How (not) to align genomes

Martin C. Frith\textsuperscript{1} \hspace{1cm} Michiaki Hamada\textsuperscript{2} \hspace{1cm} Paul Horton\textsuperscript{1}

martin@cbrc.jp \hspace{1cm} hamada-michiaki@aist.go.jp \hspace{1cm} horton-p@aist.go.jp

\textsuperscript{1} Computational Biology Research Center, Institute for Advanced Industrial Science and Technology, Tokyo 135-0064, Japan
\textsuperscript{2} Mizuho Information & Research Institute, Inc., Tokyo 101-8443, Japan

Keywords: tandem repeats, scoring scheme, LAST

1 Introduction

Genome alignments form the basis of much research. The accuracy of these alignments depends on various parameters, e.g. simple sequence masking and alignment scoring scheme, that have not been rigorously assessed.

2 Method and Results

2.1 Simple sequence masking

Simple sequences such as \texttt{atatatatatat} cause high-scoring but spurious (non-homologous) alignments. We tested simple sequence masking, by aligning two genomes after reversing (but not complementing) one of them, so that there are no real homologies (Figure 1). Standard masking methods fail to eliminate high-scoring (low-evalue) alignments. There is one method that works remarkably well: TRF \textsuperscript{1} with new parameter settings (mismatch=5, delta=5, minscore=30, maxperiod=200) and hard-masking.

2.2 Alignment scoring schemes

We tested alignment accuracy using 97 combinations of: match score, transition cost, transversion cost, gap existence cost, gap extension cost (e.g. 2:1:2:16:1). We also tested two scoring schemes used by UCSC \textsuperscript{2}: the HoxD70 and HoxD55 matrices, with gap existence cost = 400 and gap extension

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{alignment_plot.png}
\caption{E-values of local alignments between the fugu genome and the reversed stickleback genome, using several repeat-masking methods. The alignments were repeated using 20 different scoring schemes, so each panel has 20 lines.}
\end{figure}
Figure 2: Genome alignment accuracies with 495 combinations of score parameters. Each point represents one genome alignment with one combination of score parameters. True positives and false positives were counted with reference to either TreeFam (top row) or Rfam (bottom row).

cost = 30. Each scoring scheme was combined with 5 X-drop values, for a total of 495 tests. As “gold standards”, we used partial genome alignments implied by multiple alignments of proteins in the TreeFam database and of structural RNAs in the Rfam database. The HoxD55 scheme produces alignments with poor accuracy (many false positives and/or few true positives), the HoxD70 scheme is mediocre, and the simple scheme 1:1:1:7:1 produces rather accurate alignments (Figure 2).

3 Discussion

Widely used genome alignments have inaccuracies, which can be reduced by changing the alignment protocol. The HoxD55 scheme underlying many alignments in UCSC seems especially poor. Alignment accuracy in non-functional genomic regions remains unclear, because gold standards are not available for these regions. This study required thousands of whole genome alignments, which was made possible by our new alignment software, LAST [3].

References

