研究論文

ゲノム配列解析ツール giv の開発†

竹内敬一朗*，中村建介*

Development of a genome information viewer - giv

Keiichiro Takeuchi* and Kensuke Nakamura*

We here describe gve, a new CUI tool for analyzing genome sequence.

Key words： genome sequence, bioinformatics, next generation sequencer, FASTA format, CUI

1 はじめに

生物の遺伝情報は DNA 分子を構成する 4 種類の塩基（ATGC）の並び順により保持されている．1970 年代に Sanger 法が開発され、ゲノムと呼ばれるそれぞれの生物に固有の遺伝情報を読み取る技術が確立された．これにより 1995 年にインフルエンザ菌 [1]とマイコプラズマ [2]のバクテリアゲノムがはじめて解読され、2001 年にはヒトゲノム計画が完了した [3,4]．さらに 2005 年頃より次世代シーケンサーと呼ばれるゲノム塩基配列の高性能な読み取り装置が普及し始めたことで、ゲノム配列に関する情報の蓄積が飛躍的に増大しつつある．2018 年 2 月現在で NCBI に登録され、ダウンロード可能な染色体またはアセンブリ配列のある生物種数は 34,874 種にものぼる．現在進行中のプロジェクトも数多く、さらに配列解析技術もここ数年の PacBio、ナノポアなど、一分子シーケンシングの急速な発展により、研究室レベルで特定の生物種のゲノム解読も可能な時代になりつつある [5]．こうしたゲノムに関するデータは NCBI や JGI といったウェブサイトで公開されており自由にダウンロードすることが出来る．また、配列データを視覚的に解析するツールとして、NCBI の Genome View や UCSC Genome Browser, IGV[6]など多くのものが公開されている（Fig. 1）。これらのツールはそれぞれのデータベースサイトにあるゲノムデータを可視化するため、グラフィカルなユーザーインターフェース（GUI）によりゲノム内の遺伝子領域の分布などの全体像を把握することに優れている．

Fig. 1 GUI Genome Viewers (Left: NCBI Genome View, Right: UCSC Genome Browser)

GUI ツールは直感的に理解しやすく初心者にも取り扱い易い一方で、ある程度の経験を積んだユーザーによる特定の目的を伴う解析については、マウスを使わずキーボード上だけで操作が完結する CUI（キャラクターベースドユーザーインターフェース）によるツールの方が作業効率の高い場合が多い．Linux を含む Unix 環境がこの考え方に基づいて設計されており、典型的なアプリケーションとして、テキストエディターである vi（vim）や emacs が挙げられる．

本稿で紹介するゲノム配列解析ツール giv (genome information viewer)はシンプルなゲノム情報ビューワーであり、vi に準じたキーストロークの少ない操作により、ゲノム解析を効率的に進められることを目的としている．

我々の研究室で日常的に行っている次世代シーケンサーのデータ解析においては、リード配列と参照配列との比較が主要な作業のひとつとなっている．マッピングやアセンブル等のルーチン的な解析はアプリケーションにより自動的に行うことができるが、その後の詳細な解析においては参照ゲノムの部分配列やその位置を一塩基単位で観測することがしばしばあり、我々自身が研究を進めていく上で、こうしたツールの必要性を感じたことにより開発を始めたものである．

2 方法

2・1 入力フォーマット

特定の生物種のゲノムは 1 本あるいは複数の DNA 塩基配列から構成される．それぞれの塩基配列に関するデータは、>印から始まる配列 ID、生物種名、株名などを含むタイトル行に続いて、DNA 塩基を記述する A,T,G,C の文字の並びを記した行で表される FASTA（ファスタ）形式と呼ばれるフォーマットで記述されている．この

† 原稿受理 平成３０年２月２８日 Received February 28, 2018

* 生命情報学科 (Department of Life Science and Informatics)

FASTA あるいはマルチファスタファイルが giv の塩基配列入力フォーマットとなる.

FASTA フォーマットはテキストファイルであり、塩基配列の部分はエディターで開いた際の視認性のために、横幅 50 塩基から 80 塩基程度の行に分割され、各行の終端には改行コードが入っている。そのため vi など通常のテキストエディターで特定の塩基配列を検索しても、その配列が行をまたいで改行を挟む場合にはヒットしない. また、通常のテキストエディターでは、塩基配列の検索時には必須となる相補逆鎖の検索が容易に行えないほか、カーソルのある位置が、改行コードやタイトル行の影響によりゲノム配列上の塩基位置に対応しないため、塩基配列の解析には専用のプログラムが必要であると考え giv の開発を行った.

ゲノム解析を行なう際の情報としては塩基配列情報の他に遺伝子領域等の機能領域に関する情報も有用である。現時点の giv では塩基配列の操作が主であり遺伝子領域等に関する情報表示については試作段階であるが、これらの情報を記述するフォーマットとしては Genbank フォーマットにより記述されたファイルを用いることができる.

2・2　開発環境

プログラムは C により記述し MacOSX 上で開発した。UNIX 系システムで動作する端末制御ライブラリ curses を用いることで、他のコンソールウィンドウを持つ UNIX 系プラットホームにも比較的容易に移植可能である.

3　結果

3・1 プログラム

3・1・1 起動と終了

対象とするゲノム配列を含む FASTA ファイル（genome.fasta）に対して端末ソフトウェア（ターミナル）上で `giv genome.fasta` とタイプすることで起動する. ゲノムが複数の DNA 配列からなる場合にはタイトル行と塩基配列行の並びが単純に複数回繰り返すマルチファスタと呼ばれる形式のファイルを用意する。終了時には `cntl-c` をタイプする.

3・1・2 操作画面

起動後の操作画面の概観を Fig. 2 に示す.

Fig.2 Appearance of giv

最上段には複数の DNA 配列を読み込んだ場合に、操作対象とする配列を切り替えるためのタブが表示される. また、画面最下段は配列検索等を行なう際にコマンド・検索配列等ユーザーからのキーボード入力を表示する画面であり、下から二段目のステータスバーの行には、ファイル名、タブの総数と表示されているタブ番号、配列の全長と表示範囲、塩基位置とカーソル位置が表示される. 操作中にマウス操作によりウィンドウサイズを変更することも可能で、サイズの変化に追従して表示画面の行数および列数も変化する.

3・1・3 カーソル操作

カーソルの移動は vi と同様に h（左）l（右）k（上）j（下）とすることで、ホームポジションから手を離さずに操作できる. タブ間の移動は ctrl-w を押すことで次のタブへ移動するか、タブ番号のファンクションキーを押すことで直接切替えることも出来る. 画面を広く取るとカーソル位置を見失うことがあるが、c キーを押すとカーソル位置が強調表示される. 現時点で実装されているその他のキーバインディングによる操作のリストを Table 1 に示す. Table 1 には記載されていないが、数値を入力すると Fig.2 の③の行に表示され、続けて打つコマンドがその回数繰り返される. 例えば 3j と続けて打つと 3 行下へ移動する。また、特定の塩基位置へ移動したい場合、数値に続けて g を入力することで移動できる. たとえば 123 塩基目にカーソルを移動したい場合、123g と入力する. カーソル位置にマークをすることも出来る. 現在のカーソル位置にマークをするためには m の後にマークする文字を打つ. たとえば ma と打てばその位置が文字 a でマークされる. 任意の場所からシングルクォートに続けて 'a の様に打つことでマーク位置に戻ることができる。

Table 1 List of Key Bindings

	キー操作	結果
カーソル	<ESC>	操作のキャンセル
	h j k l	カーソルの移動（左下上右）
	Ctrl-Y Ctrl-E	行単位の移動　（上下）
	Ctrl-U Ctrl-D	半画面移動　（上下）
	Ctrl-B Ctrl-F	全画面移動　（上下）
	H M L	画面の上・中・下段にカーソル移動
	^ $	左端・右端にカーソル移動
	|	画面左上隅にカーソル移動
	zt zz zb	カーソル行を画面の上・中・下段に
	g G	タブ内先頭・末尾へカーソル移動
	n N	バッファ内文字列を順・逆方向へ検索
	m{char} '{char}	マークとマーク位置への移動
	c	カーソル位置の強調表示
検索	/sequence ?sequence	順・逆方向へ検索
ページ	Ctrl-w	次のタブへ移動
	<F(n)> ファンクションキー	n 番目のタブへ移動

3・1・4 配列検索

タブ内の塩基配列の検索には vi と同様にスラッシュ記号 (/) を入力することでコマンド入力行にカーソルが

移動するので、その後に検索したい塩基配列を入力しEnter キーを押すことで、現在のカーソル位置の直後に現れる当該文字列位置にカーソルが移動する(前方検索). カーソル位置よりも前にある文字列を検索したい場合にはスラッシュの代わりにクェスチョンマーク (?) を用いる (後方検索)。画面内に同じ文字列のマッチが複数箇所ある場合にはすべての一致箇所が強調表示され、順方向にマッチした箇所と、相補逆差として逆方向にマッチした箇所は異なる色で強調表示される。いちど検索した文字列はバッファに入り、続けて同じ文字列を検索したい場合には n キーを押せば前方検索、N キーを押せば後方検索を何度でも続けることが出来る.

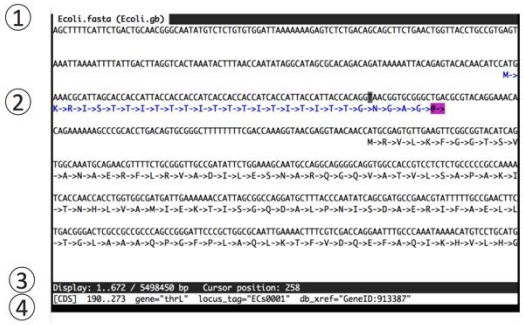
3・1・4 範囲指定

viエディターにはビジュアルモードという部分文字列の選択機能がある。v キーを押せばそのカーソル位置から文字列の選択が始まりカーソルを移動させることで選択したい範囲を指定できる。giv でも同様の範囲指定方法が利用可能である。範囲を指定後 スラッシュキー (/) を押すことで、コマンド入力行にカーソルが移り、続けて Enter キーを押すことで、選択した文字列とおなじ配列領域が検索される。その後は通常の配列検索と同様 n あるいは N キーにより連続して検索を続けることが出来る.

4 現状と今後の展望

以上、ゲノム塩基配列ビューワー giv の機能について概説してきた。シンプルだが簡便に利用できる実用性の高いツールとして、我々自身研究に活用している. 以下、今後予定している改善点について幾つかを述べる.

現状で特定の配列の塩基位置を素早く特定するのに役立てられているが、必要に応じて遺伝子領域の情報を表示することができれば有用であると考えられる。そのために、現行のバージョンには未実装であるが Genbank などの機能アノテーションファイルを読み込み、遺伝子領域および、そのアミノ酸翻訳配列を表示する仕組みを試作中である (Fig. 3).



① タブ (染色体切替え)
② 塩基配列・コード領域・翻訳アミノ酸配列表示
③ ページ・塩基位置・カーソル位置表示
④ カーソル位置の遺伝子情報表示

Fig. 3 Display of Coding Sequences
(to be implemented in the future version)

現時点で giv は編集機能を持たないビューワーである. 次世代シーケンサー解析ではリシーケンシングの結果による SNP (一塩基多型) や INDEL (挿入・欠失) に関する情報から、参照ゲノム配列を編集することもあるが、INDEL は当該箇所以降の塩基番号を変化させることから、手作業で直接編集するよりも、リシーケンシングデータから変更箇所をリストアップし一括して変換する方法を取った方が、間違いが起きにくく、手動による編集は実用上の必要性が比較的低いと考えて実装を後回しにしてきた. しかし、3.1.4 で述べた様な範囲指定の機能により様々な可能性が見えてきたことから、今後のアップデートで組み込んでいきたい機能の一つである.

塩基配列の解析でもう一つ考慮したい機能に、あいまい検索がある、シーケンサーデータにはエラーが含まれることが多く [7]、リードデータを参照配列上に探索する場合に数箇所のミスマッチを許容することが出来ると解析を効率的に行なうことができることが多い。

最新版のプログラムは MacOSX の実行形式として我々のウェブサイト(http://metalmine.mydns.jp/giv/)からダウンロード可能である。

最初にも述べたように、CUI によるインターフェースは使いこなすためにある程度の習熟が必要で、やや敷居が高いかもしれないが、ゲノム配列の解析が必要な研究では、カーソル操作と、g で先頭に戻るなどのいくつかの基本操作が指に馴染んで来ると快適に利用できる。giv は開発をはじめてまだ半年ほどであり改善の余地も未だ多い。プログラムの改善点・不備などについてご報告いただければ今後のアップデートで対応していきたい (knakamura@maebashi-it.ac.jp).

参考文献

1) R. D. Fleischmann *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd, Science, **269 (5223)**, 496-512 (1995).
2) R. Himmelreich, *et al.*, Complete Sequence Analysis of the Genome of the Bacterium *Mycoplasma pneumoniae*, Nucleic Acids Research, **24 (22)**, 4420-4449 (1996).
3) International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature, 409, 860-921(2001).
4) J. C. Venter et al. , The Sequence of the Human Genome, Science, **291(5507)**, 1304-1351 (2001).
5) H. Sakai *et al.*, The power of single molecule real-time sequencing technology in the *de novo* assembly of a eukaryotic genome, Scientific Reports, **5**, 16780, (2015).
6) H. Thorvalsdottir, J.T. Robinson, J. P. Mesirov, Integrative Genomics Viewer(IGV): high-performance genomics data visualization and exploration, 14(2), 178-192 (2013).
7) K. Nakamura *et al.*, Sequence specific error profile of Illumina sequencers, Nucleic Acids Research, **39**, e90 (2011).