次世代シーケンサによる比較ゲノムに向けた可視化手法の開発

理学専攻 情報科学コース　伊澤 亜紀子

1　はじめに

DNA を読む技術が急速に発展し，新型のシーケンサ (以下 NGS) が生まれたことで，多様な生物のゲノム配列が解読されている．ゲノム配列が分かることで，今までゲノム既知の種限定で行われてきた遺伝子の発現量を得る RNA-seq[1] や遺伝子発現の制御を調べる ChIP-seq[2] などが，多様な種で可能になった．実験結果をこれらの種を超えて比較することで，進化や生物学的なメカニズムの発見が期待されている．

ChIP-seq による種間の実験結果比較を考えよう．ChIP-seq は，遺伝子発現のスイッチである転写因子が結合した配列を調べることで遺伝子発現の制御をおこなう部位を発見する実験である．NGS はこの配列調査に用いられ，読まれた配列断片 (リード) とゲノムの対応を取る (アラインメントと呼ぶ．図 1(A) 参照) ことで遺伝子発現の制御に関わる領域を調べることが可能である．実際には実験結果のゆらぎや転写因子の弱い結合を考慮するため，図 1(B) に表した様なヒストグラムを考え，ピークの位置から転写因子の結合部位を判断する．図 2 は，種 A, B の二種での実験を示している．異なる種から読まれた ChIP-seq のリードは，その実験をおこなった種のゲノムに対応付けられる．2 種のヒストグラムを比較すると，種 A のヒストグラムのピークが種 B に比べ，1 つ少ないことがわかる．このように種間比較から，発現の制御の相違がわかる．

しかし，この解析をおこなうには 2 つの問題点が挙げられる．1 つは，NGS は数千万本の配列を読むことができ，現在公開されているソフトウエアでは，この大量の配列を扱い，比較することは困難だということ．そして，複数種を同時に可視化し，NGS の実験結果を種間で比較することは，現在公開されているソフトウエアでは容易ではないということである．

このような問題点に対し本手法では， 複数種で読まれた NGS のリードのアラインメント結果を同時に表示し，比較可能にするためのツール群 REad COordinate Transformer(RECOT) を開発した．また，RECOT は NGS の計算結果で最も一般的な形式である SAM 形式を出力する．これにより，ユーザが慣れた可視化ソフトを選ぶことが可能である．

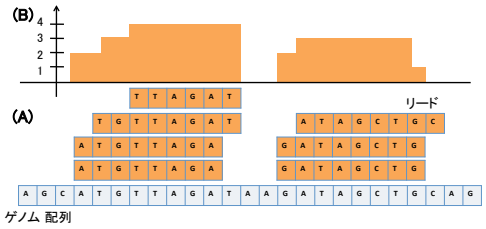

図 1: NGS で読まれたリードのアラインメント例

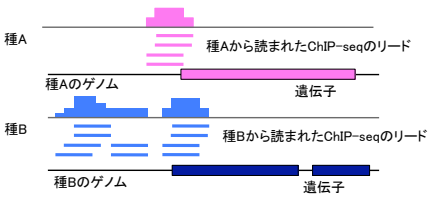

図 2: NGS による種間の遺伝子発現制御の比較

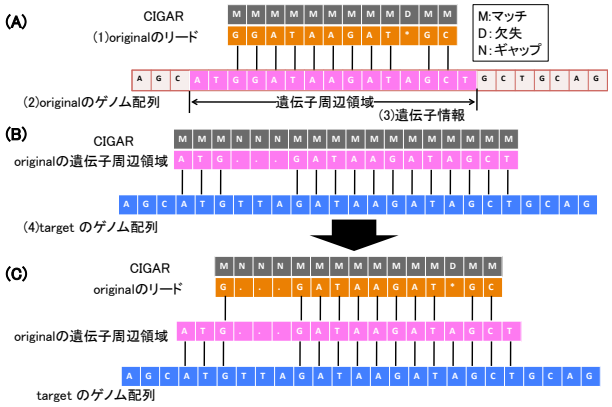

図 3: 提案手法の概観

2　提案手法

本研究では，リードが読まれた種を original species (以下 original) と呼び，リード位置を変換したい種を target species (以下 target) と呼ぶ．手法の概観を図 3 に示す．RECOT は 4 つのファイルを入力とする．(1)NGS で読まれた original のリード (2)original のゲノム配列 (3) 遺伝子情報 (4)target のゲノム配列である．RECOT はこれらの入力を基に，(1) と (2) の対応関係，(2) と (4) の対応関係を求め，(1) のリードを (4) のゲノムに対応付けるツールである．これにより，(4) のゲノム上で (1) のリードを見ることができる．

本手法では，2 つのアラインメント結果が必要になる．1 つは original の遺伝子配列と制御領域を含む周辺配列を target のゲノム配列にアラインメントした結果．もう 1 つは，リードと original のゲノム配列をアラインメントした結果である．まず，original の遺伝子配列と制御領域を含む周辺配列を target のゲノム配列にアラインメントする．遺伝子周辺領域の配列は提供されないことも多いため，ゲノム配列と遺伝子情報から，遺伝子周辺領域の配列を得るスクリプトを作成した．遺伝子周辺領域とゲノムのアラインメントには， SAM 形式で結果を出力する GMAP[3] などが利用可能である．図 3(B) は，ゲノムに遺伝子周辺領域をアラインメントした例である．2 本の配列をつなぐ黒い線は対応する塩基を表す．SAM 形式は塩基の対応を CIGAR という形式で表す．この例では遺伝子の CIGAR は 3M3N12M となる．M はマッチ・ミスマッチ (以下マッチ) であり，マッチとは塩基が一致する部分，ミスマッチは塩基が一致しない部分を指す．また D は欠失を，N はアラインメント時に挿入されたギャップを指す．これにより，1 から 3 番目の塩基はマッチ，4 から 6 番目はギャップ，7 から 18 番目は塩

基がマッチしていることがわかる．

しかし，相同性のある遺伝子が存在するため，1 つまたは複数の target の領域に複数の遺伝子がアラインメントされてしまうかもしれない．これは，1 つの遺伝子に複数の遺伝子が対応づけられることを意味し，比較する上で混乱を招く．このような問題を避けるため，同じ領域に複数の遺伝子がアラインメントされた場合，その複数の遺伝子から 1 つを選択するスクリプトを作成した．このスクリプトでは以下の 2 つの選択方法が用意されている．(1) ユーザ指定の遺伝子の対応表により，遺伝子に優先順位をつける．(2)1 番アラインメントスコアが高い遺伝子を優先する．これにより，original の遺伝子と制御領域を含む周辺配列を，target のゲノムにアラインメントした結果が得られる．

次に，リードと original のゲノムをアラインメントしたもう 1 つの結果を得る．リードと original のゲノムのアラインメントは，SAM 形式で出力可能なショートリード用のアラインメントツール，BWA[4] などが利用可能である．図 3(A) はその結果の例である．図の例は original のゲノムの，ある遺伝子周辺領域にリードをアラインメントした結果を示している．以上のように，2 つのアラインメントした結果を得る事ができる．

最後に 2 つのアラインメント結果から，リードを target のゲノムにアラインメントしていく．このアラインメントにより，original のゲノムにアラインメントされた結果を，target のゲノム上で解析することができるようになる．アラインメントには，(1) target のゲノムにアラインメントされた original の遺伝子周辺領域と，その遺伝子周辺領域にアラインメントされたリードの対応関係を表す表を作成，(2) (1) の対応表を使用し，リードの CIGAR を書き換え，の二段階のステップが必要である．図 3(A) は，original の遺伝子周辺領域とリードの，図 3(B) は original の遺伝子周辺領域と target のゲノムのアラインメント結果であり，図 3(A)(B) の遺伝子は同一のものである．(A)(B) の遺伝子が同一であるということは，ステップ (1) の対応表を調べる事でわかる．次に CIGAR の書き換えをおこなう (ステップ (2))．まず，リードの CIGAR を書き換えるために，遺伝子周辺領域でのリードのアラインメント開始位置を求める．図 3(A) の場合，リードは遺伝子周辺領域の 3 塩基目からアラインメントされている．また，図 3(B) では遺伝子周辺領域は，ゲノムの 4 塩基目からアラインメントされている．よって，リードはゲノムの 6 塩基目からアラインメントされればよく，遺伝子周辺領域の 3 塩基目からリードの CIGAR を変換しながらアラインメントすれば良い．図 3(C) はリードを変換し，ゲノムにアラインメントしたものである．遺伝子周辺領域はゲノムにアラインメントされる際，配列の 3 から 5 塩基目にギャップがある．そのため，リードの CIGAR にもギャップが挿入され，1M3N8M1D2M へと変換される．

3　実行結果

NGS の実データに本手法を適用し，有用性を検証するため，*D. simulans*(ERR020078) とキイロショウジョウバエ (*D. melanogaster*(ERR020066)) から読まれた ChIP-seq の比較をおこなった．これら 2 種はおよそ 500 万年前に分岐した近縁種である．結果の可視化にはゲノムビューア Integrative Genomics Viewer(IGV)[5] を使用した．図 4(A) は，*D. simulans* のリードを，BWA で *D. simulans* のゲノムにアラインメント後，original を *D. simulans*，target を *D. melanogaster* として，RECOT を実行した結果である．図 4(B) は，GMAP によって *D. melanogaster* の遺伝子周辺領域を D. melanogaster のゲノムにアラインメントした結果である．図 4(C) は，*D. melanogaster* のリードを，BWA で *D. melanogaster* のゲノムにアラインメントした結果である．図 4(D) は *D. melanogaster* のゲノム配列，図 4(E) はその遺伝子である．

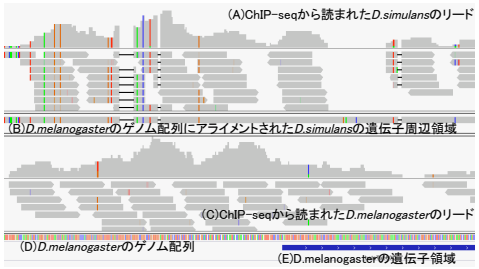

図 4: 実データによる実行結果

二種のヒストグラムを比較すると，発現制御部位の変化を読み取ることができる．このように本手法は種間を比較するために有用である．

4　まとめ

本研究では，NGS によって読まれた配列をアラインメントし，種間で比較を可能にする RECOT を開発した．RECOT により種間の変化を既存のゲノムビューアを使用し，可視化することができる．本手法の有用性を，ショウジョウバエの近縁 2 種の ChIP-seq データを利用し，発現の制御部位の変化を見ることで示した．

謝辞

本研究実施にあたり，ご指導頂きました東京工業大学大学院情報理工学研究科 瀬々潤准教授に深く感謝致します．

参考文献

[1] Wang Z. *et al.* RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009; **10**:pp. 57-63.

[2] Jothi *et al.* Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucl Acids Res.*2008; **36**:pp. 5221-5231.

[3] Wu TD, Watanabe CK. GMAP:a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005; **21**:pp. 1859-1875.

[4] Li H, Durbin R.Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; **25**:pp. 1754-1760.

[5] Robinson *et al.* Integrative Genomics Viewer. *Nature Biotechnology*. 2011; **29**:pp. 24-26.