BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 9 2005, pages 2093–2094
doi:10.1093/bioinformatics/bti224

Sequence analysis

SITEBLAST—rapid and sensitive local alignment of genomic sequences employing motif anchors

Morris Michael*, Christoph Dieterich† and Martin Vingron

Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, D-14195 Berlin, Germany

Received on November 10, 2004; revised on December 6, 2004; accepted on December 13, 2004
Advance Access publication December 14, 2004

ABSTRACT

Motivation: Comparative sequence analysis is the essence of many approaches to genome annotation. Heuristic alignment algorithms utilize similar seed pairs to anchor an alignment. Some applications of local alignment algorithms (e.g. phylogenetic footprinting) would benefit from including prior knowledge (e.g. binding site motifs) in the alignment building process.

Results: We introduce predefined sequence patterns as anchor points into a heuristic local alignment strategy. We extended the BLASTZ program for this purpose. A set of seed patterns is either given as consensus sequences in IUPAC code or position-weight-matrices. Phylogenetic footprinting of promoter regions is one of many potential applications for the SITEBLAST software.

Availability: The source code is freely available to the academic community from http://corg.molgen.mpg.de/software

Contact: christoph.dieterich@molgen.mpg.de

1 INTRODUCTION

We expand on the idea of heuristic alignment algorithms like BLAST (Altschul *et al*., 1997). These tools follow a strategy where a collection of 'matching' n-mers defines anchor points for building local pairwise alignments. Especially in the context of phylogenetic footprinting, it is often desirable to include prior information to guide the alignment building process. Comparative approaches in the domain of promoter analysis will benefit from an approach where alignments are extended from seed pairs whose identity is known a priori. In the context of promoter analysis, prior information is given as consensus patterns or weight matrices, which model a regulatory unit (e.g. a transcription factor binding site). The main idea is to generalize the rigid concept of seeds from identical or similar n-mers to a controlled set of seeds. We implement our concept on top of the existing rapid alignment solutions, namely the BLASTZ software by Schwartz *et al*. (2003).

2 APPROACH

Initially, our SITEBLAST software identifies all potential seeds. The user has two possibilities to specify the patterns that are considered as seeds. First, a list of consensus sequences in IUPAC code can be given together with a distance D. Each occurrence of an n-mer, which matches a consensus sequence with at most D errors is considered as seed. Second, a list of position weight matrices (PWMs) can be used. Here, the user is free to set constraints on the proportion of false positives (P-value) or true positives (power). To avoid unspecific matrix matches, there is an option to set a power-Limit given a P-value cut-off or a P-value limit given a power cut-off. Matrices that do not meet these criteria are excluded from the seed finding process. For example, a matrix with power t and a preset cut-off level T (power-Limit) for some fixed P-value cut-off is selected only if $t \geq T$. Seeds may occur in any orientation (also reversed, complemented or reverse complemented).

After all possible seeds in both sequences are found, alignments are computed by BLASTZ (Schwartz *et al*., 2003). Therefore, each seed in the first sequence for each consensus sequence or PWM is paired with each equivalent occurrence in the second sequence, respectively. These seed pairs are used as anchor points for BLASTZ.

Computed alignments are annotated with all matching seed pairs.

2.1 Finding the seeds

2.1.1 Specified by consensus sequences If seed patterns are given by a list of IUPAC consensus sequences, the user can choose between two different search strategies. The first one needs no preprocessing and scans the two sequences for matches of any consensus sequences in any orientation in a trivial way. This option is faster for short sequences and high distance D. The second strategy generates all matching patterns for all consensus sequences and inserts them in all desired orientations in an Aho–Corasick pattern matching machine (Aho and Corasick, 1975). This matching machine is used to rapidly retrieve all seeds in the two input sequences. This is considerably faster for long sequences and low distance D.

2.1.2 Specified by PWMs Two position specific score matrices (PSSMs) are computed for each PWM (for details see Rahmann *et al*., 2003). One PSSM is tailored to the GC-content of the first sequence and another to the GC-content of the second sequence. We compute the score threshold S for a fixed P-value or power respectively. Then, the attainable power or P-value is computed to test whether they match a given power or P-value limit. For these computations, we employ a modified version of the `PATSER (version 3b)` algorithm by Hertz and Stormo (1999). If the two PSSMs do not meet the power or P-value limits, the pair is discarded. With the remaining PSSMs the two input sequences are scanned. All corresponding

*To whom correspondence should be addressed.

†Present address: Department of Computer Science, PO Box 68 (Gustav Hällströminkatu 2b), FIN-00014 University of Helsinki, Finland.

```
if commandline option I, i or P is defined:
  1. finding seeds in both sequences
  2. combining seeds to seeding pairs
  3. for all seeding pair:try building HSP
  4. process HSPs
  5. print alignments and try to include
     additional matching seed pairs
  6. repeat for each input sequence pair.

otherwise
  1. perform a usual BLASTZ run.
```

Fig. 1. Overview of the anchor point finding procedure. Details to each step are given in the text.

```
378  .    :    .    :    .    :     .    :    .    :    .
GCGAGCAGTTCCCGTCAATCCCTCCCCCCTT-ACACAGGATGTCCATATTAGGACATCTG
GCGAGCTGTTCCCGTCAATCCCTCCCTCCTTTACACAGGATGTCCATATTAGGACATCTG
201  .    :    .    :    .    :    .    :    .    :    .
  ^^^^^^^^^^P53 (sense) dist:2|2
  ^^^^^^^^^^P53 (rev. cmpl) dist:2|2
         ^^^^^^^^CREB (sense) dist:2|2
         ^^^^^^^^CREB (rev. cmpl) dist:2|2
           ^^^^^^^^E2F (rev. cmpl) dist:2|2
           ^^^^^^^^ELK1 (sense) dist:2|2
              ELK1 (sense) dist:2|2^^^^^^^^
               P53 (sense) dist:2|2^^^^^^^^^^
           P53 (rev. cmpl) dist:2|2^^^^^^^^^^
                       SRF (sense) dist:1|1^^^^^^^^^^
                  SRF (rev. cmpl) dist:1|1^^^^^^^^^^
                      CREB (rev. cmpl) dist:2|2^^^^^^^^
                              P53 (sense) dist:2|2^^^^^^^^^
                          P53 (rev. cmpl) dist:2|2^^^^^^^^^
                               CREB (rev. cmpl) dist:2|2^^^
                                 AP1 (rev. cmpl) dist:2|2^^

437  .    :    .    :    .    :    .    :      .    :    .
CGTCAGCAGGTTTCCACGGCCTTTCCCTGTAGCCCTGGGGGGAG--CCATCCCCGAAACC
CGTCAGCAGGTTTCCACGGCCGGTCCCTGTTGTTCTGGGGGGGGGACCATCTCCGAAATC
261  .    :    .    :    .    :    .    :    .    :    .
^P53 (sense) dist:2|2
^P53 (rev. cmpl) dist:2|2
^^^^^CREB (rev. cmpl) dist:2|2
^^^^^^^AP1 (rev. cmpl) dist:2|2
   ^^^^^^^^^^P53 (sense) dist:2|2
   ^^^^^^^^^^P53 (rev. cmpl) dist:2|2
```

Fig. 2. Man–mouse comparison of *c-fos* promoter region as in the CORG database (Dieterich *et al*., 2003). The alignment covers the transcriptional activator region, which was studied by Treisman (1985). A set of consensus sequences (part of the software distribution) was used to scan the input sequences for anchor points. Seed points were allowed to differ by two substitutions from the consensus. Seed pairs need not be identical (e.g. first p53 anchor).

sequence subwords of each sequence are tested for matches to the set of given PSSMs. A match exists if the tested subword attains a score $\geq S$. The score of each subword is computed by adding up the PSSM score for each position of the subword until it is clear that the threshold score for this PSSM cannot be reached. A summary on the seed finding procedure is given in Figure 1.

2.2 Processing the seeds

All subsequences in each input sequence that match a specific motif description as given by a single PWM or a single consensus sequence are paired. Seed pairs need not be identical but have to match the same motif description. The similarity of seed pairs is adjusted by command line parameters (e.g. the maximal number of allowed substitutions of a subsequence to the consensus or the P-value and power cut-off for PWMs). These seed pairs are then used by BLASTZ to compute alignments.

Additional seed pairs are merged with computed alignments if they are consistent with the alignment. A pretty print option exists that decorates all seed pairs to the alignment (Fig. 2). An accompanying website (http://corg.molgen.mpg.de/software/siteblast) gives further details on the program and the corresponding options.

3 CONCLUSION

We present an efficient software solution for computing local alignments of genomic sequences. Prior knowledge on biological signatures like binding site motifs can be readily incorporated either via IUPAC consensus sequences or PWMs.

Figure 2 demonstrates the utility of our software in terms of 'on-the-fly' annotation of promoter regions with transcription factor binding sites.

We are aware of a similar effort (CONREAL) by Berezikov *et al*. (2004). However, the CONREAL software only computes the best global path of matching seed pairs through two input sequences. This is substantially different from our local alignment approach, which employs seed pairs as 'condensation nuclei' of the alignments building process.

REFERENCES

Aho,A.V. and Corasick,M.J. (1975) Efficient string matching: an aid to bibliographic search. *Commun. ACM*, **18**, 333–340.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berezikov,E., Guryev,V., Plasterk,R.H. and Cuppen,E. (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.*, **14**, 170–178.

Dieterich,C., Wang,H., Rateitschak,K., Luz,H. and Vingron,M. (2003) CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res.*, **31**, 55–57.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Rahmann,S., Mueller,T. and Vingron,M. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol.*, **2**, 7.

Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human–mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.

Treisman,R. (1985) Transient accumulation of c-fos RNA following serum stimulation requires a conserved 5′ element and c-fos 3′ sequences. *Cell*, **42**, 889–902.