



Title	光合成遺伝子の同定
Author(s)	横野, 牧生; 田中, 亮一; 田中, 歩
Description	6章 バイオインフォマティクス 3
Relation	光合成研究法. 北海道大学低温科学研究所, 日本光合成研究会共編
Citation	低温科学, 67, 669-672
Issue Date	2009-03-31
Doc URL	https://hdl.handle.net/2115/39209
Type	departmental bulletin paper
File Information	67-093.pdf



3. 光合成遺伝子の同定

横野 牧生¹⁾, 田中 亮一¹⁾, 田中 歩¹⁾

全ゲノム配列情報が利用できる生物種数は急速に増加している。次世代シーケンサーの利用などにより、この数は今後ますます増加することが予想される。本章では、この情報を利用した遺伝子の同定法について紹介する。

Identification of genes by using whole genome sequence

Makio Yokono, Ryouich Tanaka, Ayumi Tanaka

Recently, the number of available whole genome sequences is rapidly increasing. In order to identify the genes responsible for photosynthesis, we developed a bioinformatics tool, CCCT (Correlation Coefficient Calculation Tool), which calculates the correlation coefficient between the distributions of a certain phenotype and genes among a group of organisms. By using this method, we identified cyanobacterial DVR gene responsible for chlorophyll synthesis. This method can also be applied to the identification of non-photosynthetic genes.

3.1 はじめに

近年シーケンサーの高速化に伴い数多くの生物種の全ゲノムの配列情報が公開されている。一方でそこに含まれる遺伝子の機能についての情報（アノテーション）は不足しており、機能が不明とされているものが数多く存在する。遺伝子の機能を決定するために、関連する変異体を作成することやコードされているタンパク質を単離することが行われる。しかしこれらの作業には多くの時間がかかるため、どの遺伝子が目的とする表現型の発現に関与しているのかを予測することが必要とされる。

機能が不明な遺伝子の機能を予測するためによく用いられるのは相同性検索であり、BLAST や FASTA が有名である[1, 2]。これを使うと、機能が既知のある遺伝子に類似した別の遺伝子を探すことができる。しかしながら単に機能が既知の遺伝子と類似のものを検索するだけでは、それらと相同性が無く機能が未知の遺伝子に対応できない。

未知の遺伝子の機能を推定する上で、その遺伝子がどのような生物種に存在しているかを知ることは大きな助けになる。言い換えると、ある生物群が特定の表現型を示している場合、その原因となる遺伝子（群）もそれらの生物全てに共通して存在する可能性が高い。本章では、この考えに基づいて筆者らが開発した CCCT (Correlation Coefficient Calculation Tool) による遺伝子の同定

を紹介する。この手法は多くの生物種の全ゲノム情報を利用できることが前提となっているため、今後さらに精度は向上すると予想できる。

3.2 解析対象にする生物群の選定

まず始めに未知の遺伝子を持っていることが確実な生物種をクエリー種として選定する。これは例えば、ある色素を蓄積するのにその合成に関与する遺伝子が未知であるような種である。次にクエリー種に近縁で、未知の遺伝子を持っていると予想される種と、持っていないと予想される種を解析対象に選ぶ。

この解析を行うためには準備段階として、解析対象にする生物群の持つ全遺伝子を総当たりで相同性検索を行うことが必要となる。筆者らはすでに NCBI において公開されている真核生物を含む約 850 種の生物の全遺伝子に対して総当たりで Blast 検索を完了しており、結果のファイルを統合データベースプロジェクトの協力を得て公開する予定である。しかしながら、非公開になっている生物種を用いる場合や、Blast 以外の相同性検索の結果を用いて CCCT を用いる場合のために、解析に用いる相同性一覧の表のフォーマットを示す[表 1]。一行目は解析対象にする生物の名前であり、左端がクエリー種で、順次ターゲット種を任意の順で並べる。二行目はそれぞれの生物が、目的とする未知の遺伝子を持つはず(−180)なのか、持たないはず(1)なのかを示す。ここでは検索方法に Blast を採用し相同性の指標としては E-value の指

1) 北海道大学低温科学研究所

表 1: 相同性一覧の例

<i>A. thaliana</i>	<i>A. thaliana</i>	<i>C. reinhardtii</i>	<i>E. coli</i> E22	...
	-180	1	-180	...
NP_974945.2	-180	-15.04575749	0.146128036	...
NP_974899.2	-168	-19.15490196	0.491361694	...
NP_974897.2	-17.52287875	0.698970004	0.146128036	...
NP_974807.2	-172	-65	0.681241237	...
...	...	...	...	...

数部分を用いているので、値が小さいほど類似した遺伝子を持つことを表している。一列目の三行目以降はクエリー種（この場合は *Arabidopsis thaliana*）の全遺伝子それぞれのアクセッション番号を示す。二列目以降の三行目以降は、各列の生物種の全遺伝子をターゲットにして、クエリー種の各遺伝子の相同性を検索したときに最も相同性が高かった遺伝子の相同性の値である。例えば二列目の三行目は *A. thaliana* の遺伝子 NP_974945.2 をクエリーにして *A. thaliana* の全遺伝子を検索したときにベストヒットした遺伝子の相同性の値である。これは採用した相同性検索の方法で、各遺伝子と全く同じ遺伝子がヒットしたときに表示される相同性の値となる。また、三列目の三行目は *A. thaliana* の遺伝子 NP_974945.2 をクエリーにして *Chlamydomonas reinhardtii* の全遺伝子を検索したときにベストヒットした遺伝子の相同性の値である。

このように相同性の値を並べると、各行は各遺伝子の生物間の分布パターンを示すことになる。例えば三列目の NP_974945.2 は *C. reinhardtii* と *E. coli* には良く似ているものは無く、六行目の NP_974807.2 は *C. reinhardtii* には少し似ているものがあるが *E. coli* には似ていないものが無いことがわかる。ここで理想とするパターン（二行目）と各行との相関をとれば、クエリーの持つ全遺伝子のうちどれが理想とする分布パターンに近いかがわかる。また各行のパターンをそれぞれ各行のパターンに対して相関をとれば、クエリーの持つ全遺伝子とその生物間分布パターンに基づいて自動的に分類することができる。筆者らは Mathematica を用いて Stand-alone Blast の制御、相同性一覧表の作成、および相関係数の算出を行ったが、統合データベースのホームページ上で Firefox にて動作する CCCT のスクリプトを公開予定である。

この解析で最も重要なのは、解析対象の生物種の選定と、理想とする分布パターンの設定である。未知の遺伝子を持つクエリー種に比較的近縁な種で、その遺伝子を持たないと予想される生物種が存在する場合や、クエ

リー種と遠い種でその遺伝子を持つと予想される種がある場合には、候補を絞り込みやすくなる。ここでは筆者らが特定に成功した光合成関連遺伝子を例にとり、その手順を示す。

3.3 クロロフィル合成に関する遺伝子

クロロフィルはその合成過程においてジビニル体からモノビニル体へ変換される。高等植物においてはその反応に関与する遺伝子 (dvr) がすでに同定されているが [3], 他の多くの光合成生物にはその酵素が無くどのような変換しているのか不明であった。筆者らは CCCT を用いてシアノバクテリアの持つ dvr を特定することができた。

3.4 解析対象の生物種の選定

ほとんどのシアノバクテリアと紅藻は高等植物の持つ dvr を持たないにも関わらずモノビニル体のクロロフィルを光合成に用いている [表 2]。このことは新規の dvr があることを示している。モノビニル体のクロロフィルを持つ生物のうち高等植物と類似の dvr を持たない cyanobacteria および紅藻はすべて未知の dvr を持つと仮定した。これらの生物のうち *Synechocystis* sp. PCC 6803 は遺伝子操作が可能であり比較的良くアノテーションされていることから、これをクエリー種とし他の生物はターゲット種に加えた。一方でシアノバクテリアでも *Prochlorococcus* はジビニル体のクロロフィルを持ち、高等植物と類似の dvr も持たないことから、未知の dvr も持たないといえる。これらも全てターゲット種に加えた。高等植物、緑藻、珪藻は既知の dvr を持つため解析対象には加えなかった。

3.5 総当たりの相同性検索

Synechocystis sp. PCC 6803 の遺伝子は約 3600 個あ

表 2：解析対象の生物群と各生物の持つクロロフィルのタイプ

クロロフィルはモノビニル 高等植物の dvr は無い 未知の dvr を持つはず	<i>Synechocystis</i> sp. PCC 6803 (クエリー種として使用) <i>Anabaena variabilis</i> ATCC 29413 <i>Cyanothece</i> sp. CCY0110 <i>Lyngbya</i> sp. PCC 8106 <i>Nostoc punctiforme</i> PCC 73102 <i>Synechococcus elongatus</i> PCC 6301 <i>Synechococcus</i> sp. JA-2-3B'a (2-13) <i>Synechococcus</i> sp. RCC307 <i>Synechococcus</i> sp. RS9917 <i>Synechococcus</i> sp. WH 7803 <i>Thermosynechococcus elongatus</i> BP-1 <i>Cyanidioschyzon merolae</i>	<i>Crocospaera watsonii</i> WH 8501 <i>Gloeobacter violaceus</i> PCC 7421 <i>Nodularia spumigena</i> CCY9414 <i>Nostoc</i> sp. PCC 7120 <i>Synechococcus elongatus</i> PCC 7942 <i>Synechococcus</i> sp. JA-3-3Ab <i>Synechococcus</i> sp. RS9916 <i>Synechococcus</i> sp. WH 5701 <i>Synechococcus</i> sp. WH 7805 <i>Trichodesmium erythraeum</i> IMS101
クロロフィルはジビニル 高等植物の dvr は無い 未知の dvr も無いはず	<i>Prochlorococcus marinus</i> str. AS9601 <i>Prochlorococcus marinus</i> str. MIT 9301 <i>Prochlorococcus marinus</i> str. MIT 9312 <i>Prochlorococcus marinus</i> str. MIT 9515 <i>Prochlorococcus marinus</i> str. NATL2A <i>Prochlorococcus marinus</i> CCMP1986	<i>Prochlorococcus marinus</i> str. MIT 9211 <i>Prochlorococcus marinus</i> str. MIT 9303 <i>Prochlorococcus marinus</i> str. MIT 9313 <i>Prochlorococcus marinus</i> str. NATL1A <i>Prochlorococcus marinus</i> CCMP1375
クロロフィルはモノビニル 高等植物の dvr を持つ 未知の dvr を持つかは不明	<i>Synechococcus</i> sp. BL107 <i>Synechococcus</i> sp. CC9605 <i>Synechococcus</i> sp. WH 8102 <i>Osterococcus tauri</i>	<i>Synechococcus</i> sp. CC9311 <i>Synechococcus</i> sp. CC9902 <i>Arabidopsis thaliana</i> <i>Thalassiosira pseudonana</i>

る。これらをひとつずつクエリー配列として、各ターゲット種の全遺伝子を対象に Blast サーチを行い、最も類似しているターゲット種の遺伝子を抜き出して表 1 のような類似度 (期待値：E-value) の行列を作成した。ただしクエリー種は *Synechocystis* sp. PCC 6803 とし、ターゲット種は表 2 に赤および青で示した生物種を用いた。

3.6 相関係数による並べ替え

表 2 に示した条件をグラフに表した物が図 1 a) である。縦軸は E-value を対数表示にしたもので小さいほど相関性が高い遺伝子を持つことを意味する。このパターンとクエリー種の持つ全遺伝子それぞれの分布パターンとの相関をとり、相関の高い順にグラフとして表示した [図 1 b~d]。最も相関が高かったのはシアノバクテリアのアンテナ (フィコビリソーム) をチラコイド膜に接続するリンカータンパク質であった。ほとんどのシアノバクテリアと紅藻はフィコビリソームをアンテナとして用いるが、*Prochlorococcus* はフィコビリソームを持たないことから、これは予期された結果である。次に相関係数が高かったのはアノテーションのない slr1923 であり、これは未知の dvr である可能性が高いと考えられた。他に相関が 90% 以上の範囲でアノテーションがない遺伝

子は slr1302 だけであった。この 2 つをクエリーにして NCBI の protein database に対して Blast 検索を行ったところ、slr1923 は *Methanococcus* の coenzyme F420 hydrogenase / dehydrogenase family protein と弱い相関性のあることがわかった。最後に *Synechocystis* sp. PCC 6803 で slr1923 を破壊した変異体がジビニル型のクロロフィルを蓄積したことから、slr1923 がシアノバクテリアでモノビニル型のクロロフィルの合成に関与していることが確認できた。

3.7 まとめ

本解析方法は単純な原理に基づいており、始めの相関性検索さえ終わらせてしまえば、相関による並び替えは数分で完了する。候補を絞り込めるかどうかは解析に用いる生物種の種類と数にかかっており、十分な生物種を用いることができれば候補を相当絞り込むことができる。実際シアノバクテリアの dvr では候補が 2 つに絞り込まれた。

一方で課題もいくつか存在する。ひとつは、目的の遺伝子の理想とする分布パターンを設定するには解析に用いる生物種の特徴をよく知っている必要がある点である。これについては生物の特徴を集めたデータベースの

ような物があればそれと連動させることである程度解決が可能である。もうひとつは、複数のタンパク質が共同で何かの反応に関わっている場合、それら全てが同じような生物間分布を示すとは限らないことである。この場合 CCCT では、異なる生物間で保存性が高いと思われる鍵となるタンパク質を見つけることはできるが、補う形で働くタンパク質すべてを検出することはできない可能性がある。補うタンパク質の分布パターンを複数想定し組み合わせて解くことは可能ではあるが、計算量がかなりのものになる。例えば鍵となるタンパク質は共通だが、補うタンパク質は生物によって2種類の組み合わせがある場合、クエリ一種を2つ組み合わせることになるため、それぞれが3000の遺伝子を持っていた場合 3000×3000 通りの比較をすることになり、およそ丸一日計算にかかる。現時点では見つかった鍵となるタンパク質を手がかりに分子生物学的手法で相互作用するタンパク質を探す

ことがより容易かもしれない。なお、今回紹介したデータおよびツールは非営利目的に限り希望者に提供が可能である。

引用文献

- 1) W. R. Pearson, D. J. Lipman, Improved Tools for Biological Sequence Comparison, *Proceedings of the National Academy of Sciences* **85** (1988) 2444-2448.
- 2) S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool (BLAST), *J Mol Biol* **215** (1990) 403-410.
- 3) N. Nagata, R. Tanaka, S. Satoh, A. Tanaka, Identification of a vinyl reductase gene for chlorophyll synthesis in *Arabidopsis thaliana* and implications for the evolution of *Prochlorococcus* species, *Plant Cell* **17** (2005) 233-240.