平成 12 年度知能システム科学専攻修士論文

決定木を用いた異種外来タンパク質の発現予測

吉良 聡

Predicting a foreign gene expression by using decision trees

Satoshi Kira

提出年月日 平成 13 年 2 月 27 日 (火)

主査教官：山村 雅幸 助教授
審査教官：小林 重信　教授
審査教官：新田 克己　教授
審査教官：松崎 尹雄　教授
審査教官：安倍 直樹 助教授

決定木を用いた異種外来タンパク質の発現予測

吉良 聡

Predicting a foreign gene expression by using decision trees

Satoshi Kira

abstract: Recently many biotechnologies have been heated up. Among them there is a special research field about foreign gene expression. Useful proteins are produced by using a host cell(e.g. E.coli, yeast). However, this technology is unstable and inadequate, thinking of practical situations. One reason which makes it difficult is that each host cell has its own likes. Experts try the huge number of experiments when they face these situations. These are really hard work and basically they depend only their intuitions. So we propose a prediction system of foreign gene expression by using decision tree techniques for the purpose of reducing their efforts. Using instances very few, so we apply Cross-Validation and Bootstrap methods for error estimation and model selection, and use transfomed instances into binary classes. Finally We apply and evaluate it on some real problems offered by AGC(Asahi Glass Company). The results agreed with usual researches and these were useful informations for experiments with a foreign gene expression.

1 はじめに

近年，ゲノム解析の急速な発展に伴いバイオテクノロジーや遺伝子情報を計算機を用いて解析する遺伝子情報処理の分野が脚光を浴びている．そのバイオテクノロジーの一分野に，異種外来タンパク質の生産 [1] の分野がある．異種外来タンパク質の生産とは，大腸菌や酵母などの微生物 (宿主細胞) を用いることによって，有用なタンパク質を生産する方法である．実際の異種外来タンパク質の生産は，以下の手順に沿って行われる．

1. タンパク質の遺伝子の準備
2. 宿主細胞の選択
3. 発現ベクターの作成
   - プロモータ配列の挿入 (オプション)
   - シグナル配列の挿入 (オプション)
4. ベクターを宿主細胞に導入
5. 宿主細胞の培養
6. タンパク質の発現・生産

まず，生産したいターゲットタンパク質をコードした遺伝子を用意する．そして使用する宿主細胞を選ぶ．次に，遺伝子単体では宿主細胞内に導入してもタンパク質は発現しないことが多いので通常，遺伝子の上流に制御配列である “プロモータ” や “シグナル” などを付加する必要がある．このように制御配列を付加した塩基配列のことをベクターと呼ぶ．そして，発現ベクターを宿主細胞へ導入・培養し目標タンパク質の発現を待つ (図 1)．

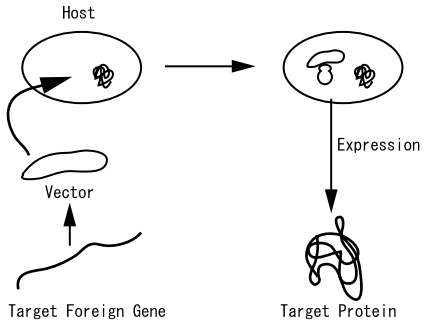

図 1: 異種外来タンパク質の生産

この一連の作業を行うには，通常数週間の期間が必要であり，発現実験は常に成功するわけではない．も

し失敗したときは，実験者の経験と勘により原因を考察し実験を繰り返し試行するので，発現させるまでに数ヵ月の期間を要するのが普通である．

発現に関係する要因はいくつか考えられ，

- 発現ベクターの設計
- 宿主細胞の選択・培養条件
- タンパク質の性質
- 宿主細胞のコドンの好み

が挙げられる．この内，興味深い要因として宿主細胞のコドンの好みがある．コドンとは，塩基配列 (遺伝子) からアミノ酸 (タンパク質) へ翻訳される際の遺伝暗号の基本単位のことであり，3 連続した塩基の通称であるが，コドン 64 種類に対しアミノ酸は 20 種類しか存在しないので，異なるコドンが同じアミノ酸を暗号化してる場合が多い (このようなコドンを同義コドンと呼ぶ)．同義コドンの存在より，同じタンパク質 (アミノ酸配列) を暗号化するとき複数の塩基配列が対応することになる．このコドンの研究を通して，池村らはコドン使用において生物間に特徴的な差異が存在することを示している [2]．この中で，細胞内に多量に存在する Ribosomal Protein などの遺伝子では，コドン使用のパターン (コドン使用の偏り) が明確である一方，外来性的の遺伝子ではそうではないことが分かっている．このことから，宿主細胞内で大量にタンパク質を発現させるには，宿主細胞が好むコドンを用いてタンパク質を暗号化すれば良いと考えられる．

本研究では発現実験の効率化を目的に，過去の異種外来タンパク質発現事例を用いた異種外来タンパク質の発現予測システムを実現する．今まで実験者の経験や勘を頼りに試行錯誤し膨大なコストが掛かっていた発現実験を，過去の発現実験データから生成した予測システムを用いることによりコストの削減が期待できる．予測システムの実装にあたっては，実際の発現実験事例からコドン使用の偏りを手がかりに決定木の帰納学習を行ない予測モデルを得る．決定木の帰納学習などの機械学習を用いるメリットは，対象問題が人の手に負えないくらい複雑で，その問題のエキスパートですら仮説の構築が困難な場合でも，対象問題を機械学習に適用可能な形に変換できれば，何らかの仮説を導くことが出来ることである．また，機械学習で得られる仮説は人の手を通らないことから，エキスパートの仮説と比較し，より偏りの少ない仮説が得られる可能性があり，時として人の手で考え出した仮説よりも優れる期待がある．本稿での予測システムの性能評価は，Cross-Validation 法と Bootstrap 法を用いて推定正答率を算出し，かつ発現量の程度の異なる予測を行なう決定木同士の構造比較によって行なう．また，従来から得られている知見と本研究で得られた結果の整合性を確認することにより木の信頼性を評価する．

以下，2 章では異種外来タンパク質の発現問題についてより詳細に説明をすると共に，生物種間におけるコドン使用率の偏りについて述べることにより，従来法の問題点と本研究の着眼点を明らかにする．3 章では，2 章で述べる着眼点に基づき本手法の提案を行い，4 章において予測システムの実装を行なう．5 章では計算機実験の方法・結果を提示し考察する．6 章において本研究をまとめ，今後の方針について述べる．

2 異種外来タンパク質の発現問題

異種外来タンパク質の発現を予測するにあたり，生物におけるタンパク質の合成過程を知ることは重要である．本章では，タンパク質の合成過程を紹介し，本研究において重要になる生物種間でコドン使用に偏りが存在することを述べる．そして最後に，従来から調べられている宿主細胞におけるコドン使用パターンを用いて発現実験を改善する際の問題点を示す．

2.1 タンパク質の合成過程

生物におけるタンパク質合成の過程を知るのは，異種外来タンパク質の生産において非常に重要である。生物細胞内におけるタンパク質の合成は，セントラルドグマの流れに沿って行われる (図 2)．

図 2 が示すように遺伝子 DNA は 4 種類の塩基 ATCG の配列として形成されている．DNA は mRNA へと転写 (transcription) され，その塩基配列は 3 つの並びを一単位として tRNA によってアミノ酸へと翻訳 (translation) される．そして，そのアミノ酸配列が物理的に安定な形に折り畳まれて (folding) タンパク質になるのである．

この合成の過程において，以下のような合成が成功するか否かの重要な要素がいくつか存在する．

- 制御配列の種類 (プロモータ，シグナル等)
- mRNA の 2 次構造

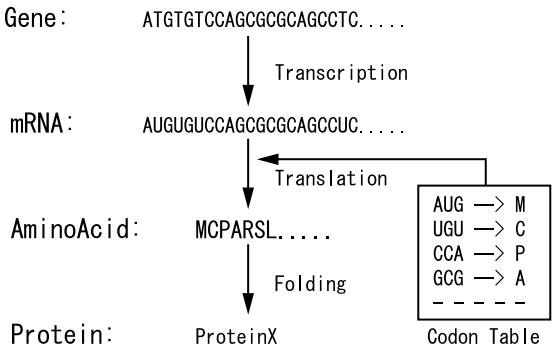

図 2: タンパク質合成過程

first letter	second letter U(T)	second letter C	second letter A	second letter G	therd letter
U(T)	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA 終止 UAG	UGU Cys UGC UGA 終止 UGG Trp	U(T) C A G
C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U(T) C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U(T) C A G
G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U(T) C A G

図 3: 遺伝暗号表

- タンパク質の性質
- コドン使用の偏り

数ある要因の中で，塩基配列からアミノ酸配列へと変換される翻訳機構に着目する．翻訳は，タンパク質合成過程において最も根本的であり，重大な影響を与えると考えられる．そこで，遺伝子の翻訳において非常に重要と思われる生物におけるコドン使用パターンについて次で述べる．

2.2 生物間におけるコドン使用パターンの相違

遺伝子の翻訳過程ではコドンを基本単位としてアミノ酸へと変換されて行く．塩基は 4 種類でコドンがそれを 3 連続したものなので，コドンは $4^3 = 64$ 種類存在することになる．そのうちの 3 つが翻訳の終了を表す終止コドン (stop codon) として使われているので，残りの 61 種類でアミノ酸を表現することになる．しかし，コドンがコードすべきアミノ酸は 20 種類しか存在しない．これはどういうことか？答えは図 3 である．すなわち，複数のコドンが同じアミノ酸を表現しているのである (但し，Met，Trp は除く)．このように，同じアミノ酸を表現するコドンのことを同義コドン (synonymous codon) と呼び，この同義コドンの存在によりコドン使用パターンというものが生物種を特徴付ける指標になる．

同義コドンが存在するということは，ひとつのタンパク質を塩基配列で表現するにあたり，複数の塩基配列の種類が許容されるということである．この自由度があることで，異なる生物間では異なるコドン使用パターンを持つことが知られている．つまり，生物毎に各コドンの使用率に好みが存在するのである．また池村らは，いくつかの生物種において細胞内に存在するタンパク質量と，その生物が好んで使用するコドンの数との間に強い相関関係があることを示している [2].

このことから，宿主細胞が外来遺伝子を発現させるときも，遺伝子中に “好みのコドン” の存在量が多ければスムーズに翻訳が行われ，宿主細胞がターゲットのタンパク質の生産を大量に行うことが予想できる．

2.3 コドン使用率の定義

ここでコドン使用率というものを定義する．コドン使用率とは，遺伝子中に含まれる各コドンの存在率のことで，2 種類のコドン使用率が考えられる．一つ目は遺伝子中に含まれる各コドンの個数を遺伝子中に存在する全コドンの総数で割ったもので，二つ目は遺伝子中に含まれる各コドンの個数を配列中に含まれるそのコドンと同じアミノ酸をコードする同義コドンの総数で割ったものである．本研究では前者を global codon usage(gcodon) と，後者 local codon usage(lcodon) と呼ぶこととする．アミノ酸 m をコードする i 番目のコドンのコドン使用率 $G_i(m)$(gcodon)，$L_i(m)$(lcodon) の定義を以下に示す．

$$G_i(m) = \frac{f_i(m)}{\sum_{m=1}^{20}\sum_{i=1}^{M(m)} f_i(m)} \quad (1)$$

$$L_i(m) = \frac{f_i(m)}{\sum_{i=1}^{M(m)} f_i(m)} \quad (2)$$

但し，$f_i(m)$ はアミノ酸 m をコードする i 番目のコドンの個数を表し，$M(m)$ はアミノ酸 m の同義コドンの数を表す．lcodon では同義コドン間での情報しか含んでいないのに対し，gcodon では同義コドン間の情報と配列中のアミノ酸組成の情報も含んでいる．

2.4 従来法における問題点

異種外来タンパク質の発現予測を陽に目的として行われた研究は非常に少なく，これは非常に問題である．関連の研究として，宿主細胞の遺伝子を調べ高発現のタンパク質をコードする遺伝子のコドンの使用パターンを調べたもの [2, 3] や，生物の全ゲノムを対象にコドンの使用パターンに対し多変量解析を行ったもの [4] がある．

前者では，まず始めにその生物中で重要かつ高発現な遺伝子 (タンパク質) をいくつか選びだし，その数種の遺伝子中のコドンを数えあげコドン使用率を算出している．そしてその使用率に明らかな偏りが存在するのを発見した．後者では，大腸菌の全ゲノムを対象に，各遺伝子配列をコドン使用率による 61 次元空間上 (終止コドンは除く) にプロットし，その分布に対して多変量解析を行う．その解析結果として，Ribosomal Protein のような高発現のタンパク質は全体の分布中のある部分に集中しており，このことから，高発現の遺伝子のコドン使用率には明らかな偏りがあることが示されている．

しかし，上記の研究はどちらも宿主細胞のコドン使用率の偏りを示しただけであり外来タンパク質の発現予測を行なうものではないので，実験室に蓄えられている外来タンパク質の発現実験情報は全く用いていない．このため，得られる情報は外来タンパク質の発現にとっては間接的な情報であるし，外来タンパク質特有のコドン使用パターンが存在しそれが発現を助けているのなら，宿主細胞のコドン使用パターンを調べるだけでは不十分である．また，どちらの研究もすべてのコドン使用率を足し合わせた末の結果を出すので，情報が一次元的になり，コドン使用率間の依存関係 (相関関係) などの情報が落ちてしまっている．実験の現場では，発現量を上げようとターゲット遺伝子のコドンを，宿主細胞に適したコドンへと変更するために宿主のコドンの偏りを用いるのだが，従来のようなテーブル式の情報では，どのコドンを変更したらよいか絞り切れずかつ，配列中のどの部分のコドンを変更したら良いか判断できず，結局実験は試行錯誤的になり効率は上がらない．

以上の考察より，発現予測を行なうには次の項目をサポートすべきである．

1. 実際の発現実験情報の活用
2. コドンの依存関係の表現
3. 発現予測・支援に相応しい高次の知識表現・利用法

本研究においては，1 の着眼点から過去の発現実験事例を用いた異種外来タンパク質の発現実験に発現予測・支援システムの実現を図り，2，3 の着眼点より，知識の表現方法として決定木と呼ばれる木構造の知識表現を用いる．

3 発現予測システム

本章では，異種外来タンパク質の発現問題に適した発現予測・支援システムの提案を行なう．予測システムは過去の発現事例から導き出すことによって，従来法とは異なりより直接的な情報を扱うことが出来る．本章では予測システムの概念を紹介し，次章において本システムの実装を行なう．

3.1 発現予測システムの提案

上記に示した異種外来タンパク質の発現実験において，人手と時間のコストが非常に掛かるという問題を解決するために，本研究では，過去の発現実験の事例を基に診断モデルの帰納学習を行ない発現予測を行なう予測システムの提案をする．提案法の全体図を図 4 に示す．本システムは，

- 学習フェーズ
- 予測フェーズ
- 実験フェーズ

の 3 つの局面を持つ．学習フェーズでは，事例からの診断モデルの帰納学習を行ない，予測フェーズでは，診断モデルを用いて新規配列の発現予測及び発現実験へのサジェストを出力する．実験フェーズでは，実験結果及び事例のフィードバックを学習フェーズへと行なう．以下では，各フェーズそれぞれのコンポーネントの機能，処理の流れを説明する．

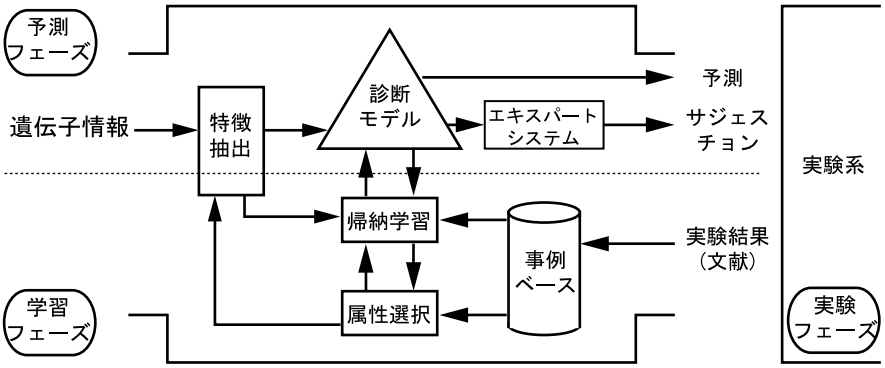

図 4: 提案手法の全体図

3.1.1 学習フェーズ

学習フェーズでは過去の発現実験の事例より診断モデルの獲得を行なう．実際の処理の流れは以下の様になる．

1. 過去の発現実験の事例を収集
2. 事例からの特徴抽出
3. 診断モデルの帰納学習
4. 診断モデルの評価

まず，**事例ベース**から実際に行なわれた発現実験の事例を得る．この実験事例から手がかりとなる特徴量を決定する (**特徴抽出**)．その例として，遺伝子配列中の制御配列の種類，アミノ酸組成，コドン使用率などが考えられる．そして，その手がかりとなる変数から発現に関係する規則を自動的に獲得し (**帰納学習**)，それを発現の**診断モデル**とする．最後に，得られた診断モデルの評価を行なう．もし得られた診断モデルが不適切と判断されたときは，適切なモデルが得られるまで特徴抽出及び帰納学習の段階を繰り返す．

3.1.2 予測フェーズ

予測フェーズでは，未知のターゲットタンパク質に対し発現するか否かの予測を行なう．更に，発現しないタンパク質には発現を起こさせるサジェストを，発現が小量なら大量に発現するようなサジェストを出力する．予測フェーズの処理の流れは以下のようになる．

1. 新規配列を入力
2. 新規配列から特徴抽出
3. 診断モデルによる診断
4. 診断結果に基づく予測・サジェスト

まず，ターゲットタンパク質の**遺伝子情報**から診断モデルで用いられる特徴量を得る．次にターゲットタンパク質の特徴量を入力として，**診断モデル**から診断結果を得る．ここで診断結果が，発現実験上思わしくない結果なら，診断結果を入力として**エキスパートシステム**から (大量) 発現に関する**サジェスチョン**を得る．エキスパートシステムでは，診断モデルの診断と過去の発現実験の結果に基づき総合的に判断を下す．

3.1.3 実験フェーズ

実験フェーズでは，実際の異種外来タンパク質の発現実験を行ない，発現実験事例の収集，予測システムの評価を行なう．従来の発現実験の方法と異なるのは以下の 2 点である．

1. 過去の実験データの収集
2. 発現予測システムの利用・評価

従来の発現実験の現場では，文献で紹介されているような発現に成功した事例だけではなく，成功例と同数かそれ以上の失敗の事例が存在している．これは非常に情報量が多いデータ集合といえる．そこで，過去に行なわれていた実験の事例を整理･収集し，発現予測システムの学習フェーズへと送る．この事例集合が整って初めて予測システムが機能する．

4 提案法の実装

本稿では，提案手法のうち 3.1.1 で示した学習フェーズのみを対象として計算機上に実装した．本章では今回対象とした学習フェーズの実装方法について説明する．予測フェーズと実験フェーズは今後の課題とする．学習フェーズを実装するためには，用いる事例の形式を考察することが重要であるので，まず用いる発現実験事例について紹介を行ない，それに続いて学習フェーズのアルゴリズムの説明を行なう．

4.1 提供を受ける発現実験事例

発現実験事例は，*Schizosaccharomyces Pombe*(以下 *S.Pombe* と略記) と呼ばれる分裂酵母を宿主として，外来タンパク質の発現実験を行なったものである．分裂酵母は，分裂をしない酵母や大腸菌と比較して，より人間などの高等生物の細胞と類似しており，複雑なタンパク質などの発現を期待できる．また，培養の手間も動物細胞などを用いる場合と比べれば非常にかからず，外来タンパク質の発現に使用する宿主細胞として適していると考えられている．

事例の特徴を表 1 にまとめる．また，実験データは共同研究を行なっている旭硝子 (株) からの協力を受けている．

事例数	48
発現量	0〜4
遺伝子配列	塩基配列
シグナル配列	塩基配列

表 1: 過去の発現実験の事例の内容

学習フェーズを実装するにあたり，以下の 3 つの点に注目する．

- 遺伝子配列が含まれる
- 得られる事例数が少ない
- 発現量が 5 段階の離散値で得られる

遺伝子配列を事例の特徴量として直接扱うのは難しい．従来から用いられている機械学習の手法に，この発現実験事例を適用するのならば配列情報から何らかの数値情報やカテゴリー情報に変換するのが好ましいと考えられる．

事例数は 48 個と，異種外来タンパク質の発現現象を解析するには非常に少ない数と思われる．機械学習では一般に，学習に用いる事例数が少ないと学習結果が非常に不安定になることが知られている [7]．そのため，学習アルゴリズムを設計する際，学習結果を安定化させる手法の導入が必要である．

最後に注意しなければならないのが，予測すべき発現量が 5 段階で得られるが，その 5 段階で発現量の予測を行なうには事例数が少ないことである．しかし，単純に 2 値化してしまうと重要な情報を失ってしまう危険性があるので工夫が必要と思われる．

4.2 決定木を用いた学習フェーズの実装

前述の注意点を考慮し学習フェーズを構築する．診断システムには決定木表現を用い，決定木の帰納学習には既存のアルゴリズムである ID3[5] を利用する．その帰納学習に用いられる属性値はコドン使用率を採用する．学習結果の安定化には，Cross-Validation 法・Bootstrap 法を組み合わせた手法を用いる．

4.2.1 決定木による診断システムの構築

診断システムに求められる機能は，

- 発現 (量) の予測
- 予測に対する理由付け

である．生物実験者は，未知のタンパク質が発現するか否か，発現しないならどの様にすれば発現させることができるのか，そのサジェストを求めて本システムを使用する．そのため，ニューラルネットを用いた予測のように予測に意味付けをしにくいシステムは適さない．そこで診断システムには決定木による表現を用いる．決定木の例を図 5 に示す．決定木は “ノード” と “エッジ” の集合として表現される．ノードには事例を分割する “属性”，特に末端のノード (リーフ) には事例の属する “クラス” が記述される．エッジには条件が記述されている．この木を上から辿って行くと if...then ルールの集合になっていることが分かる．仮に 20 代の女性に結婚観の意識調査をして，その結果をまとめると図 5 のような決定木に表現出来たとする．すると，世の未婚の男性は自分が結婚できそうかどうか決定木を用いて予想が出来る．また “N” と判断され

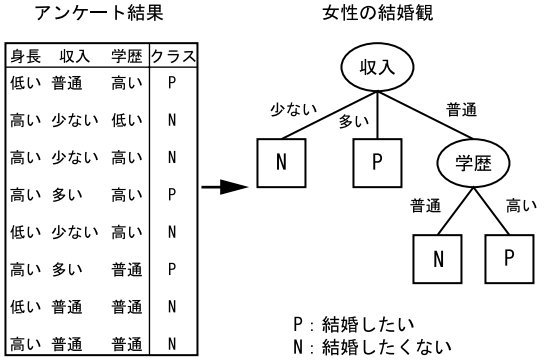

図 5: 決定木の例

た場合，“P” と判断されるにはどうしたら良いかが決定できる．このように決定木は，過去の事例から未知の事例を予測することができ，予測の理由も明確に表現されるので本研究に非常に適している．

決定木の生成に関しては，事例集合から決定木を自動的に構築する多くのアルゴリズムが提案されている．そのうち最も一般的な ID3 と呼ばれるアルゴリズムを本研究では使用する．決定木の帰納学習の基本アルゴリズムを以下に示す．基本アルゴリズムでは，ある選択基準に基づき例空間を分割統治法によって再帰的に分割しながら決定木を構築して行く．

```
begin
  while(単一クラスでない事例集合が存在)
    if(事例が単一クラス)
      クラスノードを生成
    else
      ある選択基準より属性選択
      ある属性のノードを生成
      その属性で事例を分割
    end
  end
end
```

ID3 では，基準に情報ゲイン $G_c(a,E)$ の最大化を用いてコンパクトな木の生成を目指す．この基準は分割前のクラス情報量と分割後のクラス情報量の差で定義されており，分割によって得られる情報量の期待値を表している．

$$G_c(a,E) \equiv info_c(E) - infox_c(a,E) \quad (3)$$

$$info_c(E) = -\sum_i \frac{|E_{c=i}|}{|E|} log_2(\frac{|E_{c=i}|}{|E|}) \quad (4)$$

$$infox_c(a,E) = \sum_i \frac{|E_{a=i}|}{|E|} info_c(|E_{c=i}|) \quad (5)$$

但し，a は属性を表し，$E_{a=i}$ は例集合 E のうち属性値 $a=i$ に属する例の部分集合，$E_{c=i}$ は E のうちクラス $c=i$ に属する部分集合を表す．$|E|$ は E の例数である．

本研究で用いる ID3 は数値属性も扱える様に，C4.5[6] で使用されている機能を拡張したものである．これは，事例をある属性において一つの閾値で分割を行なうもので，属性・閾値の決定には情報ゲイン最大化を指標にする．

本研究では，発現量が各事例の属するクラスを，遺伝子情報が属性を表している．発現量がクラスになるので，クラスは 5 段階あることになる．しかし，事例数が 48 個と非常に少ないのでクラスは 2 値化することにする．また，遺伝子配列はコドン使用率に変換し 61 個の連続値属性として扱う．

クラス (発現量) 5 段階 → 2 段階 (発現する，しない)

属性 (遺伝子配列) 塩基配列 → コドン使用率

クラスを 2 値化するとき，色々な 2 値化の組合せが考えられる．本研究では発現量に応じた 2 分割を考える (表 2)．このようにクラスの分割方法の異なる 4 つの

発現量	事例数	事例集合名 C0	C1	C2	C3
0	20	20	22	26	37
1	2	28			
3	4		26		
4	11			22	
5	11				11

N

P

表 2: 事例集合のクラス分割方法

事例集合 $C0, C1, C2, C3$ を用いることにする．この分割方法では，$C0 \rightarrow C3$ に変化するに連れて，発現量の多いタンパク質のみが発現したと判断する事例集合になる．

クラスが 5 段階のままでは学習を行なうのが困難なため，2 段階のクラスに落し前述のように $C0, C1, C2, C3$

事例集合を生成する．そのそれぞれで学習結果を出し，その学習結果である決定木を統合的に比較・検討し挙動を観測することで，5 段階で予測するのに近い情報の獲得が期待される．

4.2.2 学習結果の評価・選択法

決定木の帰納学習などの機械学習では，学習用の事例が少ないとき学習事例の一部が変化しただけで，学習結果が大きく変わることがある．これは学習アルゴリズムの不安定性と呼ばれており大きな問題になっている [7]．そこで学習結果の安定化，もしくは有用なモデルの選択法として，Cross-Validation 法や Bootstrap 法 [8] のリサンプリングに基づく手法が提案されている．

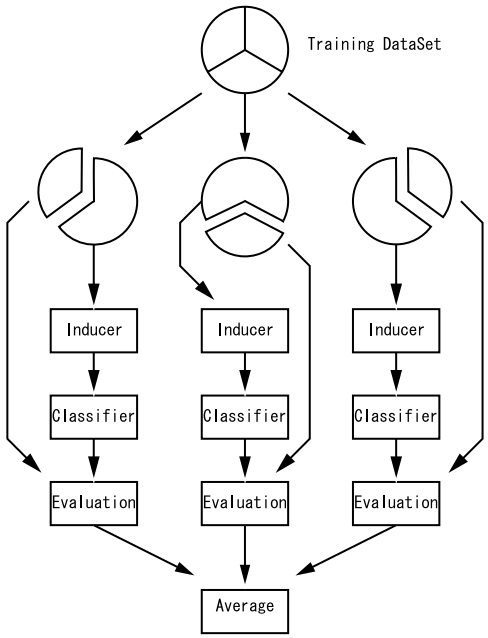

図 6: k-hold Cross-Validation ($k = 3$)

Cross-Validation 法 事例集合をほぼ同様の大きさの k 個の集合に分割し，k-1 個の事例集合を学習に用いて残りを学習結果の検定に使う．この作業を k 回繰り返し，その平均を最終的な評価とする (k-hold-Cross-Validation)．モデル選択の場合は，評価が最大のモデルを選択する (図 6)．

Bootstrap 法 元の事例集合から重複を許す再抽出により事例集合に良く似た新たな事例集合を生成する．事例数は同じでなくても良い．この集合に対して学習を行ない，これを何度か繰り返し Cross-Validation などにより有効なモデルを選択する．

上記の 2 つの方法は非常に有名で，実際に良く使われている．最後に，本研究で用いるアルゴリズムの全体図を図 7 に示し，その説明を行なう．まず，(A) において生の発現実験事例集合からクラスの 2 値化・コドン使用率の算出をし，学習に用いる事例集合に変換する．次に (B) において k-hold Cross-Validation に基づき事例集合をテスト用と学習用に分割する．その学習事例を Bootstrap して Bootstrap 学習集合を生成し (C)，その集合から ID3 で決定木を生成する (D)．得られた決定木 (E) をテスト用の事例で評価，正答率の算出を行なう (F)．正答率とはテスト用の事例の何%のクラスを正しく予測できたかの値である．この (C) → (F) を何度か繰り返す (Bootstrap 回数)．そして，(G) において生成された決定木群の中から正答率最大の木をいくつか残す (Bootstrap 生き残り数)．ここまでの試行を，Cross-Validation の hold 数である k 回繰り返し，(H) でそれぞれの生き残った決定木を集める．そして，(I) において決定木集合の中から良く現れる構造を多く含む決定木を選択する (本稿ではこの過程を人の手で実施)．(I) で選ばれた決定木を最終の学習結果として出力する (J)．簡単なアルゴリズムを以下に記す．

```
begin
  accuracyTreeSet ← null
  for(CV の繰り返し回数)
    for(k-hold-CV)
      trainingSet の決定
      testSet の決定
      bsTreeSet ← null
      for(Bootstrap 回数)
        Bootstrap 集合の生成
        bsTree ← ID3(Bootstrap 集合)
        testSet による bsTree の評価
        bsTreeSet.add(bsTree)
      end
      accuracyTree ← bsTreeSet.get(評価最大)
      accuracyTreeSet.add(accuracyTree)
    end
  end
  accuracyTreeSet より頻出の決定木を選択
```

```
end
```

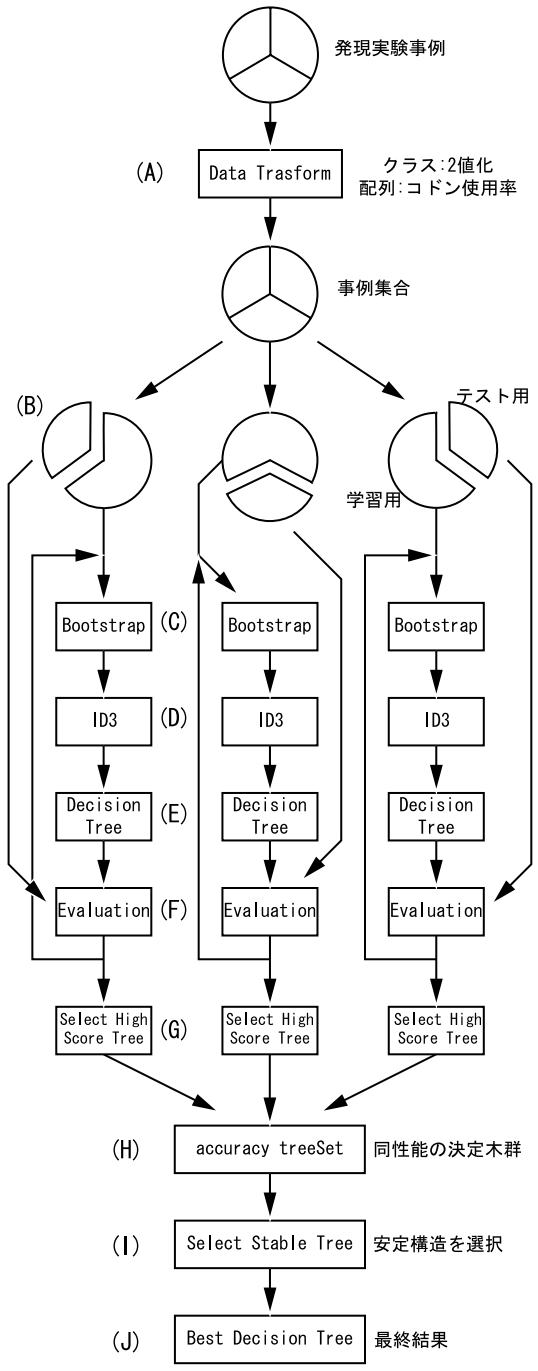

図 7: 実装の概要図 (3-hold Cross-Validation の場合)

実際には，この方法で $C0$〜$C3$ すべての事例集合において決定木を生成する．その 4 つの事例集合から生成した決定木の共通部分などを見出すことで興味深い発現ルールを発見することが期待される．

5 発現実験事例への適用

本章では過去の発現実験事例に本システムを採用し，その結果を考察する．用いる事例は前章で説明した通り，宿主細胞に *S.pombe* と呼ばれる分裂酵母を用いて外来タンパク質の発現実験を行ったものであり，旭硝子 (株) から提供される現実の事例である．まず，実行の手順を説明し従来法との比較により評価及び考察を行なう．

5.1 クラス・属性の決定

クラス 前述の通り発現量がクラスにあたる．しかし発現量は 5 段階の離散値で表現されているので，

- 発現しやすい (Positive : P)
- 発現しにくい (Negative : N)

と 2 値化をする．表 2 にように 4 種類の 2 値化をして 4 種類の事例集合 $(C0, C1, C2, C3)$ を考える．

属性 属性の選択は予測システムの性能を決める重要な過程である．本研究では,「生物的意味の深さ」「発現へのサジェストとの適合性」「実装の簡単さ」の 3 つの観点から，コドン使用率を属性に採用し，前述の gcodon，lcodon の両方を試行する．各々のコドン使用率は，発現とは直接関係しない終止コドンを除くため，61 個の連続値の属性となる．

コドン使用率は生物学の観点から見て十分に意味があり，また (大量) 発現のサジェストをする際も，同義コドン内の変更なら発現するタンパク質は変化しないので非常に都合が良い．また，gcodon ではアミノ酸組成の情報と同義コドン間の情報の両方が含まれているのに対し，lcodon では同義コドン間での情報だけに限定されている．この差がどのように学習結果に反映されるかは興味深い．

5.2 計算機シミュレーション

予測システムの学習フェーズを計算機上に実装し，用いる事例集合からの知識の獲得を行なう．対象の事例集合は，クラスの分割が $C0, C1, C2, C3$ の 4 通り，属性が gcodon,lcodon の 2 通りの 8 種類を用いることにする (表 3).

-	C0	C1	C2	C3
gcodon	g-c0	g-c1	g-c2	g-c3
lcodon	l-c0	l-c1	l-c2	l-c3

表 3: 用いる事例集合

CV の hold 数	6
CV の繰り返し回数	10
BS 集合の事例数	400
BS の繰り返し回数	300
CV 各回での生き残り数	1

表 4: シミュレーションにおけるパラメータ

使用パラメータは表 4 の通りである．CV とは Cross-Validation，BS とは Bootstrap のことである．

このパラメータならば CV + BS で $6 \times 10 = 60$ 個のほぼ同等の性能を持った決定木が選択されるので，そこから安定して出現する決定木を選択し，それを学習結果として用いる．

5.3 シミュレーション結果

属性が gcodon のとき得られた決定木を図 8〜図 11 に示す． 決定木の見方は，各ノードの上部が分岐の属性として選ばれたコドンを表していて，下部が閾値になっている．その閾値以下なら左のパスへ，閾値以上なら右のパスへ流れる．つまり，g-c2 の根ノードなら，AAG-Lys が 4.57 %以上なら P(発現する)，以下なら ACT-Thr のノードへ流れ ··· と読んで行く．また，P・N の横に書かれている数値が，そのノードに分類された事例の数である (Bootstrap を行なっているため P と N の比が一致しないことがある).

今回のシミュレーションにおいて，属性として lcodon を用いた事例では，学習が不安定であらゆる結果が生成され結果の解釈が困難だったため，その結果は付録に記載し考察は次回に託すことにする．

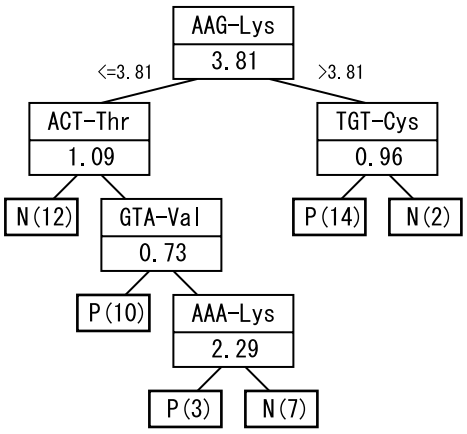

図 8: g-c0 事例における学習結果

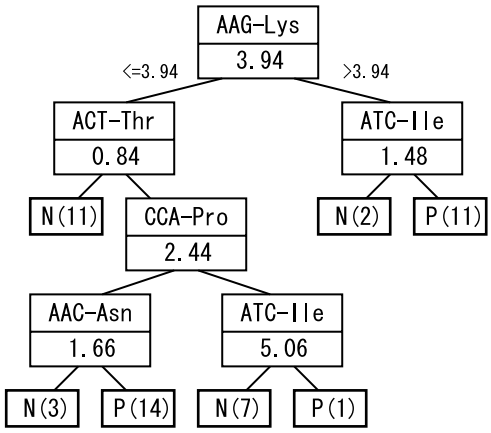

図 9: g-c1 事例における学習結果

5.4 考察

5.4.1 シミュレーション結果の考察

属性として gcodon を選んだときの学習結果はかなり安定しており，生成される決定木のサイズも lcodon を用いたときと比較しコンパクトになっているので，属性の選択としては gcodon が優良であると考えられる．gcodon 側では興味深い学習結果が見られる．$C0$ 〜$C3$ のどの学習結果も “AAG-Lys” を根ノードに持つ決定木になっているのである．発現のルールとして見ると，

Lo	Hi	Codon	Amino acid	Lo	Hi	Codon	Amino acid	Lo	Hi	Codon	Amino acid	Lo	Hi	Codon	Amino acid
72	24	TTT	Phe	39	41	TCT	SER	67	13	TAT	Tyr	60	7	TGT	Cys
28	76	TTC	PHE	19	52	TCC	SER	33	88	TAC	TYR	40	93	TGC	CYS
26	5	TTA	Leu	18	0	TCA	Ser								
23	34	TTG	LEU	9	0	TCG	Ser							TGG	Trp
32	39	CTT	LEU	56	50	CCT	PRO	74	19	CAT	His	32	88	CGT	ARG
8	22	CTC	LEU	18	50	CCC	PRO	26	81	CAC	HIS	12	12	CGC	Arg
8	0	CTA	Leu	19	0	CCA	Pro	76	88	CAA	GLN	21	0	CGA	Arg
8	0	CTG	Leu	7	0	CCG	Pro	24	12	CAG	Gln	11	0	CGG	Arg
66	51	ATT	ILE	47	41	ACT	THR	63	13	AAT	Asn	10	4	AGT	Ser
18	49	ATC	ILE	30	59	ACC	THR	37	87	AAC	ASN	9	4	AGC	Ser
16	0	ATA	Ile	19	0	ACA	Thr	66	2	AAA	Lys	17	0	AGA	Arg
		ATG	Met	8	0	ACG	Thr	34	98	AAG	LYS	7	0	AGG	Arg
53	42	GTT	VAL	48	50	GCT	ALA	76	32	GAT	Asp	40	87	GGT	GLY
12	58	GTC	VAL	22	47	GCC	ALA	24	68	GAC	Asp	16	13	GGC	Gly
21	0	GTA	Val	22	4	GCA	Ala	67	23	GAA	Glu	27	0	GGA	Gly
14	0	GTG	Val	7	0	GCG	Ala	33	78	GAG	GLU	16	0	GGG	Gly

表 5: 文献 [3] におけるコドンテーブル

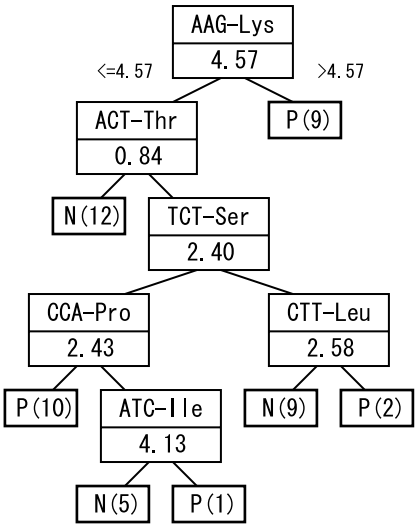

図 10: g-c2 事例における学習結果

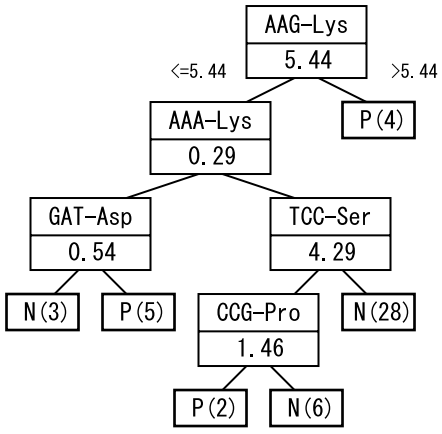

図 11: g-c3 事例における学習結果

コドン AAG が配列中にある比率以上含まれると発現する

ということを意味している．しかも，$C0 \rightarrow C3$ に移行するに連れて，ノード “AAG-Lys” で分岐するときの閾値が大きくなっている．このことから次の発現ルールが推測される．

コドン AAG が配列中にある比率以上含まれると発現し，その比率が増加する程大量に発現する

上記の結果に対し，発現実験のエキスパートに意見を求めたところ，今までは宿主細胞コドンテーブルを手がかりにする場合，重要なコドンが複数存在しどれを選択したら良いか決め兼ねていたのが,「まず AAG コドンを増やせば良い」そして,「どの程度増やせば良いかが明確」など，変更すべきコドンの順位付けと変更量の提示をしてくれるので非常に面白い，という意見を得た．

本研究では，このように決定木の表現を用いることで，テーブル形式の表現とは異なり，よりユーザーの使いやすい知識の提供を実現できている．しかしこの結果は決定木の根ノードのみの比較・考察であって，

コドン同士の依存関係の存在は，この結果だけでは判断できない．そこで，依存関係を立証するには，決定木で得られたコドンの依存関係の知識を利用した生物(発現) 実験を行なう必要があると考えられ，それは今後実現して行く予定である．

表5に文献 [3] に示されている *S.pombe* のコドンテーブルを抜粋する．このコドンテーブルは発現量の低いと思われる遺伝子 (Lo) のコドン使用率と，発現量の高いと思われる遺伝子 (Hi) のコドン使用率を調べたものである．Hi で偏りが顕著になっているコドンのアミノ酸の欄は，文字が大文字になっている (このコドンが多いと発現量が多くなると推測される)．このテーブルと本研究で得られた決定木を意味的な比較したところ，ほぼコドンテーブルの内容と一致しており (色付きのところが決定木の知識と一致したところである)，決定木が示す結果は従来の知見を支持するものとなっている．このことから，獲得した決定木は意味のある結果を学習していると推測され，かつ信頼できると考えられる．

本稿では属性を lcodon にしたときの結果の考察を省略したが，gcodon では同義コドンの情報とアミノ酸組成の情報の足し合わせの情報であるのに対し，lcodon では同義コドンの情報を直接表現しいるので，こちらの考察も重要であると思われる．本稿の結果では lcodon の結果は解釈の行ない難いものであったが，事例数の増加に伴い結果が安定してくる可能性もあるので今後の解析に期待したい．

5.4.2 属性選択における考察

本稿では，決定木に用いる属性にコドン使用率を使用したが，遺伝子配列からは他にも多くの新規属性の提案が考えられる．近年，2 つのコドンを一単位とした dicodon による遺伝子解析が考案されている．また，隣り合うコドンで上流にある 3 文字目の塩基と下流の 1 文字目の塩基間の依存関係 (context effect) が注目されており，隣り合うコドン間には何らかの重要な関係があると考えられる．このような配列の連続性を活かした新たな属性を配列より抽出し，予測システムに適用することで新たな結果の獲得が期待できる．

また，プロモータやシグナルといった，発現実験において付加される制御配列の選択の問題の解決も重要である．本稿で使用した事例以外に，制御配列を色々試行した事例が多量に存在するので，それを用いて制御配列も選択してくれる診断モデルの提案が可能と思われる．制御配列の選別を同時に考えるとき，本稿で用いた決定木の属性として導入するか，他のモデルを用いて導入するかは迷うところである．

5.4.3 診断モデルに対する考察

本研究では実験事例から決定木の帰納学習を行なったが，これは生物内部の構造・現象を特に考慮に入れずに発現予測上都合のよい形で予測モデルを提案した感がある．そこで，より宿主細胞内で起きている事象を詳細に表現するために，遺伝子の翻訳過程を確率モデルとして表現することが考えられる．コドンがアミノ酸へ翻訳されるのは，コドンに対応する tRNA が細胞内に満たされているからである．つまりあるコドンの翻訳される確率は，対応する tRNA の細胞内での濃度に比例するものと考えられる．tRNA 濃度が低くなかなか翻訳されない場合，翻訳の読み枠を外れてフレームシフトを起こすといわれている．そうなると，本来読まれない読み枠なのですぐに終止コドンが現れ翻訳が停止したり，合成が進んでも全く異なるタンパクが生成されターゲットタンパク質の発現は失敗するものと思われる．つまり，宿主細胞内の tRNA 濃度が測定できれば，それを元に確率モデルを構築することが出来，過去の発現実験事例によって学習を行ない宿主細胞内の翻訳過程を模擬することが可能になる．

6 まとめと今後の展望

異種外来タンパク質の発現実験の現場において問題になっている人件費と時間の膨大なコストを削減するために，タンパク質の発現を予測するシステムの実現を行なった．実際には，決定木の帰納学習を異種外来タンパク質の発現実験事例に適用することにより，従来にはなかった発現予測システムを実装し，過去の外来タンパク質の実験事例の有効な利用を可能し，異種外来タンパク質の発現問題に有効と思われる仮説の生成を行なった．決定木を知識の表現方法として適用したことで，コドンなど特徴量の依存関係を表現することが可能となった．また，発現予測システムに相応しい知識表現は，決定木の適用で可能となったが，その有効な知識利用法については触れなかった．本研究で

得られた結果は，エキスパートにとっても非常に興味深くかつ，従来の知見とほぼ一致しており，外来タンパク質の発現問題に対する有効な知識を獲得出来たと考えられる．しかし，コドン依存関係の存在の立証は行なっていないので，その立証には生物実験での確認が必要と思われる．

今後の展望として，まず，コドンの依存関係が存在するか否かの生物実験を行なう予定である．そして，決定木の帰納学習で得られた木構造で表現された知識 (仮説) の利用方法の提案を考えている．本研究では，従来法ではフラットになっていた結果から木構造で結果を得られるようになったことで，新しい視点が開けると考えられる．まず行ないたいのが，発現しないと判定された遺伝子を最小のコドン変更で，発現へと移行する方法である．これは，決定木上の最短パスと遺伝子のコドン使用率を考慮に入れれば簡単に実現できると思われる．また，遺伝子の上流ではコドン使用率の偏りが顕著になるという話がある．これを利用するれば，更なる最小のコドン変更で発現を促すことが可能になるかも知れない．全く異なった視点として，決定木表現による新しい生物間のコドン使用パターン解析を考えている．生物毎のコドン使用パターンの解析は長年行なわれてきたが，コドン同士の依存関係を調べ上げた研究は少ない．決定木表現を用いたコドン解析ではコドンの依存関係を容易に表現できることから，今まで発見されてこなかった階層的なコドン使用パターンの発見・解析が可能である．そこから導かれる生物種間の階層的なコドン使用パターンの比較は興味深い．

このように，異種外来タンパク質の発現予測は様々な発展が期待でき，かつ現場での需要も高く今後の発展が期待される分野である．

謝辞

本研究を行なうにあたり終止多大なる御指導ならびに御教示を頂きました山村雅幸助教授に深く感謝の意を表します．また，貴重な実験データならびに御教示を頂きました旭硝子 (株) の磯合敦博士と旭硝子 (株) の皆様に深く感謝すると同時に，興味深い御意見ならびに御指摘を頂いたかずさ DNA 研究所の中村保一氏と山形大学の金谷重彦助教授に深く感謝致します．本研究を進める上で，色々と相談に乗って頂いた山村研究室の皆様に深く感謝致します．最後に学生生活を暖かく見守ってくれた両親に心から感謝します．

参考文献

[1] Y.Giga-Hama,H.Kumagai, Foreign Gene Expression in Fission Yeast Schizosaccharomyces pombe, 1997 by Springer.

[2] 美宅成樹, 金久實, ヒトゲノム計画と知識情報処理, 培風館, 1995.

[3] Molecular Biology of The Fisson Yeast, 1989 by Academic Press, Inc.

[4] Shigehiko Kanaya, Yoshihiro Kudo, Yasukazu Nakamura, and Toshimichi Ikemura, Detection of genes in Escherichia coli sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage, CABIOS Vol12, no.3, pp213-225, 1996.

[5] J.R.Quinlan, Induction of Decision Trees, Machine Learning 1, pp81-106, 1986.

[6] J.R.Quinlan, C4.5 : Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[7] 秋葉康弘, フセイン・アルモアリム, 金田重郎, 例からの学習技術の応用に向けて, 情報処理学会誌, Vol.39, No.3, pp245-251, 1993.

[8] Ron Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, International Joint Conference on Artificial Intelligence(IJCAI), 1995.

A　付録

lcodon を属性として選んだときの結果を図 12 図〜15 に記載する．学習結果はあまり安定しておらず，これ以外にも多くの異なる決定木が生成された．どの決定木も gcodon を属性としたときよりも比較的サイズが大きく複雑なものになっている．しかしながら，$C0$ と $C3$ に共通に見られる，コドン AGG のルールは従来のテーブル形式の表現では表現できない，決定木ならではの面白い結果になっている．

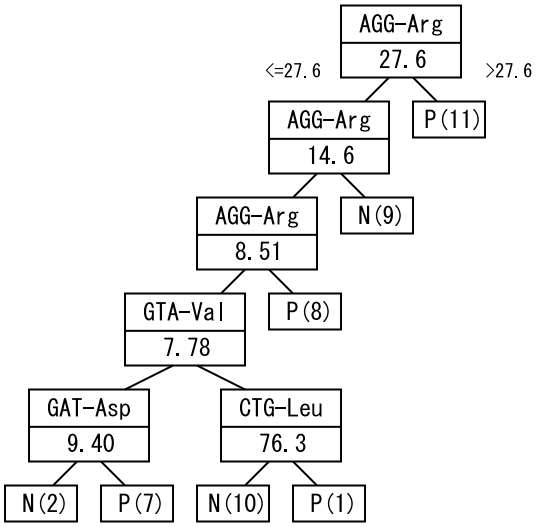

図 12: l-c0 事例における学習結果

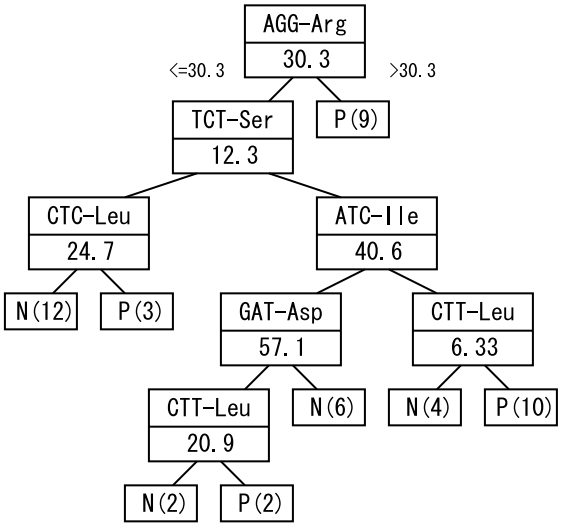

図 13: l-c1 事例における学習結果

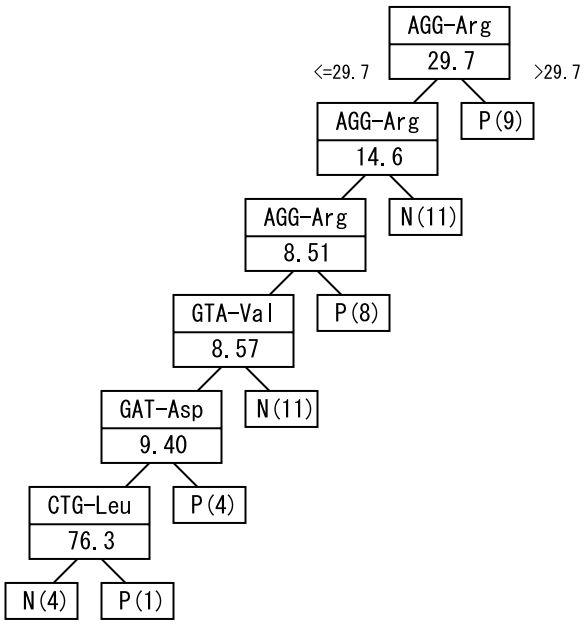

図 14: l-c2 事例における学習結果

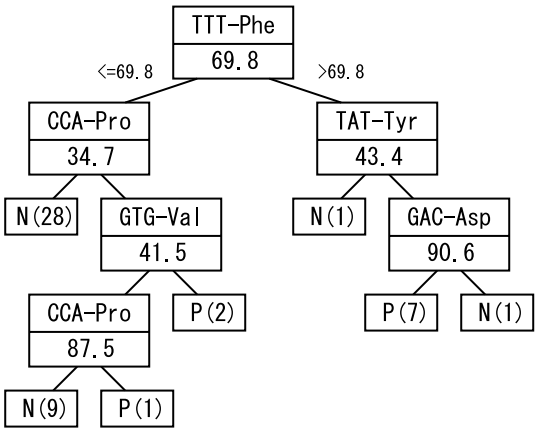

図 15: l-c3 事例における学習結果