アルゴリズム 40－6
（1994．7．22）

BONSAI Garden: 学習アルゴリズムによるアミノ酸配列からの並列知識獲得システム

宮野 悟* 篠原 歩* 内田智之+ 久原 哲‡
下薗真一† 篠原 武† 正代隆義* 有川節夫*
*九大理学部 +広島市立大学 †九工大情報工学部 ‡九大遺伝資源工学

概要

これまでに，我々は，学習アルゴリズムによる知識獲得システムBONSAIを開発し，主にアミノ酸配列からの知識獲得実験を行なってきた．このシステムは，正の例と負の例からそれらを説明する仮説を学習するもので，これまでの実験で，膜貫通領域とシグナルペプチッド配列に対して，非常に精度の高い小さな仮説を発見し，BONSAIがその能力において大きなポテンシャルをもっていることが判明した．このBONSAIを基本プロセスとして，これを複数個走らせる並列知識獲得システムBONSAI Gardenのプロトタイプを作成し，これまでBONSAIで実験を行なってきたものと同じデータを用いて，比較・検討し，その有効性を確認した．

BONSAI Garden: Knowledge Acquisition System from Amino Acid Sequences by Learning Algorithm

Satoru Miyano* Ayumi Shinohara* Tomoyuki Uchida+ Satoru Kuhara‡
Shinichi Shimozono† Takeshi Shinohara† Takayoshi Shoudai* Setsuo Arikawa
* Faculty of Science, Kyushu University † Kyushu Institute of Technology
‡ Genetic Resources Technology, Kyushu University + Hiroshima City University

Abstract

We have developed a machine learning system BONSAI which gets positive and negative examples as inputs and produces a pair of a decision tree over regular patterns and an alphabet indexing as a hypothesis. We observed that the indexing by the hydropathy indices is important in making the learning algorithms efficient and accurate. We are also implementing a system BONSAI Garden, which runs several BONSAI systems in parallel, to acquire knowledge from a hodgepodge of data. Some experiments on the prototype shall be presented.

1 背景と目的

ヒトのDNAは総計3×10^9個の塩基対からなる巨大な情報分子であり，この中にヒトの遺伝情報の全てが記述されているといわれている．生物学的・社会的見地から，この情報を解読することが強く要請されている．ゲノム解析の最終目標の一つは，大量に蓄積されるゲノムデータから，生命に関する設計図や発現に関する制御情報などの高次の情報を抽出することである．こうした背景のもとで，記号列として与えられたゲノムおよびタンパク質等のデータから，様々な学習アルゴリズムによりそれに潜在する知識を獲得する方式を構築することが重要な課題となっている．

こうした要請に応えるために，我々は，学習アルゴリズムによる知識獲得システムBONSAIを開発し，主にアミノ酸配列からの知識獲得実験を行なってきた．このシステムは，正の例と負の例からそれらを説明する仮説を学習するもので，具体的には，アミノ酸のインデックス化と正規パターン上の決定木を仮説とする．これまでの実験で，BONSAIは膜貫通領域とシグナルペプチッド配列に対して，非常に精度の高い小さな仮説を発見し，BONSAIがその能力において大きなポテンシャルをもっていることが判明した．

このBONSAIを複数のプロセスとして並列に走らせることにより，ゲノムデータからの新たな知識獲得の方式を構築することが次の課題となっている．

2 方　法

タンパク質や核酸についての配列データは，その機能や性質により分類されてデータベースに整理されている．しかし，そうしたデータにはノイズが含まれていることが多く，また1つのクラスと分類されているデータも，いくつかの未知のクラスの混ぜ合わせとなっている可能性もある．

こうしたノイズを含んだデータや混ぜ合わせのデータにBONSAIを適用すると，インデックス化と正規パターン上の決定木として知識を表現しているため，良い知識の表現が得にくくなる．そこで，BONSAIシステムを複数のプロセスとして並列に走らせ，相互にコミュニケーションをとりながら，正の例と負の例から，そのデータについての知識を，データの分類とともに獲得する方式を以下に述べるように構築した．そして，この方式のプロトタイプを，BONSAI Gardenというシステムとして実現した．その計算機実験については，次節に述べる．

2.1 並列知識獲得システムBONSAI Garden

ここでは，BONSAI Gardenの核プロセスであるBONSAIの概略と機能およびBONSAI Gardenの構成について述べる．

2.1.1 BONSAIにおける知識獲得方式

学習アルゴリズムによる知識獲得のプロセスを次の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

以下にその説明を行なう．

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$\Sigma$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$\Sigma$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獲得システムである．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獲得のパラダイムとそれを実現したBONSAIの概要である．

2.2 BONSAI Gardenの方式と概要

POSとNEGを各々正の例と負の例の集合とする．POSがいくつかの未知の概念$c_1, \cdots, c_t$からの例を集めたものであるときは，これらの正負の例の集合を1つの決定木とインデックス化によって説明することは適当ではない．このような場合，正の例の集合POSを$POS = Pos_1 \cup \cdots \cup Pos_m$

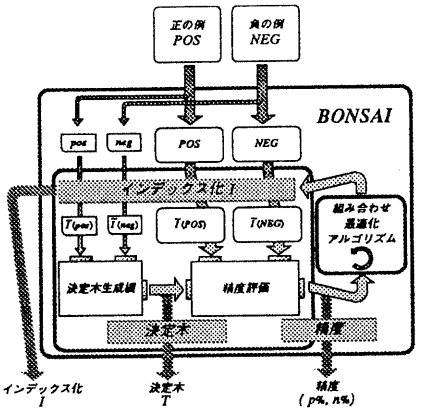

図 1: BONSAIの概念図

と分割し，各Pos_iに対して，精度の高い決定木とインデックス化を学習させることが望ましい．

BONSAI Gardenは，複数のBONSAIをプロセスとして走らせることにより，上記の目的を達成しようとするものである．この方式は，正の例の集合が数種類のアミノ酸配列の混ぜ合わせとなっているときに極めて有用である．

この方式を大きくとらえると次のように書ける．

```
begin /* POS and NEG are given as input */
  let P be a partition of POS;
  repeat
    Classify(P);
    Merge;
  until Merge is impossible
end
```

ここで，POSの分割$\mathcal{P} = (Pos_1, \ldots, Pos_m)$は，初めに適当な$m = 2^r$をとって，$POS$をほぼ同じ大きさの$m$個の部分集合にランダムに分割したものである．分割$\mathcal{P}$が与えられたとき，$Classify(\mathcal{P})$はBONSAIを並列に走らせ，以下の操作を実行する．

各BONSAIの入力は，$\langle Pos_i, NEG \rangle$である．$(p, n)$を正の例と負の例に対する精度とするとき，精度評価として$(p \cdot n^2)^{\frac{1}{3}}$を用いている．これは，ほとんどすべての$NEG$の例を排除するが，$Pos_i$に対しては，あまり高い精度を要求しないようにしているためである．さらに，決定木生成機では，深さがある定数dをこえないような決定木のみを構成するようにしている．

$B_1, \ldots, B_m$ を並列に走る BONSAI とする．このとき，次のような一連のステージを実行する．各ステージは，2つの操作からなる．第 j ステージは，次の2つの操作が行なわれる．

1. $\langle Pos_i, NEG \rangle$ を入力とする BONSAI B_i を走らせ，結果 $(T_i, I_i, (p_i, n_i))$ を得る．
2. 図 2 の Step j で組にされている BONSAI の間で正の例を交換する．組 (B_i, B_j) の間では，次のように正の例が交換される．

   $Pos_i \leftarrow (L(T_i, I_i) \cap Pos_i) \cup (Pos_j - L(T_j, I_j))$

   $Pos_j \leftarrow (L(T_j, I_j) \cap Pos_j) \cup (Pos_i - L(T_i, I_i))$

   ここで，$L(T_i, I_i)$ は，インデックス化 I_i のもとで，決定木 T_i によってクラス P（正）と判定される文字列の集合である．すなわち，$(T < j, I < j)$ によってクラス N（負）と判定される Pos_j の例と $(T < i, I_i)$ によってクラス N（負）と判定される Pos_i の例が交換される．

これは第 r ステージまで実行される．こうして得られる新たな分割 $\mathcal{P} = (Pos_1, \ldots, Pos_m)$ と決定木とインデックス化の組 $(T_1, I_1, (p_1, n_1)), \ldots, (T_m, I_m, (p_m, n_m))$ が *Classify*($\mathcal{P}$) の出力となる．

Merge は，こうして得られた分割 $\mathcal{P}$ をより小さな分割 $\mathcal{P}'$ に変換する．形式的には，*Merge* は次のようになる．

```
procedure Merge
  begin /* P and (T_i, I_i, (p_i, n_i)) (i = 1, ..., m)
      are given */
    J ← {1, ..., m};
    s ← true;
    while s = true do begin
      for each (i, j) in J × J do
        p_i^j ← |Pos_i ∩ L(T_j, I_j)| / |Pos_i|;
      if there is (i, j) ∈ J × J with i ≠ j satisfying
        (a) p_i^i ≤ p_j^j ≤ p_i^j and
        (b) n_i ≤ n_j
        then
          begin
            let (i, j) be a pair satisfying
              (a) and (b) with thelargest p_i^j;
            Pos_j ← Pos_j ∪ Pos_i; /* merge */
            J ← J − {i};
          end
        else s ← false
      end;
    return P' = (Pos_j)_{j∈J}
  end
```

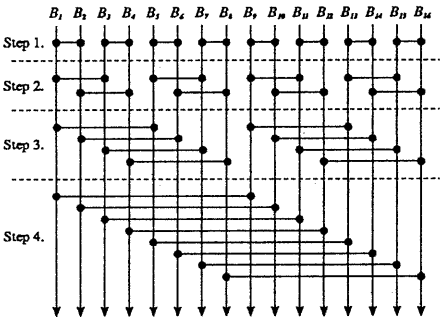

図 2: 正の例の交換方法

条件 (a) は，Pos_i と Pos_j をマージした後では，$Pos_i \cup Pos_j$ に対する (T_j, I_j) の精度は，Pos_j に対する (T_j, I_j) および Pos_i に対する (T_i, I_i) との精度よりも低くないことを保証している．条件 (b) は，NEG に対する (T_j, I_j) の精度が少なくとも (T_i, I_i) の精度以上であることを保証している．したがって，*Merge* を実行した後では，分割の個数は減少し，仮説の各精度は決して悪くはなっていない．定理の形でまとめると次のように述べられる．

定理 2 $\mathcal{P} = (Pos_1, \ldots, Pos_m)$ と $(T_i, I_i, (p_i, n_i))$ $(i = 1, \ldots, m)$ を *Merge* の入力とし，その出力を $\mathcal{P}' = (Pos'_{j_1}, \ldots, Pos'_{j_{m'}})$ とする．このとき，次のことが成り立つ．

(1) $m' \leq m$.

(2) 分割 $\mathcal{P}$ の任意の Pos_i に対し，Pos_i がマージされてできる $\mathcal{P}'$ の集合を Pos'_j とする．すなわち，$Pos_i \subseteq Pos'_j$ となっている．このとき，次の不等式が成り立つ．

$$\frac{|Pos_i \cap L(T_i, I_i)|}{|Pos_i|} \leq \frac{|Pos'_j \cap L(T_j, I_j)|}{|Pos'_j|}.$$

3 結　果

並列知識獲得システム BONSAI Garden を使った計算機実験とその結果について述べる．実験のためのデータは，これまでの研究で用いてきた膜貫通領域およびシグナルペプチッドのアミノ酸配列データを使用した．

BONSAI は計算機パワーを多く必要とするためワークステーション上で数多くの BONSAI を

走らせることには無理があるため，実験では4つのBONSAIを走らせるのに留めた．従って$m=2$である．

以下に実験の方式とその実験結果について述べる．

3.1 膜貫通領域

*POS*としては，PIRデータベースから取り出した181個の膜貫通領域のアミノ酸配列を使い，*NEG*としては膜貫通領域になっていない部分から400個の長さが30前後のアミノ酸配列をランダムにPIRデータベースから取り出したものを使った．

図3は，その実験結果である．　等分割された*POS*は，最終的にはほぼ1つのクラスにまとまっている．わずかの膜貫通領域配列が3つの残りのクラスに「ゴミ」のように集められていることがわかる．

3.2 シグナルペプチッド

*POS*しては，GenBankデータベースのprimate, bacteria, virusの3つのファイルから合わせて121個のシグナルペプチッドのアミノ酸配列を取りだし，*NEG*としてはN端からMで始まるシグナルペプチッドになっていない長さが30のアミノ酸配列をランダムに同じファイルから244個取り出したものを使った．

図4がその実験結果である．正の例の4つのクラスは大きな1つクラスとそれ以外の小さなクラスに分類されているが，その分類のされ方は膜貫通領域の実験のときのように顕著ではないことがわかる．

4 考　察

PAC学習アルゴリズムによる知識獲得システムBONSAIを基本プロセスとして，これを複数個走らせる並列知識獲得システムBONSAI Gardenのプロトタイプを作成し，これまでBONSAIで実験を行ってきたものと同じデータを用いて，比較・検討をした．想定されていた結果とほぼ同じものが得られ，並列知識獲得方式が有効に働いていることを確認できた．

前にも述べたように，基本プロセスBONSAIは計算機パワーを必要とし，またBONSAI Garden の構成から，そのBONSAIを複数個走らせるBONSAI Gardenは，MIMDタイプの並列計算機上で実働化することが望ましい．もしくは，ワークステーションのネットワーク上で実働化することが考えられる．現在，その実働化に向けて準備を行っている．

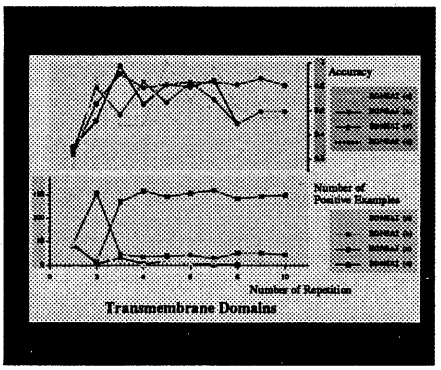

図3: BONSAI Gardenの膜貫通領域データについての実験結果

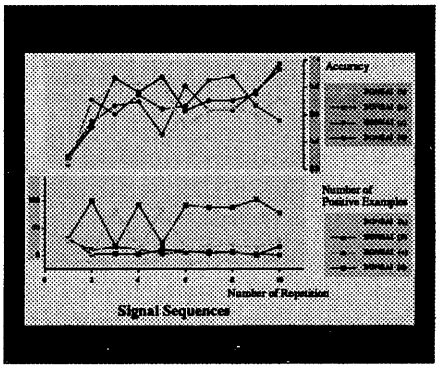

図4: BONSAI Gardenのバクテリアのシグナルペプチッドデータに対する実験結果

さらに，BONSAIの中の，インデックス化と正規パターンを遺伝的アルゴリズムを用いて高速化することも検討中である．また，より曖昧さを許した仮説の表現を追求することも今後の大きな課題として残っている．

謝辞　本稿の内容は，文部省重点領域研究「ゲノム情報」の一部として筆者等が行なってきた研究について述べたものである．

5 論文発表

1. Shinohara, A., Shimozono, S., Uchida, T., Miyano, S., Kuhara, S. and Arikawa, S. [1993], Running learning systems in parallel for machine discovery from sequences, Proceedings of Genome Informatics Workshop IV, 74–83.

2. Miyano, S. [1993], Learning theory toward genome informatics, Proceedings of the 4th International Workshop on Algorithmic Learning Theory, 19–36.

3. 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.

4. 宮野 悟 [1994], 機械学習理論を適用したアミノ酸配列からの知識獲得, 計測と制御, **33**, 40–45.

5. 宮野 悟, 篠原 歩, 有川 節夫, 下薗 真一, 篠原 武, 久原 哲 [1993], 学習アルゴリズムによるアミノ酸のインデックス化とタンパク質データからの知識獲得実験, 九州大学大型計算機センター計算機科学研究報告第 10 号, 47–52.

6. Shimozono, S. and Miyano, S. [1992], Complexity of finding alphabet indexing, Technical Report RIFIS-TR-CS-61, Research Institute of Fundamental Information Science, Kyushu University, August.

7. Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S. [1993], Finding alphabet indexing for decision trees over regular patterns: an approach to bioinformatical knowledge acquisition, *Proc. 26th Hawaii International Conference on System Sciences*, 763–772.