情 報 学 基 礎 23－1
（1991. 9. 10）

EFSの学習可能性と膜蛋白領域予測への応用

有川 節夫† 久原 哲‡ 宮野 悟† 篠原 歩† 篠原 武††

† 九州大学理学部附属基礎情報学研究施設
‡ 九州大学大学院農学研究科遺伝子資源工学専攻
†† 九州工業大学情報工学部知能情報工学教室

EFS(elementary formal system)を用いた膜蛋白領域の学習について論じる．EFSはif-then規則からなる論理プログラムの一種である．この枠組を用いてアミノ酸配列の中から膜蛋白領域を予測するアルゴリズムを実働化した．計算資源の制約のため，我々は仮説の候補として用いるEFSを正則パターンに制限した．しかし，我々のアルゴリズムは70個の膜蛋白データから膜蛋白領域を同定するいくつかの妥当な仮説を生成した．各仮説は高々10個の正則パターンで表されている．PIRデータベースを利用した検証から，これらの仮説は膜蛋白領域のうちの90%をカバーし，またそれ以外の部分の80%を排除することが確かめられた．

Learnability of EFS and its Application for Identifying Transmembrane Domains

Setsuo Arikawa† Satoru Kuhara‡ Satoru Miyano† Ayumi Shinohara† Takeshi Shinohara††

† Research Institute of Fundamental Information Science, Kyushu University 33, Fukuoka 812
‡ Graduate School of Genetic Resources Technology, Kyushu University, Fukuoka 812
†† Department of Artificial Intelligence, Kyushu Institute of Technology, Iizuka 820

We propose a method for algorithmic learning of transmembrane domains based on EFS(elementary formal systems). An EFS is a kind of a logic program consisting of if-then rules. With this framework, we have implemented the algorithm for identifying transmembrane domains in amino acid sequences. Because of the limitations on computational resources, we restrict candidate hypotheses to EFSs defined by collections of regular patterns. However, from 70 sequences of transmembrane domain data, our algorithm has produced several reasonable hypotheses that can identify transmembrane domains. Each of the hypotheses consists of at most ten regular patterns. Experiments with the database PIR show that most of these hypotheses can cover 90% transmembrane domain sequences and exclude 80% negative data.

1 はじめに

筆者らは [6] において EFS(elementary formal system) のある部分クラスが多項式時間 PAC 学習可能 [2, 7, 13] であることを証明している．EFS は Smullyan[10] によって導入されたもので，文字列の集合を定義する，if-then 規則からなる論理プログラムの一種である．その記述力はある制限のもとでも極めて大きく [1]，またその意味論も深く研究されている [14]．さらに EFS は帰納推論のための統一的な枠組としても注目されている [1, 12]．

本稿では EFS の学習アルゴリズムをアミノ酸配列中の膜蛋白領域を予測する問題に適用した結果を報告する．このアルゴリズムはある制約された形の EFS を仮説の表現として用い，与えられた正の例と負の例に対し，矛盾しない仮説を見つけ出す．EFS は論理プログラムの構造を持っているため，出力された仮説はどの事実と規則が正の例を導き出し，また負の例を除外しているのかを明解に表している．

学習アルゴリズムとして [6] で与えた基本的なアイディアを実働化し計算機実験を行なった．ただし，[6] のアルゴリズムは EFS のある部分クラスに対して多項式時間で走るとはいえ，入力長が長くなると現実的ではなくなる．そこでここでは仮説の表現する EFS を正則パターン [9, 11] と呼ばれる特殊な形に制限することで現実的な時間と領域で仮説を見つけ出すことに成功した．その結果，70 個の膜蛋白領域のデータ（正の例）と，100 個の負の例から，いくつかの仮説が得られた．この仮説を PIR データベース [8] から得られる約 600 個の正のデータと約 16,000 個の負のデータを用いて検証したところ，ほとんどの仮説が正のデータ（膜蛋白領域）の 90% を説明し，また負のデータの 80% を排除することが確かめられた．このことから，EFS の学習アルゴリズムがアミノ酸配列中の膜蛋白領域の同定に対して十分有用であると結論できる．

2 Elementary Formal System と PAC 学習可能性

Σ を有限アルファベット，$X = \{x, y, z, x_1, x_2, \cdots\}$ を変数の集合，$\Pi = \{p, q, r, s, p_1, p_2, \cdots\}$ を述語記号の集合とし，Σ, X, Π は互いに交わりをもたないとする．Σ^* を Σ 上のすべての文字列の集合，Σ^+ を空語 ε 以外の文字列の集合，さらに $n \geq 0$ に対し $\Sigma^{\leq n}$ を長さ n 以下の文字列の集合とする．$(\Sigma \cup X)^+$ の要素をパターンという．パターン π の中に各変数が高々 1 回ずつしか現れないとき，π は**正則**という．**アトム**とは $p(\tau_1, \ldots, \tau_n)$ の形の式をいう．ここに p は n 引数の述語記号，$\tau_1, \ldots, \tau_n$ はパターンである．**確定節**とは $A \leftarrow B_1, \ldots, B_m$ の形の式をいう．ここに $m \geq 0$ であり，A および $B_1, \ldots, B_m$ はアトムである．A を確定節の**頭部**，$B_1, \ldots B_m$ を**本体**という．**Elementary formal system(EFS)** とは確定節の有限集合である．EFS Γ の確定節を Γ の**公理**という．

例 1. $\Sigma = \{a, b\}$ とし，次のような EFS を考える．

$$\Gamma = \left\{ \begin{array}{l} p(x) \leftarrow q(x), r(x) \\ q(ax) \leftarrow q(bx) \\ q(bbxby) \leftarrow \\ r(xaxb) \leftarrow \end{array} \right\}.$$

パターン $bbxby$ は正則であるが，$xaxb$ は正則ではない．確定節 $p(x) \leftarrow q(x), r(x)$ の頭部はアトム $p(x)$ であり，本体は $q(x), r(x)$ である．

代入 θ はパターンからパターンへの準同型写像で，各 $a \in \Sigma$ に対し $\theta(a) = a$ となるものをいう．特に変数を空語 ε に対応させる代入を ε 代入とよび，以下では特に断らない限り使用しない．代入 θ によるパターン π の像を $\pi\theta$ で表す．アトム $A = p(\pi_1,\ldots,\pi_n)$ および確定節 $C = A \leftarrow B_1,\ldots,B_m$ に対しては，それぞれ $A\theta = p(\pi_1\theta,\ldots,\pi_n\theta)$，$C\theta = A\theta \leftarrow B_1\theta,\ldots,B_m\theta$ と定義する．

確定節 C が EFS Γ から**証明可能**であるとは，Γ の公理から代入と modus ponens を有限回適用して C が得られることをいう．1引数の述語記号 $p \in \Pi$ に対し，$L(\Gamma,p) = \{w \in \Sigma^+ \mid p(w) \leftarrow$ はΓから証明可能$\}$ と定義する．このような Γ と p が存在するとき，言語 $L \subseteq \Sigma^+$ は **EFS で定義可能**あるいは **EFS 言語**という．パターン π に対し，パターン言語 $L(\pi)$ とは集合 $\{w \in \Sigma^+ \mid$ ある代入θに対して$w = \pi\theta\}$ である．ここでパターン言語 $L(\pi)$ は $\Gamma = \{p(\pi) \leftarrow\}$ なる EFS 言語 $L(\Gamma,p)$ であることに注意しよう．

パターン π の文字列としての長さを $|\pi|$ で表す．アトム $p(\pi_1,\ldots,\pi_n)$ に対し，$||p(\pi_1,\ldots,\pi_n)|| = |\pi_1| + \cdots + |\pi_n|$ と定義する．確定節 $A \leftarrow B_1,\ldots B_m$ が**長さ限定**であるとは，任意の代入 θ に対して $||A\theta|| \geq ||B_1\theta|| + \cdots + ||B_m\theta||$ であることをいう．例えば，確定節 $q(bx) \leftarrow q(ax)$ は長さ限定であるが，$p(axbc) \leftarrow q(ax), r(xb)$ は長さ限定ではない．EFS Γ が長さ限定であるとは，Γ のすべての公理が長さ限定であることをいう．

確定節 $q(\pi_1,\ldots,\pi_n) \leftarrow q_1(\tau_1,\ldots,\tau_{t_1}), q_2(\tau_{t_1+1},\ldots,\tau_{t_2}),\ldots,q_l(\tau_{t_{l-1}+1},\ldots,\tau_{t_l})$ が**継承的** (*hereditary*) であるとは，本体に現れるパターン τ_j が，それぞれ1つ以上の変数を含み，頭部のいずれかのパターン π_i の部分文字列になっているときをいう．例えば，確定節 $p(axbc) \leftarrow q(ax), r(xb)$ は継承的であるが，$q(bx) \leftarrow q(ax)$ は継承的ではない．EFS Γ のすべて公理が継承的であるとき，Γ は**継承的**であるという．

自然数 $m,k \geq 1$ に対し，頭部の変数の出現が高々 k である，高々 m 個の公理からなる長さ限定継承的 EFS によって定義される言語の族を LB-H-EFS(m,k) と表す．LB-H-EFS(m,k) は任意の m および k に対して無限に多くの言語を含む．特に，任意の文脈自由言語は，ある $m \geq 1$ に対し LB-H-EFS(m,2) に含まれ，任意の正規言語は，ある $m \geq 1$ に対し LB-H-EFS(m,1) に含まれる．さらに，LB-H-EFS(m,k) は高々 k 個の変数の出現をもつパターンによって定義されるパターン言語の m 個の和を含む．

以下に長さ限定継承的 EFS とそれによって定義される言語の例を示す．

例 2. $\{a^{3n} \mid n \geq 1\} \in$ LB-H-EFS(2,1).

$$\Gamma = \left\{ \begin{array}{l} p(aaax) \leftarrow p(x) \\ p(aaa) \leftarrow \end{array} \right\}.$$

例 3. $\{a^n b^n \mid n \geq 1\} \in$ LB-H-EFS(2,1).

$$\Gamma = \left\{ \begin{array}{l} p(axb) \leftarrow p(x) \\ p(ab) \leftarrow \end{array} \right\}.$$

例 4. $\{a^{2^n} \mid n \geq 0\} \in$ LB-H-EFS(2,2).

$$\Gamma = \left\{ \begin{array}{l} p(xx) \leftarrow p(x) \\ p(a) \leftarrow \end{array} \right\}.$$

例 5. $\{ww \mid w \in \Sigma^+\} \in$ LB-H-EFS(1,2).

$$\Gamma = \{p(xx) \leftarrow\}.$$

例 6. $\{ww^R \mid w \in \{a,b\}^+\} \in$ LB-H-EFS(4,1).

$$\Gamma = \left\{ \begin{array}{l} p(axa) \leftarrow p(x) \\ p(bxb) \leftarrow p(x) \\ p(aa) \leftarrow \\ p(bb) \leftarrow \end{array} \right\}.$$

例 7. $\{a^n b^n c^n \mid n \geq 1\} \in$ LB-H-EFS(3,3).

$$\Gamma = \left\{ \begin{array}{l} p(xyz) \leftarrow q(x,y,z) \\ q(ax,by,cz) \leftarrow q(x,y,z) \\ q(a,b,c) \leftarrow \end{array} \right\}.$$

概念学習の枠組においては，Σ^* の部分集合 c を概念とよぶ．概念 c を，$x \in c$ のとき $c(x) = 1$，そうでないとき $c(x) = 0$ となる関数 $c : \Sigma^* \to \{0,1\}$ とみなすこともできる．概念の空でない集合 $C \subseteq 2^{\Sigma^*}$ を概念クラスという．概念 c の例とは組 $\langle x, c(x)\rangle$ である．

概念クラス C が多項式時間学習可能 [2, 7] であるとは，以下の条件をみたすアルゴリズム $\mathcal{A}$ が存在するときをいう：

(1) $\mathcal{A}$ は入力長についての多項式時間で走る．

(2) ある多項式 $p(\cdot,\cdot,\cdot)$ が存在して，任意の自然数 $n \geq 0$，任意の概念 $c \in C$，任意の実数 ε, δ $(0 < \varepsilon, \delta < 1)$，および $\Sigma^{\leq n}$ 上の任意の確率分布 P に対し，P にしたがって互いに独立に得られた $p(n, \frac{1}{\varepsilon}, \frac{1}{\delta})$ 個の例を入力として与えると，$\mathcal{A}$ は少なくとも $1-\delta$ の確率で $P(c \oplus h) < \varepsilon$ をみたす仮説 $h \in C$ の表現を出力する．

定理 1. [6] LB-H-EFS(m,k) は任意の $m, k \geq 1$ に対して多項式時間学習可能である．

この定理の証明に用いた学習アルゴリズムの実行時間は入力長に関しては多項式的であるが，定数 m と k に関しては指数的に増加するため，$m = 10$, $k = 10$ 程度であっても莫大な時間と領域を要し，そのままでは我々の問題に用いることはできない．したがって実際に学習を行なうためには，変数の出現数 k と公理の個数 m をどちらも小さく制限しなければならない．また，m の値をどのように設定しておけば学習が可能になるのかを前もって知ることもできない．

そこで我々は EFS の公理を，π を正則パターンとして $p(\pi) \leftarrow$ という形に制限し，以下のような方法をとった．まず，与えられた正の例の集合 Pos のうちの少なくとも一つを含み，負の例の集合 Neg の要素をすべて除外できる公理の集合 S を枚挙する．次に S の部分集合で，Pos をすべて説明できる小さな部分集合 Γ を取り出す．このとき Γ のサイズができる限り小さいことが望まれるが，最小集合被覆問題は NP 完全であることが知られている [3]．したがって，集合の包含関係に関して必ずしも最小ではないが極小な S の部分集合 Γ をみつけることにする．我々の用いたアルゴリズムの概略を示す．

input Pos, Neg;
$S := \emptyset$;
foreach pattern π with $\pi\theta = w$ for some $w \in Pos$ and θ
    if $L(\pi) \cap Neg = \emptyset$
        then $S := S \cup \{p(\pi) \leftarrow\}$;
Find $\Gamma \subseteq S$ such that $L(\Gamma, p) \supseteq Pos$ and Γ is minimal
    with respect to set-inclusion;
output Γ;

最小集合被覆問題に対する近似アルゴリズムとして Johnson[4] を用いた．この近似アルゴリズムは与えられた集合の被覆に必要な最小のサイズを M とすると高々 $M \log M$ の集合被覆を出力することが保証されている．

3　膜蛋白領域予測の実験

この計算機実験の目標は，アミノ酸配列を表す文字列だけを用いて，その蛋白の性質を次の 3 種類に分類できる仮説を見つけることである．

C1. Membrane proteins （膜蛋白），

C2. Secretary proteins,

C3. Cytosolic proteins.

我々は PIR データベース [8] から 37 個の膜蛋白データを取り出した．その一例を図 1 に示す．

```
MDVVNQLVAGGQFRVVKE(PLGFVKVLQWVFAIFAFATCGSY)TGELRLSVECANKTESALNIEVEFEYPFRLHQVVYFDA
PSCVKGGTTKIFLVGDYSSSAE(FFVTVAVFAFLYSMGALATYIFL)QNKYRENNK(GPMMDFLATAVFAFMWLVSSSAW
A)KGLSDVKMATDPENIIKEMPMCRQTGNTCKELRDPVTS(GLNTSVVFGFLNLVLWVGNLWFVF)KETGWAAPFMRAPP
GAPEKQPAPGDAYGDAGYGQGPGGYGPQDSYGPQGGYQPDYGQPASGGGGYGPQGDYGQQGYGQQGAPTSFSNQM
```

図 1: ある膜蛋白のアミノ酸配列．括弧で囲まれた 4 つの部分は膜蛋白領域を表す．

しかしながら，各アミノ酸配列は非常に長い (35 ~ 4544) ため，学習アルゴリズムを直接適用して適当な仮説を表現する EFS を見つけるのは計算量的に困難である [6].

分子生物学において，膜蛋白は α ヘリックス構造をもつ膜蛋白領域 (transmenbrane domain) を含んでいるとされており，もしもアミノ酸配列中に膜蛋白領域に対応する部分が見つかれば，その蛋白が膜蛋白である可能性は大きいといえる．したがって，アミノ酸配列から膜蛋白領域を学習するアルゴリズムを考えることにする．膜蛋白領域の長さは通常 20 ~ 30 程度であるため，現実的な時間で学習を行なえる可能性がある．

アミノ酸を表す 20 文字を直接使うかわりに，各アミノ酸の親水度 (hydropathy index) [5] によって，それを 3 種類の新しい文字に置き換えた．具体的には，次のような変換を行なっている．

アミノ酸		新しい文字	親水度
A, C, I, L, M, F, V	→	*	1.8 ~ 4.5
G, S, T, W, Y, P	→	+	−1.6 ~ −0.4
R, N, D, Q, E, H, K	→	−	−4.5 ~ −3.2

例えばこの変換により，図 1の列は図 2のようになる．

```
*-**--***++-*-**--(+*+**-**-+********+*+++)++-*-*+*-**--+-+**-*-*-*-++*-*--**+*-*
++**-++++-****+-++++*-(***+*******++*+***++***)---+-----(++**-***+******+**+++*+
*)-+*+-*-**+-+--**--*+**--++-+*--*--+*++(+*-++***+**-***+*+-*+***)--+++**+**-*++
+*+---+*++-*++-*++-++++++++--++++-++++-+-++-+*++++++++-+-++--+++--+*+++*+--*
```

図 2: 変換後の列

この変換によって膜蛋白領域の性質がより顕著になる．図 3に示す 70 個のデータは，37 個の膜蛋白から得られた膜蛋白領域に上記の変換を行なったものである．

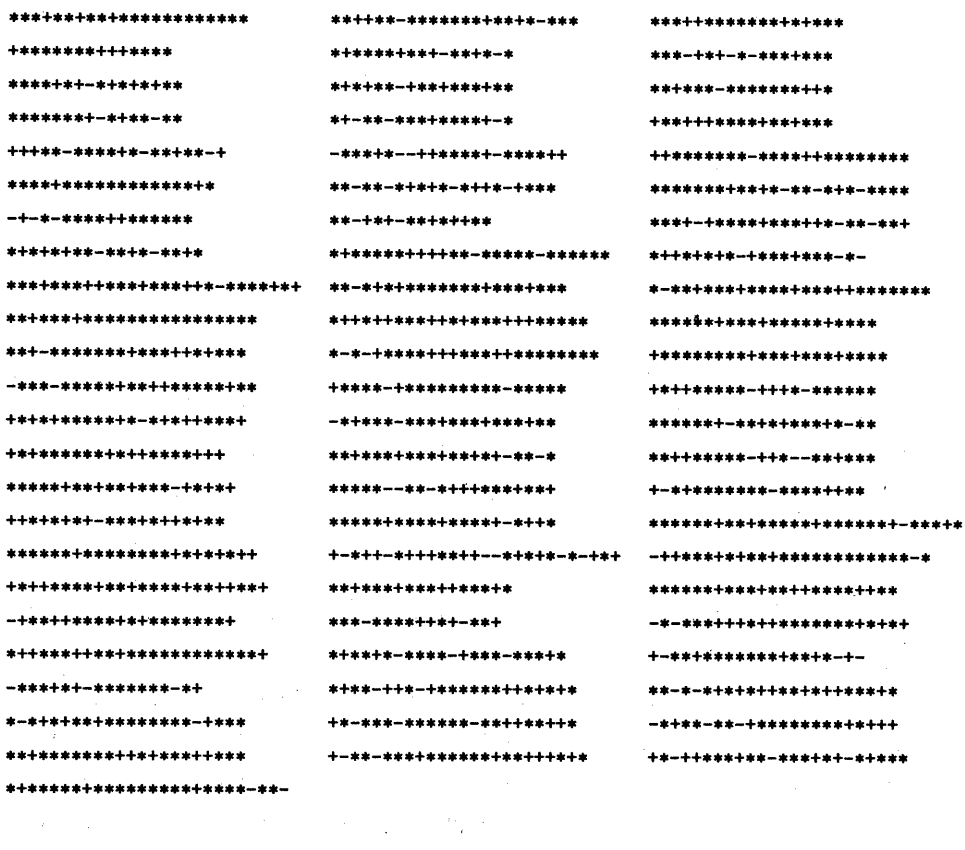

図 3: 変換後の膜蛋白領域

図3に示した70個のデータを正の例とし，膜蛋白領域以外の部分から長さ25前後で切り出した100個の列を負の例として採用した．学習アルゴリズムの目標は与えられた正の例・負の例に矛盾しない仮説の生成である．すなわち，正の例はすべて覆うが，負の例はすべて排除する正則パターンの集まりを見つけ出すことである．

ところが一般には正則パターンのクラスは $\mathbf{RP} \neq \mathbf{NP}$ の仮定のもとで多項式時間学習可能ではない[6]ので，実際に計算を行なうためには変数の個数を制限する必要がある．そのため，使うパターンの形を $x\alpha_1 z, x\alpha_1 y\alpha_2 z, x\alpha_1 y_1\alpha_2 y_2\alpha_3 z$ の3種類に制限した．ここに $\alpha_1, \alpha_2, \alpha_3 \in \{*, +, -\}^+$ であり，x, y_1, y_2, z は相異なる変数である．また，変数 x, z に対しては ε 代入を許している．このような形の有限個の正則パターンからなるEFSを仮説の表現として用いる．

このアルゴリズムをSun SPARC station 2上で実働化した．まず，上記の形のパターンを約244万個生成し，その中から妥当な仮説をいくつか抽出した．このために要した計算時間の総和は約3時間である．これらの仮説を検証するため，PIRデータベースを用い，新たに357個の膜蛋白から599個の膜蛋白領域を，それ以外の部分から長さ25前後の列を16761個取り出した．図4に我々の学習アルゴリズムが出力した仮説とその検証の結果を示す．この結果から，我々の得た仮説は十分に良い仮説であるといえる．

(H1)

******	58.8	4.4
+***X**X***+	34.6	3.9
+****X***+	18.9	2.5
+**X***X+***	15.0	3.0
-X*X***+	12.0	3.6
*-X******	16.2	3.1
X***X*+	13.4	1.8
total	89.1%	16.8%

(H2)

*X******	54.6	3.9
******X*X*	54.3	3.3
+****+	35.7	7.6
X**X***+	50.1	4.3
+XX**	15.2	2.3
X+*X**+	21.9	5.8
*-X******	16.2	3.1
*X***X****+	20.7	4.5
total	94.9%	22.9%

(H3)

******X*	53.1	3.8
*X****X****+	43.9	3.1
X**X***+	50.1	4.3
X*X*+	36.9	3.6
+X******	47.9	3.6
*-X******	16.2	3.1
X+*X**+	21.9	5.8
*X***X****+	20.7	4.5
X*X****+	55.3	4.4
total	93.3%	20.0%

(H4)

X*X****+	55.3	4.4
*X**X*****	56.4	5.8
X*X*+	36.9	3.6
X**X***+	50.1	4.3
*****X*X+**	35.2	2.6
X+*X**+	21.9	5.8
*-X******	16.2	3.1
X*X*	57.1	6.0
*X***X****+	20.7	4.5
total	93.6%	23.1%

(H5)

+***X***X+	54.8	5.9
*****X**X***	41.6	1.8
X*X*+	36.9	3.6
X**X***+	50.1	4.3
X+*X**+	21.9	5.8
+X**	49.4	3.8
*****X*X+**	35.2	2.6
+X******X-	18.7	2.9
*-X******	16.2	3.1
*X***X****+	20.7	4.5
total	93.4%	22.5%

(H6)

X*X****+	55.3	4.4
+****+	35.7	7.6
X***X**	19.4	2.4
-**X****X*	11.4	6.0
-+X*X	5.8	2.2
*+X*****	21.0	4.2
+**X***-*	6.0	2.6
+**+X***X+**	15.0	3.0
X*+X***	31.6	2.1
-X*****X+	11.7	4.7
total	82.6%	26.3%

図 4: 70 個の正の例と 100 個の負の例から作られた仮説とその検証結果．簡単のために両端の変数は省略し，変数の出現位置を X で表している．第 2, 第 3 欄はそれぞれ，第 1 欄のパターンが検証用の新たな 599 個の正のデータ，16761 個の負のデータを含んだ割合を示す．

4 おわりに

本稿では蛋白データから知識の獲得を行なう新しい枠組を与え，計算機実験を行なった．その中でEFSの制限されたクラスが膜蛋白領域の同定に有効であることを示した．より一般的なクラスLB-II-EFS(m,k)に対する学習アルゴリズムも理論的には多項式時間で走ることが証明されている[6]が，そのアルゴリズムは莫大な時間を要するため，このままでは現実的でない．より一般的な設定に対処するためにはアルゴリズムの本質的な改良が必要である．膜蛋白領域予測のための他の方法との比較は今後の課題として残されている．

参考文献

[1] S. Arikawa, T. Shinohara and A. Yamamoto, Elementary formal system as a unifying framework for language learning, In *Proc. 2nd Workshop on Computational Learning Theory*, 312-32, 1989 (to appear in *Theoretical Computer Science*).

[2] A. Blumer, A. Ehrenheucht, D. Haussler and M.K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *JACM* **36**, 929-965, 1989.

[3] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman, 1979.

[4] D.S. Johnson, Approximation algorithms for combinatorial problems, *JCSS* **9**, 256–278, 1974.

[5] J. Kyte and R.F. Doolittle, A simple method for displaying the hydropathic character of protein, *J. Mol. Biol.* **157**, 105-132 1982.

[6] S. Miyano, A. Shinohara and T. Shinohara, Which classes of elementary formal systems are polynomial-time learnable?, Technical Report RIFIS-TR-CS-37, Research Institute of Fundamental Information Science, Kyushu University, 1991 (to appear in Proc. 2nd ALT'91).

[7] B.K. Natarajan, On learning sets and functions, *Machine Learning* **4**, 67-97, 1989.

[8] Protein Identification Resource, National Biomedical Research Foundation.

[9] T. Shinohara, Polynomial time inference of pattern languages and its applications, In *Proc. 7th IBM Symp. on Mathematical Foundations of Computer Science*, 191-209, 1982.

[10] R.M. Smullyan, *Theory of Formal Systems*, Princeton University Press, 1961.

[11] T. Shinohara, Polynomial time inference of extended regular pattern languages, In *Proc. RIMS Symposia on Software Science and Engineering* (Lecture Notes in Computer Science **147**), 115-127, 1983.

[12] T. Shinohara, Inductive inference from positive data is powerful, In *Proc. 3rd Workshop on Computational Learning Theory*, 97-110, 1990.

[13] L. Valiant, A theory of the learnable, *Commun. ACM* **27**, 1134-1142, 1984.

[14] A. Yamamoto, Elementary formal system as a logic programming language, In *Proc. Logic Programming Conference '89*, 123-132, 1989.