ライフサイエンス新着論文レビュー・DBCLS からの成果発信

FIRST AUTHOR'S

統合遺伝子検索 GGRNA：遺伝子を Google のように検索できるウェブサーバ

2012 年 5 月 28 日

内藤雄樹・坊農秀雅
（ライフサイエンス統合データベースセンター）

email：**内藤雄樹，坊農秀雅**

GGRNA: an ultrafast, transcript-oriented search engine for genes and transcripts.
Yuki Naito, Hidemasa Bono
Nucleic Acids Research, **40** (Web Server issue), W592–W596 (2012)

要 約

GGRNA（http://GGRNA.dbcls.jp/）は，遺伝子や転写産物を Google のようにすばやく検索できるウェブサーバである．検索キーワードとして，遺伝子名やアクセッション番号など各種の ID をはじめ，遺伝子の機能やタンパク質のドメイン名，さらには，塩基配列やアミノ酸配列など，あらゆる語句を単一の検索窓に入力するだけで RefSeq に登録された転写産物をすばやく探し出すことができる．とくに，塩基配列の検索においてはあいまいな塩基を含むパターンやミスマッチを含む配列にも対応し，一般的な配列類似性検索サイトでは検索の困難な 10 塩基ほどの短い配列でも高速な検索が可能である．本稿では，GGRNA ウェブサーバの概要を解説するとともに，GGRNA の具体的な活用事例を紹介する．GGRNA のすべての機能は無償で自由に利用できる．

はじめに

公共データベースから遺伝子を検索するという作業は，多くの生命科学系あるいは医学系の研究者にとり日常的なものであろう．しかしながら，既存のデータベースを利用して目的の遺伝子の情報をすばやく探し出すことは必ずしも容易でない．ユーザが入力する検索キーワードとしては，遺伝子名，アクセッション番号などの各種の ID，遺伝子の機能に関する語句，タンパク質のドメイン名，関連する疾患，さらには，塩基配列やアミノ酸配列など，実に多様な内容が想定される．従来であれば，遺伝子名や各種 ID は GenBank [1] などの塩基配列データベースから検索し，タンパク質の機能，細胞内局在，ドメイン名，疾患などのキーワードは Gene Ontology や文献情報をたよりに探す，塩基配列やアミノ酸配列は BLAST [2] や BLAT [3] のような配列類似性検索ツールを用いるなど，検索キーワードの種類により複数のデータベースやウェブツールを使い分ける必要があった．しかし，これは煩雑なうえ，個々の検索に適したデータベースやウェブツールを把握していなければ効率的に情報を得ることはできない．また，GenBank などの既存のデータベースは，ひとつの遺伝子に対し複数のエントリーが存在するなど冗長であることも多く，たとえば，単純に遺伝子名で検索した場合でさえも多数のエントリーがヒットし目的の情報になかなかたどりつけない場合も多い．さらに，現在，塩基配列やアミノ酸配列の検索において広範に利用されている BLAST のウェブサーバ [4] は，結果が表示されるまで十秒から数十秒もかかるなど，ユーザが不便を感じることも多かった．

そこで，筆者らは，あらゆるキーワードを単一の検索窓に入力するだけで高速に遺伝子や転写産物を探せるようなウェブサービスを構築したいと考え，GGRNA（http://GGRNA.dbcls.jp/）を開発した（図 1）．

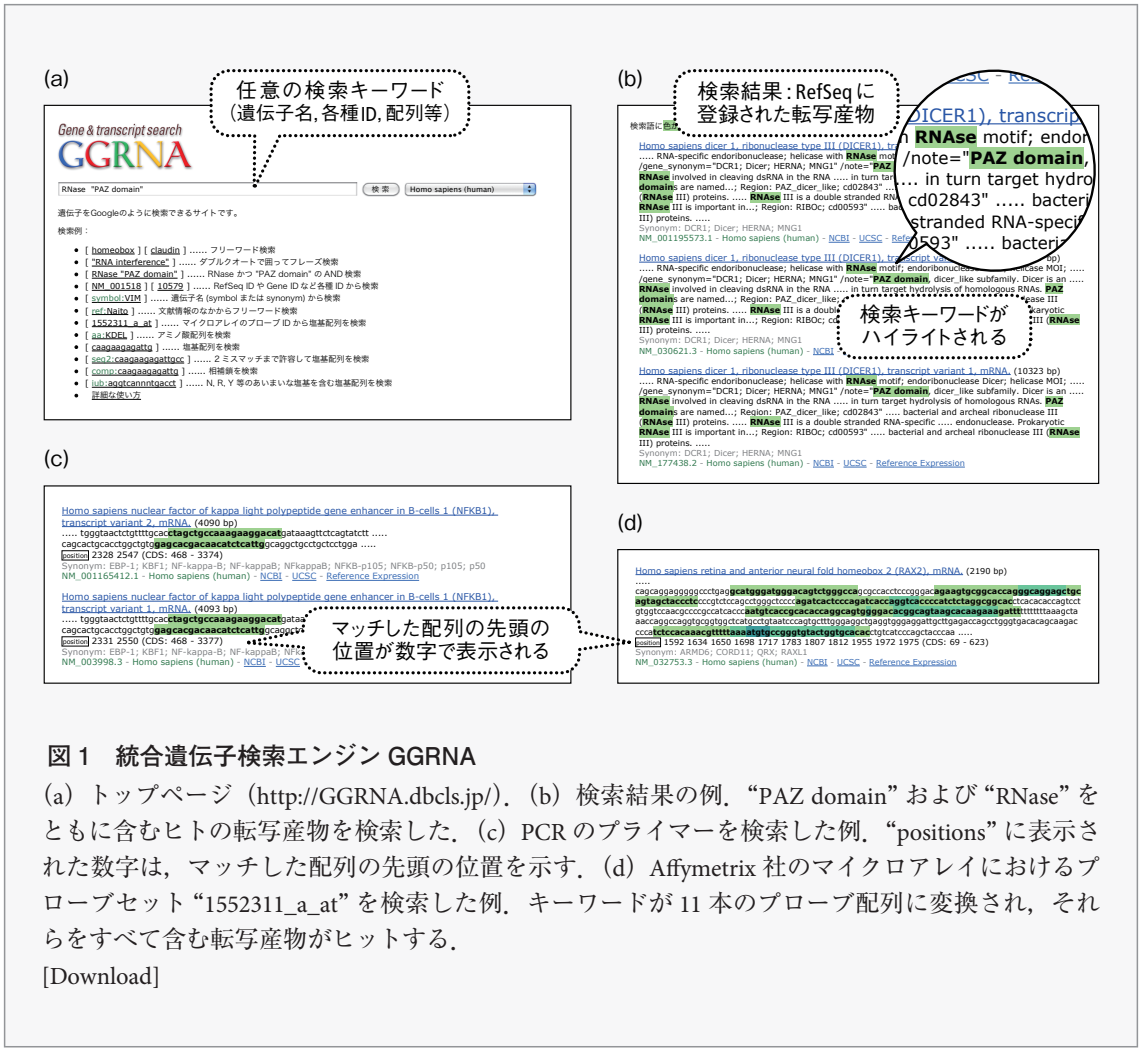

図 1　統合遺伝子検索エンジン GGRNA

（a）トップページ（http://GGRNA.dbcls.jp/）．（b）検索結果の例．“PAZ domain” および “RNase” をともに含むヒトの転写産物を検索した．（c）PCR のプライマーを検索した例．“positions” に表示された数字は，マッチした配列の先頭の位置を示す．（d）Affymetrix 社のマイクロアレイにおけるプローブセット “1552311_a_at” を検索した例．キーワードが 11 本のプローブ配列に変換され，それらをすべて含む転写産物がヒットする．

[Download]

1．GGRNA 検索エンジンの概要

GGRNA では，多様な検索キーワードを用いて遺伝子や転写産物を探せるよう，複数の公共データベースに由来する情報を RefSeq [5] の転写産物にひもづけて整理した “GGRNA データベース ” を構築している．RefSeq とは米国 NCBI（National Center for Biotechnology Information）の提供する重複のない塩基配列およびアミノ酸配列のデータベースで，GenBank/EMBL/DDBJ 国際塩基配列データベース [6]（INSDC）に登録された配列のなかから代表となるものを NCBI のスタッフが選択し注釈をつけたデータセットである．RefSeq にはゲノム，mRNA，非コード RNA，タンパク質の情報が登録されているが，GGRNA ではこのうち mRNA および非コード RNA の情報を用いている．重複のない RefSeq を基盤とすることにより検索の際に多数の転写産物が重複してヒットしてしまうことを防いでいる．さらに，これらの転写産物に対して，Gene Ontology の ID/term や酵素 EC 番号に関する情報などを NCBI Entrez [7] の情報を利用してひもづけることで GGRNA データベースとしている．現時点で，GGRNA データベースはヒト，マウス，ラット，ニワトリ，ツメガエル，ゼブラフィッシュ，ショウジョウバエ，線虫，ホヤ，シロイヌナズナ，イネ，出芽酵母，分裂酵母について構築されており，RefSeq のアップデートにあわせ 2 ヶ月に 1 回の頻度で再構築を行っている．

この GGRNA データベースを高速に検索するため，遺伝子名や ID など一部のキーワードは MySQL により検索，それ以外のキーワードは Sedue（Preferred Infrastructure 社）により全文検索を行っている．Sedue は圧縮接尾辞配列 [8] のインデックスをメモリに保持することによりもれのない検索をきわめて高速

に実行できるソフトウェアで，テキストの検索のみならず，塩基配列やアミノ酸配列の検索にも適している．たとえば，GGRNA において 10 塩基からなる塩基配列 “GACCTTGAAC” や 4 アミノ酸残基からなるアミノ酸配列 “IETD” を検索すると，どちらも 1 秒以内に結果が返ってくる．検索時間はほとんどのキーワードにおいて数秒であるが，ヒット件数が極端に多い場合は圧縮接尾辞配列の解凍に時間を要するため 10 秒以上かかる場合もある．

GGRNA のトップページ（図 1a），および，検索結果（図 1b）を示す．“PAZ domain” および “RNase” をともに含むヒトの転写産物を検索すると，検索結果は，この検索キーワードがハイライトされて表示される．

GGRNA では，Google と同様の方法によりいくつかの検索オプションを指定することができる．たとえば，“VIM” をキーワードとして検索すると，ビメンチン（VIM）の遺伝子，アミノ酸配列としての “VIM”（Val-Ile-Met），文献の著者としての “Kivimaki” などがすべてヒットするが，“symbol:VIM”，“aa:VIM”，“ref:Kivimaki” のように検索タグをつけることによりキーワードの種類を特定して検索することができる．また，“seq3:CAAGGAGAGATGGGACAC” として 3 塩基までのミスマッチを許容して検索したり，“iub:YYAAGGNNNAGACAC” としてあいまい塩基（N，R，Y，S など）を展開して検索したりすることも可能である．GGRNA で使用可能なタグの一覧はヘルプページを参照してほしい．

また，これらの検索タグを覚えていなくても同等の検索を簡単に行えるよう，別に Advanced search というページを用意している．このページのそれぞれの欄に検索キーワードを入れることで，検索タグと同様にキーワードの種類を限定して検索することができる．

2．GGRNA の活用事例

以下に，GGRNA の活用事例を紹介する．なお，使い方を解説した動画をライフサイエンス統合データベースセンターが提供する統合 TV [9] から公開しているので，そちらも参考にされたい．

・PCR のプライマー配列から増幅遺伝子や増幅領域を確認する　CTAGCTGCCAAAGAAGGACAT comp:CAATGAGATGTTGTCGTGCTC のようにして，foward プライマーの配列と，reverse プライマーの相補鎖の配列とを同時に検索すれば，PCR で増幅する遺伝子や増幅領域を確認できる．“comp:” は相補鎖検索のオプションである．この例をヒトにおいて検索すると，NFKB1 遺伝子の 2 つの転写産物（NM_001165412，NM_003998）がヒットする（図 1c）．塩基配列やアミノ酸配列がヒットした場合は先頭の位置が数字で表示されるので，PCR 産物の長さも計算できる．

・マイクロアレイのプローブ ID から塩基配列を検索する　マイクロアレイのプローブ ID を GGRNA で検索すると，自動的にプローブの配列へと変換したうえでその結合部位を検索してくれる．たとえば，Affymetrix 社のマイクロアレイのプローブセット ID“1552311_a_at” を検索すると，これに対応する 25 mer のプローブ配列× 11 本に変換され，それらの配列すべてにマッチする転写産物がヒットする（図 1d）．Agilent 社のマイクロアレイのプローブ ID“A_23_P101434” の場合は，60 mer のプローブ配列× 1 本に変換されて検索が行われる．

・タンパク質のモチーフ検索　タンパク質の C 末端に存在する KDEL 配列は小胞体係留シグナルとして機能することが知られている [10]．そこで，GGRNA において “aa:KDEL” と検索すると，RefSeq release 52（2012 年 3 月）の時点で 359 件がヒットした．しかし，このなかには C 末端以外の部位に KDEL という配列をもつものも多数含まれる．一方，Gene Ontology において小胞体を表す “GO:0005783” を検索すると，1985 件がヒットした．この 2 つの AND 検索 aa:KDEL GO:0005783 では 28 件がヒットし，このうち 13 件で C 末端に KDEL 配列が存在した．現時点では，C 末端の KDEL という配列を検索する方法は提供していないが，配列とそれ以外のキーワードとを組み合わせて思いついたことを気軽に検索できる点は，GGRNA の強みであると思われる．

3. データ出力機能と API の活用

GGRNA の検索結果を外部のソフトで利用しやすいよう，タブ区切りテキストを出力する機能を用意している．検索結果の最下部にあるテキストボックスからタブ区切りテキストをコピー＆ペーストするか，あるいは，ダウンロードボタンによりファイルを保存することができる．

また，GGRNA は REST API を提供しておりウェブ検索と同じ結果をタブ区切りテキストまたは JSON 形式にて得ることができる．詳細はヘルプページを参照してほしい．

おわりに

GGRNA という名称は“Google ライクな RNA 検索エンジン”を意味するが，筆者らは，開発の初期から非公式に GGRNA を“ぐぐるな”と発音している．Google のように強力で誰にでも使いやすい検索エンジンをめざし，遺伝子の検索においてはググらなくてもこのサービスにより効率的に情報を得られるようにしたいという目標がある．

現時点で，GGRNA は RefSeq で提供されている生物種のうち 13 種に対応しているが，2012 年度中に RefSeq の全体，さらには，GenBank/EMBL/DDBJ 国際塩基配列データベースに登録されたゲノム以外の配列全体を検索できるよう，DDBJ（DNA Data Bank of Japan，日本 DNA データバンク）と協力しながら開発を進めていきたいと考えている．GGRNA が読者の日々の研究に少しでも役だてば幸いである．

文 献

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

著者プロフィール

内藤 雄樹（Yuki Naito）
略歴：2007 年 東京大学大学院理学系研究科博士課程 修了，同年 東京大学大学院理学系研究科助教を経て，2011 年よりライフサイエンス統合データベースセンター 特任助教．

坊農 秀雅（Hidemasa Bono）
ライフサイエンス統合データベースセンター 特任准教授．