4- An Alternative Method for Detecting Conserved Elements in Multiple Sequence Alignments

Pavol Hanus1, Janis Dingel1, Joachim Hagenauer1 and Jakob C. Mueller2

1Munich University of Technology, Institute for Communications Engineering (LNT), Arcisstr. 21, 80290 Munich, Germany
2Jakob C. Mueller, Munich University of Technology, Institute for Medical Statistics and Epidemiology, Germany
Email: pavol.hanus@tum.de, hagenauer@tum.de, janis.dingel@mytum.de, jakob.mueller@imse.med.tu-muenchen.de

We present an alternative method for detecting conserved regions in multiple sequence alignments. In contrast to earlier approaches, we avoid using assumptions about neutral substitution rates. Our method is based on the Kullback-Leibler distance of the actual to the absolute conservation. Maximum Likelihood method is used to estimate the conservation related evolutionary parameters. A comparison of our results to the popular PhastCons method is provided.

I. CALCULATION OF CONSERVATION SCORES

Given a multiple sequence alignment of several species (see Fig. 1), we use the common approach of modeling evolution by the parameter set = {R,_, _ , ak_0} [1]. The rate matrix R describes a continuous stationary Markov process that models the substitutions occurring in the course of evolution.

The background equilibrium distribution is denoted as _ =[_A, .., _T]T . The topology of the phylogenetic tree relating the species is specified by _ . The relative branch lengths are denoted by _0. Different selection pressure on various regions of the DNA leads to different degrees of conservation. This substitution rate heterogeneity is accounted for by introducing a scalar ak acting as a multiplicative factor to the relative branch lengths of the phylogenetic tree _ = ak_0. We propose a definition of conservation which, opposed to earlier methods, is completely free of any assumptions about the neutral substitution
rate. Our method rather relies on the Kullback-Leibler distance to the well defined maximum possible conservation that does not allow for any mutations to occur. The pmf of a column x in the alignment X generated by the process satisfying the maximum conservation condition has only four entries corresponding to the four possible realizations of a column fcons x = _A _(x − [A..A]) + .. + _T _(x − [T..T]).

In order to calculate the pmf ˆ fx of the process that generated the actual column x in the alignment, we need to estimate the parameter set of the evolutionary process that led to
this column. First we fix all the a priori information in our parameter set and estimate the remaining free parameters in an alignment window XM of size M using a Maximum
Likelihood approach: ˆ = max{log(P(XM;))}. Based on our estimate, we can now compute the pmf ˆ fx of a column in the alignment. The conservation score sk assigned to the base k in the center of XM is the Kullback-Leibler distance of the estimated pmf to the pmf of absolute conservation:

sk = Px fcons
x log _fcons
x
ˆ fx _. Note that a zero score means that the base has been maximally conserved. The higher the score, the less conserved the base.

II. SIMULATION RESULTS

We get similar simulation results to those obtained with the PhastCons method [1]. However, the distance based conservation score calculated using our approach provides a more sensitive, better resolved measure of the degree of conservation than the probability score used in PhastCons. Figure 1 shows our score signal (LNTcons) in a conserved region of the human Chromosome 1. The evolutionary parameters R, _, _ and _0
were fixed as a priori information. A sliding window of size 21 was used to obtain the Maximum Likelihood estimate of ak. Below the plot, the corresponding alignment is depicted. For comparison purposes we also show the conservation score function obtained by PhastCons. Here the values correspond to the probability of being conserved for each base and thus converge to one for conserved bases. The distance like char-

Fig. 1. Comparison between the conservation scores obtained by PhastCons and by our method in a conserved region on human chromosome 1.

acter of our method allows to identify regions with different degree of conservation. Thus, e.g. by setting a very restrictive threshold the region starting at around 118 and ending at 143
is detected as highly conserved.

REFERENCES

[1] Adam Siepel, Gill Bejerano, and Jakob S. Pedersen et al., “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes,” Genome Res., vol. 15, no. 8, pp. 1034–1050, 2005