DNA reference alignment benchmarks based on tertiary structure of encoded proteins

Hyrum Carroll\textsuperscript{1,*}, Wesley Beckstead\textsuperscript{2}, Timothy O’Connor\textsuperscript{2}, Mark Ebbert\textsuperscript{1,2}, Mark Clement\textsuperscript{1}, Quinn Snell\textsuperscript{1} and David McClellan\textsuperscript{1}

\textsuperscript{1}Computer Science Department and \textsuperscript{2}Biology Department, Brigham Young University, Provo, Utah 84602, USA

Received on April 7, 2007; revised on July 19, 2007; accepted on July 22, 2007
Advance Access publication August 8, 2007

ABSTRACT

\textbf{Motivation:} Multiple sequence alignments (MSAs) are at the heart of bioinformatics analysis. Recently, a number of multiple protein sequence alignment benchmarks (i.e. BAliBASE, OXBench, PREFAB and SMART) have been released to evaluate new and existing MSA applications. These databases have been well received by researchers and help to quantitatively evaluate MSA programs on protein sequences. Unfortunately, analogous DNA benchmarks are not available, making evaluation of MSA programs difficult for DNA sequences.

\textbf{Results:} This work presents the first known multiple DNA sequence alignment benchmarks that are (1) comprised of protein-coding portions of DNA (2) based on biological features such as the tertiary structure of encoded proteins. These reference DNA databases contain a total of 3545 alignments, comprising of 68581 sequences. Two versions of the database are available: \texttt{mdsa\_100s} and \texttt{mdsa\_all}. The \texttt{mdsa\_100s} version contains the alignments of the data sets that TBLASTN found 100\% sequence identity for each sequence. The \texttt{mdsa\_all} version includes all hits with an \textit{E}-value score above the threshold of 0.001. A primary use of these databases is to benchmark the performance of MSA applications on DNA data sets. The first such case study is included in the Supplementary Material.

\textbf{Availability:} The databases, further details and the Supplementary Material are publicly available at \url{http://csl.cs.byu.edu/mdsas/}

\textbf{Contact:} hyrumdc@gmail.com

\textbf{Supplementary information:} Supplementary data are available at Bioinformatics online.

\section{INTRODUCTION}

Multiple sequence alignments (MSAs) provide the foundation for much of the analysis in bioinformatics. They are the first step for everything from annotation of genomes to evolutionary studies. Because of this, it is crucial for automated alignment programs to generate highly accurate and biologically meaningful MSAs to ensure accuracy in subsequent steps in the research process.

Recently, a number of protein sequence databases have been presented to provide a benchmark for alignment algorithms: BAliBASE (Thompson \textit{et al.}, 2005) OXBench (Raghava \textit{et al.}, 2003) PREFAB (Edgar, 2004b) and SMART (Ponting \textit{et al.}, 1999). These databases leverage structural-alignments to provide a suite of ‘gold standard’ alignments. They are assumed to be the ‘true’ alignments, and calculated alignments are evaluated by comparing against them. They have been well accepted by the scientific community and used in numerous studies to compare the quality of protein alignments generated by MSA programs (Do \textit{et al.}, 2005; Edgar, 2004a, b; Karplus and Hu, 2001; Lasmann and Sonnhammer, 2002, 2005; Thompson \textit{et al.}, 1999; Van Walle, 2004). These multiple protein sequence alignment (MPSA) benchmarks are limited to the evaluation of protein alignment applications.

Rarely is a novel alignment technique assessed for its ability to align nucleotide data accurately. The shortage of assessments of MSAs with DNA data may be due to the lack of DNA reference alignments. Applications that work well on amino acid sequences may not be as accurate on DNA data sets. One solution to this problem would be to compare calculated nucleotide alignments against reference nucleotide alignments that are based on the biological features used in protein benchmarks.

Work has been done to address this lack of reference DNA alignments. Pollard \textit{et al.} (2004) created a benchmarking tool for the alignment of non-protein coding DNA using simulated data. While this benchmark gives researchers a starting point to evaluate DNA alignments, the degree to which the simulated sequences reflect those in nature is uncertain.

A ‘gold standard’ benchmark of DNA alignments that is (1) comprised of protein-coding portions of DNA and (2) based on biological features such as the tertiary structure of encoded proteins can help researchers assess the quality of DNA alignment algorithms. This article presents the first known collection of protein-coding DNA benchmark alignments that meet this criteria. A computational tool, MPSA2MDSA, was developed and utilized to convert the following MPSAs into multiple DNA sequence alignment (MDSAs): BAliBASE, OXBench, PREFAB and SMART.

\section{MATERIALS AND METHODS}

Estimating an MDSA from an MPSA is a straight forward procedure that requires three steps. The first step is to find the best analogous DNA sequence (hit) from a protein sequence (query). We queried the
Table 1. Q score, TC score and CPU time ranks

<table>
<thead>
<tr>
<th>Program</th>
<th>DNA data sets</th>
<th>Amino acid data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q score rank</td>
<td>TC score rank</td>
</tr>
<tr>
<td>CLUSTALW</td>
<td>6.35</td>
<td>5.94</td>
</tr>
<tr>
<td>DIALIGN</td>
<td>4.35</td>
<td>4.71</td>
</tr>
<tr>
<td>Kalign</td>
<td>6.77</td>
<td>6.05</td>
</tr>
<tr>
<td>MAFFT-GINSI</td>
<td>9.22</td>
<td>9.59</td>
</tr>
<tr>
<td>MAFFT-LINSI</td>
<td>9.31</td>
<td>9.73</td>
</tr>
<tr>
<td>MAFFT-NSI</td>
<td>7.70</td>
<td>6.96</td>
</tr>
<tr>
<td>MAFFT-NSI</td>
<td>8.68</td>
<td>8.73</td>
</tr>
<tr>
<td>MUSCLE-Default</td>
<td>6.63</td>
<td>7.43</td>
</tr>
<tr>
<td>MUSCLE-Fast</td>
<td>5.22</td>
<td>5.06</td>
</tr>
<tr>
<td>POY</td>
<td>4.03</td>
<td>3.94</td>
</tr>
<tr>
<td>ProbCons</td>
<td>6.24</td>
<td>6.38</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>3.51</td>
<td>3.52</td>
</tr>
</tbody>
</table>

For each category, the ranks according to the Friedman test are given. For the Q and TC scores, the higher the rank indicates higher accuracy. For the times, a lower rank indicates better performance in comparison to other programs. The alignment programs that ranked the best in each column are presented in bold face.

September 2006 version of GenBank’s nt database (Benson et al., 2005) with each of the protein sequences using the TBLASTN algorithm (Altschul et al., 1990). TBLASTN provides the accession number of the best hit. The DNA sequences are then retrieved from the nt database with fastacmd, an NCBI tool. The second step is to account for the occasional gaps introduced by the similarity search. The final step is to apply the alignment from the MPSA to the MDSA. This is done by inserting gaps that correspond with the gaps in the protein alignment. This step is important to preserve the alignment features obtained by higher-order methods (e.g. secondary and tertiary structure or chemical properties) or in other words, to preserve the higher order benchmark alignment. By preserving the biological information, the DNA alignment can be considered a reference alignment. Each step is covered in more detail in the Supplementary Material.

Two versions of each database are publicly available. The first version, mdsa_all100bs, includes only those data sets with all perfect matches (100% sequence identity). This version ensures the highest level of integrity in the conversion. The second version, mdsa_all, includes all hits with an E-value score above the threshold of 0.001. This version retains more of the MPSAs and aids in comparisons with the original MPSAs.

For any heuristic, it is important to quantify the accuracy. Here, the accuracy can be measured by the sequence identity of the hit sequence. In general, as the sequence identity increases, so does the likelihood that the two sequences share the same tertiary structure. For this work, sequences that share 100% sequence identity are assumed to have the same tertiary structure. Sequences with the same tertiary structure will have the same alignment.

Using the nt database, 97.4% of the protein queries found a match with an E-value score above the threshold of 0.001. Furthermore, 69.0% of these hits have 100% sequence identity with the query. While the tool finds a high percentage of exact matches with a current database, databases are growing at an exponential rate, thereby increasing the number of analogous hits of protein queries.

In total, 3545 DNA reference alignments, comprising of 68 581 DNA sequences and 35 600 958 bases are publicly available at http://csl.cs.byu.edu/mdsas/.

To illustrate the usefulness of the reference DNA databases, a case study of the performance and ranks of alignment programs on DNA data sets is included in the Supplementary Material (see also Table 1). Alignments and their respective scores were calculated for seven different multiple sequence alignment applications for each of the 3545 alignments.

7 CONCLUSION

In this work, the first known databases of reference protein-coding DNA alignments are presented. These databases are constructed by leveraging the popular BLAST program to find DNA sequences corresponding to those found in multiple protein sequence alignments. The alignments of the protein sequences (which reflect higher-order information) are applied to the DNA sequences to qualify them to be reference alignments. High-quality hits were obtained from public databases. Over two-thirds of the queries found a perfect match in the nt database. Two versions of the converted databases are available, the first only contains hits that perfectly matched the query, and the comprehensive second version includes all hits above the cut off threshold. These DNA reference alignment databases are publicly available. This benchmark will be extremely useful in evaluating the quality of DNA alignments generated by existing and forthcoming MSA techniques. Finally, the first case study of DNA alignments evaluated by these reference alignments is included in the Supplementary Material.

ACKNOWLEDGEMENTS

We would like to thank Keith Crandall for his review and critiques of this work. This material is based upon work supported by the National Science Foundation under Grant No. 0120718.

Conflict of Interest: none declared.

REFERENCES