DNAfan: a software tool for automated extraction and analysis of user-defined sequence regions

Andreas Gisel1, Maria Panetta1, Giorgio Grillo1, Vito Flavio Licciulli1, Sabino Lioni1, Cecilia Saccone1,2 and Graziano Pesole3,*

1Sezione di Bioinformatica e Genomica di Bari, Istituto di Tecnologie Biomediche, CNR, via Amendola 122/D, 70126 Bari, Italy, 2Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Via Orabona 4, 70125 Bari, Italy and 3Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Via Celoria 26, 20133 Milano, Italy

Received on February 25, 2004; revised on June 16, 2004; accepted on July 8, 2004
Advance Access publication July 15, 2004

ABSTRACT
Summary: DNAfan (DNA Feature ANalyzer) is a tool combining sequence-filtering and pattern searching. DNAfan automatically extracts user-defined sets of sequence fragments from large sequence sets. Fragments are defined by annotated gene feature keys and co- or non-occurring patterns within the feature or close to it. A gene feature parser and a pattern-based filter tool localizes and extracts the specific subset of sequences. The selected sequence data can subsequently be retrieved for analyses or further processed with DNAfan to find the occurrence of specific patterns or structural motifs. DNAfan is a powerful tool for pattern analysis. Its filter features restricts the pattern search to a well-defined set of sequences, allowing drastic reduction in false positive hits.

Availability: http://bighost.ba.itb.cnr.it:8080/Framework
Contact: graziano.pesole@unimi.it

The increasing availability of annotated sequence data requires the concurrent development of adequate bioinformatic tools for mining and deciphering information contained therein. Only the correct interpretation of such data can produce new insights into sequence–function relationships and into transcriptional and post-transcriptional regulatory mechanisms. A major challenge of the so-called ‘post-genomic’ era is to determine the biological functions of non-coding sequences—which are known to play crucial roles in modulating the chronology and cytological location of gene expression.

The expression regulation of a given gene is determined by the specific assortment of transcriptional and post-transcriptional regulatory motifs in the appropriate context. Transcription of a gene is regulated using short DNA fragments (6–20 bp) located close to the gene and targeted by specific transcription factors (TFs). Owing to the high level of degeneracy of individual transcription factor binding sites (TFBSs) (Moses et al., 2003), genome-wide searches for them are often ineffective due to the large number of false positive hits. Indeed, functional promoters consist of multiple TFBSs in specific arrangements and suitable genomic contexts (e.g. the distance and upstream or downstream position with respect to the transcription start site, etc.).

After transcription, the fate of mRNAs can be controlled by a variety of sequence motifs generally located in the 5′-untranslated region (5′-UTR) or the (3′-UTR) (Mignone et al., 2002). The functional activity of these RNA motifs depends both on their sequence and secondary structure. They modulate mRNA stability, subcellular localization and translation efficiency via interactions with specific RNA-binding proteins. Software tools facilitating searches for any combination of sequence patterns, including ‘positional weight matrices’ (PWMs) and structural sequence motifs (e.g. stem–loop structures) in specific genome locations are thus desirable. Such tools could also be useful to identify genome regions containing (or not containing) such defined sequence motifs.

The software presented here, DNAfan, possesses the above-mentioned properties and allows the extraction of sequence fragment subsets from large sequence data collections defined by the surrounding annotated sequence feature keys (i.e. 5′-UTR, exon, intron, mRNA, CDS, etc. and/or by close co-occurring, known patterns).

DNAfan consists of three major functions: ‘Sequence selection’, ‘Sequence feature key recognition’ and ‘Pattern recognition’, designed to extract a user-defined, specific sequence subset. ‘Sequence selection’ allows the user to choose a set of sequences as the input dataset for DNAfan. ‘Sequence feature key recognition’ allows the user to find

*To whom correspondence should be addressed.
DNAfan: a feature/pattern-based analysis tool

**Fig. 1.** Program structure: DNAfan is structured in two work stages 'Sequence Extraction' and 'Sequence Analysis'. Sequence Extraction filters sequence from an initial set of sequences according to the user-defined gene feature key and a possible co-/non-occurrence of a pattern. Sequence Analysis executes first Sequence Extraction and runs on the final subset of sequence fragment a pattern analysis.

a feature key in the gene annotation, while 'Pattern recognition' searches for a pattern motif in the input sequences. The three functionalities can be combined into two different tasks, 'Sequence Extraction' and 'Sequence Extraction and Analysis' (Fig. 1).

'Sequence Extraction' returns the list of sequence fragments with the user-defined extraction conditions. 'Sequence Analysis' uses the list of sequence fragments as input data for a user-defined, extended pattern analysis and returns a list of pattern sequences and their position within the initial input sequence.

The workflow of DNAfan (Fig. 1) is designed so that the user first chooses the database from which DNAfan retrieves the initial sequences. DNAfan offers two datasets, EMBL database and its divisions (Kulikova et al., 2004) and Ensembl (Birney et al., 2004) with all currently annotated species. In case of Ensembl as the sequence repository, the user is directed to enter or upload a list of gene identifiers that DNAfan will analyse. All gene identifiers maintained by Ensembl, including Affymetrix probe and GO IDs, are accepted by DNAfan. This 'Sequence selection' drastically restricts the search data and improves the result quality. Instead of uploading gene identifiers, DNAfan can also extract sequences from a genomic region specified by chromosome number and start and end coordinates.

The next step is independent of the database chosen and consists of the choice of filter parameters according to which DNAfan extracts the set of sequence fragments. The parameters include the type of molecule (DNA or RNA), all major feature keys (e.g. mRNA, intron, CDS, etc.), the position within the feature and the possible occurrence of a nearby pattern.

The last step, in the case of a 'Sequence Extraction' is the definition by the user of the output format of the extracted sequence fragments. Alternatively, in the case of a 'Sequence Extraction and Analysis' task, a pattern recognition analysis is carried out on the set of extracted fragments using the Pat-Search software (Grillo et al., 2003). Figure 2 shows a typical DNAfan output obtained in the case of a 'Sequence Extraction and Analysis' task.

DNAfan allows the user to define, extract and analyse a highly specific set of sequences with similar functional properties—thus facilitating the rapid focusing of intensive sequence analyses. To make DNAfan user-friendly, pop-up help instructions direct the user in all the required steps (help windows open clicking when a '?' appears in correspondence of required fields).

Another on-line tool for the extraction of sequence fragments according to gene feature keys is EnsMart (Kasprzyk et al., 2004). EnsMart is a generic data warehousing for fast and flexible data querying but, unlike DNAfan, it does not offer the use of either extensive parameters to describe the extracted sequence fragments or pattern recognition features to localize co-occurring patterns.

The power of DNAfan can best be demonstrated with an example. It has been shown that the downstream promoter element (DPE) occurs with the same frequency in TATA-less promoters as the TATA motive occurs in DPE-less promoters in Drosophila (Kutach and Kadonaga, 2000), whereas TATA-less DPE containing promoters are rare in humans (Vijh et al., 2002). Two such configured promoters drive the expression of the genes CCR3 and IRF1, a chemokine receptor and a transcription factor, respectively. DNAfan was asked to find all genes from a set of 10952 human genes—defined by an approved HUGO gene symbol which contain a DPE but not a TATA-box. The search was specified in such a way that the DPE should be found in the first 50 nt downstream of the transcription start site (TSS) since in CCR3 and IRF1 the DPE is located approximately at position +30 from the TSS.
Fig. 2. Sample output of a ‘Sequence Extraction and Analysis’ task. Two RefSeq accession numbers were used as input for the sequence retrieval from Ensembl. The filter used ‘Intron’ as filter key and the sequence fragments were processed by PatSearch (Grillo et al., 2003) to match the sequence pattern described by COMMAND. The output shown in order from left to right: sequence ID, input gene ID, number of feature (all separated by underscore), position of the pattern relative to the feature sequence (square brackets), genomic position of the feature and gene ID (round parentheses) and after the colon the pattern sequence structured according to the pattern syntax.

To exclude all sequences that do not have a TATA-box until −60 from TSS (Audic and Claverie, 1998), we used the pattern option with the PWM for the TA TA consensus with a score cut-off of −8.16 (Bucher, 1990). In accordance with the published DPE consensus of rat CCR3, human CCR3 and human IRF1 (Vijh et al., 2002), we defined the DPE consensus as TGASRC allowing the motives T-G-A-(C/G)-(A/G)-C. DNA-fan matched 468 DPE motives corresponding to 426 genes containing a DPE in the region from +1 to +50 with respect to the TTS and missing the TA TA-box in the 60 bp upstream of TTS. Assuming an exact position of the DPE motive and tolerating a variation at 1 nt the search recovered 30 genes that could belong to the TA TA-less DPE promoter class. We could also confirm the DPE/TA TA-less constellation in the human CCR and IRF1 genes. Such a search reveals much more information but further analysis is beyond the scope of this publication.

ACKNOWLEDGEMENTS
We thank David Horner for valuable comments on the manuscript. This work was supported by FIRB project ‘Bioinformatica per la Genomica e la Proteomic’ (Ministero dell’Istruzione e Ricerca Scientifica, Italy) and by Telethon.

REFERENCES


