APPLICATIONS NOTE

fragrep: *Efficient Search for Fragmented Patterns in Genomic Sequences*

Axel Mosig, Katrin Sameith and Peter F. Stadler

Bioinformatics Group, Department of Computer Science, University of Leipzig, Kreuzstrasse 7b, Leipzig, D-04103, Germany and Interdisciplinary Center for Bioinformatics, University of Leipzig, Kreuzstrasse 7b, Leipzig, D-04103, Germany

ABSTRACT

Summary: Many classes of non-coding RNAs (including yRNAs, vaultRNAs, RNAse P and MRP RNA, as well as a novel class recently discovered in *Dictyostelium discoideum*) can be characterized by a pattern of short but well conserved sequence elements that are separated by poorly conserved regions of sometimes highly variable length. Local alignment algorithms such as `blast` are therefore ill-suited for the discovery new homologs of such ncRNAs in genomic sequences. The `fragrep` tool instead implements an efficient algorithm for detecting pattern fragments that occur in a given order. For each pattern fragment a mismatches tolerance and bounds on the length of the intervening sequences can be specified separately.

Availability: The program `fragrep` can be downloaded from http://www.bioinf.uni-leipzig.de/Software/fragrep/.

Contact: Axel Mosig, Tel: ++49 341 14951 31,
Fax: ++49 341 14951 19, axel@bioinf.uni-leipzig.de

Methods for detecting non-coding RNAs (ncRNAs) in genomic sequence data has been a topic of intense research. While techniques for detecting protein-coding genes can rely on universal characteristics such as start and stop codons, the triplet amino acid code or ribosome binding sites, there are no corresponding characteristics known in ncRNAs. Computational tools for ncRNAs detection are therefore restricted to one or few particular classes of RNA. Some classes, such as yRNAs and vaultRNAs, contain stem or loop regions with well-conserved *sequence* patterns (Farris *et al.*, 1999; Teunissen *et al.*, 2000; Kickhoefer *et al.*, 2003)). These characteristics are used by `fragrep`.

Suppose that our ncRNA of interest contains k conserved sequence fragments, denoted by $C_1, \ldots, C_k$, which occur in a given order in a set of known examples. In practice, the C_i are obtained as the consensus sequences of conserved blocks in a multiple alignment. Scanning a genome T for these blocks, we expect to find a non-conserved sequence segment X_i between any two fragments C_i and C_{i+1}. The `fragrep` solves he problem of determining whether there are sequences $X_1, \ldots, X_{k-1}$ such that $C_1X_1C_2X_2 \ldots X_{k-1}C_k$ is a substring of T. Additionally, `fragrep` can take into account two further aspects:

Gap length bounds: For each X_i, the user can specify lower and upper bounds, denoted by ℓ_i and u_i, respectively, for the length of X_i, so that only matches satisfying $\ell_i \leq |X_i| \leq u_i$ will be taken into account by `fragrep`.

Mismatches: The fragments C_i do not need to match the corresponding sequence part of T exactly; the user can specify a number of mismatches m_i. Denoting C_i' as the fragment C_i modified by at most m_i many arbitrary mismatches, `fragrep` will report occurrences of some $C_1'X_1C_2'X_2 \ldots X_{k-1}C_k'$ as well.

We start by computing all occurrences of the *most informative fragment* C_a, i.e., the fragment that is least likely to occur as a random subsequence of the genome T. Then a neighborhood defined by the bounds ℓ_i on the lengths of the intervening sequences X_i is searched for the other fragments C_i, $i \neq a$. From this position information a graph G is constructed such that paths of length k in G correspond to occurrences of $C_1, \ldots, C_k$ in the given order under the specified mismatch and gap length constraints. These paths can be found easily by means of dynamic programming. Starting with the most informative sequence C_a rather than C_1 increases the efficiency of the search and in practice leads to a significant speedup, in particular when short or ambiguous fragments are part of the pattern. The `C` implementation of `fragrep` has been optimized in several algorithmic details to improve the runtime.

We used `fragrep` to studying the evolution of a class of ncRNAs in the the slime mold *Dictyostelium discoideum* that was discovered in an experimental survey by Aspegren *et al.* (2004). We searched the genomic sequence (Fey *et al.*, 2004) for *type-I ncRNAs* using the following simple pattern:

```
0   0  GTTGRCCTTACAGCAA   2
0 120  GTCAACTG           2
```

The first two columns contain the minimal and maximal

distance before a the pattern fragment (always 0 for the first fragment, of course), the last column is the maximal number of mismatches that is tolerated in each fragment. We recovered 45 candidates of which 34 are sufficiently similar to the experimentally determined sequences to be alignable. 11 very divergent sequences were not included in the further analysis. A neighborjoining tree summarizing both known sequences and the novel candidates detected by `fragrep` is displayed as Fig. 1. We find the the class-I ncRNAs are located in small clusters in all 6 chromosomes. Interestingly, there are two subclasses, denoted by A and B, that alternate in the larger clusters, even though their direction on the chromosomes does not seem to follow a simple rule.

In order to evaluate the performance of the algorithm underlying `fragrep`, we used a query derived from vaultRNA A-, B1-, and B2 box consensus structures in Kickhoefer *et al.* (2003) to scan the whole human genome. The query consisted of three fragments, each of which was 11 nucleotides long. Scanning all chromosomes of the human genome took less than 8 minutes on a standard desktop computer with a 2.4GHz processor and 1GB main memory; further results from scanning the human as well as the mouse, dog and rat genome are given in the following table.

Genome	**H.sap.**	**M.musc.**	**R.nov.**	**Dog**
size in Mb	2,980	2,561	2,640	2,454
runtime (mm:ss)	7:18	8:04	6:25	7:01
# cand. matches	21	53	31	1

These examples demonstrate that `fragrep` can be used for systematic surveys of eukaryotic genomes. The application of standard multiple alignment tools such as `ClustalW` or `dialign` to a relatively small set of representatives of an ncRNA class can be used to determine conserved sequence patterns, which can be turned into `fragrep` queries in a straightforward manner. The `fragrep` tool can then be employed to find additional members of the ncRNA family in related genomes. This approach yields significant matches where other sequence search tools such as `blast` fail to report useful results, while structure based approaches, such as `infernal` are too costly. Of course, `fragrep` is not limited to ncRNA detection; the search for specific constellations of transcription factor binding sites is another potential application.

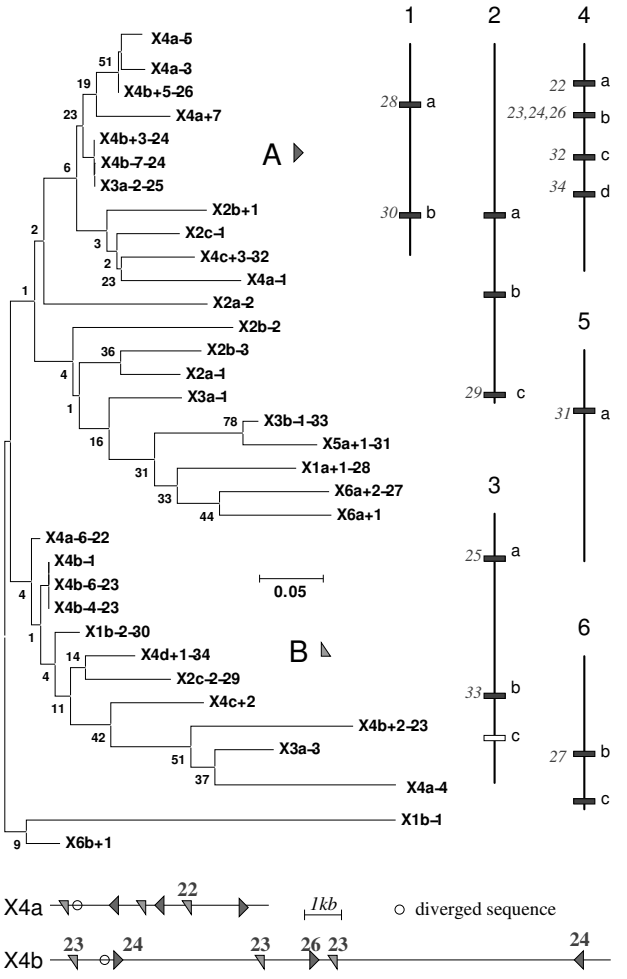

Fig. 1. Type-I ncRNAs from *Dictyostelium discoideum*. Red numbers are the DdR- numbers of the expressed RNAs from the experimental survey by Aspegren *et al.* (2004). The sequences appear in clusters on all chromosomes (right). The phylogenetic tree (left, neighborjoining method) suggests that there are two major subgroups, labeled A and B. Below the organization of the two largest clusters X4a and X4b located at chromosome 4. Note that type A and type B copies alternate. The other type-I ncRNA clusters consist of not more than three sequences.

REFERENCES

Aspegren, A., Hinas, A., Larsson, P., Larsson, A. & Söderbom, F. (2004). Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development. *Nucl. Acids Res.*, **32**, 4646–4656.

Farris, A. D., Koelsch, G., Pruijn, G. J., van Venrooij, W. J. & B., H. J. (1999). Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis. *Nucl. Ac. Res.*, **27**, 1070–8.

Fey, P., Gaudet, P., Just, E. M., Merchant, S. N., Pilcher, K. E., Kibbe, W. A. & Chisholm, R. L. (2004). dictyBase. `http://www.dictybase.org/`.

Kickhoefer, V. A., Emre, N., Stephen, A. G., Poderycki, M. J. & H., R. L. (2003). Identification of conserved vault RNA expression elements and a non-expressed mouse vault RNA gene. *Gene*, **309**, 65–70.

Teunissen, S. W., Kruithof, M. J., Farris, A. D., Harley, J. B., Venrooij, W. J. & Pruijn, G. J. (2000). Conserved features of Y RNAs: a comparison of experimentally derived secondary structures. *Nucl. Ac. Res*, **28**, 610–9.