データベース・システム 80－9
人 工 知 能 73－9
（1990. 11. 9）

遺伝子情報知識ベースシステムの構想について

小長谷 明彦†, 新田 克己‡, 山西 健司†

†日本電気株式会社, ‡新世代コンピュータ技術開発機構

第五世代計算機プロジェクトでは、並列知識処理技術と知識ベース管理技術の応用の一つとして、遺伝子情報処理の専門家向け実験環境システムの開発を進めている。本稿では、この実験環境システムの中核をなす遺伝子情報知識ベースシステム構築の構想について述べる。遺伝子情報知識ベースは、遺伝子およびタンパク質に関する実験結果から得られた経験的知識と、経験的知識から学習された知識（モチーフ）からなる。モチーフの学習には確率的分類学習法を用い、また、“良い”モチーフの基準として、未知配列に対する予測精度向上の観点から記述長最小（MDL）基準を採用する。

The Conceptual Design of a Genetic Information Knowledge Base

Akihiko Konagaya†, Katsumi Nitta‡, Kenji Yamanishi†

†NEC Corporation, ‡ICOT

This paper describes the conceptual design of an experimental genetic information knowledge base system. The system is being developed in the fifth generation computer systems project, as one of the main applications of parallel knowledge information processing technology and knowledge base management technology. The knowledge base consists of two types of knowledge; experimental data and induced knowledge (motifs). The experimental data contain sequence data, structure data, and features concerning genes and proteins. A method for learning stochastic rules is used for extracting the induced knowledge from the experimental data. In the learning method, the minimum description length (MDL) criterion plays an essential role in evaluating a good motif.

1 はじめに

近年、遺伝子工学の発展により、DNA配列、アミノ酸配列、タンパク質などの遺伝子関連情報が急増し、高性能かつ高精度の解析手法が求められている。このような解析手法の一つとして、第五世代計算機プロジェクトでは、遺伝子およびタンパク質に関する経験的知識と、帰納的知識からなる知識ベースの構築を進めている [7, 16, 8]。ここで、経験的知識とは、生物学実験から得られたデータおよび知見であり、帰納的知識とは、経験的知識から統計的手法を用いて抽出された知識をいう。また、本稿では、帰納的知識として特にモチーフに注目する [6, 4]。モチーフは、ある特定の集団(カテゴリ)において共通に見い出せる規則であり、本稿のアプローチの特徴は、モチーフをカテゴリを抽出するための確率的な分類規則 [19, 18] とみなす点にある。また、モチーフの抽出においては筆者らが以前から提案しているように、未知データに対する予測精度の向上が理論的に保証されている記述長最小(MDL)基準を適用する [8]。

なお、本稿の構成は以下の通りである。2節では、遺伝子情報処理の背景として、遺伝子ならびにタンパク質の概要、遺伝子情報処理研究の重要性、遺伝子情報処理の現状と課題、知識情報処理の有用性および遺伝子情報知識ベースの位置付けについて述べる。次に3節において、配列モチーフを例に遺伝子情報知識ベースで扱う経験的知識について説明する。また、4節では、記述長最小(MDL)基準を用いて帰納的知識が有効に抽出できることを配列モチーフを例として示し、その理論的根拠を付録に示す。

2 遺伝子情報処理の背景

2.1 遺伝子と蛋白質

生物の体は細胞からできている。真核生物の場合、各細胞の核には染色体がある。染色体はDNAから構成され, D NAは4種類の核酸 (A,T,G,C) からなる2重らせん構造を持つ。この核酸の配列の中に遺伝子が含まれている。

DNAのらせんは、ほどけて遺伝子情報がmRNAに転写される。このとき、不要な部分が切り落とされて必要な部分だけがつなぎ直されるという編集操作 (スプライシング) がおきる。mRNAの配列情報は細胞内のリボソームにて翻訳され、アミノ酸(20種類)の列が合成される。アミノ酸の列は[らせん]や[シート]などの部分構造を形成し、さらに折り畳まれて複雑な立体構造をとり蛋白質となる。

蛋白質の性質(機能)はその立体構造に依存する。例えば、酵素は特定の化学反応を促進する触媒作用を持つが、これは酵素の特定の部位に、他の分子がはまりこむことにより物理的に接近し、反応を起こしやすくするからである [17]。

遺伝子情報処理の課題は、これらの一連の処理を解明し、未知の遺伝子情報が与えられたときに遺伝子としての機能および遺伝子情報から生成されるタンパク質の構造および機能を予測することにある。

2.2 遺伝子情報処理研究の重要性

遺伝子情報処理の研究は現実に重要な応用がたくさん存在し、社会的にもその意義がある。 その例を以下に挙げる。

1) 新しい蛋白質の設計に使える可能性がある。例えば、AIDSワクチンなどの新物質の設計ができる可能性がある。

2) 遺伝病や癌などの原因解明や治療の糸口の発見につながる。例えば、鎌型赤血球貧血病の原因は、ヘモグロビンを構成するたった一つのアミノ酸が突然変異したために赤血球の形状が鎌型に変形することに起因する。

3) 生命の起源の解明につながる。例えば、DNAや蛋白質の比較により、生物の進化のメカニズムを研究することが可能である。

2.3 遺伝子情報処理研究の現状と課題

近年、遺伝子操作技術の発展により、多数の遺伝子の配列情報ならびに蛋白質のアミノ酸配列(1次構造)が文献および配列データバンクに発表されるようになった [12, 2]。文献に発表されるDNA配列を集めたデータベースとして、GenBank(Los Alamos研究所)やEMBL(欧州分子生物学研究所)などがあり、我が国でもDDBJ(国立遺伝学研究所)が配列情報の収集、データバンク化を進めている。また、蛋白質データベースとしては、蛋白質の1次構造のPIR(アメリカ生物医学研究財団)がある。

配列情報に関する最大の課題は、そのデータ数の増加に対する対応である。例えば、GenBankに含まれるDNAの1次構造のデータは、塩基数にして1982年に50万であったものが、1985年には500万となり、1990年には約43000配列(4250万塩基)と増えている。また、近年、ヒトゲノム解析プロジェクト[3]に代表されるように生物種の持つDNA(ゲノム)を全て解析しようというプロジェクトが活発化してきている。ヒトの持つゲノムはおよそ30億塩基あると言われており、今後配列情報はさらに増加することが見込まれるため、より効率的なデータ管理モデルならびに検索手法が求められている。

一方、構造情報に関しては、PDB(Brook haven国立研究所)がRNAならびにタンパク質の立体構造を提供している。立体構造情報は、配列と構造および機能を結び付けるうえで貴重な情報であるが、X線解析のための精度の高い結晶の作成ならびに構造解析の手間が大きいため、現在までに高々数百件のデータしか得られていない。ただし、比較的分子量の小さいタンパク質に関しては核磁気共鳴装置(NMR)を利用した立体構造情報解析手法が開発されており[9]、今後の成果が期待される。構造情報に関しては、実験からのデータの入手が困難なため、データ数の多い配列データからの予測が最大の課題となっている。

2.4 知識処理技術の有用性

遺伝子情報処理における知識処理技術の有用性は、知識の利用による計算量の減少および高機能化にある。その具体例として類似配列検索(ホモロジー検索)と構造予測問題をあげる。ホモロジー検索は、未知配列を与えたときに共通の先祖を持つ配列をデータバンクから捜し出す操作である。従来手法では、動的プログラミング法等、配列情報を直接比較する方法[5]がとられているため、配列データ数の増加に伴い検索時間が非効率になりつつある。これに対し、共通の祖先を持つ配列の集合から規則性を帰納的知識として抽出した場合には、帰納的知識を検索するだけで済むため検索時間の大幅な改善が期待できる。また、帰納的知識は遺伝子やタンパク質の持つ機能情報や構造情報などの経験的知識とを関連づけることができ、従来のアミノ酸の並びによる検索よりもより精密で、生物学的に意味のある検索が期待できる。

構造予測問題とは、配列情報からRNAあるいはタンパク質の構造を予測する問題であり、分子軌道法、分子動力学[13]といった数値解法を利用した計算アプローチと生物実験の経験的知識を利用した知識アプローチがある。計算アプローチでは、条件さえ整えば原理的には化学反応を厳密にシミュレーションすることが可能であるが、分子数の多いタンパク質の構造予測に関しては計算量が膨大となり現在のコンピュータの計算能力では十分とは言い難い。

これに対し、知識アプローチでは、配列情報や構造情報に関する帰納的知識あるいは、アミノ酸の性質に関する経験的知識を用いて立体構造を予測する。知識アプローチにおいて、帰納的知識の抽出には学習アルゴリズムが適用でき、予測にはプロダクションシステムなどの推論技法を適用することができる。構造予測問題は、非常に難しく、知識情報処理技術を利用したからといってすぐに解決する訳ではないが、経験的知識および帰納的知識を活用することにより、現在の計算機の能力の範囲内でより良い近似解を求めることが期待できる。

2.5 遺伝子情報処理プロジェクト

第五世代計算機プロジェクトでは、遺伝子情報知識ベースを中心に遺伝子情報処理専門家向け実験環境の実現を目指している。このような環境を実現するために、以下の研究開発を行なう。

- 配列データ、構造データなどの経験的知識からの配列モチーフ、構造モチーフ、機能モチーフなどの帰納的知識の抽出法。
- 経験的知識と帰納的知識を統合化し、より高度な情報(演繹情報)を提供するためのデータベースモデル。
- 経験的知識および帰納的知識を活用した、ホモロジー検索システム、2次構造予測システムなどの解析システム。

これらのシステムの特徴は、モチーフを確率的な分類規則として抽出することにより、推論結果を定量的に評価できる点にある。しかしながら、これらのシステムが本当に有用となるか否かは、遺伝子情報からどのような知識が抽出できるかに大きく依存する。このような知識の例を以下に示す。

3 遺伝子情報処理における知識

遺伝子情報処理における知識には、大きくわけて、個々の遺伝子あるいはタンパク質に関する生物実験から得られた経験的知識と、経験的知識を統計的手法を用いて解析することにより得られる帰納的知識がある。

経験的知識は、その大半は文献の形でしかアクセスすることはできないが、近年、配列情報、立体構造情報、系統樹情報およびその他特徴的な機能については、電子化が進められつつある。この、経験的知識を統計的手法を用いて解析することにより得られる帰納的知識の代表としてモチーフがある。モチーフは、共通の先祖あるいは機能を持つ個々の遺伝子あるいはタンパク質の間で進化的に保存されてきた情報であり、保存する対象に応じて配列モチーフ、構造モチーフ、機能モチーフと呼ばれる。次に、シトクロムCを題材にモチーフの例を示す。シトクロムCは、多種の生物について調べられており、多数のヌクレオチド配列データ、アミノ酸配列データ、タンパク構造データが現在揃っている。

3.1 配列の保存部位と機能部位

シトクロムは呼吸における一連の酸化還元処理において電荷を運ぶタンパク質のグループの総称であり、シトクロムCは、C型のヘムと結合するタンパク質をいう。また、多細胞動物のシトクロムCはミトコンドリア内に存在するため、ミトコンドリアシトクロムCと呼ばれる。

一般に、突然変異は全てのアミノ酸に対して等確率で発生しているといわれているが[10]、タンパク質の機能を変化させるような突然変異は淘汰されてしまうため、結果的に、重要な役割を果たしているアミノ酸ほど進化的に保存される傾向を持つ。

実際、図1に示すように、共通の先祖を持つタンパク質の間には共通のアミノ酸のパターン(保存部位)を見い出すことができる。ただし、図1はミトコンドリアシトクロムCのアミノ酸配列のマッチング結果(配列アライメント)の一部分を示したものであり、分子生物学の慣例により個々のアミノ酸はアルファベット1文字で表されている。

配列アライメントにより発見される保存部位の多くは、タンパク質として作用するために必要な部位(機能部位)と対応づけることができる。例えば、ミトコンドリアシトクロムCでは、*"CXXCH"*、*"GPXLXG"*、*"PGTKM"*という3つの保存部位を見い出すことができ、これらは機能部位とは以下の関係を持つ[1]。*"CXXCH"*において、2つのシステイン(C)は結合部位(binding site)と呼ばれ、側鎖がヘムのビニル基の間でチオエステル結合を形成することによりヘムCと結合している。また、ヘムCは中央に電荷を運ぶための鉄原子を持つが、鉄原子はヘムCに結合する以外に、1つは保存部位(*"CXXCH"*)のヒスチジン(H)のイミダゾール環の窒素原子(第5配位子)と、もう一つは保存部位(*"PGTKM"*)のメチオニン(M)の硫黄原子(第6配位子)と結合している[11]。したがって、ミトコンドリアシトクロムCにおいては、*"CXXCH"*および*"PGTKM"*の2つの保存部位が電荷を運ぶ(ヘムと結合する)という機能を実現するために保存されたと推測することができる。

3.2 配列モチーフ

配列モチーフの定義には様々なものがあるが[6, 4]、我々のアプローチの特徴は、配列モチーフを特定のカテゴリ(例えばシトクロムC)を識別するための確率的分類規則とみなす点にある[8, 19, 18]。このように配列モチーフを定義した際の第一の課題は、分類条件として何をどこまで採用するかであり、第二の課題は、突然変異情報の扱いにある。

前節で述べたように、タンパク質の作用において特定のアミノ酸(機能部位)が特別な役割を果たしている。したがって、機能部位はモチーフの分類条件として極めて重要な意味を持つ。実際、ミトコンドリアシトクロムCの場合、機能部位(2つのシステイン(C)とヒスチジン(H))から構成される保存部位、すなわち、*"CXXCH"*を持つか否かでアミノ酸配列データバンク(PIR21.0版)を検索すると、データバンク全体(6158例)中、保存部位*"CXXCH"*を含むアミノ酸配列は189例存在し、そのうちミトコンドリアシトクロムCに属する配列は67例ある。また、ミトコンドリアシトクロムCで*"CXXCH"*を含まないアミノ酸配列は3例存在する。すなわち、*"CXXCH"*を含めば確率 $\frac{67}{189}$ でミトコンドリアシトクロムCであり、含まなければ確率 $\frac{6158-189-3}{6158-189}$ でミトコンドリアシトクロムCではないという確率的な対応関係が得られる。

[1] X は任意のアミノ酸を表す。

```
CCFS     --ASFAEAPAGDPTTGAKIFKTKCAQCHTVEKGAGHKQGPNLNGLFGRQSGTTAGYSYSAANKNMAVIWEENTLYDYLLNPKKYIPGTKMVFPGLKKPQERADL
CCLK     --ATFSZAPPGBZKAGQKIFKLKCAQCHTVEKGAGHKQGPNLNGLFGRQSGTAAGYSYSAANKNMAVVWZZBTLYDYLLNPKKYIPGTKMVFPGLKKPQDRADL
CCNG     --ASFAEAPAGDAKAGEKIFKTKCAZCHTVZKGAGHKQGPNLNGLFGRQSGTTAGYSYSAANKNKAVAWZZBSLYDYLLNPKKYIPGTKMVFPGLKKPZZRADL
CCSP     --ATFSEAPPGWKDVGAKIFKTKCAQCHTVDLGAGHKQGPNLNGLFGRQSGTAASYSYSAANKNKAVIWSEDTLYEYLLNPKKYIPGTKMVFPGLKKPQDRADL
CCND     --ASFBZAPAGBSASGEKIFKTKCAZCHTVBZGAGHKZGPNLHGLFGRQSGTVAGYSYSAANKNKAVNWEEKTLYDYLLNPKKYIPGTKMVFPGLKKPZZRABL
CCGK     --ATFSEAPPGDPKAGEKIFKTKCAZCHTVZKGAGHKQGPNLHGLFGRQSGTTAGYSYSTGNKNKAVNWGZZTLYEYLLNPKKYIPGTKMVFPGLKKPZZRADL
CCEI     --STFABAPPGBPAKGAKIFKAKCAZCHTVBAGAGHKQGPNLNGAFGRTSGTAAGFSYSAABKBKTADWBZBTLYDYLLNPKKYIPGTKMVFAGLKKPZBRADL
CCEG     ----------GDAERGKKLFESRAAQCHSAQKGV-NSTGPSLWGVYGRTSGSVPGYAYSNANKNAAIVWEEETLHKFLENPKKYVPGTKMAFAGIKAKKDRQDI
CCCRCO   PPKAREPLPPGDAAKGEKIFKGRAAQCHTGAKGGANGVGPNLFGIVNRHSGTVEGFAYSKANADSGVVWTPEVLDVYLENPKKFMPGTKMSFAGIKKPQERADL
CCCRCF   .PPKARAPLPPGDAARGEKLFKGRAAQCHTANQGGANGVGPNLYGLVGRHSGTIEGYAYSKANAESGVVWTPDVLDVYLENPKKFMPGTKMSFAGMKKPQERADV

         .........G....G...F......CH..........GP.L.G...R..G......Y..........W.........L..P.K..PGTKM.F.G......R...
```

図 1: ミトコンドリア シトクロム C の配列アライメント結果の一部

この対応関係は、分類精度としては、決して悪くはないが、ミトコンドリアシトクロム C を識別する分類条件としては改良の余地がある。一つの方法は、前節で示したように、ミトコンドリアシトクロム C の配列を並べたときに見つかる保存部位の情報を分類条件に加える方法である。ここで、問題となるのは、どこまで保存部位の情報を分類条件に加えるかである。保存部位の情報を加えれば加えるほど現在のデータバンクにおける分離精度は高まる。しかしながら、現在の保存部位が見知の配列データにおいても保存されている確証はなく、必要以上に分類規則を厳しくした場合にはかえって未知配列に対する分類精度が落ちるという過剰学習の可能性がでてくる。すなわち、保存部位を分類条件に加えるには分類にあたってどこまで本当に必要かのバランスを考慮する必要があり、この問題を解決するために、分類基準として記述長最小 (MDL) 基準を用いる。

また、突然変異の問題点は、機能部位は一般的に保存性が高いが、機能部位においても突然変異が起きる可能性があることにある。実際、シトクロム C においては、CCEG(ミドリムシ)、CCRCO(Crithidia Oncopelti)、CCRCF(Crithidia Fasciculata) の 3 つは、ヘム C との結合を行なうシステイン (C) が一つしかなく、システイン (C) がアラニン (A) に置き換わっている。したがって、配列モチーフ等の抽出においては突然変異の可能性を含めて真に保存されている情報を見つける必要があり、また、配列モチーフの表現法は、このような突然変異情報を含めてカテゴリと分類条件の確率的対応関係を十分に表現できる必要がある。

3.2.1 確率的決定述語

配列モチーフは、分類条件と対象とするカテゴリとの確率的対応関係であるが、このような、確率的対応関係の表現法として確率的決定述語を提案している [8]。確率的決定述語は、配列モチーフにみられるような分類条件の記述が容易であり、また、論理型言語での実現が容易という特徴を持つ。確率的決定述語の一般的な構造を以下に示す。

```
motif(S,C1) with p1 :- Q11,...,Q1n1.
motif(S,C2) with p2 :- Q21,...,Q2n2.
    ...
motif(S,Cm) with pm.
```

確率的決定述語は複数のクローズからなり、i 番目のクローズは、配列 S が分類条件 $Q_{i1},...,Q_{ini}$ を全て満足する時、確率 p_i でカテゴリ C_i に分類されることを示す[2]。ここで、Q_{ij} としてとり得るのは $R_1;...;R_l$ の形式の分類条件を表す述語の OR- 結合である。なお、本稿では、記述長の計算を簡潔にするために分類条件の記述として (アミノ酸配列 S が保存部位 P を含むか否かを判定する述語 *contain(P,S)* のみを用いる。分類条件としては保存部位の前後関係や距離情報およびアミノ酸の類似関係などの情報を活用できるように拡張することが可能である。また、最終クローズは配列 S がどの分類条件も満足しないとき、確率 p_m でカテゴリ C_i に分類されることを示す。確率的決定述語はデータバンクを 3 つ以上のカテゴリに分類する配列モチーフについても表現することが可能であるが、本稿では、2 つのカテゴリ (対象とするカテゴリとそれ以外) の分類の場合についてのみ述べる。

- モチーフ例 1

[2] ここで p_i は確率的決定述語の真の確率パラメタを意味するが以下に述べる例では p_i の推定量として Bayes 確率を用いる。

```
motif(S,mitocondoria_cytochrome_c)
   with 68/191
   :- contain("CXXCH",S).
motif(S,others) with 5967/5971.
```

S が *"CXXCH"* と一致する部位を含めば確率 $\frac{68}{191}$ で S はミトコンドリアシトクロムCであり、そうでなければ、確率 $\frac{5967}{5971}$ で *others* である。

- モチーフ例2

```
motif(S,mitocondoria_cytochrome_c)
   with 68/75
   :- contain("CXXCH",S),
      contain("PGXKM",S).
motif(S,others) with 6083/6087.
```

S が *"CXXCH"*、*"PGXKM"* の両方と一致する部位を含めば S は確率 $\frac{68}{75}$ でミトコンドリアシトクロムCであり、そうでなければ確率 $\frac{6083}{6087}$ で *others* である。カンマで区切られた分類条件は *AND*-結合を表す。

- モチーフ例3

```
motif(S,mitocondoria_cytochrome_c)
   with 71/78
   :- (contain("CXXCH",S);
       contain("AAQCH",S)),
      contain("PGXKM",S).
motif(S,others) with 6083/6084.
```

S が *"CXXCH"* または *"AAQCH"* と一致する部位を含み、かつ、*"PGXKM"* と一致する部位を含めば確率 $\frac{71}{78}$ でミトコンドリアシトクロムCであり、そうでなければ確率 $\frac{6083}{6084}$ で *others* である。セミコロンで区切られた分類条件は *OR*-結合を表す。

4 MDL基準を用いた配列モチーフの抽出

遺伝子情報知識ベース構築のための最大の課題は、不確実性を含むデータからいかにして最良な知識(モチーフ)を抽出するかにある。モチーフ抽出においては、すでに与えられたデータバンクをいかに精密に分類するかという分類精度よりも、未知のデータが与えられたときにそれを正しく分類するための予測精度が重要となる。このような観点から、モチーフ抽出に関しては、予測精度を最大にすることが理論的に保証されている記述長最小(MDL)基準[15, 14, 18]を採用する。以下、MDL基準の配列モチーフ抽出への適用について文献[8]に従って述べる。

4.1 MDL基準

MDL基準は、不確実性を含む学習セットからの確率モデルの推定を行なう際に有効な基準であり、確率モデルの記述長と確率モデルを用いたときの学習セットの記述長(本稿では不確実性の記述長と呼ぶ)の和を最小にする確率モデルを最良の確率モデルとみなす考え方である。すなわち、MDL基準を配列モチーフ抽出に適用すると、

確率的決定述語の記述長 + 不確実性の記述長

を最小にするような確率的決定述語が最良の配列モチーフであるという選択基準が得られる[3]。ここで、記述長とは一意に復号可能な(Kraftの不等式を満たす)符号化をした際のビット長をいう。また、不確実性の記述長とは確率的決定述語による分類の後に残る不確実性を確率的に表現した場合の記述長であり、確率的決定述語の適合度が大きいほど小さな値を示す。一般に、複雑な確率的決定述語ほど記述長が長くなるような自然な符号化法を用いると、MDL基準により、与えられたデータ(配列データとカテゴリの組)に対する適合度と分類条件の複雑さのトレードオフを考慮した配列モチーフが抽出されることになる。なお、MDL基準に基づく確率的決定述語の学習の理論的根拠については付録を参照されたい。次に、記述長の具体的な計算方法を示す。

4.2 不確実性の記述長の計算法

不確実性の記述長の計算においては、確率的決定述語が配列モチーフが与えられたときのカテゴリに関する条件付き確率分布を定義することに注目すると、最短の不確実性の記述長はこの確率分布を用いて

$-\log$(確率的決定述語より定まるカテゴリの尤度)

として求めることができる(付録参照)。ただし、計算に必要な確率パラメタの値は学習に用いた配列データからの推定値を用いる。なお、本稿では、対数の底は全て2とする。

今、クローズの総数をm、i番目のクローズの分類条件を満足する配列の個数を N_i、このうち、これに対応させるカテゴリを C_i として、C_i に属していた配列の個数を N_i^+ ならびに C_i に属していない配列の個数を N_i^-、た

[3]生物学者はできるだけ長いモチーフを見つけようとする傾向がある。しかしながら、MDL基準に従えば、同程度の分離精度ならば短いモチーフの方が良いということが導かれる。

だし、$(N_i = N_i^+ + N_i^-)$ とする。また、i番目のクローズの分類条件を満足する配列がi番目のクローズで分類しようとしているカテゴリに属する真の確率を p_i とする。i番目のクローズの分類条件を満足する配列データの発生確率(尤度)は配列データの独立性を仮定すると、$p_i^{N_i^+} * (1-p_i)^{N_i^-}$ なので、この尤度を符号化するために必要な記述長は

$$-\log(p_i^{N_i^+} * (1-p_i)^{N_i^-})$$

ビットとなる。これを全てのクローズについて総和をとれば、次式を得る。

$$\sum_{i=1}^{m} -\log(p_i^{N_i^+} * (1-p_i)^{N_i^-})$$

ここで、p_i の推定量を $\hat{p}_i$, $\tilde{p}_i = \frac{N_i^+}{N_i}$ とすると、上記式は以下のように書き換えることができる。

$$UL = \sum_{i=1}^{m} N_i * \{H(\tilde{p}_i) + D(\tilde{p}_i||\hat{p}_i)\}$$

ただし、第1項のHはエントロピー関数であり、第2項のDはKullback-Leibler情報量であり、それぞれ次式で定義される。

$H(\tilde{p}_i) = -\tilde{p}_i \log \tilde{p}_i - (1-\tilde{p}_i)\log(1-\tilde{p}_i)$

$D(\tilde{p}_i||\hat{p}_i) = \tilde{p}_i * (\log(\tilde{p}_i) - \log(\hat{p}_i)) +$
$(1-\tilde{p}_i) * (\log(1-\tilde{p}_i) - \log(1-\hat{p}_i))$

また、$\hat{p}_i$ としては、通常は最尤推定量 $\hat{p}_i = \frac{N_i^+}{N_i}$ を用い、このとき第2項が0となるが、実際の適用においては N_i^+ あるいは N_i^- が0のとき、エントロピーが0となり、記述長が極端に減少するという問題が生じる。これを避けるため、Bayes推定法から導かれる偏りのある最尤推定量 $\hat{p}_i = \frac{N_i^+ + 1}{N_i + 2}$ を用いる。

4.3 確率的決定述語の記述長の計算法

確率的決定述語の記述長は大きく分けて確率パラメタの記述長と配列モチーフ自体の記述長からなる。配列モチーフ自体の記述長はさらにカテゴリの記述長と保存部位の記述長からなり、保存部位の記述長はcontain述語の記述長の和として計算できる。

確率パラメタの記述長は、推定量 $\hat{p}_i (i = 1, ..., m)$ の記述に必要な記述長である。確率パラメタは実数で与えられるので、記述長としては有効精度分だけあれば良い。$\hat{p}_i$ の精度はよく知られた最尤推定量の分散のオーダ評価を用いるとi番目のクローズの分類条件を満足する配列の個数 N_i を用いて $\mathcal{O}(1/\sqrt{N_i})$ で与えられ、$\frac{1}{2} * \log(N_i)$ ビットで記述できる。したがって、全てのクローズについては、

$$PL = \sum_{i=1}^{m} \frac{1}{2} * \log(N_i)$$

ビットが必要となる。

配列モチーフ自体の記述長は以下のようにして計算する。まず、とり得るカテゴリの総数をMとすると、各クローズ毎にカテゴリを指定するために $\log M$ ビットが必要となる。

次に、contain述語の記述長の求め方を以下に示す。保存部位に変数(すなわちX)が含まれていない場合には、保存部位の長さをLとすると20のN乗個の文字列から一つの文字列を選ぶことになるので、その記述長は $L * \log(20)$ ビット必要になる。保存部位中に変数が含まれる場合には変数の出現位置に関する記述長と保存部位の複雑さに関する記述長から求める。今、i番目のクローズのj番目のcontain述語の保存部位の長さを L_i^j、この中に含まれる変数の個数を X_i^j とする。保存部位中の変数の出現位置は L_i^j 個の位置から X_i^j 個を指定する組合せの個数だけある。一方、アミノ酸の文字の種類は20種類なので、符号化すべき文字列の総数は20の $(L_i^j - X_i^j)$ 乗個となる。したがって、全体では

$$ML = \sum_{i=1}^{m}\{\sum_{j=1}^{B_i}\{\log\left(\begin{array}{c} L_i^j \\ X_i^j \end{array}\right) + (L_i^j - X_i^j) * \log 20\} + \log M\}$$

ビットの記述長が必要となる。ただし、B_i はi番目のクローズに含まれるcontain述語の数である。また、述語のAND-結合およびOR-結合に必要な記述長は無視している。

以上により、与えられたデータに対して

$$UL + PL + ML$$

の値が小さな値を持つ配列モチーフほど、MDL基準の意味で良いモチーフである評価することができる。

4.4 適用例

ミトコンドリアシトクロムCの配列モチーフを表す4つの確率的決定述語について、各クローズ毎に検索対象となった配列数、分類条件を満足した照合配列数および正しくカテゴリを分類した正例数を表1に示す。また、この情報から計算した配列モチーフの記述長(ML)、確率

表 1: ミトコンドリアシトクロム C のモチーフを表現する確率的決定述語のクローズ毎の分類結果

保存部位	対象配列数	照合配列数	正例数
CXXCH	6158	189	67
others	5969	5969	5966
CXXCH and GPXLXG and PGTKM	6158	71	67
others	6087	6087	6084
CXXCH and PGTKM	6158	73	67
others	6085	6085	6082
(CXXCH or AAQCH) and PGTKM	6158	76	70
others	6082	6082	6082

表 2: ミトコンドリアシトクロム C を分別する配列モチーフから計算される記述長

配列モチーフ	ML	PL	UL	総計		$\hat{p}_1$	$\hat{p}_2$
CXXCH	36.2	10.1	225.5	271.7		0.356	0.9993
CXXCH and GPXLXG and PGTKM	80.0	9.4	73.6	163.0		0.932	0.9993
CXXCH and PGTKM	55.8	9.4	80.7	145.9		0.906	0.9993
(CXXCH or AAQCH) and PGTKM	77.4	9.5	47.1	134.0		0.910	0.9998

パラメタの記述長 (PL)、不確実性の記述長 (UL) および各クローズにおける確率パラメタの推定値 ($\hat{p}_1$,$\hat{p}_2$) を表 2に示す。ミトコンドリアシトクロム C の配列モチーフ抽出における MDL 基準による判断は以下の通りである。

保存部位 *"CXXCH"*を含むかどうかだけで分類する配列モチーフは ML が 36.2 ビットと少ないのに対して UL は 225.5 ビットと非常に大きく、配列モチーフとして単純すぎることを示している。また、保存部位 *"PGTKM"* を加えると、ML は 19.6 ビット増えるが、UL は 144.8 ビットも減少し、総記述長としては 125.2 ビット減少し、より理想的な配列モチーフに近づいたことがわかる。一方、保存部位 *"GPXLXG"*をさらに加えた場合には、ML が 24.2 ビット増加しているのに対し、UL はわずか 7.1 ビットしか減少しておらず、総記述長は逆に 17.1 ビット増加し、MDL 基準からは過剰適合の可能性があることが示唆されている。

すなわち、ミトコンドリアシトクロム C の分類する 3 つの配列モチーフにおいて、最尤推定法を用いた場合には 3 つの保存部位を全て含むかどうかを調べる配列モチーフが最良の配列モチーフとなるが、MDL 基準を用いた場合には生物学的に裏付けのある機能部位を含む保存部位 (*"CXXCH"*と *"PGTKM"*) の両方とも含むかどうかを調べる配列モチーフが最良となる。

また、表 2に示すように、ミトコンドリアシトクロム C で保存部位 *"CXXCH"*では分類できなかった配列に共通な保存部位 *"AAQCH"*を OR- 結合として加えると、記述長をさらに短くすることができる。この理由として、不確実性の記述長が照合配列数の多いクローズの正例数の変化に敏感なこと、確率的決定述語では保存部位の OR- 結合を別クローズに展開しないで表現できることがあげられる。保存部位 OR- 結合を別クローズに展開した場合には分類条件がコピーされるため ML が必要以上に増加する恐れがある。配列モチーフでは突然変異情報を表現する上で保存部位の OR- 結合を多用する傾向にあり、この意味からも、確率的決定述語は配列モチーフに適した構造となっている。

5 まとめ

知識処理を応用した遺伝子情報処理は世界的に見てほとんど例がなく、今後大いに研究される分野である。そこでは、経験的知識と帰納的知識の有機的な結合が必要であり、帰納的知識の抽出においては、確率的モデルの学習理論が本質的な役割を果たす。この意味で、遺伝子情報処理は知識処理技術の有効性を実証する最適な場といえる。さらに、大量のデータ (知識) を高速に処理する必要があることから、並列処理応用としても位置付けられ、知識の保持、検索の観点からはオブジェクト指向データベースあるいは演繹データベースとの研究とも関連づ

けられる。

また、帰納的知識の一つである配列モチーフの抽出に関して、記述長最小(MDL)基準が有効であることを示した。遺伝子情報処理においては、このような確率的な特徴抽出は不可避であり、今後は、配列モチーフ抽出法の改良を進めるとともに、MDL基準を構造モチーフの抽出や、立体構造の予測についても適用してゆく予定である。

謝辞

本研究は第五世代計算機プロジェクトの一環として行なわれたものである。本研究の機会を与えてくれたICOTの内田部長ならびに日本電気株式会社C&Cシステム研究所小池部長、横田課長、C&C情報研究所中村部長、金子課長に深謝致します。また、本研究をサポートして頂いた遺伝子情報処理プロジェクト関係者に感謝致します。

参考文献

[1] Barron, A.R., "Logically smooth density estimation", *PhD dissertion, Dept. of Electrical Eng., of Stanford Univ.,* Aug. (1985).

[2] Bishop,A.M. and Rawlings,C.J.(ed.), "Nucleic Acid and Protein Sequence Analysis", Oxford Univ. Press Limited, (1987).

[3] Congress of the United States (Office of Technology Assessment), "Mapping Our Genes - Genome Projects: How Big, How Fast?" The Johns Hopkins Univ. Press, (1998).

[4] Hamilton,O.S., Thomas,M.A. and Srinivasan, C., "Finding sequence motifs in groups of functionally related proteins", in *Proc. Natl. Acad. Sci. USA*, vol.87, (1990),pp.826-830.

[5] von Heijne,G., "Sequence Analysis in Molecular Biology - Treasure Trove or Trivial Pursuit", Academic Press, Inc., (1987).

[6] Kanehisa,M. and Seto, Y., "Unique Amino Acid Patterns Identified as Possible Functional Elements of Protein Molecules",(1989) private communication.

[7] 小長谷, 横田, "DNA配列知識ベースシステム(KNOA)-論理型言語アプローチ-", 信学会技報 AI89-37,(1989).

[8] 小長谷, 山西, "記述長最小基準の遺伝子情報処理への適用について", 日本ソフトウエア科学会第7回大会論文集,B3-3,(1990).

[9] 稲垣, 神田, "NMR解析(I)", 細胞工学 vol.7, no.11, (1988),pp.91-96.

[10] 木村, 大沢(編), "生物の歴史", 岩波講座 - 分子生物学3,(1989).

[] Lehman, E.L., "Testing statistical hypotheses". *Wiley,* (1959).

[11] 勝部, 京極, 崎山, 高木, 中川(編), "タンパク質II構造と機能編", 東京化学同人,(1988).

[12] Lesk,M.A.(ed), "Computational Molecular Biology", Oxford Univ. Press, (1988).

[13] 岡田, 大澤(編), "分子シミュレーション入門", 海文堂,(1989).

[14] Quinlan, J.R. and Rivest,R.L. "Inferring decision trees using the minimum description length criterion", in *Inform. and Comput.* vol.80, no.3,(1989), pp.227-248.

[15] Rissanen,J., "Stochastic complexity in statistical inquiry", World Scientific Series in Computer Science, vol.15, (1989).

[16] 田中, "代謝反応データベース", 情処研究報告 90-DBS-78,78-14,(1990).

[17] Watson,J.D.,Hopkins,N.H.,Roberts,J.W., Steitz,J.A. and Weiner,A.M., "Molecular Biology of the Gene", Fourth Edition, The Benjamin/Cummings Publishing Company, Inc.,(1987).

[18] Yamanishi, K., "A learning criterion for stochastic rules", in *Proc. of the 3rd Annual Workshop on Computational Learning Theory*,(1990),pp.67-81.

[19] 山西, "階層的パラメータ構造を持つ確率的分類規則の帰納的推論と学習基準", 第12回情報理論とその応用シンポジウム,(1989),pp.707-712.

付録 MDL基準に基づく確率的決定述語の学習の理論的根拠

今、確率的決定述語を M, 確率パラメータを要素とするベクトルを $\theta = (p_1, \cdots, p_m)$ とすると、配列モチーフの定める確率モデルは M と θ により規定される。M を固定したときの θ の事前分布を $v(\theta \mid M)$ とかき、θ, M 及び配列 X を固定したときの カテゴリ Y の発生確率を $P(Y \mid X : \theta \prec M)$、配列 X の発生確率を $Q(X)$、M の事前分布を $P(M)$ とかく。$P(Y \mid X : \theta \prec M)$ の θ に関する mixture を $P(Y \mid X : M) \stackrel{\text{def}}{=} \int P(Y \mid X : \theta \prec M)v(\theta \mid M)d\theta$ (積分区間は $[0,1]^m$ にわたるとする) と定めると、データが N 個の独立な配列とカテゴリの対 $D^N = (X_1, Y_1), \cdots, (X_N, Y_N)$ として与えられたときの尤度は

$$\prod_{i=1}^{N} P(Y_i \mid X_i : M)Q(X_i)$$

と計算できるから、これを $P(D^N \mid M)$ で表すと、M の事後確率 $P(M \mid D^N)$ は Bayes の定理を用いて次式のように計算できる。

$$P(M \mid D^N) = \frac{P(D^N \mid M)P(M)}{\sum_M P(D^N \mid M)P(M)} \qquad (1)$$

そこで、Bayes 推定の考え方を用いて、$P(M \mid D^N)$ を最大化するような M を求めることを考えると、(1) の右辺で分母は M に依らないから、結局、

$$P(D^N \mid M)P(M)$$

を最大化させることに帰着でき、これはさらに、

$$-\log P(D^N \mid M) - \log P(M)$$

を最小化させる M を求めることに等価である。しかも漸近的に

$$-\log P(D^N \mid M) \sim -\log P(D^N \mid \hat{\theta} \prec M) + \frac{m \log N}{2}$$

と近似でき [15]、右辺の第1項、第2項はそれぞれ $O(N)$, $O(\log N)$ のオーダである。但し、$\hat{\theta}$ は D^N より求められる θ の最尤推定値、m は確率パラメータの個数である。従って、以上の Bayes 推定の問題は結局、

$$-\log P(D^N \mid \hat{\theta} \prec M) + \frac{m \log N}{2} + \{-\log P(M)\} \quad (2)$$

を最小化する M を求めることに等価である。ここで、一般に確率分布 $\{P(x)\}$ に従う x に対しては、$-\log P(x)$ の符号長で一意復号可能な符号化ができて、しかもそのときの平均符号長は下限値 (エントロピー) を達成することに注意する。すると、(2) の第1項は θ の代わりに最尤推定値 $\hat{\theta}$ を用いて計算されるデータ D^N の記述長に等しく、これは、確率的決定述語の分類に伴う D^N の不確実性の記述長とみなすことが出来る。また、この項は、定義から

$$-\sum_{i=1}^{N} \log P(Y_i \mid X_i : \hat{\theta} \prec M) - \sum_{i=1}^{N} \log Q(X_i)$$

とかけて、2番目の項は M に無関係であるから無視することにより、(2) の第1項としては

$$-\sum_{i=1}^{N} \log P(Y_i \mid X_i : \hat{\theta} \prec M)$$

だけを評価すれば良い。一方、(2) の第2項は $O(1/\sqrt{N})$ の精度をもつ $\hat{\theta}$ の記述長に等しく、第3項は M 自体の記述長に等しい。すなわち、第2項、第3項は合わせて配列モチーフの確率モデルを記述するのに必要な記述長を表している。よって、(2) を最小化させる試みは、確率モデル (M と $\hat{\theta}$ の記述長とこれを用いて記述される配列データの記述長 (不確実性の記述長) を合わせて最小化する営み、すなわち MDL 基準そのものに他ならない。以上より、MDL 基準の適用は確率パラメタに関する mixture を用いた Bayes 推定に根拠をもつものであることが分かる。

また、Bayes 解は統計的決定理論の意味で誤り率最小解であることが知られており []、さらに、真の確率モデル (データの発生分布) の存在を仮定すれば、MDL モデルは真のモデルに漸近的に収束すること等が明らかにされている [1, 18]。また、その収束速度についても詳細に評価されており [18]、最尤推定モデルの収束速度よりも速いことが理論的に確かめられる。以上のように、MDL 基準は Bayes 推定の考え方に基礎をもちながら、統計学的にもその良い性質が保証された確率モデルを選択する規範にもなっている。