Ultra-fast functional classification of short reads using UProC with Pfam and KEGG

Manuel Landesfeind, Robin Martinjak, Heiner Klingenberg and Peter Meinicke
Department of Bioinformatics, Institute of Microbiology and Genetics, University of Göttingen
peter@gobics.de

As introduced in [Mei15] and [LM14a] we here present a novel tool UProC for large-scale sequence classification and show its application to functional analysis of metagenomic and transcriptomic data.

The functional and metabolic characterization of organisms and organism communities based on massive sequencing is central in genome and metagenome studies. With the current next-generation sequencing techniques the number of reads per sample has dramatically increased while in comparison to Sanger-sequencing the average read-length has substantially decreased. Because the vast amount of short read data renders a classical BLAST-based analysis infeasible, novel tools have to be developed to cope with the already existing computational bottleneck. In metagenomics and metatranscriptomics the functional classification of sequencing reads based on assignments to annotated protein sequences or families is usually the computationally most expensive task. Using a sensitive BLASTX search [AGM+90] against a large protein sequence database, the processing of millions of short reads on an actual desktop computer can take years and makes massive parallelization on large computer clusters inevitable. Using HMMER profile HMMs [Edd98] is about one order of magnitude faster but would be restricted to protein families that can be well represented by multiple sequence alignments. Therefore, we have developed the **U**ltra-fast **Pro**tein domain **C**lassification (UProC) tool which is about 10000 and 1000 times faster than BLASTX and HMMER3, respectively. The UProC algorithm features a novel sequence scoring approach that we refer to as "Mosaic Matching". Although, UProC has been designed to assign sequences to protein domain families we have also used it with the KEGG database of full-length gene families (KEGG Orthologs) and found it well-suitable to predict the functional repertoire of an organism from unassembled RNAseq reads [LM14a]. We have used UProC for the development of several tools that require the computation of functional profiles [KALM13, LM14b, AWDM15] and it makes up the core of our CoMet [LASM11] and CoMet-Universe [AKLM14] web servers. The UProC source code is available at `https://github.com/gobics/uproc` with the latest release package (including precompiled binaries for Windows) and databases provided at `http://uproc.gobics.de/`.

1 UProC algorithm and Pfam domain classification results

The UProC "Mosaic Matching" algorithm comprises several essential elements which involve the scoring and classification of protein sequences as well as the prior database construction. In this extended abstract we will only give a short overview of the most basic steps and would like to point the interested reader to the original publication [Mei15] for a full description.

The algorithm first extracts all oligopeptides ("words") of length 18 from a query protein sequence and for every word identifies the nearest neighbour in a sorted database array of labelled reference words. All residues of a nearest neighbour word are then scored with respect to their similarity to the query word using a position specific scoring matrix that has been optimized by a machine learning approach. Finally all position specific scores from reference words with the same protein family label are combined in a Mosaic Match to yield the final score of the sequence with respect to the particular family. If the score is above a length-dependent threshold a significant match is reported.

We evaluated UProC with the Pfam 24 database [FMT+10] on real and simulated metagenome data of varying complexity and read length, comparing it with HMMER3 and RPS-BLAST. We found that on the shortest read length (100 bp) UProC outperformed the profile-based tools in terms of sensitivity

at a comparable specificity which was around 95% in all cases. The sensitivity of UProC varied from 88.9% on a human microbiome dataset to 68.5% on marine metagenome data, down to 50.1% on data from a microbial mat community. The corresponding sensitivity of HMMER (RPS-BLAST) for these datasets was 52.3 (48.5), 47.5 (44.8) and 42.8 (39.6) percent, respectively. The results indicate that the computationally more expensive profile methods which constitute the state-of-the-art for full length protein sequences might not be optimal for this kind of short read data. This has also been found in a recent study using transcriptomic data [ZSC13] which exhibited a substantial sensitivity loss of HMMER and other profile-based methods for the classification of protein domains in short reads. As expected, for increasingly longer reads, at some point HMMER becomes the most sensitive tool. For a good classification performance, UProC requires a large sequence database that covers much of the variation within different protein families. On the other hand, UProC does not require the protein families to be representable in terms of multiple alignments. At the UProC homepage we offer a precompiled database for a recent version of the KEGG orthologs [KG00] which are widely used for metabolic profiling in metagenomics and metatranscriptomics. An application of UProC to KEGG-based classification of short reads is reported in the following.

2 Using UProC with KEGG to predict functional repertoires from unassembled RNA-Seq data

In the annotation of *de novo* sequenced organisms the inference of potential gene functions is a fundamental step. If a genome sequence can be assembled at sufficient quality, for an automatic annotation of predicted genes, putative functions are usually identified using homology search techniques. Without the genomic sequence a *de novo* transcriptome assembly can be used to assess major parts of the functional repertoire where the achievable coverage strongly depends on the experimental setup and the organism under investigation. This strategy has been adopted as a valuable alternative for certain organisms which for example provide large or hybrid genomes. Although many tools have been proposed for *de novo* transcriptome assembly, the risk of misassemblies remains and also depends on the organism. In addition, the computational effort in terms of RAM storage requirements for the assembly can be demanding. Finally, the result of the analysis is highly dependent on several parameters, in particular on a suitable threshold for the homology search step, such as a BLAST E-value cut-off, which is necessary to decide on the presence or absence of a particular function.

In a recent study [LM14a] we have investigated to what degree it is possible to reconstruct the functional inventory of an organism using only unassembled transcriptome data. The short read data was directly mapped to KEGG functions by searching for homologies to the corresponding KEGG Ortholog families. For the evaluation we used a large RNA-Seq data set from *Arabidopsis thaliana* and removed all sequences of that organism and close relatives from the database. To obtain a reliable prediction on the presence of a function on the basis of short reads, it is important to evaluate the aggregated evidence that is generated by all reads showing similarity to reference sequences of the same family. The similarity scores calculated at the homology search step were combined in a family-specific evidence measure which was finally used for the prediction of the corresponding function. We found that over the whole range of possible functions the distribution of the evidence measure typically shows a bimodal distribution that reflects the dichotomy of strong and weak similarities with respect to different organisms in the database. This bimodality makes it possible to automatically adjust the prediction threshold using a mixture model for analysis of the evidence distribution.

Our results show a high sensitivity of up to 94 percent for the prediction of biomolecular functions in KEGG. The low false positive rate of 4 percent indicates that the automatic threshold calibration is highly effective even providing a better performance than prediction on the basis of a *de novo* transcriptome assembly. In our study we also compared the impact of different homology search tools, including several pairwise approaches and UProC. We found that the application of UProC provides the fastest solution and at the same time the highest detection performance (F1-measure) for this particular task.

Thereby, the UProC memory requirements of approximately 16 GB RAM are clearly higher than with BLAST but much lower than for transcriptome assembly tools.

In metatranscriptomics, not only the functional characterization but also the phylogenetic classification of sequencing reads is required. Although, UProC can be used for taxonomic profiling of metagenomes by means of the Taxy-Pro mixture model [KALM13] for evaluation of protein domain counts, the taxonomic binning of reads is currently not possible. This is a clear advantage of BLASTX-based approaches (see e.g.[GS11, HMR$^+$11]) that can provide both, functional and phylogenetic classification of single reads. Currently, we are working on a UProC version that integrates both kinds of classification.

References

[AGM$^+$90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, Oct 1990.

[AKLM14] K. P. Aßhauer, H. Klingenberg, T. Lingner, and P. Meinicke. Exploring Neighborhoods in the Metagenome Universe. *International Journal of Molecular Sciences*, 15(7):12364–12378, July 2014.

[AWDM15] K. P. Asshauer, B. Wemheuer, R. Daniel, and P. Meinicke. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, May 2015.

[Edd98] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, January 1998.

[FMT$^+$10] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res.*, 38:D211–222, Jan 2010.

[GS11] W. Gerlach and J. Stoye. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research*, 39(14):e91–e91, August 2011.

[HMR$^+$11] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, September 2011.

[KALM13] H. Klingenberg, K. P. Aßhauer, T. Lingner, and P. Meinicke. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics*, 29(8):973–980, April 2013.

[KG00] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28(1):27–30, Jan 2000.

[LASM11] T. Lingner, K. P. Aßhauer, F. Schreiber, and P. Meinicke. CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research*, 39(Web Server issue):W518–523, July 2011.

[LM14a] M. Landesfeind and P. Meinicke. Predicting the functional repertoire of an organism from unassembled RNAseq data. *BMC Genomics*, 15(1):1003, November 2014.

[LM14b] Lingner, T. and Meinicke, P. Characterizing metagenomic novelty with unexplained protein domain hits. In *German Conference on Bioinformatics 2014*, GI-Edition : lecture notes in informatics, Proceedings, pages 69–78, 2014.

[Mei15] P. Meinicke. UProC: tools for ultra-fast protein domain classification. *Bioinformatics*, 31(9):1382–1388, 2015.

[ZSC13] Y. Zhang, Y. Sun, and J. R. Cole. A Sensitive and Accurate protein domain cLassification Tool (SALT) for short reads. *Bioinformatics*, 29(17):2103–2111, January 2013.