TM-0965

記述長最小基準の遺伝子情報処理への適用について

小長谷　明彦、山西　健司（日本電気）

November, 1990

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

日本ソフトウエア科学会第7回大会 (1990)

記述長最小基準の遺伝子情報処理への適用について
The Application of Minimum Description Length Criterion to Genetic Information Processing

小長谷 明彦, 山西 健司
Akihiko Konagaya, Kenji Yamanishi
日本電気株式会社 C&C システム研究所, C&C 情報研究所

概要

不確実性を伴う大量の遺伝子情報の特徴抽出問題において記述長最小 (MDL) 基準を適用する方法を示し、その有効性を実データを用いて示す。具体的には、アミノ酸配列データバンクを分類する配列モチーフの抽出問題を考え、配列モチーフの抽出を確率的な分類規則の学習問題とみなし、配列モチーフの記述に適した分類規則の記述法として新たに確率的決定述語を提案する。また、確率的決定述語の評価基準として、MDL 基準を適用することにより、生物学的に意味があり、安定で汎用性の高い配列モチーフを抽出できることを示す。

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、蛋白質などの遺伝子関連情報に関するデータバンクが急増し、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、遺伝子情報の特徴部分を抽出した知識ベースの構築を進めている [2]。配列データの場合、このような特徴部分は「配列モチーフ」と呼ばれ、高速かつ信頼性の高い相同検索 (ホモロジーサーチ)[1] 手法として注目を集めつつある [1]。

しかしながら、配列モチーフの抽出においては、不確実性の取り扱いと、データに対する過剰適合の回避という２つの大きな問題がある。配列データの不確実性は生物の多様性 (個体差)、実験エラー、入力エラーなどに起因する。このことは完全な配列モチーフを求めることは困難であり、何らかしらの曖昧性を表現する確率モデルとして扱うことが必要なことを意味する。また、配列モチーフの過剰適合は、与えられた配列データの偏りから、本来抽出してはならない規則性を抽出した際に生じる。このことは、配列モチーフの良否の判定に最尤推定法をあてはめるだけでは不十分であり、配列モチーフの複雑さとデータへの適合性のトレードオフを考慮した、新たな確率モデルの推定方式が必要なことを示す。

これらの問題を解決するために、本稿では、配列モチーフの抽出を確率的な分類規則の学習問題 [6] とみなし、配列モチーフの記述に適した確率的な分類規則の記述法として新たに確率的決定述語を提案する。そして、確率的決定述語の最適化の基準に記述長最小 (Minimum Description Length (MDL)) 基準 [5] を適用した配列モチーフ評価法を提案する。MDL 基準の分類学習への適用は文献 [4, 6] にも見られるが、ここでは、配列モチーフ固有の符号化法に関する記述長の計算法を与える。そして、アミノ酸配列データバンクにおいて共通の先祖を持つタンパク質のグループを分類する配列モチーフの抽出を例にして、本手法により生物学的に意味があり、安定で汎用的な配列モチーフを抽出できることを示す。

[1] 未知の配列が与えられたとき配列の類似性より共通の先祖を持つ配列を求める検索。

本稿の構成は以下のとおりである。はじめに、２節において、配列モチーフ抽出の具体例を示し、３節で配列モチーフを記述するための確率的決定述語を提案する。次に、４節において、確率的決定述語の記述長の計算法ならびに MDL 基準に基づく確率的決定述語の評価法を提案し、５節において、アミノ酸配列データバンクを例にして、本稿で提案する配列モチーフ抽出法の効果を示す。

2 配列モチーフの抽出

配列モチーフは、生物の活動において極めて重要なため共通の先祖をもつグループ内で進化的に保存された部位を表す。逆に、このような共通の配列モチーフを持つかどうかを調べることにより、タンパク質の相同検索を行うことが可能となる。

本稿で述べる配列モチーフ抽出の手順は以下の通りである。(1) 配列のマッチングにより共通部分を求める。(2) 機能部位に関連する共通部分を中心に配列パターンを生成する。(3) 各配列パターンの組合せについて妥当性を検査し、最適な配列モチーフを決定する。

配列の共通部分は DP マッチングなどの良く知られたアルゴリズムを用いて求めることができる。図 1 に、酵素の一種であるミトコンドリアシトクロム C のマッチング結果の一部を示す。ただし、分子生物学の慣例によりアミノ酸をアルファベット１文字で表す。また、共通部分を最下段に示す。

配列パターンの生成においては、いかにして、生物学的に意味のある部位を含めるかが重要となる。ミトコンドリアシトクロム C では、配列前部にある２つのシステイン (C) が電荷を運ぶヘム補分子との結合部位に、その直後のヒスチジン (H) ならびに配列後部にあるメチオニン (M) がヘム補分子の鉄原子との接合部となっている [3]。これらの部位を含む共通パターンとして *CXXCH* および *PGTKM* を得る[2]。さらに、配列の中央部に *GPXLXG* という共通パターンがあり、これらの３つの配列パターンが配列モチーフの候補となる。ここで、最初のシステインは全てのミトコンドリアシトクロム C に共通ではないが、機能部位として配列パターンに含める点に注意されたい。

次に、配列パターンの組合せを考える。生物学的な知見を優先するとすれば、下記の組合せのみを考えればよい。

1. *CXXCH*
2. *CXXCH* and *PGTKM*
3. *CXXCH* and *GPXLXG* and *PGTKM*

ここで重要なことは、配列モチーフとミトコンドリアシトクロム C のような分類対象となるカテゴリとの関係は決定的ではなく、確率的な対応関係になっていることである。例えば、アミノ酸配列データバンク (PIR21.0 版) において、データバンク全体 (6158 例) 中、配列パターン (*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

[2] X は任意のアミノ酸を表す。

```
CCFS    --ASFAEAPAGDPTTGAKIFKTKCAQCHTVEKGAGHKQGPNLNGLFGRQSGTTAGYSYSAANKNMAVIWEENTLYDYLLNPKKYIPGTKMVFPGLKKPQERADL
CCLK    --ATFSZAPPGBZKAGQKIFKLKCAQCHTVEKGAGHKQGPNLNGLFGRQSGTAAGYSYSAANKNMAVVWZZBTLYDYLLNPKKYIPGTKMVFPGLKKPQDRADL
CCNG    --ASFAEAPAGDAKAGEKIFKTKCAZCHTVZKGAGHKQGPNLNGLFGRQSGTTAGYSYSAANKNKAVAWZZBSLYDYLLNPKKYIPGTKMVFPGLKKPZZRADL
CCSP    --ATFSEAPPGNKDVGAKIFKTKCAQCHTVDLGAGHKQGPNLNGLFGRQSGTAASYSYSAANKNKAVIWSEDTLYEYLLNPKKYIPGTKMVFPGLKKPQDRADL
CCND    --ASFBZAPAGBSASGEKIFKTKCAZCHTVBZGAGHKZGPNLNGLFGRQSGTVAGYSYSAANKNKAVNWEEKTLYDYLLNPKKYIPGTKMVFPGLKKPZZRABL
CCGK    --ATFSEAPPGDPKAGEKIFKTKCAZCHTVZKGAGHKQGPNLEGLFGRQSGTTAGYSYSTGNKNKAVNWGZZTLYEYLLNPKKYIPGTKMVFPGLKKPZZRADL
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 の配列マッチング結果の一部

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となる。配列データは不確実性を伴うため、この確率的な対応関係は本質的であり、配列モチーフとカテゴリとの関係を示す分類規則はこのような確率的対応関係を十分に表現できるものでなければならない。

また、与えられた配列データに対して最適な確率的対応関係を求める手法としては最尤推定法が知られているが、配列データに偏りがある場合には、抽出した配列モチーフが過剰適合を起こす可能性があり、必ずしも未知データに対して最良の分類を与えるとは限らないという問題がある。

これらの問題を解決するために配列モチーフとカテゴリとの対応関係の表現形式として確率的決定述語を提案し、MDL 基準に基づいて配列モチーフの評価を行なう方式を提案する。以下、確率的決定述語の表現法と MDL 基準の適用に必要な具体的な記述長の計算法について述べる。

3 確率的決定述語

前節でみた配列モチーフとカテゴリとの確率的対応関係に適した確率的分類規則の表現法として「確率的決定述語」を新たに提案する。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。確率的決定述語の構造を以下に示す。

```
motif(S,C1) with p1 :- Q11,...,Q1n.
motif(S,C2) with p2 :- Q21,...,Q2n.
     ...
motif(S,others) 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 $Q_{i1},...,Q_{in}$ 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。ここで、Q_{ij} としてとり得るのは $R_1;...;R_l$ の形式のパターン述語の OR-結合である。なお、本稿では、パターン述語として *contain(P, S)*(アミノ酸配列 S が配列パターン P を含むか否かを判定する述語) のみを用いる。パターン述語に関しては、配列パターン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 *others* 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

```
motif(S,mitocondoria_cytochrome_c) with p1
   :- contain("CXXCH",S).
motif(S,others) with p2.
```

S が *"CXXCH"* と一致する部位を含めば確率 p_1 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 p_2 で *others* である。

- モチーフ例 2

```
motif(S,mitocondoria_cytochrome_c) with p1
   :- contain("CXXCH",S),
      contain("PGXKM",S).
motif(S,others) with p2.
```

S が *"CXXCH"*、*"PGXKM"* の両方と一致する部位を含めば *S* は確率 p_1 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 p_2 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 p1
   :- (contain("CXXCH",S);contain("AAQCH",S)),
      contain("PGXKM",S).
motif(S,others) with p2.
```

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 p_1 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 p_2 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 基準に基づく配列モチーフの評価

4.1 MDL 基準

配列モチーフの抽出は事例データからの確率的分類規則 (この場合は確率的決定述語) の学習問題とみなすことができる [6]。記述長最小 (MDL) 基準は、このような学習問題において与えられた事例データに対する過剰適合を避け、より安定で信頼性の高い確率的決定述語を選択するための基準となっている [5, 4, 6]。MDL 基準を確率的決定述語の学習に適用すると、

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を選択せよという、規則選択基準が得られる。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バランスがとれ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を付録において簡潔に示す。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

$$-\log(\text{確率的決定述語より定まるカテゴリの尤度})$$

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、求めるカテゴリに属していた配列の個数をN_i^+ ならびに求めるカテゴリに属していない配列の個数を$N_i^-(N_i = N_i^+ + N_i^-)$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$\hat{p}_i$, $\hat{p}_i = \frac{N_i^+}{N_i}$ 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。

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

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。

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

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

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パターンの記述長からなり、パターンの記述長は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$O(1/\sqrt{N_i})$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、カテゴリの記述長は、対象としている配列の集合をM個に分割し、その中の一つのカテゴリを選ぶとすると各クローズ毎に$\log M$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。配列パターンに変数(すなわちX)が含まれていない場合には、配列パターンの長さをN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$N * \log(20)$ビット必要になる。配列パターン中に変数が含まれる場合にはパターン変数の出現位置に関する記述長と配列パターン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述語の配列パターンの長さをN_i^j、この中に含まれるパターン変数の個数をX_i^jとする。配列パターン中のパターン変数の出現位置はN_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$(N_i^j - X_i^j)$乗個となる。したがって、全体では

$$ML = \sum_{i=1}^{m} \{\sum_{j=1}^{B_i} \{\log(\begin{pmatrix} N_i^j \\ X_i^j \end{pmatrix}) + (N_i^j - X_i^j) * \log 20\} + \log M\}$$

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。

以上により、与えられたデータに対して

$$- \qquad UL + PL + ML$$

の値を比較することにより(小さいほど良い)、配列モチーフを評価することができる。

5 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)、確率パラメタの記述長(PL)および不確実性の記述長(UL)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。

配列パターン *"CXXCH"* 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、配列パターン *"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、配列パターン *"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。以上より、ミトコンドリアシトクロムCにおいては、生物学的に裏付けのある機能部位を含むパターンの組合せがMDL基準の観点から配列モチーフとしてより適していると判断することができる。

また、表2に示すように、ミトコンドリアシトクロムCで配列パターン *"CXXCH"* では分類できなかった配列に共通な配列パターン *"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パターン変数の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。OR-結合の配列パターン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配列パターンのOR-結合を多用する

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

モチーフ	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
CXXCH	36.2	10.1	225.5	271.7
CXXCH and GPXLXG and PGTKM	80.0	9.4	73.6	163.0
CXXCH and PGTKM	55.8	9.4	80.7	145.9
(CXXCH or AAQCH) and PGTKM	77.4	9.5	47.1	134.0

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。

6 まとめ

配列モチーフの抽出を確率的な分類規則の学習とみなすことにより、生物学的に意味があり、安定かつ信頼性の高い配列モチーフを記述長最小(MDL)基準を用いて求めることができることを示した。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、良否の判定においてMDL基準は極めて有効な手法といえよう。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、新田室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、岡本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] 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.

[2] 新田, "並列推論マシンを用いた遺伝子情報処理", 第五世代コンピュータに関するシンポジウム予稿集,(1990),pp.5-6.

[3] 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).

[4] 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.

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

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

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、確率的決定述語の配列モチーフを 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。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 $P(Y \mid X : M) \stackrel{def}{=} \int P(Y \mid X : \theta \prec M)v(\theta \mid M)d\theta$ (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 [5]、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、

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

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項は合わせて確率的モチーフ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、規則と規則に伴う不確実性の両者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速い収束速度をもって漸近的に収束すること [6] 等が明らかにされている。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。