NAIST-IS-MT0551004

修士論文

遺伝子配列の相同性に基づく
ドメイン探索アルゴリズムの開発

新井 美紗子

2007年 2月 1日

奈良先端科学技術大学院大学
情報科学研究科 情報生命科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (理学) 授与の要件として提出した修士論文である。

新井 美紗子

審査委員：

金谷 重彦 教授　(主指導教員)
植村 俊亮 教授　(副指導教員)
黒川 顕 助教授　(副指導教員)

遺伝子配列の相同性に基づく ドメイン探索アルゴリズムの開発*

新井 美紗子

内容梗概

遺伝子には, 他の遺伝子と相同性を示す 1 つまたは複数の高く保存されている配列領域であるドメインが存在する. 個々のドメインは機能をもっており, 遺伝子がタンパク質として働く際に深く関係している. そのため, ドメインは遺伝子アノテーションや遺伝子の進化, タンパク質立体構造の予測など多様な解析で役立つ. 本研究では, 遺伝子の配列相同性からドメインの探索を行うアルゴリズムの開発を行った. 本研究で開発したドメイン探索のアルゴリズムは, 始めに, 遺伝子について異・同生物間で比較解析を行い相同な遺伝子であるオーソログ遺伝子・パラログ遺伝子の検出を行なった. 次に, オーソログ遺伝子・パラログ遺伝子の相同配列データを基に, ドメインを探索した. また, ドメインの探索と探索結果の可視化を行うソフトウェアの開発も行った. 開発したソフトウェアを使い, 333 種の生物における全遺伝子について, ドメインの探索を行った. 植物 (シロイヌナズナ) と微生物 (真菌, 古細菌, 細菌: 332 種) 間のオーソログ遺伝子, 植物と植物間のパラログ遺伝子からそれぞれドメインの探索を行い, さらに, 主要な機能をもつ遺伝子についてドメインの解析を行った. 解析結果から, 微生物から植物への進化の過程で受け継がれたドメインを明らかにすることができた.

キーワード

ドメイン, オーソログ遺伝子, パラログ遺伝子, シロイヌナズナ, 微生物

* 奈良先端科学技術大学院大学 情報科学研究科 情報生命科学専攻 修士論文, NAIST-IS-MT0551004, 2007 年 2 月 1 日.

Detection of common domains in homologous gene sequences*

Misako Arai

Abstract

Genes are composed of domains. Domains are homologous regions in different genes and are conserved evolutionary units that often also correspond to functional units. Domains represent one of the most useful levels at which to understand protein function, and gene evolution. In this study, we developed an algorithm for detecting common domains among homologous gene sequences. First, homologous gene sequences are detected by comparing the genomes of two species. If two species are different, homologous genes are orthologous genes. If two species are the same, homologous genes are paralogous genes. Second, domains are detected among orthologous or paralogous gene sequences. In this work, we developed software for detecting and visualizing domains. We detected the domains of all genes in 333 species by this software. We analyzed these domains by comparing plant (*Arabidopsis thaliana*) and microorganism genomes, and by classifying plant domains on the basis of gene functions. These results report the identification of domains which derived from microorganism in genes of plant.

Keywords:

Domain, Orthologous gene, Paralogous gene, *Arabidopsis thaliana*, Microorganism

* Master's Thesis, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT0551004, February 1, 2007.

目 次

図 目 次

表 目 次

1. 序論

現在, 多様な生物種のゲノム配列が決定している. ゲノムの決定により, その生物に存在する全てのタンパク質, 遺伝子, ドメインなどの解析が可能となった. そのため, 多様な生物種間で, 遺伝子配列を比較解析することにより, 異生物種間の遺伝子において相同性をもつオーソログ遺伝子や同生物種間の遺伝子において相同性をもつパラログ遺伝子を検出することが可能となっている. また, 局所的に他の遺伝子と相同性を示す領域であるドメインをもつ遺伝子が存在することが知られている [1].

ドメインは生物が進化をしていく過程で, 生命活動に必要な機能を保持し, より複雑な機能を得るために変化しながら, 異・同生物種間で伝播していったと考えられる. ドメインは大きく分け, 1 つの遺伝子において, ドメインが 1 つのみ存在するシングルドメインと 2 つ以上存在するマルチドメインがある. 遺伝子がマルチドメインを形成していることは, 1 つの遺伝子において複数の機能が存在している可能性が考えられる. 実際に, 遺伝子は, bi-function や multi-function といわれるもののように, 1 つの遺伝子において複数の機能が存在しているものがあることが報告されている [2]-[5]. 進化の過程においても, ある生物種のゲノムでは 2 つの異なる遺伝子にコードされているタンパク質 (ドメイン) が, 別の生物種のゲノムでは, 融合した 1 つの遺伝子としてみいだされることがある. このようなタンパク質はロゼッタストーンタンパク質 [6], [7] といわれ, その 2 つの遺伝子産物は機能的に関連していることが多い. そのため, ドメインは遺伝子のアノテーションやタンパク質間相互作用の解析などに用いられている. また, ドメインはタンパク質の立体構造にも重要な役割を果たしている. AFU(Autonomously Folding Units) といわれるタンパク質の立体構造の情報を保存しているドメインが遺伝子配列にあるといわれている. AFU を探索する方法として, PASS [8] や DOMAINATION [9] などがある.

このように, ドメインは, 生命活動の解明に必要とされている遺伝子やタンパク質関連の研究と大変深い関係がある. ドメインを異・同生物種のゲノム間で比較解析することは, 生物種間において高く保存されているドメインを同定することができ, 機能が既存のものと比較することで遺伝子の機能予測や, 比較した生

物種における保存度の違いからドメインがどの生物に由来するのかを予測できると考えられる. また, 遺伝子におけるドメインのコードの違いから, ドメインの融合・解離の様子をみることができると考えられる. そのため, ドメインについてみていくことは, 今後のバイオインフォマティクス研究のためにも大変有用であると考えられる.

これまで, ドメインの研究については, Pfam [10] や SCOP [11] などドメインのデータベースの構築や InterPRO [12] や SMART [13] など既存のドメインのデータベースから, 配列パターンを照合し, ドメインを探索することが行われてきた. しかし, InterPRO や SMART では, 異・同生物間でどのようにドメインが保存されているのか, また未知な配列パターンのドメインについては検出することができない. そのため本研究では, 遺伝子配列の相同性に基づくドメイン探索のアルゴリズムの開発を試みた. ドメインは, 遺伝子配列において, 他の遺伝子と相同性がある領域である. そのため, ある遺伝子について, ある生物種のゲノムからオーソログ遺伝子またはパラログ遺伝子を検出し, ドメインを探索する. オーソログ遺伝子からドメインを探索した場合, 解析対象とした生物種間で保存されているドメインが検出できる. また, パラログ遺伝子からドメインを探索した場合, その生物種固有のドメインが検出できる. 特に, 対象生物種を微生物と植物のように進化のレベルが異なるものにすることで, 進化の過程において保存されているドメインを検出できる. また, 既存のドメインの配列パターンとの照合からでは検出できなかった未知のドメインについても検出できると考えられる.

本研究で開発したアルゴリズムによりドメインの探索を行い, その探索結果及び, 探索の過程を可視化するソフトウェアの開発も行った. 本ソフトウェアはある遺伝子において, どの遺伝子が相同であるか, ドメインがどのように保存されているかなど, 遺伝子とドメインの情報を統合的に解析できるように開発を試みた.

本研究では, 植物に注目したドメイン解析を行った. 高等植物であるシロイヌナズナ (*Arabidopsis thaliana*) は, 微生物に比べ多様で複雑な機能をもっている. そのため, 遺伝子も微生物より複雑な構造をとる. 本ソフトウェアを用い, 植物 (シロイヌナズナ) と微生物 (真菌, 古細菌, 細菌) 間, 植物と植物間でドメインの探索を行うことで, ドメインを保存している生物種の違いから, 探索されたドメインが

微生物由来か植物固有かを明らかにするため, 植物の遺伝子に注目し解析を行った. また, 植物の遺伝子機能ごとに, 遺伝子におけるドメインの構造を解析することで, 遺伝子機能の違いにおけるドメインの解析・検討を行った.

2. 材料及び方法

2.1 ゲノムデータ

本研究では, 植物 1 種 (*Arabidopsis thaliana*), 真菌 2 種 (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*), 古細菌 27 種 (ナノ古細菌門 1 種, クレン古細菌門 5 種, ユーリ古細菌門 21 種), 細菌 303 種 (フソバクテリウム門 1 種, プラクトミセス門 1 種, アクイフェックス門 1 種, テルモトガ門 1 種, 緑色非硫黄細菌門 2 種, 緑色硫黄細菌門 3 種, デイノコックス・テルムス門 3 種, バクテロイデス門 5 種, スピロヘータ門 6 種, クラミディア門 11 種, シアノバクテリア門 17 種, 放線菌門 21 種, 発酵菌門 75 種, プロテオバクテリア門 156 種) の計 333 種の生物のゲノムデータを用いた. これら 333 種の生物の内訳とゲノムの遺伝子数, accession number を付録 A に表で示す. ゲノム配列データは全て NCBI の RefSeq(http://www.ncbi.nlm.nih.gov/) から入手した.

2.2 オーソログ遺伝子とパラログ遺伝子

異なる生物の遺伝子間で, 相同である 2 つの遺伝子をオーソログ遺伝子と定義する. 一方, 同じ生物の遺伝子間で相同である 2 つの遺伝子をパラログ遺伝子と定義する. オーソログ遺伝子, パラログ遺伝子は, 通常, 遺伝子のアミノ酸配列の類似性に基づいて対応づけられる. 本研究では, この配列類似性の評価を BLAST(Altschul et al., 1990, 1997 [14], [15]) により行った. BLAST プログラムの BLASTP 解析によりオーソログ遺伝子の検出には E-value は 10^{-5} 以下のものを, パラログ遺伝子の検出には, E-value は 10^{-30} 以下のものを類似性のある配列とした.

一般的には, オーソログ遺伝子・パラログ遺伝子は, BLAST による配列類似性の評価が双方向に最も高い 1 組の遺伝子のことをいう. しかし, 本研究では, BLAST で配列類似性の評価を行った 2 つの遺伝子について, どちらの遺伝子からも BLAST の結果で得た類似配列領域が等しく, かつ十分な類似度を示す遺伝子対は相同であると考え, この 2 つの遺伝子を異生物間においてはオーソログ遺伝子, 同生物間においてはパラログ遺伝子として検出した. これにより, 2 つの遺伝子間での正確な類似配列領域を検出でき, 2 つの生物間での配列パターンによる保存性の違いを調べることができる.

オーソログ遺伝子・パラログ遺伝子の検出方法の例を, 生物 S_1 の遺伝子 $gene_{1i}$ ($i = 1, \ldots, N_1$, N_1: 生物 S_1 に存在する全遺伝子数) のうち, $gene_{11}$ を問い合せ遺伝子 (query gene), 生物 S_2 における全遺伝子 $gene_{2j}$ ($j = 1, \ldots, N_2$, N_2: 生物 S_2 に存在する全遺伝子数) を対象 (subject gene) として以下に述べる (図 1 参照).

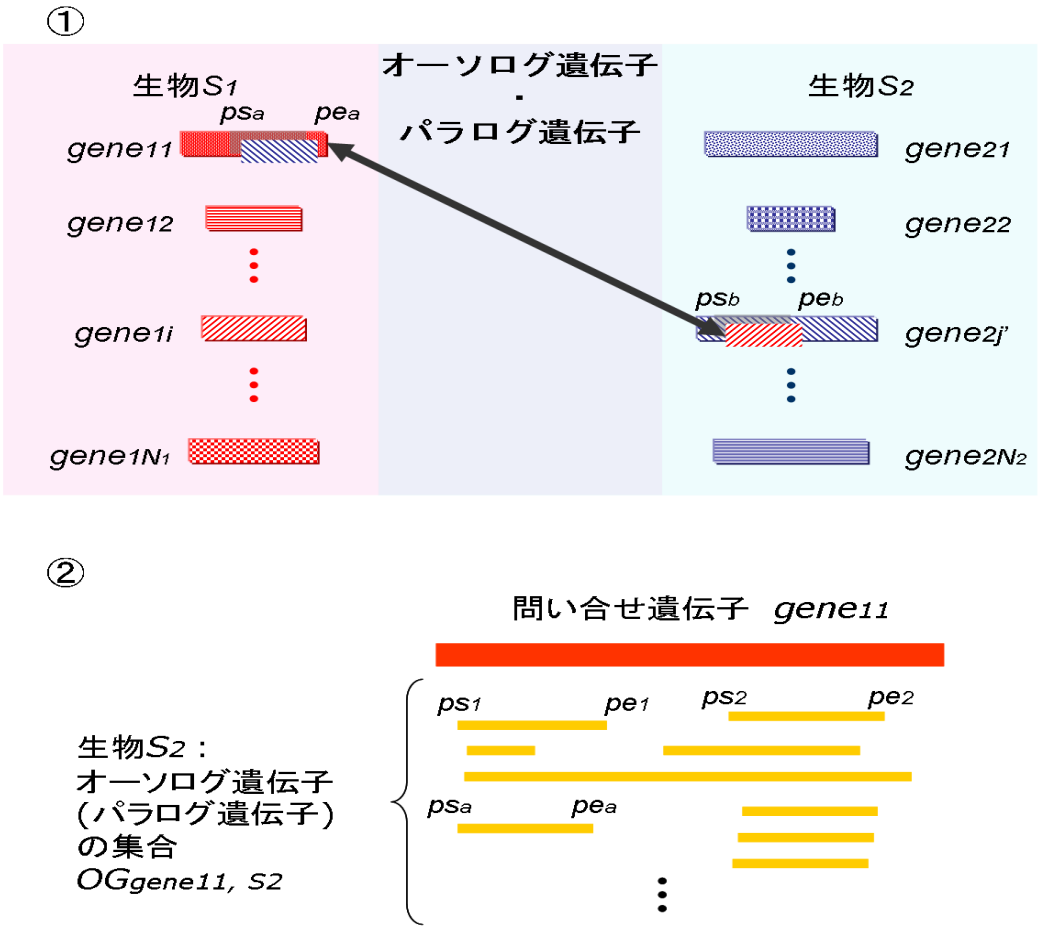

図 1 オーソログ遺伝子・パラログ遺伝子の検出方法

異生物間 $S_1 \neq S_2$ の場合: オーソログ遺伝子 ; 同生物間 $S_1 = S_2$ の場合: パラログ遺伝子

1: bi-directional BLAST の結果から HSP を求め, さらに類似領域も双方に等しいものをオーソログ遺伝子・パラログ遺伝子とする.

2: $gene_{11}$ と S_2 におけるオーソログ遺伝子・パラログ遺伝子の配列の集合

2.2.1 BLAST による配列類似性の評価

1. 生物 S_1 のある遺伝子 $gene_{11}$ を問い合せ遺伝子とする．$gene_{11}$ と生物 S_2 における全ての遺伝子の中から閾値以上の類似度がある配列対 (High Scoring Pair) の集合 (式 (1)) を探索する．

   $$HSP_{gene_{11},gene_{2j'}} = \{(gene_{11}(ps_a, pe_a), gene_{2j'}(ps_b, pe_b))\} \quad (1)$$

   $HSP_{querygene,subjectgene}$

   $gene_{2j'}$: 生物 S_2の全遺伝子 $gene_{2j}$のうち $gene_{11}$と HSP となる遺伝子

   $gene_{11}(ps_a, pe_a)$: $gene_{11}$における開始残基 ps_aから終了残基 pe_aまでの配列領域

   $a = 0, \ldots, NR_{11}$

   NR_{11} : 式 (1) によりみつかった $gene_{11}$での類似配列領域の数

   $gene_{2j'}(ps_b, pe_b)$: $gene_{2j'}$における開始残基 ps_bから終了残基 pe_bまでの配列領域

   $b = 0, \ldots, NR_{2j'}$

   $NR_{2j'}$: 式 (1) によりみつかった $gene_{2j'}$での類似配列領域の数

2. 1. で $gene_{11}$ と HSP となった生物 S_2 における遺伝子 $gene_{2j'}$ を問い合せ遺伝子とする．$gene_{2j'}$ と生物 S_1 の遺伝子 $gene_{11}$ から閾値以上の類似度がある配列対の集合 (式 (2)) を探索する．

   $$HSP_{gene_{2j'},gene_{11}} = \{(gene_{2j'}(ps_{b'}, pe_{b'}), gene_{11}(ps_{a'}, pe_{a'}))\} \quad (2)$$

   $HSP_{querygene,subjectgene}$

   $gene_{2j'}(ps_{b'}, pe_{b'})$: $gene_{2j'}$における開始残基 $ps_{b'}$から終了残基 $pe_{b'}$までの配列領域

   $b' = 0, \ldots, NR_{2j'}$

   $NR_{2j'}$: 式 (2) によりみつかった $gene_{2j}$での類似配列領域の数

   $gene_{11}(ps_{a'}, pe_{a'})$: $gene_{11}$における開始残基 $ps_{a'}$から終了残基 $pe_{a'}$までの配列領域

   $a' = 0, \ldots, NR_{11}$

   NR_{11} : 式 (2) によりみつかった $gene_{11}$での類似配列領域の数

2.2.2 遺伝子相同性の評価

2.2.1 項で求めた $HSP_{gene_{11},gene_{2j'}}$ と $HSP_{gene_{2j'},gene_{11}}$ ($gene_{2j'}$: 生物 S_2 の全遺伝子 $gene_{2j}$ のうち $gene_{11}$ と HSP となる遺伝子) について, それぞれの集合に存在する配列を比較する. 双方向の BLAST 探索において, 互いに十分な配列類似性を示し, かつ式 (3) のように類似領域の範囲が双方向に等しい配列を 1 つ以上もつ場合, 2 つの遺伝子は相同である.

$$\begin{aligned} gene_{11}(ps_a, pe_a) &= gene_{11}(ps_{a'}, pe_{a'}) \\ &\text{かつ} \\ gene_{2j'}(ps_b, pe_b) &= gene_{2j'}(ps_{b'}, pe_{b'}) \end{aligned} \tag{3}$$

このような遺伝子をオーソログ遺伝子 ($S_1 \neq S_2$), またはパラログ遺伝子 ($S_1 = S_2$) と定義する.

生物 S_1 の遺伝子 $gene_{1i}$ ($i = 1, \ldots, N_1$, N_1: 生物 S_1 に存在する全遺伝子数) を問い合せ遺伝子 (query gene), 生物 S_2 の全遺伝子を対象 (subject gene) とし, みつかったオーソログ遺伝子またはパラログ遺伝子の配列の集合を OG_{gene_{1i},S_2} ($OG_{querygene,subjectspecies}$) とする.

同様に, 生物 S_2 の遺伝子 $gene_{2j}$ ($j = 1, \ldots, N_2$, N_2: 生物 S_2 に存在する全遺伝子数) を問い合せ遺伝子, 生物 S_1 の全遺伝子を対象とし, みつかったオーソログ遺伝子またはパラログ遺伝子の配列の集合を OG_{gene_{2j},S_1} とする.

2.3 ドメイン

ある遺伝子において, 他の遺伝子と相同性が保存されている領域部分をドメインと定義する. ドメインは, ある遺伝子において 1 つ以上存在する場合がある. ドメインが 1 つだけ存在する場合をシングルドメインと呼び, 2 つ以上存在する場合をマルチドメインと呼ぶ. 本研究では, このドメインに注目した解析を行うために, 2.2 節で求めたオーソログ遺伝子・パラログ遺伝子の配列の集合を用いてドメインの探索を行った. 本研究のドメインの探索方法は, オーソログ遺伝子とパラログ遺伝子とも同じである. そのため, 生物 S_1 の遺伝子 $gene_{11}$ を問い合せ遺伝子として, 生物 S_2 の全遺伝子を対象として得たオーソログ遺伝子の配列の集合 OG_{gene_{11},S_2} を用いて, ドメインを探索する例を以下に述べる.

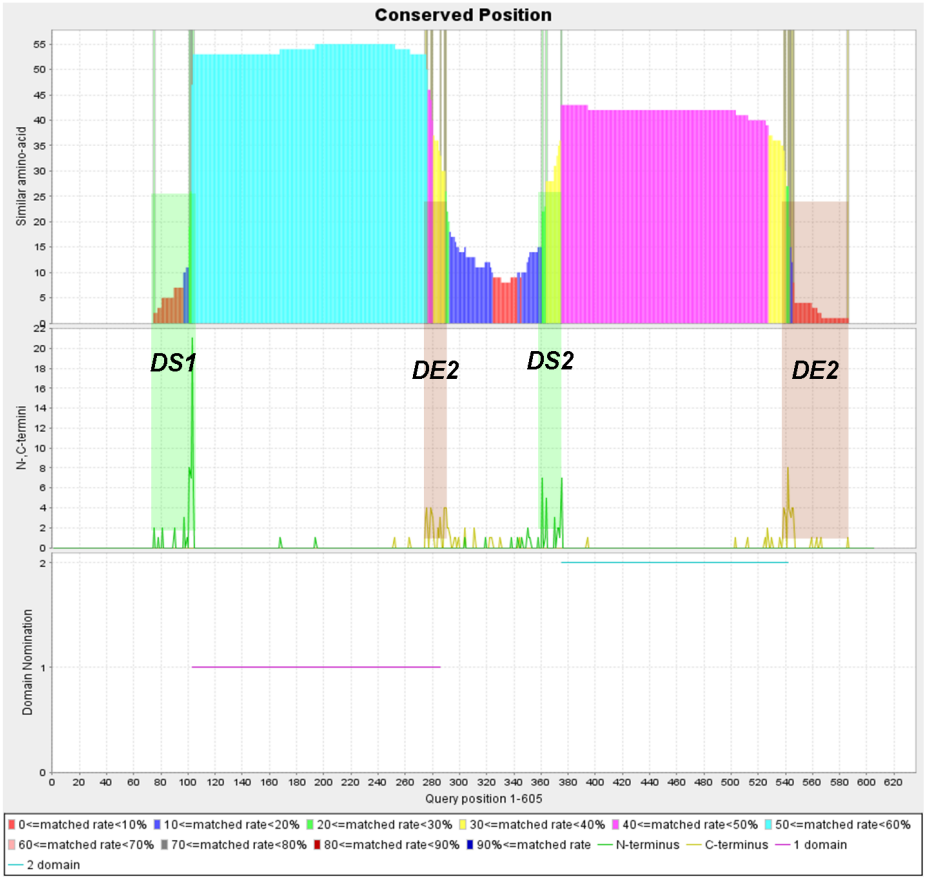

図 2　オーソログ遺伝子の配列集合からのドメイン探索

(a): 問い合せ遺伝子 (query gene) の各残基におけるオーソログ遺伝子の集合の全配列の保存度とドメインの開始・終了残基位置の候補

(b): オーソログ遺伝子の全配列の開始・終了残基位置の出現回数

(c): ドメイン領域

2.3.1 ドメインと非ドメイン領域の境界

ドメインと非ドメイン領域の境界を定めるため, 始めにドメインの開始残基位置と終了残基位置の候補をオーソログ遺伝子の配列の集合 OG_{gene_{11},S_2} における各配列の開始・終了残基位置によって決める (図 2(a), (b) 参照).

1. OG_{gene_{11},S_2} に存在する全ての配列領域 $gene_{11}(ps_a, pe_a)$ (式 (1) に同じ) の開始残基位置 ps_a の出現回数 $Cps(ps_a)$ を求める.

2. 開始残基位置の出現回数 $Cps(ps_a)$ が, 式 (4) 以上のものをドメインの開始残基位置の候補 dps_{k_s} $(k_s = 1, \ldots, dns,\ dns$: ドメインの開始残基位置の候補数) とする. dps_{k_s} は, 問い合せ遺伝子 $gene_{11}$ の N-末端に近いものから $dps_1, dps_2, \ldots, dps_{dns}$ と残基順に並べる.

3. 1. から 2. の手順と同様に終了残基位置 pe_a の出現回数 $Cpe(pe_a)$ を求め, 式 (4) 以上のものをドメインの終了残基位置の候補 dpe_{k_e} $(k_e = 1, \ldots, dne,$ dne: ドメインの終了残基位置の候補数) とする. dpe_{k_e} は, 問い合せ遺伝子 $gene_{11}$ の N-末端に近いものから $dpe_1, dpe_2, \ldots, dpe_{dne}$ と残基順に並べる.

$$
\begin{aligned}
AvgC &= \frac{\sum_{l=1}^{q} C(l)}{ON} \qquad (4)\\
C(l) &: Cps(ps_a) \text{ または } Cpe(pe_a)\\
q &: \text{問い合せ遺伝子 } gene_{11} \text{の長さ}\\
ON &: \text{オーソログ遺伝子の配列の集合 } OG_{gene_{11},S_2} \text{の全配列数}
\end{aligned}
$$

次に, ドメインの開始・終了残基位置の候補から, ドメインごとの開始・終了残基位置を求める.

1. ドメインの開始残基位置の候補 dps_{k_s} が存在する範囲 DS とドメインの終了残基位置の候補 dpe_{k_e} が存在する範囲 DE が一部でも重複する場合, 問い合せ遺伝子 $gene_{11}$ は,
   $ND|DS_1|DS_x\&DE_y|DE_{dn}|ND$

(dn: 最終的に探索されたドメインの数) と領域が分割される. DS_1 には, 開始残基位置の候補のみが存在し, DE_{dn} には, 終了残基位置の候補のみが存在する.
DS と DE が重複しない場合は, 問い合せ遺伝子 $gene_{11}$ は,
$ND|DS_1|DE_1|ND$
と領域が分割され, 7. と 8. が行われ, ドメインの探索が終了する.

2. 前のステップで得られた $gene_{11}$ の領域 $DS_x\&DE_y$ において, ドメインの終了残基位置の候補 dpe_{k_e} が存在する範囲 DE_y を求める. 領域 $DS_x\&DE_y$ から DE_y を除いた領域 DS_{dn} には, 開始残基位置の候補のみが存在する. よって, 問い合せ遺伝子 $gene_{11}$ は,
$ND|DS_1|DS_x\&DE_y|DS_{dn}|DE_{dn}|ND$
と領域が分割される.

3. 前のステップで得られた $gene_{11}$ の領域 $DS_x\&DE_y$ において, ドメインの開始残基位置の候補 dps_{k_s} が存在する範囲 DS_x を求める. 領域 $DS_x\&DE_y$ から DS_x を除いた領域 DE_1 と DE_{dn-1} には, 終了残基位置の候補のみが存在する. よって, 問い合せ遺伝子 $gene_{1i}$ は,
$ND|DS_1|DE_1|DS_x\&DE_y|DE_{dn-1}|DS_{dn}|DE_{dn}|ND$
と領域が分割される.

4. 前のステップで得られた $gene_{11}$ の領域 $DS_x\&DE_y$ において, ドメインの終了残基位置の候補 dpe_{k_e} が存在する範囲 DE_y を求める. 領域 $DS_x\&DE_y$ から DE_y を除いた領域 DS_2 と DS_{dn-1} には, 開始残基位置の候補のみが存在する. よって, 問い合せ遺伝子 $gene_{11}$ は,
$ND|DS_1|DE_1|DS_2|DS_x\&DE_y|DS_{dn-1}|DE_{dn-1}|DS_{dn}|DE_{dn}|ND$
と領域が分割される.

5. 3., 4. を問い合せ遺伝子の領域が分割されなくなるまで行う.

6. 最終的に, 問い合せ遺伝子 $gene_{11}$ は,
$ND|DS_1|DE_1|\ldots|DS_{dn}|DE_{dn}|ND$

と分割される.

7. それぞれのドメインの開始残基位置の候補範囲 DS_{id} において, 残基位置の出現回数 $Cps(dps_{k_s})$ が最大なものをドメイン $domain_{id}$ の開始残基位置 $dstart_{id}$ とする.

8. それぞれのドメインの終了残基位置の候補範囲 DE_{id} において, 残基位置の出現回数 $Cpe(dpe_{k_e})$ が最大なものをドメイン $domain_{id}$ の終了残基位置 $dend_{id}$ とする.

2.3.2 問い合せ遺伝子における相同遺伝子の類似配列保存性の評価

ドメイン領域は, 他の遺伝子領域よりもオーソログ遺伝子と相同性が高く保存されている. そのため, ドメインと非ドメイン領域を分けるために, 保存性の評価を行う.

本研究では, 始めに問い合せ遺伝子の各残基ごとに保存度を求める (図 2(a) 参照).

オーソログ遺伝子の配列の集合 OG_{gene_{11},S_2} に存在する全ての要素における, 各残基位置の出現回数 $Cp(m)$ $(m = ps_a, \ldots, pe_a)$ を求め, 式 (5) から, 問い合せ遺伝子 $gene_{11}$ の各残基の保存度 $Cons(n)$ $(n = 1, \ldots, q,\ q$: 問い合せ遺伝子 $gene_{11}$ の長さ) を求める.

$$
\begin{aligned}
Cons(n) &= 100 \times \frac{Cp(n)}{ON} \quad \% \qquad (5)\\
Cp(n) &: \text{残基位置 } n \text{ の出現回数 } (n = 1, \ldots, q)\\
ON &: \text{オーソログ遺伝子の配列の集合 } OG_{gene_{11},S_2} \text{の全配列数}
\end{aligned}
$$

2.3.3 ドメイン領域の決定

2.3.1 項で求めたドメイン $domain_{id}$ の開始残基位置 $dstart_{id}$ と終了残基位置 $dend_{id}$ 間の長さ $dlength_{id} = |dend_{id} - dstart_{id}| + 1$ が, 30 残基以上でかつ 2.3.2 項で求めた保存度から, ドメインとされる領域の残基位置ごとの保存度の平均 $AvgCons(id)$ (式 (6) 参照) が 25%以上のものをドメインと決定した (図 2(c)).

$$
\begin{aligned}
AvgCons(id) &= \frac{\sum_{m=dstart_{id}}^{dend_{id}} Cons(m)}{dlength_{id}} \quad \% \\
Cons(m) &: \text{残基位置 } m \text{ の保存度 (\%)} \\
dlength_{id} &= |dend_{id} - dstart_{id}| + 1
\end{aligned}
\tag{6}
$$

ドメインは, 30 残基以下の長さのものはほとんど存在しないことが報告されている (Jones S et al., 1998 [16]). また, 本研究で使用した全ゲノムデータにおいて遺伝子の長さを図 3 にヒストグラムで表す. 遺伝子数が 30 残基付近で急激に増加しており, 比較的短い長さのプラスミドなどの染色体外遺伝子を除去する意味でも, ドメインの長さを 30 残基以上とした.

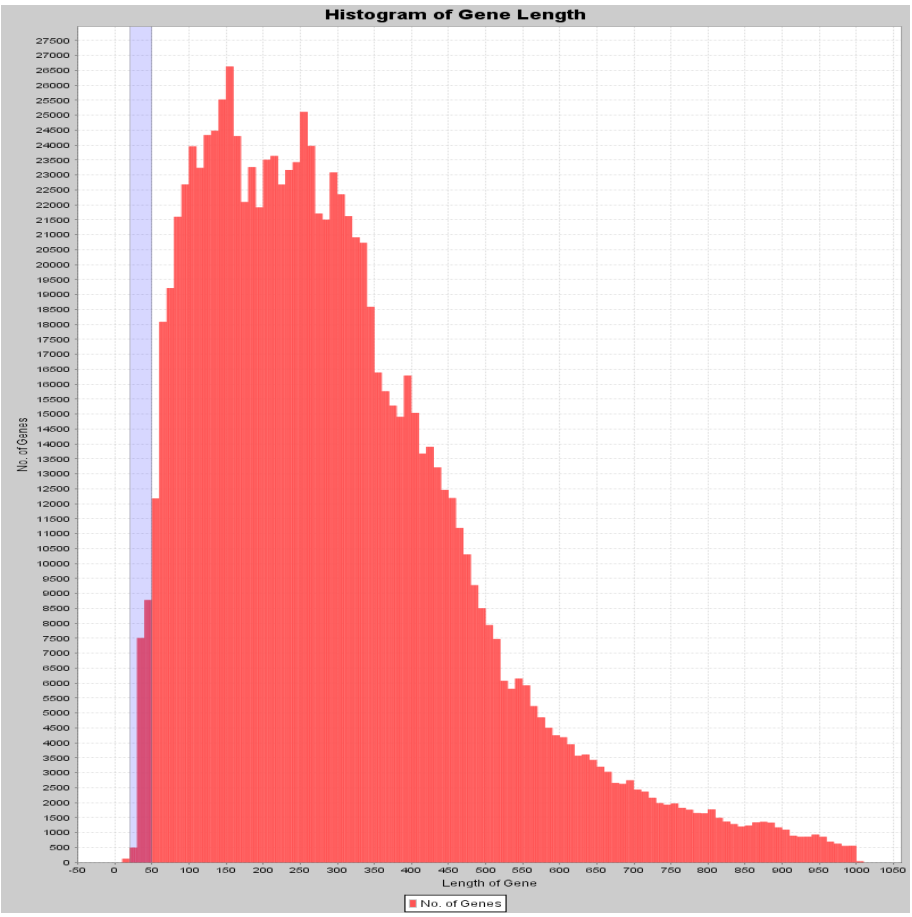

図 3 遺伝子の長さのヒストグラム (Length of Gene $\leq$ 1000)

遺伝子の長さが 1,000 残基以下のものをヒストグラムで表す.
横軸: 遺伝子の長さ (10 残基間隔で尺度をとる) ; 縦軸: 遺伝子の累積数

また, 配列保存性を考慮せずドメイン探索すると図 4 のように, ドメインの数が 5 個以上の遺伝子は極めて少ない. 1 つの遺伝子にドメインが 4 つ存在する場合, 1 つのドメインは最低オーソログ遺伝子の配列と $\frac{1}{4}$ の保存性を示す. そのため, 25%以上の保存度があるものをドメインとした.

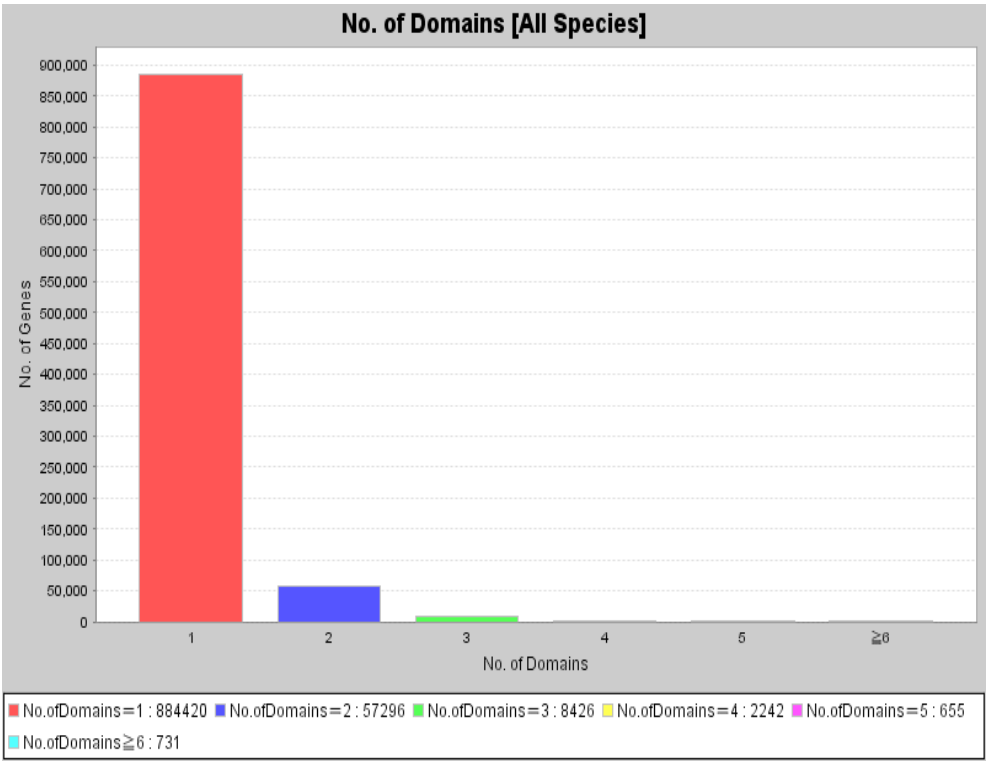

図 4　ドメイン数の割合 (保存度 0)

式 (5) $\geq 0\%$ においてのドメイン探索結果
横軸: 遺伝子におけるドメインの数 (6 個以上のドメインが存在する遺伝子については, カテゴリーをまとめた) ;
縦軸: 各ドメイン数における遺伝子累積数

2.4 ドメイン探索結果の可視化ソフトウェア

2.3 節の本研究で開発したアルゴリズムにより, ドメインの探索を行い, 探索の過程及び結果を可視化することで視覚的に遺伝子とドメインを統合的に理解するためのソフトウェア (図 5 参照) の開発を行った.

本ソフトウェアは, ある遺伝子におけるオーソログ遺伝子またはパラログ遺伝子の類似配列領域のデータを入力することで, ドメインの探索から可視化までを行う. 入力データのソフトウェアへのロードについては, ファイルかデータベースかを使用者が選択できる. ファイルの場合は使用者が独自に検出したオーソログ遺伝子・パラログ遺伝子のデータを規定の形式にそってファイルを作成し, ファイルを選ぶことでデータがロードされる. また, 2.2 節の手法により, 333 種全ての生物の遺伝子について, 333 種を対象生物として検出されたオーソログ遺伝子・パラログ遺伝子のデータを保管するデータベースを構築した. データベースの場合は, このデータベースから解析したい問い合せ遺伝子 (query gene) と対象生物 (subject speceis) を GUI の選択画面から選ぶことで必要なデータがロードされる. データが入力されると, ソフトウェアでドメインが探索され, 問い合せ遺伝子におけるオーソログ遺伝子・パラログ遺伝子の類似配列領域とドメインの探索結果が可視化される. また, 問い合せ遺伝子, オーソログ遺伝子・パラログ遺伝子の機能情報やアミノ酸配列によるゲノム情報も得られる.

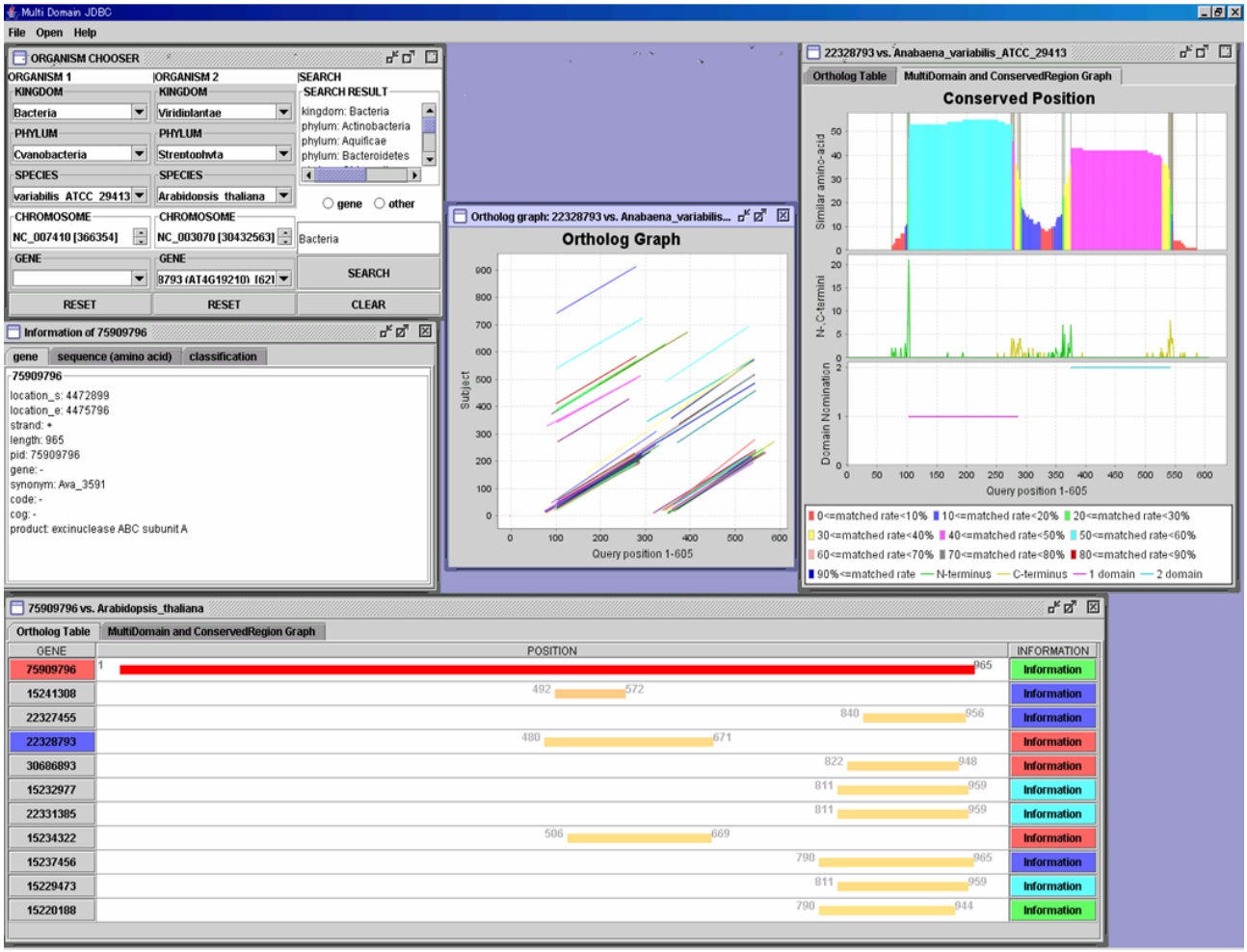

図 5　ドメイン探索ソフトウェア

3. 結果及び考察

本研究で使用した 333 種の生物における全ての遺伝子 (1,034,233 個) について, ドメインの探索を行った. ある遺伝子 $gene_{xx}$ におけるドメインは, 生物種ごとにドメインを探索したのち (図 6 参照), それらのドメインを 2.3 節でのオーソログ遺伝子の配列の集合と同様に考え, ドメインの集合を $OG_{gene_{xx},S_{all}}$ とし, 333 種全ての生物とのオーソログ遺伝子・パラログ遺伝子からのドメインとした. 各生物種ごとに求めたドメインは十分な保存性があると考え, $OG_{gene_{xx},S_{all}}$ からドメインを探索する際は, 保存度は 0%以上とした.

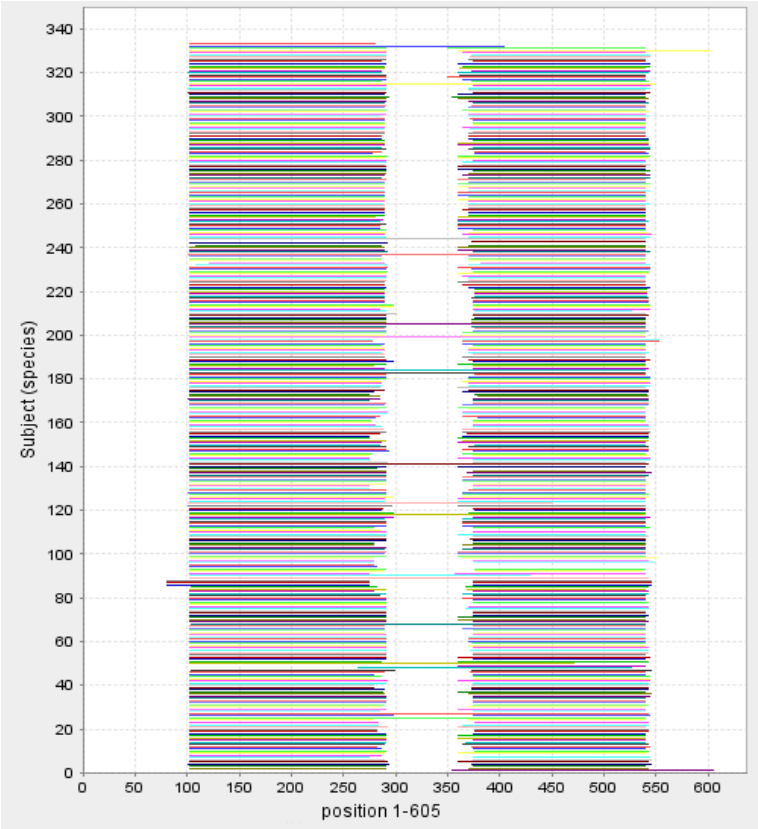

図 6 生物種ごとのドメイン探索結果

横軸: 問い合せ遺伝子の長さ ; 縦軸: ドメインが探索された生物

3.1 全生物種の遺伝子におけるドメインの統計

図 7 は, それぞれの遺伝子から探索されたドメインの数について遺伝子を分類した結果である. 全遺伝子の 92%において 1 つ以上のドメインが存在している. このことからもドメインは, 異・同生物種間において高く保存されていることがわかる. 特に, シロイヌナズナ (*Arabidopsis thaliana*) の遺伝子 (26,536 個) と他の 332 種の生物におけるオーソログ遺伝子とのドメインに注目し解析することで, 単細胞生物から多細胞生物への進化の過程をドメインによりみることができると考える. 3.3 節以降では, シロイヌナズナに注目した解析結果を述べる.

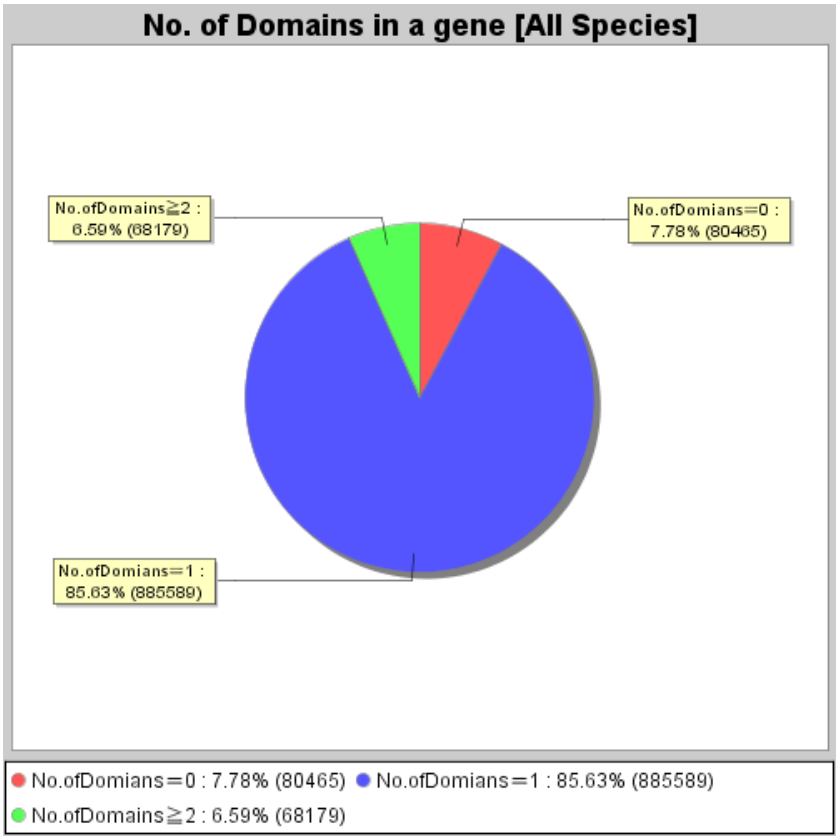

図 7 全遺伝子におけるドメイン数の割合

red: ドメイン数=0 ; blue: ドメイン数=1 ; green: ドメイン数 ≥ 2

3.2 植物 (シロイヌナズナ) と微生物 (シアノバクテリア) のドメインの統計

シアノバクテリアは, 植物と同じ酵素発生型の光合成を行う細菌である. シアノバクテリアの祖先は進化上はじめて酵素発生型光合成の能力を獲得した生物であり, その能力は細胞内共生によって藻類や植物へと受け継がれたと考えられている. したがって, シアノバクテリアと植物は, 光化学系や CO_2 代謝, エネルギー伝達系, シグナル伝達系などで多くの類似点や共通点をもっている. そのため, 植物のモデル生物であるシロイヌナズナとシアノバクテリアにおけるドメインを比較することは微生物から植物への進化の過程をみる糸口となると考えられる.

図 8 で, シロイヌナズナとシアノバクテリア門に属する 17 種の生物のマルチドメインとシングルドメインの割合を比較した. シロイヌナズナは, シアノバクテリアに比べ, マルチドメインとなる遺伝子数が増加している. また, ドメインを有する遺伝子全てのうちマルチドメインを形成する遺伝子の割合もシロイヌナズナの方がシアノバクテリアよりも高い. このことから微生物から植物への進化の過程で, 植物は 1 つの遺伝子に複数のドメインが入り込み, bi-function や multi-function のように複雑な機能を獲得したと考えられる.

そのため, 本研究では植物が保持している微生物由来のドメインを明らかにし, 微生物から植物への遺伝子の進化の過程や bi-function や multi-function の遺伝子構造の解明の新たな知見を得るため解析を行う.

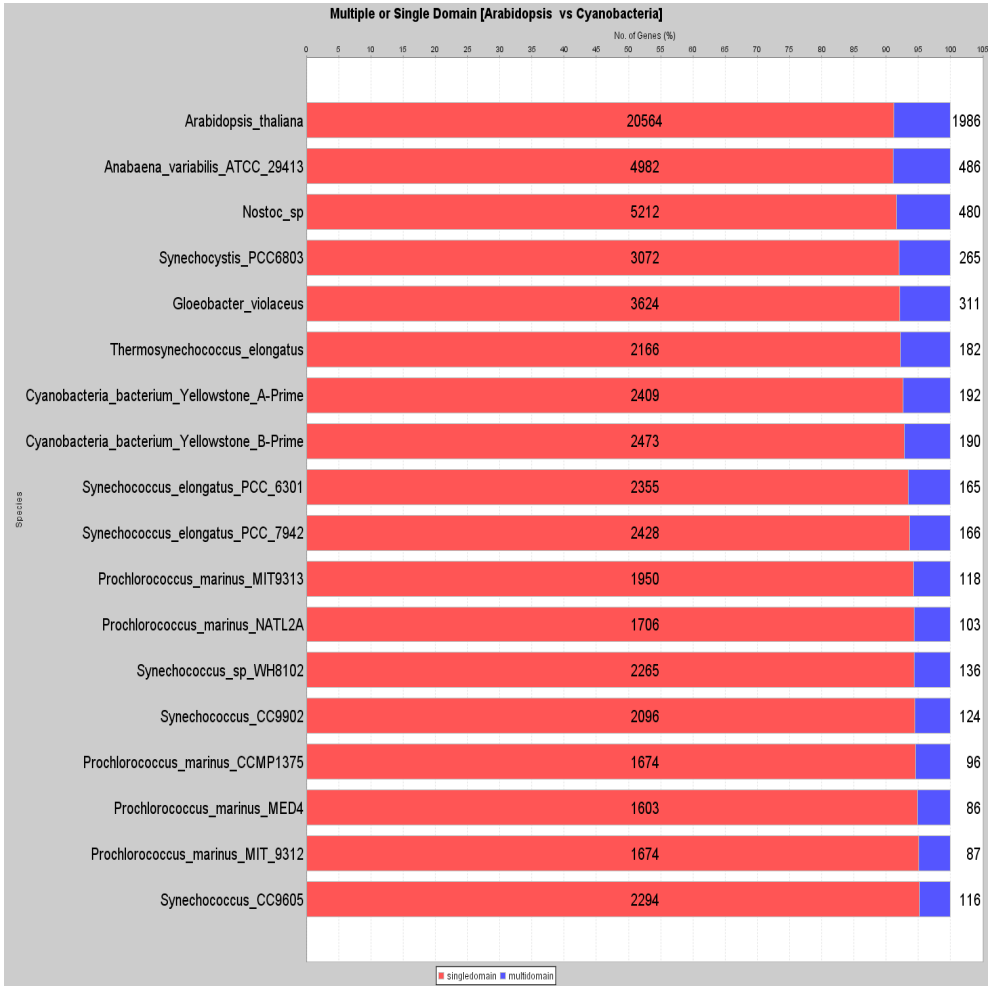

図 8　シロイヌナズナとシアノバクテリアにおけるマルチドメインとシングルドメインの割合

横軸: 生物種における全遺伝子を 100%とした各ドメインについての遺伝子数の割合 ; 縦軸: 生物種名

red: singledomain ; blue: multidomain

3.3 シロイヌナズナのオーソログ遺伝子とパラログ遺伝子

シロイヌナズナの全遺伝子について, 333 種全ての生物とのオーソログ遺伝子とパラログ遺伝子の探索結果を図 9 に示す. シロイヌナズナの遺伝子のうち重複遺伝子 (Duplicate Gene) を用いて, ドメインが探索される. シロイヌナズナの全遺伝子 (26,536 個) において, 重複遺伝子は 22,566 個 (85%) みつかった. そのうち, 他の生物種と相同性があるオーソログ遺伝子は 16,058 個 (71%) であった.

ここでの解析におけるオーソログ遺伝子は, 微生物 (真菌, 古細菌, 細菌) の遺伝子との相同性から求めたものである. そのため, シロイヌナズナは, 進化の過程で微生物からドメインを 71%受け継いでいる. また, パラログ遺伝子のみをもつ遺伝子に着目することで, 植物 (シロイヌナズナ) 固有のドメインを予測できる.

3.4 節以降では, シロイヌナズナの遺伝子について, 本手法により検出したオーソログ遺伝子・パラログ遺伝子から, 機能ごとにドメインの探索と解析をすることで, その遺伝子機能に特徴的な微生物から植物へ受け継がれたドメインの検出を行った.

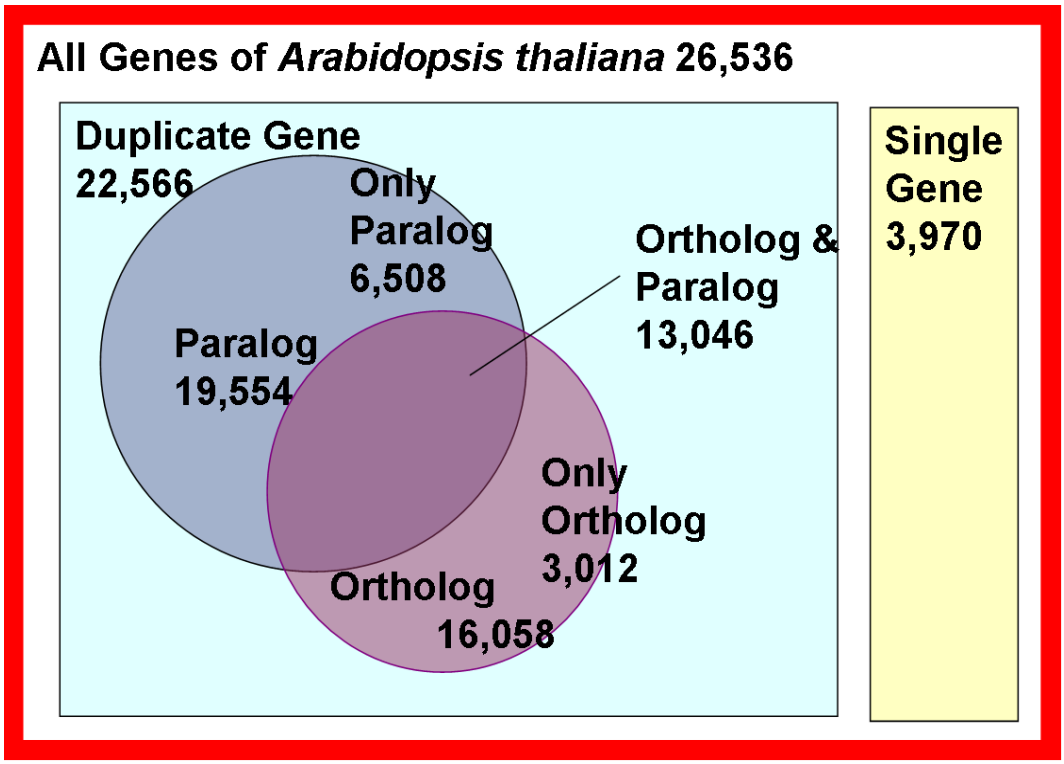

図 9 シロイヌナズナの遺伝子の分類

Duplicate Gene: オーソログ遺伝子またはパラログ遺伝子が少なくとも 1 つ以上存在する遺伝子
Single Genne: Duplicate Gene 以外の遺伝子
Ortholog: オーソログ遺伝子が 1 つ以上存在する遺伝子
Paralog: パラログ遺伝子が 1 つ以上存在する遺伝子
Only Ortholog: オーソログ遺伝子が 1 つ以上存在し, かつパラログ遺伝子が存在しない遺伝子
Only Paralog: パラログ遺伝子が 1 つ以上存在し, かつオーソログ遺伝子が存在しない遺伝子
Ortholog&Paralog: オーソログ遺伝子が 1 つ以上存在し, かつパラログ遺伝子も 1 つ以上存在する遺伝子

3.4 シロイヌナズナのドメイン解析

3.4.1 シロイヌナズナの遺伝子機能ごとのドメイン解析

シロイヌナズナの遺伝子機能は, TAIR (The Arabidopsis Information Resource) (http://www.arabidopsis.org/) データベースの 20 種類の機能分類にしたがう. 各機能に分類される遺伝子とドメインの数について, 表 1 と図 10 に示す. 表 1 と図 10 から, 微生物由来の遺伝子にマルチドメインが多く検出された. このことからも微生物から植物への遺伝子の進化の過程をドメインにより解析することは, 有用である. 表 1 と図 10 の遺伝子機能のうち, 遺伝子数が多くマルチドメインの割合が多い以下の 4 つの機能に分類される遺伝子について詳しくみていく.

- METABOLISM (代謝機能)
- TRANSCRIPTION (転写機能)
- CELLULAR COMMUNICATION/
  SIGNAL TRANSDUCTION MECHANISM
  (細胞内情報伝達機能/シグナル伝達機構)
- TRANSPORT FACILITATION (膜輸送機能)

表 1 シロイヌナズナの遺伝子機能における遺伝子とドメイン数

gene function	uniq arabidopsis			derived microbial		
	gene	domain		gene	domain	
		multi	single		multi	single
METABOLISM	199	2	197	2,117	200	1,917
TRANSCRIPTION	439	5	434	705	88	617
CELLULAR COMMUNICATION/ SIGNAL TRANSDUCTION MECHANISM	80	0	80	920	173	747
CELL RESCUE, DEFENSE AND VIRULENCE	183	12	171	401	98	303
TRANSPORT FACILITATION	40	0	40	438	89	349
PROTEIN FATE	31	0	31	362	86	276
ENERGY	35	0	35	237	17	220
PROTEIN SYNTHESIS	1	0	1	229	38	191
CELL CYCLE AND DNA PROCESSING	13	0	13	160	26	134
CONTROL OF CELLULAR ORGANIZATION	33	0	33	109	10	99
SUBCELLULAR LOCALISATION	12	0	12	111	19	92
CELLULAR TRANSPORT AND TRANSPORT MECHANISMS	7	0	7	113	12	101
CELL FATE	24	0	24	41	5	36
DEVELOPMENT	30	0	30	29	4	25
TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS	31	2	29	5	1	4
SYSTEMIC REGULATION OF / INTERACTION WITH ENVIRONMENT	15	0	15	23	6	17
STORAGE PROTEIN	8	0	8	18	0	18
REGULATION OF / INTERACTION WITH CELLULAR ENVIRONMENT	18	0	18	10	1	9
PROTEIN ACTIVITY REGULATION	0	0	0	6	1	5
PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT	2	0	2	3	0	3
UNCLASSIFIED PROTEINS	5,307	80	5,223	10,021	1,061	8,949

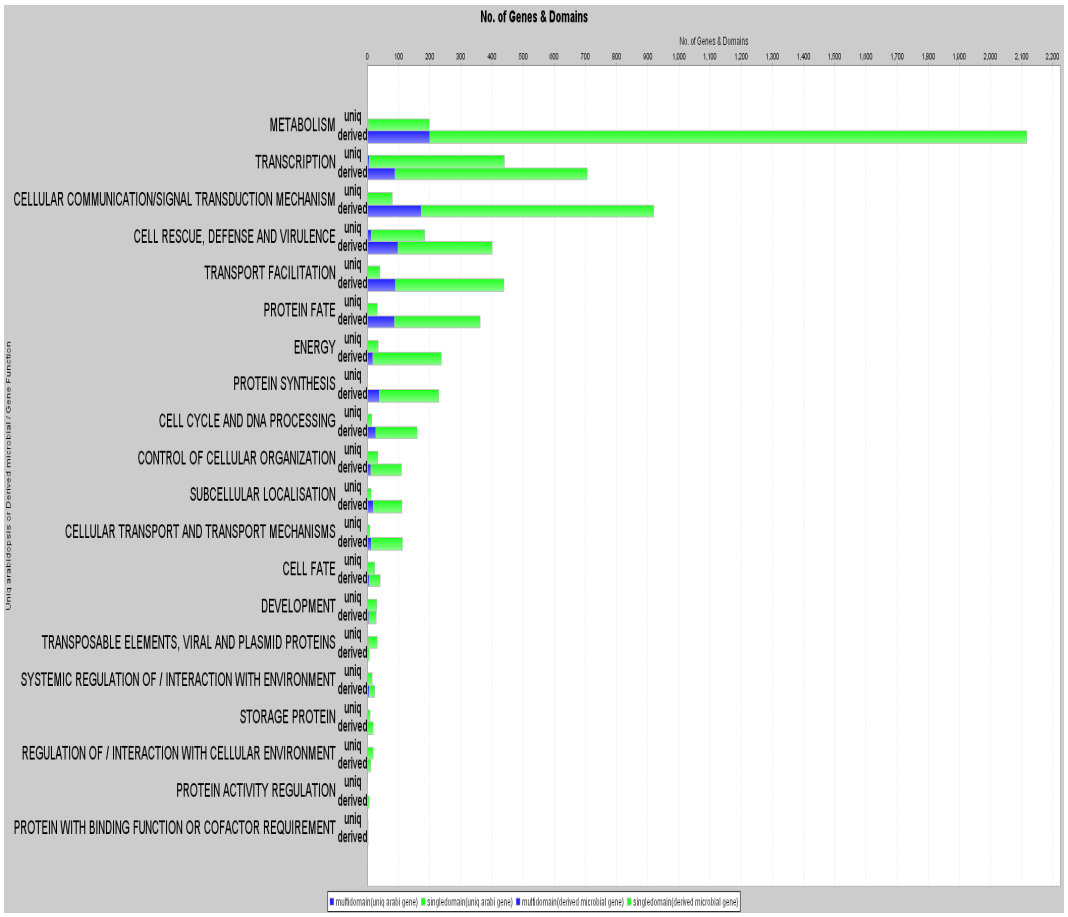

図 10 シロイヌナズナの遺伝子機能における遺伝子とドメインの分類

横軸: 遺伝子数 ; 縦軸: メインカテゴリー: 20 種の遺伝子機能, サブカテゴリー: uniq: シロイヌナズナ固有遺伝子と derived: 微生物由来の遺伝子

bule: multidomain ; green: singledomain

ドメインを有する遺伝子におけるマルチドメインとシングルドメインの割合

METABOLISM (代謝機能)

代謝機能に分類される遺伝子は, 表1と図10から微生物由来の遺伝子が植物固有の遺伝子に比べ圧倒的に多い. そのため, 微生物由来の遺伝子に注目する. 微生物から植物への進化の過程で, bi-functional遺伝子となったと報告されているシロイヌナズナの8個の遺伝子についてのドメインの探索結果をみていく. AT1G31230と AT4G19710 の遺伝子については, 2005 年に出された Curien G. らの論文 [17] を参照した. AT1G31860 と AT4G26900 の遺伝子については, 1998 年に出された Fujimori K. らの論文 [18], [19] をした. AT2G16370 の遺伝子については, 1993 年に出された Lazar G. らの論文 [20] を参照した. AT3G06860 の遺伝子については, 1999 年に出された Richmond TA. らの論文 [21] を参照した. AT3G18000 の遺伝子については, 2000 年に出された Bolognese CP. らの論文 [22] を参照した. AT4G21470 の遺伝子については, 2005 年に出された Sandoval FJ. らの論文 [23] を参照した. これら8個の各遺伝子の機能については, 表2に示す.

表 2 代謝機能: 解析対象遺伝子の詳細機能

atg code	gene name	gene function
AT1G31230	AK-HSDHI	aspartate kinase-homoserine dehydrogenase
AT4G19710	AK-HSDHII	aspartate kinase-homoserine dehydrogenase
AT1G31860	At-IE	Phosphoribosyl-AMP cyclohydrolase(PRA-CH)(hsiI)/ Phosphoribosyl-ATP pyrophosphohydrolase(PRA-PH)(hisE)
AT4G26900	HisH/HisF	glutamine amido-transferase/cyclase
AT2G16370	DHFR-TS	dihtdrofolate reductase and thymidylate synthase
AT3G06860	AIM1	Abnormal Inflorescence Meristem1
AT3G18000	AtNMT1	S-adenosyl-Met:phospho-base N-methyltransferase assays
AT4G21470	AtFMN/FHy	riboflavin kinase/FAD synthetase

表 3 には, 8 個の遺伝子についての本手法によるドメインの探索結果を示す. また, それぞれの遺伝子についてのドメイングラフを図 11-図 18 に示す.

表 3 代謝機能遺伝子におけるドメイン

atg code	gene length	no. of homo seq.	no. of homo species	no. of domain	domain1			domain2		
					start	end	length	start	end	length
AT1G31230	911	711	295	2	86	550	465	553	904	352
AT4G19710	916	707	296	2	91	555	465	558	909	352
AT1G31860	281	301	236	2	56	157	102	175	267	93
AT4G26900	592	549	245	2	64	262	199	280	592	313
AT2G16370	519	447	234	2	24	198	175	239	519	281
AT3G06860	725	1,975	246	2	16	205	190	312	711	400
AT3G18000	491	895	258	2	44	134	91	275	392	118
AT4G21470	379	838	274	2	14	193	180	238	363	126

これら 8 個の遺伝子は表 2 の遺伝子機能 (gene function) からもわかるように, 1 つの遺伝子において, 2 つの機能をもつ bi-functional 遺伝子である. このことを踏まえ表 3 のドメイン数 (no. of domain) に注目すると, 8 個全ての遺伝子がマルチドメインを形成していることがわかる. これは, 遺伝子において bi-function であることとマルチドメインとなることは, 深く関わっているといえる. また, オーソログ遺伝子の配列がみつかった生物種の数が本研究に用いた全体の約 70%でどの遺伝子においても保存されている. よって, 本手法の探索から得られたドメインは微生物に広域に存在し, かつよく保存されたドメインである. そのため, 8 個の遺伝子がそれぞれもつ 2 つのドメインは, 微生物由来であり, 各遺伝子において bi-function 遺伝子を形成していることが明らかである.

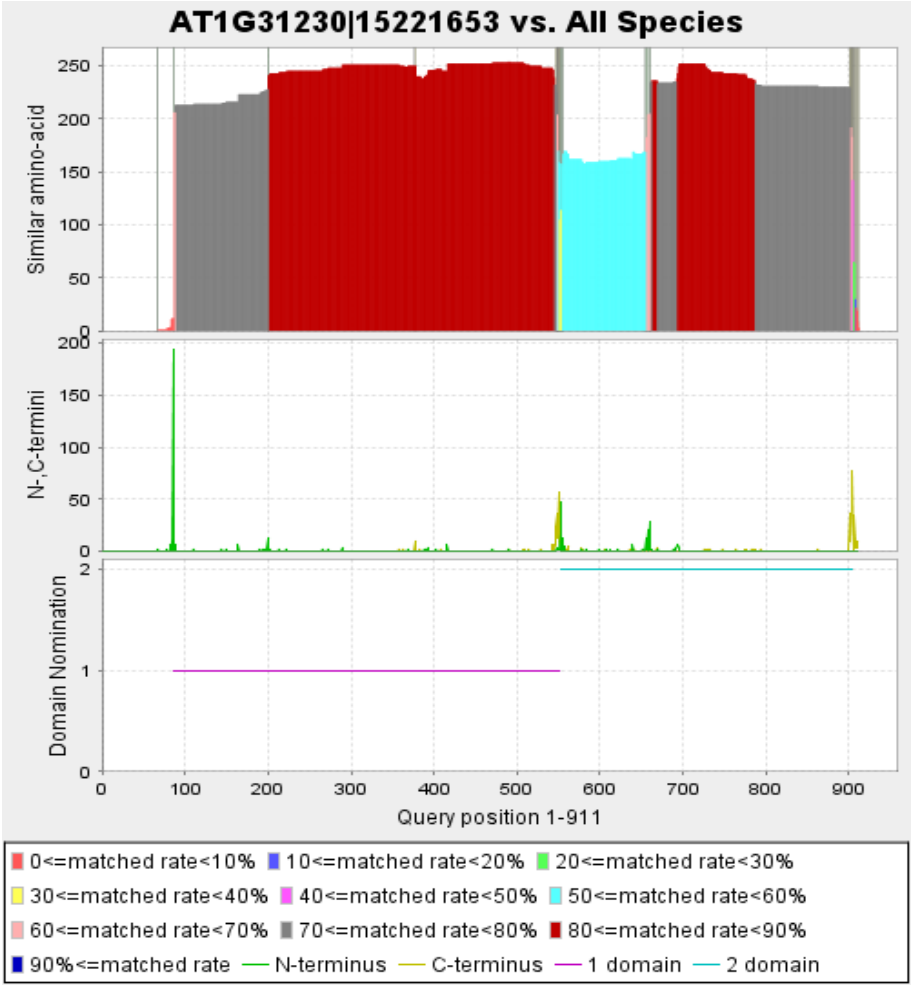

図 11 METABOLISM: AT1G31230

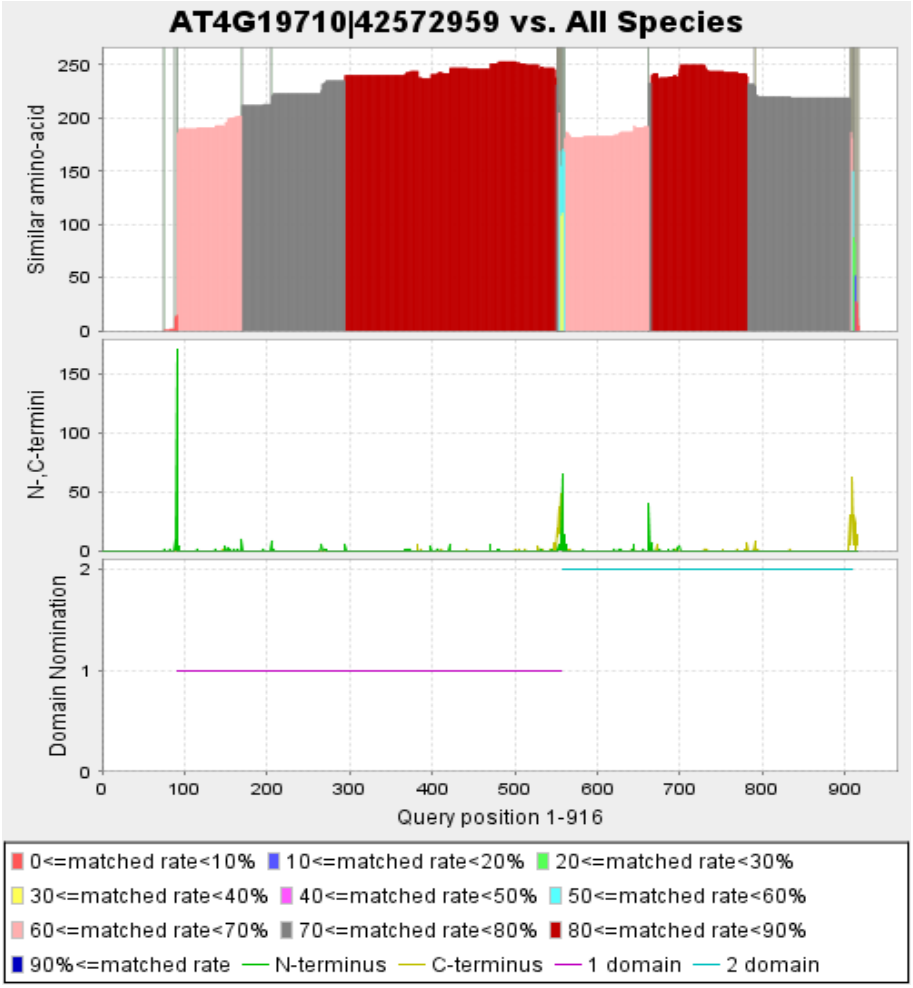

図 12 METABOLISM: AT4G19710

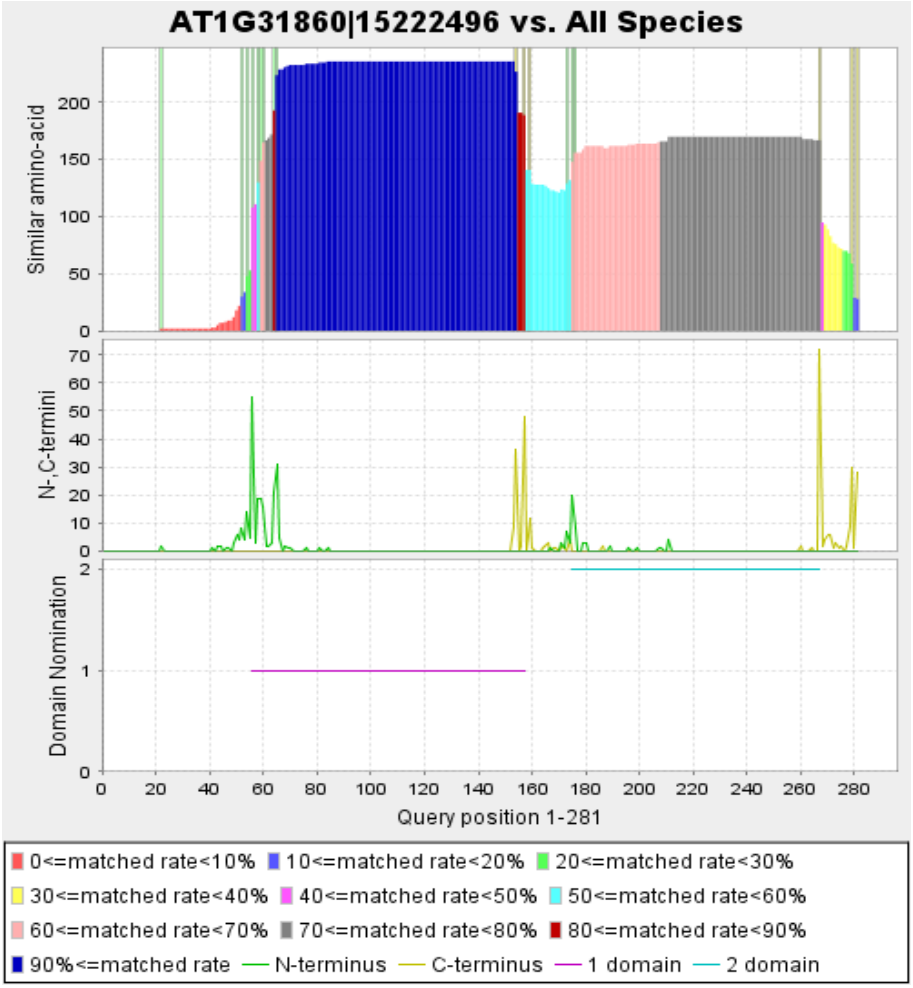

図 13 METABOLISM: AT1G31860

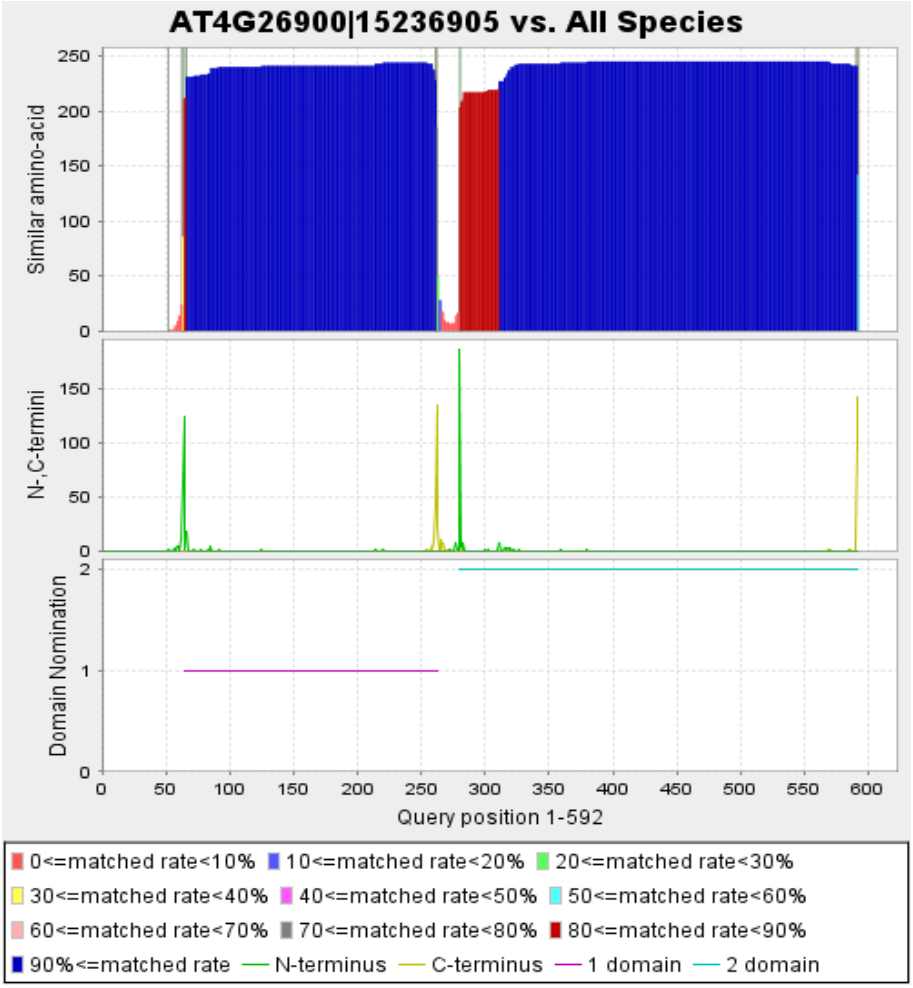

図 14 METABOLISM: AT4G26900

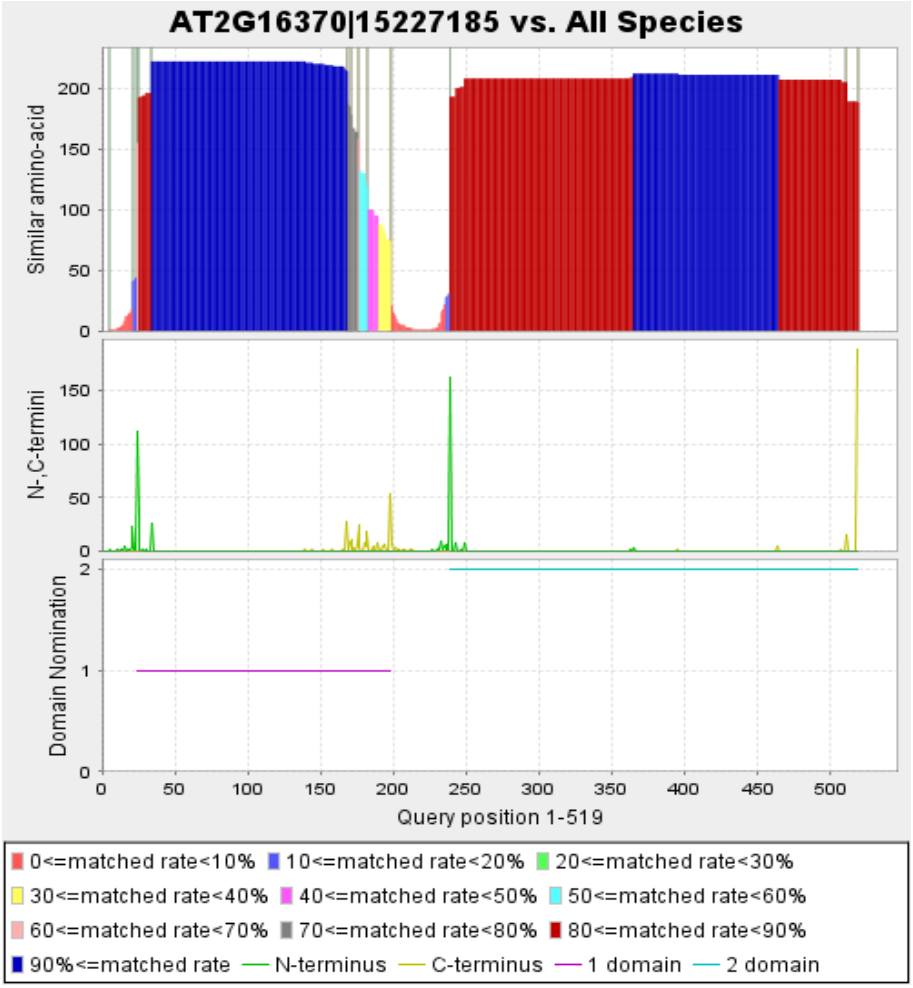

図 15 METABOLISM: AT2G16370

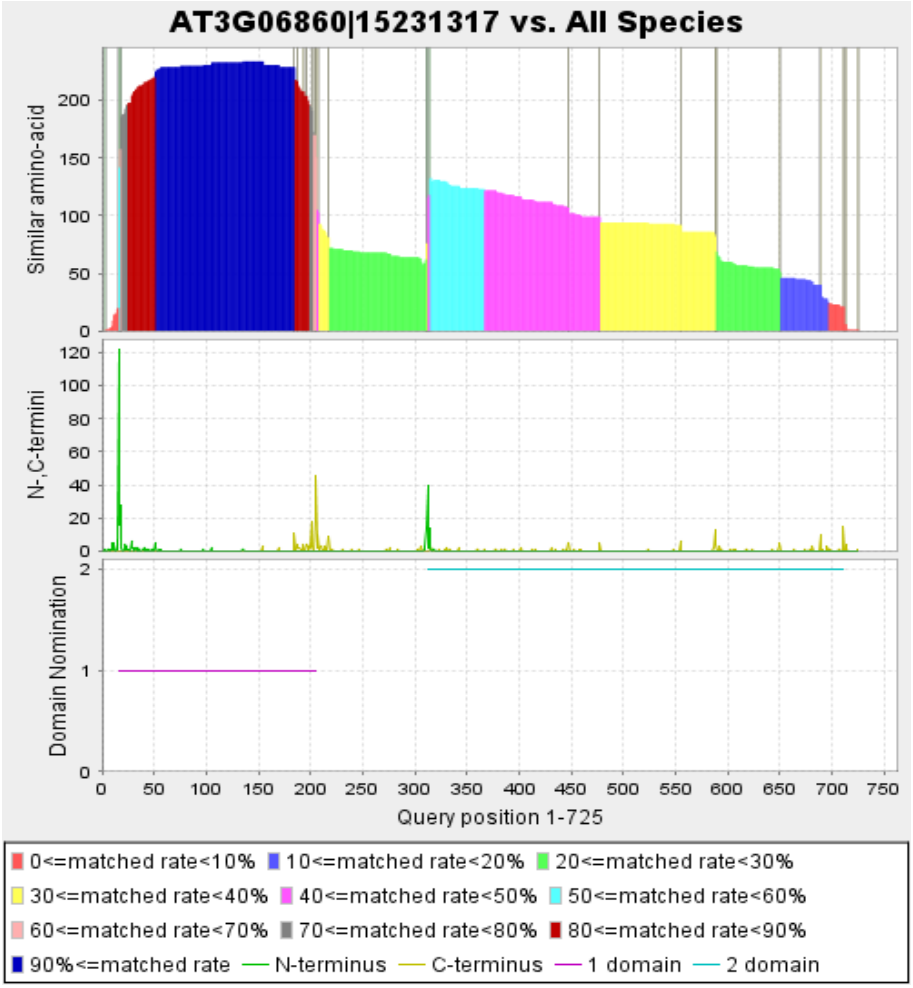

図 16 METABOLISM: AT3G06860

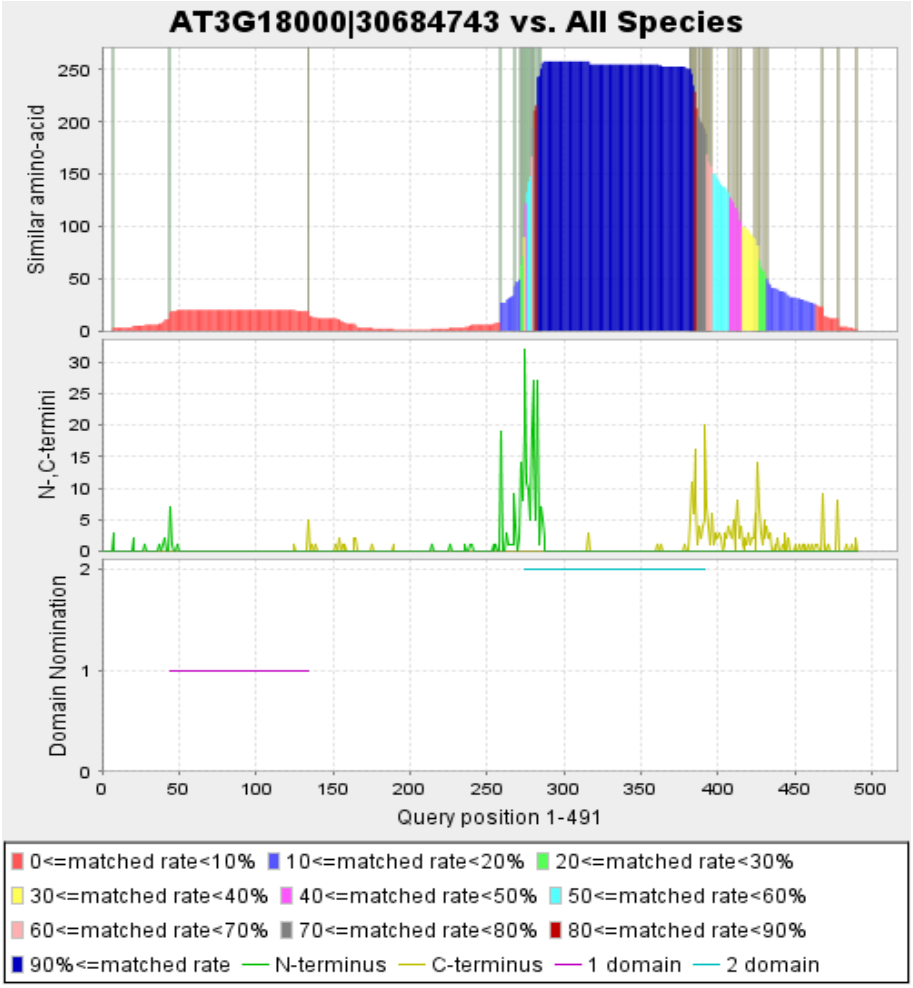

図 17 METABOLISM: AT3G18000

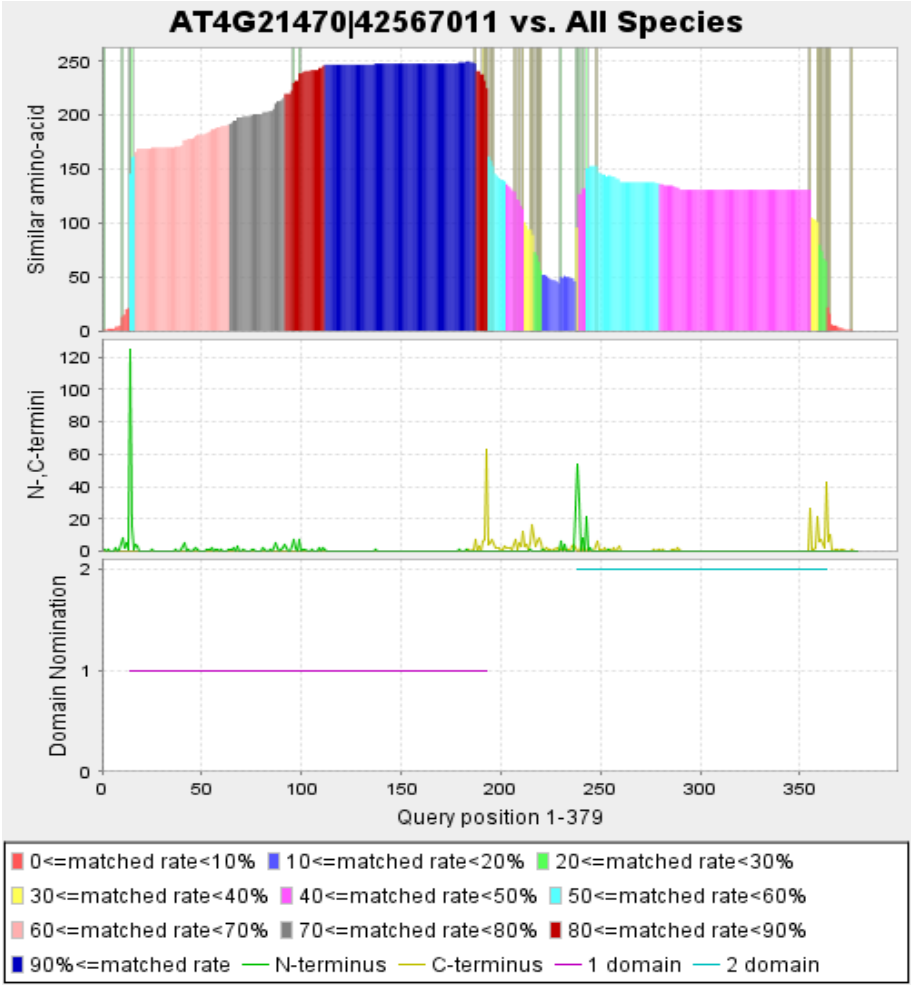

図 18 METABOLISM: AT4G21470

TRANSCRIPTION (転写機能)

表 1 と図 10 から転写機能に分類される遺伝子は, 他の機能に比べ植物 (シロイヌナズナ) 固有の遺伝子が多いことがわかる. そのため, 転写機能の遺伝子については, 植物固有のドメインに注目する.

転写因子 (transcription factor) 遺伝子は, シロイヌナズナゲノムの 5%以上を占めている. Riechmann JL. らは, 2000 年に植物, 動物, 真菌の 3 つの界に存在すると予想される各転写因子遺伝子の構成について比較解析を行った [24]. Riechmann JL. らは, 植物のゲノムデータには, シロイヌナズナを使用しており, 解析結果からシロイヌナズナには, 植物特異的な転写因子ファミリーが存在していることがわかった. 本研究では, この植物特異的な転写因子ファミリーの中で遺伝子数が多く報告されている, 5 個に注目し, シロイヌナズナの転写因子遺伝子におけるドメインの探索を行った. 5 個の転写因子ファミリー名, 遺伝子数については, 表 4 に示す. また, データはシロイヌナズナの転写因子ファミリーのデータベース RARTF (RIKEN *Arabidopsis* Transcription Factor database) (http://rarge.gsc.riken.jp/rartf/)[25] から最新 (2007 年 1 月) のものを入手した.

表 4 シロイヌナズナ における転写因子ファミリーと遺伝子数

superfamily	no. of gene	subfamily	no. of gene
AP2/EREBP	145	ABSCISIC ACID-INSENSITIVE4	144
		AINTEGUMENTA	145
		APETALA2	144
		CBF1	145
		DREB2A	137
ARF	119	MONOPTEROS/ARF5	74
		NPH4/ARF7	119
		ETTIN/ARF3	45
NAC	106	CUP-SHAPED COTYLEDON2	106
WRKY(Zn)	72		
Aux/IAA	49		

各転写因子ファミリーに分類される遺伝子について, 本手法によるドメインの探索結果を表 5 と図 19 に示す.

表 5　シロイヌナズナの転写因子ファミリーごとの遺伝子とドメイン数

superfamily	uniq arabidopsis			derived microbial		
	gene	domain		gene	domain	
		multi	single		multi	single
AP2/EREBP	111	0	111	14	2	12
ARF	89	1	88	9	2	7
NAC	94	0	94	5	3	2
WRKY(Zn)	48	0	48	16	1	15
Aux/IAA	40	0	40	5	2	3

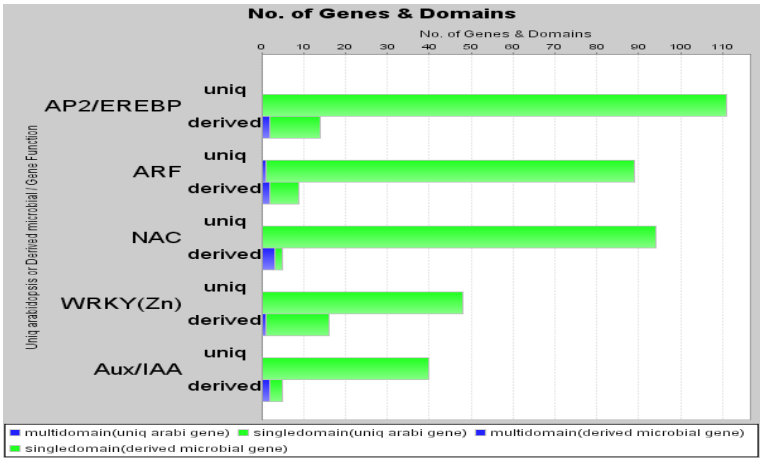

図 19　シロイヌナズナの転写因子ファミリーごとの遺伝子とドメインの割合

bule: multidomain ; green: singledomain
ドメインを有する遺伝子におけるマルチドメインとシングルドメインの割合

表 5 と図 19 から, 明らかに 5 個の転写因子ファミリー (AP2/EREBP, NAC, WRKY(Zn), ARF, Aux/IAA) は, シロイヌナズナのみと相同性を示すものが圧倒的に多い. これら 5 個の転写因子ファミリーは, Riechmann JL. らによって植物特異的なファミリーと位置づけられている. したがって, これらのファミリーに属している遺伝子とドメインは植物固有であることがいえる.

表 5 から, 植物固有で, かつマルチドメインとなる遺伝子が転写因子ファミリー ARF に 1 つのみ検出されたことがわかる. マルチドメインは, 進化の過程で異なる遺伝子に存在していた複数のドメインが 1 つの遺伝子に入り込むことが原因で形成されることがよく知られている. しかし, 同生物種内のみだけでのマルチドメイン化をみていくことは, その生物種特異的な何らかの構造をみることが可能になると考えられる. そのため, ARF でみつかった植物に固有の遺伝子で, かつマルチドメインを形成している遺伝子 AT2G24650 についてみていく. AT2G24650 についてシロイヌナズナの全遺伝子からみつかったパラログ遺伝子の配列を図 20 に示す. また, 図 20 のパラログ遺伝子から探索したドメインの結果を図 21 に示す.

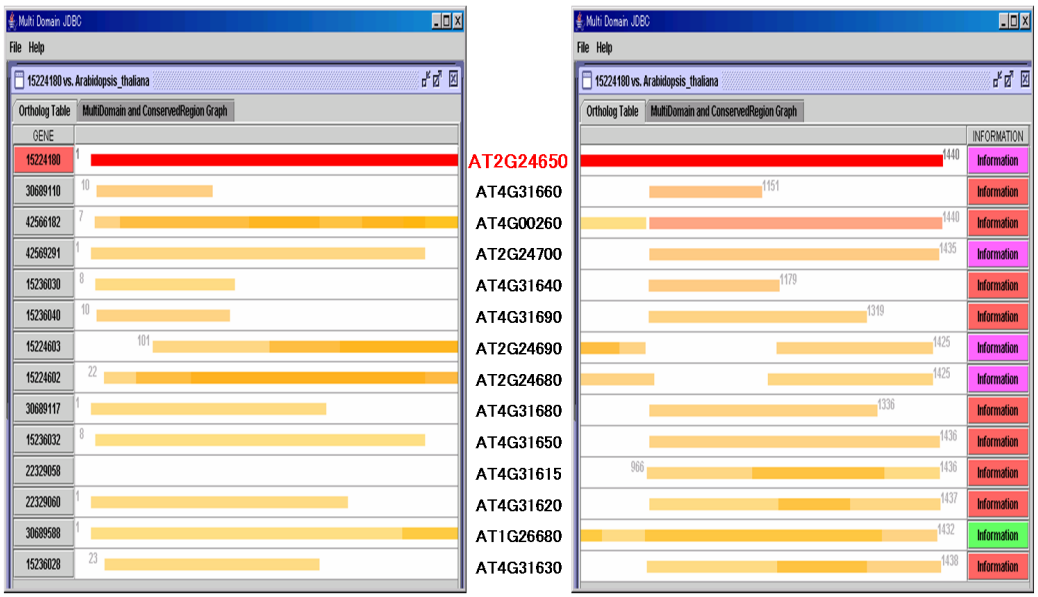

図 20　TRANSCRIPTION: AT2G24650 におけるパラログ遺伝子

red line: 問い合せ遺伝子 AT2G24650 ; orange lines: パラログ遺伝子の問い合せ遺伝子における類似配列領域

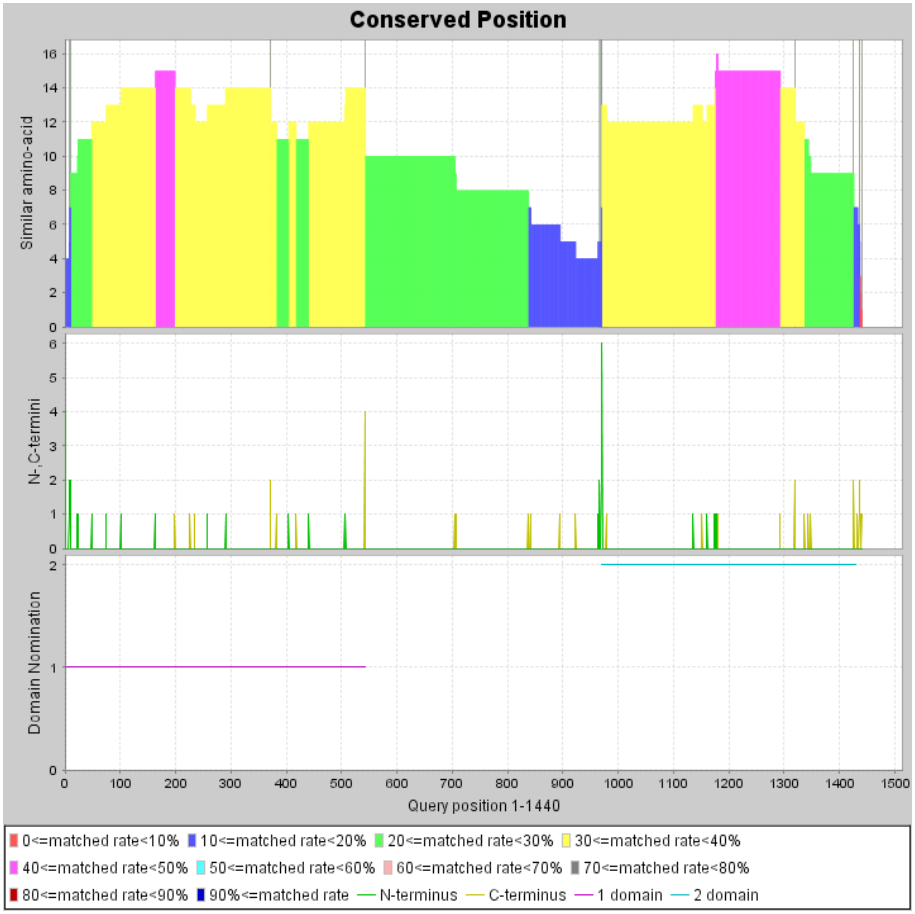

図 21　TRANSCRIPTION: AT2G24650

図 20 でみつかったパラログ遺伝子がそれぞれどの転写因子ファミリーに分類されているかを調べた．その結果, AT4G31620 以外は全て AT2G24650 と同じ転写因子ファミリー ARF に分類された．また, AT4G31620 は RARTF には登録されていなかったが, NCBI の RefSeq から, 転写因子の機能をもつことがわかった．このことから, AT4G31620 は転写因子ファミリー ARF に分類されると予測される．

転写因子ファミリー ARF は, Riechmann JL. らの研究によるとシロイヌナズナの他の転写因子ファミリーとドメインシャッフリングを起こすことがわかっている．そのため, AT2G24650 はドメインシャッフリングが原因でマルチドメインとなったと考えられる．

CELLULAR COMMUNICATION/ SIGNAL TRANSDUCTION MECHANISM (細胞内情報伝達機能/シグナル伝達機構)

細胞内情報伝達機能において, 代表的な遺伝子機能であるキナーゼ (kinase) についてみていく [26]-[28]. キナーゼは, ATP の γ-リン酸基を糖, アミノ酸, タンパク質, リン脂質などに転移する反応を触媒する酵素の総称である. 特にタンパク質を基質にするものはプロテインキナーゼと呼ぶ. 逆にリン酸化タンパク質から脱リン酸化を行う酵素はホスファターゼと呼ばれる. プロテインキナーゼには, 細胞内情報伝達を担うセカンドメッセンジャーにより活性化されるものや, 受容体キナーゼ群のようにキナーゼ自身が細胞外から情報を受け取り自分自身をリン酸化 (自己リン酸化) することで活性を調節するものが含まれる. 細胞内にはホスファターゼが存在し, 多くのキナーゼ類は迅速で可逆的な活性調節が行われ, 細胞の形態変化, 増殖, 物質の膜透過など基本的な生体機能に重要な役割を果たすと考えられる. ある種のタンパク質はリン酸化によってコンホメーション変化や, 複合体の形成などが引き起こされ, 活性が調製される.

図 22 と図 23 は, シロイヌナズナの STY(Serine/Threonine/Tyrosine) タンパク質キナーゼの系統樹とファミリーごとのドメインのタイプを表したものである. これらは, 2006 年に Rudrabhatla P. らによって発表された論文 [28] から引用した. 図 22 に記載されているシロイヌナズナの遺伝子それぞれについて, 本手法によるドメインの探索をおこなった. ドメインの探索結果については, ファミリーごとに 1 つの遺伝子を代表として図 24-図 32 に示す. 探索されたドメインと図 23 のドメインを対応させると, 図 23 で図示されている全ての遺伝子に表れているチロシンキナーゼ (tyrosine kinase) ドメインは, 本手法からも検出された. また, Family 2.2 のほとんどの遺伝子に存在する PAS ドメイン, Group III のほとんどの遺伝子に存在するアンキリン (ankyrin) ドメインも検出された. このことから, チロシンキナーゼ, PAS, アンキリンのドメインは, 微生物由来であることがわかる. 次に, Family 2.1 に注目する. Rudrabhatla P. らの解析結果によると Family 2.1 のほとんどの遺伝子には PB1 ドメインが存在する. 本手法の結果 (図 28) では, PB1 のドメインは探索されなかった. このことは, 本研究においてシロイヌナズナのド

メインを決めるのは, 微生物との相同配列から得るオーソログ遺伝子と, シロイヌナズナとの相同配列から得られるパラログ遺伝子である. そのため, PB1 ドメインは, 本研究に用いた 333 種の生物以外に由来するドメインと予測できる.

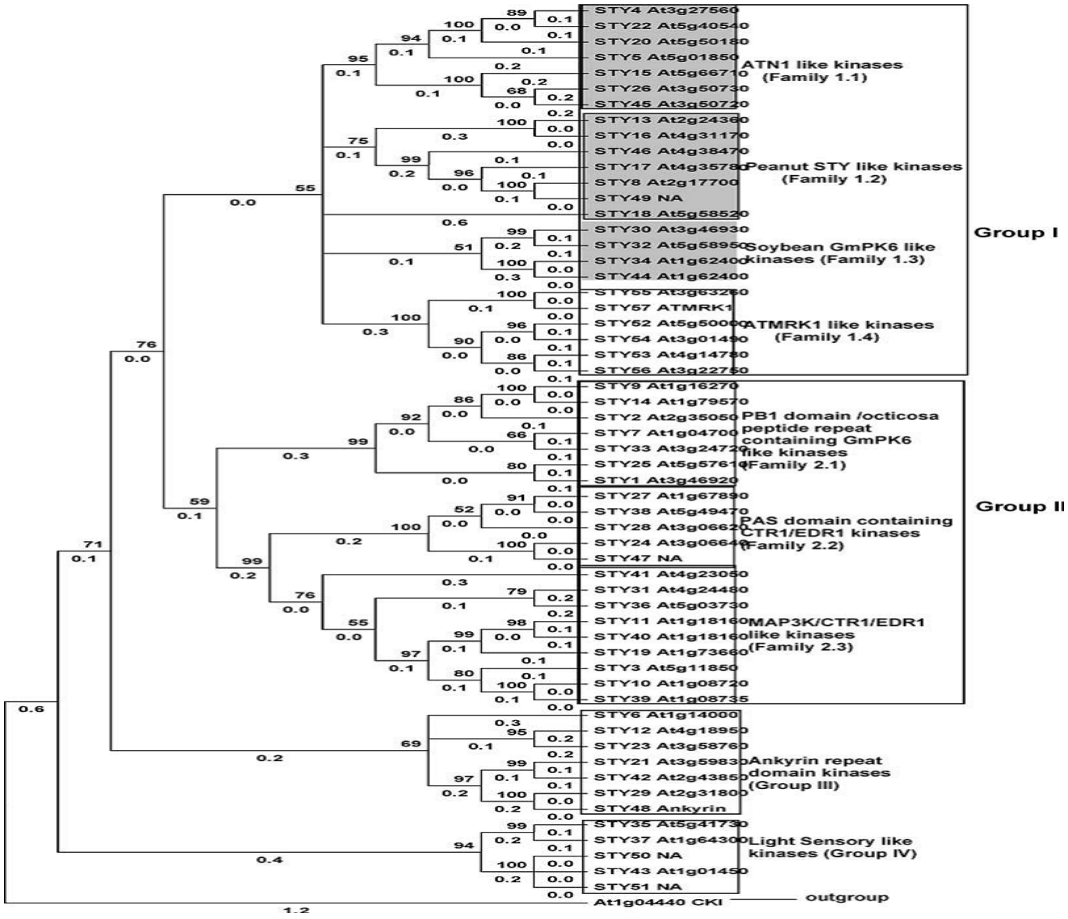

図 22 シロイヌナズナの STY キナーゼファミリーの遺伝子における系統樹

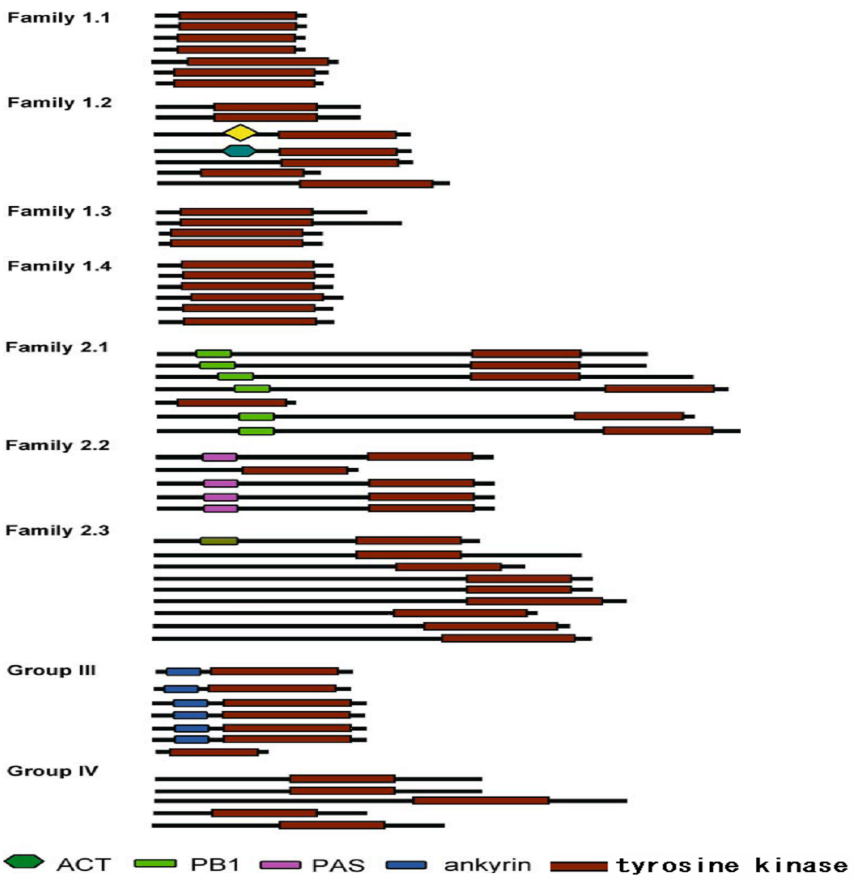

図 23 シロイヌナズナの STY キナーゼのファミリーごとのドメインタイプ

red: tyrosine kinase catalytic domain ; green: ACT domain ; light green: PB1 domain ; pink: PAS domain ; blue: ankyrin

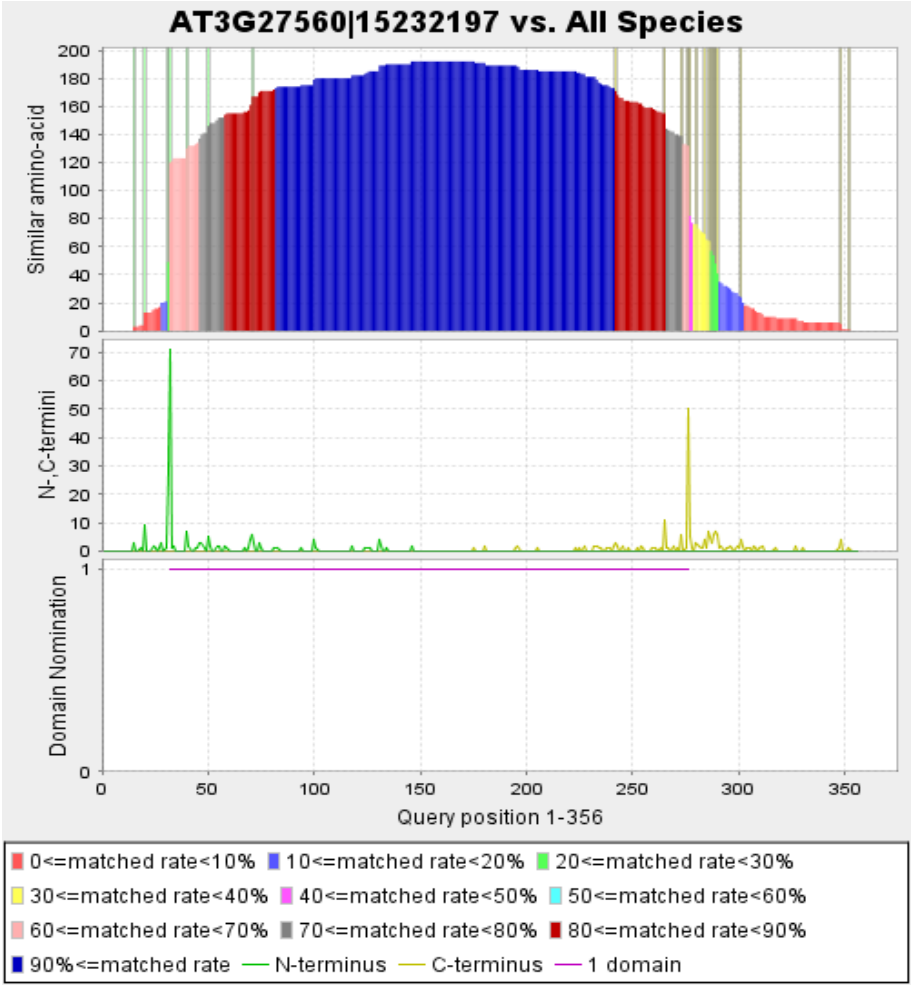

図 24 CELL COM: Family1.1

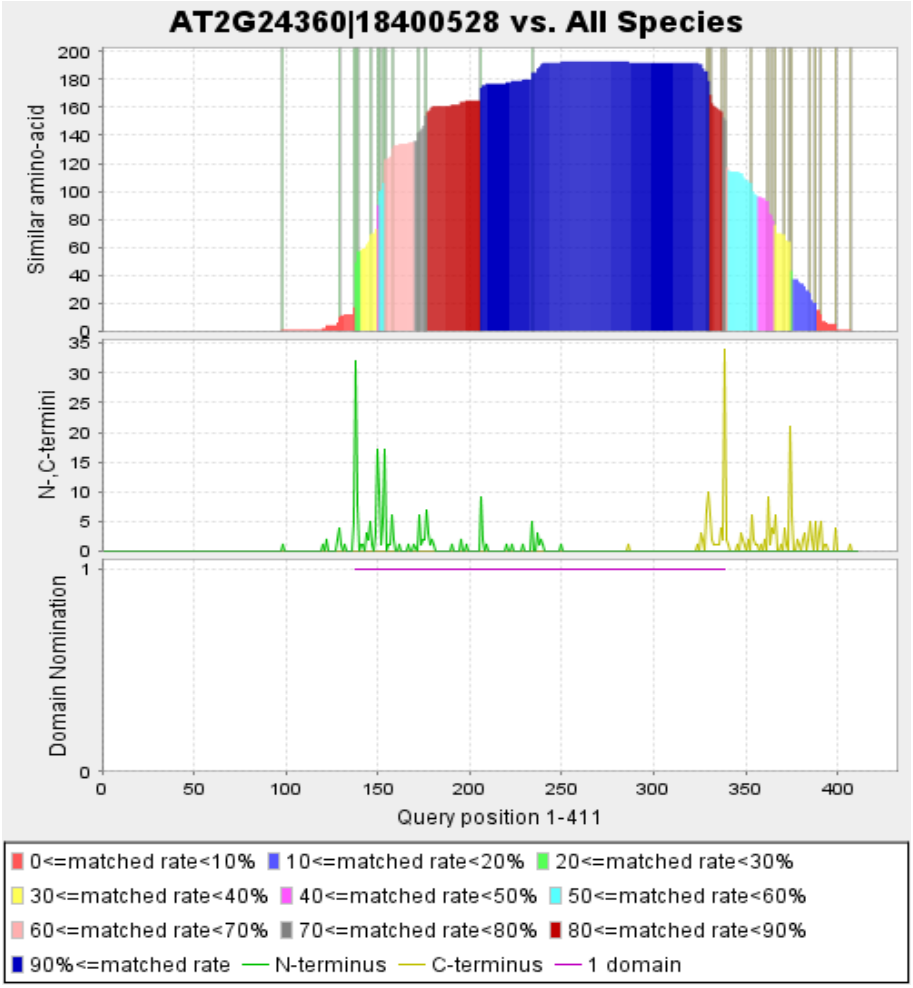

図 25 CELL COM: Family1.2

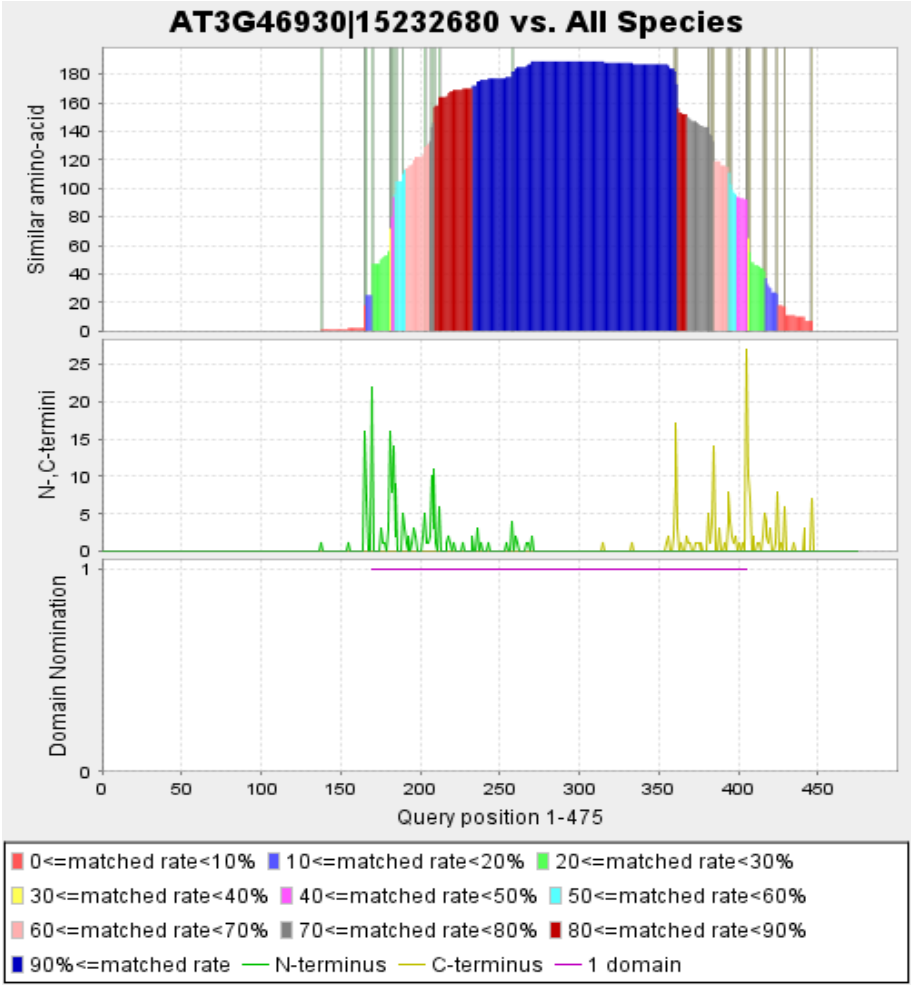

図 26 CELL COM: Family1.3

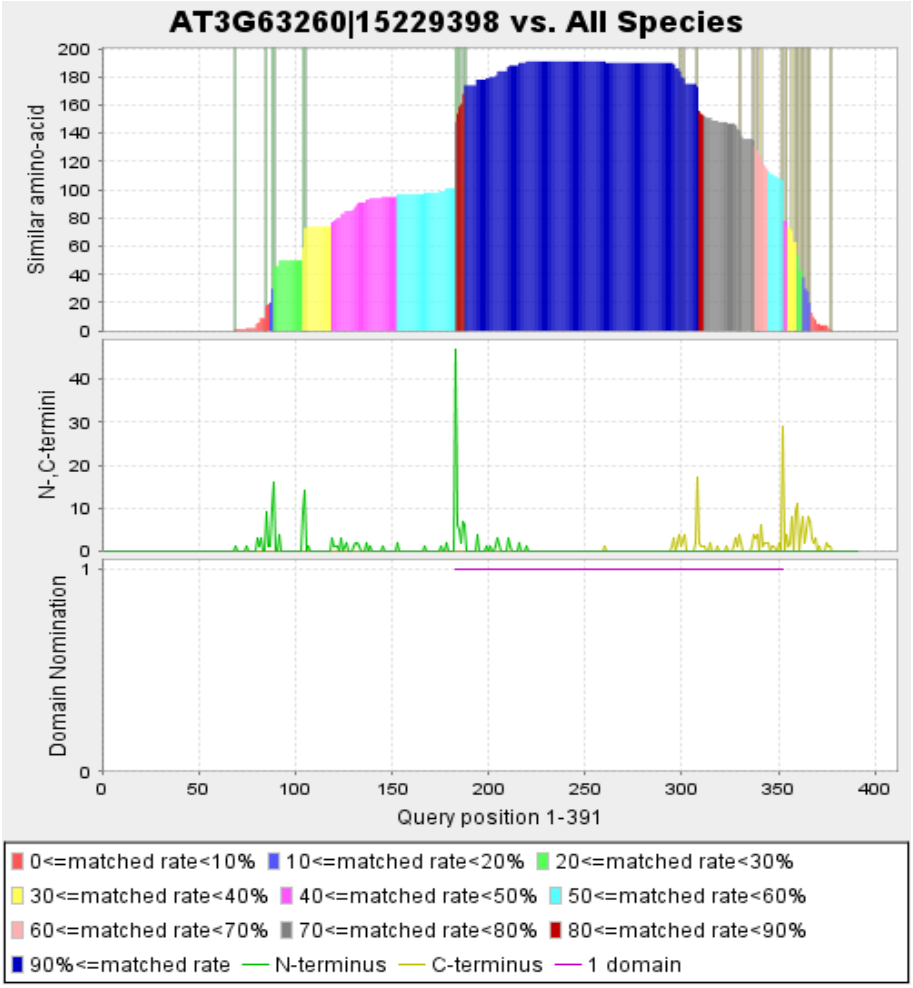

図 27 CELL COM: Family1.4

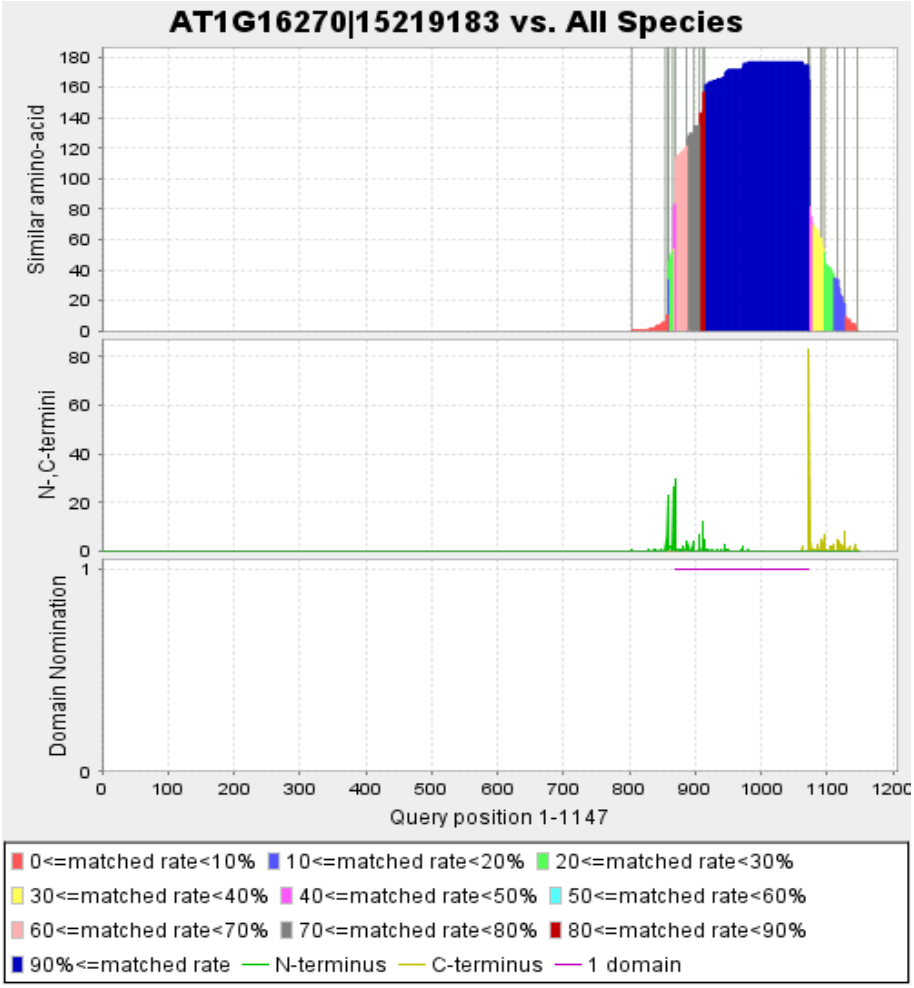

図 28 CELL COM: Family2.1

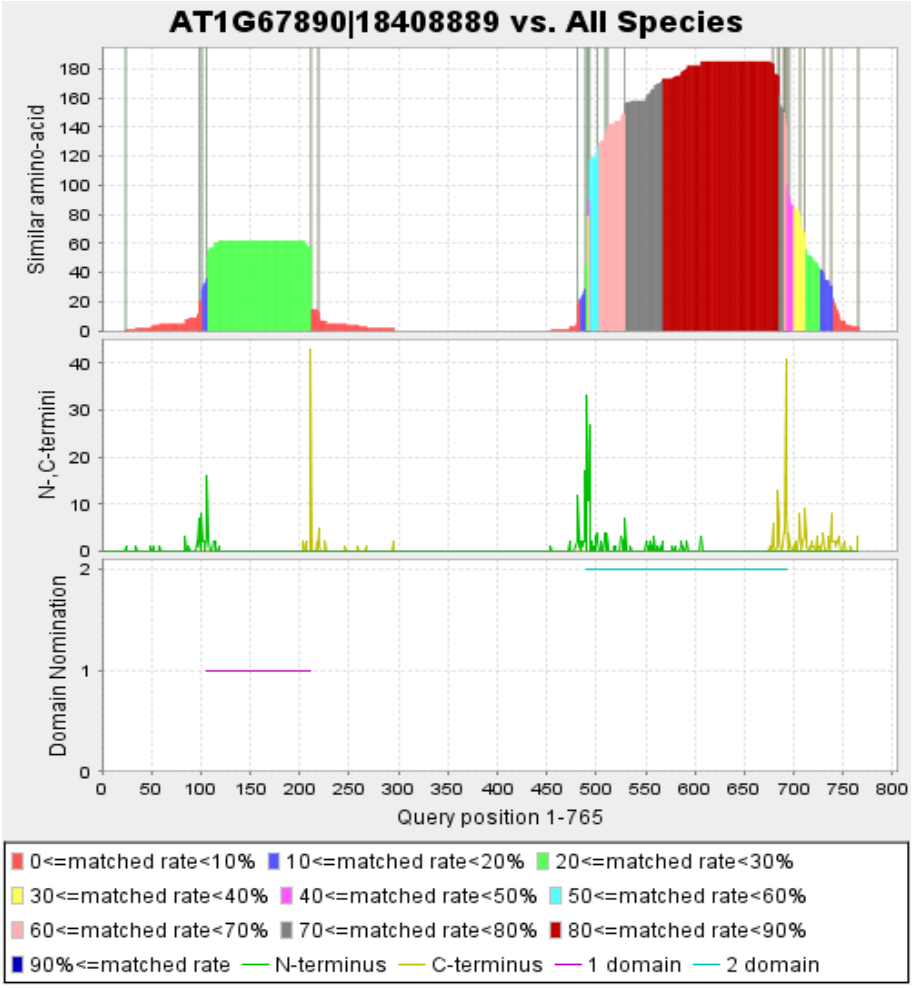

図 29 CELL COM: Family2.2

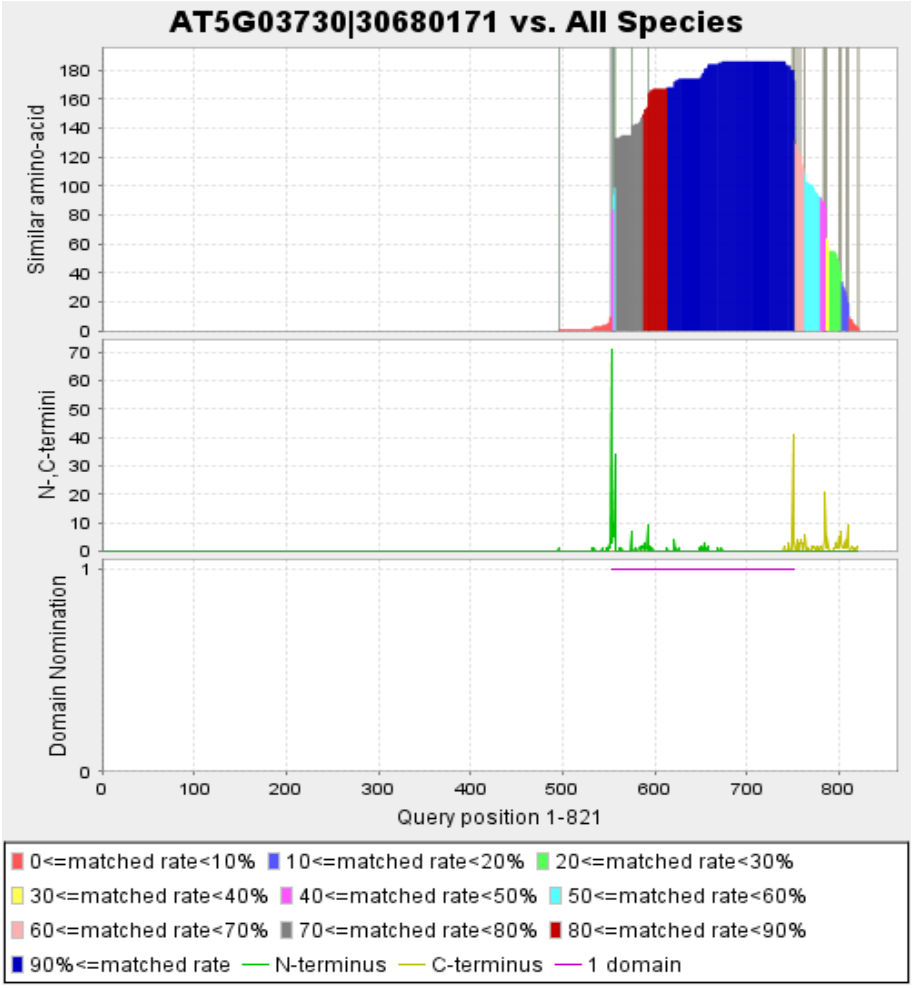

図 30 CELL COM: Family2.3

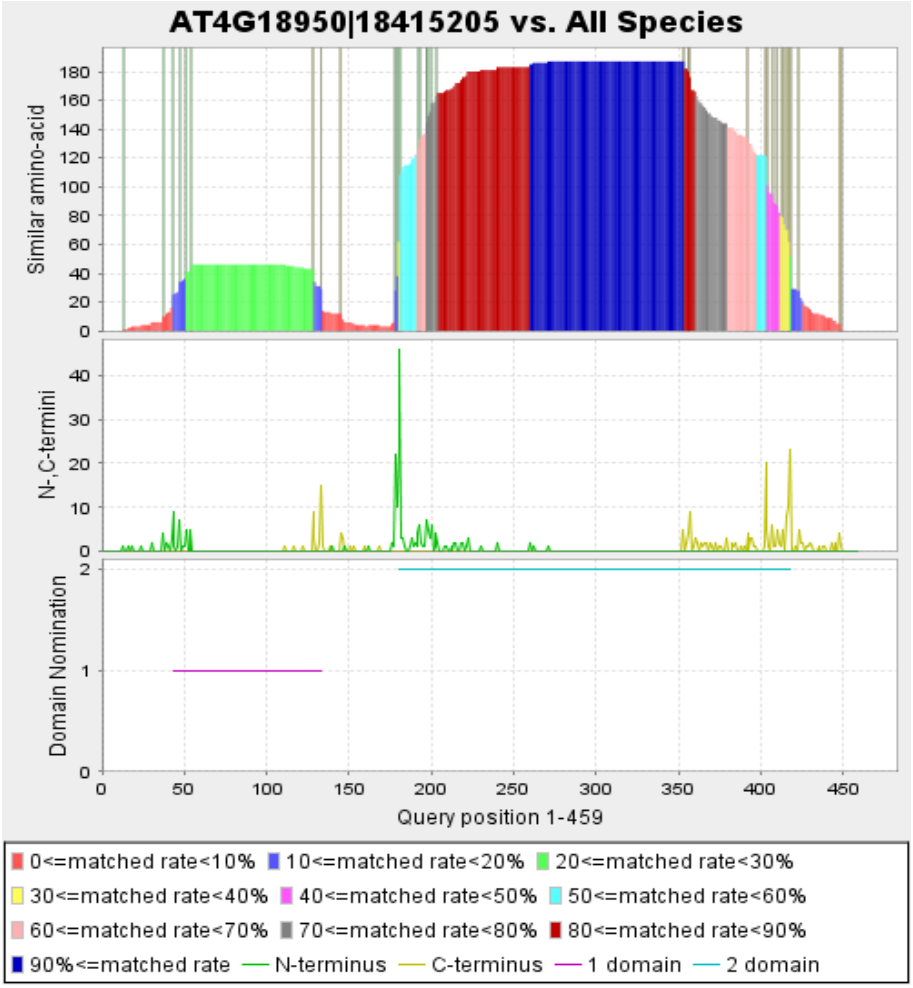

図 31 CELL COM: GroupIII

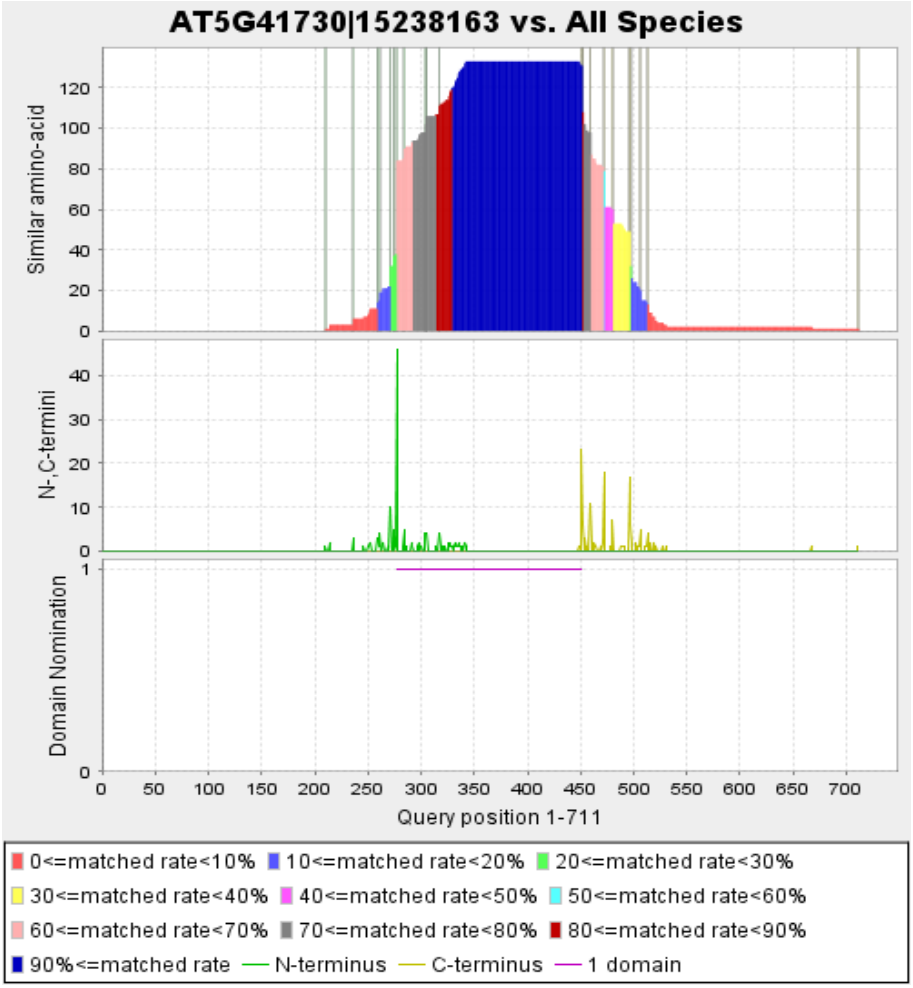

図 32 CELL COM: GroupIV

TRANSPORT FACILITATION (膜輸送機能)

シロイヌナズナゲノムの約5%の遺伝子が膜輸送機能を有している [29], [30]. また, 膜輸送タンパク質は, 46 個のファミリーに分けることができる. このファミリーのうち, ABC 輸送体 (ABC transporter) のサブファミリーについての解析を 2001 年に Sanchez-Fernandez R. ら [29] が行った. ABC 輸送体は, 2 個の ATP 結合部位を有し, ATP-結合カセット輸送タンパクと総称される. 生体に侵入した毒物や薬物の排出 (肝臓, 腎臓, 大腸), 塩素イオンの輸送などの機能がある.

図 33 は, ABC 輸送体のサブファミリー 13 個 (MDR, MRP, PDR, AOH, PMP, WBC, ATH, ATM, TAP, RLI, GCN, SMC, NAP) それぞれについて, Sanchez-Fernandez R. らが研究で明らかにしたドメインである. また, Sanchez-Fernandez R. らは, ABC 輸送体のファミリーに属す遺伝子配列を用いて, 系統解析を行っている (図 34 参照). 図 34 の系統樹からわかるように, それぞれの遺伝子は, ABC 輸送体のサブファミリーごとに系統的に分類される. Sanchez-Fernandez R. らの ABC 輸送体遺伝子の系統樹と, 本手法でのドメイン探索の結果を照らし合せたものもまた, 図 34 に示す. 本手法でのドメイン探索においても, ABC 輸送体のサブファミリーごとに遺伝子におけるドメインの構造がよく似ていた. 図 33 と図 34 を本手法で得たドメイン構造と比較すると, 全てのサブファミリーで, NBFs (nucleotide-binding folds) がよく保存されていた. このことから, NBFs は ABC 輸送体において重要なドメインと考えられる. また, NBFs はシロイヌナズナ以外の生物種の遺伝子とも相同であり, 配列が高く保存されていることが, 本手法によるドメインの探索結果から得られた. そのため, NBFs は, 微生物から植物への進化の過程で, 微生物から受け継がれたドメインであるとわかる.

次に, 図 34 から ABC 輸送体の配列を用いた系統関係と遺伝子におけるドメインの形状についての関連性をみていく. 図 34 のマルチドメインとなっている遺伝子をもつファミリーに注目すると, これらマルチドメインを形成しているファミリー同士は, ある程度まとまって存在している. そのため, 遺伝子におけるドメイン構造は, 進化や機能などの系統分類において重要な役割があると考えられる.

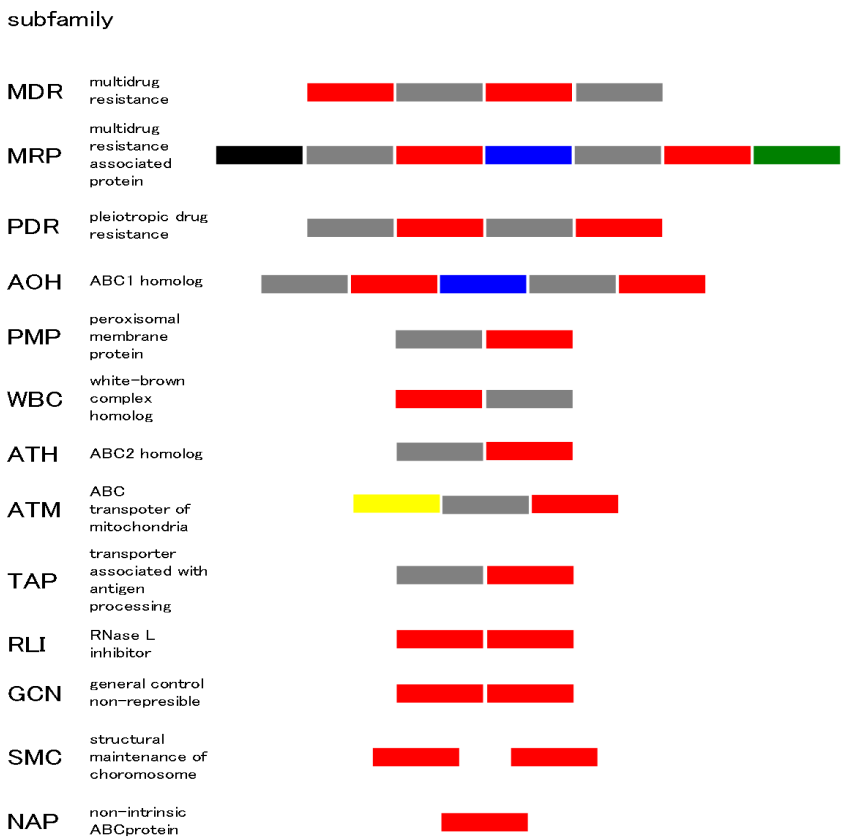

図 33　シロイヌナズナの ABC 輸送体遺伝子におけるサブファミリーごとのドメイン

red: NBFs 1 and 2 ; black: TMD0 ; gray: TMDs 1 and 2 ;
blue: linker domain ; green: C-terminal extension ; yellow: amphipathic signal peptide

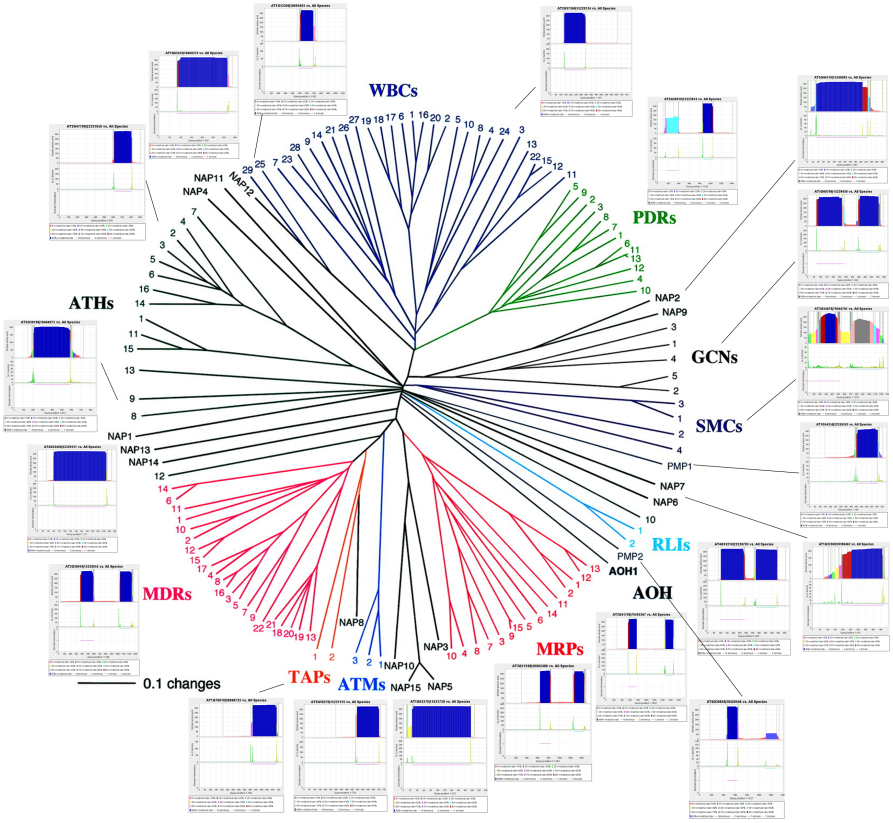

図 34　シロイヌナズナの ABC 輸送体遺伝子における系統樹とドメイン探索結果

Protein sequences were aligned using ClustalX and subjected to phylogenetic analysis by the distance with neighbor-joining method using PAUP4.04a. The reliabilities of each branch point, as assessed by the analysis of 1000 computer-generated trees (bootstrap replicates), were in excess of 90% except for those discussed in the text.

3.5 微生物由来のドメインと植物固有のドメイン

3.3 節と 3.4 節から, 本手法により探索されたドメインは, オーソログ遺伝子・パラログ遺伝子を検出する際の生物種によって, どの生物種由来のドメインかを確認できることが明らかである. 図 35 にシロイヌナズナの遺伝子及びドメインが微生物由来か植物固有かを分類した統計結果を示す. 本研究で植物 (シロイヌナズナ) と微生物 (真菌, 古細菌, 細菌) のゲノムを用い, 植物の遺伝子上のドメインの探索をおこなったことで, 微生物から植物の進化の過程において, 受け継がれたドメインを明らかにすることができた. 微生物から植物の遺伝子の進化では, ドメインや遺伝子の融合・解離などが起こりながら, 遺伝子が複雑化, 多様化している. そのため, 植物は微生物から受け継いだドメインと独自に生成したドメインが存在している. 本手法では, 植物固有のドメインの予測も行うことができた.

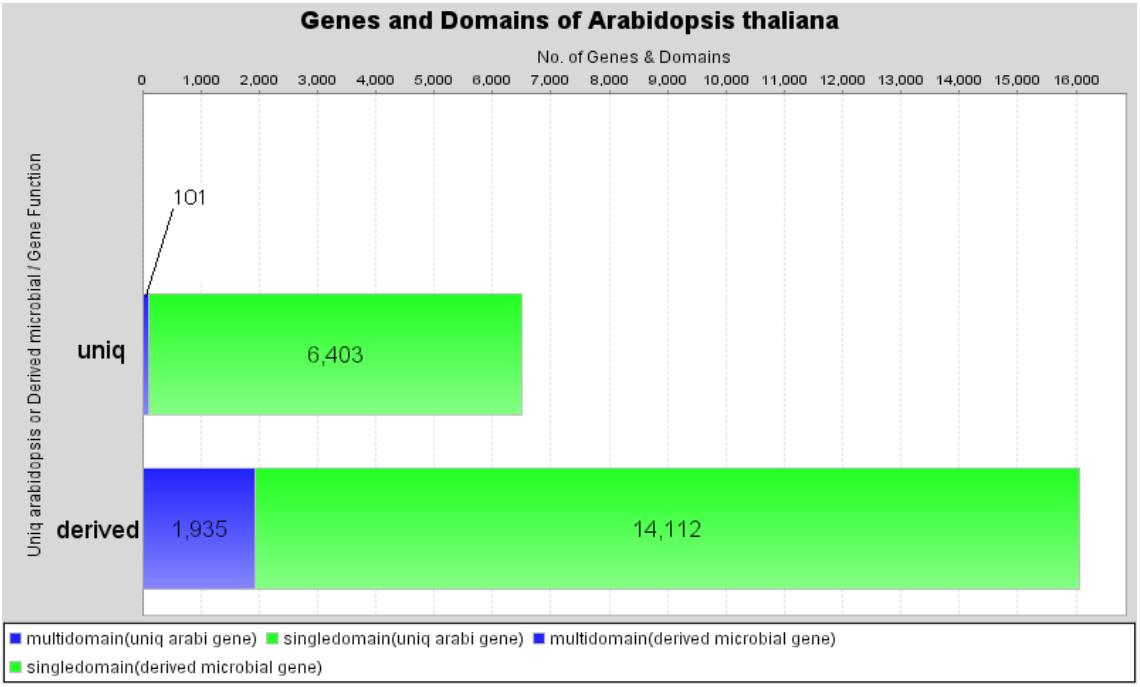

図 35 微生物由来・植物固有の遺伝子とドメイン

横軸: 遺伝子数 ; 縦軸: uniq: シロイヌナズナ固有遺伝子と derived: 微生物由来の遺伝子
bule: multidomain ; green: singledomain

4. 結論

本研究では, オーソログ遺伝子・パラログ遺伝子の配列からドメインを探索するアルゴリズムの開発を行った. 始めに, オーソログ遺伝子・パラログ遺伝子の検出を BLAST と本研究独自のオーソログ遺伝子・パラログ遺伝子の定義に基づいて行った. 本研究では, どちらの遺伝子側からも BLAST の結果で得た類似配列領域が等しく, 互いに十分な類似度を示す遺伝子対は相同であるとし, この 2 つの遺伝子を異生物間においてはオーソログ遺伝子, 同生物間においてはパラログ遺伝子と定義した. 次に, ある遺伝子について検出されたオーソログ遺伝子・パラログ遺伝子の類似配列領域から, 遺伝子におけるオーソログ遺伝子・パラログ遺伝子との高保存配列領域を求める. また, 各類似配列領域における開始残基位置と終了残基位置の保存度からドメインの開始残基位置と終了残基位置の候補を選択した. 開始残基位置の候補と終了残基位置の候補から, ドメインが重複しないように, 各ドメインの開始残基位置と終了残基位置を決める, この開始残基位置と終了残基位置で挟まれた領域が高保存配列領域で, かつ十分な長さがある場合, その配列領域をドメインとし, ドメインの探索を行った.

遺伝子配列の相同性に基づく本アルゴリズムにより, ドメインを探索することで, 解析対象とした生物種間で保存されているドメインの検出や配列パターンが未知なドメインについても検出可能である. また, 探索されたドメインだけでなく, 探索の過程にも重要な生物情報があるため, ドメインの探索過程で出るデータ及び遺伝子におけるドメインを可視化するソフトウェアの開発も行った.

本研究で開発したソフトウェアは, ある遺伝子におけるオーソログ遺伝子またはパラログ遺伝子の類似配列領域のデータを入力することで, ドメインの探索から可視化までを行う. 入力データのソフトウェアへのロードについては, ファイルかデータベースかを使用者が選択できる. ファイルの場合は使用者が独自に検出したオーソログ遺伝子・パラログ遺伝子のデータを規定の形式にそってファイルを作成し, ファイルを選ぶことでデータがロードされる. データベースの場合は, 本手法で検出したオーソログ遺伝子・パラログ遺伝子のデータから, 解析したい問い合せ遺伝子 (query gene) と対象生物 (subject speceis) を GUI の選択画面から選ぶことで必要なデータがロードされる. 可視化を行わず, ドメインの探索結

果をテキストファイル, 画像ファイルのみで出力することも可能である.

本研究では, 開発したソフトウェアを使い, ドメインの解析を行った. 始めに, 333 種全ての生物における全遺伝子についてドメインの探索を行い, 全遺伝子におけるマルチドメインとシングルドメインの比率を求めた. 全体の結果とシロイヌナズナの遺伝子におけるドメインの探索結果を比較することで, 高等植物であるシロイヌナズナでは, 微生物に比べ, マルチドメインとなる遺伝子が増えていることが確認できた. それゆえ, 微生物から植物への進化の過程で, ドメインが重要な役割を担っていると考え, 植物 (シロイヌナズナ) の遺伝子について, 微生物由来のドメインと植物固有のドメインの検出行った.

シロイヌナズナの全遺伝子について, 332 種の微生物とのオーソログ遺伝子とパラログ遺伝子を検出し, オーソログ遺伝子をもつシロイヌナズナの遺伝子から探索されたドメインは微生物由来のドメイン, パラログ遺伝子のみから探索されたドメインは植物固有のドメインとした. また, シロイヌナズナの遺伝子機能ごとに, 本手法で探索されたドメインを解析することで, 微生物由来のドメインを明らかにすることができた.

代謝機能の遺伝子からは, 微生物由来の異なるドメインが, 植物の遺伝子で bi-functional 遺伝子を形成していることが確認できた. 細胞内情報伝達機能/シグナル伝達機構の遺伝子からは, キナーゼのファミリーに注目することで, キナーゼの遺伝子で高く保存されているチロシンキナーゼドメイン, PAS ドメイン, アンキリンドメインが微生物由来であることがわかった. 膜輸送機能の遺伝子からは, ABC 輸送体のファミリーに注目することで, ABC 輸送体の遺伝子で高く保存されている NBFs ドメインが, 微生物由来であることがわかった. 逆に, 植物特異的であるといわれている機能に注目することで, 転写機能から, 異なる転写因子ファミリーの遺伝子間で, ドメインシャッフリングが起こり, 同生物内で, 遺伝子がマルチドメインを形成することがわかった. また, 遺伝子配列の系統分類の結果とドメイン構造を照らし合わせることで, 遺伝子は機能と構造に深い関係性があることがわかった.

これらのことは, 遺伝子について微生物から植物への進化の過程を解明する新たな知見を与えると考える. 今回の研究では, シロイヌナズナの遺伝子における

ドメインのみの解析を行ったが, 他の生物種の遺伝子に注目したドメイン解析をすることで, さらに生物種特異的なドメインやドメイン単位での遺伝子の進化についての理解に役立つと期待できる.

謝辞

本研究の遂行ならびに論文の作成に当たって, 情報科学研究科 金谷 重彦 教授, 黒川 顕 助教授, 並びに MD. Altaf-Ul-Amin 助手には常に親身になって多大な御指導, 御助言を請け賜りました. 心より感謝いたしております.

情報科学研究科 植村 俊亮 教授には, 総合的にご指導していただき深く感謝いたします.

そして, 本研究室の皆様方には研究生活そのものをあらゆる面で協力していただきました. 心よりお礼を申し上げます.

参考文献

[1] Wetlaufer DB. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci U S A. 70, 697-701.

[2] Efimov I, Kuusk V, Zhang X, McIntire WS. (1998) Proposed steady-state kinetic mechanism for Corynebacterium ammoniagenes FAD synthetase produced by Escherichia coli. Biochemistry. 37, 9716-9723.

[3] Mack M, van Loon AP, Hohmann HP. (1998) Regulation of riboflavin biosynthesis in Bacillus subtilis is affected by the activity of the flavokinase/flavin adenine dinucleotide synthetase encoded by ribC. J Bacteriol. 180, 950-955.

[4] Manstein DJ, Pai EF. (1986) Purification and characterization of FAD synthetase from Brevibacterium ammoniagenes. J Biol Chem. 261, 16169-16173.

[5] Mayhew SG, Wassink JH. (1980) Continuous fluorescence assay, partial purification and properties of flavokinase from Megasphaera elsdenii. Methods Enzymol. 66, 323-327.

[6] Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature. 402, 86-90.

[7] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science. 285, 751-753.

[8] Kuroda Y, Tani K, Matsuo Y, Yokoyama S. (2000) Automated search of natively folded protein fragments for high-throughput structure determination in structural genomics. Protein Sci. 9, 2313-2321.

[9] George RA, Heringa J. (2002) Protein domain identification and improved sequence similarity searching using PSI-BLAST. Proteins. 48, 672-681.

[10] Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. (2000) The Pfam protein families database. Nucleic Acids Res. 28, 263-266.

[11] Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226-D229.

[12] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. (2005) InterPro, progress and status in 2005. Nucleic Acids Res. 33, D201-D205.

[13] Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 34, D257-D260.

[14] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment serch tool. J Mol Biol 215, 403-410.

[15] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

[16] Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci. 7, 233-242.

[17] Curien G, Ravanel S, Robert M, Dumas R. (2005) Identification of six novel allosteric effectors of Arabidopsis thaliana aspartate kinase-homoserine de-

hydrogenase isoforms. Physiological context sets the specificity. J Biol Chem. 280, 41178-41183.

[18] Fujimori K, Ohta D. (1998) Isolation and characterization of a histidine biosynthetic gene in Arabidopsis encoding a polypeptide with two separate domains for phosphoribosyl-ATP pyrophosphohydrolase and phosphoribosyl-AMP cyclohydrolase. Plant Physiol. 118, 275-283.

[19] Fujimori K, Ohta D. (1998) An Arabidopsis cDNA encoding a bifunctional glutamine amidotransferase/cyclase suppresses the histidine auxotrophy of a Saccharomyces cerevisiae his7 mutant. FEBS Lett. 428, 229-234.

[20] Lazar G, Zhang H, Goodman HM. (1993) The origin of the bifunctional dihydrofolate reductase-thymidylate synthase isogenes of Arabidopsis thaliana. Plant J. 3, 657-668.

[21] Richmond TA, Bleecker AB. (1999) A defect in beta-oxidation causes abnormal inflorescence development in Arabidopsis. Plant Cell. 11, 1911-1923.

[22] Bolognese CP, McGraw P. (2000) The isolation and characterization in yeast of a gene for Arabidopsis S-adenosylmethionine:phospho-ethanolamine N-methyltransferase. Plant Physiol. 124, 1800-1813.

[23] Sandoval FJ, Roje S. (2005) An FMN hydrolase is fused to a riboflavin kinase homolog in plants. J Biol Chem. 280, 38337-38345.

[24] Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science. 290, 2015-2110.

[25] Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K. (2005) RARTF: Database and Tools for Complete Sets of

Arabidopsis Transcription Factors. DNA Res. 12, 247-256.

[26] Hwang I, Chen HC, Sheen J. (2002) Two-component signal transduction pathways in Arabidopsis. Plant Physiol. 129, 500-515.

[27] Bogre L, Okresz L, Henriques R, Anthony RG. (2003) Growth signalling pathways in Arabidopsis and the AGC protein kinases. Trends Plant Sci. 8, 424-431.

[28] Rudrabhatla P, Reddy MM, Rajasekharan R. (2006) Genome-wide analysis and experimentation of plant serine/ threonine/tyrosine-specific protein kinases. Plant Mol Biol. 60, 293-319.

[29] Sanchez-Fernandez R, Davies TG, Coleman JO, Rea PA. (2001)The Arabidopsis thaliana ABC protein superfamily, a complete inventory. J Biol Chem. 276, 30231-30244.

[30] Maser P, Thomine S, Schroeder JI, Ward JM, Hirschi K, Sze H, Talke IN, Amtmann A, Maathuis FJ, Sanders D, Harper JF, Tchieu J, Gribskov M, Persans MW, Salt DE, Kim SA, Guerinot ML. (2001) Phylogenetic relationships within cation transporter families of Arabidopsis. Plant Physiol. 126, 1646-1667.

付録

A. 解析に用いた生物の遺伝子数とaccession number

kingdom	phylum	species	accession number	no. of gene
Viridiplantae	Streptophyta	Arabidopsis thaliana	NC_003070	6,836
			NC_003071	4,164
			NC_003074	5,286
			NC_003075	4,057
			NC_003076	6,193

kingdom	phylum	species	accession number	no. of gene
Fungi	Ascomycota	Saccharomyces cerevisiae	NC_001133	94
			NC_001134	406
			NC_001135	160
			NC_001136	755
			NC_001137	273
			NC_001138	126
			NC_001139	526
			NC_001140	281
			NC_001141	207
			NC_001142	356
			NC_001143	312
			NC_001144	508
			NC_001145	460
			NC_001146	393
			NC_001147	536
			NC_001148	461
			NC_001224	19
		Schizosaccharomyces pombe	NC_003421	887
			NC_003423	1,796
			NC_003424	2,235

kingdom	phylum	species	accession number	no. of gene
Archaea	Nanoarchaeota	Nanoarchaeum equitans	NC_005213	536
	Crenarchaeota	Pyrobaculum aerophilum	NC_003364	2,605
		Sulfolobus acidocaldarius DSM 639	NC_007181	2,223
		Sulfolobus solfataricus	NC_002754	2,977
		Sulfolobus tokodaii	NC_003106	2,825
		Aeropyrum pernix	NC_000854	1,841
	Euryarchaeota	Archaeoglobus fulgidus	NC_000917	2,420
		Halobacterium sp	NC_001869	176
			NC_002607	2,075
			NC_002608	371
		Natronomonas pharaonis	NC_007426	2,661
			NC_007427	125
			NC_007428	36
		Haloarcula marismortui ATCC 43049	NC_006389	36
			NC_006390	42
			NC_006391	40
			NC_006392	51
			NC_006393	131
			NC_006394	166
			NC_006395	362
			NC_006396	3,131
			NC_006397	281
		Thermoplasma acidophilum	NC_002578	1,482
		Thermoplasma volcanium	NC_002689	1,499
		Picrophilus torridus DSM 9790	NC_005877	1,535
		Pyrococcus abyssi	NC_000868	1,896
			NC_001773	2
		Pyrococcus furiosus	NC_003413	2,125
		Pyrococcus horikoshii	NC_000961	1,955
		Thermococcus kodakaraensis KOD1	NC_006624	2,306
		Methanopyrus kandleri	NC_003551	1,687
		Methanospirillum hungatei JF-1	NC_007796	3,139
		Methanosarcina acetivorans	NC_003552	4,540
		Methanosarcina barkeri fusaro	NC_007349	18
			NC_007355	3,606
		Methanosarcina mazei	NC_003901	3,370
		Methanococcoides burtonii DSM 6242	NC_007955	2,273
		Methanosphaera stadtmanae	NC_007681	1,534
		Methanobacterium thermoautotrophicum	NC_000916	1,873
		Methanococcus maripaludis S2	NC_005791	1,722
		Methanococcus jannaschii	NC_000909	1,729
			NC_001732	45
			NC_001733	12

kingdom	phylum	species	accession number	no. of gene
Bacteria	Fusobacteria	Fusobacterium nucleatum	NC_003454	2,067
	Planctomycetes	Pirellula sp	NC_005027	7,325
	Aquificae	Aquifex aeolicus	NC_000918	1,529
			NC_001880	31
	Thermotogae	Thermotoga maritima	NC_000853	1,858
	Chloroflexi	Dehalococcoides ethenogenes 195	NC_002936	1,580
		Dehalococcoides CBDB1	NC_007356	1,458
	Chlorobi	Chlorobium tepidum TLS	NC_002932	2,252
		Chlorobium chlorochromatii CaD3	NC_007514	2,002
		Pelodictyon luteolum DSM 273	NC_007512	2,083
	Deinococcus-Thermus	Thermus thermophilus HB27	NC_005835	1,982
			NC_005838	228
		Thermus thermophilus HB8	NC_006461	1,973
			NC_006462	251
			NC_006463	14
		Deinococcus radiodurans	NC_000958	146
			NC_000959	39
			NC_001263	2,629
			NC_001264	368
	Bacteroidetes	Salinibacter ruber DSM 13855	NC_007677	2,801
			NC_007678	32
		Porphyromonas gingivalis W83	NC_002950	1,909
		Bacteroides fragilis YCH46	NC_006297	47
			NC_006347	4,578
		Bacteroides fragilis NCTC 9434	NC_003228	4,184
			NC_006873	47
		Bacteroides thetaiotaomicron VPI-5482	NC_004663	4,778
			NC_004703	38

kingdom	phylum	species	accession number	no. of gene
Bacteria	Spirochaetes	Leptospira interrogans serovar Copenhageni	NC_005823	3,394
			NC_005824	264
		Leptospira interrogans serovar Lai	NC_004342	4,360
			NC_004343	367
		Treponema pallidum	NC_000919	1,036
		Borrelia garinii PBi	NC_006128	26
			NC_006129	74
			NC_006156	832
		Borrelia burgdorferi	NC_000948	42
			NC_000949	45
			NC_000950	44
			NC_000951	42
			NC_000952	44
			NC_000953	43
			NC_000954	35
			NC_000955	11
			NC_000956	72
			NC_000957	6
			NC_001318	851
			NC_001849	23
			NC_001850	29
			NC_001851	26
			NC_001852	32
			NC_001853	37
			NC_001854	43
			NC_001855	51
			NC_001856	48
			NC_001857	76
			NC_001903	29
			NC_001904	11
		Treponema denticola ATCC 35405	NC_002967	2,767

kingdom	phylum	species	accession number	no. of gene
Bacteria	Chlamydiae	Parachlamydia sp UWE25	NC_005861	2,031
		Chlamydia muridarum	NC_002182	7
			NC_002620	904
		Chlamydia trachomatis	NC_000117	895
		Chlamydia trachomatis A HAR-13	NC_007429	911
			NC_007430	8
		Chlamydophila abortus S26 3	NC_004552	932
		Chlamydophila caviae	NC_003361	998
			NC_004720	7
		Chlamydophila felis Fe C-56	NC_007899	1,005
			NC_007900	8
		Chlamydophila pneumoniae AR39	NC_002179	1,112
		Chlamydophila pneumoniae CWL029	NC_000922	1,052
		Chlamydophila pneumoniae J138	NC_002491	1,069
		Chlamydophila pneumoniae TW 183	NC_005043	1,113
	Cyanobacteria	Anabaena variabilis ATCC 29413	NC_007410	344
			NC_007411	31
			NC_007412	243
			NC_007413	5,039
		Nostoc sp	NC_003240	186
			NC_003241	5
			NC_003267	90
			NC_003270	31
			NC_003272	5,366
			NC_003273	66
			NC_003276	386
		Cyanobacteria bacterium Yellowstone A-Prime	NC_007775	2,760
		Cyanobacteria bacterium Yellowstone B-Prime	NC_007776	2,862
		Synechococcus CC9605	NC_007516	2,638
		Synechococcus CC9902	NC_007513	2,304
		Synechococcus elongatus PCC 6301	NC_006576	2,525
		Synechococcus elongatus PCC 7942	NC_007595	50
			NC_007604	2,611
		Synechococcus sp WH8102	NC_005070	2,517
		Synechocystis PCC6803	NC_000911	3,167
			NC_005229	132
			NC_005230	106
			NC_005231	49
			NC_005232	110
		Thermosynechococcus elongatus	NC_004113	2,475
		Prochlorococcus marinus MED4	NC_005072	1,712
		Prochlorococcus marinus MIT9313	NC_005071	2,265
		Prochlorococcus marinus MIT 9312	NC_007577	1,809
		Prochlorococcus marinus NATL2A	NC_007335	1,890
		Prochlorococcus marinus CCMP1375	NC_005042	1,882
		Gloeobacter violaceus	NC_005125	4,430

kingdom	phylum	species	accession number	no. of gene
Bacteria	Actinobacteria	Symbiobacterium thermophilum IAM14863	NC_006177	3,337
		Corynebacterium diphtheriae	NC_002935	2,272
		Corynebacterium efficiens YS-314	NC_004369	2,950
		Corynebacterium glutamicum ATCC 13032 Bielefeld	NC_006958	3,057
		Corynebacterium glutamicum ATCC 13032 Kitasato	NC_003450	2,993
		Corynebacterium jeikeium K411	NC_003080	16
			NC_007164	2,104
		Nocardia farcinica IFM10152	NC_006361	5,683
			NC_006362	160
			NC_006363	93
		Tropheryma whipplei Twist	NC_004572	808
		Tropheryma whipplei TW08 27	NC_004551	783
		Leifsonia xyli xyli CTCB0	NC_006087	2,030
		Streptomyces coelicolor	NC_003888	7,769
			NC_003903	351
			NC_003904	34
		Streptomyces avermitilis	NC_003155	7,577
			NC_004719	96
		Thermobifida fusca YX	NC_007333	3,110
		Mycobacterium avium paratuberculosis	NC_002944	4,350
		Mycobacterium bovis	NC_002945	3,920
		Mycobacterium leprae	NC_002677	1,605
		Mycobacterium tuberculosis CDC1551	NC_002755	4,189
		Mycobacterium tuberculosis H37Rv	NC_000962	3,989
		Propionibacterium acnes KPA171202	NC_006085	2,297
		Frankia CcI3	NC_007777	4,499
		Bifidobacterium longum	NC_004307	1,727
			NC_004943	2

kingdom	phylum	species	accession number	no. of gene
Bacteria	Firmicutes	Aster yellows witches-broom phytoplasma AYWB	NC_007716	671
			NC_007717	5
			NC_007718	4
			NC_007719	7
			NC_007720	6
		Onion yellows phytoplasma	NC_005303	754
		Mycoplasma capricolum ATCC 27343	NC_007633	812
		Mycoplasma gallisepticum	NC_004829	726
		Mycoplasma genitalium	NC_000908	484
		Mycoplasma hyopneumoniae 232	NC_006360	691
		Mycoplasma hyopneumoniae 7448	NC_007332	663
		Mycoplasma hyopneumoniae J	NC_007295	665
		Mycoplasma mobile 163K	NC_006908	633
		Mycoplasma mycoides	NC_005364	1,016
		Mycoplasma penetrans	NC_004432	1,037
		Mycoplasma pneumoniae	NC_000912	689
		Mycoplasma pulmonis	NC_002771	782
		Mycoplasma synoviae 53	NC_007294	672
		Ureaplasma urealyticum	NC_002162	614
		Mesoplasma florum L1	NC_006055	682
		Thermoanaerobacter tengcongensis	NC_003869	2,588
		Moorella thermoacetica ATCC 39073	NC_007644	2,465
		Carboxydothermus hydrogenoformans Z-2901	NC_007503	2,620
		Desulfitobacterium hafniense Y51	NC_007907	5,060
		Clostridium perfringens	NC_003042	63
			NC_003366	2,660
		Clostridium tetani E88	NC_004557	2,373
			NC_004565	59
		Clostridium acetobutylicum	NC_001988	176
			NC_003030	3,672

kingdom	phylum	species	accession number	no. of gene
Bacteria	Firmicutes	Listeria innocua	NC_003212	2,968
			NC_003383	75
		Listeria monocytogenes	NC_003210	2,846
		Listeria monocytogenes 4b F2365	NC_002973	2,821
		Staphylococcus aureus COL	NC_002951	2,615
			NC_006629	3
		Staphylococcus aureus MW2	NC_003923	2,632
		Staphylococcus aureus Mu50	NC_002758	2,697
			NC_002774	34
		Staphylococcus aureus N315	NC_002745	2,588
			NC_003140	31
		Staphylococcus aureus NCTC 8325	NC_007795	2,892
		Staphylococcus aureus RF122	NC_007622	2,515
		Staphylococcus aureus USA300	NC_007790	5
			NC_007791	3
			NC_007792	36
			NC_007793	2,560
		Staphylococcus aureus aureus MRSA252	NC_002952	2,656
		Staphylococcus aureus aureus MSSA476	NC_002953	2,579
			NC_005951	19
		Staphylococcus epidermidis ATCC 12228	NC_004461	2,419
			NC_005003	11
			NC_005004	22
			NC_005005	16
			NC_005006	8
			NC_005007	6
			NC_005008	3
		Staphylococcus epidermidis RP62A	NC_002976	2,494
			NC_006663	32
		Staphylococcus haemolyticus	NC_007168	2,676
		Staphylococcus saprophyticus	NC_007350	2,446
			NC_007351	45
			NC_007352	23

kingdom	phylum	species	accession number	no. of gene
Bacteria	Firmicutes	Bacillus anthracis Ames	NC_003997	5,311
		Bacillus anthracis Ames 0581	NC_007322	204
			NC_007323	104
			NC_007530	5,309
		Bacillus anthracis str Sterne	NC_005945	5,287
		Bacillus cereus ATCC14579	NC_004721	21
			NC_004722	5,234
		Bacillus cereus ATCC 10987	NC_003909	5,603
			NC_005707	241
		Bacillus cereus ZK	NC_006274	5,134
			NC_007103	430
			NC_007104	5
			NC_007105	54
			NC_007106	8
			NC_007107	10
		Bacillus clausii KSM-K16	NC_006582	4,096
		Bacillus halodurans	NC_002570	4,066
		Bacillus licheniformis ATCC 14580	NC_006270	4,152
		Bacillus licheniformis DSM 13	NC_006322	4,196
		Bacillus thuringiensis konkukian	NC_005957	5,117
			NC_006578	80
		Bacillus subtilis	NC_000964	4,105
		Oceanobacillus iheyensis	NC_004193	3,500
		Geobacillus kaustophilus HTA426	NC_006509	42
			NC_006510	3,498

kingdom	phylum	species	accession number	no. of gene
Bacteria	Firmicutes	Lactobacillus acidophilus NCFM	NC_006814	1,864
		Lactobacillus johnsonii NCC 533	NC_005362	1,821
		Lactobacillus plantarum	NC_004567	3,009
			NC_006375	3
			NC_006376	4
			NC_006377	43
		Lactobacillus sakei 23K	NC_007576	1,879
		Lactobacillus salivarius UCC118	NC_006529	27
			NC_006530	51
			NC_007929	1,717
			NC_007930	222
		Streptococcus agalactiae 2603	NC_004116	2,124
		Streptococcus agalactiae A909	NC_007432	1,996
		Streptococcus agalactiae NEM316	NC_004368	2,094
		Streptococcus mutans	NC_004350	1,960
		Streptococcus pneumoniae R6	NC_003098	2,043
		Streptococcus pneumoniae TIGR4	NC_003028	2,094
		Streptococcus pyogenes M1 GAS	NC_002737	1,697
		Streptococcus pyogenes MGAS10394	NC_006086	1,886
		Streptococcus pyogenes MGAS315	NC_004070	1,865
		Streptococcus pyogenes MGAS5005	NC_007297	1,865
		Streptococcus pyogenes MGAS6180	NC_007296	1,894
		Streptococcus pyogenes MGAS8232	NC_003485	1,845
		Streptococcus pyogenes SSI-1	NC_004606	1,861
		Streptococcus thermophilus CNRZ1066	NC_006449	1,915
		Streptococcus thermophilus LMG 18311	NC_006448	1,889
		Lactococcus lactis	NC_002662	2,321
		Enterococcus faecalis V583	NC_004668	3,113
			NC_004669	72
			NC_004670	18
			NC_004671	62

kingdom	phylum	species	accession number	no. of gene
Bacteria	Proteobacteria	Chromohalobacter salexigens DSM 3043	NC_007963	3,298
		Hahella chejuensis KCTC 2396	NC_007645	6,778
		Francisella tularensis tularensis	NC_006570	1,603
		Francisella tularensis holarctica	NC_007880	1,754
		Thiomicrospira crunogena XCL-2	NC_007520	2,192
		Idiomarina loihiensis L2TR	NC_006512	2,628
		Saccharophagus degradans 2-40	NC_007912	4,008
		Pseudoalteromonas haloplanktis TAC125	NC_007481	2,940
			NC_007482	546
		Shewanella denitrificans OS217	NC_007954	3,754
		Shewanella oneidensis	NC_004347	4,324
			NC_004349	148
		Colwellia psychrerythraea 34H	NC_003910	4,910
		Vibrio cholerae	NC_002505	2,742
			NC_002506	1,093
		Vibrio fischeri ES114	NC_006840	2,575
			NC_006841	1,172
			NC_006842	55
		Vibrio parahaemolyticus	NC_004603	3,080
			NC_004605	1,752
		Vibrio vulnificus CMCP6	NC_004459	2,926
			NC_004460	1,562
		Vibrio vulnificus YJ016	NC_005128	69
			NC_005139	3,259
			NC_005140	1,696
		Photobacterium profundum SS9	NC_005871	67
			NC_006370	3,416
			NC_006371	2,008
		Xanthomonas campestris	NC_003902	4,181
		Xanthomonas campestris 8004	NC_007086	4,273
		Xanthomonas campestris vesicatoria 85-10	NC_007504	2
			NC_007505	22
			NC_007506	43
			NC_007507	172
			NC_007508	4,487
		Xanthomonas citri	NC_003919	4,312
			NC_003921	42
			NC_003922	73
		Xanthomonas oryzae KACC10331	NC_006834	4,080
		Xylella fastidiosa	NC_002488	2,766
			NC_002489	2
			NC_002490	64
		Xylella fastidiosa Temecula1	NC_004554	2
			NC_004556	2,034

kingdom	phylum	species	accession number	no. of gene
Bacteria	Proteobacteria	Nitrosococcus oceani ATCC 19707	NC_007483	43
			NC_007484	2,974
		Methylococcus capsulatus Bath	NC_002977	2,960
		Pseudomonas aeruginosa	NC_002516	5,567
		Pseudomonas fluorescens Pf-5	NC_004129	6,137
		Pseudomonas fluorescens PfO-1	NC_007492	5,736
		Pseudomonas putida KT2440	NC_002947	5,350
		Pseudomonas syringae	NC_004578	5,470
			NC_004632	70
			NC_004633	67
		Pseudomonas syringae phaseolicola 1448A	NC_005773	4,983
			NC_007274	127
			NC_007275	60
		Pseudomonas syringae pv B728a	NC_007005	5,089
		Psychrobacter arcticum 273-4	NC_007204	2,120
		Acinetobacter sp ADP1	NC_005966	3,325
		Mannheimia succiniciproducens MBEL55E	NC_006300	2,380
		Pasteurella multocida	NC_002663	2,015
		Haemophilus influenzae	NC_000907	1,657
		Haemophilus influenzae 86 028NP	NC_007146	1,791
		Haemophilus ducreyi 35000HP	NC_002940	1,717
		Legionella pneumophila Lens	NC_006366	56
			NC_006369	2,878
		Legionella pneumophila Paris	NC_006365	139
			NC_006368	3,027
		Legionella pneumophila Philadelphia 1	NC_002942	2,942
		Coxiella burnetii	NC_002971	2,016
			NC_004704	36
		Buchnera aphidicola	NC_004545	504
		Buchnera aphidicola Sg	NC_004061	546
		Buchnera sp	NC_002252	3
			NC_002253	7
			NC_002528	564
		Candidatus Blochmannia floridanus	NC_005061	583
		Candidatus Blochmannia pennsylvanicus BPEN	NC_007292	610
		Photorhabdus luminescens	NC_005126	4,683
		Erwinia carotovora atroseptica SCRI1043	NC_004547	4,472
		Shigella boydii Sb227	NC_007608	148
			NC_007613	4136
		Shigella dysenteriae	NC_007606	4,274
			NC_007607	223
		Shigella flexneri 2a	NC_004337	4,182
			NC_004851	261
		Shigella flexneri 2a 2457T	NC_004741	4,068
		Shigella sonnei Ss046	NC_007384	4,223
			NC_007385	238

kingdom	phylum	species	accession number	no. of gene
Bacteria	Proteobacteria	Salmonella enterica Choleraesuis	NC_006855	51
			NC_006856	170
			NC_006905	4,441
		Salmonella enterica Paratypi ATCC 9150	NC_006511	4,093
		Salmonella typhi	NC_003198	4,395
			NC_003384	235
			NC_003385	128
		Salmonella typhi Ty2	NC_004631	4,318
		Salmonella typhimurium LT2	NC_003197	4,425
			NC_003277	102
		Yersinia pestis CO92	NC_003131	72
			NC_003132	9
			NC_003134	101
			NC_003143	3,885
		Yersinia pestis KIM	NC_004088	4,086
			NC_004838	116
		Yersinia pestis biovar Mediaevails	NC_005810	3,895
			NC_005813	85
			NC_005814	30
			NC_005815	122
			NC_005816	10
		Yersinia pseudotuberculosis IP32953	NC_006153	95
			NC_006154	42
			NC_006155	3,901
		Sodalis glossinidius morsitans	NC_007712	2,432
			NC_007713	54
			NC_007714	23
			NC_007715	7
		Wigglesworthia brevipalpis	NC_003425	6
			NC_004344	611
		Escherichia coli CFT073	NC_004431	5,379
		Escherichia coli O157H7	NC_002127	3
			NC_002128	85
			NC_002695	5,253
		Escherichia coli O157H7 EDL933	NC_002655	5,324
		Escherichia coli UTI89	NC_007941	145
			NC_007946	5,066
		Escherichia coli W3110	AC_000091	4,227
		Escherichia coli K12	NC_000913	4,237

kingdom	phylum	species	accession number	no. of gene
Bacteria	Proteobacteria	Erythrobacter litoralis HTCC2594	NC_007722	3,011
		Zymomonas mobilis ZM4	NC_006526	1,998
		Novosphingobium aromaticivorans DSM 12444	NC_007794	3,324
		Rhodobacter sphaeroides 2 4 1	NC_007488	100
			NC_007489	82
			NC_007490	87
			NC_007493	3,022
			NC_007494	835
		Silicibacter pomeroyi DSS-3	NC_003911	3,810
			NC_006569	442
		Jannaschia CCS1	NC_007801	71
			NC_007802	4,212
		Magnetospirillum magneticum AMB-1	NC_007626	4,559
		Rhodospirillum rubrum ATCC 11170	NC_007641	50
			NC_007643	3,791
		Gluconobacter oxydans 621H	NC_006672	163
			NC_006673	29
			NC_006674	18
			NC_006675	18
			NC_006676	4
			NC_006677	2,432
		Caulobacter crescentus	NC_002696	3,737
		Ehrlichia canis Jake	NC_007354	925
		Ehrlichia ruminantium Gardel	NC_006831	950
		Ehrlichia ruminantium Welgevonden	NC_005295	888
		Ehrlichia ruminantium str. Welgevonden	NC_006832	958
		Ehrlichia chaffeensis Arkansas	NC_007799	1,105
		Neorickettsia sennetsu Miyayama	NC_007798	932
		Candidatus Pelagibacter ubique HTCC1062	NC_007205	1,354
		Wolbachia endosymbiont of Brugia malayi TRS	NC_006833	805
		Wolbachia endosymbiont of Drosophila melanogaster	NC_002978	1,195
		Rickettsia bellii RML369-C	NC_007940	1,429
		Rickettsia conorii	NC_003103	1,374
		Rickettsia felis URRWXCal2	NC_007109	1,400
			NC_007110	68
			NC_007111	44
		Rickettsia prowazekii	NC_000963	835
		Rickettsia typhi wilmington	NC_006142	838

kingdom	phylum	species	accession number	no. of gene
Bacteria	Proteobacteria	Brucella abortus 9-941	NC_006932	2,030
			NC_006933	1,055
		Brucella melitensis	NC_003317	2,059
			NC_003318	1,139
		Brucella melitensis biovar Abortus	NC_007618	2,000
			NC_007624	1,034
		Brucella suis 1330	NC_004310	2,123
			NC_004311	1,148
		Mesorhizobium loti	NC_002678	6,743
			NC_002679	320
			NC_002682	209
		Nitrobacter hamburgensis X14	NC_007959	239
			NC_007960	172
			NC_007961	111
			NC_007964	3,804
		Nitrobacter winogradskyi Nb-255	NC_007406	3,122
		Rhodopseudomonas palustris BisB18	NC_007925	4,886
		Rhodopseudomonas palustris BisB5	NC_007958	4,397
		Rhodopseudomonas palustris CGA009	NC_005296	4,813
			NC_005297	7
		Rhodopseudomonas palustris HaA2	NC_007778	4,683
		Bradyrhizobium japonicum	NC_004463	8,317
		Bartonella quintana Toulouse	NC_005955	1,142
		Bartonella henselae Houston-1	NC_005956	1,488
		Agrobacterium tumefaciens C58 Cereon	NC_003062	2,715
			NC_003063	1,833
			NC_003064	547
			NC_003065	193
		Agrobacterium tumefaciens C58 UWash	NC_003304	2,785
			NC_003305	1,876
			NC_003306	543
			NC_003308	198
		Sinorhizobium meliloti	NC_003037	1,294
			NC_003047	3,341
			NC_003078	1,570
		Rhizobium etli CFN 42	NC_007761	4,035
			NC_007762	175
			NC_007763	163
			NC_007764	232
			NC_007765	455
			NC_007766	567
		Anaplasma marginale St Maries	NC_004842	949
		Anaplasma phagocytophilum HZ	NC_007797	1,264

kingdom	phylum	species	accession number	no. of gene
Bacteria	Proteobacteria	Ralstonia eutropha JMP134	NC_007336	512
			NC_007337	88
			NC_007347	3,439
			NC_007348	2,407
		Ralstonia solanacearum	NC_003295	3,440
			NC_003296	1,676
		Burkholderia 383	NC_007509	1,209
			NC_007510	3,334
			NC_007511	3,174
		Burkholderia mallei ATCC 23344	NC_006348	2,996
			NC_006349	2,029
		Burkholderia pseudomallei 1710b	NC_007434	3,736
			NC_007435	2,611
		Burkholderia pseudomallei K96243	NC_006350	3,399
			NC_006351	2,329
		Burkholderia thailandensis E264	NC_007650	2,358
			NC_007651	3,276
		Burkholderia xenovorans LB400	NC_007951	4,430
			NC_007952	2,960
			NC_007953	1,312
		Rhodoferax ferrireducens DSM 15236	NC_007901	248
			NC_007908	4,170
		Polaromonas JS666	NC_007948	4,817
			NC_007949	326
			NC_007950	310
		Bordetella bronchiseptica	NC_002927	4,994
		Bordetella pertussis	NC_002929	3,436
		Bordetella parapertussis	NC_002928	4,185
		Nitrosospira multiformis ATCC 25196	NC_007614	2,757
			NC_007615	17
			NC_007616	16
			NC_007617	15
		Nitrosomonas europaea	NC_004757	2,461
		Thiobacillus denitrificans ATCC 25259	NC_007404	2,827
		Methylobacillus flagellatus KT	NC_007947	2,753
		Neisseria gonorrhoeae FA 1090	NC_002946	2,002
		Neisseria meningitidis MC58	NC_003112	2,063
		Neisseria meningitidis Z2491	NC_003116	2,065
		Chromobacterium violaceum	NC_005085	4,407
		Dechloromonas aromatica RCB	NC_007298	4,171
		Azoarcus sp EbN1	NC_006513	4,133
			NC_006823	272
			NC_006824	194

kingdom	phylum	species	accession number	no. of gene
Bacteria	Proteobacteria	Bdellovibrio bacteriovorus	NC_005363	3,587
		Desulfovibrio vulgaris Hildenborough	NC_002937	3,379
			NC_005863	152
		Desulfovibrio desulfuricans G20	NC_007519	3,775
		Pelobacter carbinolicus	NC_007498	3,119
		Geobacter sulfurreducens	NC_002939	3,446
		Geobacter metallireducens GS-15	NC_007515	13
			NC_007517	3,519
		Desulfotalea psychrophila LSv54	NC_006138	3,116
			NC_006139	101
			NC_006140	17
		Syntrophus aciditrophicus SB	NC_007759	3,168
		Anaeromyxobacter dehalogenans 2CP-C	NC_007760	4,346
		Thiomicrospira denitrificans ATCC 33889	NC_007575	2,097
		Wolinella succinogenes	NC_005090	2,043
		Helicobacter hepaticus	NC_004917	1,875
		Helicobacter pylori 26695	NC_000915	1,576
		Helicobacter pylori J99	NC_000921	1,491
		Campylobacter jejuni	NC_002163	1,629
		Campylobacter jejuni RM1221	NC_003912	1,838