W262–W266 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki368

PatMatch: a program for finding patterns in peptide and nucleotide sequences

Thomas Yan[1,2], Danny Yoo[1], Tanya Z. Berardini[1], Lukas A. Mueller[1], Dan C. Weems[3], Shuai Weng[4], J. Michael Cherry[4] and Seung Y. Rhee[1,*]

[1]Department of Plant Biology, Carnegie Institution of Washington, 260 Panama Street, Stanford, CA 94305, USA, [2]Department of Computer Engineering, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053, USA, [3]National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA and [4]Department of Genetics, Stanford University, Stanford, CA 94305, USA

Received February 11, 2005; Revised and Accepted February 28, 2005

ABSTRACT

Here, we present PatMatch, an efficient, web-based pattern-matching program that enables searches for short nucleotide or peptide sequences such as *cis*-elements in nucleotide sequences or small domains and motifs in protein sequences. The program can be used to find matches to a user-specified sequence pattern that can be described using ambiguous sequence codes and a powerful and flexible pattern syntax based on regular expressions. A recent upgrade has improved performance and now supports both mismatches and wildcards in a single pattern. This enhancement has been achieved by replacing the previous searching algorithm, scan_for_matches [D'Souza *et al.* (1997), *Trends in Genetics*, 13, 497–498], with nondeterministic-reverse grep (NR-grep), a general pattern matching tool that allows for approximate string matching [Navarro (2001), *Software Practice and Experience*, 31, 1265–1312]. We have tailored NR-grep to be used for DNA and protein searches with PatMatch. The stand-alone version of the software can be adapted for use with any sequence dataset and is available for download at The *Arabidopsis* Information Resource (TAIR) at ftp://ftp.arabidopsis.org/home/tair/Software/Patmatch/. The PatMatch server is available on the web at http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl for searching *Arabidopsis thaliana* sequences.

INTRODUCTION

PatMatch is designed to find short (3–30 nt or amino acid) sequence matches. It can be useful for finding short patterns in nucleotide sequences such as *cis*-elements, massively parallel signature sequence (MPSS), instances of known serial analysis of gene expression (SAGE) tags, small RNA binding sites or small protein domains and motifs in protein sequences. PatMatch uses a short sequence or regular expression as input and allows ambiguous sequences to be represented by standard International Union of Pure and Applied Chemistry (IUPAC; http://www.chem.qmw.ac.uk/iupac) nomenclature. The program also allows inexact matching (mismatches) of the query sequence against literal or regular expression patterns. PatMatch requires users to explicitly enter a pattern to search for and is not meant for *de novo* pattern detection.

The original version of PatMatch was provided by the *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) (1) and was modified to be deployed at The *Arabidopsis* information resource (2,3). In this paper, we report on changes we made to the software to improve performance and support for mismatches when using wildcards in the query sequence by using Nondeterministic-Reverse grep (NR-grep) (4). In addition, the Common Gateway Interface (CGI) code has been restructured, and the auxiliary programs that displayed the results, which were written in C, were rewritten in Perl and modularized to facilitate maintenance and future extension. This new version of PatMatch is available at TAIR and is also available from SGD at http://db.yeastgenome.org/cgi-bin/PATMATCH/nph-patmatch.

*To whom correspondence should be addressed. Tel: +1 650 325 1521 ext 251; Fax: +1 650 325 6857; Email: rhee@acoma.stanford.edu
Present address:
Lukas A. Mueller, Cornell University, Emerson Hall Room 251, Ithaca, NY 14853, USA

USAGE

The web interface of PatMatch (http://arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl) allows users to search for short nucleotide or peptide sequences and patterns against *Arabidopsis thaliana* sequence datasets available at TAIR using any standard web browser and operating system. The web interface has been tested with Netscape 6.X, IE5.X(Mac)/IE6.X(Windows), Safari 1.0 and Firefox/Mozilla browsers. The datasets used at TAIR can be downloaded from ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets.

Query configuration

PatMatch supports queries for either exact or approximate sequence matches by employing a regular expression syntax that includes both ambiguous characters and pattern syntax (Table 1). Ambiguous characters that can be used in PatMatch include wildcard characters, which will match any amino acid or any nucleotide base (*), and more specialized characters that will match only hydrophilic residues (J) or hydrophobic residues (O) for peptide searches or only purine or pyrimidine bases for nucleotide searches. PatMatch supports all specialized wildcard characters represented by the standard IUPAC nomenclature. In addition, strings can be explicitly excluded in the input pattern.

In addition to choosing the regular expression syntax for their query, the web interface of PatMatch (Figure 1A) allows users to further customize their searches in several ways. These include specifying which dataset to search against and further adjusting the stringency of the search by specifying the number and types of mismatches allowed. Types of mismatches include insertions, deletions and substitutions. For a nucleotide search, the user can specify searching of both or only one of the DNA strands. The user can also modify the output by specifying the maximum number of hits to return (250 000 maximum), and the minimum and maximum number of hits allowed in each sequence within the chosen dataset.

Output

PatMatch returns the results of the user's query to a web page (Figure 1B) that contains the query parameters as well as a table of results. The results table can also be downloaded as a tab-delimited text file. The HTML page provides the results in a table; each match to a sequence is shown as a separate row that includes the name of the sequence, the number of matches of the query sequence to the matching sequence, the sequence of the pattern that matched the query and the coordinates of the match within the matching sequence. The last column is a hyperlink to a display that shows the sequence with the position of the hit highlighted in red (Figure 1C). If the match is to the complementary strand of a DNA sequence, the hit pattern is the reverse complement of the query sequence. For hits within transcript sequences, the coding sequence is shown in uppercase and the UTRs are shown in lowercase fonts. More detailed information about a sequence is provided by the hyperlink in the sequence name column of the results table. For instance, if the sequence name is a TAIR locus identifier, it is hyperlinked to the corresponding TAIR locus detail page, which shows the functional annotations and gene features for that locus along with other details curated by the TAIR staff.

Stand-alone version

The PatMatch software used by the PatMatch web server is also available for download and local installation at ftp://ftp.arabidopsis.org/home/tair/Software/Patmatch/. Users can download and install this version and configure it for use with their own datasets or for processing a large amount of data locally. The software available on the FTP site also includes a Perl script that is needed to unjustify FASTA files that are to be used by PatMatch. This simple script takes a FASTA file, with a single or multiple sequences, as input and outputs a file with each individual sequence on a single line. This is necessary because NR-grep looks for matches line by line in a file. Documentation for installing and running PatMatch is also included in the FTP site (ftp://ftp.arabidopsis.org/home/tair/Software/Patmatch/README.txt). At this time, only the command line version can be downloaded from the FTP site. The CGI version of the software is not currently available for download. We anticipate that the software will function on any Unix system and has been extensively tested on both Solaris 8 and Linux 2.4.21.

METHODS

Tailoring NR-grep for nucleotide and peptide searches

NR-grep is a general tool for approximate string matching, thus a Perl wrapper called scan_pipeline was written around NR-grep in order to tailor the program towards nucleotide and peptide pattern matching. The wrapper script checks the input pattern for errors and translates the degenerate patterns represented by the standard IUPAC nomenclature into the set of

Table 1. Pattern syntax supported by PatMatch

Pattern	Meaning	Example	Example explanation
[ ]	A subset of elements	AT[TC]ATA	AT, followed by T or C, followed by ATA
[^]	An excluded subset of elements	GC[^TA]G	GC, followed by C or G, followed by G
( )	A subpattern	IF(YPT)SV	IF, followed by YPT, followed by SV
{*m*,*n*}	{*m*} = exactly *m* times {*m*,} = at least *m* times {,*m*} = 0 to *m* times {*m*,*n*} = between *m* and *n* times	L{3,5}W{5}DG	3 to 5 L's, followed by 5 W's, followed by DG
<	Constrains pattern to N-terminus (peptide) or 5′ end (DNA)	<MNTD	Matches MNTD, but only if it occurs at the N-terminus of the peptide sequence
>	Constrains pattern to C-terminus (peptide) or 3′ end (DNA)	TGA>	Matches TGA, but only if it occurs at the 3′ end of the nucleotide sequence

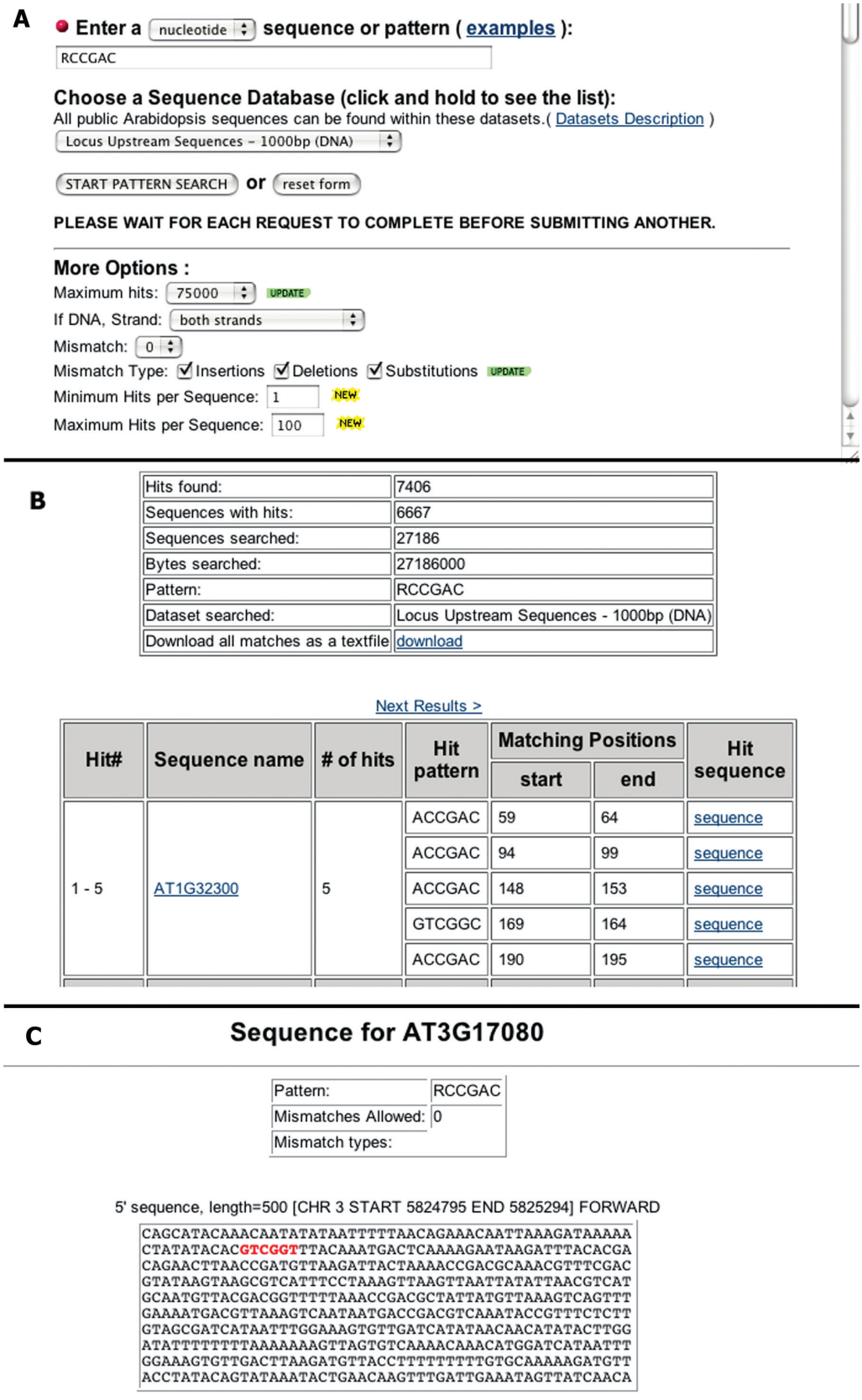

Figure 1. (**A**) The PatMatch input web interface. This screen capture shows how PatMatch is used to find the DREB binding site (12), RCCGAC, where R stands for any purine base. One of the locus upstream sequence datasets is used to find sequences containing this *cis*-element. (**B**) The PatMatch results page. This screen capture shows the output of the query of the pattern, RCCGAC, after searching the 1000 bp locus upstream dataset on both strands. (**C**) A page showing a single match (highlighted in red) of the query in a sequence. The pattern, mismatch options of the search and information about the sequence from its FASTA header are shown.

nucleotides or amino acids they represent. The input pattern in PatMatch syntax is also converted into a different regular expression syntax that is used by NR-grep. We felt that the regular expression syntax used by NR-grep was too cumbersome for certain types of patterns. For example, the pattern to search for three to five occurrences of the MWA subsequence in a peptide sequence is (MWA){3,5} in PatMatch syntax and [(MWA)(MWA)(MWA)(MWA)?(MWA)?] in NR-grep syntax. In addition to checking and converting the user's input, the scan_pipeline script also enables searching for patterns on the reverse strand of datasets containing nucleotide sequences. This script also prunes the output of NR-grep to associate each match with a sequence name and the coordinates of the match relative to the sequence rather than the location in the entire dataset file, which is the default output of NR-grep.

Running PatMatch on analysis servers

Computationally, the PatMatch program can potentially consume a significant amount of CPU time depending on the length and type of the sequence, the size and type of the sequence datasets being searched and other search parameters. The original configuration of PatMatch on the TAIR website was to execute the program directly on the web server, but occasionally the execution of the program would compete excessively with the web server for computer resources. Therefore, we redesigned the program to execute the computationally intensive search algorithm on a remote system of Linux computers. This expandable system currently consists of three independent nodes where one is responsible for balancing the requests between each node. Advantages of this system include load balancing, expandability, stability and ease of maintenance.

CGI interface modifications

The web interface was also updated. Where the old versions of PatMatch used C programs to display the results on the web, the new PatMatch uses Perl CGI scripts that are easier to maintain for updates such as changing links on the results pages.

DISCUSSION

Rationale for improvements

For PatMatch to function as desired, a string matching tool that could efficiently handle searches for patterns containing regular expressions, wildcard characters and inexact matching to a degree specified by the user was required. The previous version of PatMatch at TAIR and SGD used scan_for_matches (5), a program that is capable of searching for complex patterns in DNA and protein sequences using a brute-force backtracking search algorithm. Queries in scan_for_matches are based on patterns that allow ambiguous codes as well as substitutions, insertions and deletions. In addition, users are able to specify weight matrices as patterns, although this feature was not used by PatMatch.

While powerful, scan_for_matches does have some limitations. Its patterns do not support repetitions of subpatterns, a feature desired in PatMatch. The previous version of PatMatch that used scan_for_matches was able to support repetitions of subpatterns only by translating a query that includes repetitions into multiple queries. In addition, scan_for_matches allows for inexact matching of simple patterns, but it is unable to apply inexact matching on a query pattern that is made up of several simple patterns.

A grep-like tool for nucleotide and peptide sequences seemed to be an appropriate choice to replace scan_for_matches. The agrep string matching tool (6) meets these requirements, but has the undesirable effect of widely different search times for patterns of different complexity (4). NR-grep, a free approximate string matching program based on the Backward Nondeterministic Dawg Matching (BNDM) algorithm (7), was chosen to search for pattern hits in PatMatch. NR-grep is able to search for regular expressions exactly or allowing errors in the match using a bit-parallel simulation of nondeterministic suffix automation for pattern matching. It has the advantage over agrep when searching for complex patterns due to its smoothness in search time. In addition, NR-grep's bit parallel suffix automation is faster than the backtracking algorithm used by scan_for_matches when searching for patterns with mismatches, although both programs could take a long time to return results if the user enters a loose pattern that matches many subsequences in the dataset.

Comparison with similar tools

Two types of pattern matching algorithms commonly used in biology are scan_for_matches and grep-like programs. PatScan (5) provides another program where scan_for_matches is used to search a dataset for matches against a query pattern. PatSearch (8,9) is a program based on scan_for_matches that has added features such as the assessment of the statistical significance of pattern hits using a Markov chain simulation. Currently, PatMatch does not assess the statistical significance of hits returned by NR-grep nor does it support weight matrices in the query pattern.

Another group of software for finding user-specified patterns in DNA and protein sequences uses tools from the grep family of string matching algorithms or are based on grep. A grep-like tool called tacg (10) supports regular expressions, IUPAC degeneracies, searching with errors and probability matrices. While PatMatch does not support searches with probability matrices, tacg does not support searches that allow for insertions and deletions. In addition, tacg has the disadvantage of being slower than the grep family of tools as well as an inefficient algorithm for finding degenerate matches (10). eMOTIF-SCAN is a program using the agrep tool that supports approximate matching and regular expression searches against the eMOTIF (11) database of protein sequence motifs. PatMatch has an advantage over eMOTIF-SCAN in that it uses NR-grep, which has been shown to be the fastest string matching tool for complex searches (4).

The fuzznuc and fuzzpro programs available from the European Molecular Biology Open Source Software Suite (EMBOSS; http://emboss.sourceforge.net) provide pattern searches for nucleotide and protein sequences. We considered using these programs to replace scan_for_matches in PatMatch. However, fuzznuc and fuzzpro do not support repetitions of groups of nucleotides or amino acids. This was a desired feature of PatMatch and was a reason why NR-grep was chosen over fuzznuc and fuzzpro.

Limitations of PatMatch and future plans

The changes made to PatMatch have improved performance and allowed for pattern searching using a flexible regular expression syntax, including wildcard characters and mismatches within a single pattern. The main limitation of PatMatch is that results are returned without any evaluation of their significance. Patterns are returned in the order that they were found by NR-grep. In addition, PatMatch does not support searches with probability matrices. For complex queries, scan_for_matches allows simple patterns that can be used as variables in a complex pattern made up of several simple patterns, which may be easier to use than the PatMatch regular expression syntax for complex patterns.

We are continuing to make improvements to PatMatch in response to user requests to make the software more useful. Future work includes the ability to query using Boolean logic such as '<pattern A> and <pattern B>'. More sophistication in pattern matching algorithms will be useful in extending our knowledge about the complexity of organizations, architectures, and patterns found in DNA and protein sequences.

ACKNOWLEDGEMENTS

We thank Drs Leonore Reiser and Eva Huala for reading the manuscript and providing us with valuable feedback. We thank the Carnegie Summer Research Internship Program for providing a stimulating environment for T.Y. while carrying out part of this work. This work was supported, in part, by the National Science Foundation grant number DBI-9978564. Funding to pay the Open Access publication charges for this article was provided by NSF grant number DBI-9978564.

Conflict of interest statement. None declared.

REFERENCES

1. Cherry,J.M., Ball,C., Weng,S., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K. and Botstein,D. (1997) Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, **387**, 67–73.
2. Huala,E., Dickerman,A., Garcia-Hernandez,M., Weems,D., Reiser,L., LaFond,F., Hanley,D., Kiphart,D., Zhuang,J., Huang,W. *et al.* (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.
3. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
4. Navarro,G. (2001) NR-grep: a fast and flexible pattern matching tool. *Software Practice and Experience*, **31**, 1265–1312.
5. D'Souza,M., Larsen,N. and Overbeek,R. (1997) Searching for patterns in genomic data. *Trends Genet.*, **13**, 597–498.
6. Wu,S. and Manber,U. (1992) Agrep—a fast approximate pattern-matching tool. In *Proceedings of USENIX Technical Conference*. USENIX Association, Berkeley, CA, pp. 153–162.
7. Navarro,G. and Raffinot,M. (1998) A bit-parallel approach to suffix automata: fast extended string matching. *9th International Symposium on Combinatorial Pattern Matching (CPM'98)*, ***LNCS* 1448**, pp. 14–33.
8. Pesole,G., Liuni,S. and D'Souza,M. (2000) PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, **16**, 439–450.
9. Grillo,G., Licciulli,F., Liuni,S., Sbisa,E. and Pesole,G. (2003) PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequence. *Nucleic Acids Res.*, **31**, 3608–3612.
10. Mangalam,H.J. (2002) tagc—a grep for DNA. *BMC Bioinformatics*, **3**, 8.
11. Huang,J.Y. and Brutlag,D.L. (2001) The EMOTIF database. *Nucleic Acids Res.*, **29**, 202–204.
12. Stockinger,E.J., Gilmour,S.J. and Thomashow,M.F. (1997) *Arabidopsis thaliana* CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a *cis*-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. *Proc. Natl Acad. Sci. USA*, **94**, 1035–1040.