BMC Bioinformatics

Methodology article

Kangaroo – A pattern-matching program for biological sequences

Doron Betel[1,2] and Christopher WV Hogue*[1,2]

Address: [1]Department of Biochemistry, University of Toronto, Toronto, Ontario, M5S 1A8, Canada and [2]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Ave., Toronto, Ontario, M5G 1X5, Canada

E-mail: Doron Betel - betel@mshri.on.ca; Christopher WV Hogue* - hogue@mshri.on.ca

*Corresponding author

Published: 31 July 2002

BMC Bioinformatics 2002, **3**:20

Received: 5 July 2002

Accepted: 31 July 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/20

© 2002 Betel and Hogue; licensee BioMed Central Ltd. This article is published in Open Access: verbatim copying and redistribution of this article are permitted in all media for any non-commercial purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: Biologists are often interested in performing a simple database search to identify proteins or genes that contain a well-defined sequence pattern. Many databases do not provide straightforward or readily available query tools to perform simple searches, such as identifying transcription binding sites, protein motifs, or repetitive DNA sequences. However, in many cases simple pattern-matching searches can reveal a wealth of information. We present in this paper a regular expression pattern-matching tool that was used to identify short repetitive DNA sequences in human coding regions for the purpose of identifying potential mutation sites in mismatch repair deficient cells.

Results: Kangaroo is a web-based regular expression pattern-matching program that can search for patterns in DNA, protein, or coding region sequences in ten different organisms. The program is implemented to facilitate a wide range of queries with no restriction on the length or complexity of the query expression. The program is accessible on the web at [http://bioinfo.mshri.on.ca/kangaroo/] and the source code is freely distributed at [http://sourceforge.net/projects/slritools/].

Conclusion: A low-level simple pattern-matching application can prove to be a useful tool in many research settings. For example, Kangaroo was used to identify potential genetic targets in a human colorectal cancer variant that is characterized by a high frequency of mutations in coding regions containing mononucleotide repeats.

Background

Significant progress has been made in search and homology detection algorithms for DNA and protein sequences. Many of these algorithms are geared toward heuristic searches where the program assumes that the end user is interested in sequences that may be distantly related to the query sequences. As a result, low complexity query patterns, such as repetitive DNA sequences, are often filtered out using various filtering masks, such as SEG [1], since they represent sequences with low information content and therefore, add additional "noise" to the search results. Some applications provide specific functional annotation through pattern-matching such as, domain mapping, intron/exon boundaries, or finding statistically significant patterns in sequences [2–5]. Other matching programs are specific to one organism or search through a specific subset of sequence data. In spite of these advanced search techniques there is still a need for a simple, unassuming low-level pattern-matching program when looking for very specific motifs in DNA or protein records. Such motifs may be novel protein binding signatures, repetitive sequences, transcription factor binding sites, protein

domains and functional genomic sequences. Researchers sometimes misuse powerful homology searching programs, such as BLAST [6], to perform low-level pattern-matching where a simple binary (yes or no) search will suffice.

In this work we present a new web-based pattern-matching program that identifies protein or DNA records containing patterns of interest in a number of model organisms. A novel feature of this program allows matching patterns in coding region sequences. The program reports back all GenBank records that match the regular expression along with the sequence coordinates without any filtering or post processing of the results.

Results

Kangaroo is an *ad hoc* program that performs basic regular expression searches through DNA, protein and manually annotated coding regions for a user-entered query expression. The program retrieves annotated GenBank records from our in-house SeqHound database (K. Michalickova et al., manuscript in preparation) and performs a regular expression search on those records. In cases where the user has specified a coding region search, the program first extracts the coding region information from GenBank ASN.1 structures and then carries out a regular expression search on those sequences exclusively.

The web interface contains a text window where the user can enter a sequence of amino acids or nucleotides (Figure 1). Using the simple rules and metacharacters of regular expressions, the user can compile a complex pattern that might represent a functional sequence in protein or DNA. For example, the regular expression "[ST]X [VIL]$" represents potential PDZ binding sites at the C-terminus of a protein. This example illustrates the use of special metacharacter symbols that extend the query expression beyond a single pattern to an expression that can represents a wide range of queries. In the above example, the "$" symbol is used to specify that all matches must be at the C-terminus, " [ST]" allows for either Ser or Thr at first position, and "X" represents any amino acid. Biologists use letters other than A, C, T, and G, to designate more than one possible DNA base at a given position within a sequence (e.g. K means G or T). Kangaroo supports the use of these IUPAC ambiguous symbols in the pattern searches. For protein searches, however, there are only a few symbols that code for more than one amino acid. In those cases, it is easy to specify a choice of amino acids at a given position by using regular expression rules. Another useful feature of regular expressions is the ability to specify variable length patterns within an expression. For example, the pattern "GGT{5,8} AC" specifies all sequences where "GG" and "AC" are separated by a linker of minimum 5 to a maximum of 8 "T".

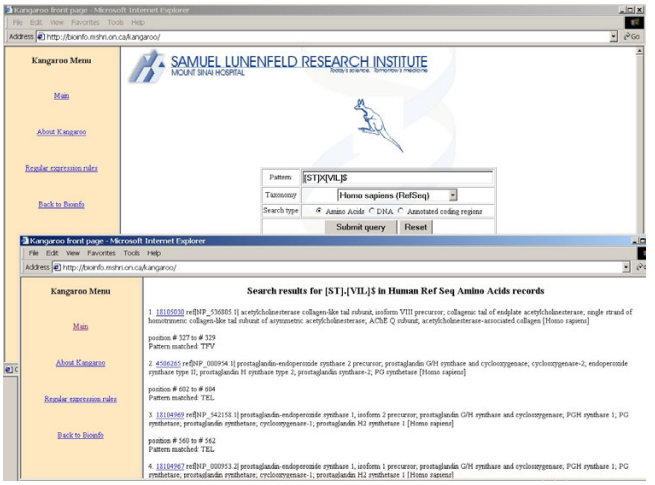

Figure 1

Kangaroo screenshot. A screenshot of Kangaroo query web interface and results page for a search for potential PDZ binding sites. Top panel shows the Kangaroo user interface. The user can select to match the expression to protein, DNA or coding region sequences from 10 different organisms. The bottom panel shows sample search results. Each hit contains a hyperlink to the full GenBank flatfile record. Note the use of regular expression rules and wildcard character "X" to specify the wide range of potential PDZ binding motifs.

A pull-down menu offers users a choice of 10 model organisms to search for a pattern of interest. The search results display the GenBank records for the selected organism that matched the query pattern. A FASTA definition line appears for every matched record along with the location of the matches in the sequence and the exact pattern that was matched to that location (Figure 1). The search algorithm reports all patterns that were matched in a single record with the exception of overlapping hits. For example, the pattern "AAA" will match the sequence "CAAAAAG" only once even though this pattern appears three times in that sequence. This restriction is meant to avoid reporting multiple matches in a region that contains long repeat sequences. On the other hand, for some applications it may be preferable to identify overlapping patterns within a region. In future versions of the program users will be able to select between these two modes of pattern-matching. If the user is interested in additional information about the hits, hyperlinks connect the matched records to the full flatfile record. Due to the size of the database the maximal number of reported matches is restricted to 10,000 hits. For sequence retrieval, Kangaroo relies on our in-house SeqHound integrated database that is similar to the NCBI Entrez system and the NCBI MMDB structure database. To speed up searches a pre-computed table contains lists of sequence identifiers of large taxonomies for fast retrieval of sequences, a second pre-comput-

ed table contains the human coding region sequences. All other coding region sequences are computed per search request. To ensure that the data is current, both pre-computed tables and all other sequence sources are integrated into SeqHound and updated on a regular basis.

Discussion

Kangaroo was initially implemented for the purpose of searching short repeat patterns in human coding regions. A number of genes containing mononucleotide repeats were implicated in a distinct type of colorectal cancer (CRC), which is characterized by increased rates of mutations in those repeat units [7]. Using Kangaroo, we searched human coding sequences for genes that contain any of the four possible mononucleotide repeats ranging from 6 to 13 bases in length in an effort to identify more genes that might be involved in this CRC pathway. A number of genes identified in this search were shown to have increased mutation rates in mononucleotide repeats in tissue samples taken from patients with this type of CRC [8]. The search results also reveal that the human genome contains more mononucleotide repeats than was originally predicted, among them, adenine repeats are most common. The abundance of adenine repeats in human coding regions might be attributed to the high lysine content (coded by AAA and AAG codons) in nuclear localization signals [9].

It stands to reason that natural selection processes will disfavour repetitive DNA segments due to their increased rates of mismatches during DNA replication. Specifically, we expect that evolution will select against codon arrangements that contain mononucleotide repeats. We postulate that the observed frequencies of such codon combinations would to be much lower than would be expected by their overall frequency in the genome. To confirm this hypothesis we are using Kangaroo to search for occurrences of three tandem codons that code for the same amino acid and that produce a stretch of 6 to 9 mononucleotide repeats (manuscript in preparation). Kangaroo has been used in other research settings, such as identification of novel domains and searches for potential phosphorylation sites.

Conclusions

Kangaroo has proven to be a useful low-level pattern-matching program. The simplistic user interface and the absence of any scoring function make it an easy-to-use database mining tool. This program can be used to search for short, low complexity DNA sequences. By using a relatively small set of symbols and simple regular expression rules, the user can perform a powerful search for a wide variety of protein and DNA fingerprint sequences, such as novel domain regions, binding motifs and other elements of interest.

Materials and Methods

Kangaroo was written entirely in the C programming language using the NCBI toolkit (Ostell, J. 1997) and developed on a dual Pentium II processor Linux machine. The web-based application runs on a four processor Sun Solaris server. Kangaroo is accessible at [http://bioinfo.mshri.on.ca/kangaroo] and the source code is available at [http://sourceforge.net/projects/slritools]. All gene and protein records are retrieved from our in-house SeqHound database, which mirrors NCBI's latest GenBank release, the NCBI taxonomy database and MMDB. All human records are retrieved from the GenBank primate division, which excludes all high throughput sequencing data such as, EST and STS. The search algorithm is based on regular expression functions that are part of the NCBI toolkit and POSIX UNIX. Annotated coding region information is parsed from GenBank ASN.1 files.

Authors' contributions

Doron Betel developed Kangaroo and performed the database searches for coding regions for mononucleotide repeats. Chris Hogue conceived the program and participated in its design.

Acknowledgements

The authors wish to thank Jane Park for her role in the identification of potential genes involved in the CRC study and Dr. Mark Redston for his fruitful collaboration. This project was supported by the National Cancer Institute of Canada.

References

1. Wootton JC, Federhen S: **Statistics of local complexity in amino acid sequences and sequence database.** *Computational Chemistry* 1993, **17:**149-163
2. Appel RD, Bairoch A, Hochstrasser DF: **A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server.** *Trends Biochem Sci* 1994, **19:**258-260
3. Pesole G, Liuni S, D'Souza M: **PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance.** *Bioinformatics* 2000, **16:**439-450
4. Pesole G, Prunella N, Liuni S, Attimonelli M, Saccone C: **WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences.** *Nucleic Acids Research* 1992, **20:**2871-2875
5. Dsouza M, Larsen N, Overbeek R: **Searching for patterns in genomic data.** *Trends Genet* 1997, **13:**497-498
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410
7. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, *et al*: **A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer.** *Cancer Res* 1998, **58:**5248-5257
8. Park J, Betel D, Gryfe R, Michalickova K, Di Nicola N, Gallinger S, Hogue CW, Redston M: **Mutation profiling of mismatch repair-deficient colorectal cancers using an in silico genome scan to identify coding microsatellites.** *Cancer Res* 2002, **62:**1284-1288
9. Cokol M, Nair R, Rost B: **Finding nuclear localization signals.** *EMBO Rep* 2000, **1:**411-415