

Dataset	Number	#positives	Alphabet size	Average length
Protein	3,238	96	20	607
DNA	3,238	96	4	1,827
Music	10,261	9,022	61	329
Sports	296,337	253,017	63	307
Compound	1,367,074	57,536	44	53