Poster G-23
Alignment-Independent Homology Estimator

Authors:
Avihay Apatoff (*The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University*)
Eddo Kim (*The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University*)
Yossef Kliger (*Compugen Ltd.*)

Short Abstract: We present a method that estimates the fraction of homologs in a set of protein alignments. The method requires an abundant protein feature that is conserved among homologs (e.g. the N-terminal signal peptide). The method is beneficial for the development and assessment of homology search methods.

Long Abstract:
Alignment-Independent Homology Estimator
Avihay Apatoff1,2, Eddo Kim1,2, and Yossef Kliger2

1The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel. 2Compugen Ltd, Tel Aviv 69512, Israel.

Identification of homologous proteins provides a basis for protein annotation. Protein alignment tools can reliably identify close homologs. Lowering the cutoff values enables identification of distant homologs, but also increases false pairing of non-homologous proteins. The ability to estimate the fraction of homologs in a set of protein pairs can be helpful in development and quality assessment of protein alignment methods.

Herein, we present an alignment-independent method, the Fhom-Estimator, to estimate the amount of homologs in a set of protein pairs. This method requires an abundant protein feature that is conserved among homologs. Protein pairs are classified into four subgroups according to the presence (or absence) of the chosen feature in both proteins. The observed distribution of protein pairs to the four subgroups can be extracted from the data (e.g. protein alignment hits). The expected distribution is calculated according to the protein feature prevalence. Comparison of the two distributions yields an estimation of the real fraction of homologs (Fhom) in the paired proteins set. We show that this estimation is always an underestimation, thus, it keeps us on the safe side.

We tested the method using the HomoloGene database [1], as a standard of truth, and the signal peptide as the protein feature. The signal peptide is a suitable protein feature for our method, since it is abundant, can be reliably detected [2], and is conserved among homologs.

To examine whether our method is alignment-independent, we divided the HomoloGene homologs into three groups according to their identity level. The estimated fractions of homologs in the three groups exceeded 90% (see figure). Hence, the Fhom-Estimator is

alignment-independent, and can be used in estimating the fraction of homologs in a set of protein pairs that may contain close homologs as well as distant homologs.

We have also tested the dynamic range of our method by adding random protein pairs, while maintaining the original signal peptide prevalence. The results confirmed that Fhom estimates well the real fraction of homologs between 0.01 and 1.

In conclusion, we present a simple method to estimate the fraction of homologs in a set of paired proteins. The Fhom-Estimator suits for the assessment of protein alignment methods.

[1] Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant et al. (2006). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 34: D173-180.
[2] Bendtsen, J.D., H. Nielsen, G. von Heijne, and S. Brunak. (2004). Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783-795.