生物情報実験法2019 岩崎先生

05-195512 生物情報科学科3年
山口尚人

2019年8月14日

第1章　課題2 微生物ゲノムの解析

1.1　目的

広大な人類未踏の領域である微生物ゲノムを自由な発想で解析することで、その解析手法を学ぶと共に、新規の発見を目指す。

1.2　解析内容

極限環境微生物のゲノムを比較し、dN/dS 解析をおこなった。その微生物に特徴的な極限環境で重要な役割を果たす遺伝子は正の自然選択を受けているという仮説を元に、その遺伝子を配列から発見することを目指し、2 種類の微生物ゲノムを比較し、dS/dN 解析をおこなった。松井先生にお話を伺う中で、強い放射線耐性を持つ微生物に興味を持った。現在までに見つかっている強い放射線耐性を持つ微生物の多くは、放射線によるゲノムの損傷を予防するのではなく、損傷したゲノムを修復する能力が高いということを知り、この遺伝子の機能やその応用可能性を魅力的に感じた。特に、よく解析されている Deinococcus Radiodurans をはじめとする、Deinococcus 属ではその属に特有の、DNA の結びつき二本鎖切断の修復効率をあげるとされる PprA タンパク質の立体構造が決定されている [2]。

高い放射線耐性を持つ微生物は、Deinococcus Radiodurans が有名であるが、それ以外にも多く存在しており、Deinococcus 属に属する微生物の他、Ruburobacter Radiotolerans、Kineococcus radiotolerans、Halobacterium salinarum NRC-1 などがあげられる。

今回の解析の主な目的は、極限環境微生物の中でも特に、高放射線耐性微生物のゲノムに着目することで、高い放射線耐性をどのように獲得してきたかを発見すること、もしくは、ゲノムの比較によって、放射線耐性に寄与している遺伝子を見つけることができるのではないかという仮説を検証することである。

1.3　具体的な手順

dN/dS 解析の方法、手順に関しては [1] を参考にした。

1. 2 種の微生物ゲノムを選択した

本解析では、2 つのゲノムのアノテーションされたゲノム領域を用いた。これらの微生物のコドン表は 11 番であるが、例外の含まれていたため、各遺伝子の塩基配列だけでなく、アノテーションされた翻訳後のアミノ酸配列も必要であったことから、両者とも genomic の Genbank ファイルを用いた。

2. 2 つのゲノムから、オーソログを抽出

   [1] で紹介されているように、複数の方法が考えられる。代表的なものは、BLAST を RBB(Reciprocal Best BLAST hit) method である。それ以外には、OMA (Orthologous MAtrix) や OrthoMCL といったオーソログ推定データベースがあげられる。また、BLAST スコアに関連する遺伝子長の偏りを考慮に入れることによってオルソログの精度を高めた OrthoFinder というソフトウェアも考慮に入れた。しかし、今回は、インストールの手間や、API ドキュメントの簡潔さの観点から、OMA を用いることにした。OMA は、REST API を提供しており、ドキュメントも非常にわかりやすく、適度にアップデートされており、すぐに使える上、それなりに信頼できると考え、選択した。

   比較したい 2 つのゲノムの taxonomy id を取得し、それらをクエリとして GET リクエストを送ることで、2 つのゲノムのオーソログのタンパク質情報を得ることができる。それらには各ゲノムでの開始地点、終了地点、また UniProt などでのユニークな ID などが含まれる。

3. 各オーソログのアミノ酸配列のアライメントをおこなった

   次のステップで後述するように、dN/dS 解析にあたってコドンベースのアライメントが必要であり、その前段階としてアミノ酸配列のアライメントが必要である。そこで、アライメントツールとして Clustal OMega を用いた。コドンベースのアライメントをオプションと入力の変更によっておこなえる PRANK も考慮に入れたが、Clustal OMega が使い易かったこと、後述の PAL2NAL と組み合わせることでも目的を達成できそうであったため、前者を選択した。

   1 つ前のステップで得られたオーソログのアミノ酸配列を入力として、アライメントをおこなった。

4. 各オーソログに対してコドンベースのアライメントをおこなった

   次のステップは各オーソログの塩基配列をアライメントすることであるが、当然オーソログ同士の配列長は異なることが多い。また、今回の解析の目的は、非同義置換、同義置換の比を求めることであるから、各アミノ酸に対応するコドンの関係がアライメントによってずれてしまう（フレームシフト）と、正しい dN/dS 値を算出することができない。そこで、dN/dS 解析にあたっては、コドンベースでの塩基配列のアライメントをおこなう必要がある。これはタンパク質のアミノ酸配列同士をアライメントし、その後それに対応

するコドンを元の塩基配列を参照しながら塩基配列をアライメントする方法である。これによってコドンの対応が崩れることなく、非同義置換、同義置換の解析をおこなうことができる。

そこで、今回は、このコドンベースのアライメントをおこなうソフトウェアである、PAL2NAL を用いた。これは perl スクリプトであり、Web 上からローカルに download することで、jupyter notebook からコマンドライン実行で実行できたため、利用した。PAL2NAL は 2019 年現在、2011 年から更新されていないようであった。このコドンベースのアライメントはさほど複雑ではないと考えられるので自分でプログラムを書くこともできたとは思うが、時間の関係上、このソフトウェアを信頼し利用した。

PAL2NAL に対する入力は、OMA の REST API から得られたオーソログのタンパク質のアミノ酸配列と、その各ゲノム上での位置を利用して Genbank のゲノムデータ上から取得した、対応する遺伝子領域の塩基配列である。(これを 2 セット用意した)

5. アライメントされた塩基配列を入力とし、dN/dS の値を算出した

   dN/dS 値の算出では、PAML(Phylogenetic Analysis by Maximum Likelihood) に含まれる CodeML を用いた。アライメントされた塩基配列と比較するゲノムの系統関係を元に、非同義置換、同義置換の数、率、それらの比率を算出するプログラムである。CodeML は、枝モデル (branch models), サイトモデル (site models), 枝サイトモデル (branch-site models) の 3 種類モデルに基づいた解析が可能である。

   枝モデルは，遺伝子系統樹のある枝で正の自然選択が働いたかどうかを検定する解析モデルであり [3]、今回のような、2 種類のみのゲノムのペアワイズの比較には不適当であると考え、サイトモデルは、枝間の ω を変化させず，サイト間での ω の変異を検定するモデルである [3]。

1.4 結果

以下に、Deinococcus Radiodurans R1 と Thermus Thermophilus HB の 85 個のオーソログ遺伝子について、dN/dS の値を算出し、プロットした結果を示した。

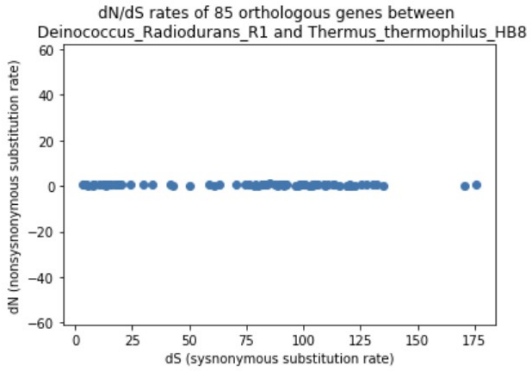

図 1.1: Deinococcus Radiodurans と Thermus Thermophilus の 85 個のオーソログ遺伝子の dN/dS を 1:1 スケールでプロットしたもの

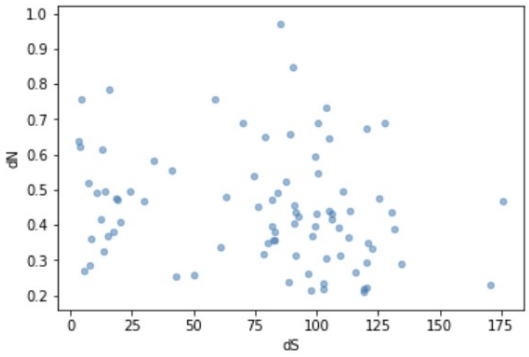

図 1.2: Deinococcus Radiodurans と Thermus Thermophilus の 85 個のオーソログ遺伝子の dN/dS のプロットを拡大したもの

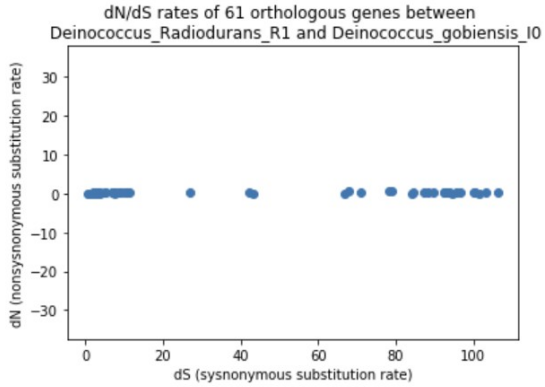

図 1.3: Deinococcus Radiodurans と Deinococcus Gobiensis の 61 個のオーソログ遺伝子の dN/dS を 1:1 スケールでプロットしたもの

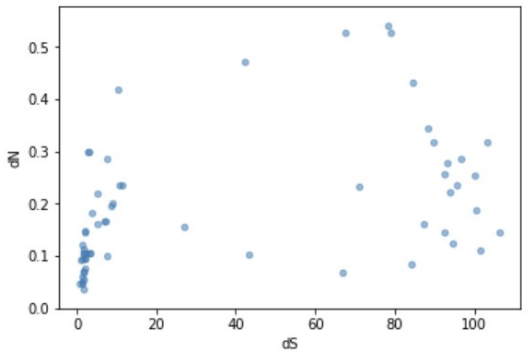

図 1.4: Deinococcus Radiodurans と Deinococcus Gobiensis の 61 個のオーソログ遺伝子の dN/dS のプロットを拡大したもの

また、同様のプログラムを用いて、Deinococcus Radiodurans R1 と Deinococcus Gobiensis I0 の個のオーソログ遺伝子についても、dN/dS の値を算出し、プロットした結果を示した。

圧縮して提出したプログラムファイル、入出力ファイルは、GitHub にもアップロードされている。[https://github.com/Naoto-Yamaguchi/3s-iwasaki-kadai2]

1.5 考察

結果の図のどちらを見てもわかるように、各オーソログ遺伝子の dN/dS の値は、0.01 よりも小さいものがほとんどであり、dN/dS は 1 以下であるばかりか、同義置換率が非同義置換率に比べて非常に大きい値を取っているということがわかった。また、同義置換率が桁違いに大きいものが多数ある。これにはいくつかの理由が考えられる。

1. オーソログの抽出によるもの

   OMA データベースから抽出されてオーソログは、種間での保存度合いが高いもののみに偏っていた可能性が考えられる。

2. 比較する 2 種類のゲノムが dN/dS 解析に適していない

   今回、Deinococcus Radiodurans にまず着目し、その種と進化的に近く、かつ放射線耐性がそこまで高くない Thermus Thermophilus との比較をおこなった。また、その結果からより近い種での解析も試そうと考え、Deinococcus Gobiensis との比較もおこなった。しかし、この比較対象の検討が十分であったとは言えず、また目的を達成するために相応しいものではなかった可能性もあり、再検討が必要である。

3. 正しい解析がおこなえていなかった

   CodeML のモデル選択が適切でなかった

1.6 利用したデータ

- Deinococcus Radiodurans R1

  RefSeq データ

- Thermus Thermophilus HB

  RefSeq データ

- Deinococcus Gobiensis I0

  アノテーションされた GenBank ファイル

1.7 参考文献

[1] Daniel C.Jeffares et all.(2015)「A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome」

[2] M.Adachi (2019)「Extended Structure of Pleiotropic DNA Repair-Promoting Protein PprA from Deinococcus radiodurans」
[3] Ziheng Yang(2017)「PAML Manual User Guide」