SUPPLEMENTARY MATERIAL 1: Datasets

Bioinformatics: Application Note

Dragon PolyA Spotter: Predictor of poly(A) motifs within human genomic DNA sequences

Manal Kalkatawi[1,#], Farania Rangkuti[1,#], Michael Schramm[1,#], Boris R. Jankovic[1,#], Allan Kamau[1], Rajesh Chowdhary[2], John A.C. Archer[1], Vladimir B. Bajic[1,*]

[1] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

[2] Biomedical Informatics Research Center, MCRF, Marshfield Clinic, 1000 North Oak Avenue, Marshfield, WI 54449, USA

We used human mRNA sequences from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/ as of 30/6/2011 and selected from these only the polyadenylated ones. From the polyA tail we used the most 5' four A nucleotides and extended these by 96 nucleotides upstream. We mapped these 100 nucleotides fragments back to the human genome version hg19 using consensus CDS information version 37.2 and applying stringent BLASTN mapping criteria (e-value 10e-20 or better, and matching sequence of at least 96 nucleotides in length from the 3' end of the matched portion). Negative records were selected from human chromosome 21 randomly. Within candidate sequences so obtained, we selected those where the poly(A) motif (any of the 12 variants we considered) is found at locations conforming to the positional distributions reported in (Beaudoing et al., 2000). In the case of multiple motifs found, we selected the one that was most close to the most frequent position as reported in (Beaudoing et al., 2000). We then flanked such poly(A) motifs by 100 nucleotides upstream and 100 nucleotides downstream, resulting in sequences of total length of 206. This process resulted in 14799 sequences for 12 motif variants. Number of positive samples (we generated an equal number of negative samples) for each of the poly(A) motifs are given in Supplementary Table 1. All sequences that we used can be downloaded from http://cbrc.kaust.edu.sa/dps/code/DataToBuildModel.tar.gz.

Supplementary Table 1: Summary of data used for training and testing of our prediction model

Number of positive samples	Poly(A) motif variant
1402	AAAAAG
1391	AAGAAA
5164	AATAAA
724	AATACA
332	AATAGA
392	AATATA
553	ACTAAA
697	AGTAAA
2433	ATTAAA
474	CATAAA
481	GATAAA
755	TATAAA

Supplementary References

Emmanuel Beaudoing, Susan Freier, Jacqueline R. Wyatt, Jean-Michel Claverie, and Daniel Gautheret (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Res.* 10: 1001-1010.