情 報 学 基 礎 38－6
（1995．7．28）

アミノ酸配列データからの機械発見システムBONSAI Garden

宮野 悟

〒812 福岡市東区箱崎 6-10-1
九州大学理学部基礎情報学研究施設

概要

正の例と負の例から仮説を発見する機械発見システムBONSAIは，そのデータを説明するアルファベットのインデックス化と決定木を生成するものである．このシステムは，これまでの計算機実験により，膜貫通領域の配列データやシグナルペプチッド配列データから分子生物学的に意味のある知識を発見することに成功している．しかし，データが様々の配列データの混合物であるときは，精度の高い小さな仮説を発見することをBONSAIに期待することには無理がある．このような状況に対応するため開発したものがBONSAI Gardenである．BONSAI Gardenでは，*Gardener*というコーディネータプログラムのもとで，複数個のBONSAIがネットワーク上で並列に走り，データの分類と知識発見が行なわれる．本講演では，この一連の研究を解説する．

BONSAI Garden: Macine Discovery System from Amino Acid Sequences

Satoru Miyano

Research Institute of Fundamental Information Science
Kyushu University 33, Fukuoka 812, Japan.
Email: miyano@rifis.kyushu-u.ac.jp

Abstract

We have developed a machine discovery system BONSAI which receives positive and negative examples as inputs and produces as a hypothesis a pair of a decision tree over regular patterns and an alphabet indexing. This system has succeeded in discovering reasonable knowledge on transmembrane domain sequences and signal peptide sequences by computer experiments. However, when several kinds of sequences are mixed in the data, it does not seem reasonable for a single BONSAI system to find a hypothesis of a reasonably small size with high accuracy. For this purpose, we have designed a system BONSAI Garden, in which several BONSAI's and a program called *Gardener* run over a network in parallel, to partition the data into some number of classes together with hypotheses explaining these classes accurately. This paper surveys a series of our researches on this topic.

1 はじめに

遺伝やタンパク質についての重要な情報は，それぞれ核酸配列やアミノ酸配列にコードされているといわれている [20]. PIR, GenBank, PDB, EMBL, PROSITE などのデータベースはこうした配列データをその機能や構造と共に蓄積し，検索等のサービスを行なっている．タンパク質に関しては，現在こうしたデータベースに約 50,000 本の配列データがあり，数年後には 100,000 本を越える状況にある．

ヒトゲノム計画 (HGM) は，アメリカ合州国において 1990 年ごろ開始され，続いてイギリス，日本，フランス及びその他のヨーロッパ先進国で立ち上がり，各種の分野の研究者の協力のもとで遂行されている．HGM は，2005 年から 2010 年ごろまでにヒトの遺伝子の全配列情報を決定すべく，第二ステージに入ろうとしている．その中で，ゲノム情報 (Genome Informatics) または分子生物情報学 (Molecular Bioinformatics) とよばれる分野は，ゲノム及びそれに関連する情報の収集・体系化及びその情報解析と利用方式を集中的に研究しながら，急速な展開をしている．

ヒトゲノム計画は，分子生物学や情報科学ばかりでなく，本来文系の学問が対象としてきた分野にも関係し，学問のユニバース的な性格をもっている．そしてゲノム情報が人工知能技術に関わっているところでは，この本講演のテーマである「大規模データベースからの知識発見」が重要な課題としてあげられる．そして，タンパク質のアミノ酸配列データや核酸配列データから，有用な知識を抽出し，それによって生物学的実験・検証への方向を提示する方式の構築が急務となっている．

ここでは，ゲノム情報に計算論的学習理論や計算量等の考察に基づいて，このデータベースからの知識発見にどのように取り組んできたかを，我々のこれまでの研究 [2, 9, 10, 11, 14, 15, 17, 18] にしたがって解説する．計算機実験の結果については講演の中で報告する．

2 データベース

アミノ酸配列及び核酸配列データをそれらの性質や機能，文献情報とともに供給してくれるデータベースには次のようなものがある．こうしたデータベースには機能や性質が明らかにされた配列が蓄積されている．配列データとその機能や性質を対応させてとらえて例とみなし，機械学習などの手法を用いて，配列データの中に潜んでいる機能や性質を説明する規則を統一的に見つけだすことがゲノム情報における大きなトピックとなっている．

まず，**GenBank** は，ジャーナル等に掲載されたもの及び著者からの直接の寄稿による核酸配列のデータベースであり，データベースの各レコードは配列データの出典，キーワード，コード領域の翻訳等を含んでいる．**PDB** は，タンパク質及び核酸の 3 次元構造データを含んでいる．**PIR** も，ジャーナル等に掲載されているタンパク質のアミノ酸配列と関連情報を集めている．**PROSITE** は，タンパク質データのモチーフのデータベースである．その他にも，**EMBL** や **SWISS-PROT** などのデータベースがある．

3 構造と機能の予測

アミノ酸配列や核酸配列において，特殊な性質や機能をもっていると思われる部分の位置やその範囲などの個々のデータが，第 2 節で紹介したようなデータベースに蓄えられいる．そうしたデータから，特徴等を表現した知識を発見したり，データについての概念を形成することが重要な課題である．こうした課題の情報科学的解決を通して，より汎用な “Scientific Discovery” の理論や方式を構築することが情報科学としての目的である．

以下の議論では，分かり易さを優先し，生物学的表現の正しさを多少無視しているところがある．Watson et al. [20] や Lewin [8] などの教科書は，分子生物学における基本的な事柄を扱っている．

3.1 タンパク質の構造

タンパク質の**一次構造** (primary structure) とは，20 種類あるアミノ酸を文字とした文字列である．このために，文字列を対象としている計算学習の適用の可能性がある．このアミノ酸の文字列が，これを**アミノ酸配列**とよんでい

るが，一般に，適当な環境において，それが表しているタンパク質の形や機能を決めるとされている．タンパク質の**二次構造** (secondary structure) は，タンパク質の局所的な形を表すもので，主に，**α-helix, β-sheet, turn** の3種類の構造がある．α-helix という構造は，らせん状の構造をしたもので，1つの α-helix は5～30個のアミノ酸から構成されている．Turn とよばれる構造は，ちょうどヘアピンのような形をしており，β-sheet は，波打ったような形をしたシート状の構造である．**二次構造予測問題** (secondary structure prediction problem) とは，タンパク質のアミノ酸配列における α-helix, β-sheet, turn の位置を決める問題である．タンパク質の**三次構造** (tertiary structure) とは，アミノ酸配列の「ひも」が三次元空間にどのように位置しているかを表したものである．タンパク質の構造予測の最終目標は，タンパク質の一次構造からこの三次構造を決めることで，**三次構造予測問題**といわれている．タンパク質の三次構造データが知られ公開されているものはあまり多くないが，PDB データベースがこの種類のデータを蓄えている．

3.2 膜貫通領域予測

タンパク質の構造予測には次のような特殊な問題もある．タンパク質のなかには，細胞膜の内と外を数回縫うようにはまりこんでいるものがあり，膜タンパク質とよばれている．この膜の中にはまりこんでいる部分は**膜貫通領域** (transmembrane domain) とよばれ，その長さはアミノ酸にして 20 前後の α-helix 構造をしたものといわれている．タンパク質のアミノ酸配列の中に，この α-helix 構造をした膜貫通領域にを表すアミノ酸配列がいくつかみいだせるとき，そのタンパク質は**膜タンパク質**である可能性が大きいとされている．**膜貫通領域同定問題**とは，タンパク質のアミノ酸配列が与えられたとき，それが膜貫通領域を含んでいるかどうかを調べ，その膜貫通領域の位置を同定する問題である．PIR データベースは，アミノ酸配列と膜貫通領域の位置についての情報を提供してくれる．

3.3 シグナルペプチッド予測

タンパク質のなかには細胞膜を通り貫けて細胞の外に送り出されるものがある．この種のタンパク質は１５個から３０個のアミノ酸からなる precursor を N 端というアミノ酸配列の左端につけた形で合成される．シグナルペプチッド (signal peptide) とは，この N 端についている purecursor のことである．この配列は通常 Methionine (M) で始まっている．

GenBank は，核酸配列とともにこの precursor 部分の情報を提供してくれる．**シグナルペプチッド同定問題**とは，DNA 配列が与えられたとき，その配列がシグナルペプチッドに対応する部分を含んでいるかを調べ，その位置を決定する問題である．

3.4 プロモータ領域

DNA の転写開始位置の上流には，プロモータ (promoter) とよばれる領域があり，RNA ポリメラーゼによって認識され転写の開始位置が決められるといわれている．DNA 配列においてプロモータとその位置を特徴付けることも重要な課題である．GenBank などがこのための情報を提供している．

4 BONSAI における知識発見方式

ここでは，BONSAI Garden の核プロセスである BONSAI の概略と機能について述べる．

学習アルゴリズムによる知識発見のプロセスを次の4つの段階によりとらえた．

- ビューの設定
- 仮説空間の設定
- 学習アルゴリズムの開発
- 計算機実験

ビューの設定　アミノ酸配列は単なる記号列であるため，その記号列を説明するビューが必要となる．そしてそのビューを通してアミノ酸配列を説明することを考える．例えば，アミノ酸配列を各アミノ酸の頻度でとらえるというのも一つのビューである．PROSITE データベースなどでは，アミノ酸配列を特徴づけるためにモチーフという概念を利用している．この見方

を一般化し，正規パターンというビューでアミノ酸配列をとらえることにした．

Σ を有限アルファベット，$x_1, x_2, \cdots$ を変数記号とする．Σ 上の正規パターンとは，$\alpha_1, \cdots, \alpha_n$ を Σ 上の記号列，$x_1, \cdots, x_n$ を互いに異なる変数記号とするとき，$\alpha_1 x_1 \alpha_2 x_2 \cdots x_n \alpha_n$ の形をした記号列で定義される．これは記号列 $\alpha_1, \cdots, \alpha_n$ をこの順に含んでいる Σ 上の記号列全体を定義する．

膜貫通領域の同定問題では，20 種のアミノ酸に親水度を表す −4.5 から +4.5 の間の実数値を与え，hydropathy plot [7] という方法が有用であることが知られている．本研究では，アミノ酸を親水度により 3 つのカテゴリーに分類し，20 種のアミノ酸を *, +, - につぶし，それにより学習の効率化とより鮮明な知識の獲得に成功している．この経験から得られた概念にアルファベットのインデックス化がある．このアルファベットのインデックス化とは，入力データに使われている文字を，あらかじめ設定された，より少ない個数の文字へ変換する対応づけである．例えば，アミノ酸配列のデータを入力とした場合，20 種類のアミノ酸を親水性かそうでないかに分類したりすることに対応する．こうした変換により正負の情報が失われないことがインデックス化を考える上で重要となる．インデックス化をとおしてアミノ酸配列の集合を見ることも一つのビューである．

仮説空間の設定　正規パターンというビューによってアミノ酸配列をとらえ，こうした正規パターンを用いて概念を説明する規則（仮説）を作ることを考えた．この規則にあたるものとして，**正規パターン上の決定木**という概念を導入した．正規パターン上の決定木は，与えられた記号列をクラス N（負）とクラス P（正）に分類する手続きを記述したものである．

インデックス化というビューを取り入れて，インデックス化の写像 $\psi : \Sigma \to \Gamma$ $(|\Sigma| > |\Gamma|)$ と Γ 上の正規パターンをノードのラベルとする決定木 T を考える．このとき組 (T, ψ) に対して，Σ 上の言語を $L(T, \psi) = \{w \in \Sigma^* : \psi(w) \in L(T)\}$ で定義する．そして，このような写像と正規パターン上の決定木を概念の表現とし，その定義する言語からなる概念クラスを仮説空間とした．

学習アルゴリズムの開発　前述の仮説空間に対して，確率的近似学習 (PAC 学習) の観点から次の知見を得た．正規パターンに現れる変数の個数を定数 k で限定し，決定木の深さを定数 d で限定したような正規パターン上の決定木で定義される概念クラス DTRP(d, k) とするとき，次の定理が成り立つ．

定理 1. DTRP(d, k) は多項式時間 PAC 学習可能である．

この結果の証明は，DTRP(d, k) が多項式次元であることを示すこと，および，例の列 $\langle x_1, a_1 \rangle, \cdots, \langle x_n, a_n \rangle$ を入力するとこれらの例を正の例と負の例に完全に分類する上に述べたような正規パターン上の決定木（もし存在するならば）を構成する多項式時間アルゴリズムを与えることからなる．

この結果は，機械発見への応用の観点から以下のように解釈できる．例えば，ある機能をもつアミノ酸配列を表す概念を変数が k 個以下で，深さが d 以下である正規パターン上の決定木で説明できるときには（こうした説明が可能か否かはだれも前もって知っていないが），あまり多くの例を用いずにほどほどの時間で精度の良い仮説を高い確率で得られることを保証している．逆に何度もサンプルをとり多くの時間をかけても良い精度の仮説が得られないならば，この概念は DTRP(d, k) には属していないと強く信じさせてくれることになる．その意味で仮説空間 DTRP(d, k) を棄却できる．

定理 1の学習アルゴリズムは，多項式時間で走るとはいえ多くの時間を必要とし，実用に耐えられるものではない．また決定木の深さ d や変数の個数 k をどのように設定すればよいかも前もってわかっているわけではない．そこで我々は，Quinlan の ID3 のアイディアを用い，与えられた例に無矛盾な小さな仮説を非常によく見つける効率の良いアルゴリズムを開発した．ID3 は高速で，また実験によると多くの場合，十分に小さい決定木が得られることが知られている．ID3 では，あらかじめ決定木に使われる属性を仮定している．しかし，このアルゴリズムでは，決定木を構成する最中に最適な正規パターンを見つけるので，ID3 のようにデータの属性を選

択しその属性値によってデータを前もって特徴づける必要がない．入力としては，単に，正の例と負の例を表す記号列を与えればよい．

$P \cup N$ に属するすべての記号列の部分記号列を定数記号列として使って構成される正規パターンの集合を $\Pi(P,N)$ とする．正規パターン現にれる変数を 4 ～ 5 個以下としても $\Pi(P,N)$ は巨大である．このため実験では正規パターンを $x_1\alpha_1x_2$ の形に限定して実験を行っている．

インデックス化を見つけることは計算量的に困難であることが判明したので，この決定木の構成アルゴリズムと連動してより良いインデックス化を局所探索法により見つける方式を開発した．ただし，完全なインデックス化を見つけることは，計算量的に困難であるため，インデックス化により変換された記号列の集合にオーバーラップを許している．

こうした方式で開発したものがBONSAIである．BONSAIは，文字列データからの知識発見システムである．BONSAIシステムの概要を図1に与える．BONSAIへの入力は，正の例の集合 *POS* と負の例の集合 *NEG* である．このシステムは，正の例と負の例からなるこれらの記号列の集合が与えられると，それらを分類する仮説として，アルファベットのインデックス化と正規パターン上の決定木を探索する．正規パターンに現れている記号はインデックス化により変換されたものである．

BONSAIに用いられている主なアルゴリズムは二つからなる．一つは決定木生成機で，もう一つは組み合わせ最適化に用いられている局所探索アルゴリズムである．インデックス化に対応する写像をはじめにランダムに設定しておく．1回の試行において，正の例と負の例のサンプル *pos* と *neg* をランダムにいくつか取り出す．そのサンプルを 100% の精度で分類する正規パターン上の決定木を仮説として作り出す．この決定木は，小さなものほど大きな知識を含んでいるという原理にしたがって，できるだけノードの数の小さなものが探索される．次に現在得られているインデックス化 I のもとで *POS* と *NEG* を変換し，その集合に対してこの決定木の精度評価を行う．そして局所探索アルゴリズムによりインデックス化の変更を行う．この

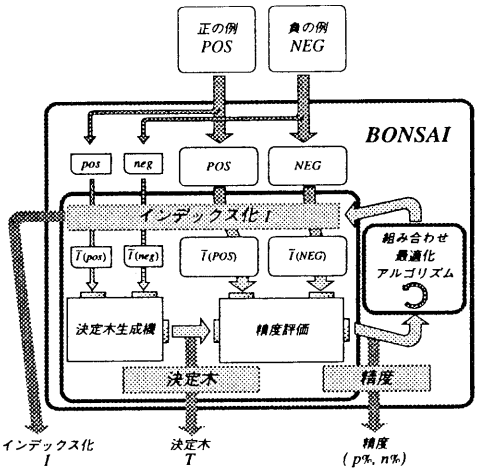

図 1: BONSAIの概念図

プロセスを *pos* と *neg* を固定したままで局所解に落ちるまで続け，その時点でインデックス化と決定木とその精度を出力する．この試行を可能な限り行い，より小さい精度の高い決定木とインデックス化を探索している．

計算機実験 計算機実験により，方式の有効性を確認することが最後のプロセスとなる．

以上が学習アルゴリズムによる知識発見のパラダイムとそれを実現したBONSAIの概要である．

5 BONSAI Garden

知識発見の対象となっているデータが，ノイズを含んでいたり多様な種類の配列の混合体であるとき，BONSAIをそのまま適用したのでは適切な知識の発見を期待できないことがある．そこでこのBONSAIシステムを基本プロセスとして，それを複数個並列に稼働させることにより，次の目標を達成し，より多様性に富んだ知識発見を行うことができるようその方式を追求したシステムがBONSAI Garden[17]である．

- ノイズを含んでいたり多様な種類の配列から構成されるデータを対象とし，これらの配列データをいくつかのクラスに分類する

と同時にこれらの各クラスのデータを高い精度で説明する仮説を発見する．

そこで，このBONSAIを複数個並列に走らせて知識獲得を行うシステムBONSAI Gardenを設計・開発した．そのためにまず以下に述べるような単純な並列知識獲得モデルを構築し，これに基づいてシステムを設計・開発した．

BONSAI Gardenは，複数個のBONSAIとそのコーディネータであるプログラム *Gardener* とからなる（図2を参照）．

まず，各BONSAIのタスクを説明しよう．BONSAI B_i には正の例の集合 POS_i と負の例の集合 NEG_i がデータとして与えられているとする（$i = 0, ..., m-1$）．このときBONSAI B_i は (POS_i, NEG_i) から仮説 (T_i, ψ_i) を作りだす．そして，仮説 (T_i, ψ_i) によってネガティブと分類された POS_i の配列を $POS.TRASH_i$ に入れ，ポジティブと分類された NEG_i の配列を $NEG.TRASH_i$ に入れる．POS_i と NEG_i は (T_i, ψ_i) によってポジティブおよびネガティブに分類された配列で更新する．これがBONSAIのジョブサイクルである．各BONSAIはこのジョブサイクルを繰り返す．

Gardener は，BONSAI $B_0, \ldots, B_{m-1}$ が作り出す仮説や，分類するデータの状況を見ながら，次の仕事をする．（*印のついている部分は，負の例も分類する際に行う作業である．負の例を分類しないときは，NEG_i は固定されたままである．）

1. **Watch:** *Gardener* は，ジョブサイクルを終了する2つのBONSAI B_i と B_j を見つけ出す．B_i と B_j の出力を

$$\begin{array}{lll} (T_i, \psi_i) & POS_i & POS.TRASH_i \\ & NEG_i & NEG.TRASH_i \end{array}$$

$$\begin{array}{lll} (T_j, \psi_j) & POS_j & POS.TRASH_j \\ & NEG_j & NEG.TRASH_j \end{array}$$

とする．

2. **Compare:** *Gardener* は，仮説 (T_i, ψ_i) と (T_j, ψ_j) のサイズを比較し，どちらが小さいかを決める．このためには，仮説のサイズを定義しておく必要がある．BONSAI Gardenでは，決定木を表現したときに記号列としての長さをサイズとしている．次の説明のために，仮説 (T_i, ψ_i) のサイズが仮説 (T_j, ψ_j) のサイズよりも小さいとしよう．

3. **Classify:** *Gardener* は，BONSAI B_j の出力 POS_j に入っている配列を，小さな仮説 (T_i, ψ_i) により以下のように再分類する．

   (a) $POS.NEW_i$ を，(T_i, ψ_i) によってポジティブに分類される POS_j の配列の集合とする．(*: $NEG.NEW_i$ を，(T_i, ψ_i) によってネガティブに分類される NEG_j の配列の集合とする．)

   (b) $POS.NEW_j$ を，(T_i, ψ_i) によってネガティブに分類される POS_j の配列の集合とする．(*: $NEG.NEW_j$ を，(T_i, ψ_i) によってポジティブに分類される NEG_j の配列の集合とする．)

4. **Merge:** *Gardener* は，POS_i と POS_j (*:NEG_i と NEG_j) を次のように更新する．

   (a) $POS_i \leftarrow POS_i \cup POS.NEW_i$
   (*:$NEG_i \leftarrow NEG_i \cup NEG.NEW_i$)

   (b) $POS_j \leftarrow POS.NEW_j$
   (*:$NEG_j \leftarrow NEG.NEW_j$)

5. **Distribute Trash:** $POS.TRASH_i$ (*: $NEG.TRASH_i$) をBONSAI B_{i+1} に渡す．これは次の B_{i+1} のジョブサイクルの開始時点において POS_{i+1} にマージされる．また $POS.TRASH_j$ (*: $NEG.TRASH_i$) も同様にBONSAI B_{j+1} に渡す．ただし $i+1 = m$ $(j+1 = m)$ のときは0とする．

上記の *Gardener* の仕事によると，より小さな仮説を生成したBONSAIにはより多くのデータがいき，大きな仮説を生成したBONSAIの

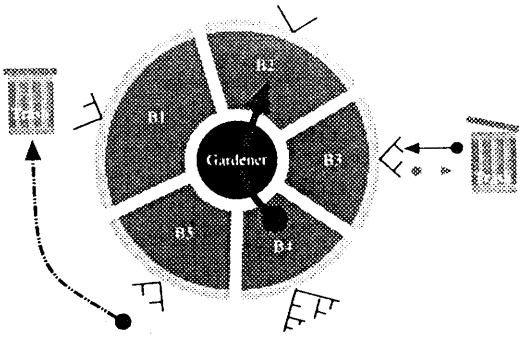

図 2: BONSAI Garden

データは減少することになる．また，分類に失敗したデータは，どれかのBONSAIに受理されるまでBONSAIの間を循環することになる．そして，すべての $POS.TRASH_k$ (*: $NEG.TRASH_i$) が空になれば，*Gardener* の仕事は終わるが，空になるという保証は何もない．そのため，BONSAI Gardenを終了するプロセスが用意されている．

このような考えで設計されたBONSAI Gardenは，現在ワークステーションのネットワーク上で実働化されている．これにより，全体として大きな計算量に対応できるようになっている．

また，*Gardener* の仕事を以上の述べた以外に定義することにより，別の機能をもったBONSAI Gardenを構築することが可能となる．これについては，カスケード方式などいくつかの方式を実働化して実験中であるが，理論的な考察は深く行えていない．

6 おわりに

BONSAIでは，正規パターンに制約を設けていたが，現在のバージョンではlongest common subsequenceのアルゴリズムを適用した方法により，複雑な形のパターンの抽出が可能となっている．Arimura et al. [4]では，正の例だけからそれらをカバーする正規パターン集合を探す比較的効率のよいヒューリスティックアルゴリズムを提案している．こうした方法も今後システムの効率化と機械発見に有効に利用できる可能性がある．

これまでの研究ではモチーフを正規パターンとしてとらえているため，変数に代入できる記号列の長さに制約がない．しかし，PROSITEのデータベースに登録されているモチーフのほとんどは変数に代入できる文字列の長さや文字の種類に制約がついたものがほとんどである．Tateishi et al. [18]は，こうしたさらに表現力の大きくなっているモチーフを対象として，モチーフ発見を最適化問題としてとらえ，**Best Consensus Motif問題**として定式化し，その計算量のほぼ全容を解明し，また一部の問題に対して多項式時間近似アルゴリズムをその近似率の解析とともに与えている．

謝 辞

本研究を協同して遂行していただいた九大理学部有川節夫，正代隆義，篠原 歩，九大農学研究科遺伝資源工学専攻久原 哲，琉球大学岡崎威生，広島市立大内田智之，九工大情報工学部篠原 武，下薗真一，Universität Paderborn Michael Lappeの諸氏に深く感謝いたします．また，九大大学院総合理工学研究科情報システム学専攻大学院生の古川直広君には，BONSAIおよびBONSAI Gardenでの多くの実験を行なっていただきました．

参考文献

[1] Angluin, D.[1980], Finding patterns common to a set of strings, *J. Comput. System Sci.* **21**, 46–62.

[2] Arikawa, S., Kuhara, S., Miyano, S., Mukouchi, Y., Shinohara, A., and Shinohara, T. [1993], Machine discovery of a negative motif from amino acid sequences by decision trees over regular patterns, *New Generation Computing* **11**, 361–375.

[3] Arikawa, S., Kuhara, S., Miyano, S., Shinohara, A., and Shinohara, T. [1992], A learning algorithm for elementary formal systems and its experiments on identification of transmembrane domains, *Proc. 25th Hawaii International Conference on System Sciences*, 675–684.

[4] Arimura, H., Fujino, R., Shinohara, T. and Arikawa, S. [1994], Protein motif discovery

from positive examples by minimal multiple generalization over regular patterns, in *Proc. Genome Informatics Workshop 1994*, Universal Academy Press, 39–48.

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

[6] Gribskov, M. and Devereux, J. [1991], *Sequence Analysis Primer*, UWBC Biotechnical Resource Series, Macmillan Publishers Inc.

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

[8] Lewin, B. [1987], *Genes: Third Edition*, John Wiley & Sons, Inc.

[9] 宮野悟，篠原歩，有川節夫 [1994]，ゲノム情報における機械学習の計算量 — 理論と実際 —，人工知能学会誌 **9**, No. 3, 350–356.

[10] Miyano, S. [1995], Learning theorey towards Genome Informatics, *IEICE Trans. Information and Systems*, **E78-D**, No. 5, 560–567.

[11] Miyano, S., Shinohara, A., and Shinohara, T. [1993], Learning elementary formal systems and an application to discovering motifs in proteins, Technical Report RIFIS-TR-CS-37, Research Institute of Fundamental Information Science, Kyushu University, revised in April, 1993 (former version: Proc. 2nd Algorithmic Learning Theory, 139–150, 1991).

[12] Natarajan, B.K. [1991], *Machine Learning - A Theoretical Approach,* Morgan Kaufmann Publishers.

[13] Quinlan, J.R. [1986], Induction of decision trees, *Machine Learning*, **1**, 81–106.

[14] Shimozono, S. and Miyano, S. [1995], Complexity of finding alphabet indexing, *IEICE Trans. Information and Systems*, **E78-D**, No. 1, 13–18.

[15] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S. [1994], Knowledge acquisition from amino acid sequences by machine learning system BONSAI, *Trans. Information Processing Society of Japan*, **35**, No. 10, 2009–2018.

[16] Shinohara, T. [1983], Polynomial time inference of extended regular pattern languages, *Proc. RIMS Symp. Software Science and Engineering* (Lecture Notes in Computer Science, **147**, 115–127.

[17] Shoudai, T., Lappe, M., Miyano, S., Shinohara, A., Okazaki, T., Arikawa, S., Uchida, T., Shimozono, S., Shinohara, T., and Kuhara, S. [1995], BONSAI Garden: parallel knowledge discovery system for amino acid sequences, to appear in *Proc. 3rd International Conference on Intelligent Systems for Molecular Biology*, AAAI Press.

[18] Tateishi, E., Maruyama, O. and Miyano, S. [1995], Extracting motifs from positive and negative sequence data, RIFIS-TR-CS-115, Research Institute of Fundamental Information Science, Kyushu University.

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

[20] Watson,J.D., Hopkins, N.H., Robets, J.W., Steitz, J.A., and Weiner, A.M. [1987],*Molecular Biology of The Gene: Fourth Edition*, The Benjamin/Cummings Publishing Company, Inc.