二次構造予測法における不確実性推論

2L-9

山本 秀樹　古田 香代里　椎野 努
沖電気工業(株)

1．はじめに

蛋白質の二次構造予測支援エキスパート・システム[古田86]では、システムに対して入力された任意のアミノ酸配列から二次構造を予測する。この予測には不確実性推論が必要になる。現在、合理的な不確実性推論としてDempster & Shafer理論が有力とされているが、この問題は適当とはいえない。本稿では、二次構造予測法をエキスパート・システムに実現する際に用いた不確実知識による推論について述べる。

2．構造予測における不確実性

エキスパート・システムに入力される任意のアミノ酸配列と、その二次構造とを完全に照合できるような知識ベースは現在のところ作成できない。なぜなら、構造の解析されている蛋白質の数は限られており、入力するアミノ酸配列が、構造解析結果と完全に一致することは期待できないからである。また、入力するアミノ酸配列の起りうる組み合わせは、アミノ酸配列の長さをｎとした場合20^nのオーダになるが、今後解析技術が飛躍的に進歩したとしてもこれだけのアミノ酸配列の二次構造データを集めるのは現実的ではない。従って二次構造予測では、不確実ではあるが利用可能な大きさの知識を扱う必要が生じる。

知識工学では、真偽の２値で表せないような不確実な知識や事実の取り扱いが研究されている。従来、不確実性はBayes理論に従う確率量として表されてきたが、Bayes確率では、無知(ignorance)な部分と不信用(disblief)を区別して表現できないという問題がある。それゆえ、MYCINのCF(Certainity Factor)が開発され、主観的Bayesの方法と呼ばれる方法が提示された。その後、不確実性推論の理論的なアプローチとしてDempster & Shafer理論が注目されている。

Dempster & Shafer理論は、有限な事象の全集合の部分集合に対して、部分集合内を自由に移動する基本確率量を表現することで無知の程度を表す[石塚83]。すなわち、問題領域に置ける命題の集合のすべての部分集合を考え、それに基本確率量を割り当てる必要がある。蛋白質の二次構造予測問題について考えてみると、アミノ酸配列に対してαヘリックスになる基本確率量、βシートになる基本確率量、折れ曲り構造になる基本確率量を割り当てる必要がある。そうすることで入力アミノ酸配列について基本確率量の結合が可能である。しかし、Dempster & Shafer理論の結合法則を使用するには、入力アミノ酸配列のすべての部分配列を数え上げる必要があり配列の長さに応じて指数関数的に計算量が増大してしまう。従って、Dempster & Shafer理論をこの問題領域に適用することは困難だと考えられる。

3．不確実知識の表現形式

二次構造予測問題を以下のような副問題の集りとしてとらえる。副問題への分割例を図１に示す。

（ａ）ある記号列を複数に分割する問題。
図１では、記号列'EEKLKK･･･'を'EEK･･K'と'IIFVV'などに分割している。

（ｂ）分割された記号列に対し、ある仮説をたて不確実性推論を適用した場合の確実度を求める問題。
図１では、'EEKL･･'に対して仮説1～3、'IIFVV'に対して仮説4,5をたて、その確実度(P_1)～(P_5)を求めている。

（ｃ）分割された記号列に対する最適な仮説を決定する問題。
図１では、仮説2と仮説5が採択されている。

これらの問題を解くために、記号列の分割、分割された記号列に対する仮説、およびその仮説の確実度の３つを一組に表す

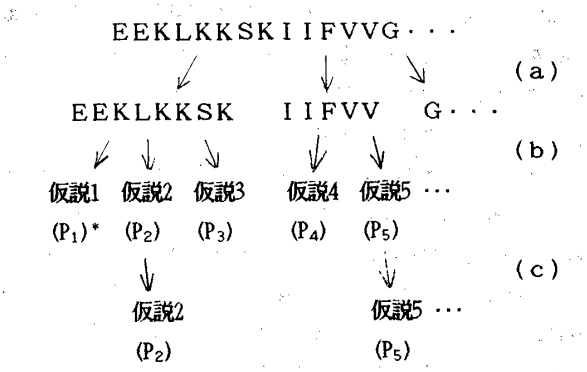

図１　問題の分割例
注*：仮説に対する確実度

Inference with uncertainty on secondary structural predictions

Hideki YAMAMOTO Kayori FURUTA Tutomu SHIINO
OKI Elec. Ind. Co., Ltd.

```
MEEKLKKSKIIF····
```

（ａ）入力アミノ酸配列

```
[ αヘリックス,  1,  9,  1.333 ]
```

（ｂ）二次構造リスト表現例

```
[ αヘリックス,  1,  6,  0.000 ]
[ βシート,      1,  5,  0.000 ]
```

（ｃ）

```
[ αヘリックス,  1,  6,  1.333 ]
[ αヘリックス,  2,  7,  1.285 ]
```

（ｄ）

図２　二次構造リスト例

知識表現が必要になる。そこで、知識表現として[H, I_s, I_e, P]のリストを使用する。ここで、Hは仮説、I_s、I_eは分割された記号列の入力記号列上の位置、そしてPはI_s、I_e間の記号列に仮説Hを仮定したときの確実度を表す。

エキスパート・システム上に二次構造予測法を実現するにあたり、ここで述べた知識表現を使用した。二次構造予測問題を上記の副問題の集りとしてとらえ不確実知識の扱いを説明する。説明は、Chou & Fasman法に基づいて行う。この際、上記の[H, I_s, I_e, P]のリストを二次構造リストとよぶ。入力アミノ酸配列の例を図２（ａ）に、その第１アミノ酸から第９アミノ酸までがαヘリックスである場合の表現を図２（ｂ）に示す。図２（ａ）でアルファベット一文字は一つのアミノ酸に対応している。

３．１　入力アミノ酸配列の分割

入力されたアミノ酸配列は、Chou & Fasman法によって得られる二次構造の最小長さ（αヘリックスは６、βシートは５）を切出し二次構造リストに入れる。例えば図２（ａ）の入力配列を左から図２（ｃ）の様に二次構造リストに入れる。二次構造リスト中の確実度は初期値０がセットされる。

３．２　不確実性推論の適用

二次構造リストの確実度Ｐを求める。確実度は、I_s、I_e間のアミノ酸一つ一つがもつ不確実データを結合して求める。不確実データは、構造の解析されている蛋白質について、各々のアミノ酸が二次構造領域に入る頻度から決められる。これらの頻度は、各アミノ酸が二次構造に表われる傾向を示す。このデータは図３のようにPROLOGのファクト形式で各アミノ酸の確実性を表現する。（ａ）はαヘリックス、（ｂ）はβシート、そして（ｃ）は折れ曲り構造に関する知識を表している。（ａ）～（ｃ）の第１引数はアミノ酸名をアルファベット一文字で表記している。（ａ）と（ｂ）の第２引数はアミノ酸の確実度、第３引数は確実度の記号表示である。（ｃ）の引数は、左からアミノ酸名、アミノ酸の確実度、第３引数から第６引数は折れ曲り構造の二次構造の最小長さ４で分割されたアミノ酸配列中

```
alpha('A',1.42,'H').
alpha('C',0.70,'i').
alpha('D',1.01,'I').
        :
```

（ａ）αヘリックスに関する知識例

```
beta('A',0.83,'i').
beta('C',1.19,'h').
beta('D',0.54,'B').
        :
```

（ｂ）βシートに関する知識例

```
rt('A',0.66,0.060,0.076,0.035,0.058).
rt('C',1.19,0.149,0.053,0.117,0.128).
rt('D',1.46,0.147,0.110,0.179,0.081).
        :
```

（ｃ）折れ曲り構造に関する知識例

図３　不確実な知識の表現例

の位置に関係した確実度である。

アミノ酸配列の確実度は、この予測法では、一つ一つのアミノ酸の該当する確実度の平均で結合している。αヘリックス、βシートについては、相加平均を使用している。折れ曲り構造については、相加平均と相乗平均を使用している。このような平均による結合は、構造の解明された蛋白質からの統計的、経験的知識に基づいている。

３．３　最適仮説の決定

アミノ酸配列中のある部分に対する異なる仮説は、確実度Ｐの大きい仮説を最適とする。例えば、図２（ｄ）では、確実度Ｐの大きいαヘリックスの仮説[αヘリックス, 1, 6, 1.333]を採択する。

４．まとめ

蛋白質の二次構造予測問題を分析し、この種の問題に有効な知識表現を与えた。蛋白質の二次構造予測問題は不確実性を伴うものでありその取り扱いについて述べた。

今後は、より幅広い不確実な知識の表現と利用の方法について検討していく必要がある。

[参考文献]

[CHOU74] CHOU P.Y.& FASMAN G.D.,"Prediction of Protein Conformation",Biochemistry VOL.13,NO.2,1974.

[石塚83] 石塚 「Dempster & Shaferの確率理論」信学会誌 Vol.66.No.9,pp900-903.1983

[古田86] 古田 他「蛋白質の二次構造予測支援エキスパートシステム」第33回情処全国大会論文集