Benchmarks results

We used the *OrthoBench* benchmark set [1] to validate the clusters of FASTA Herder. We evaluated the level of compression and misclassifications for the default parameters, for varying thresholds of comparable lengths, when low-complexity regions (LCRs) are detected and masked from the original sequences. We have as well compared with the levels of compression of PISCES [3] a tool with a similar capability.

1. The *OrthoBench*

The *OrthoBench* is a phylogeny-based benchmark set of clusters of orthologous groups (COGs). The set consists of 70 protein families that were defined based on domain content, manual inspection of the alignments and previous published description of the families from 12 metazoans species [1]. These protein families, referred as reference orthologous groups (RefOGs), are classified according to the rate of evolution (fast- vs. slow-evolving families), domain architecture (single domain vs. multiple repeated domains), low complexity/repeats (LCR), lineage-specific loss/duplication (single copy families vs. multiple duplication events), and alignment quality (high- vs. low-quality alignment). The RefOGs include a total of 1,677 proteins.

Supplementary Table 1: OrthoBench_benchmark.xls

2. Compression level and missclassifications for the OrthoBench & detailed results for each family of RefOGs

We identified the number of sequences that are clustered, the number of singletons as well as the number of misclassified sequences for the full Orthobench and for each family (RefOG). When compressing the full OrthoBench with the default parameters, our algorithm achieved 35.24% of the total size of the FASTA file with 0 errors.

Supplementary Table 2: RefOGs_OrthoBench.xls

3. Compression level and missclassifications for the OrthoBench for varying parameters of LCRs detection

LCRs in protein sequences are identified by the SEG program [2]. SEG uses a window parameter (W) to specify the size of regions to be detected with a complexity (K1) that is equal or less than a cutoff. We tested multiple sets of parameters of SEG to test for compression and missclassifications for all the sequences of the whole OrthoBench. In general, compressions are slightly larger when LCR detection is

more strict.

Supplementary Table 3: LCR.xls

4. Compression level and missclassifications for the OrthoBench for varying parameters of comparable lengths

FASTA Herder uses a set of thresholds limiting the maximum difference allowed between sequences to be clustered together. We investigated the effect on compression and the number of misclassified sequences when varying these thresholds. This experiment was performed using all the OrthoBench benchmark sequences with and without the use of dropping LCRs. Results indicate that more liberal thresholds can be applied to remove redundancy.

Supplementary Table 4: threshold_comparable_lengths.xls

5. Comparison to PISCES

We compared the compression of FASTA Herder and PISCES [3] with the default parameters using the OrthoBench. The OrthoBench is too large to run in the online PISCES tool; for that reason we split the file in 4 portions. FASTA Herder takes an average of 10 seconds to process the data while PISCES takes 29.25 minutes. As PISCES only takes into account the similarity between sequences, it achieved much greater compressions than FASTA Herder (see results in Supplementary Table 5). However PISCES did often misclassified a large number of sequences.

Supplementary Table 5: PISCES_and_FASTA_Herder.xls

6. References

[1] K. Trachana, T.A. Larsson, S. Powell, W.-H. Chen, T. Doerks, J. Muller, P. Bork, Orthology prediction methods: a quality assessment using curated protein families, BioEssays, 33 (2011), pp. 769–780.

[2] Wootton,J. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. Methods in Enzymol., 266, pp. 554–571.

[3] Wang G. and Dunbrack, R. (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research,* 33, W94.