W592–W596 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 28 May 2012*
doi:10.1093/nar/gks448

GGRNA: an ultrafast, transcript-oriented search engine for genes and transcripts

Yuki Naito and Hidemasa Bono*

Database Center for Life Science, Research Organization of Information and Systems, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032 Japan

Received February 25, 2012; Revised April 23, 2012; Accepted April 29, 2012

ABSTRACT

GGRNA (http://GGRNA.dbcls.jp/) is a Google-like, ultrafast search engine for genes and transcripts. The web server accepts arbitrary words and phrases, such as gene names, IDs, gene descriptions, annotations of gene and even nucleotide/amino acid sequences through one simple search box, and quickly returns relevant RefSeq transcripts. A typical search takes just a few seconds, which dramatically enhances the usability of routine searching. In particular, GGRNA can search sequences as short as 10 nt or 4 amino acids, which cannot be handled easily by popular sequence analysis tools. Nucleotide sequences can be searched allowing up to three mismatches, or the query sequences may contain degenerate nucleotide codes (e.g. N, R, Y, S). Furthermore, Gene Ontology annotations, Enzyme Commission numbers and probe sequences of catalog microarrays are also incorporated into GGRNA, which may help users to conduct searches by various types of keywords. GGRNA web server will provide a simple and powerful interface for finding genes and transcripts for a wide range of users. All services at GGRNA are provided free of charge to all users.

INTRODUCTION

Searching for genes and transcripts from public databases is a routine task for biologists. However, it requires users to select a suitable database or web service according to the search terms; e.g. gene names, accession numbers, gene descriptions, annotations of gene or nucleotide/amino acid sequences. Searches by gene names, accession numbers or certain types of keywords can be performed on GenBank (1) to obtain comprehensive results. However, searching from GenBank usually returns a huge number of results from various organisms with redundant entries, because GenBank is an archival repository of all the original sequences submitted to, and exchanged among GenBank/EMBL/DDBJ (2). Users are required to narrow down their search results by specifying the organism and give additional keywords in order to reach the content that they are really interested in. On the other hand, searches by nucleotide or amino acid sequences can be performed using BLAST (3) searches on the web (e.g. http://blast.ncbi.nlm.nih.gov/) (4), but the searches are usually queued instead of returning the results immediately. For searching genomic sequences, faster web services have been proposed, such as BLAT (http://genome.ucsc.edu/) (5) and TDSE (http://www.dnasoso.com/) (6), but no web services are available for searching gene and transcript sequences very quickly.

In this article, we present GGRNA (http://GGRNA.dbcls.jp/), a GooGle-like search engine for RNA molecules, which can efficiently find genes and transcripts by utilizing the compressed suffix array (7). The server accepts various words, phrases, and sequences in one simple search box and quickly returns relevant RefSeq (8) transcripts with the queried keywords highlighted. RefSeq is a curated, non-redundant source of sequence information maintained by NCBI. RefSeq includes genomic, mRNA, non-coding RNA (ncRNA) and protein records. Of these, GGRNA uses RefSeq mRNA (accession starts with NM/XM) and ncRNA (NR/XR) as the main sources of database because RefSeq provides only a single entry for each transcript. Our system can search sequences as short as 10 nt or 4 amino acids, which cannot be handled easily by popular sequence analysis tools, such as NCBI BLAST and UCSC BLAT. GGRNA will provide the fastest and easiest way to search genes and transcripts for a wide range of users. All services at GGRNA are provided free of charge to all users.

INTEGRATED DATABASE OF TRANSCRIPTS

Since it is common for users to search using heterogeneous words and phrases that are not acceptable to existing search engines, we mapped multiple types of information such as Gene Ontology (GO) terms/IDs and Enzyme Commission (EC) numbers, onto each RefSeq transcript

*To whom correspondence should be addressed. Tel: +81 3 5841 6754; Fax: +81 3 5841 8090; Email: bono@dbcls.rois.ac.jp

using Entrez Gene (9) data and created a transcript-oriented, integrated database (GGRNA database). The GGRNA database is a text-based database that includes data on human, mouse, rat, chicken, frog, zebrafish, fly, worm, *Ciona*, *Arabidopsis*, rice, budding yeast and fission yeast. We plan to incorporate more species when sufficient numbers of transcripts are available in RefSeq. The GGRNA database will be updated every two months in pace with RefSeq releases.

SEARCH ENGINE

When a search is performed, GGRNA first identifies the query type of the terms. A term identified as being an accession number, Gene ID or gene symbol is searched using MySQL relational database, which separately stores these fields of the GGRNA database. Otherwise, full-text searches of terms and phrases are made against the GGRNA database, using Sedue software (Preferred Infrastructure), which utilizes the compressed suffix array algorithm. The software compactly stores the index in main memory to perform fast and accurate text retrieval, and is well suited for searching not only words and phrases but also nucleotide/amino acid sequences. For example, the 10 nt sequence 'GACCTTGAAC', or 4 amino acid sequence 'IETD', can be exhaustively searched from human RefSeq in <1 s through the GGRNA web server. A typical search takes less than a few seconds, but will take longer when there are a large number of total hits. GGRNA can search nucleotide sequences allowing up to three mismatches, or the query sequences may contain degenerate nucleotide codes (e.g. N, R, Y, S) by specifying an option described below. These are especially useful for handling short sequences, such as oligonucleotide primers/probes and protein motifs. The remarkably fast response of GGRNA should enhance the usability of routine searching.

WEB SERVER IMPLEMENTATION

Overview

GGRNA web server accepts arbitrary keywords in one simple search box (Figure 1A). The server quickly returns relevant RefSeq transcripts, with the queried keywords highlighted (Figure 1B). If the query matches nucleotide or amino acid sequences, matched positions are indicated with numbers (Figure 1C and D). This function is useful for checking the target sequence positions of PCR primers, probes or siRNA/miRNAs. Probe IDs from catalog microarrays (e.g. 1552311_a_at, A_23_P101434) are converted into corresponding probe sequences and searched for their binding sites. For Affymetrix probe set ID, GGRNA searches for the probe set sequences, i.e. a set of eleven or more 25-mer probe sequences all together (Figure 1D). For Agilent probe ID, a single 60-mer probe sequence is searched.

Search operators

Users can refine their query using the search operators listed in Table 1. For example, searching for 'VIM' without an operator will return a number of gene transcripts including the VIM (vimentin) gene, the amino acid sequence VIM (Val-Ile-Met) and the cited references containing Kivimaki. However, using the search operator 'symbol:VIM' tells GGRNA that the user is searching by gene name only. Similarly, using the operator 'aa:VIM' will restrict the results to those matching the amino acid sequence only. Search operators are used not only for restricting the search results but also for activating certain options. Nucleotide sequences can be searched allowing 1- to 3-mismatches using seq1:, seq2: or seq3: operator. Degenerate bases represented by IUB code letters (e.g. N, R, Y, S) are expanded using iub: operator. For searching complementary nucleotide sequences, comp: operator can be used.

Advanced search

Alternatively, queries can be easily refined using the 'Advanced search' form (Figure 1E). This search interface provides additional fields that may help to qualify searches by various criteria. All of the search terms entered into the form are transformed automatically into a single query string containing search operator(s), which is shown at the bottom of the page (Figure 1E). Copy–pasting this query string in the GGRNA regular search box will give the same results.

DATA EXPORT

Search results can be exported as tab-delimited text from the bottom of the result page. Users can copy–paste the results into a spreadsheet application or a text editor for downstream analysis. The results can also be downloaded as a separate file by clicking the 'download' button. Alternatively, CSV or JSON output can be obtained via GGRNA Application Programming Interface (API), as described below.

GGRNA REST API

GGRNA provides a simple Representational State Transfer (REST) API that enables users to perform searches with their client programs in an automated manner. The search results can be retrieved through the following URI:
http://GGRNA.dbcls.jp/api/*SPECIES*/*QUERY*[.*FORMAT*]
SPECIES: 'hs' (*Homo sapiens*), 'mm' (*Mus musculus*), etc.
QUERY: a simple keyword or a URI-encoded string.
FORMAT: select txt or json as output format. (optional)
Currently available *SPECIES* and *FORMAT* types are listed in the following URI:
http://GGRNA.dbcls.jp/api/
For example, a search for the string 'caagaagagattg' in human is represented as follows:
http://GGRNA.dbcls.jp/api/hs/caagaagagattg
A JSON formatted output is retrieved by adding '.json' suffix in URI:
http://GGRNA.dbcls.jp/api/hs/caagaagagattg.json
For a phrase search, note that a double quote is encoded as '%22', and a space character is encoded as '+' in the URI. A search for the phrase ''RNA interference'' in *Caenorhabditis elegans* is represented as follows:
http://GGRNA.dbcls.jp/api/ce/%22RNA+interference%22

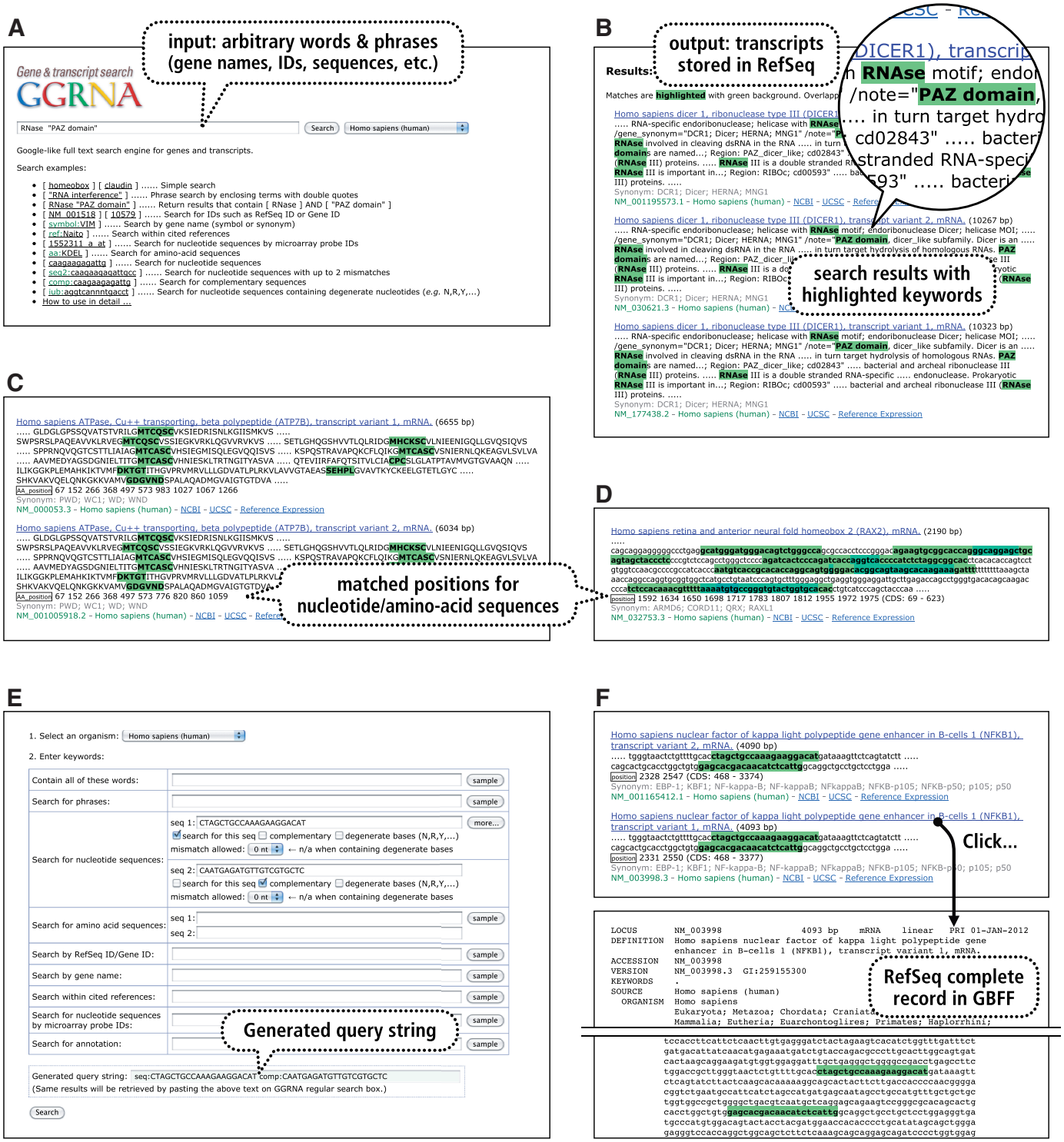

Figure 1. Screenshot from GGRNA. (**A**) Top page (http://GGRNA.dbcls.jp/). (**B**) Typical search output of GGRNA. The phrases 'PAZ domain' and 'RNase' are searched in human transcripts. (**C**) Amino acid sequence search. The sequences MTCQSC, MHCKSC, MTCASC, CPC, DKTGT, SEHPL and GDGVND are searched simultaneously. (**D**) Affymetrix microarray probe set ID, 1552311_a_at, is automatically expanded into eleven corresponding probe sequences and searched for their binding sites. (**E**) Advanced search page. All of the search terms are transformed into the single query string shown at the bottom of the page. (**F**) An example of searching PCR primer binding sites. Note that each number corresponds to the first base in the matched sequence. Clicking on a transcript displays the complete record from RefSeq in GenBank Flat File (GBFF) format.

Table 1. Search operators in GGRNA

Search operators	Description	Alias
refid:NM_001518	Search by RefSeq ID. Version number following a dot is ignored: [refid:NM_003380.2] and [refid:NM_003380] will return the same results. Words starting with NM_, XM_, NR_ or XR_ are automatically treated as refid: search without operator.	refseqid: refseq: id:NM_, id:XM_, id:NR_, id:XR_
geneid:10579	Search by Gene ID. An integer is automatically treated as geneid: search without operator.	gene:integer id:integer
symbol:VIM	Search for gene symbols and synonyms which partially match to the query. For example, query EIF2C will return EIF2C1, 2, 3 and 4.	name:
aa:KDEL	Search for amino acid sequence.	
ref:Naito ref:1327585	Full text search within cited references. PubMed ID can also be queried.	reference:
probe:1552311_a_at probe:A_23_P101434	Search for nucleotide sequences by microarray probe ID. Words ending with _at, _st (Affymetrix ID) and starting with A_ (Agilent ID) are automatically treated as probe: search without operator. When probe ID is not converted into sequences, the probe ID is subjected to a regular search.	probeid:
anot:GO:0006915 anot:[apoptosis] anot:"EC 2.3.1.51"	Search for annotation. - Search by Gene Ontology ID and term - Search by Enzyme Commission (EC) number	annotation: annot:
seq:caagaagagattg seq1:caagaagagattg seq2:caagaagagattgcc seq3:caaggagagatgggacac	Search for nucleotide sequence. Query containing letters A, T, G, C and U only will automatically be treated as seq: search without the operator. U and T will be treated identically. seq1:, seq2: and seq3: will return results with 1-, 2- and 3-nt mismatch tolerance.	sequence: sequence1: sequence2: sequence3:
comp:caagaagagattg comp1:caagaagagattg comp2:caagaagagattgcc comp3:caaggagagatgggacac	Search for complementary sequence. comp1:, comp2: and comp3: will return results with 1-, 2- and 3-nt mismatch tolerance.	complementary: complementary1: complementary2: complementary3:
both:caagaagagattg both1:caagaagagattg both2:caagaagagattgcc both3:caaggagagatgggacac	Simultaneously retrieve sense and antisense nucleotide sequences corresponding to the query. both1:, both2: and both3: will return results with 1-, 2- and 3-nt mismatch tolerance.	bothseq: bothseq1: bothseq2: bothseq3:
iub:yyaaggnnnagacac iubcomp:yyaaggnnnagacac iubboth:yyaaggnnnagacac	Search for nucleotide sequence containing IUB code letters (e.g. N, R, Y, S). iubcomp: will return complementary sequences to the query; iubboth: will return both strands.	iubseq: → iub:

SEARCH EXAMPLES

The following two examples show typical applications of GGRNA. A video tutorial introducing more uses of GGRNA (http://GGRNA.dbcls.jp/en/togotv/) is available at TogoTV (http://togotv.dbcls.jp/en/) (10), a collection of freely available tutorial videos for bioinformatics resources maintained by Database Center for Life Science, Japan.

Searching PCR primer binding sites

GGRNA can quickly map the location of PCR primer binding sites within a target gene and estimate the expected size of the PCR product. For example, searching for 'CTAGCTGCCAAAGAAGGACAT comp:CAATG AGATGTTGTCGTGCTC' in human will return two transcript variants of NFKB1 (NM_001165412 and NM_003998, as of RefSeq release 52, March 2012) as the target gene (Figure 1F). Note that the results contain all keywords entered into the search box, separated with spaces or commas; also note that the 'comp:' operator searches for complementary sequences of the reverse primer. Alternatively, inputting the two primer sequences in the 'Advanced search' page, as shown in Figure 1E, will return the same results. The expected size of the PCR product can be estimated using the matched positions indicated in numbers (Figure 1F). Clicking on a transcript displays the complete record from RefSeq in GenBank Flat File (GBFF) format with the queried keywords highlighted.

Searching short amino acid sequence motifs

The amino acid sequence 'KDEL' at the C-terminus serves as the endoplasmic reticulum (ER) retention signal (11). To search for the KDEL motif in GGRNA, enter 'aa:KDEL' in the search box. An operator 'aa:' restricts the search to within amino acid sequences only. Searching 'aa:KDEL' in human will retrieve 359 results (as of RefSeq release 52, March 2012), but these results contain KDELs that are not at the C-terminus. On the other hand, transcripts annotated as GO:0005783 (endoplasmic reticulum) can be retrieved by searching 'GO:0005783', which returns 1985 results. An intersection of these two searches can be obtained by entering the two keywords separated by a space: 'aa:KDEL

GO:0005783', which returns 28 results. Of these, 13 results contain the KDEL motif at the C-terminus. Searching by sequences and other keywords simultaneously is one of the unique advantages of GGRNA.

ACKNOWLEDGEMENTS

We thank Dr Takeru Nakazato, Dr Hiromasa Ono and Mr Tazro Ohta for helpful discussions and comments, Mr Masamichi Chichii for making the tutorial video in TogoTV. We also thank users of GGRNA for providing the feedback, which has greatly improved the service.

FUNDING

Life Science Database Integration Project, National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST). Funding for open access charge: Life Science Database Integration Project.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
2. Karsch-Mizrachi,I., Nakamura,Y. and Cochrane,G. on behalf of the International Nucleotide Sequence Database Collaboration (2012). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **40**, D33–D37.
3. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
4. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.
5. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
6. Liang,W. and Bo,F. (2011) How to build a DNA search engine like Google? *J. Comput. Sci. Syst. Biol.*, **4**, 81–86.
7. Grossi,R. and Vitter,J.S. (2000) Compressed suffix arrays and suffix trees with applications to text indexing and string matching. In: *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 397–406.
8. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
9. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
10. Kawano,S., Ono,H., Takagi,T. and Bono,H. (2012) Tutorial videos of bioinformatics resources: online distribution trial in Japan named TogoTV. *Brief. Bioinform.*, **13**, 258–268.
11. Munro,S. and Pelham,H.R. (1987) A C-terminal signal prevents secretion of luminal ER proteins. *Cell*, **48**, 899–907.