKDWRIHVDQMHQHRSGIESALKETKGFLDKLHNEITRTLEKISSREKYINNQLENLVQEYRAAQAQLSEAKERYQQGNL',
'MTAALAVVTTSGLEDGVPRSRGEGTGEVVLERGPGAAYHMFVVMEDLVEKLKLLRYEEEFLRKSNLKAPSRHYFALPTNPGEQFYMFCTLAAWLINKAGRPFEQPQEYDDPNATISNILSELRSFGRTADFPPSKLKSGYGEHVCYVLDCFAEEALKYIGFTWKRPIYPVEELEEESVAEDDAELTLNKVDEEFVEEETDNEENFIDLNVLKAQTYHLDMNETAKQEDILESTTDAAEWSLEVERVLPQLKVTIRTDNKDWRIHVDQMHQHRSGIESALKETKGFLDKLHNEITRTLEKISSREKYINNQLENLVQEYRAAQAQLSEAKERYQQGNL',
'MTAALAVVTTSGLEDGVPRSRGEGTGEVVLERGPGAAYHMFVVMEDLVEKLKLLRYEEEFLRKSNLKAPSRWVPASPSARVPASPQGDALLGPAGNQFRESGHYFALPTNPGEQFYMFCTLAAWLINKAGRPFEQPQEYDDPNATISNILSELRSFV',
'MTAALAVVTTSGLEDGVPRSRGEGTGEVVLERGPGAAYHMFVVMEDLVEKLKLLRYEEEFLRKSNLKAPSRWVPASPSARVPASPQGDALLGPAGNQFRESGHYFALPTNPGEQFYMFCTLAAWLINKAGRPFEQPQEYDDPNATISNILSELRSFVS']

LAB4 – Take home message

- The most common configuration for a gene fusion is “promoter-end”, although multiple configurations are possible (refer to the recording).
- Classify gene fusions is usually really challenging: an algorithm that works well on a training set can fail con the test set.



Questions?

Remember:
no question is
stupid