Comprehensive Redundancy Analysis for Complete Proteomes

Abbas Idris, Jonathan Patsenker, Leonid Peshkin, Marc Kirschner
Department of Systems Biology, Harvard Medical School, Boston, MA 02115

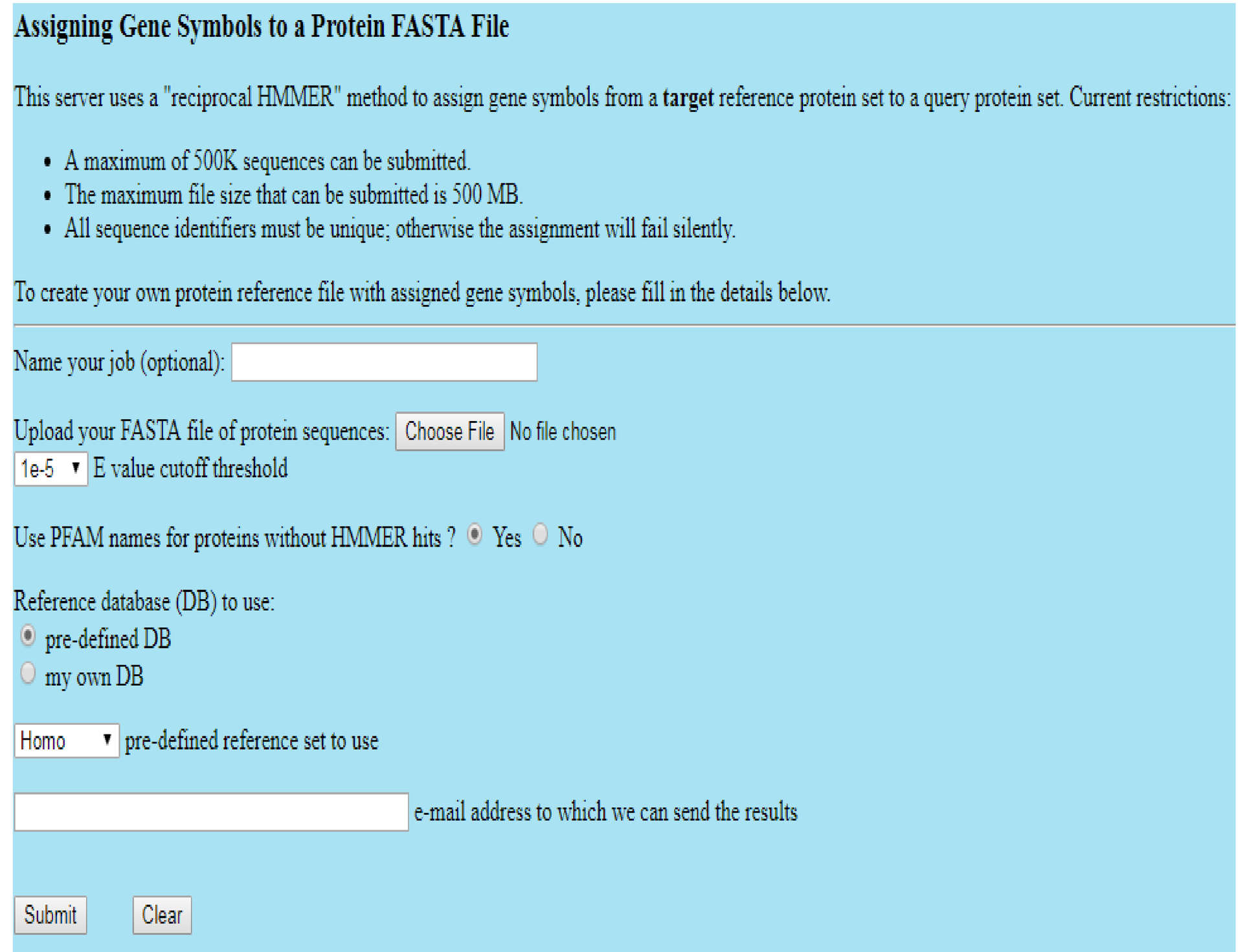
FAS Center for Systems Biology

Introduction

Advances in sequencing technology, software algorithms, and computing power have allowed researchers to sequence organisms from all branches of life - from humans to invertebrates (worm and fly) and bacteria. [19 mammalian genomes (out of 56 eumetazoan) enjoyed a lot of attention fueled by hope of characterizing functional genetic variation, on the other end of the spectrum ~1500 bacteria were sequenced mostly to arm our defenses against parasites. Remarkably, same technology can be used without modification on all branches of life. The hope is that from comparison we will understand the principles of evolution as well as human development, physiology, anatomy and ultimately disease. Problems with proteins may arise at two levels: the issues may be with the genome assembly or there may be problems with creating the gene models. In the case of poor gene models, one may fail to completely mask repeat regions or to properly train gene prediction software, resulting in low complexity and high redundancy.

CRAP Analysis Pipieline

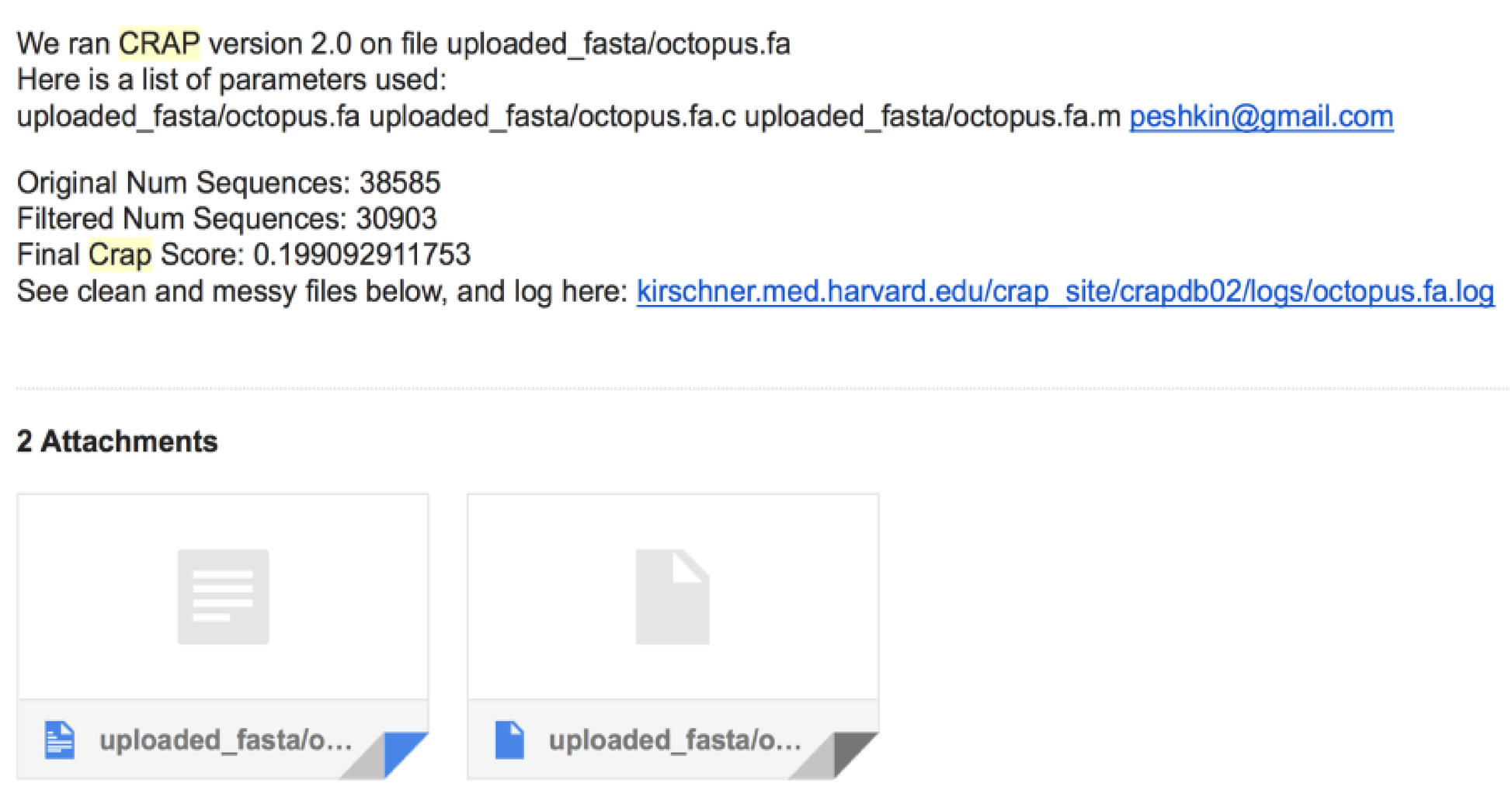
A.



B.

Figure 1: Analysis on user supplied input is carried out from CRAP server input. **A**. gene symbol assignment for proteomes using HMMER and PFAM. **B**. CRAP filtration and analysis pipeline which implements gene symbol analysis as a core step.

A.


We ran CRAP version 2.0 on file uploaded_fasta/octopus.fa
Here is a list of parameters used:
uploaded_fasta/octopus.fa uploaded_fasta/octopus.fa.c uploaded_fasta/octopus.fa.m peshkin@gmail.com

Original Num Sequences: 38585
Filtered Num Sequences: 30903
Final Crap Score: 0.199092911753
See clean and messy files below, and log here: kirschner.med.harvard.edu/crap_site/crapdb02/logs/octopus.fa.log

2 Attachments

uploaded_fasta/o...
uploaded_fasta/o...

B.

```
>Ocbimv22004408m.p Sequence Has Too Many Xs
MLLVRFTMLLYNFQPLVKVGLKTKRILAQGYDSLVISFSCCCCCYYCCSCCTEKNNSFGGGGXXXXXX
XXXXXXXXXGVDKFKF

>Ocbimv22034104m.p Sequence Too Short: Length = 29
MNGVVPTQQTKDVSQYTGRVYLWNSGMIF

>Ocbimv22000013m.p Too Low Complexity 0.367088607595 < 0.9
HIFHARRELRITGTERGWKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK
KPISMKQSPKE

>Ocbimv22029368m.p Sequence Is Redundant Fragment
MYFRVAGAKREGRSLAFGRRLRSVGWLVGFGLDFVRVFLWLG

>Ocbimv22022521m.p Sequence is Fission Fragment
MKAKQSRLQRDDFETLKIIGRGAFGEVAVVKLKGTEEVYAMKILNKWEMLKRAE
```

Figure 2: The output of the pipeline includes **A.** Email body corresponding to compression success of the proteome, including a CRAP score. Two output files are sent to the user: the file on the left corresponds to the compressed proteome and the right most file is a diagnostics file of bad protein sequences. **B**. This is the content contained within the diagnostics file. This file gives you the sequence ID, the problem detected with the sequence and often a value corresponding to the filtration done on the sequence.

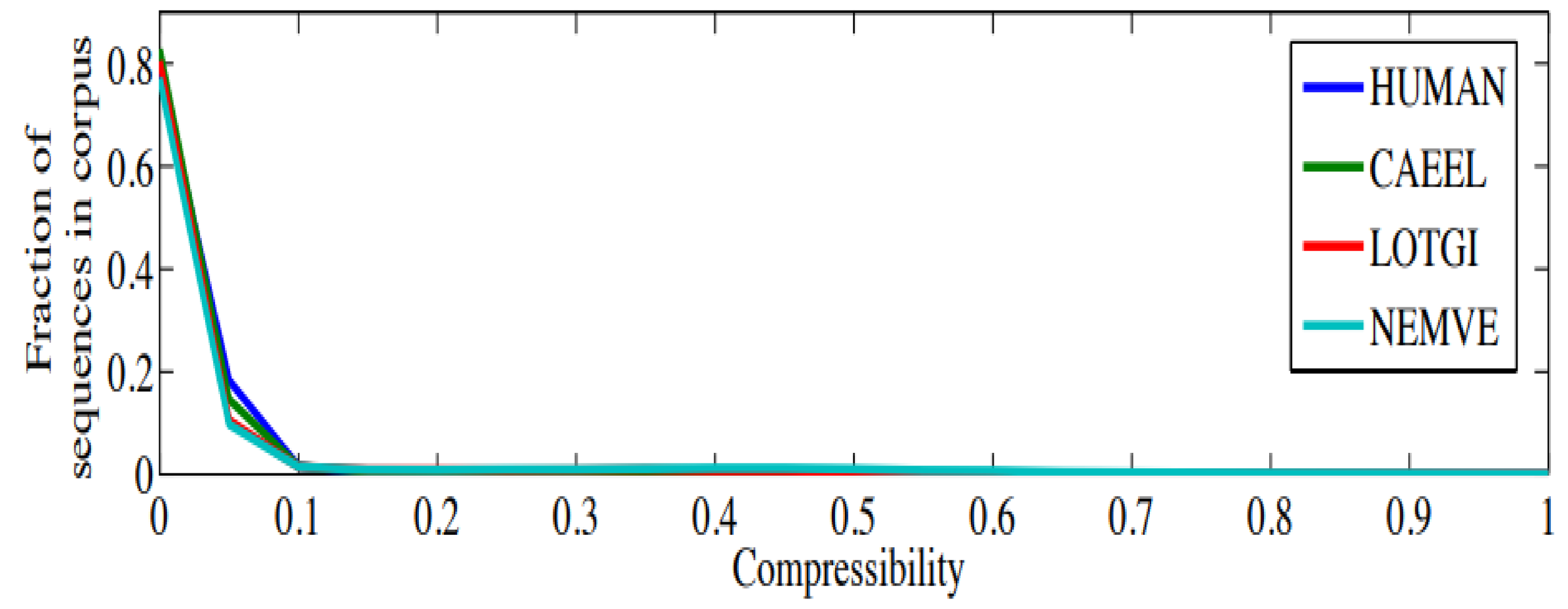
Methods

The CRAP pipeline operates on a system of filters and analyzers which take a user supplied input of an organism's full proteome and processes it, returning a successfully compressed proteome. The pipeline deals with problems of low complexity and high redundancy within the proteome, through the implementation of filters that eliminate junk sequences (unidentified amino acids, short sequences, and repeated amino acids), repeated sequence motifs, corpus redundancy, potential fusion and fission errors induced by poor sequencing, and the presence of genes identified as essential genes stratified through all living organisms. The pipeline then returns results corresponding to the compressibility of the proteome.

Redundancy and Compressibility

Proteomes are classified according two factors with respect to complexity: the inter-sequence complexity and the intra-sequence complexity. Intra-sequence complexity (corpus complexity) reflects the average quality of an individual sequence judged on its own. The inter-sequence complexity (corpus redundancy) reflects the average quality of an individual sequence judged with respect to the entire proteome. Our assumption is that low complexity and high redundancy reflect a poor proteome rather than a biologically encoded characteristic of the proteome.

A.



B.

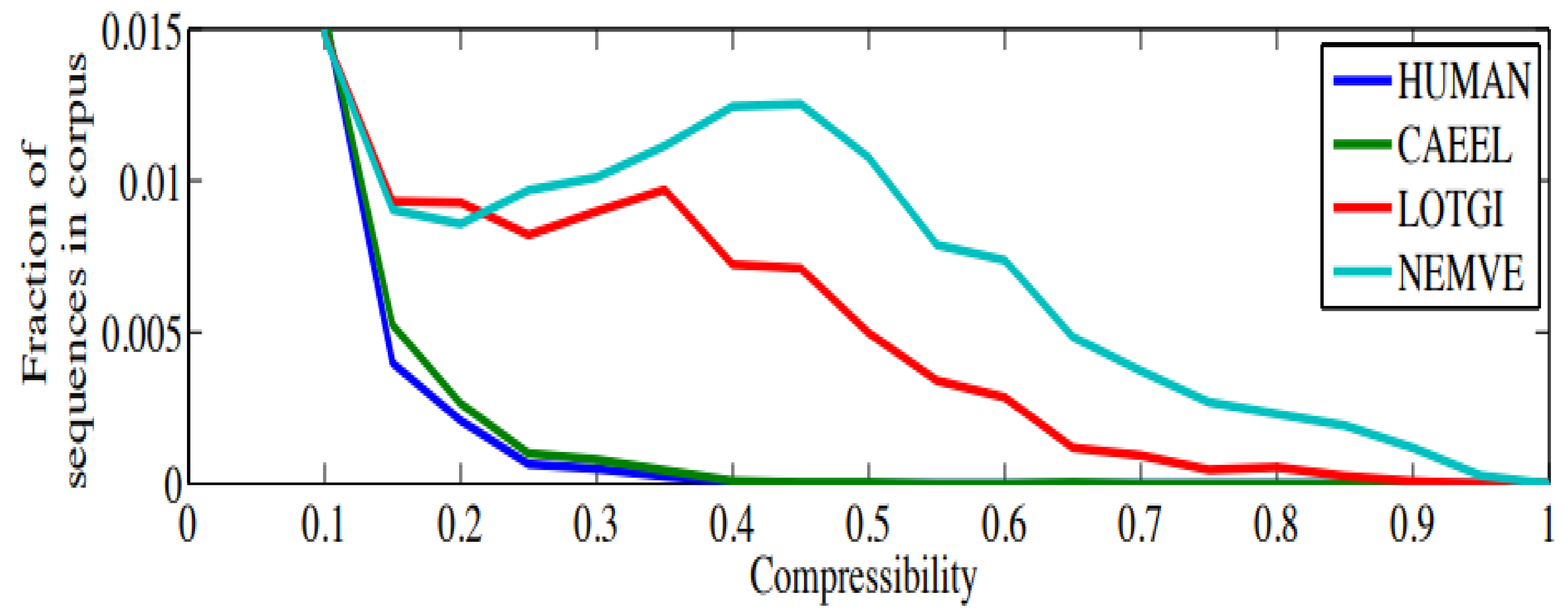


Figure 3. Sequence compressibility **A.** Plot of the fraction of sequences in the corpus of each organism (human, nematode, mollusks, and sea anemone) that can be compressed. **B.** Closer look at the characteristics of compressibility, note that it is a predictable trend where humans have more sequences at higher percent compressibility

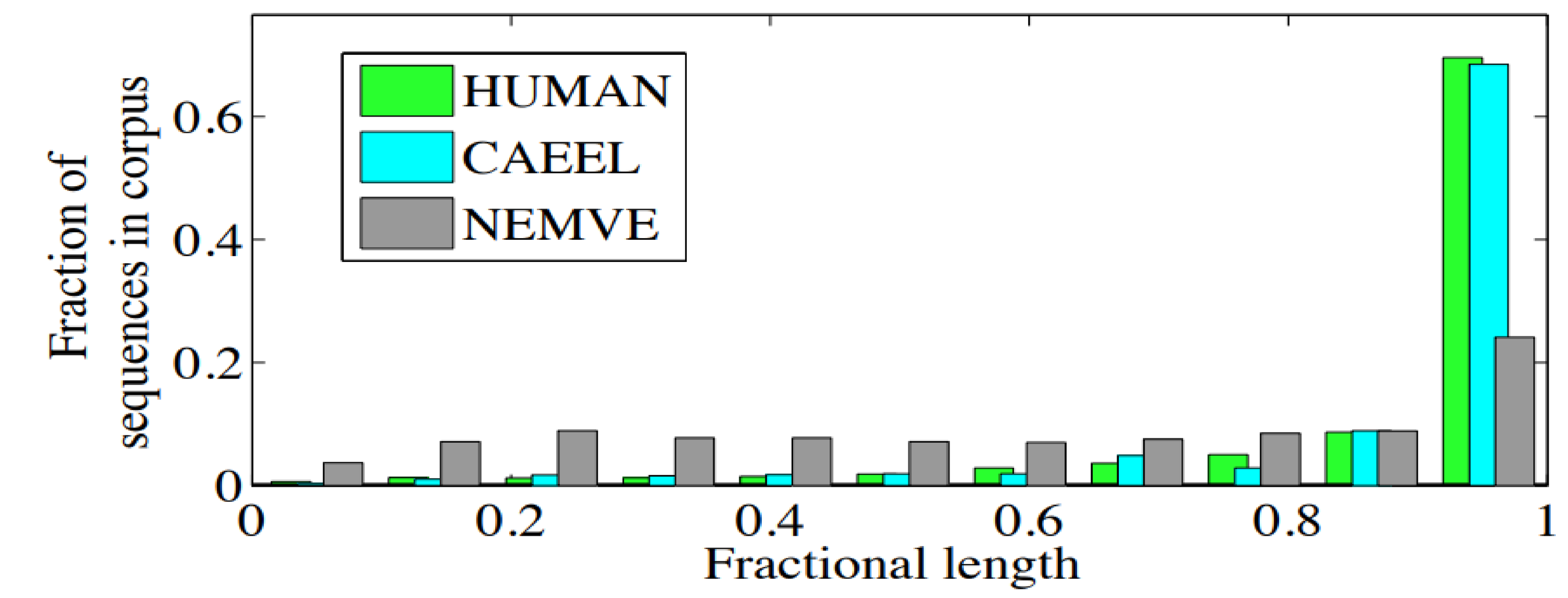


Figure 4. Histogram of fractional length at the sequence identity of 70 percent or higher. Human and worm genomes have very similar signature while in sea anemone large percentage of sequences is shifted, indicating short pieces of other genes

Conclusion and future works

In conclusion, (point out the improvements in the human and fly) we call for standardization of the quality control for the complete genome sequences. Reasoning about evolution of function critically depends on high quality data across the phylogeny. The "curse of availability" of technological advances has led to heavily favoring obtaining more data at the expense of careful curation and re-analysis of existing data, as well as obtaining more data to improve existing genomes. Genomics field is spread thin across the phylogenetic landscape

Acknowledgment and references

Work was supported by and conducted at Harvard University department of Systems Biology, Marc Kirschner Laboratory, under the mentorship and Advising of Leon Peshkin.

1. Khenoussi, W., Vanhoutrève, R., Poch, O. & Thompson, J. D. SIBIS: a Bayesian model for inconsistent protein sequence estimation. Bioinforma. Oxf. Engl. 30, 2432–2439 (2014). 2. Yang, Y., Gilbert, D. & Kim, S. Annotation confidence score for genome annotation: a genome comparison approach. Bioinforma. Oxf. Engl. 26, 22–29 (2010). 3. Nagy, A. et al. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics 9, 353 (2008). 4. Prosdocimi, F., Linard, B., Pontarotti, P., Poch, O. & Thompson, J. D. Controversies in modern evolutionary biology: the imperative for error detection and quality control. BMC Genomics 13, 5 (2012). 5. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma. Oxf. Engl. (2015). doi:10.1093/bioinformatics/btv351 6. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. Nucleic Acids Res. 37, 289–297 (2009). 7. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinforma. Oxf. Engl. 23, 1061–1067 (2007). 8. Pagani, I. et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40, D571–579 (2012). 9. Blomberg, P., Nordström, K. & Wagner, E. G. Replication control of plasmid R1: RepA synthesis is regulated by CopA RNA through inhibition of leader peptide translation. EMBO J. 11, 2675–2683 (1992). 10. Shinzato, C. et al. Using the Acropora digitifera genome to understand coral responses to environmental change. Nature 476, 320–323 (2011). 11. Katti, M. V., Sami-Subbu, R., Ranjekar, P. K. & Gupta, V. S. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. Protein Sci. Publ. Protein Soc. 9, 1203– 1209 (2000). 12. Wise, M. J. 0j.py: a software tool for low complexity proteins and protein domains. Bioinforma. Oxf. Engl. 17 Suppl 1, S288–295 (2001). 13. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinforma. Oxf. Engl. 26, 680–682 (2010).