IDBA-MTP: A Hybrid MetaTranscriptomic Assembler Based on Protein Information

Henry C.M. Leung, S.M. Yiu, Francis Y.L. Chin

Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
{cmleung2, smyiu, chin}@cs.hku.hk

Abstract. Metatranscriptomic analysis provides information on how a microbial community reacts to environmental changes. Using next-generation sequencing (NGS) technology, biologists can study microbe community by sampling short reads from a mixture of mRNAs (metatranscriptomic data). As most microbial genome sequences are unknown, it would seem that de novo assembly of the mRNAs is needed. However, NGS reads are short and mRNAs share many similar regions and differ tremendously in abundance levels, making de novo assembly challenging. The existing assembler, IDBA-MT, designed specifically for the assembly of metatranscriptomic data only performs well on high-expressed mRNAs.

This paper introduces IDBA-MTP, which adopts a novel approach to metatranscriptomic assembly that makes use of the fact that there is a database of millions of known protein sequences associated with mRNAs. How to effectively use the protein information is non-trivial given the size of the database and given that different mRNAs might lead to proteins with similar functions (because different amino acids might have similar characteristics). IDBA-MTP employs a similarity measure between mRNAs and protein sequences, dynamic programming techniques and seed-and-extend heuristics to tackle the problem effectively and efficiently. Experimental results show that IDBA-MTP outperforms existing assemblers by reconstructing 14% more mRNAs. **Availability:** www.cs.hku.hk/~alse/hkubrg/

Keywords: metatranscriptomic reads, assembling, next-generation sequencing, protein sequence alignment

1 Introduction

The traditional approach for studying microorganisms is to isolate and cultivate each single microorganism and then study its behavior, such as gene expression levels, under different environments. As different microbes usually live together to form a microbial community, isolating a single microbe is usually impossible [4] and, even possible, changes the microbe's living behavior in a microbial community. Metatranscriptomic studies in the past have been based on microarrays or cDNA clone libraries [2,23,29]. The microarray-based approach [17] requires knowledge of target mRNA sequences, which limits its usefulness in relation to novel mRNAs.

adfa, p. 1, 2011.

cDNA clone libraries, on the other hand, can be applied to novel mRNAs, but the approach is labor-intensive and the estimations of expression levels of mRNAs are inaccurate.

High-throughput next-generation sequencing (NGS) technology [3,7,22,18] introduces a new and better approach for studying metatranscriptomic data. By sequencing reads from mRNA sequences of a sample, scientists can reconstruct novel mRNA sequences by assembling reads and can estimate the expression levels of each mRNA by the number of reads aligned to the mRNA sequence. Currently, there are two main NGS technologies for metatranscriptomic data: pyrosequencing technology and synthesis technology. Pyrosequencing technology [6,8,24,30] produces long reads (of length about 400 bp) with relatively higher cost (over 40 times higher for the same throughput). Since the read length is long, no or limited assembly is required. Pyrosequencing technology has achieved promising results for soil samples [30] and marine samples [6,8]. Synthesis technology, on the other hand, produces relatively short reads (of length varying from 75 bp to 150 bp) at much lower cost. Since the length of reads produced by synthesis technology is much shorter than the length of the mRNA sequence (about 1000 bp), the reads need to be assembled into longer sequences (contigs) before analysis.

Compared with assembling genomic, transcriptomic or metagenomic data, assembling metatranscriptomic data is much more difficult because of the following reasons.

1. *Repeat patterns across different mRNAs.* Repeat patterns usually introduce ambiguity during assembly and are a common problem in all types of assembling. However, the problem is more serious in metatranscriptomes than in other data. Many genes exist in multiple species with similar functions and the resultant proteins share common protein domains [9]. As a result, in the metatranscriptomic data, many different mRNAs have similar patterns. According to analysis of genBank [1], based on known gene information, 24.53% of bacteria genes contain at least one repeat pattern of length longer than 100 bp (note that, in this analysis, different versions of the same genes from the same bacteria were ignored and only the repeat patterns in genes from different bacteria were considered). In these circumstances, assemblers, not specially designed for metatranscriptome data, produce either short contigs or chimeric contigs that merge mRNA sequences from more than one gene [15]. This is consistent with our experiments (see Table 2): these assemblers can either only recover 31% of mRNAs with average contig length of 194bp and 4.14% error rate (Oasis), or recover more mRNAs (59.29%) with longer average contig length (395bp) but the error rate is increased to 10.73% (IDBA-UD).
2. *Extreme differences in abundances.* For the DNA genome assembly problem of a single species, this is not a problem because there is one abundance only. On the other hand, in transcriptomic data and metagenomic data, since the abundances of different mRNAs and the number of genomes vary (can be 100 times and 1,000 times different, respectively [25]) because of different expression levels and abundances of species, erroneous reads cannot be identified easily by sampling

rates. In metatranscriptomic data, this problem becomes more serious. Since both the abundances of species and the expression levels of mRNAs from the same species may vary, the abundances of different mRNAs can vary much more significantly (over 100,000 times). Thus low-expressed mRNA sequences are very difficult to reconstruct as correct reads from these sequences and erroneous reads are very difficult to distinguish. As Table 1 shows for our experiments on low-expressed mRNAs, the performance of existing assemblers suffers.

Thus, existing assemblers for genomic, transcriptomic and metagenomic data do not perform well on metatranscriptomic data especially for the low-expressed transcriptomes [15]. To our best knowledge, IDBA-MT [15] is the only assembler designed for metatranscriptomic data. IDBA-MT aims at solving the repeat pattern problem. By applying information from paired-end reads, IDBA-MT can resolve some of the chimeric contigs (See Table 2, IDBA-MT can recover more mRNAs while decreasing the error rate from about 10% to 5% when compared to IDBA-UD). However, this approach can only work for high-expressed transcriptomes with high sequencing depths as it relies on paired-end data and fails when there are insufficient sampling reads from the mRNAs (i.e., low-expressed mRNAs).

Similar to genome assembly, besides de novo assembly, one can apply the reference-based approach. Existing work tries to reconstruct mRNAs by aligning metatranscriptomic reads to known genomes or gene DNA sequences. However, this approach has had only limited success [32] as the genomes of most microbes are still unknown [4] and the microbe gene sequences mutate frequently.

Our observations on the reference-based approach: Although the aforementioned reference-based approach has limited success, about 60% to 70% of the proteins in bacteria have similar sequences as some known proteins [5, 30], thus known reference protein sequences could help in the assembling of novel mRNAs. There are two difficulties to resolve in order to make use of the protein sequences. First, we need to consider amino acid instead of nucleotides. Even if we consider amino acid, it is not trivial due to the following. For proteins with similar functionalities, even though their structures are similar and their sequences share some conserved regions, the amino acid sequences corresponding to these conserved regions might not be exactly the same. Second, to consider amino acid, the information contained in a single read becomes much less (3 nucleotides converted to 1 amino acid). Since one read only corresponds to about 25 amino acids (aa), it is difficult to have a confident alignment [32]. Another approach is to align contigs, instead of reads, to proteins. However, as the performance of existing assemblers is not good, the resultant contigs are short or incorrect and not many confident alignments can be obtained.

Our contributions: To overcome the first problem of amino acid similarity, we found that even though the amino acid sequences may not be exactly the same, it is known that some amino acids, though different, have similar chemical properties and functionality [11]. Consequently, the mRNA can be reconstructed using the approach of first decoding the reads into peptide sequences and then aligning these peptide sequences to protein sequences based on the similarity of amino acids (e.g. Blosom

62). Thus, we incorporate the similarity of amino acids into our alignment algorithm. To solve the problem of short reads and low-expressed mRNAs, we make use of the paths of the de Bruijn graph with a small k.

Our proposed assembler, IDBA-MTP, reconstructs mRNA sequences from metatranscriptomic reads, especially for low-expressed mRNAs, using the information of known microbial protein sequences to guide the construction of contigs as follows. IDBA-MTP first constructs a de Bruijn graph from the input reads using a relatively small k (k = 21 bp) to compensate for the missing long k-mers in low-expressed mRNAs. Since k is small, the de Bruijn graph, though connected, has many branches representing repeat regions in the mRNA sequences (due to problem 1 and 2) and with each mRNA represented by one of its paths. In order to determine whether a path represents an mRNA sequence or not, IDBA-MTP will decode the path into a peptide sequence and then align it with known protein sequences. Those paths, which can be aligned to known protein sequences, should be potential candidates for mRNA sequences depending on their similarity and sequencing depths. However, since the number of paths is huge (many paths will not represent any mRNA sequences) and the alignment with the protein sequences is not straightforward because of the similar chemical properties of amino acids, a dynamic programming approach with a seed-and-extend (with the seed derived from the known protein sequences) heuristics is employed to reduce the complexity of the problem.

Since the candidate mRNA sequences are constructed by aligning known protein sequences, mRNA sequences for novel proteins cannot be reconstructed using this approach. An intuitive idea is to run IDBA-MT for novel mRNAs, then combine the results of IDBA-MT and the output from our reference-based approach. However, some mRNAs sequences may be reconstructed by both approaches, which results in redundant or similar contigs. To prevent having redundant contigs, IDBA-MTP will treat those mRNAs sequences reconstructed by alignment of known reference proteins as long input reads for IDBA-MT, i.e. the output of the first approach will be the input of the second approach. Experiments on simulated data show that even though 48% regions of the mRNAs can be aligned to known reference proteins, existing assemblers can only reconstruct contigs representing at most 62.9% of these regions. IDBA-MTP can reconstruct contigs covering 77.6% of these regions and some novel mRNAs using protein reference sequences. As a result, IDBA-MTP can reconstruct 14% more mRNAs (in term of the total length of mRNAs) than existing assemblers.

The paper is organized as follows. The IDBA-MTP algorithm is described in Section 2. Experimental results for IDBA-MTP and other existing assemblers on both simulated and real metatranscriptomic data are presented in Section 3. Conclusions are drawn on the performance of IDBA-MTP in Section 4.

2 Methodology

Given a set of reads sampled from a set of mRNA sequences (with nucleotides A, C, G and U), we can construct a de Bruijn graph where each vertex v represents a length-k substring (k-mer) of the reads and where an edge connects vertex u to vertex v if and only if the corresponding k-mers for vertex u and vertex v overlap at $k-1$ positions and appear in a read. An mRNA sequence can be represented by a path of k-mers in the de Bruijn graph. Since there are many paths in the de Bruijn graph and most of them do not represent any mRNA, a correct mRNA sequence R can be reconstructed from the de Bruijn graph if some known protein sequence P can be aligned to the path. If the alignment similarity between R and P is high, R will likely be an mRNA sequence in the sample.

A protein or peptide sequence is represented by a sequence of amino acids (of which there are 20 kinds). Given a length-3m mRNA sequence R, we can decode it into a length-m sequence $D(R)$ of amino acids by converting each non-overlapping coden (length-3 substring) in R into an amino acid character. Given a protein sequence P and an mRNA sequence R, P and $D(R)$ can be aligned by inserting space characters in P and $D(R)$ to form P' and $D(R)'$ of equal length respectively, and the similarity score based on this alignment is defined as follows:

$$score_a(P', D(R)') = \sum \delta(P'[i], D(R)'[i]) + p_{\text{open}} \cdot \text{number of gaps} \quad (1)$$

where $P'[i]$ and $D(R)'[i]$ are the i-th amino acid in P' and $D(R)'$ respectively, $\delta(x,y)$ is the similarity score between amino acids x and y (which depends on their chemical properties and roles in the protein's functionality), p_{open} is the gap penalty and a gap is defined as consecutive space characters in P' or $D(R)'$ (the gap penalty can be refined to take the gap size into consideration). Note that the similarity score $\delta(x,y)$ can be negative and is $-\infty$ whenever a stopping coden in $D(R)'$ is compared to space or any amino acid in P'. The optimal global similarity score between P and $D(R)$ is defined as the highest similarity score of all alignments of P and $D(R)$.

$$score_g(P, D(R)) = max_{\text{all alignment } P' \text{ and } D(R)'}\{score_a(P', D(R)')\} \quad (2)$$

Since the decoded protein from an mRNA usually does not exist in the protein database but some part of the decoded protein sequence might match with some regions of some proteins in the database because of their functional similarity, instead of aligning the whole sequence of P and $D(R)$, the optimal local alignment between all substrings of P and $D(R)$ is considered in IDBA-MTP and this information, in terms of contigs, will be needed for mRNA assembly later (see Section 2.3). The optimal local similarity score is defined as:

$$score_l(P, D(R)) = max_{\text{all substrings } p_s \text{ and } d_s \text{ of P and } D(R)}\{score_g(p_s, d_s)\} \quad (3)$$

The **Protein-Graph Alignment (PGA) Problem** can be defined as follows: given a de Bruijn graph G and a protein P, find a path in G (representing a substring in an mRNA sequence R) such that $score_l(P, D(R))$ is maximized.

2.1 Dynamic Programming

The PGA problem can be solved by dynamic programming based on the principle of optimality. Consider an optimal global alignment *Opt* of the substring d_s (represented by a path $Q(d_s)$) of the decoded protein $D(R)$ for an mRNA sequence R with the substring p_s of protein sequence P. The same alignment *Opt* for any subpath of $Q(d_s)$ and the corresponding substring of p_s should also be optimal.

Let $S(v, i)$ define the maximum global similarity score between a substring of P ending at $P[i]$ and all decoded sequences $D(R)$ for path R in the de Bruijn graph G ending at vertex v. Similarly, we define $S_M(v, i)$, $S_P(v, i)$ and $S_R(v, i)$ to be the maximum global similarity score with the following restrictions respectively: (1) $P[i]$ is aligned with the last amino acid of the corresponding protein sequence decoded from the path ending at vertex v, (2) $P[i]$ is aligned with the space character and (3) the last amino acid of the corresponding protein sequence decoded from the path ending at vertex v is aligned with the space character. The value of $S(v, i)$ is the maximum of 0 (alignment of two null substring), $S_M(v, i)$, $S_P(v, i)$ and $S_R(v, i)$. The value of $S_M(v, i)$, $S_P(v, i)$ and $S_R(v, i)$ can be calculated by considering the alignment of the last coden, any length-3 path $s \rightarrow v$ with $D(s \rightarrow v)$ represent the decoded amino acid of path $s \rightarrow v$, and the subproblem of alignment ending as vertex s.

$S(v, i)$, $S_M(v, i)$, $S_P(v, i)$ and $S_R(v, i)$ can be calculated as follows:

$$S(v,i) = \begin{cases} 0 & \text{no path ending at } v \text{ can be decoded to an amino acid} \\ \max\{0, S_M(v,i), S_P(v,i), S_R(v,i)\} & \text{otherwise} \end{cases}$$

$$S_M(v,i) = \begin{cases} -\infty & \text{no path ending at } v \text{ can be decoded to an amino acid} \\ S(s, i-1) + \delta\big(P[i], D(s \rightarrow v)\big) & \text{otherwise} \end{cases}$$

$$S_P(v,i) = \begin{cases} -\infty & i = 0 \\ \max\{S_P(v, i-1), S_M(v, i-1) + p_{\text{open}}\} + \delta(P[i], space) & \text{otherwise} \end{cases}$$

$$S_R(v,i) = \begin{cases} -\infty & \text{no path ending at } v \text{ can be decoded to an amino acid} \\ \max\{S_P(s,i), S_M(s,i) + p_{\text{open}}\} + \delta(space, D(s \rightarrow v)) & \text{otherwise} \end{cases}$$

If $D(s \rightarrow v)$ represents the stopping coden, $\delta(D(s \rightarrow v), x) = -\infty$. $\max_{v,i}\{S(v, i)\}$ represents the optimal local similarity score and the corresponding aligned mRNA sequence can be obtained by backtracking. Note that care should be taken for the starting vertex of the path. Since the starting vertex of a path in de Bruijn graph represents the length-k prefix of an mRNA and each subsequent vertex represents an extra nucleotide of the mRNA, we modify zero in-degree vertices in the de Bruijn graph implicitly such that each vertex only represents one single nucleotide (the last nucleotide of the k-mer) of an mRNA. Note that since the protein sequence P is fixed, the dynamic programming is correct even there is loop in the de Bruijn graph.

Since there are at most $4^3 = 64$ length-3 paths $s \rightarrow v$ to a vertex v, each entry $S(v, i)$, $S_M(v, i)$, $S_P(v, i)$ and $S_R(v, i)$ can be computed in constant time by preprocessing. The time complexity for aligning a length-$|P|$ protein P is $O(n|P|)$ and for a set of protein sequences with total length m is $O(nm)$, where n is the total number of vertices in the de Bruijn graph.

2.2 Seed-and-Extend Heuristic

Although the dynamic programming approach can solve the PGA problem in $O(nm)$ time, n and m are usually large for real biological data (in the order of millions and thousand millions respectively) and the running time for the above dynamic programming approach is too long for practical use. In order to speed up the running time, IDBA-MTP applies on seed-and-extend heuristic to speed up the process. Assume that the optimal local alignment of an mRNA and a protein has at least one aligned region with t consecutive matches of amino acids (with similarity score larger than a predefined threshold), the PGA problem can be solved by a seed-and-extend heuristic. Given a simple path (a path with all intermediate vertices have exactly one incoming and one outgoing edge) or a k-mer in the de Bruijn graph representing a length-t peptide (sequence of amino acids), the reference protein sequences containing this peptide can be obtained in constant time after $O(m)$ preprocessing, where m is the total length of the reference proteins. By considering these positions as the starting alignment positions (seeds) and extending the alignment in both forward and backward directions using dynamic programming, a small subset of paths containing the seed as a subpath will be considered and the running time can be greatly reduced in practice.

2.3 Preventing Redundant mRNAs

As some reference proteins could have similar sequences, these similar proteins might align to overlapping paths in the de Bruijn graph and similar mRNA sequences may be obtained. Among these similar mRNA sequences, it is likely that only one of them is correct while the others are only artifacts caused by sequencing errors or misalignment. However, duplicate genes and genes with similar functions in different species may also introduce similar mRNA sequences. IDBA-MTP applies two techniques to remove artifacts. The first approach is to prevent aligning multiple proteins with seeds on the same simple path in the de Bruijn graph. Simple paths in the de Bruijn graph are sorted in decreasing order of lengths and are considered one by one. Once a protein is aligned to a path R (with the maximum alignment score among all proteins) in the de Bruijn graph, all substrings in R are removed from the seed table and will not be considered as starting positions for alignment. Note that these simple paths could still be considered when extending the alignment of other proteins using dynamic programming. Although the first approach can determine some redundant contigs represent the same mRNAs, sequence error could introduce error paths in the de Bruijn graph result as alignment of similar proteins to overlapped but similar paths in the de Bruijn graph. In our experiment, there can be 50 similar

paths represented by the correct and erroneous paths corresponding to the same mRNA. Thus, we should not output the aligned mRNAs directly. The second approach was considering these mRNAs as long reads and treating them as input to IDBA-MT for de novo assembly. By using these extra long reads, paired-end reads and sequencing depths information, IDBA-MT avoids assembling redundant mRNAs and can reconstruct novel mRNAs with no similar reference proteins.

3 Experiments

We compared the performances of Oases [26], Trinity [10], IDBA-UD [21], IDBA-MT[15] and IDBA-MTP on a real dataset from mouse gut [32] and two simulated datasets generated from known bacteria gene sequences obtained from genBank [1]. Oases and Trinity were designed for assembling transcriptomic data, IDBA-UD for assembling metagenomic data, and IDBA-MT for assembling metatranscriptomic data. All bacteria gene sequences with known sources in the genBank were downloaded. To prevent selecting mRNAs from the same species (either from the same or different strains), duplicated sequences were removed and only one version was kept. Note that similar mRNAs obtained from different bacteria would be kept. Similar to [15], mRNAs sharing at least half of the sequences with other mRNAs were selected for generating a difficult dataset (mRNAs which do not share common sequence regions with others would be isolated in the de Bruijn graph and can be assembled easily). The resultant 658 mRNA sequences were used to generate the simulated data. Although the number of mRNA sequences selected is small compared with the real experiments, this small subset of mRNAs sequences with long repeats represents the most difficult part of assembling metatranscriptomic data. The reference bacteria protein sequences for IDBA-MTP was downloaded from NCBI database and we used the Blosum-62 scoring matrix, open gap penalty = -10 – (-1) = -9 and gap extend penalty = -1 for calculating the similarity scores of protein sequences. In all experiments on simulated data, all the corresponding protein sequences of the 658 mRNA sequences were removed from the reference protein sequences for testing the performance of IDBA-MTP.

For each simulated dataset, we randomly picked length-75 bp paired-end reads from the RNA sequence with 1% sequencing error according to the predefined abundances. The mean insert distance of each paired-end read was 200 bp with a standard deviation of 10 bp. Two sets of simulated data were generated: (1) Low abundance - 100 mRNAs were sampled with 3x sequencing depth for evaluating the performances of the assemblers for mRNAs with low expression levels. (2) Mixture abundance - 658 mRNAs were sampled from 1000x to 3x sequencing depth with the number of mRNAs following the power law (number of mRNAs with a certain abundance is directly proportional to the negative of abundance ratio) for evaluating the performances of the assemblers for mRNAs with different expression levels.

All assemblers were tested on simulated data using default parameters. Each contig produced by the assemblers was aligned to the 658 mRNAs in the samples using Blat [13]. A contig was considered correct if and only if at least 95% of the contig region

Table 1. Experimental Result on simulated data with low abundance ratios

Softare	Coverage	Max. Len.	Avg. Len.	# of wrong contig (len.)	# of correct contig (len.)	Error Rate
Oases	25.99%	524 bp	172 bp	9 (1,063 bp)	149 (25,690 bp)	3.97%
Trinity	9.85%	497 bp	287 bp	17 (7,362 bp)	34 (9,837 bp)	42.80%
IDBA-UD	48.26%	783 bp	342 bp	8 (4,425 bp)	83 (28,480 bp)	13.45%
IDBA-MT	52.68%	900 bp	279 bp	8 (3,194 bp)	136 (37,993 bp)	7.75%
IDBA-MTP	66.00%	916 bp	273 bp	5 (1,057 bp)	156 (42,771 bp)	2.40%

Table 2. Experimental Result on simulated data with mixed abundance ratios.

Softare	Coverage			Max. Len.	Avg. Len.	# of wrong contig (len.)	# of correct contig (len.)	Error Rate
	total	≤5x	> 5x					
Oases	31.00%	22.46%	8.45%	676 bp	194 bp	63 (8,471 bp)	1009 (196,162 bp)	4.14%
Trinity	15.10%	11.28%	3.80%	1,270 bp	319 bp	106 (75,713 bp)	310 (99,603 bp)	43.18%
IDBA-UD	59.29%	42.74%	16.38%	1,430 bp	395 bp	43 (28,837 bp)	606 (239,887 bp)	10.73%
IDBA-MT	64.07%	46.53%	17.37%	1,511 bp	310 bp	37 (18,023 bp)	1005 (312,500 bp)	5.45%
IDBA-MTP	69.62%	51.29%	18.33%	1,615 bp	368 bp	41 (23,461 bp)	1127 (415,813 bp)	5.34 %

could be aligned to the mRNA sequence with 95% similarity. Some short, even correct, contigs which could not align confidently to the 658 mRNAs were considered incorrect. Regions of mRNAs aligned by correct contigs were considered covered and the coverage of an assembler was calculated as the ratio of regions in the mRNAs covered by the contigs produced by the assembler. Although Oases, Trinity, IDBA-UD could produce scaffolds using paired-end reads, the scaffolds performed worse than the contigs in all simulated data because these assemblers connected contigs wrongly and produced long but incorrect scaffolds. Thus, we compared the performances of the assemblers based on the resultant contigs and the experimental results are shown in Table 1 and 2.

3.1 Low Abundance mRNAs

When the abundances of mRNAs were low, Oases and Trinity did not perform well in assembly because of the low sequencing depths and the similarity of mRNAs. Oases tended to produce confident but shorter contigs. As a result, it had a low error rate (3.97%) but the lengths of contigs were short (average length = 172 bp) and the coverage was not high (25.99%). Since Trinity was designed for assembling transcriptomic data for eukaryotic mRNAs and was not suitable for assembling prokaryotic mRNAs, the error rate of Trinity was high (42.80%) and the coverage was low (9.85%). IDBA-UD, which was designed for assembling metagenomic data, performed better than Oases and Trinity because it applied various technologies, e.g. multiple *k*-mers, local assembling and local coverage of contigs for assembling reads sampled from low abundance genomes (mRNAs in this case). However, since the mRNAs had many similar sequences, IDBA-UD could not determine these chimeric contigs and the error rate was high (13.45%) but the coverage was acceptable (48.26%). IDBA-UD has such high error rate because it merged two or more mRNA sequences incorrectly to produce chimeric contigs. IDBA-MT, which was designed

Table 3. Experimental Result on real mouse gut data

Softare	Maximum Length	Average Length	Contigs number	Total Length	# of contig aligned to known proteins (length)
Oases	693 bp	127 bp	99,611	12,655,199 bp	489 (84,044 bp)
Trinity	15,857 bp	500 bp	19,721	9,862,469 bp	7,188 (2,994,588 bp)
IDBA-UD	10,741 bp	490 bp	18,951	9,287,101 bp	9,510 (4,178,162 bp)
IDBA-MT	8,863 bp	490 bp	18,972	9,301,484 bp	9,515 (4,181,949 bp)
IDBA-MTP	9,070 bp	477 bp	20,062	9,581,626 bp	10,429 (4,712,857 bp)

for assembling metatranscriptomic data, outperformed IDBA-UD because it used paired-end reads information to resolve chimeric contigs. It achieved a relatively high coverage (52.68%) with low error rate (7.75%). With the information from known protein sequences, IDBA-MTP further improved the coverage to 66.00% and had the lowest error rate (2.40%).

3.2 mRNAs with different abundances

For the simulated data with mixed abundances, the overall performance of the assemblers improved because of the mRNAs with high abundances. We have also analysed the coverage of low-abundance mRNAs (76% mRNA with sequencing depth $\leq$ 5x) and high-abundance mRNAs (24% mRNA with sequencing depth > 5x). As expected, the high-abundance mRNAs had better overall results than the low-abundance mRNAs. Again, Oases produced short but confident contigs, achieved higher coverage (31.00%) than Trinity and had the lowest error rate (4.14%). Trinity, which assembled many long and wrong contigs, had the lowest coverage (15.10%) and the highest error rate (43.18%). IDBA-UD had higher coverage (59.29%) and moderate error rate (10.73%). By resolving some chimeric contigs, IDBA-MT had slightly higher coverage (64.07%) and lower error rate (5.45%) than IDBA-UD. IDBA-MTP had the highest coverage (69.62%) and a low error rate (5.34%). Considering the performance of mRNAs with different abundances, IDBA-MTP could reconstruct 5% and 1% more mRNAs with low and high abundances respectively than the best existing assembler IDBA-MT. By using protein reference information, the performance of IDBA-MTP improved not only for the low-abundance mRNAs, but also for the high-abundance mRNAs.

3.3 Real metatranscriptomic data

Xiong et al. [32] isolated mRNAs from the lumen of the cecum and colon of 4 mice at 12 weeks old, colonized with an Altered Schaedler flora (ASF) containing eight known species without reference genomes. A total of 3.3 million paired-end reads were generated using Illumina sequencing technology. The read length was about 75 bp and the insert distance was about 300 bp. Similar to [15], we merged the reads sampled from the 4 mice into a single dataset as the number of reads in each sample was small. The reads were inputted to existing assemblers for comparison. Since there were no reference genomes, we evaluated the accuracy of output contigs by aligning them to

known protein sequences using Blastx with default parameters. A contig was considered "correct" if at least 90% of the contig sequence could be aligned to a single protein sequence. We used number of aligned contigs instead of number of aligned proteins to evaluating the result because a contigs can be aligned to hundred of similar proteins and it is difficult to evaluate the softwares based on the number of discovered proteins. Noted that IDBA-UD, IDBA-MT and IDBA-MTP consider each *k*-mer in the de Bruijn should belong to at most one contigs, they should not output redundant contigs represents the same protein or protein regions.

Similar to simulated data, Oases produced very short contigs. Since it was difficult to obtain confident alignment for short contigs, only 489 (out of 99,611) contigs produced by Oases could be aligned to known protein sequences. Trinity produced longer contigs than other assemblers. However, over half of them (7,188 out of 19,721 can be aligned) could not be aligned to known protein sequences although the contigs were long enough for confident alignment. The performances of IDBA-UD and IDBA-MT were similar with half of the contigs aligned to known protein sequences. IDBA-MTP produced a thousand more contigs than IDBA-UD and IDBA-MT. Since the extra contigs constructed mainly due to using protein reference sequences, most of these extra contigs could be aligned to known protein sequences.

4 Conclusions

Existing assemblers do not perform well on metatranscriptomic data, especially on low-expressed mRNAs. In this paper, we have proposed IDBA-MTP to assemble mRNAs, making use of information from the database of millions of known protein sequences. In particular, dynamic programming technique with a seed-and-extend heuristics was introduced to reconstruct mRNA sequences from paths in the de Bruijn graph with maximum similarity scores when aligned with the known protein sequences. Experimental results on both simulated and real biological data showed that IDBA-MTP outperformed existing assemblers on metatranscriptomic data.

However, when applying IDBA-MTP on metatranscriptomic data, there is an issue of running time when compared with existing assemblers. Since the reference proteins databae is big and highly redundant, i.e. many proteins with very similar sequences exist, IDBA-MTP takes one or two days for aligning reference proteins to de Bruijn graph even using the seed-and-extend heuristic. This is much longer that existing assemblers which takes one or two hours to assemble the reads. Although it may not be a problem at current state because it takes weeks to generate a metatranscriptomic dataset, further research should be performed to increase the speed of IDBA-MTP by preprocessing the reference proteins or parallel processing.

The techinque of assembly based on known protein sequence information is applicable not only on metatranscriptomic data. It can also improve the performance on transcriptomic data of single species. We plan to study the usage of protein reference sequence information on transcriptomic assembly of single species.

5 Acknowledgement

This work was supported by Hong Kong GRF HKU 7111/12E, HKU 719709E and 719611E, Shenzhen basic research project (NO.JCYJ20120618143038947) and NSFC(11171086).

References

1. D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, B. Rapp and D. Wheeler, "GenBank," Nucleic Acids Research, 2000, 28(1), pp. 15-18.
2. C. Booijink, J. Boekhorst, E. Zoetendal, H. Smidt, M. Kleerebezem and W. de Vos, "Metatranscriptome Analysis of the Human Fecal Microbiota Reveals Subject-Specific Expression Profiles, with Genes Encoding Proteins Involved in Carbohydrate Metabolism Being Dominantly Expressed," Appl. Environ. Microbiol, 2010, 76(16), pp. 5533–5540.
3. J. ten Bosch and W. Grody, "Keeping up with the next generation: massively parallel sequencing in clinical diagnostics," J Mol Diagn, 2008, 10, pp.484-492.
4. J. Eisen, "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes," PLoS biology, 2007, 5(3), pp. e82.
5. R. Finn, J. Tate, J. Mistry, et al., "The Pfam Protein Families Database, " Nucleic Acids Research, 2000, 28 (1), pp. 263–266.
6. J. Frias-Lopez, Y. Shi, G. Tyson, et al., "Microbial community gene expression in ocean surface waters," Proc Natl Acad Sci, 2008, 105, pp. 3805–3810.
7. M. Fullwood, C. Wei, E. Liu and Y. Ruan, "Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses," Genome Res, 2009, 19, pp. 521-532.
8. J. Gilbert, D. Field, Y. Huang, et al, "Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities," PLoS ONE, 2008, 3, pp. e3042.
9. A. Glazer and K. Kechris, "Conserved Amino Acid Sequence Features in the α Subunits of MoFe, VFe, and FeFe Nitrogenases", PLoS ONE, 2009, 4(7), pp. e6136.
10. M. Grabherr, B. Haas, M. Yassour M, et al., "Full-length transcriptome assembly from RNA-seq data without a reference genome," Nat Biotechnol, 2011, 29(7), pp. 644-652.
11. S. Henikoff and J. Henikoff, "Amino Acid Substitution Matrices from Protein Blocks," PNAS, 1992, 89(22), pp 10915–10919.
12. X. Huang, J. Wang, S. Aluru, S. Yang and L. Hillier, "PCAP: AWhole-Genome Assembly Program," Genome Research, 2003, 13, pp. 2164–2170.
13. J. Kent, "BLAT--the BLAST-like alignment tool," Genome Research, 12 (4), pp. 656–664.
14. S. Leininger, T. Urich, M. Schloter, et al., "Archaea predominate among ammonia-oxidizing prokaryotes in soils," Nature, 2006, 442, pp. 806–809.
15. H. Leung, S. Yiu, J. Parkinson, F. Chin, "IDBA-MT: de novo assembler for metatranscriptomic data generated from next-generation sequencing technology," Journal of Computational Biology, 2013, 20(7), pp. 540–550.
16. Z. Khachatryan, Z. Ktsoyan, G. Manukyan, D. Kelly, K. Ghazaryan, and R. Aminov, "Predominant role of host genetics in controlling the composition of gut microbiota," PLoS One, 2008, 3(8), pp. e3064.
17. V. Parro, M. Moreno-Paz and E. Gonzalez-Toril, "Analysis of environmental transcriptomes by DNA microarrays," Env Microbiol, 2007, 9, pp. 453–464.

18. O. Morozova and M. Marra, "Applications of next-generation sequencing technologies in functional genomics," Genomics, 2008, 92, pp. 255-264.
19. J. Mullikin and Z. Ning, "The Phusion Assembler," Genome Research, 2003, 13, . 81–90.
20. Y. Peng, H. Leung, S. Yiu and F. Chin, "Meta-IDBA: a de Novo assembler for metagenomic data," Bioinformatics, 2011, 27(13), . i94-i101.
21. Y. Peng, H. Leung, S. Yiu and F. Chin, "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth," Bioinformatics, 2012, 28(11), pp. 1420-1428.
22. E. Pettersson, J. Lundeberg and A. Ahmadian, "Generations of sequencing technologies," Genomics, 2009, 93, pp.105-111.
23. R. Poretsky, N. Bano, A. Buchan, et al, "Analysis of microbial gene transcripts in environmental samples," Appl Environ Microbiol, 2005, 71, pp. 4121–4126.
24. R. Poretsky, S. Sun, X. Mou, M. Moran, "Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon," Environ Microbiol, 2010, 12, pp. 616–627.
25. J. Qin, R. Li, J. Raes, et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010. 464(7285): p. 59-65.
26. M. Schulz, D. Zerbino, M. Vingron and E. Birney, "Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels", Bioinformatics, 2012, 28(8), pp. 1086-1092.
27. J. Simpson and R. Durbin, "Efficient construction of an assembly string graph using the FM-index," Bioinformatics, 2010, 26(12), i367-i373.
28. J. Simpson, K. Wong, S. Jackman, J. Schein, S. Jones and I. Birol, "Assembly By Short Sequences - a de novo, parallel, paired-end sequence assembler," Genome Res., 2009, 19(6), pp.1117-1123.
29. A. Tartar, M. Wheeler, X. Zhou, M. Coy1, D. Boucias1 and M. Scharf, "Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite Reticulitermes flavipes," Biotechnology for Biofuels, 2009, 2, pp. 25
30. R. Tatusov, E. Koonin and D. Lipman, "A Genomic Perspective on Protein Families," Science, 1997, 278(5338), pp. 631 – 637.
31. T. Urich, A. Lanzen, J. Qi, D. Huson, C. Schleper, S. Schuster, "Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome," PLoS ONE, 2008, 3(6), pp. e2527.
32. X. Xiong, D. Frank, C. Robertson, et al., "Generation and Analysis of a Mouse Intestinal Metatranscriptome through Illumina Based RNA-Sequencing," PLoS ONE, 2012, 7(4), pp. e36009.
33. D. Zerbino, E. Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," Genome Research, 2008, 18(5), pp. 821–829.