NonPEST: a nonparametric method for Transcription Start Site prediction

Tatiana Tatarinova,
Laboratory of Applied Pharmacokinetics, University of Southern California, Los Angeles, USA,
Tatiana.tatarinova@lapk.org
Alyona Chubatiuk
Department of Mathematics, University of Southern California, Los Angeles, California, USA, achubatiu@usc.edu
Michael Neely
Laboratory of Applied Pharmacokinetics, University of Southern California, Los Angeles, USA, mneely@usc.edu
Alan Schumitzky
Department of Mathematics, University of Southern California, Los Angeles, California, USA
schum@usc.edu

We present NonPEST - a novel tool for analysis of EST distributions and Transcription Start Site (TSS) prediction. The method combines two nonparametric methods of estimation of unknown probability distribution using Maximum Likelihood and Bayesian approaches. Accurate identification of TSS is an important genomics task, since position of regulatory elements with respect to the TSS affects gene regulation, and performance of promoter motif-finding methods depends on correct identification of TSSs. Our probabilistic approach expands recognition capabilities to multiple TSS per locus and enhances the understanding of alternative splicing mechanisms.

Positions of 5' EST can be considered as noisy experimental evidences of the location of TSS. If the total number of ESTs for a given locus is N, any upstream position can have from *0* to *N* ESTs mapped to it. In case when all *N* ESTs are mapped to the same position, we have a single reliable prediction of the TSS. Other cases are more complex. Since each locus may have one or more real TSS, we have a mixture model with unknown number of components, corresponding to an unknown number of TSS per locus. We used well annotated genome of *A. thaliana,* whose loci may have thousands of ESTs per locus mapped to any given promoter region. The true positions of the TSS are determined by an unknown parameter θ. The task is to determine the probability distribution of θ based on the distribution positions of 5' ESTs on the genome. Nonparametric maximum likelihood (NPML) and nonparametric Bayesian (NPB) are used for this task. Validity and utility of NPB and NPML methods for nonlinear mixture models in pharmacokinetics has been demonstrated by us earlier [1].

We assumed $Y_i \sim p_i(Y_i \mid \theta_i),\ i = 1,...,N$, $\theta_i \sim F$, $F \sim DP(G_0)$. We use Binomial distribution $P(y|n,p) = \binom{n}{y} p^y (1-p)^{n-y}$, where n is the length of the upstream region, *N* is the number of ESTs corresponding to a given locus, and *p* is the probability to extend sequence

by one nucleotide. From TAIR we obtained genome annotation files and upstream sequences. The sequences were truncated based on the position of the nearest upstream locus. 234,213 ESTs were downloaded from NCBI and mapped onto the upstream sequences. NonPEST was able to predict positions of TSS for 13,208 loci.

We compared performance of NonPEST and TAIR annotations. The number of predicted TSS ranged from 1 to 15, with 63% of analyzed loci had one TSS predicted, 22% had two TSSs, 8% had three TSSs. Pearson's correlation coefficient between the number of ESTs per locus and the number of predicted TSS is 0.037, hence there is no significant relationship between the two values. Presence of multiple TSSs per locus has a weak association with alternative splicing: those loci that have open reading frame (ORF) predicted, have 37% chance of multiple TSS, as compared to 45% chance for those with multiple ORFs.

We used known statistical features of core promoter regions to compare performance of TSS-prediction methods. TATA box, located around position -30 from the TSS is the most over-represented sequence pattern [2]. Another conserved feature is initiator (Inr), located at TSS, commonly containing dinucleotide sequence of CA [4]. We compared frequencies of canonical TATA-box 4-nucleotide sequence (TATA) in promoters predicted by TAIR and by NonPEST: 30% of TAIR-predicted promoters contain TATA at positions -40..-20, as compared to 38% of NonPEST-predicted promoters. Counting less common forms of the TATA-box (such as TAAA and CTAT) are also more prevalent in NonPEST-predicted promoters (61% vs. 55%). There is stronger nucleotide consensus at TSS (44% of C and 35% of T followed by 63% of A) for NonPEST then for TAIR (35% of C and 39% of T followed by 53% of A).

The database of predicted promoters is available at http://www.glacombio.net.

1. TATARINOVA, T. et al. Two General Methods for Population Pharmacokinetic Modeling: Non-Parametric Adaptive Grid and Non-Parametric Bayesian. **JPP**, 2013.

2. SMALE, S. Core promoters: active contributors to combinatorial gene regulation. **Genes and Development**, v. 15, p. 2503-2508, 2001.

3. BERENDZEN, K. W. et al. Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. **BMC Bioinformatics**, v. 7, n. 522, 2006.