たかがアセンブリ？―次世代シーケンス解析はじめの一歩―

末永　光*・小山　芳典

超並列DNA配列決定装置（いわゆる次世代シーケンサー）の加速度的な進展の恩恵を受け，研究者の誰もが手軽にゲノム情報を得ることが可能となった．そして研究プロジェクト全体に占めるコストと研究開発の重点は，配列決定そのものよりも，その後のデータ解析に移行しつつある[1]．

次世代シーケンスデータ解析は大きく次の三つの工程からなる．（1）読まれた配列データ（リード）のアセンブリ，（2）タンパク質をコードしている領域（ORF）の同定や機能の推定（アノテーション），（3）遺伝子発現ネットワーク解析や比較ゲノム解析などの高次解析．アセンブリは，その後のゲノム解析の精度を決定づける重要なはじめの一歩である．にもかかわらず，一見意味のある解析データを次々に生み出す工程の（2），（3）に比べると，いくぶん地味な作業で，それゆえ「たかがアセンブリ」と軽視されているのではないだろうか．そこで本稿においては，普段あまり陽の目を見ないアセンブリ工程に焦点をあててみたい．なお，アセンブリには，未知のゲノム配列を再構築する*de novo*と，既知のゲノム配列にリードを重ね合わせるマッピングの二つに分けられるが，本稿でいうアセンブリとは，*de novo*アセンブリに特化した記述とする．

アセンブリは，しばしば，びりびりに破れた本を復元する作業に例えられる．紙片は細かく膨大である（100万～1億以上のリード）．また，本には誤植がつきものである（読み取りのエラー）．しかも内容は一角獣の生態である（これまで読んだことがないということ：*de novo*）．感覚的に*de novo*アセンブリ工程の困難さが想像いただけたと思う．この復元作業の担い手がアセンブラーと呼ばれるプログラムであり，復元された文字（DNA）配列をContigと呼ぶ．ところで，このアセンブラーにはNewblerやVelvet以外にもさまざまなものが存在していることはご存じであろうか？ Wikipediaの“Sequence assembly”の項を確認すると，現在我々は40以上ものアセンブラーを利用できるらしい[2]．シーケンサーの機種やライセンスなどがとりあえずの選択基準となろう．しかしPCRを行う際にポリメラーゼが重要なファクターになるのと同様に，使用するアセンブラーによって得られる配列結果がまったく異なるのである．

Roche社の454システムから得られたtrascriptomeの配列データ（741,387本のリード，約2億塩基）を6種のプログラム（CAP3，CLC，MIRA，Newbler ver. 2.3，Newbler ver. 2.5，SeqMan）でアセンブリし，その結果を比較した論文が報告された[3]．結果をいかに評価するか実のところは難しい課題であるが，「より長いContig，既知リファレンスゲノム配列との良好なアライメント」などの指標を用いた場合，「総合優勝はNewbler 2.5で，その他はまあ同程度に優秀，ただしNewbler 2.3はイケテナイ」という結論だ．ただし，ここで重要なのは順位ではなく，「異なるアセンブラーを用いると異なる配列が生まれる」という（当たり前の）結果である．産出された全Contigの長さの指標（N50で表す）はもちろんばらばらであるが，Contigの数も12,000から36,000，合計塩基数は14.5 Mbから21.4 Mbと壮大なばらつきぶりを示した．これは，重複されたContigや不完全長なContigが含まれている可能性も示唆している．ただし繰り返しになるが，未知配列を復元する（*de novo*）のであるから，各々のContigの信頼性を評価することは難しい．逆にスタートの時点でこれだけ異なると，ゴールではそれぞれどんな考察が出来上がるのかむしろ比べてみたい興味に駆られる．いずれにせよ，取り扱う配列データと目的によって最適なアセンブラーは異なるということを認識し，研究者はその都度ベストなアセンブラーを選択していくべきである．

さらに同論文では，Contigの信頼性の確立という悩ましい課題への対策についても示されている．「異なるアルゴリズムをもつアセンブラーから共通のContigができたら，それはすごく信頼性が高いのではないか」という発想のもとで，“二次アセンブリ”を推奨している．つまり，（i）まず異なる2種のプログラムでアセンブリを行う．（ii）次に得られた2グループのContigどうしのアセンブリを，第3のプログラムを用いて行う．（iii）こうして得られた“Robust contig”はより長く良質である．

最近，Miniaというメモリ使用量が極端に少ないアセンブラーが，フランスの学生によって開発された[4]．我々にとっては，Minia自体よりも，これで使用されたアルゴリズムが他のプログラムにも応用され，一人ひとりの研究者が，普通のデスクトップパソコンで利用できる時代がもうすぐきそうなことに意味があると思う．

次世代シーケンス解析がもっとも力を発揮する研究対象のひとつは，環境試料（メタゲノム）だと思う．しかし，リファレンスとなる微生物の種類も数もまったく不明というかなりやっかいな相手である．そういった五里霧中の状況の中で，上記を参考にアセンブリ戦略を立て，可能な限り良質のContigを準備して解析のスタート地点に立つことが，現在我々ができる最善の策ではないだろうか．

1) Sboner, A. *et al.*: *Genome Biol.*, **12**, 125 (2011).
2) http://en.wikipedia.org/wiki/Sequence_assembly
3) Kumar, S. and Blaxter, M. L.: *BMC Genomics*, **11**, 571 (2010).
4) http://minia.genouest.org

***著者紹介**　産業技術総合研究所生物プロセス研究部門（主任研究員）　E-mail: suenaga-hikaru@aist.go.jp