Vol.2011-MPS-86 No.11
Vol.2011-BIO-27 No.11
2011/12/1

符号理論を用いた薬剤耐性菌の金属トランスポーター遺伝子の同定

佐藤　勲興[†]　權 娟大[††]　宮崎 智[††]

グラム陰性桿菌である *Stenotrophomonas maltophilia* は、抗生物質の *β*-lactam 環を破壊する *β*-ラクタマーゼ酵素を産生することで、抗生物質の効果を無力化する。生化学的実験によって、*S.maltophilia* の薬剤耐性における重金属の取り込みに関連したトランスポーター遺伝子の重金属取り込み機構が、酵素の活性化と深く関係していることが明らかになっている。しかし、ほとんどの金属トランスポーター遺伝子に対しては、ホモロジー検索やモチーフ検索のような一般的なバイオインフォマティクス手法を適用することができない。そこで本研究では、情報理論の１つである符号理論を応用し、トランスポーター遺伝子の膜貫通領域の親水性アミノ酸を符号理論における冗長部分とみなすことにより、金属トランスポーター遺伝子の同定を試みた。

Identification of Metal Transporter Genes Using Coding Theory

Hiroki Sato[†], Yeondae Kwon[††] and Satoru Miyazaki[††]

Stenotrofomonas maltophilia, gram-negative bacteria, produces the enzyme beta-lactamase which breaks open the beta-lactam ring of the antibiotic, rendering the antibiotic ineffective, and hence, has a remarkable capacity for drug and heavy metal resistance. Biochemical experiments have revealed that there is a relationship between the enzyme content and the metal content and that the uptake of metal ions by transporter genes is associated with the resistance against metal ions of *S.maltophilia*. However, few metal transporter genes make general bioinformatics tools such as homology and motif search inappropriate. In this study, we assume hydrophilic amino acids in hydrophobic transmembrane domains of a transporter gene to be redundant amino acids and present a novel method for identifying metal transporter genes with coding theory.

1. 背景・目的

薬剤耐性菌 *Stenotrophomonas maltophilia*（以下、*S.maltophilia*）は γ-proteobacteria に属し、自然環境や臨床に幅広く生息している[1]。*S.maltophilia* は、幅広い抗菌スペクトラムの β-lactam 系抗生物質に対する耐性を有し、その耐性は活性中心に亜鉛を持つ特有の L1 型 metallo-β-lactamase によるものである。この酵素が活性化するとき、細胞内の亜鉛の濃度は通常の数十倍になることが観測され、亜鉛の取り込みと β-lactam 系抗生物質に対する耐性機構の関連性が生化学的実験により示唆された。重金属は、生物の生命維持には必要不可欠でありながら、過剰な濃度は有害であるため、細胞ではその濃度の微妙な調節を巧みな機構で行っている。その機構の中心はトランスポーター遺伝子であり、トランスポーター遺伝子は、様々な生物において多種多様な排出、取り込みに関わっている[2]。そして、*S.maltophilia* においても特に耐性機構と関わりのある亜鉛トランスポーターの存在が示唆された。しかし、*S.maltophilia* の全ゲノム情報はわかっているものの、トランスポーター遺伝子は未だに同定されていない。

トランスポーター遺伝子が同定されていない理由として、トランスポーター遺伝子の特徴が挙げられる。一般的なバイオインフォマティクスの機能予測手法として、塩基配列やアミノ酸配列の相同性検索、配列モチーフを利用した機能予測手法がある。しかし、これらの手法は配列相同性を基に探すことで機能を特徴づけるため、ファミリー間の配列相同性が低く、繰り返し配列を多く含むトランスポーター遺伝子に用いることができない。そこで、本研究では *S.maltophilia* の重金属トランスポーター遺伝子を同定することを目的に、情報理論の一つである符号理論を応用した配列類似性によらない新規のバイオインフォマティクス手法を開発した。解析事例として、薬剤耐性菌で未同定の重金属トランスポータータンパク質の遺伝子同定を試みた。

2. 準備

2.1 符号理論

情報を送り手が受け手に送る際に、通信路上での雑音による情報の誤りを検出し、訂正できるような機能を持たせるため送る情報に冗長性を与える。この情報に冗長性を加える作業を符号化という。そして誤りを検出し、訂正することをシンドローム（誤り訂正検査能力）という。符号理論とは、情報を符号化して通信を行う際の効率と信頼性を向上させる理論である。

[†] 東京理科大学大学院　薬学研究科　薬科学専攻
Graduate School of Pharmaceutical Sciences, Tokyo university of Science
[††] 東京理科大学　薬学部　生命創薬科学科
Faculty of Pharmaceutical Sciences, Tokyo University of Science

Vol.2011-MPS-86 No.11
Vol.2011-BIO-27 No.11
2011/12/1

2.2 ガロア体

符号理論の数学的基礎をなすガロア体について述べる。

集合 F において、加減乗除の四則演算が可能なとき、この集合を体と呼ぶ。さらに、この集合 F が有限のとき、これをガロア体と呼ぶ。元の個数が q のとき、ガロア体を $GF(q)$と表し、特に q が素数であるガロア体を素体と呼ぶ。この素体の和及び積の演算は通常の整数演算を行い、q よりも大きい数になったとき、q で余剰をとることによって求められる。

本研究では、ガロア体の二次拡大体を用い、塩基配列を構成する 4 つの塩基 A、T、G、C を 0、1、2、3 と置換することで数量化した（表 1）。

表 1　$GF(2)$の二次拡大体

+	0	1	2	3
0	0	1	2	3
1	1	0	3	2
2	2	3	0	1
3	3	2	1	0

×	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	3	1
3	0	3	1	2

2.3 新規手法開発の必要性

多くのゲノム情報が公開され、その配列情報がデータベース上に登録されつつある。しかし、*S.maltophilia* の金属トランスポーター遺伝子の配列情報は未だにデータベース上に登録されていない。その主な理由として、トランスポーター遺伝子特有の配列の特徴により、同定するのが難しいというのが挙げられる。

遺伝子とその産物の機能を特定するバイオインフォマティクスの代表的な同定手法に、相同性検索とモチーフ配列を利用した機能予測法がある。相同性検索は、ターゲットとしているタンパク質に類似した機能既知タンパク質を、配列相同性を基に探すことで機能を特徴づけることができる。しかし、この手法はトランスポーター遺伝子のファミリー間の配列相同性の低さのために、類似タンパク質を検索することが困難である。配列モチーフは類似の機能をもつ相同なタンパク質の間で保存されている塩基またはアミノ酸配列パターンであり、特定の機能や構造と強く関連していることが多い。一般的な配列モチーフ探索方法として、マルチプルシークエンスアラインメントなどから、配列間で保存された領域を見出す方法がとられてきた。しかし、この手法もトランスポーター遺伝子はファミリー間での相同性が低い、配列長が大きく異なる、繰り返し配列が含まれるといった理由から既存の方法ではうまくアラインメントできず、モチーフがうまく抽出できない欠点がある。

数多くの情報理論から符号理論を選択した理由として、遺伝子と符号間の類似性、またトランスポーター遺伝子と符号間との類似性が挙げられる。符号理論は雑音のある通信路を通して情報を送る際に、情報が損なわれないために検査記号と呼ばれる冗長性を付加するための理論である。遺伝子においても符号に似た冗長性が数多く存在する。例えば、DNA は 4 つの塩基 A、T、G、C から構成されている。本来、3 つの塩基（コドン）が 1 つのアミノ酸に対応しているため、アミノ酸の種類は 4^3=64 通りあるはずであるが、タンパク質を構成する主要なアミノ酸は 20 種類程しかない。コドンとアミノ酸の対応表より、コドンの第 3 番目の塩基に変異が起きても、アミノ酸は変化しない場合が多いことから、第 3 番目の塩基はアミノ酸指定に重要な役割を果たしていないことが分かる。このことから、コドンの 3 番目の塩基は符号の冗長性に似た役割をしていると考えられる。

また、トランスポーター遺伝子自身においても符号に似た冗長性がある。トランスポーター遺伝子の膜貫通領域は疎水性の高い繰り返し構造になっている。この領域は疎水性が一定以上高くないと形成できないため、この領域には疎水性アミノ酸が多く存在する。本来疎水性アミノ酸が多数存在するはずの領域に親水性アミノ酸が存在することを考えると、膜貫通領域における親水性アミノ酸も冗長性とみなすことができる[3]。

そこで、本研究では遺伝子を一種の符号と見なし、符号構造という観点から機能遺伝子を分類し、金属トランスポーター遺伝子の同定を試みた。

3. 方法・実験

3.1 データセットの作成

学習データとして、TCDB (Transporter Classification Database)[4]から原核生物を対象とした金属トランスポーター遺伝子 69 件、金属以外のトランスポーター遺伝子 69 件のデータを取得した。

また、テストデータとして、国際塩基配列データベース DDBJ（DNA Databank of Japan）[5]から *Stenotrofomonas maltophilia K279a* 株と *Stenotrofomonas maltophilia R551-3* 株の全 ORF（Open Reading Frame）情報を取得した。これらの ORF 情報から location、product、gene、protein_id、アミノ酸配列、塩基配列、配列長を抽出し、テストデータセットを作成した（表 2）。

Vol.2011-MPS-86 No.11
Vol.2011-BIO-27 No.11
2011/12/1

表 2　*S.maltophilia K279a* 株のテストデータセットの抜粋例

location	product	gene	protein_id	アミノ酸配列	塩基配列	配列長
1..1332	putative chromosomal replication initiator protein DnaA	dnaA	CAN85568.1	MDAWSRSLEF	atggatgctg	1332
1608..2708	putative DNA polymerase III beta subunit	dnaN	CAQ43620.1	MRFTLQREAF	atgcgtttcac	1101
complement(2957..3274)	hypothetical protein		CAQ43621.1	MDKKALLPDL	gtggataaaaa	318
3704..4798	putative DNA replication and repair protein RecF	recF	CAQ43622.1	MLIRRLALHQI	atgcttatccg	1095
4911..7370	putative DNA gyrase subunit B	gyrB	CAQ43623.1	MSDEQNTPAI	atgagcgacg	2460
7438..8283	putative transmembrane protease		CAQ43624.1	MSASAPVSAF	atgtcagcatc	846
8355..9161	putative peptidase		CAQ43625.1	MKQLLLALVI	atgaaacaac	807
9294..10481	conserved hypothetical TPR repeat family protein		CAQ43626.1	MIFRNHKAVL	atgattttccg	1188
10625..11293	putative proline-rich TonB dependent receptor protein		CAQ43627.1	MTEQLVVHRY	atgacggaac	669
11388..12149	putative biopolymer transport exbB protein	exbB1	CAQ43628.1	MLQEIFIAAAA	atgctgcagg	762

location:配列上の位置情報; product:タンパク質の名称や機能; gene:配列に対応する遺伝子シンボル; protein_id: 翻訳される CDS (Coding Sequence) feature に対して国際塩基配列データベース INSDC (International Nucleotide Sequence Database Collaboration)が発行する識別子

3.2 シンドロームを利用した解析

本解析では､シンドロームという誤り訂正検査能力を DNA 配列に応用することで、遺伝子の符号構造を評価した。線形符号を満たす (6,3) 線形ブロック符号を用い、符号を形成するために最低限必要となる条件のみを考慮して、できるだけ多くの生成行列を算出した。(6,3) 線形ブロック符号は、符号長 6、情報記号 3、検査記号 3 の符号である。つまり、塩基配列に応用した場合、塩基配列を長さ 6 で区切り、前半の 3 塩基を情報記号、後半の 3 塩基を検査記号と仮定して解析を行う。

シンドロームは、符号系列Wに検査行列Hを乗算することで求められる。Wは以下の式で表される。

$$W = xG$$

ここで、xは情報系列、Gは生成行列を示す。以下に、シンドロームを算出するために必要な生成行列Gの算出方法と検査行列Hの算出方法について説明する。

(1)　生成行列の算出方法

本研究で用いた (6,3) 線形ブロック符号の生成行列の算出方法を図 1 に示す。まず、長さ 6 の文字列の全通りのリストを作成し、それらのリストの中から 3 行 6 列の行列 M を ${}_4{}^6\mathrm{C}_3$ 通り作成する。次に、${}_4{}^6\mathrm{C}_3$ 通りの M のうち、情報記号長にあたる 3 列が一次独立である場合のみ、M を既約標準形に変形することにより生成行列 G を算出する。

(2)　検査行列の算出方法

シンドロームを計算する際に必要となる検査行列の算出例を図 2 に示す。まず、3×3 の単位行列と残りの 3×3 行列を置換し、その左半分の 3×3 行列を転置することにより検査行列 H を算出する。

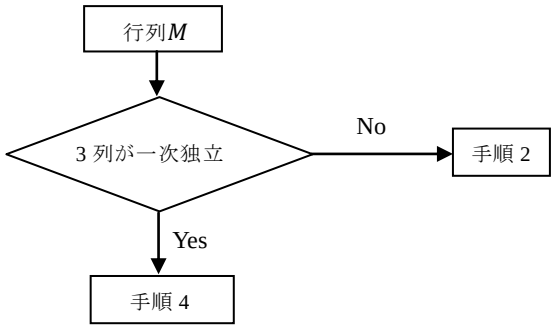
手順 1　*GF*(4)上の元{0、1、α、α^2}による長さ 6 の文字列全通りのリストを作成

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ & & & \vdots & & \\ \alpha^2 & \alpha^2 & \alpha^2 & \alpha^2 & \alpha^2 & \alpha^2 \end{matrix} \quad \Big\} \; 4^6\text{通り}$$

手順 2　リストの中から 3 つの行を取り出し、3 行 6 列の行列Mを ${}_4{}^6\mathrm{C}_3$ 通り作成

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

手順 3　${}_4{}^6\mathrm{C}_3$ 通りの行列Mに対して、3 以上の列が 1 次独立である行列を選択



手順 4　行列Mを基本行変形により既約標準形に変形し、生成行列を獲得

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{\text{基本行変形}} G = \begin{pmatrix} 1 & 0 & 0 & \mathrm{P1} & \mathrm{P2} & \mathrm{P3} \\ 0 & 1 & 0 & \mathrm{P'1} & \mathrm{P'2} & \mathrm{P'3} \\ 0 & 0 & 1 & \mathrm{P"2} & \mathrm{P"2} & \mathrm{P"3} \end{pmatrix}$$

図 1　(6,3)線形ブロック符号の生成行列の算出方法

Vol.2011-MPS-86 No.11
Vol.2011-BIO-27 No.11
2011/12/1

(3) シンドロームの算出方法

DNA 配列からシンドロームを計算する工程を図 3 に示す。本研究では、DNA 配列を受信後の符号語と仮定し、数量化した DNA 配列に検査行列を乗算することでシンドロームを算出した。また、配列から得られたシンドロームのパターンをカウントし、その頻度を配列長で除算することでベクトル化した。

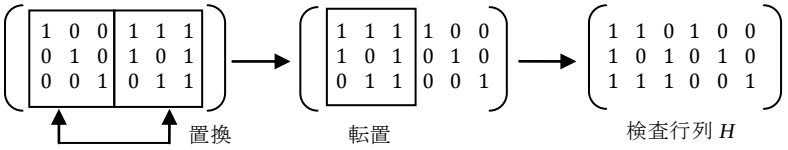


図 2　検査行列の算出例

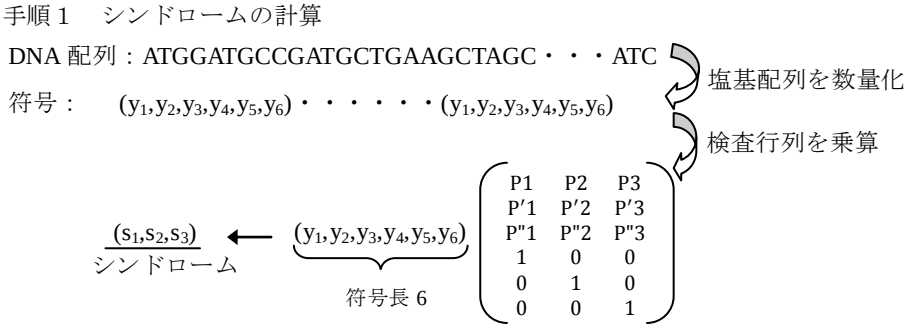


図 3　DNA 配列におけるシンドロームの算出方法

3.3　判別分析による機能遺伝子の分類方法

統計解析手法の一つである判別分析を用いて、金属トランスポーター遺伝子と金属以外のトランスポーター遺伝子の分類を試みた。判別分析は、クラスの分類を行うための統計的学習アルゴリズムであり、学習データをクラスごとに分ける最適な境界線を引くことで、未知のデータがどちらのクラスに属するかを識別することができる[6]。本研究では、判別式の妥当性を leave-one-out 交差検定を用いて検証した。判別分析による分類器の構築手順を以下に示す。

(1) TCDB から取得した金属トランスポーター遺伝子 69 件を陽性データ、金属以外のトランスポーター遺伝子 69 件を陰性データに割り当てる。
(2) 上記の 138 件のデータから 1 件のデータを検証用データとして抜き出す。
(3) 残りの 137 件のデータに対して判別分析を行い、判別曲線を学習させる。
(4) 学習済みの分類器に、抜き出した 1 件の検証用データを入力し、その分類結果が、真陽性、偽陰性、偽陽性、真陰性のうち何に当てはまるかを見る。
(5) (2)から(4)までの作業を 138 回繰り返す。
(6) 最後に、学習させた判別曲線に *S.maltophilia K279a* 株と *S.maltophilia R551-3* 株の全遺伝子を適用することで、*S.maltophilia K279a* と *S.maltophilia R551-3* 株の金属トランスポーター遺伝子を同定する。

3.4　判別分析の精度の指標

判別分析における精度の評価は、特異性（Specificity、以下 SP）、精度（Accuracy、以下 AC）、感度（Sensitivity、以下 SE）、相関係数（Correlation coefficient、以下 CC）の 4 つの指標を用いて行った。

$$SP = \frac{TP}{TP + FP} \qquad AC = \frac{TP + TN}{TP + FN + TN + FP}$$

$$SE = \frac{TP}{TP + FN} \qquad CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

特異性 SP はデータセットにおける金属トランスポーター遺伝子のうち、判別分析で正しく判定された割合を示し、感度 SE は、判別分析によって金属トランスポーター遺伝子と判定されたうち、実際にその判定が正しかった割合を示す。また、相関係数 CC は金属とそれ以外のトランスポーター遺伝子の分類における相関を表し、精度 AC は全遺伝子中、金属トランスポーター遺伝子・それ以外のトランスポーター遺伝子と正しく判別された割合を示す。

Vol.2011-MPS-86 No.11
Vol.2011-BIO-27 No.11
2011/12/1

4. 結果・考察

TCDB から取得した機能既知のトランスポーター遺伝子を用いた判別分析の結果を表 3 に示す。この結果は最も判別精度の高かった符号の判別結果である。この結果から、例えば、実測値の金属トランスポーター遺伝子 69 件のうち、判別分析によって正しく金属トランスポーター遺伝子に分類されたのが 64 件であることが分かる。

表 3　leave-one-out cross validation の結果

		予測値	
		Metal Transporter	Other Transporter
実測値	Metal Transporter	64	4
	Other Transporter	5	65

Leave-one-out 交差検定によって、TP、TN、FP、FN と判定された遺伝子の数から算出した SP、SE、AC、CC の結果を表 4 に示す。SP 値から、金属トランスポーター69 件中の 94.2%の遺伝子が判別分析により金属トランスポーター遺伝子として正しく分類されたことが分かる。また、SE 値より、金属トランスポーター遺伝子と判定された 64 件中の 92.8%の遺伝子が正しく判定されたことが分かる。さらに、CC の値が 0.865 であることから、両遺伝子群に強い相関関係があることが読み取れる。総合的な判別精度の指標である AC 値より、93.4%の確率で金属とそれ以外のトランスポーター遺伝子に正しく判別されたことが分かる。

表 4　分類器の精度

	Metal Transporter	Other Transporter	両遺伝子群
SP	0.942	0.927	
SE	0.928	0.941	
CC			0.865
AC			0.934

この学習器の判別曲線を用いて、*S.maltophilia K279a* 株と *S.maltophilia R551-3* 株の金属トランスポーター遺伝子の同定を試みた。*S.maltophilia K279a* 株に関しては、全 ORF の 4,386 件中、1,708 件が金属トランスポーター遺伝子として判別された。また、1,708 遺伝子について注釈情報を調べた結果、「transporter」と記載されている遺伝子が 58 件、「transmembrane」と記載されている遺伝子が 215 件含まれていた。*S.maltophilia R551-3* 株は、全 ORF の 4039 件中、1674 件が金属トランスポーター遺伝子と判別され、「transporter」と記載されているものが 50 件、「transmembrane」と記載されているものが 8 件含まれていた。

本解析により、情報数理理論の一つである符号理論（シンドローム）が金属トランスポーター遺伝子を特定するのに有用な手段であることが示唆された。また、この結果から *S.maltophilia* の薬剤耐性に関与している亜鉛を取り込むトランスポーター遺伝子特有の符号を特定することが可能であることも示唆された。

5. まとめと今後の課題

本研究では、従来のバイオインフォマティクス手法ではその配列の特徴により同定が難しかった *S.maltophilia K279a* と *S.maltophilia R551-3* 株の重金属トランスポーター遺伝子を同定するために、情報理論の一つである符号理論の応用を試みた。具体的には、符号理論における符号の冗長性とコドン表の 3 番目の塩基の役割が似ていることから、タンパク質を構成する ORF の DNA 配列を長さ 6 で区切り、前半の 3 塩基を情報記号、後半の 3 塩基を検査記号とみなして配列の解析を行った。

符号を形成するために最低限必要となる条件のみを考慮して、(6,3)線形ブロック符号を作り、13 万件以上の生成行列を解析に用いた。また、各機能遺伝子のシンドロームを指標とした判別分析により、高精度に金属トランスポーター遺伝子を同定する符号を特定した。金属トランスポーター遺伝子と金属以外のトランスポーター遺伝子を最も区別する符号を用いて、*S.maltophilia K279a* と *S.maltophilia R551-3* 株の金属トランスポーター遺伝子の同定を試みた。

今後の課題として、本解析で *S.maltophilia K279a* と *S.maltophilia R551-3* 株の金属トランスポーター遺伝子と判定された遺伝子について、どの遺伝子が亜鉛の取り込みに関与する金属トランスポーター遺伝子なのかを特定する予定である。また、今回構築した学習器によって誤判別された金属トランスポーター遺伝子について、原因を調査し、その結果を判別分析に反映させることで、学習器の判別精度をさらに高める予定である。

Vol.2011-MPS-86 No.11
Vol.2011-BIO-27 No.11
2011/12/1

参考文献

[1] McGowan, J.: Resistance in non-fermenting gram-negative bacteria: multidrug resistance to the maximum, *AM. J. Med.*, 119:S29-36, 2006.
[2] 辻彰総集編: トランスポーター科学最前線, pp.73-91.
[3] Takashima, S., Okihara, R., Suzuki, T., and Miyazaki, S.: Identification of gene function from application of coding theory, *Proc. of the 2008 Annual Conference of Japanese Society for Bioinformatics* (JSBi2008), P012, 2008.
[4] http://www.tcdb.org/
[5] http://www.ddbj.nig.ac.jp/
[6] Seber, G.A.F.: Multivariate Observations, Wiley, New York, 2004.