METADOMAIN: A PROFILE HMM-BASED PROTEIN DOMAIN CLASSIFICATION TOOL FOR SHORT SEQUENCES

YUAN ZHANG and YANNI SUN

Dept. of Computer Science and Engineering, Michigan State University,
East Lansing, MI 48824, U.S.A.
E-mails: zhangy72@msu.edu
yannisun@msu.edu

Protein homology search provides basis for functional profiling in metagenomic annotation. Profile HMM-based methods classify reads into annotated protein domain families and can achieve better sensitivity for remote protein homology search than pairwise sequence alignment. However, their sensitivity deteriorates with the decrease of read length. As a result, a large number of short reads cannot be classified into their native domain families. In this work, we introduce MetaDomain, a protein domain classification tool designed for short reads generated by next-generation sequencing technologies. MetaDomain uses relaxed position-specific score thresholds to align more reads to a profile HMM while using the distribution of alignment positions as an additional constraint to control false positive matches. In this work MetaDomain is applied to the transcriptomic data of a bacterial genome and a soil metagenomic data set. The experimental results show that it can achieve better sensitivity than the state-of-the-art profile HMM alignment tool in identifying encoded domains from short sequences. The source codes of MetaDomain are available at http://sourceforge.net/projects/metadomain/.

Keywords: Protein domain classification; metagenomics; short reads; profile HMM.

1. Introduction

With the advent of next-generation sequencing and culture-independent methods, an enormous amount of metagenomic data have been sequenced from microbial communities from different habitats. In order to understand the phylogenetic complexity and biological functions of microbial communities, as well as their interactions with the host, automatic annotation tools such as CAMERA,[1] MG-RAST,[2] and MEGAN[3] are being used for annotating metagenomic data sets. As an important component of these metagenomic annotation tools, protein homology search provides basis for identifying putative genes and assigning those genes to annotated functional categories (e.g. protein domain families). There are two major methods for protein homology search. The first method is based on pairwise sequence alignment. Putative genes can be identified by comparing metagenomic reads against annotated protein databases using BLASTX.[4] Although BLAST is one of the most efficient protein homology search tools, probabilistic model-based methods have better sensitivity for remote protein homology recognition. In particular, using profile hidden Markov models (HMMs) to represent a protein family greatly improves homology search sensitivity between highly diverged sequences.[5] Thus it is desirable to conduct protein domain classification using profile HMM-based tools such as HMMER.[6] In conjunction with a fast-growing protein domain family database Pfam,[7] HMMER is able to classify sequences into different domain families with high accuracy. In addition, the latest implementation of profile HMM-based domain classification tool HMMER 3.0[6] has achieved comparable speed to BLAST, making it suitable for large-scale protein compositional

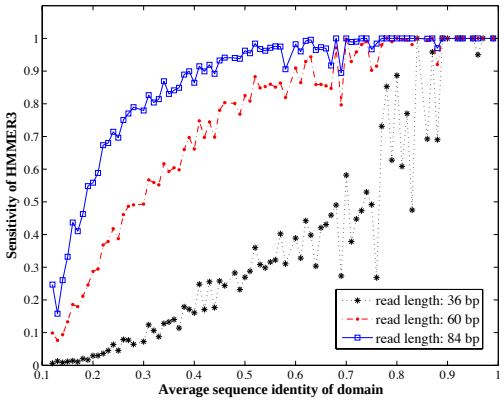

Fig. 1. Change of the read classification sensitivity of HMMER over read length and the average sequence identity of domain families

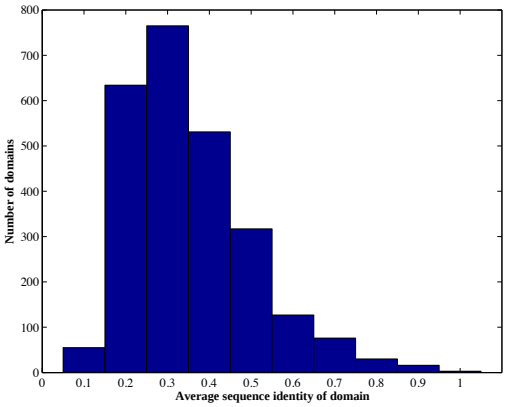

Fig. 2. Histogram of the average pairwise sequence identity for 2558 domains

analysis. For the convenience of discussion, we use HMMER to refer to HMMER 3.0 hereafter unless otherwise specified.

Low sensitivity of HMMER on classifying short reads. Because of the high sensitivity of remote homology recognition, HMMER has been successfully applied to genome-wide domain analysis. However, its sensitivity is significantly limited by the short reads of metagenomic data sets and poorly conserved domains. In order to investigate how read length and domain identity affect the sensitivity of HMMER, we randomly sampled 200 peptides with lengths of 12, 20, and 28 amino acids from the seed sequences of each of the 2,558 Pfam domains, which contain the word "Bacteria" in their descriptions. The peptides were aligned with the domain families using HMMER. We used the E-value cutoff 1000 in order to boost the sensitivity. For each domain, the read classification sensitivity of HMMER is measured as the ratio of the number of aligned reads to the total number of sampled reads. We sort all data points by domain identity in ascending order and plot them in Figure 1. For domains with the same identity, their average sensitivity is reported.

Figure 1 shows that the sensitivity of HMMER deteriorates with the decrease of the query sequence length and domain identity. The sensitivity is decreased from 90% to 65-70% when the lengths of reads change from 28 residues (i.e., 84 bp for corresponding DNA reads) to 20 residues (i.e., 60 bp for DNA reads) for domains with identity around 40%. Although next-generation sequencing technologies are producing longer reads and assembly tools may be available to assemble short reads into longer contigs, there is still a need for a protein domain analysis tool for short reads. First, many finished or on-going metagenomic sequencing projects contain reads with lengths from 35 to around 400 bp depending on the chosen sequencing technologies. In addition, peptide sequences encoded in individual metagenomic sequence reads may share only small overlaps with existing domain families. Thus, a sizable portion of many available data still contains short reads. Second, the sheer amount of data and the complexity of many metagenomic data sets pose a great challenge for assembly tools.[8] A large portion of short reads cannot be correctly assembled into longer contigs. Third, many domain families

exhibit low average sequence identity, which poses a challenge for short and medium-sized reads. Figure 2 shows the histogram of pairwise sequence identity for domains related to bacteria. Of 2558 domains, there are about 43% domains with average identity no greater than 0.3. For these domains, the sensitivity of HMMER is between 0.7 and 0.8 for reads of length 84 bp, between 0.4 and 0.6 for reads of length 60 bp, and smaller than 0.1 for reads of length 36 bp. As a result, although a large number of reads are sequenced from genes, which are highly compact in microbial genomes, only a small percentage of the short reads can be classified into their native domains using existing tools.

In this work, we introduce *MetaDomain*, a protein domain classification tool designed for short reads in metagenomic data sets. MetaDomain provides a complementary protein analysis tool to HMMER on assigning short reads into their native families.

2. Related Work

Profile HMM-based protein homology search is widely used for mining microbial genomes. For example, Ellrott et al.[9] investigated the distribution of protein families in the available human gut genomic and metagenomic data. As the data set contains assembled contigs, using HMMER is expected to achieve high sensitivity. Schlüter et al.[10] used HMMER to understand the genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. The reads have an average length of 104 bp, which is also adequate for HMMER to achieve high sensitivity.

Besides providing a basis for functional profiling, profile HMM-based homology search was also used for phylogenetic complexity analysis in metagenomic data. The phylogenetic algorithm CARMA[11] uses all Pfam domain and protein families as phylogenetic markers to identify the source organisms of environmental DNA fragments as short as 80 bp. As we show in Figure 1, profile HMM-based tools have sensitivity of at least 0.9 in classifying reads of 80 bp into domains with average sequence identity above 40%. However, for poorly-conserved domains, a significant number of reads might be missed. A similar but faster tool Treephyler[12] conducted community profiling in metagenomics and metatranscriptomics based on Pfam domain assignments. Treephyler was applied to a data set with average read length of 200 bp. It is unclear how shorter reads affect its performance.

Our previous work designed a tool HMM-FRAME,[13] which can identify and correct frame-shift errors in pyrosequencing reads during protein domain classification using profile HMM-based alignment. However, it was not specifically designed to handle short reads.

Finally, we note that the method used in MetaDomain shares a similar rationale to the recent work by Weng et al.[14] Weng et al. reported that taxonomic binning tools for metagenomes discard 30-40% of Sanger sequencing data due to the stringency of BLAST cut-offs. Thus, they re-analyzed the discarded reads using less stringent cut-offs. In order to control the false positive matches introduced by the relaxed cut-offs, they used the evolutionary conservation of adjacency between neighboring genes as an additional criterion.

3. Method

HMMER uses E-values as the discrimination threshold to determine the membership of a query sequence. However, short reads may only generate low alignment scores and thus in-

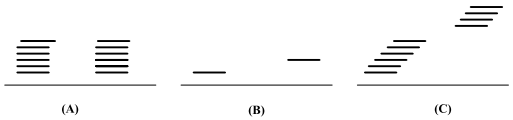

Fig. 3. Three types of alignment distributions

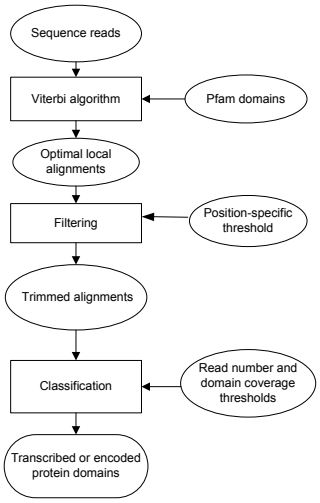

Fig. 4. Pipeline of MetaDomain

significant E-values. In particular, the conservation across the entire length of a domain family can be highly variable, posing a great challenge for classifying reads sequenced from poorly conserved sub-regions. In order to increase the sensitivity of aligning remotely-related short reads, we propose position-specific score cutoffs, by which poorly conserved regions allow more relaxed discrimination thresholds than well-conserved regions. However, the low thresholds can easily incur random matches. In order to control the false positive rate, we examine the position distribution of read alignments. The position distribution of read alignments on a truly encoded domain is expected to be more uniform than a domain that incurs random read alignments.[15,16] Figure 3 shows the schematic representations of three types of distributions of read alignments along a domain. The alignments in (A) and (B) are more likely to be random. Thus the domains may not be encoded in the data set. The alignment distribution in (C) exhibits a much more uniform distribution, providing strong evidence for the existence of the underlying domain in the data set. Thus, by using relaxed position-specific score cutoffs and inspecting the distribution of alignments, we expect to classify more short reads into the correct domain families while not falsely reporting domains that are not characterized in the data.

3.1. *Pipeline of MetaDomain*

The input to MetaDomain includes sequence reads and a list of protein domains. The output is a list of domains encoded in the underlying data set and the number of aligned reads. Figure 4 shows a schematic representation of the pipeline of MetaDomain.

MetaDomain consists of three main stages: short read alignment, filtering, and classification. In the alignment stage, we use the Viterbi algorithm[5] to search for the best local alignment between a query sequence and a profile HMM-represented domain family. In the filtering stage, we first apply a position-specific score threshold to eliminate insignificant alignments. Then we remove stacked alignments with the same alignment positions inside a poorly conserved region. In the final stage, we use the number of aligned reads and the distribution of alignment positions to determine whether a domain is encoded.

3.2. *The Viterbi algorithm*

The Viterbi algorithm aligns a query sequence to a profile HMM by searching for the most probable state path in the model. Unlike HMMER, MetaDomain directly aligns a DNA sequence to a profile HMM. To do so, we implicitly align translated peptides under different reading frames with a profile HMM. Let π be a state path in a profile HMM M and let x be a query DNA sequence. The Viterbi algorithm searches for the most probable path π^* such that $\pi^* = argmax_\pi(x, \pi)$. The output of the Viterbi algorithm includes the optimal alignment and its score. As Viterbi is a standard algorithm designed for HMMs, we refer readers to Durbin et al.[5] for a detailed illustration of the dynamic programming equations for finding π^*. The major difference between our implementation and the standard Viterbi algorithm includes : 1) our implementation accepts a DNA rather than a peptide sequence as input; 2) a local alignment can start and end with any state without incurring insertion or deletion penalties.

3.3. *Alignment Filtering*

MetaDomain employs two filtering mechanisms to increase its sensitivity in aligning short reads while maintaining a low false positive rate: position-specific thresholds (PSTs) and trimming.

3.3.1. *Position specific threshold*

PST allows different alignment thresholds for well conserved and poorly conserved regions. Let the length of a query DNA sequence be L (in bp). Denote the profile HMM as M. Let $M_{i,j}$ be a sub-model formed by all consecutive states from the ith match state M_i to the jth match state M_j. The upper bound of the alignment score against $M_{i,j}$ is the maximum score that can be generated by aligning any input sequence of length $j - i + 1$ with $M_{i,j}$. Let $a_{i,j}$ denote the transition probability from state M_i to state M_j. Let $e_i(a)$ denote the probability of state M_i emitting amino acid a. Then the upper bound $U_{i,j}$ for sub-model $M_{i,j}$ is calculated as follows:

$$U_{i,j} = \prod_{k=i}^{j} a_{k,k+1} \times max(e_k(a))$$

where $a_{j,j+1}$ is set to 1 because j is the ending state of the sub-model.

We define PST for the submodel $M_{i,j}$ as:

$$PST_{i,j} = \gamma U_{i,j}$$

where the coefficient γ is a user-specified parameter in the range of [0,1]. It can be flexibly adjusted to control the trade-off between sensitivity and false positive rate of MetaDomain. The default value is 0.6, which is used in our experiments.

3.3.2. *Alignment trimming*

Alignment with scores larger than their corresponding PSTs will pass the first filtering stage. As each domain has various conservation along the entire length of the model, well-conserved

sub-regions have high PSTs while poorly-conserved sub-regions yield low PSTs. Thus, random sequences tend to be aligned to poorly-conserved regions by MetaDomain, incurring a high FP rate. Our empirical experiments show that dozens of reads that are not sequenced from the underlying domain can be aligned to the same position in a poorly-conserved subregion. In order to minimize the effects of noise, we discard stacked alignments that have the same alignment positions.

3.4. *Protein domain classification*

In this stage we extract two features from the collected read alignments for each domain: the number of aligned reads and the domain coverage. The domain coverage is the fraction of positions covered by at least one read alignment in a domain. MetaDomain then applies a simple decision tree to classify all the target domains into two classes: encoded domains and non-encoded domains. If both features of a domain are equal to or bigger than their corresponding thresholds, this domain will be classified as encoded. Otherwise it is not encoded in the sample. By default, the cutoff for domain coverage is 30%. Ideally, the cutoff for the number of aligned read should be determined based on the properties of data such as sequencing depth. If users do not specify this value, we use 20 by default.

4. Experimental Results

In order to evaluate the performance of MetaDomain on real data generated by next-generation sequencing technologies, we applied MetaDomain to protein domain analysis in two data sets. The first one is the transcriptome generated using RNA-seq for *Burkholderia cenocepacia*. As both the reference genome and its domain annotations are available, we can quantify the sensitivity and false positive (FP) rate of MetaDomain. The second one is metagenome data sequenced from soil. We applied MetaDomain to identify domains encoded in the underlying data. In addition, we compared HMMER and MetaDomain in both applications.

4.1. *Identifying transcribed protein domains in transcriptome*

In this experiment, we conducted transcribed domain analysis in the transcriptome from one strain of *B. cenocepacia* named AU1054.[17] By using Illumina RNA-seq, the authors generated multiple samples for AU1054 in two growth media. We used one replicate of cDNA sample of AU1054 in the growth medium cystic fibrosis. In total, 3,361,008 reads of a length of 41 bp were downloaded from the website provided by the authors. We evaluated the performance of read classification and domain identification of MetaDomain and HMMER.

4.1.1. *Performance of read classification*

The performance of read classification is quantified using both read classification sensitivity and FP (false positive) rate. In this experiment, the read classification performance is computed on reads that can be mapped to annotated domains. Below we sketch the main steps to obtain mapped reads for a domain using the reference genome and the domain annotations. First, we downloaded the genome of AU1054 and the annotated genes and domains from the IMG website.[18] There are 2,181 annotated Pfam domains. Second, the reads were mapped to

the reference genome using Bowtie[19] with two mismatches allowed. Third, we compared the positions of read mapping and annotated domains. For a domain, all reads that fall into it are defined as "mapped" reads. Denote the set of mapped reads as M. All other (unmapped) reads constitute set U. For a domain classification tool, let the set of aligned reads for a domain be A. Thus, the sensitivity and FP rate of read classification for a domain are $\frac{A \cap M}{M}$ and $\frac{A-M}{U}$, respectively. Sensitivity of 100% indicates that all mapped reads can be aligned. A zero FP rate indicates that only mapped reads can be aligned to a domain.

Of the 2,181 annotated families, we evaluated the performance of HMMER and MetaDomain on 1406 families which have at least one mapped read. Of the 1406 tested domains, HMMER could not align any read to 1150 domains, resulting in zero sensitivity and FP rate. For the rest 256 domains, all aligned reads by HMMER are non-mappable reads, resulting in zero sensitivity and a positive FP rate. The comparison between HMMER and MetaDomain is summarized using a bubble chart in Figure 5. The biggest bubble indicates that HMMER has zero sensitivity and zero FP rate for 1150 domains. This experiment shows that it is highly difficult for HMMER to correctly align reads as short as 41 bp. There are two possible reasons for the low sensitivity of HMMER on short reads. First, the parameter training in E-value calculation of HMMER is based on much longer reads (100 amino acids). Thus, the small alignment scores generated by the short reads yield large E-values and cannot pass the E-value threshold. Second, the small alignment scores of short reads may not pass the filtration stage of HMMER.

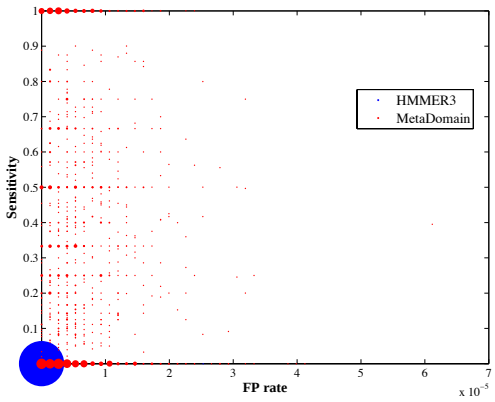

Fig. 5. Read classification sensitivity and FP rate of HMMER and MetaDomain. The size of each bubble represents the number of data points (i.e., domains) with the same sensitivity and FP rate.

4.1.2. *Identifying transcribed domains in the transcriptome*

Figure 5 only shows the read classification performance. MetaDomain uses both aligned read number and domain coverage as thresholds for domain identification. We expect that the additional constraint will reduce the false positive rate in domain identification. Because of the low read classification sensitivity, we speculate that HMMER will have low sensitivity in identifying transcribed domains.

Positive and negative test sets In order to quantify the performance of domain identification, we need to build positive and negative test sets, which include transcribed and non-transcribed domains based on mapped reads. There is no commonly accepted criterion to define transcribed genes using the number of mapped reads. Various expression scores such as an average coverage depth across the entire length of each gene[20] and reads per kilobase of exon model per million mapped reads (RPKM) are used to quantify transcriptional level. In addition, the cutoffs of defining highly transcribed, lowly transcribed, or non-transcribed genes are variable in different applications.[21] In this work, we define transcribed domains based on the rationale that a truly transcribed domain should be mapped by a number of reads at different positions. Correspondingly, we use the following criteria to determine whether a domain inside a gene is transcribed: 1) at least **N** reads are mapped to a domain; 2) at least 30% of positions in a domain are mapped by reads. A domain is labeled "non-transcribed" if the number of mapped read is zero. For domains that fall between the criteria for transcribed and non-transcribed domains, they are labeled "unknown" and are excluded from the test sets. Table 1 shows the size change of the positive and negative test sets over the cutoff N. Intuitively, bigger N creates an easier case for domain classification than smaller N.

Table 1. Number of transcribed and non-transcribed domains using different cutoffs (N) for the number of mapped reads

N	transcribed	unknown	none-transcribed
10	318	1317	546
15	262	1373	546
20	226	1409	546
25	195	1440	546
30	169	1466	546

Domain analysis using MetaDomain and HMMER We align all reads to the transcribed and non-transcribed domains using MetaDomain and HMMER. The "unknown" domains are removed due to their ambiguity. For HMMER, we first translated the short reads into peptide sequences using 6-frame translations. We then aligned the domains with the translated sequences using 1000 as the E-value threshold, which is chosen to maximize the sensitivity. For MetaDomain we directly aligned the short reads with the domains. The pipeline in Figure 4 was used to output a list of transcribed domains for MetaDomain. Let D^+ and D^- be the number of transcribed and non-transcribed domains identified using the read mapping results in Section 4.1.2. Let M^+ and M^- be the predicted number of transcribed and non-transcribed domains by MetaDomain or HMMER. The sensitivity and FP rate of domain classification tools are defined using the following equations:

$$\text{Sensitivity} = \frac{D^+ \cap M^+}{D^+}$$
$$\text{FP rate} = \frac{D^- \cap M^+}{D^-}$$

The values of D and M are affected by several options. First, D^+ and D^- can change over the cutoff N as shown in Table 1. Second, we used both the domain coverage and the number of aligned reads to determine whether a domain is encoded or transcribed. In this experiment, the cutoff for domain coverage is 30%, which we found reasonable across different

experiments. Thus, M^+ and M^- mainly change over the required number of aligned reads to a domain. For simplicity, we denote the cutoff as τ. Increasing τ implies a more stringent constraint for defining transcribed domains, and thus might result in lower sensitivity and a smaller FP rate. Decreasing τ is likely to increase the sensitivity while incurring a higher FP rate. In order to compare the performance of MetaDomain and HMMER under different τ, we plotted the ROC curves by changing τ from 1 to N for N=10, 20, and 30 in Figure 6.

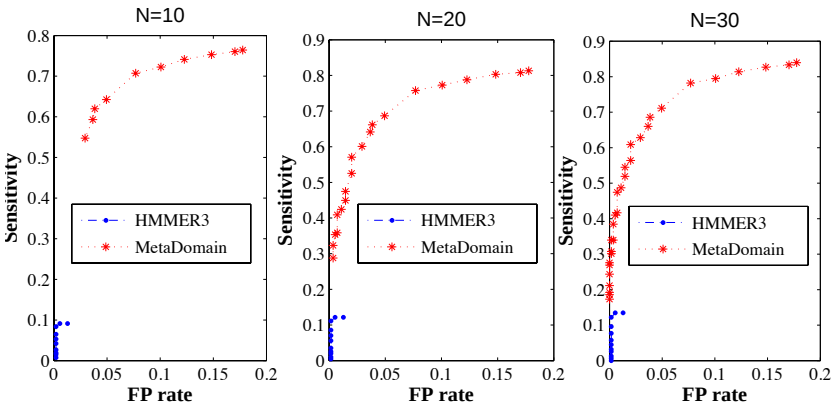

Fig. 6. ROC curves of HMMER and MetaDomain

Figure 6 shows that HMMER is highly specific (FP rate $\leq 1.3\%$). However, as we speculated, its sensitivity is low, with the highest sensitivity being only 0.135. HMMER misses a large portion of short reads that can be mapped to protein domains even when we use a very relaxed E-value cutoff. When both tools incur an FP rate of 0.02, the sensitivity of MetaDomain is 0.53 vs. 0.13 for HMMER. When N decreases from 30 to 10, the size of the positive test set D^+ becomes larger and the sensitivity of both HMMER and MetaDomain decreases. Note that the sensitivity and FP rate of HMMER keep the same for many different thresholds (i.e., τ), resulting in compact ROC curves. Overall, the ROC curves show that MetaDomain can achieve higher sensitivity while keeping a similar FP rate as HMMER for domain classification in this experiment. In addition, Figure 6 provides guidance on determining appropriate τ for MetaDomain in order to achieve desired sensitivity and FP rate.

On average, it took MetaDomain 280 seconds to align 752,156 reads with one domain on a 2.2GHz dual-core AMD Opteron machine. There are 2181 domains in this experiment.

4.2. *Protein domain analysis in a soil metagenomic data set*

In the first experiment, we demonstrated the accuracy of MetaDomain in classifying short RNA-seq reads into their native domain families. In this section, we present the utility of MetaDomain in identifying encoded protein domains in a complicated metagenomic data set, which is sequenced from the microbes dwelling in the soil from a long term cultivated corn site at Iowa using Illumina HiSeq platform [a]. There are 520,346,510 sequence reads of various lengths, ranging from 31 bp to 114 bp. The average length of the reads is $\sim$ 73 bp. Figure 7

[a]Sequenced by James Tiedje Lab at Michigan State University. Unpublished yet.

shows the distribution of the read lengths in this data set. The sheer amount of data and the complexity of this data set pose a great challenge for read assembly programs. Thus, we directly apply MetaDomain and HMMER to unassembled reads.

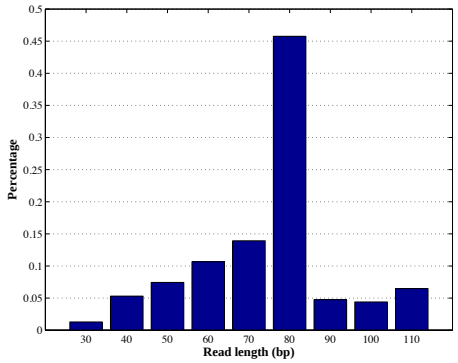

Fig. 7. Read length distribution in the soil data set.

We first used HMMER to align all the reads against the 2558 Pfam domains that contain the word "Bacteria" in their descriptions. Using E-value 1000, HMMER classified 34,602,784 reads into 2558 Pfam domains, accounting for 6.65% of all reads in the data set. The classifiable reads have an average length of $\sim$ 80 bp. A large number of reads shorter than 60 bp were not classified. By conducting a complementary domain analysis on short reads using MetaDomain, we expect to classify more reads to their native families. As many domains have been aligned with a large number of reads by HMMER, it is highly likely that they are encoded by the bacterial species in the soil data set. We thus excluded them from further screening. There are 80 domains with less than 20 reads aligned or with a smaller domain coverage than 30%. The left panel of Figure 9 shows an example of such domain. The small number of aligned reads and their biased distribution do not support the representation of this domain in this data set. We thus applied MetaDomain to the 80 domains and investigated whether they are encoded. On average, it took MetaDomain about 31 CPU minutes to align 11,194,176 sequences that were not classified by HMMER and are shorter than 60 bp against one domain family.

Figure 8 presents the number of aligned reads and the domain coverage output by HMMER and MetaDomain. Note that MetaDomain was only applied to reads that were not classifiable by HMMER. Thus, the total number of reads that can be classified into each domain should be the sum of the output of HMMER and MetaDomain. This figure shows that significantly more reads can be classified into the corresponding domains. We need to specify thresholds for domain coverage and the number of aligned reads in order to define an encoded domain. Similar to previous experiments, the domain coverage cutoff is 30%. According to Figure 6, the sensitivity and FP rate of MetaDomain are 0.34 and 0.004 when τ is 20. We thus choose 20 as the cutoff for the number of aligned reads. Out of the 80 Pfam domains, 24 have less than 20 reads aligned or a domain coverage no greater than 30%. So these 24 domains are not likely to be encoded in this data set. For the other 56 domains, their average domain coverage by MetaDomain alone is 97.25%. The number of aligned reads by MetaDomain is 169.52 versus 15.27 by HMMER. This provides strong evidence that these protein domains are actually

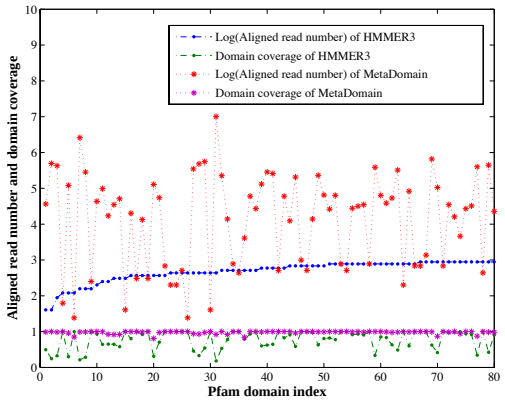

Fig. 8. Reads aligned by HMMER and MetaDomain.

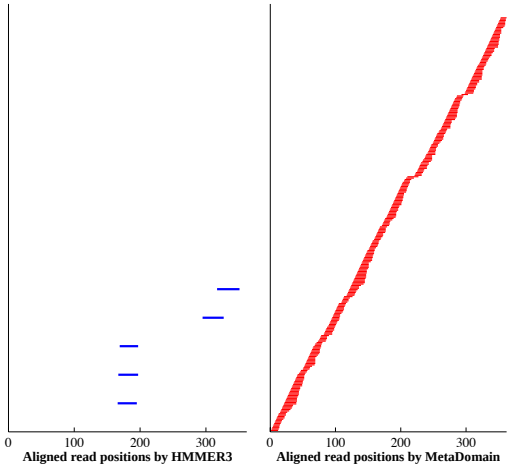

Fig. 9. The distributions of aligned reads for PF09703 by HMMER and MetaDomain.

encoded. The average read length aligned by MetaDomain is 38 bp. As an example, Figure 9 shows the distribution of aligned reads by HMMER and MetaDomain for the domain PF09703. In summary, by using MetaDomain, we are able to identify 56 more domains encoded in this data set. 130,930 (0.025%) more reads are classified into these domain families.

Among these 56 protein domains, 21 have unknown functions. 6 domains are CRISPR-associated domains. These kinds of domains are found in the genomes of approximately 40% of bacteria and 90% of archaea. More detailed analysis is needed to understand whether the functions of these domains are important to the specific habitat.

5. Conclusion and future work

In this work, we introduce MetaDomain, a protein domain classification tool for short reads produced by next-generation sequencing technologies. It provides a complementary domain classification tool to HMMER on classifying short reads into domain families with low sequence identity. Our experimental results show that it can achieve a better tradeoff between sensitivity and FP rate than HMMER in classifying short sequences. Its current version is based on a faithful implementation of Viterbi and is slow when applied to thousands of millions of reads and the whole Pfam database. We plan to improve its efficiency by using filtration strategies such as ungapped alignment and parallel programming. In addition, we plan to improve the method of designing position-specific score thresholds in order to achieve a better discrimination power.

6. Acknowledgements

We would like to thank Dr. James Tiedje and Dr. James Cole for providing us the soil metagenomic data set from the cultivated corn site. The sequencing efforts were supported by JGI and DOE.

References

1. "CAMERA: Community cyberinfrastructure for advanced microbial ecology research and analysis." `http://camera.calit2.net/`.

2. F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. Edwards, "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, no. 1, p. 386, 2008.
3. D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377–386, 2007.
4. S. F. Altschul, G. Warren, W. Miller, E. W. Myers3, and D. J. Lipmanl, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.
5. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*. UK: Cambridge University Press, 1998.
6. "HMMER3: a new generation of sequence homology search software." `http://hmmer.janelia.org/`.
7. 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. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D211–D222, 2010.
8. S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nat Meth*, vol. 5, no. 1, pp. 16–18, 2008.
9. K. Ellrott, L. Jaroszewski, W. Li, J. C. Wooley, and A. Godzik, "Expansion of the protein repertoire in newly explored environments: Human gut microbiome specific protein families," *PLoS Comput Biol*, vol. 6, no. 6, p. e1000798, 2010.
10. A. Schlüter, L. Krause, R. Szczepanowski, A. Goesmann, and A. Pühler, "Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant," *Journal of Biotechnology*, vol. 136, no. 1-2, pp. 65–76, 2008.
11. L. Krause, N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye, "Phylogenetic classification of short environmental DNA fragments," *Nucleic Acids Research*, vol. 36, no. 7, pp. 2230–2239, 2008.
12. F. Schreiber, P. Gumrich, R. Daniel, and P. Meinicke, "Treephyler: fast taxonomic profiling of metagenomes," *Bioinformatics*, vol. 26, no. 7, pp. 960–961, 2010.
13. Y. Zhang and Y. Sun, "HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors," *BMC Bioinformatics*, vol. 12, no. 1, p. 198, 2011.
14. F. Weng, C. H. Su, M. T. Hsu, T. Y. Wang, H. K. Tsai, and D. Wang, "Reanalyze unassigned reads in sanger based metagenomic data using conserved gene adjacency," *BMC Bioinformatics*, vol. 11, no. 1, p. 565, 2010.
15. J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic Acids Research*, vol. 36, no. 16, p. e105, 2008.
16. K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Research*, 2010.
17. D. R. Yoder Himes, P. S. G. Chain, Y. Zhu, O. Wurtzel, E. M. Rubin, J. M. Tiedje, and R. Sorek, "Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing," *Proceedings of the National Academy of Sciences*, 2009.
18. "IMG: Integrated microbial genomes." `http://img.jgi.doe.gov/`.
19. B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, p. R25, 2009.
20. K. D. Passalacqua, A. Varadarajan, B. D. Ondov, D. T. Okou, M. E. Zwick, and N. H. Bergman, "Structure and complexity of a Bacterial transcriptome," *J. Bacteriol.*, vol. 191, no. 10, pp. 3203–3211, 2009.
21. B. A. Williams, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat Meth*, vol. 5, no. 7, pp. 621–628, 2008.