博 士 論 文

機械学習を用いたエピジェネティクス関連領域の予測と属性選択

指導教官　池田 満 教授

北陸先端科学技術大学院大学
知識科学研究科知識システム基礎学専攻

東原 正智

2011 年 9 月 30 日

要 旨

本研究の目的は、近年活発に研究されているエピジェネティクス現象を示すヒストンの化学修飾データである遺伝子の配列を対象に、遺伝子の発現が活性化または不活性の２値判別を機械学習による判別分析で行うことである。エピジェネティクス現象とは、遺伝子の発現においてセントラルドグマによる発現ばかりではなく、化学的な作用で遺伝子の発現が制御される現象である。配列は、n-gram による sliding window で特徴ベクトルを作成し、判別に寄与する属性を RandomForest の Gini 係数 (varibale importance) によって、属性をランキング (feature ranking) した。そのランキングを基に判別に対しての重要な属性部分集合を探索する実用的な近傍探索 (Local Search) アルゴリズムを提案した。さらに、特徴ベクトルとして配列解析では、sliding window を用いた頻度ベクトルを使うが、そのほか位置的な情報を考慮した特徴ベクトルも考えられる。本研究では、そうした様々な特徴ベクトルとして表現のなかで最適な特徴ベクトルの表現も目的としている。また、本研究では、RandomForests の寄与度 variable importance の特性を調べるために SOM でのクラスタリングでの比較を行った。寄与度 variable importance のグラフと予測率の間には関連性があり、予測率が高い場合には特徴あるグラフを示すことがわかった。また、SOM クラスタリングでの可視化においてもそれを裏付ける特徴があった。さらに、本研究で提案した近傍探索アルゴリズムの有効性を測るため、一般的な機械学習のベンチマークデータに対して予備実験を行った。その結果においても予測率が高い属性の組合せの近傍により高い予測率を示す可能性のある属性の組合せがあることを示した。なお、属性選択 (feature selection) と属性部分集合選択 (feature subset selection) は混在している文献が多いが、本論文では、２者の相違点は、前者がデータの削減を目的とするが、後者は部分集合の探索を目的とするように定義する。

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 研究の背景

1.1.1 遺伝子配列解析 (高次元配列データの処理について)

ヒトゲノム計画の進展とともに、DNA の配列データを高速かつ高精度に決定する技術は急速に進歩した [1][2]。これに伴い、データベース上に大量に蓄積された遺伝子配列やアミノ酸配列などの配列データを使って生体分子の構造や機能を予測する研究が盛んに行われている。その際、配列データを固定長の特徴ベクトルとして表現するために、n-gram の頻度を用いる手法が広く用いられている。例えば、Pham らによって行われた研究 [15] では、与えられた配列におけるヒストンの化学修飾（アセチル化やメチル化）を予測するために、3-gram から 11-gram までの頻度を特徴ベクトルとして用いた。一般に n-gram を用いた特徴ベクトルでは n を増やすほど疎な特徴ベクトルになり、特徴数が大幅に増加する。例えば、塩基配列を n-gram の頻度で表現する場合、特徴ベクトルの次元数は 4^n になる。しかし、特徴数が多くなり過ぎると判別の精度を落としてしまう特徴が含まれるため、精度向上のためにはこのような特徴を除くことが重要である。本研究では、最近活発な研究がおこなわれているエピジェネティクスの研究の実験で抽出された解析対象である DNA の配列データを用いている [6]。エピジェネティクスとは、個体発生や細胞分化の過程において、DNA の遺伝情報を変更することなく化学的に遺伝子発現を制御する現象の総称として使われている。すなわち、DNA の遺伝情報のみが遺伝子発現にかかわるのではなく、化学的な修飾によっても遺伝子発現が制御されている現象のことをさしている [18]、[21]。細胞核内では、遺伝情報が書き込まれた DNA 鎖である DNA がすべて収まっている。たとえば、人間の全 DNA 配列は、23 本の DNA 鎖、染色体で構成され、2 倍体である体細胞には 46 本の染色体があり、全体で約 6000Mb、直線にすると 2 mに及び、その長さの DNA が

10 μ 程度の細胞核内に高圧縮に折畳まれている。真核生物の染色体 DNA は、クロマチンという構造をとっている。クロマチンは、ヌクレオソームの繰り返し構造がらせん状につながったものでヌクレオソームは、H2A、H2B、H3、H4 ヒストンタンパク質が 2 分子からなるヒストンオクタマーに、146 塩基対の DNA が約 2 回転巻付いている構造をとっている。ヒストンはリジンなどの塩基対アミノ酸をもつタンパク質で、酸性である DNA と堅く結合しており、ヒストンの N 末端は、ヒストンテールと呼ばれ、この部位はいろいろな化学的な修飾を受ける。　近年、転写誘導の際にヒストン修飾によるクロマチン構造変換が重要な働きをすることがわかってきている。さらにヒストンは、アセチル化、メチル化、リン酸化などの修飾をうけ、転写の制御・サイレンシング・クロマチン凝縮などを引き起こすことが知られている。その他 DNA のメチル化、クロマチン構造の形成とモデリング、転写因子のネットワークもエピジェネティクスを担う役割と考えられており、活発に研究されている分野である機械学習によるエピジェネティクス関連領域の予測の先行研究としては、Pham らによる SVM を用いた 研究がある [15]。彼らは RBF カーネルを用いて予測を行う一方で、別途 polynomial kernel で学習した際の重みを用いて特徴のランキングを行うことにより、特徴ベクトルの属性の重要性を解析している。さらに、Tran らによる研究では、Conditional Random Field を用いて予測を行い、SVM との比較を行っている [16]。

配列解析において高次元配列データの特徴ベクトルの解析が必要となる理由は、遺伝子の発現が離れた位置での複数の遺伝子が関連している場合があるためである。そのため離れた遺伝子の部位の相関性を考慮した機械学習の処理が必要となっている。配列の n-gram の長さが長くとることは、相関性のあるモチーフを特定するためである。しかし、配列の n-gram の長さが長くなると組合せも多くなり、その結果作成される特徴ベクトルも疎 (スパース) になる。また、次元の呪いのため高次元になると汎化誤差が向上しなくなる問題が生じる。そのため属性選択し不要な属性を削除する必要性が要求されるようになった。また、先行研究では、頻度ベースの特徴ベクトルを使用することが多いが、本研究では、位置特異的な情報をもつ特徴ベクトルの特性についても解析対象とした。これは、頻度データでは位置情報が欠損しており、頻度とともに位置も化学的な制御との関連も予想されるためである。

1.1.2　variable importance(寄与度での定量評価とその利用)

本研究の動機付けとして

1. データ削減 (feature selection)

2. データの全体的な傾向や特性の把握 (feature ranking)

3. 予測率が最も高い最適な特徴表現をもとめること (feature subset selection) 1

が挙げられる。上記の 1-1-1 で述べたように疎な (スパース) 高次元配列データに伴う属性選択が本研究の出発点であったが、そこでは、1. のデータ削減が目的であった。データ量の削減には、大別して属性の削減と事例の削減の 2 つの方法があるが、本研究では前者を対象とした。属性選択は、与えられたデータの属性の中から目的に対して有効な属性を選択し、余分な属性を削除しデータを削減することである。属性選択には、大別してフィルタ法とラッパ法の 2 種類がある。前者は、属性選択の評価に適当な指標を計算して用いる。後者は、学習結果そのものを用いる。属性選択の手順は（例:フィルタ法）

1. データに対して評価基準 (エントロピー値など) を計算する。

2. 評価された基準にしたがって属性をランキングする。属性数が n 個とすると 1 通りのランキングが決まる。

3. 評価基準の一番低いものから一つずつ削除していき、残った属性で学習する。（このとき探索方向は後ろ向きという。探索の戦略は全探索とする。）

4. 学習結果が一番良いものを最適な属性とする。1

属性選択の手法の分類として、上記の例のように評価基準と探索法（方向、戦略）の観点から分類できる。評価基準としては、ラッパ法では、学習結果そのものが用いられる。フィルタ法では、情報利得や Gini index などが用いられる。また、その他にも様々な有用な指標が提案されている。探索の方向とは、なにも選択されていない状態から先のランキングに従って一番有効な属性から順次追加していく前向き探索、逆に全属性から出発し一番有効ではない属性から順次削除していく後ろ向き探索、両方から探索する両方向探索、属性数が多い時に使うランダム探索などがある。探索の戦略とは、属性の探索空間 (属性数を n 個とすると全組合せは、2^n-1 となる。n が大きくなると膨大な組合せとなる。) を如何に探索するかという戦略である。大別すると、完全探索（全空間探索と部分探索)、ヒューリステック探索 (最良優先探索、ビーム探索、欲張り探索), 非決定探索に分類される。実際に提案されている代表的な属性選択アルゴリズムでは、Focus、Relief、ABB などがある [3][4]。

本研究では、RandomForest[26] という学習アルゴリズムを用い、その途中のプロセスで計算される Mean Decrease Gini index を評価基準として属性のランキングを求めた。探索の方向としては、前向き探索を用い、探索戦略として全ての部分集合を探索する完全探索を用いた。その結果、一意の決定したランキングから前向き探索を行った場合の n 個の属性の部分集合以外にも予測率が上の部分集合があることを計算機実験により示した。Zenglin Xu[17] は、属性の探索の戦略の性質に対して、単調性の定義と MKL（Multiple Kernel learning）による非単調な探索アルゴリズムを提案している。単調性に関しては、以下のような定義をしている。

定義（単調な属性選択アルゴリズム）

A:属性選択アルゴリズム

Sm:属性選択アルゴリズム A によって選択された m 個の部分集合

属性選択アルゴリズム A は単調である。

⇔属性数が k,m($k \leq m$) のとき, 常に $S_k \subseteq S_m$ となる。

本研究では、ランキングされた n 個の組合せのみではなくその近傍を探索するアルゴリズムを提案し計算機実験により比較を行った。従来の機械学習の研究では、予測率の向上に焦点が当てられていた。最新の機械学習アルゴリズムにおいても劇的な予測率の向上はみられない。そのような状況からその予測率の範囲内でデータについての知見が得られないかということが問題意識としてあった。生物学や医学などの分野では、予測率ばかりではなく説明能力をも求められることが多い。2) の feature ranking は、そういったデータの全体的な傾向を示す指標として提案したものである。多変量解析の線形判別分析においては、寄与率の計算は、基本的な手順であるが、機械学習では、その評価に言及されることが少ない [23]。属性の重要度の全体的な把握は、理論的な動機というより、実際のデータ解析のユーザーである実験系の立場からの問題提起でもあった。寄与率の相当する指標がいくつか提案されているが、それらの指標からデータの全体的な傾向の把握ができるのではないかという考えがあった。

新島ら [18] の研究では、カーネル関数を介して構成される相互作用空間において属性選択をする研究を提案している。説明変数に相関性がある場合の研究は、RandomForest においても最近研究されてきている [45]。

1.2 本研究の目的

本研究の目的は、エピジェネテイクス現象を示す遺伝子配列を対象として、活性化及び不活性化を示す判別分析を機械学習で行った。先行研究では、判別に寄与する指標が部分的ではあったため、本研究では、RandomForest の寄与度 (variable importance) を用いて全属性について寄与度を求めることである。目的としては、寄与度からの知見つまり予測率、相関性、SOM クラスタリングとの関連から対象データからの知見を求めることである。従来、属性選択において何らかの指標を求めて順序付けを行い、属性の組合せを行っている。その場合、属性集合の全探索空間を探索してはいない。そのため指標による順序づけされた属性の組合せ以外に予測率が上回る組合わせがある可能性を検証するため計算機実験を行う。機械学習の一般的なベンチマークテストにより検証を行う。最終的には、本研究の対象であるエピジェネテイクス現象を示す遺伝子配列に対して、提案する近傍探索の属性部分集合探索を行う。次に、配列解析には、sliding window を用いた頻度ベクトルが用いられることが多い。特徴ベクトルには、それ以外に様々な特徴ベクトルが考えられる。本研究では、マルチプルアライメントで使用される Position specific scoring matrix を参考にして位置情報を特徴ベクトルに表現することを提案した。予測率を指標として最適な特徴ベクトルの表現または条件を示すことが本研究の目的の 1 つである。

1.3 本論文の構成

本論文の構成は以下の次の構成となる。

第 2 章は、遺伝子配列解析の概要と本研究の対象であるエピジェネティクス、位置特異行列に関する説明をする。

第 3 章では、機械学習に関する概要、生成モデルと識別モデル、bootstrap、Random-Forests の説明である。

第 4 章では、機械学習においての属性処理についての一般的な手法、バイオインフォマティクスでの特徴選択、特に配列解析、マイクロアレイ解析での特徴選択の先行研究、また判別解析において最適で最小の属性集合を発見する Minimal Optimal Problem と目的とする変数に関連する全ての属性を発見する All Relative Problem について説明をする。

第 5 章では、主論文である RandomForest を用いたエピジェネティクス現象を示す配列に関する予測と属性選択の効果を説明した。

第 6 章では、主論文での解析で用いられた寄与度 (variable importance) の属性のランキングと SOM のクラスタリングとの比較を行った。

第 7 章では、位置特異な特徴ベクトルに注目した予測と属性部分集合選択について説明する。

第 8 章では、UCI バークレイの機械学習のベンチマークデータから一般的なデータを選び、近傍探索の裏付けとなる予備実験を行った。属性数 n の場合、探索空間は $2^n - 1$ となるが、ある評価値でランキングし、属性の組合せを考えると n 通りの組合せを考慮することになる。

第 9 章では、研究の結論と今後の研究について述べた。

第 2章

遺伝子配列解析

2.1 配列解析の先行研究

バイオインフォマティクスにおいて配列解析とは、生物遺伝子配列 (DNA、RNA、ペプチドなどの配列) に対して

1. データの格納 (データベース化)
2. 配列に対しての検索
3. 配列から機能などを予測

ことを目的とする。生物配列で最初にデータベース化されたのは、タンパク質配列であった。1951 年に Sanger と Tuppy らによってタンパク質の配列解析法が開発された。それによって一般的によく知られたタンパク質ファミリー[1] の中から代表的なアミノ酸配列が決定した。1960 年代、NBRF(国立生物医学研究財団) の Dayhoff らはこれらの配列を最初にデータベース化し、Atlas of protein sequence and structure(タンパク質配列・構造の図説) としてまとめられた。やがてそれらの配列収集センターは、タンパク質情報リソース (RIP) と名称を変更した。NBRF は、1984 年以来このデータベースを保守管理しており、1988 年には NBRF、MIPS（ミュンヘンタンパク質センター)、日本の国際蛋白質情報データベースの 3 者の協力の下、RIP-国際蛋白質情報データベースが設立された。
Dayhoff らは、配列の類似性の程度に基づいてタンパク質のファミリーやスーパーファミリー[2] を分類した。そして、類縁関係の最も近いタンパク質間で比較を行い、観察された

[1] タンパク質ファミリーとは、進化上の共通祖先に由来すると推定されるタンパク質をまとめたグループである。

[2] ファミリーの定義は研究者により異なり、またファミリーの範囲も厳密に定義されるものではない。ファミリーより広い範囲をスーパーファミリー、より狭い範囲をサブファミリーとする分類も用いられるが、いずれも厳密に定義されるものではなく相対的な概念である。

配列変化の頻度表を作成した。タンパク質の違いが大きい場合、特定のアミノ酸が 2 度以上変異したかが問題となってくる。

90 年代に遺伝子やたんぱく質の配列の自動化・高速化であるハイスループットの開発が開発されて以来、生物データベースに追加されるデータ数は飛躍的に増加した。しかし遺伝子配列データが増加したとしても、それのみでは生物の組織・機能の理解は深まらない。実験によって得られた新しい配列と既知の配列との比較することは、新しい配列の特性を知る手段である。このとき、配列解析は比較された配列間の類似性の研究によって遺伝子とタンパク質の機能を調べるのに使われる。分子生物学とバイオインフォマティクスの配列解析は、特徴のある断片（例:DNA ストランド) は自動化され、計算機実験で結果が得られる。

関連するトピックとしては、遺伝子構造の配列での比較同定で、類似性と非類似性を発見するための配列を比較すること、遺伝的なマーカーを得るために突然変異や SNP を発見すること、組織の進化と遺伝的な分布の発見、遺伝子機能のアノーテーション、化学的には、複数のモノマーを形成するポリマーを決定するために使われる技術を含む。分子生物学と遺伝学において、同様のプロセスを単位 ”シークエンス ” と呼ぶ。

Methodology(手法) については、配列アライメントとは、複数の配列間を比較することで共通する部分を抽出することである。遺伝子には、同一生物種においても突然変異によって塩基に対して削除、置換が行われ必ずしも配列は一致しない。また、個人差の示す一塩基多型性 (SNP) によっても配列の一部は置換されている。また、異種の生物種においても同一部位（目や鼻など）の配列を比較することもある。これは、人間に対し生体を用いての実験ができないため、近い生物種をの同一部位の配列を比較することでその特性を調べることが目的である。配列アライメントには、2 本の配列を比較するペアワイズと複数の配列を比較するマルチプルアライメントがある。exact match にはならないため、動的計画法などを用いスコア行列を作成する。他の手法としては、隠れマルコフモデル、ビタビ、貪欲法などを用いた手法がある。ソフトウエアも多数制作されており、代表的なソフトでは、ClustalW, PROBCONS, MUSCLE, MAFFT, DIALIGN, T-Coffee, POA, MANGO などがある。

2.2 エピジェネティクス

生物学では、エピジェネティクスという用語は、ゲノムに書かれた遺伝情報を変更することなく、個体発生や細胞分化の過程において、遺伝子発現を制御する現象の総称である[18]。より狭義には「DNA 塩基配列の変化を伴わない子孫や娘細胞に伝達される遺伝子発現機構と機能」を対象とする分野である。このエピジェネティクスに関連するものとして、タンパク質因子や RNA 分子を含めた多彩な分子が関与している。そのために多彩な細胞活動をするためには、これらの分子群が適切な枠割を果たすことを必要であり、誤ったエピジェネティクスの情報は様々な疾病をもたらす。

2.2.1 歴史的な背景

1928 年イギリスの Griffith による形質転換現象の観察報告から遺伝子が DNA であることが分かり、その後 2004 年にはヒト、マウス、ラットなどの哺乳類のゲノムが解読されるようになった。一方、エピジェネティクスの分野では、1987 年に Hollidays が DNA メチル化の重要性を指摘し、注目を集めるようになった。しかし、そのエピジェネティクスを示唆する現象は 1962 年の Lyon による X 染色体不活性化現象の報告である。これは Watson と Crick の 2 重らせん構造の発見の 10 年前である。その後, 1984 年の前核移植実験により、母親と父親に由来するゲノムが機能的には等価ではなく、個体発生にはその双方が不可欠であることが示された。これは常染色体上にゲノム刷り込みを受ける遺伝子が存在することを示唆している。これに並行して、メチル化された CpG 配列に結合するたんぱく質や DNA メチル化酵素、あるいはヒストンの修飾に関するタンパク質など多彩なエピジェノタイプ（エピジェネテイクスな情報）の構築に関する分子群が同定される。これらの分子群は、発癌や遺伝性疾患など、また、体細胞クローンで注目を集める細胞核のリプログラミングにも関与している。そのため、DNA 脱メチル化酵素やヒストン脱メチル化酵素の同定やその分子の解明は、今後の研究課題である。　ゲノムインプリンティングや X 染色体不活性化現象は、メンデルの遺伝説の例外的現象である。

2.2.2 エピジェネティクスの分子生物学的な基礎

生物が正常に発生分化するためには、組織特異的にタイミング良く一定の量だけ必要な遺伝子が発現する必要がある。染色体ゲノムから遺伝情報の発現制御機構を理解すること

はエピジェネテイクスのメカニズムの解明に役立つ。

1. DNA のメチル化修飾は、哺乳類ゲノムを直接的に修飾する唯一の仕組みであり、メチル基を付加したり、外すことによって遺伝子の発現制御を行っている。現在までに、DNA メチル化修飾機構に関与する 5 つの遺伝子が明らかになっている。

2. クロマチンは、ヒストンタンパク質がコアとなるヌクレオソームから構成されており、遺伝子発現調整をするためには、基本転写因子群をはじめとする DNA 結合タンパク質との共同作業が必要である。メチル化 DNA 結合タンパク質や HMG タンパク質などの構造的クロマチン因子は、クロマチンの再構成を伴った遺伝子の転写活性制御にかかわる。

3. ヒストン自体もアセチル化やメチル化により修飾され、エピジェネテイクスな機構の大きな役割を担っている。

4. また、最近ヒストン修飾の変化を伴ったヘテロクロマチン化を誘導する RNAi が注目されている。ヒストンの中でも、N 末端を構成する立体構造に乏しい 20〜30 のアミノ酸残基は、ヒストンテールと呼ばれ、特にアセチル化やメチル化の標的となる。また、ヌクレオソーム間をつなぐリンカーヒストンのリン酸化も遺伝子発現制御にとって重要な因子である。さらに H2A、H2B、H3 についても細胞内でリン酸化を受け、細胞周期や DNA 修飾等 l クロマチンの様々な機能制御にかかわっている。

5. 卵子と精子に由来するクロマチンは、必ずしも同一の修飾を受けるわけではなく、ある一群の遺伝子座については、その親由来の DNA メチル化やヒストンのアセチル化、メチル化が異なっている。この現象をゲノムインプリンティングと呼ばれ、エピジェネテイクスの不均等性を与える。近年、X 染色体不活性化との類似性が指摘されこの不均等性なエピジェノタイプが正常な個体発生や細胞分化を考えるうえで重要である。

6. 5 の現象が破綻した場合、腫瘍や遺伝子疾患などに発症に関与し、ゲノム刷り込みを受ける遺伝子は、染色体上で近接して存在し、また、類似した発現パターンを示すことが多く、染色体機能ドメインを形成している。このような機能ドメインを規定するための境界配列がクロマチンインスれーたであり、インプリンテイングドメインばかりではなく、ゲノム全体に散在し、ダイナミックなクロマチン構造の構築に大切な役割を果たす。

7. クロマチンの構築に関連して、non-codingRNA があげられる。これは、X 染色体の不活性化のみならず、インプリンテイングドメインにおける制御センターとしての役割をもつことが知られている。さらに smallRNA が関与する RNAi 機構は、近年、強力な遺伝子解析法として脚光を集めているが、染色体ゲノム上で転移すｒとされるトランスポゾンの不活性化にも深く関与する。これまでジャンク DNA と呼ばれていたヒトゲノムの大半がこのような転写因子に由来することからもゲノムの多様性形成を考える上で興味深い。

8. エピジェネテイクスの基盤は、クロマチン構造に基づいた遺伝子発現制御にあるキネトコアやセントロメア領域中のヘテロクロマチン形成に強く関与するなど染色体動態にもエピジェネテイクスが関わっている。エピジェネテイクスはクロマチンや染色体という構造を制御するメカニズムである。

2.2.3　エピジェネテイクスの破綻による疾病

エピジェネテイクスは、正常な発生や分化にかかわる重要なメカニズムであり、その破綻により様々な発生・分化の異常が伴う。このようなエピジェネテイクスな修飾は、基本的には、体細胞に特異的であるが、ゲノムが次世代に伝わる時にはリセットされる。これを細胞核のリプログラミングという。

1. 発生や組織あるいは細胞のプログラムが進むにつれ、DNA メチル化などエピジェネテイクスな特性（エピジェノタイプ）もダイナミックに変化する。

2. また、エピジェネテイクスは生物の多様性や生物進化を考える上で、重要な情報である。

3. 生物種によって DNA メチル化機構が異なることからも生物進化を汁手掛かりが得られる。さらに多様性も個体間には認められる。この個体差は疾病の羅患率とも関係しているため、多様性を生み出すエピジェネテイクスなメカニズムを解明する過程で、疾病の予防や診断に役立つ。癌の治療という観点では、遺伝子変異を伴わないエピジェネテイクスな変化は可逆的であり、ある程度の可塑性が見出されることからエピジェネテックな変化の修復が期待される。

4. これまで、エピジェネテイクスな変異は、癌化の 2 次的、3 次的な現象であるとされてきたが、腫瘍の初期段階にも認められることから、部位特異的補正ができればエピジェネテイクスな側面からも治療法が可能とされる [7]。

5. さらにエピジェネテイクスの破綻もゲノムのアンバランスから誘発されていることを示唆する多くの知見が得られてきた。例えば、染色体異数体をもつ細胞においてはより多くの知見が得られてきた。例えば、染色体異数性をもつ細胞においては、より多くの遺伝子変化が蓄積する。

6. これからゲノム不安定性はさらなるエピジェノタイプの破綻を誘起し、癌などの疾病をもたらすと考えれている。

7. 精神疾患との関連も明らかになりつつある。様々な生命現象にとってエピジェネテイクスは不可欠であり、種間の相違や個体差を生む原動力となっている。そのため、環境の変化に伴う適応とも深く関わり、生物進化にも役割を持っている。

2.3 頻度による特徴ベクトルと位置特異的な特徴ベクトル

配列を用いた機械学習の特徴ベクトルでは、sliding-window がしばしば用いられる。DNA の 3 塩基が最終的にタンパク質に翻訳されるため、3-gram の window を用いて配列の頻度をカウントする。この方法は有効であるが、位置的な情報は消失している。ここでは多重配列のアライメントに用いられる PSSM 行列 (position specific scoring matrix, 位置特異スコア行列) について説明をする。位置特異スコア行列とは、類縁関係にある配列間のアライメントの特定の列に見出される変動の様子を数値化した行列である。

この行列の列は、元のアライメントの列に対応し、行は特定の文字（DNA では 4 種類の塩基、タンパク質では 20 種類のアミノ酸）に対応する。行列の要素は、対数オッズとして求めされる。これは、アライメントにある列に現れる特定の文字の出現数を配列全体の組成から予想される期待値で割り、その対数をとったものである。

2.3.1 位置特異的スコア行列

位置特異的スコア行列 (Position specific scoring marix) とは、モチーフの記述に用いられるスコア行列である。通常，各要素には，各位置での各塩基（アミノ酸）の出現に対す

る対数オッズ値をあてる。モチーフとは、複数本の相同なアミノ酸配列について多重アライメントを構築したときに，配列中に強く保存されている部分 (共通) 配列である。アライメントされているタンパク質が属しているタンパク質ファミリーに特有の機能や構造のために保存されている。モチーフは 1 つの配列中に複数個存在する場合もある。DNA では、転写因子結合部位など，ゲノム中に繰り返して現れる塩基配列パターンで、通常、周囲にあまり類似性が見られない。タンパク質では、局所的な共通アミノ酸配列パターンであり、なんらかの進化的要請から保存されている機能部位やシグナル部位である可能性が高い。氷山の一角のように，大きな共通構造の中で突出した部分で、繰り返し現れる立体構造パターンを指すこともある。モチーフの記述方法としては、正規表現、重み行列（プロファイル)、隠れマルコフモデル（HMM）がある。

2.3.2 位置特異的スコア行列の計算

位置特異的スコア行列は、モチーフの列ごとに頻度を計算し、図 2.1 では、10 本の遺伝子配列が並んでいる。第 1 列は、縦に AAAAAAGCTT と並んでいる。A は 6 本、G は 1 本、C は 1 本、T は 2 本なので、頻度はそれぞれ 0.6、0.1、0.1、0.2 となる。もともとの 4 つの塩基の配列の頻度を一様分布と仮定すると、それぞれ 0.25(図では、背景的頻度配列とよぶ) となる。頻度を背景的頻度配列で割り、自然対数をとると対数オッズが計算される。これを図 2.1 では、4 列分計算しスコア行列を作成する。これが位置特異行列である。対象とするモチーフ配列を TGAGCTAA とすると第 1 番目の塩基 T から始めてスコアの対数オッズの値を加算する。TGAG まで計算できる。次に第 2 番目の塩基 G から始めてスコアの対数オッズの値を加算する。GAGC まで計算できる。これを繰り返し計算すると位置ごとの対数オッズを計算できる。最後にこれを 2 の指数として計算し、オッズが計算される。一番オッズの高い 3 番目の塩基からが対応するモチーフの候補となる。

PSSM の中での要素は次のように計算される。

$$m_{i,j} = \log(\frac{p_{i,j}}{b_i})$$

$p_{i,j}$はモチーフの位置 j でのシンボル i の出現頻度.

b_iは、そのモデルの中のシンボル i の出現頻度.

配列

…	AAAA	…
…	AGAG	…
…	AGGC	…
…	AGCT	…
…	AGCT	…
…	AGCT	…
…	GGCT	…
…	CGCT	…
…	TCCT	…
…	TTTT	…

背景的出現頻度 0.25

	頻度(正規化)	背景的出現頻度	対数オッズ	頻度(正規化)	背景的出現頻度	対数オッズ	頻度(正規化)	背景的出現頻度	対数オッズ	頻度(正規化)	背景的出現頻度	対数オッズ
A	0.6	2.4	0.9	0.1	0.4	-0.9	0.2	0.8	-0.2	0.1	0.4	-0.9
G	0.1	0.4	-0.9	0.7	2.8	1.0	0.1	0.4	-0.9	0.1	0.4	-0.9
C	0.1	0.4	-0.9	0.1	0.4	-0.9	0.6	2.4	0.9	0.1	0.4	-0.9
T	0.2	0.8	-0.2	0.1	0.4	-0.9	0.1	0.4	-0.9	0.7	2.8	1.0

対象配列 TGAGCTAA

		合計	対数オッズをオッズに変換する
1 1番目のTから計算から対数オッズを加算	-0.2+1+(-0.2)+(-0.9)	-0.3	0.8
2 2番目のGから計算から対数オッズを加算	-0.9+(-0.9)+(-0.9)+(-0.9)	-3.6	0.1
3 3番目のAから計算から対数オッズを加算	0.9+1.0+0.9+1.0	3.8	13.9
4 4番目のGから計算から対数オッズを加算	-0.9+(-0.9)+(-0.9)+(-0.9)	-0.3	0.8
5 5番目のCから計算から対数オッズを加算	-0.9+(-0.9)+(-0.2)+(-0.9)	-2.9	0.1

図 2.1: PSSM の計算

2.3.3 位置特異行列の情報量

ある PSSM が実際の配列パターンを背景から識別するのに、どの程度有効かは測定できる。測定の単位は bit である。モチーフのそれぞれの座位に相当する配列を標的の配列上から同定するために、対数オッズスコアを算出した。この表の各列に見出されるスコアの変動は、このモチーフを作成するために用いた元の訓練配列の多様性の指標である。ある列には 1 種類の塩基しかないかもしれないし、あるれるには複数の塩基が存在しているかもしれない。強く保存された列は、変動の大きい列よりも多くの情報をもち、標的配列中の合致部位を探すのにより決定的に働く。PSSM 行列を評価として情報量 (エントロピー) が使われる。

第 3 章

機械学習アルゴリズム

本章では、機械学習の分類の説明を行い、次に RandomForest で用いられるブートストラップ、RandomForest の説明を行う。樹木に基づく方法（樹木構造接近法）は、データに潜む非線形効果や交互作用構造を何らかの樹木形式に変換して理解する方法である [8]。その他 support vector machine も用いているが、これは成書がかなり出ているため省略する [10]。

3.1　機械学習の分類

機械学習の分類には数種類あり代表的な分類を以下で説明する。

3.1.1　教師あり学習、教師なし学習

観測データとそれを分類するクラスとの関係により、教師付き学習、教師なし学習、半教師付き学習に分類される。教師あり学習 (supervised learning) では、観測データと、そのデータを分類するためのクラス（ラベルともいう）が与えられている。例えば、文書分類問題であれば、観測された文書とその文書の属するカテゴリー（スポーツ、芸能、など）の対のデータ集合（これを training data と呼ぶ。) である。学習によって、観測データの持つ属性と意味の関係を推定し、未知のデータ (これを test data　と呼ぶ。）が与えられると、そのデータの意味を出力する。教師なし学習 (un-supervised learning) では、観測データだけが与えられる。観測データたちの間の距離をその属性から計算し類似するデータを１つのグループにまとめる。教師あり学習で使う training data は人手で作ることが多いので、作成コストが大きい。一方、教師なし学習は、類似したデータがまとまるだけで学習

結果の意味づけが難しい。そこで、少数の training data から学習を開始し、学習の過程で training data を拡大していく半教師あり学習 (semi-supervised learning) も有力である。

3.1.2 生成モデルと識別モデル

クラス分類の問題において手法の分類として、

1. 識別モデル (Discriminative model)
2. 生成モデル (generative model)
3. 識別関数

という分類がある [11][12][36][37][38]。これらのモデルは、それまで統一的には論じられていなかったベイズ的手法から SVM を代表とする識別手法までの手法の関連を説明する枠組みとして提案されている。また、両者を取り入れた hybrid モデルも提案されている。

入力ベクトルを x とする。c をラベルとする。ここで訓練データとしては、N 個のデータ $X = \{x_1, ..., x_N\}$ とする。またクラスラベルとして $C = \{c_1, ..., c_N\}$ とする。クラス分類を目的とする。目的は、新しい入力ベクトル $\hat{x}$ に対してクラス $\hat{c}$ を予測することである。

$$\hat{c} = argmax_c(\hat{x}, X, C)$$

パラメータ θ の集合によって支配されるパラメトリックモデルでの確率分布を、ベイズ的な設定の下で決定するためには、一般に

$$p(c|\hat{x}, X, C) = \int p(c|\hat{x}, \theta)p(\theta|X, C)d\theta$$

を計算する。$p(c|\hat{x}, \theta)$ は、モデルの違い（生成モデルかまたは識別モデル）を表し、$p(\theta|X, C)$ は訓練/テストの違いを示している。

生成モデル

生成モデルは、システムの利用可能な状態を統合するために、システムの全ての変数全体の相互作用を捉えるように構築される。これは、入力、隠れ変数、出力 z を結合してモデリングし、確率分布 p を設計することで達成される。$p(z|\ \theta\ )$ で表され、θ はモデルのパ

ラメータである。z は、異なった変数の組合せである。結合確率分布をより単純化するために、条件付き独立という条件が p を分解するため付けられる。また、不要な変数を避けるために、パラメータ θ 上の事前分布を定義することができる。モデリングの為に、生成モデルの場合、通常事前知識を入れるかどうか選択できる。

分類問題では、生成モデルは、入力はデータ x で、出力はクラス c である。確率論的な表記では、$p(x, c|\ \theta)$ として定義される。画像認識で、猫と犬を判別する問題があるとすると、生成モデルでは、「なにが猫を猫と認識させるのか？」「なぜ犬を犬として認識するのか？」ということが問われる。それは、ラベルが結合確率分布でモデリングされているため、生成モデルは、$p(c|z, \theta)$ を計算することで分類することができる。生成モデルの種類としては、ナイーズベイズモデル、Markov random fields などがある。

機械学習の問題は最適化問題で定式化される。大半の機械学習の問題は目的関数を最適化することで表せる。生成モデルでは、生成学習を使って訓練データを学習する。生成学習では、訓練データ全ての結合した尤度関数を最適化できる。$p(X, C|\ \theta)$ と表記する。結合尤度関数は、

$$L_G(\theta) = p(X, C, \theta) = p(x_\theta, c_n, \theta) = p(\theta) \prod_{n=1}^{N} p(x_n, c_n|\theta_{c_n})$$

で表される。

識別モデル

識別モデルは、入力の分布を計算せずに、システムの異なった出力の境界をを捉えるように構築される。これは、入力データ x で条件づけられたクラスラベル c の上での確率分布 p を設計することで得られる。これは、$p(c|x, \theta)$ で表記される。θ はモデルのパラメータである。注意としてこれは、確率分布ではない場合がある。その場合、関数 $f_\theta(x)$ が設計される。これはクラスラベルの c の 1 つが出力される場合である。

$p(x, c|\ \theta)$ と $p(c|x, \theta)$ の違いは本質的である。分類問題では、入力データは x で、クラスラベルは c である。そのため、入力データの分布を考慮するかわりに、現在のモデルのクラス間の境界の形を近似することを目的とする。猫の分類問題では、「猫と犬のどちらか？」が識別モデルでは問われている。代表的な識別モデルとしては、ガウス過程、SVM、ニューラルネットワークなどがあげられる。

識別モデルの学習は、識別学習をつかって訓練データを学習する。これは生成学習とは根本的に異なっている。訓練データ x, c は手動でラベル付けされる。パラメータ θ を最大

化する関数は次にように書かれる。

$$L_D(\theta) = p(C|X,\theta) = p(c_n, x_\theta, \theta) = p(\theta)\prod_{n=1}^{N} p(c_n|x_n, \theta_{c_n})$$

生成モデルと識別モデルのちがい

生成モデルと識別モデルの違いの１つは、生成モデルがそれぞれの分類から独立に計算できる点である。モデルと分類の１対１写像は、分類を付け加える際、容易に付け加えることができる。また、それは、全ての分類に対して異なったモデルをもつことをも容易にする。反対に識別モデルは境界部分に関心がもたれるために、全てのモデルは結合していることが必要とされる。そのため、新しい分類を付加する場合、また最初からやり直さなければならない。

しかし、生成モデルの場合、重要な特徴はモデリング力である。生成モデルでは、システム環境について専門家の考えを吸収して設計することができる。例えば、変数がどのように相関するかという事にに関する事前知識、どちらの変数が関連しないかという事に関する事前知識、パラメータの値の範囲に関する事前知識などである。識別モデルは分類指向であり、そのため柔軟性に欠ける。これはブラックボックスになる傾向をもつ。データは入力として与えられ、$P(\text{分類}\,|\,\text{入力})$ として返ってくるが、その理由と方法に関する理解は明確ではない。

他に生成モデルとの違いは、生成モデルは、モデリング力があるため欠損値を処理する能力がある。しかし、識別モデルでは、入力データの分布がないために欠損値の修復が容易でない場合が多い。この違いは大きく、なぜなら生成モデルが異なった種類のデータ、例えばラベル付けされたデータやラベル付けされていないデータなど、を容認するからである。生成モデルでは、ラベル付けされていないデータも上記と同様の考えで処理できる。

反対に、識別モデルでは、結合確率分布のモデルをすべて活用する。その代わりに、クラス間の境界に注目する。実際のところ、結合確率分布は、事後確率の効果がすこししかないような構造を多くもつ。そのため、結合確率分布の計算を要求しない。これが識別モデルが普及している理由である。他の識別モデルの特徴は、スピードである。実際に、新しいデータを分類することは早い、なぜなら $p(c|x, \theta)$ を直接計算するだけであるからである。

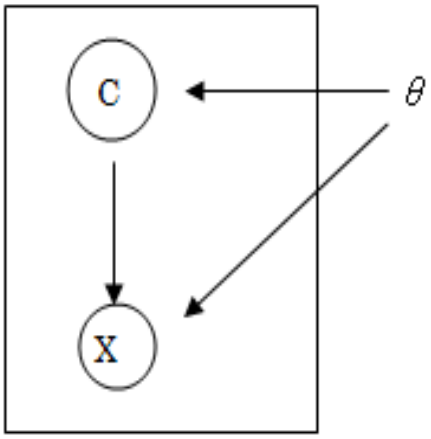


図 3.1: 生成モデル

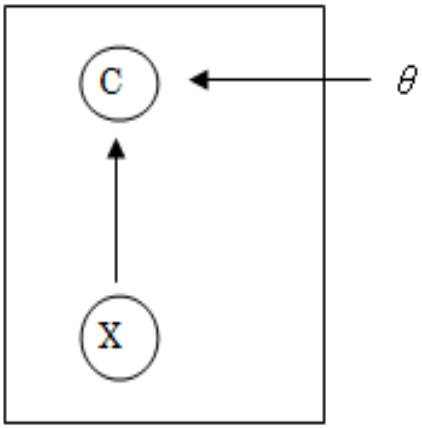


図 3.2: 識別モデル

3.2　ブートストラップ

ブートストラップ法は、複雑な理論や数式に基づく解析を、計算機を用いた大量の反復計算で置き換えて実行する統計的計算法である [46]、[47]、[48]。その特徴は、ブートストラップ法の実行プロセスの中で、解析的表現を計算機を用いた大量の反復計算によるモンテカルロ法で置き換えたところによる。これによって、緩やかな仮定のもとで、複雑な推測論の問題に適用できる柔軟な統計手法となった。

より詳しく定義するとブートストラップ法は、１つの標本から復元抽出を繰り返して大量の標本を生成し、それらの標本から推定値 $\hat{\theta}^*$ を計算し、母集団の性質やモデルの推測の誤差などを分析する方法である。ブートストラップ法では母数 θ の推定量は、標本から生成したブートストラップ標本の推定量 $\hat{\theta}^*$ を用いて推定する。１つの標本からリサンプリングを繰り返して生成される標本をブートストラップ標本と呼ぶ。

ブートストラップ標本の生成には幾つかの方法が提案されているが、確率分布型を仮定するパラメトリック・ブートストラップ法と確率分布型を仮定しないノンパラメトリック・ブートストラップ法に大別される。そのアルゴリズムの例を次に示す。

1) パラメトリック・ブートストラップ法

① 標本サイズが n である標本データ $x_1, x_2, \cdots, x_i, \cdots, x_n$ の平均 $\bar{x}$ 、標準偏差 s を計算する。

② n 個の正規乱数 z_1, z_2, ⋯ , z_i, ⋯ ,z_n を生成し、$x_i^* = -x + z_i * s$ で新しい標本 $x_1^*, x_2^*, \cdots, x_i^*, \cdots, x_n^*$ を生成する。この標本による推定値を $\hat{\theta}_1^*$ 例えば、平均 $\bar{x}_1$) とする。

2) ノンパラメトリック・ブートストラップ法

① 区間（0, 1）を n 等分した各区間の値を標本データ $x_1, x_2, \cdots, x_i, \cdots, x_n$ に１対１で対応させる。

② n 個の一様乱数 u_1,u_2, ⋯ ,u_i, ⋯ ,u_n を生成し、u_i の値が含まれる区間に対応する x_k を x_i^* とし、新しい標本データ $x_1^*, x_2^*, \cdots, x_i^*, \cdots, x_n^*$ を生成する。この標本から得られた推定値を $\bar{\theta}_1^*$ とする。

両方法ともステップ②を B 回繰り返し、B 個の標本の推定値 $\bar{\theta}_1^*, \bar{\theta}_2^*, \cdots, \bar{\theta}_3^*, \cdots, \bar{\theta}_B^*$ を求める。その推定値、標準偏差、バイアスはそれぞれ次の式で求める。

$$\bar{\hat{\theta}} = \frac{1}{B} * \sum_{i=1}^{B} \bar{\theta}_i^*$$

$$s(\bar{X}) = \sqrt{(\frac{1}{B-1} * \sum_{i=1}^{B} (\hat{\theta}_i^* - \bar{\hat{\theta}})^2}$$

$$bias(\bar{X}) = \frac{1}{B} * \sum_{i=1}^{B} (\hat{\theta}_i^* - \bar{\hat{\theta}})$$

また、確率分布関数は

$$Pr(\bar{X} \leq x) = \frac{1}{B}\{x \geq \bar{x}_i \text{の個数}\}$$

により推定できる。B 個の推測値を大小順に並べた B × α 番目の値を 100 α %点とする。繰り返しの回数 B については、推定値の標準誤差を求める場合は 100〜200 回、確率分布関数の推定値や 100 α %点を求める場合は 1,000〜2,000 回が必要であるとされている。

3.3 RandomForest

RandomForest は、bootstrap というリサンプリングという方法でサブデータを作成し、各サブデータセットで決定木を構築する [5]。RandomForest は、樹木モデルを用いているが、集団学習（アンサンブル学習）の 1 種である。決定木は、高精度の分類器ではないが、計算の速さやその結果の可読性に優れている。集団学習は、精度は高くはない分類器を複数組み合わせることで、精度を向上させることを提案している [9]。

3.3.1 アルゴリズム

RandomForest は、Bagging の提案者である Brieman が提案した。アルゴリズムは

1. 与えられたデータセットから n 組みの bootstrap サンプルを作る。

2. 各々の bootstrap データを用いて未剪定の最大の決定木・回帰木を作成する。

3. 全ての結果を統合組み合わせ (回帰問題では平均、分類問題では多数決)、新しい予測・分類器を構築する。

Bagging と RandomForest の相違点は、Bagging は全ての変数を用いるが、RandomForest は変数をランダムサンプリングした

サブセットを用いることができるので高次元のデータの計算に適している。

以下、RandomForest の長所である。

・精度が高い。

・規模の大きいデータに対応。

・分類に用いる変数の重要度を計算する。

・欠損値の推測および多数の欠損値をもつデータに対しても正確さと維持している。

・分類問題における各群の個体数がアンバランスであるデータにおいてもエラーのバランスが保たれる。

などがあげられる。

Algorithm 1 RandomForest による分類・回帰

1: **for** b=1 to B: **do**
2: (a) 訓練データからサイズ N の bootstrap サンプル Z をとる.
3: (b)bootstrap されたデータに randomForest の木 T_B を構築する.
4: 最小のノードサイズ n_{min} に到達するまで、木のそれぞれの終端ノードに対して
5: 次のステップを繰り返す.
6: i) 変数 p からランダムに変数 m を選択する.
7: ii)m の間で最良の変数の分割点を取り出す.
8: iii)2 つの娘のノードにノードを分割する.
9: **end for**
10: アンサンブルの木 $\{T_b\}$ を出力.
11: 新しいテストデータである点 x で予測するために:
12: ・回帰問題:$\hat{f}(x)_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
13: ・判別問題:$\hat{C}(x)_b$ は、b 番目の randomForest の木のクラスの予測をする.
14: 次に $\hat{C}(x)_{rf}^{B} = \hat{C}(x)_{b1}{}^{B}$ の多数決を行う.

RandomForest では、out of bag(OOB) を使う。データセットの中からランダムに一部をテスト用として取り出し、その残りを学習用とする方法もある。取り出したデータを OOB（out-of-bag）データと呼ぶ。OOB 以外の訓練データで学習を行い、OOB でテストする。最終的に複数の結果から最も高い結果を多数決によって選択する。学習とテストを繰り返す回数を多くすることで、信憑性が高い結果を得ることが可能である。

3.3.2 variable importance

RandomForest では、決定木の分割指標として、Gini index がよく用いられる。Gini index は、イタリアの経済学者 Gini が、1936 年に考案した指数で、経済学の分野では、所得格差を表すのことに使われる。多様性指数としては、集団内で復元的にランダム選択された、

任意の 2 つの要素が異なるクラスに属する確率を意味する。たとえば、データが a,b のどちらかのクラスに属する場合、データをランダムに選択して、ab(a に属する選択のあとに b に属する選択になる) あるいは ba になる確率が、そのデータの多様性を表していると考える。a になる確率を p_a、B になる確率を p_b とすると Gini index は、

1-(p_a^2+p_b^2)=1-(p_a^2+(1-p_a)^2) = 2*p$_a$(1-p_a)

また、別の定義として（第 4 章の属性選択章でも定義するが）Gini インデックスなどを指標として使う場合である。

データ集合 D とし、ランダムに選択したデータのクラスを同定するのに必要な平均情報量は、データがクラス C_k(k=,1,2,...,K) に属する確率を P(C_k) とすると、次の式で表される。

$$Info(D) = -\sum_{k=1}^{K} P(C_k) \log_2 P(C_k)$$

情報利得は、属性 A を用いた分割による情報量の差で、

$$Gain(A) = -\sum_{k=1}^{K} P(C_k) \log_2 P(C_k) + \sum_{j=1}^{J} \beta_j \sum_{k=1}^{K} P(C_{jk}) \log_2 P(C_{jk}))$$

で定義される。ここで、β_j は次式で定義される。

$$\beta = \frac{N_j}{N} (j = 1, ..., J)$$

$$\beta_j \geq 0$$

$$\sum_{j=1}^{J} \beta_j = 1$$

ただし、J は分割数、N は分割数のデータ数、N_j は分割 j のデータ数、P(C_{jk}) は、分割 j 内のデータがクラス C_k に属する確率である。情報利得は、分割数 J の大きな属性を選ぶ傾向があるので、属性 A の値を同定するのに必要な情報量（属性値の個数 J が大きいと大きい値をとる。）

$$Info(A) = -\sum_{j=1}^{J} \beta_j \log_2 \beta_j$$

で情報利得を割ったものが情報利得比である。

データ集合 D からランダムに選択したデータのクラスが誤分類される確率を Gini 関数とよび、次式で表される。

$$Gini(P(D)) = \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} P(C_i)P(C_j) = 1 - \sum_{k=1}^{K} P(C_k)^2$$

Giniインデックスは属性Aを用いた分割によるGini関数の差で、

$$Gini - Index(A) = 1 - \sum_{k=1}^{K} P(C_k)^2 - \sum_{j=1}^{J} \beta_j (1 - \sum_{k=1}^{K} P(C_{jk})^2)$$

と定義される。

寄与度の計算は決定木を構築する際、該当変数をモデルから除いた際の、予測精度の低下 (Mean Decrease Accuracy)、あるいはGini indexの減少 (Mean Decrease Gini) に基づいている。つまり対象となる属性から1つ取り除き、そのGini index(またはAccuracy)の低下の大きいものほど寄与の高い属性とする。

第 4 章

属性処理について

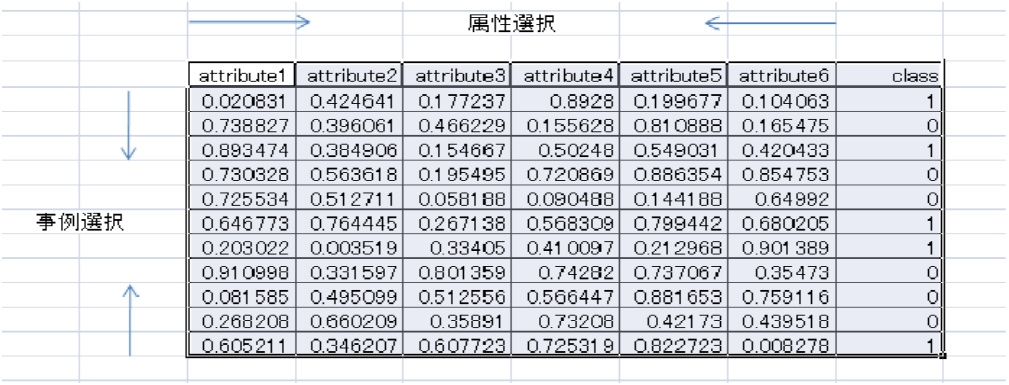
機械学習の問題において大量のデータを対象とする場合、学習アルゴリズムの高速化以外に、データそのものに対する処理の研究が行われている [3][4][25][22]。元来、データにはデータ解析の目的のために収集されているのではなく、ノイズや冗長、または誤ったデータが混入している場合も多い。そのために必要最小限の部分集合を取り出し、データ量を削減することが重要である。属性処理には、次のような処理がある。

1. 属性選択 (feaure selection) : 与えられた属性から有効なものを選択する。

2. 事例選択 (data reduction) : 与えられたデータから有効なデータを選択し、データ量を減らす。

3. 属性抽出 (feature extraction) : 与えられた属性から新しい属性を抽出する。新しく抽出された属性数は、元の属性数より少ない。

4. 属性構築 (feature construction) : 元の属性では学習アルゴリズムが作動しない場合に、新しい属性を構築すること。元の属性に必要な情報が含まれているという前提で、より望ましい属性を元の属性から機能的に構築することである。



attribute1	attribute2	attribute3	attribute4	attribute5	attribute6	class
0.020831	0.424641	0.177237	0.8928	0.199677	0.104063	1
0.738827	0.396061	0.466229	0.155628	0.810888	0.165475	0
0.893474	0.384906	0.154667	0.50248	0.549031	0.420433	1
0.730328	0.563618	0.195495	0.720869	0.886354	0.854753	0
0.725534	0.512711	0.058188	0.090488	0.144188	0.64992	0
0.646773	0.764445	0.267138	0.568309	0.799442	0.680205	1
0.203022	0.003519	0.33405	0.410097	0.212968	0.901389	1
0.910998	0.331597	0.801359	0.74282	0.737067	0.35473	0
0.081585	0.495099	0.512556	0.566447	0.881653	0.759116	0
0.268208	0.660209	0.35891	0.73208	0.42173	0.439518	0
0.605211	0.346207	0.607723	0.725319	0.822723	0.008278	1

図 4.1: 属性処理

4.1 属性選択

属性選択は、与えられた属性から目的に有効な属性を選択することである。図 3.1 では横方向に列を削除する。属性数が n のとき、属性パターン（部分集合) の数は、2 の冪乗 2^n-1 となる。n が大きくなる時、膨大な数となり、効率よく削減することが重要である。属性選択の手法は大別して、Filter(フィルタ法)、Wrapper(ラッパ法)、Embeded(埋め込み) 法がある。フィルタ法は、属性の選択に適当な指標を用いて、それを基準に属性を選択する。ラッパ法は、学習結果を用いる。フィルタ法は学習モデルを知らなくてもよいため処理時間が短い。ラッパ法は、学習したモデルの結果そのものを評価指標としてもちいるので選択の精度はよいが、学習アルゴリズムを内臓するので処理時間の点から実用的ではない。Embeded(埋め込み) 法は、学習アルゴリズムの中に属性選択が含まれている手法である。以下属性選択を探索法（方向・戦略）と評価基準の観点から分類する。

表 4.1: 属性選択一覧 [13]

モデルの探索	利点	欠点	先行研究
Filter	1 変数		
	処理が早い	属性の独立性は無視される	ユークリッド距離
	scalable	判別器間の相関は無視される	χ^2
	判別器は独立		i-test
			Information gain,
			Gain ratio (Ben-Bassat, 1982)
	多変数		
	モデルの属性依存	1 変数の技術より遅い	Correlation-based feature
	判別器は独立	1 変数の技術より Scalable ではない	selection (CFS) (Hall, 1999)
	ラッパ法より計算量はすくない	判別器間の相関は無視される	Markov blanket filter (MBF)
			(Koller and Sahami, 1996)
			Fast correlation-based
			feature selection (FCBF)
			(Yu and Liu, 2004)
Wrapper	決定論的		
	単純	over fiting のリスクあり	Sequential forward selection
	判別器間で相関あり	局所最適なスタックになるために	(SFS) (Kittler, 1978)
	モデルの属性に依存		Sequential backward elimination
	ランダマイズの手法より	ランダマイズアルゴリズムになる傾向がある	(SBE) (Kittler, 1978)
	計算量的が少ない	(greedy search)	Plus q take-away r
			(Ferri et al., 1994)
			Beam search (Siedelecky
			and Sklansky, 1988)
	ランダマイズ		
	局所最適にないにくい	計算量的に高価	Simulated annealing
	判別器間で相関あり	判別器は選択に依存	Randomized hill climbing
	モデルの属性に依存	決定論的手法より計算量が大きい	(Skalak, 1994)
			Genetic algorithms
			((Holland, 1975))
			Estimation of distribution
			algorithms (Inza et al., 2000)
Embedded	判別器間のインタラクション		
	判別器間で相関あり	判別器は選択に依存	Decision trees
	ラッパ法より計算量は少ない		Weighted naive Bayes
	モデルの属性に依存		(Duda et al., 2001)
			Feature selection using
			the weight vector of SVM
			(Guyon et al., 2002;
			Weston et al., 2003)

4.1.1 探索法

探索の方向

探索の方向には 4 種類ある.

1. 前向き探索 ··· 空集合 (なにも選択されていない状態) から一番効果的な属性を順次加えていく方向.
2. 後向き探索 ··· 全属性から出発して一番効果的ではない属性を順次削除していく方向.
3. 両方向探索 ··· 両方方向から探索し、先に見つかった属性の部分集合を解とする.
4. ランダム探索 ··· 属性が非常に多い場合、計算資源が許す限りランダムに選択し、その時点までの最良の結果の部分集合を解とする.

Algorithm 2 前向き探索 (Sequential forward feature set generation-SFG

```
SFG
Input:
    set S to hold the selected features
    set F having the original features
    evalution measure M,
Initialize:S={}
Repeat
    (1) f=FindNxt(F) (一番評価の高い属性を選択する関数)
    (2) S=S ∪ {f}
    (3) F = f − {f}
Until S satisfies M or F = {}
Output:S
```

探索の戦略

属性選択の探索空間のサイズは 2^N である。また、属性値部分集合は、束 (半順序集合 (L, ≦) であって、L のどの二元 x, y に対しても L の部分集合 {x, y} の順序 ≦に関する下限 inf{x, y} と上限 sup{x, y} が存在するもののことである) となる。探索の戦略とはこの探

Algorithm 3 後向き探索 (Sequential backward feature set generation-SBG

SBG
Input:
    set S to hold the selected features
    set F having the original features
evalution measure M,
Initialize:S={}
Repeat
    (1) f=GetNxt(F) (一番評価の低い属性を選択する関数)
    (2) S=S ∪ {f}
    (3) $F = f - \{f\}$
Until F does not satisfies M or $F = \{\}$
Output:S

Algorithm 4 ランダム探索 (Random search-RAND)

RAND
Input:F-full set,M-measure,
    S=S_{best}={ } (S は部分集合)
    C_{best}=#{F} (C:集合の要素数)
Repeat
Initialize:S={}
Repeat
S=RandGen(F)
    C=#(S)
If C $\leq C_{best} \wedge$ S satisfies M
    S_{best}=S
    C_{best}=C
    print S_{best}
endif
Until some stopping criterion is satisfied
Output:S_{best} (今までのベスト)

索空間をどのように探索するかという戦略である。探索の方向性とは別の概念である。分類すると

1. 完全探索
   (a) 全探索
      i. 深さ優先探索 ••• 深さ優先探索と幅優先探索は scan する方向が違う。利点は、メモリ消費量が少ない。欠点は束が深くなると非効率.
      ii. 幅優先探索 ••• 利点は解があれば必ず探索できること. 欠点はメモリ消費量が大きい.
      iii. 反復深化優先探索 ••• 深さ優先探索に幅優先探索の利点を加味したもので、深さの上限を制限しながら深さ探索をする.
   (b) 部分探索　分枝限定法が代表的. 評価指標が閾値を越えた時その先を探索しない. 評価手法が属性の部分集合の包含関係に関し単調性を有す場合、この手法は完全探索となる.
2. ヒューリステック探索
   (a) 最良優先探索 ••• 未展開のノードのうち評価指標の一番良いものを展開する.
   (b) ビーム探索 ••• 未展開のノードのうち評価指標の高い上位指定個数だけを残して、評価値の一番良いものを展開する.
   (c) 欲張り探索 ••• 過去のものを捨て、現時点で一番良いものを探索する.
3. 非決定的探索

本研究で用いた機械学習のソフト Weka では以下の探索戦略が実装されている.

1. BestfirstSearch
2. ExhaustiveSearch
3. FCBSearch
4. GeneticSearch
5. GreedyStepwiseSearch

Algorithm 5 深さ優先探索 (Exhaustive search:depth first-DEP)

```
1: Input:F-full set,S-stack,M-measure,
2: Initialize:
3: node=null
4: S=null
5: DEP(node)
6: If node is the best subset so far w.r.t M
7:     Set=node
8: end if
9: For all children C of node
10:    push(C,S) (行列の先頭に追加)
11: end for
12: While(notEmpty(S))
13:    node=pop(S) (行列の先頭から取り出す)
14:    DEP(node)
15: end while
16: Output:Set
```

Algorithm 6 幅優先探索 (Exhaustive search:breadth first-BRD)

Input:F-full set,Q-queue,M-measure,
Initialize:node=null,Q=null
    BRD(node)
For all children C of node
enqueue(C,Q)
end for
While(notEmpty(Q)) (行列の最後に追加)
    node=deqeue(Q) (行列の先頭から取り出す)
If node is the best subset so far w.r.t M
    Set=node
end if
    BRD(node)
end while
Output:Set

Algorithm 7 分枝限定法 (Complete search:branch & bound-BAB)

Input:F-full set,Q-queue,M-measure, (M は小さい方がよい)
$\beta - bound\ for\ some\ value\ of\ M Initialize : node = F, best = \beta$
BAB(node)
For all children C of node
IF(C's value of M $\leq \beta$) (M がβ未満の時だけ以下を実行) 条件未達の時は枝刈り
    IF(C's value of M $\leq$ best's value)
        best=C
end if
BAB(C)
end if
end for
Output:best

6. LinerFoardSelection
7. RaceSearch
8. Randomsearch
9. Ranker
10. RankSearch
11. ScatterSearchV1
12. SubsetSizeForwardSelection

4.1.2 評価基準

評価基準として代表的なものをあげる.

また、属性の良さをここで定義する。

属性の良さ

属性の部分集合 S_i、S_j
評価指標 M
要素の数 (属性数) #
とする。
M(S_i)=M(S_j), #(S_i)=#(S_j)　ならば S_i と S_j は同じ。
M(S_i)=M(S_j), #(S_i)¡#(S_j) または
M(S_i)¡M(S_j), #(S_i)¡#(S_j) の時、S_i が S_j より良い。

と定義する。

予測精度

予測制度を評価指標とする場合は、決定木など学習モデルが必要である。学習モデルのでの予制精度で評価し探索（属性の増加または減少）を行う。

情報利得

情報利得（情報利得比)、Gini インデックスなどを指標として使う場合である。

データ集合 D とし、ランダムに選択したデータのクラスを同定するのに必要な平均情報量は、データがクラス C_k(k=,1,2,...,K) に属する確率を P(C_k) とすると、次の式で表される。

$$Info(D) = -\sum_{k=1}^{K} P(C_k) \log_2 P(C_k)$$

情報利得は、属性 A を用いた分割による情報量の差で、

$$Gain(A) = -\sum_{k=1}^{K} P(C_k) \log_2 P(C_k) + \sum_{j=1}^{J} \beta_j \sum_{k=1}^{K} P(C_{jk}) \log_2 P(C_{jk}))$$

で定義される。ここで、β_j は次式で定義される。

$$\beta = \frac{N_j}{N} (j = 1, ..., J)$$

$$\beta_j \geq 0$$

$$\sum_{j=1}^{J} \beta_j = 1$$

ただし、J は分割数、N は分割数のデータ数、N_j は分割 j のデータ数、P(C_{jk}) は、分割 j 内のデータがクラス C_k に属する確率である。情報利得は、分割数 J の大きな属性を選ぶ傾向があるので、属性 A の値を同定するのに必要な情報量（属性値の個数 J が大きいと大きい値をとる。）

$$Info(A) = -\sum_{j=1}^{J} \beta_j \log_2 \beta_j$$

で情報利得を割ったものが情報利得比である。

データ集合 D からランダムに選択したデータのクラスが誤分類される確率を Gini 関数とよび、

$$Gini(P(D)) = \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} P(C_i) P(C_j) = 1 - \sum_{k=1}^{K} P(C_k)^2$$

Gini インデックスは属性 A を用いた分割による Gini 関数の差で、

$$Gini - Index(A) = 1 - \sum_{k=1}^{K} P(C_k)^2 - \sum_{j=1}^{J} \beta_j (1 - \sum_{k=1}^{K} P(C_{jk})^2)$$

と定義される。情報利得（情報利得比）も Gini インデックスも類似の挙動を示す。

距離尺度

距離尺度は 2 つ挙げる。いずれも確率分布から計算される。

1. Directed Divergence

$$DD(x_j) = \int [\sum P(c_i|x_j = x') \log \frac{P(c_i|x_j = x')}{P(C_i)}] P(x_j = x') dx'$$

2. Variance

$$V(X_j) = \int [\sum P(c_i)(P(c_i|X_j = x) - P(c_i)^2] P(X_j = x) dx$$

依存尺度

$$B(x_j) = \sum -\log[P(c_i)] \int \sqrt{P(x_j = x'|c_i) P(x_j = x' dx'}]$$

不整合度

不整合度に基づく指標について説明する。クラスが違うが、属性の値が等しいデータ同士は矛盾している。不整合度とは、この矛盾の程度を定量的に評価したものである。

まず不整合度の定義は、

不整合度＝（属性の値が同じデータ数) − (その中でクラスが同じものの最大数)

例：属性の値が同じデータが n 個あり、その中で n_1 個がクラス c_1、n_2 がクラス c_2、n_3 がクラス c_3 (n_1+n_2+n_3=n) とする。n_3 が最大なら不整合度は (n-n_3) となる。これを用いて属性集合 S_i に対するデータの不整合度 M は、

$$\text{不整合度 } M = \frac{\text{不整合度の合計}}{\text{全データ数}}$$

で定義する。

不整合度の性質

属性集合の包含関係に関して、単調性が保持されることである。

$$S_i \subset S_j \text{ならば } M(S_i) \geq M(S_j)$$

(証明)

$S_i \subset S_j$ なので、S_i の分類能力は S_j より大きくなりえない。分類能力と不整合度は、逆の関係 $\mathrm{U}(S_i) \geq \mathrm{U}(S_j)$ がある。S_k(=S_j-S_i) と置くと、S_k は、次の 3 つの場合しかない。

1.S_k は無関係 (irrelevant)

「無関係」の定義により、余分な属性は S_j の不整合度に影響をあたえないので

$$M(S_j) = M(Si)$$

2.S_k は冗長 (redundant)

「冗長」の定義により、余分な属性は S_j の不整合度に影響を与えないので

$$M(S_j) = M(S_i)$$

3.S_k は関連あり (relevant)

S_i には、S_j より関連する属性が不足している。したがって、

$$S_i < S_j \text{なら} U(S_i) \geq U(S_j).$$

属性アルゴリズム

ここでは代表的なアルゴリズムを挙げる。

・Focus ⋯ 空集合の属性集合から属性を 1 つづつ追加する前向き探索によって評価指標に不整合度を採用し、整合性を保持できる範囲で最小の属性集合を求めるものである。簡単な方法であるが、連続数値やノイズが扱えない。

・Relief ⋯ ある事例とそのニアミス (属性パターン間の距離が最小なクラスが違う事例) を区別する属性の方が、その逆の、その事例とニアヒット (属性パターンの距離が最小の方がヒューリステックを用いている。Relief はノイズに強く、混在属性 (連続数値、離散数値、名義) にも適用可能であるが、冗長性に弱く、クラスはバイナリーに限定されている。その後、距離が最小のもの一つを選択するのでなく、最小のものから k 個選択し平均をとる。各クラスの事前分布でクラスごとに重みを付けするなどの改良が加えられて、現在ノイズもさらに頑強で、多クラスにも適用可能である。

・ABB ⋯ 分枝限定法に評価尺度として単調性を有する不整合度を導入して、全属性集合に対する不整合度。不要な探索をさらに減らす為に、探索の戦略には幅優先探索を採用している。Focus と同じく、連続数値が扱えない。同じ不整合度を評価指標としている

Algorithm 8 Focus Algorithm

```
1: Focus(S)
2: For i=1 to No_of_Attributes
3:     For サイズ i の S の各部分集合 S_i
4:         If 不整合度=0
5:             解候補部分集合:=S_i
6:             Goto out
7:         end if
8:     end for
9: end for
10: label:out
11: 解候補部分集合 S_i を出力
```

Algorithm 9 Relief Algorithm

```
1: Relief(S)
2: 全ての重みを 0 に初期化
3: For j=1 to No_of_Sample
4:     ランダムにデータを一つ選択
5:     ニアヒット (hit) とニアミス (miss) を検索
6:     For 全ての属性 F_i
7:         W_i:=W_i-δ(x_ji,hit_ji)^2 + δ(x_ji,miss_ji)^2
8:     end for
9: end for
10: W_i:=W_i/No_of_Sample
11: For 全ての属性 f_i
12:     If W_i > 閾値    Then S_0 : = S_0 ∪ {f_i}
13: end for
14: S_0 を出力
```

が、探索の方向は、Focus が前向きなのにたいして、ABB は後ろ向きである。プログラムの legitmate は、すでに枝刈りされたノードの子を、別のノードを展開してテストすることを避けるための条件で、具体的には、あるノードと枝刈りされたとのハミング距離が、1でないことを確認することである。

Algorithm 10 ABB Algorithm

Inputs:
S-all features x in data D,
M-inconsistency rate as evalution measure,
Q-an empty queue,
S_1,S_2-subsets
Initialize:L={},$\delta = CalM(S, D)$
ABB(S,D)
For each feature x in S
    S_1=S-x （一度に一個すつ削除）
    enque(Q,S_1)
end for
While notEmpty(Q)
    S_2=deque(Q);
    If(S_2 is legitmateCalM(S_2,D) $\leq \delta$) $L = append$(S$_2$,L)
        ABB(S_2,D)
    end if
end while
S_{min}= the minimum subset(s) in L satisfying M.
Output:S_{min}

4.2 bioinfomaticsにおける feature selection

4.2.1 配列解析

配列解析には長い歴史があるが、属性選択という観点からみると 2 種類に分類される [13]。1 つ目は、内容分析 (content analysis) と信号解析 (signal analysis) である。content analysis

は、配列の幅広い特性に焦点をあてる。例としては、ある生物の機能をもつタンパク質の配列の傾向などである。また、signal analysis は、配列内の重要なモチーフの同定に焦点があてられる。例としては、遺伝子の構造要素や転写領域の同定が上げられる。

content analysis

Bioinfomatics の初期からタンパク質の coding 領域の予測は、研究の関心が高かった。多くの特徴が配列から抽出することができ調整位置でお互いに依存し合っているために、マルコフモデルの様々な種類が開発された。代表的なマルコフモデルとして interpolated Markov mode(IMM) がある。サンプルサイズが小さい場合、異なる順序の間でマルコフモデルで補間をし、関連する属性のみを選択する。フィルタ法を使う。さらに IMM を拡張した interpolated context mode(ICM) がある。これは、隣接していない属性の依存性を処理するために拡張されたものである。属性の相関性を考慮するためにフィルタ法を使いベイジアン決定木をクロスさせる。Markov Branket multivariate Filter(MBF) は、coding の潜在的な予測のために異なった測度のものを組合せ、相関性をあるものを残す為に使われる。

配列からのタンパク質の機能を予測するなどの第 2 の技術について述べる。rRNA の大きなサブユニットを判別するために遺伝的アルゴリズムを組み合わせた手法、SVM のカーネル関数で重みが少ないものを選択的に削除する手法、配列解析での属性選択手法は、プロモータ領域の予測、microRNA を標的とした予測手法が提案されている。

signal analysis

シグナル解析とは、配列の中でタンパク質やその複合体の転写領域のシグナルを認識するための手法である。回帰問題が転写モチーフや遺伝子発現モデルでの関連モチーフを発見するためのアプローチである。判別問題では、モチーフの判別が行われる。また、TIS など構造的な要素がある遺伝子予測領域の発見がある。

4.2.2 マイクロアレイ解析

マイクロアレイの解析では、高次元データの扱いが課題となる。属性選択問題はその問題解決の一手段である。

以下マイクロアレイ解析で開発された手法の一覧である [13]。

(a) Filter method

   i. Univarite

      A. Parametric

      B. t-test, ANOVA, Bayesian, Regression,Gamma

      C. Model-free

      D. Wilcoxon rank sum, BSS/WSS, Rank products, Random permutation, ,TNoM

   ii. Multivarite

      A. Bivarite, CFS, MRMR,USC,Markov blanket

(b) Wrapper method

   i. Seqential search, Genetic algorithms, Estimation of distribution algorithm

(c) Embeded method

   i. RandomForest, Weight vector of SVM,Weights of logistic regression

第 5 章

Random Forest を用いたエピジェネティクス関連領域の予測と属性選択

5.1 背景

遺伝子と遺伝子発現は生物を理解するために重要な概念である。ヒトゲノムプロジェクトを含む様々なゲノムプロジェクトの成功によって、今日生物が遺伝子の配列の数千や数万の遺伝子をもつということは周知となっている。遺伝子発現は（タンパク質の合成にもとづく転写、翻訳）は生命にとって重要であるが、それは必ずしもセントラルドグマに従って必ず発現するものではない。同じ種の中でさえ、遺伝子発現は、飢餓や低温衝撃を含む個々、組織、物理科学状態で制限され、さらに遺伝子発現は様々な因子で規制される。O. Hobert [49] では、転写因子（TFs) と microRNA(miRNAs) は、別々に協合したり、拮抗して規制し、次第に遺伝子の規制の複雑なネットワークを形成することを示した。ヌクレオソームは、真核生物の遺伝子の中で遺伝子の規制の因子であり、最近活発な研究が行われている。真核生物の比較的長い遺伝子の配列は、ヒストン (H2A、H2B、H3、H4) と呼ばれる 4 対のタンパク質とヒストンオクタマーの周りを包んだ DNA の 145-147 塩基対から構成されるヌクレオソームと呼ばれる単位で包まれる。クロモソームの中へ DNA の圧縮物を含んだヌクレオソームの様々な役割の中で、遺伝子調整は重要な役割を果たす。なぜなら遺伝子の転写は、DNA がヌクレオソームによって密である領域のなかでは転写されにくい状態にあるため、ヌクレオソームによる遺伝子調整は、TF や miRNA によるものよりかなり高く、DNA のヌクレオソームの占有率は各遺伝子の発現パターンを理解するための重要な手掛かりである。さらに、DNA の and/or のヒストンの化学的な組み換えもまたクロマチンの形成 (密、または、ゆるい結びつき) と遺伝子調整に関連する。ポコロップら [6] は、酵母菌のヒストン占有率と組み換えのゲノム全体に及ぶ地図で比較実験の結果を報

告している。彼らは組織的な解析を行い、ヒストンのプロファイルと遺伝子の発現の間の関係を明らかにした。配列の情報からヒストンのプロファイルを予測したなら、遺伝子の発現パターンを理解する有用な手掛かりになると考えられる。ポコロップら [6] によって公開されている 14 のデータセットのうち 10 セットを用い、 Pham らは、DNA の配列でのメチル化の占有率と組み換えの予測を試みた。RBF カーネルで SVM を用いて、ウインドウサイズ (k=3、4、5、6) でのいろいろなサイズでの配列 k-gram の特徴は、それらは、高い予測率を示した。彼らは、正例と負例を区別するための情報のある特徴を同定するため polynomial kernel で SVM をつかってランキングを行っている。SVM のかわりに、Tran は、Conditional Random Field(CRF) を用いて、同様の結果と属性のランキングを得ている。しかし、それらの両者の研究でも予測率の正確さの改良のために属性のランキングをつかっていない。本研究では、RandomForest での variable importance と呼ばれる属性ランキングを用いて、この問題への属性選択を提案した。

ノイズまたは間違った値のある属性は予測性能を減少させるため、そのような属性を除くことで全ての属性を使用より実際に正確さを改良される。ランキングに沿って、variable importance と属性選択を用いて属性のランキングをつかうことにより、予測性能の改良は小さい値ではあるが全データにおいて改良が見られた。

さらに、最高の予測率を示す属性の部分集合の近傍の組合せを考慮することで更によい正確さをもつ属性の部分集合の種類を探すことにより改良される。

予測性能の改良のほかに、その結果は最良の正確さをもつ属性の集合のまわりでの属性の選択についての洞察をもたらす。予測性能の改良に加えて、属性のランキングは属性とデータセットのグループの間の関係を明らかにする。この章の構成はの次ののとうりである。5.2 では、データセットの説明、問題の定式化、予測アルゴリズムの説明を行い、本研究で提案する近傍探索のアルゴリズムを説明する。5.3 では、実験結果の解析とその説明を示す。最終的に 5.4 でこの研究のまとめを述べる。

5.2 提案手法

5.2.1 正例と負例の準備

表 1 はヌクレオソームの占有率と組み換えに関しての 10 個のデータを表示している。それらはポコロップらによって出版されたデータの部分集合であり、本研究と先行研究に使

われた。表 1 でデータの名前の中で ,「H3」と「H4 はヒストンの型を示しており、「K」とそれに続く数字は組みかえられたアミノ酸を示している。(「K9」は、ヒストンのなかで 9 番目のアミノ酸のリジンを示す。) 1.1.2 より大きな値を持つ位置に対しては、中止に位置をとり、500 塩基の部分配列を抽出することによって正例とする。2. 同様に 0.8 以下からの位置を負例とする。他は使われない。正例と負例は表 2 に示す。生成された部分配列は各部分配列の中で k-gram でカウントしベクトルに変換する。例としては、ウインドウサイズ k=3 を適用する。頻度としては“ AAA ”、“ AAT ”、“ AAC ”、 “ AAG ”、“ ATA ” ... がカウントされる。先行研究で探索した様々なウインドウサイズと通して、ここでは、k=3 を単純に仮定してみる。例としては、3 ヌクレオチドの配列を参照している 64 次元の属性をもつベクトルとして表現される。

5.2.2 予測アルゴリズムと実装

SVM は、判別分析や回帰分析でいろいろな種類に応用される保障された教師つき学習アルゴリズムである。また、バイオインフォマティクスでは、SVM は構造解析、遺伝子発現解析、たんぱく質相互作用に十分適用された。われわれは RBF カーネルで SVM に適用し [3] の研究で使われた。われわれは予測の性能をさらに良くするためにパラメーター σ =0.05 とした。実装については、Pham は、SVM の実装を使った。再現性のために、R のカーネルパッケージに含まれる ksvm を使用した。

5.2.3 属性選択と属性ランキング

パターン認識のために、属性選択は活発に研究された。幅広い応用としてテキスト判別、たんぱく質判別、侵入発見などを含む。問題は次のように定式化される。与えられた N の属性のうち、予測の判別力を最高にする部分集合をいかに発見するか？探索空間は 2^N(または空集合を除いて 2^{N-1}) のため exhaustive search は実際の問題に適用できない。この問題を解くために、いろいろな手法が提案されている。それらは 3 種類に大別される。ラッパ手法は、学習と予測を行い、候補として残るかまたは削除されるべき属性 (属性集合) を決定するために予測の結果を活用する。反対に、フィルタ法は、学習や予測なしに前処理の段階で統計的に属性の関連性を評価する。3 番目の方法は埋め込み法である。学習器に固有であるが、学習のプロセス内で属性選択を実行する。この判別のほかに、探索アル

Dataset	説明
H3	H3 の占有率
H4	H4 の占有率
H3K9ac	H3K9ac は H3 をアセチル化
H3K14ac	H3K14ac は H3 をアセチル化
H4ac	H4ac は H4 をアセチル化
H3K4me1	H3K4me1 は H3 をモノメチル化
H3K4me2	H3K4me2 は H3 をジメチル化
H3K4me3	H3K4me3 は H3 をトリメチル化
H3K36me3	H3K36me3 は H3 をトリメチル化
H3K79me3	H3K79me3 は H3 をトリメチル化

表 5.1: ヌクレオソームデータセット

Dataset	正例	負例
H3	7,667	7,298
H4	6,480	8,121
H3K9ac	15,415	12,367
H3K14ac	18,771	14,277
H4ac	18,410	15,685
H3K4me1	17,266	14,411
H3K4me2	18,143	12,540
H3K4me3	19,604	17,195
H3K36me3	18,892	15,988
H3K79me3	15,337	13,500

表 5.2: 例の数

ゴリズムの選択は、属性選択を特徴づける重要な因子である。前方または後方削除のようなヒューリステックな方法に加えて、多くの探索アルゴリズムが属性選択で提案されている。best-first search、floating search、random search including Relief algorithm、genetic algorithm search などがある。

本研究では、属性選択のアルゴリズムは 2 つのステップに分けられる。

Algorithm 11 属性選択のアルゴリズム

1: 属性 f_1,⋯,f_n の前処理のランキングに沿って、f_1 と f_n は、ランキングのトップと一番下である。部分集合 $\{f_1\}$,$\{f_1, f_2\}$,⋯,$\{f_1, \cdots, f_n\}$ は、SVM(RBF カーネル) によって学習と予測を実行する。

2: 前のステップで最高の予測の正確さをもつ属性の集合の近傍（前後 3 近傍) をテストする。

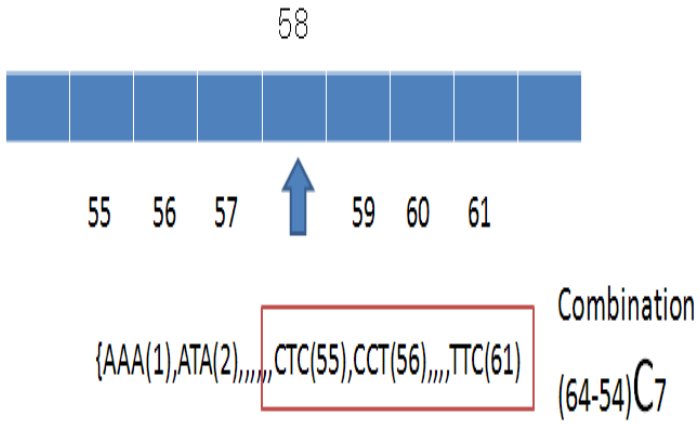


図 5.2: 近傍探索

図 5.2 では近傍探索の組合せを示している。step.1 での最高の予測率を示す属性部分集合の末尾の属性を中心に、前後 3 近傍、合計 7 属性の全組合せを考える。図 5.2 では、58 番目の属性を中心として前後 3 近傍を考えている。全組合せは全属性の 64 から 58 を引いた 10 属性から 7 つを選択する組合せとなる。step.1 では、Gini 係数を計算するために randomForest をつかった。情報利得や t -検定と同様属性ランキングの中での人気のある尺度である。Random forest は、ランダムに生成される決定木に基づくアンサブル学習の

一種である。様々な利点があるが、属性ランキングではGini係数を生成することに使われる。step.2で変化を生成するために、ランキングのなかでの8つの連結した属性の集合を考える。step.1で最高の予測率をもつ属性の集合のなかで一番低いGini係数のものもつ属性を中心にする。例えば、もし$\{f_1,..,f_k\}$が、ステップ1で最高の属性の集合であるとすると$\{f_1,..,f_{k-4}\}$と積集合$\{f_{k-3}, f_{k-2},f_{k-1}, f_k, f_{k+1}, f_{k+2}, f_{k+3}\}$の結合である。k ≦ n-4(=60)なので、$2^8$の属性の集合がテストされる。その他に、$2^4$から$2^7$までの属性がテストされる。

5.3 実験結果

5.3.1 randomForestによる属性選択

学習と予測の結果に加えて、Rのrandom Forestの関数は各属性に関してMeanDecreaseGiniと呼ばれる値を生成できる。これを使うと、我々は重要性の順序で属性をランク付けできる。（判別力）図2では、各データセットで区間[0,1]で正規化されたランキングとMeanDecreaseGiniの間の関係を示す。

すべてのデータに共通して、属性の重要性はtop2から11までの領域で急激に減少し、次に残りはゆっくりと減少する。これは、データの属性の大半が小さい役割しかもたないことを意味している。重要な属性は、図2の視覚的にわかるように表3に表している。この表では、多くのデータが”T”や”A”から構成される属性をもっていることがわかる。2つのデータセットで（H3K9acとH3K36me3)、”G”と”C”の比率は比較的高い。その他の観察としては、

・TTTとAAAではH3とH4の占有率が重要である。H3のメチル化は大半が重要であるが、しかしアセチル化はそうでもない。

・AATやATTは、H3の占有率の中で重要である。（H4では重要ではない）それらはH4acを含むすべてのアセチレンの中で共通して重要である。

・ATAとTATは、H3とH4の占有率で重要であるが、AATやATTとは反対にそれらはH3のメチル化において選択的に重要である。

・H3K79me3は、TTT、AAA、AAT、ATTがあまり重要ではないという意味で特別である。

注意：重要な属性が正例と負例のどちらが選択されるかは明確ではない。

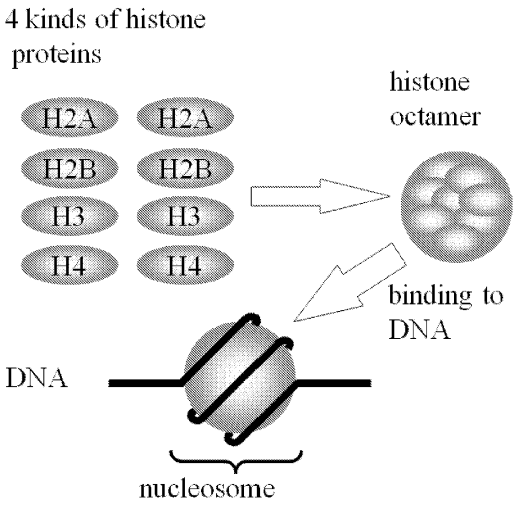


図 5.1: ヒストンとヌクレオソーム

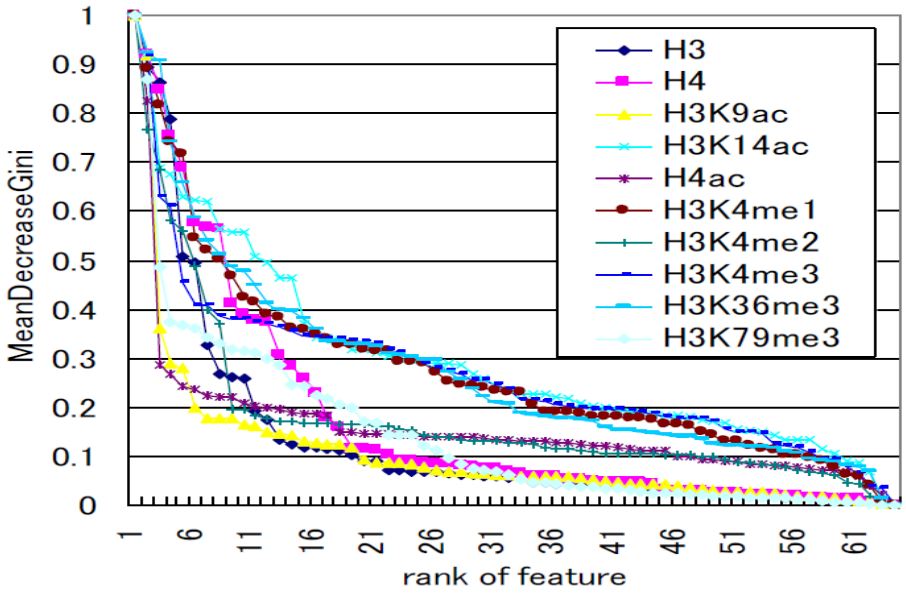


図 5.3: 属性ランキングによる MeanDecreaseGini

Dataset	ランキングの幅	ランキングの降順でリストした高い重要性の属性
H3	1〜10	TTT,AAA,TAA,TTA,ATA,TAT,AAT,CCA,ATT,TGG
H4	1〜8	TTT,ATA,AAA,TAT,CCA,ATC,TAA,TGG
H3K9ac	1〜5	ATT,AAT,CGC,GCG,TTA
H3K14ac	1〜2	AAT,ATT
H4ac	1〜2	AAT,ATT
H3K4me1	1〜5	TTT,TAT,CCA,ATA,AAA
H3K4me2	1〜8	ATT,AAT,TTA,TTT,TAT,ATA,TAA,AAA
H3K4me3	1〜5	ATT,AAT,AAA,TTT,TAT
H3K36me3	1〜11	CCA,ATA,TGG,TAT,CAA,TTT,AAA,TCA,TTG,ATC,TGA
H3K79me3	1〜3	ATA,TAT,ATC

表 5.3: 重要な属性のリスト

5.3.2 ランキングに沿って選択された属性部分集合の予測性能

5.2 で説明したステップ 1 で、ランキングに沿った 64 の異なった属性の部分集合 {f1},{f1,f2}, …,{f1,…,fn} が、RBF カーネルの SVM で各データに対してテストされた。予測の結果は図 5.3 に要約されている。最高の予測性能をもつ部分集合は表 5.4 に示されている。図 5.3 では、各データでの予測性能が属性ランキングの最下位からなめらかに増加している。与えられた属性が低い重要性を含む属性を含む場合でも SVM は、さらによい判別のため大半の属性を利用することができる。実際に、表 5.5 では、3 つのデータで、最高の属性の集合が全属性の集合 (size=64) にたいして同定されている。(H4ac, H3K4me2, H3K36me3) 他のデータセットに対しては小さいが予測性能は改良される。予測性能は次のように計算できる。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

ここで、TP は true positive、FP は false positive、TN は true negative、FN は false negative。Pham も同様に、3 クロスバリデーションで性能を評価している。

5.3.3　最高の性能をもつ属性の部分集合の周りの近傍での予測

表 5.4 の最高の属性の部分集合を使って、ステップ 2 での 4 属性の属性選択が各データに対して行った。結果は表 5.4 で示されている。第 2 列が先行研究での実験結果である。第 3 列が本研究で行った属性選択なしの実験結果である。第 4 列がアルゴリズムの step.1 での選択された属性の部分集合である。第 5 列が step.1 で選択された属性部分集合に対し SVM を行った結果である。第 6 列が step.2 で最高の予測性能を持つ属性の部分集合である、第 7 列が step.2 での SVM の予測結果である。第 8 列が step.1 と step.2 での予測性能の差異である。

H3 の場合、先行研究の予測率は 84.93 %、本研究の SVM での全属性の予測率は 86.19 %、step.1 での属性部分集合の数は 59 で、step.1 での予測率は 86.21 %、近傍探索した結果の予測率は 86.45 %で全属性の 86.19 %より 0.26 %上昇した。H4 の場合、先行研究の予測率は 85.91 %、本研究の SVM での全属性の予測率は 86.36 %、step.1 での属性部分集合の数は 55 で、step.1 での予測率は 86.64 %、近傍探索した結果の予測率は 86.36 %で全属性の 87.04 %より 0.31 %上昇した。H3K9ac の場合、先行研究の予測率は 71.04 %、本研究の SVM での全属性の予測率は 72.68 %、step.1 での属性部分集合の数は 62 で、step.1 での予測率は 72.86 %、近傍探索した結果の予測率は 72.90 %で全属性の 72.68 %より 0.22 %上昇した H3K14ac の場合、先行研究の予測率は 68.64 %、本研究の SVM での全属性の予測率は 69.98 %、step.1 での属性部分集合の数は 62 で、step.1 での予測率は 70.23 %、近傍探索した結果の予測率は 70.36 %で全属性の 69.98 %より 0.38 %上昇した。H4ac の場合、先行研究の予測率は 67.65 %、本研究の SVM での全属性の予測率は 69.26 %、step.1 での属性部分集合の数は 64 で、step.1 での予測率は 69.26 %、近傍探索した結果の予測率は 69.32 %で全属性の 69.26 %より 0.06 %上昇した。H3K4me1 の場合、先行研究の予測率は 66.21 %、本研究の SVM での全属性の予測率は 67.13 %、step.1 での属性部分集合の数は 62 で、step.1 での予測率は 67.28 %、近傍探索した結果の予測率は 67.13 %で全属性の 67.13 %より 0.44 %上昇した。H3K4me2 の場合、先行研究の予測率は 66.09 %、本研究の SVM での全属性の予測率は 68.23 %、step.1 での属性部分集合の数は 64 で、step.1 での予測率は 68.23 %、近傍探索した結果の予測率は 68.28 %で全属性の 68.23 %より 0.05 %上昇した。H3K4me3 の場合、先行研究の予測率は 62.37 %、本研究の SVM での全属性の予測率は 65.79 %、step.1 での属性部分集合の数は 63 で、step.1 での予測率は 65.87 %、近傍探索した結果の予測率は 65.92 %で全属性の 65.79 %より 0.14 %上昇した。H3K36me3 の場

合、先行研究の予測率は 71.74 %、本研究の SVM での全属性の予測率は 73.80 %、step.1 での属性部分集合の数は 64 で、step.1 での予測率は 73.80 %、近傍探索した結果の予測率は 73.80 %で予測率に変化なし。H3K79me3 の場合、先行研究の予測率は 78.25 %、本研究の SVM での全属性の予測率は 79.56 %、step.1 での属性部分集合の数は 61 で、step.1 での予測率は 79.77 %、近傍探索した結果の予測率は 79.94 %で全属性の 69.53 %より 0.38 %上昇した。３塩基での場合も近傍探索が作用し、９データで予測率が上がったの近傍探索の有効性が示せた。

ステップ１と同じように、ステップ２は小さいが性能は改良されている。予測性能は別に、高い (低い) ランキングを持つ属性は常にステップ２では最高の属性の部分集合の中に必ずしも含まれない。例えば、データセット H4 では、GGG(59) は最高の集合の中に含まれるが、一方、TAG(56), GCC(57),AGT(58) は、含まれない。(表 5.4) さらに特別なケースとしては、59 から 64 のランクでの６属性の中で、最低のランクの属性が生き残り、よりランキングが高い属性が選択されていない。属性ランキングの中でステップ１のなかでの属性選択だけは十分ではなく、ステップ２での近傍を考慮した探索によって重要な属性が拾われる。よって提案している近傍探索の有効性を示している。

Dataset	予測性能(Pham)	全属性	step1で属性選択した部分集合	step1での最高の予測性能をもつ予測	step2での最高の予測性能をもつ属性部分集合	step2での最高の属性部分集合の予測	step1とstep2との改良の合計
H3	84.93	86.19	TTT(1),...,CTA(59)	86.21	TTT(1),...,TGG(55),AAC(56),CTA(59),ACT(60),TAG(62)	86.45	0.26
H4	85.91	86.36	TTT(1),...,CCC(55)	86.64	TTT(1),...,AGC(51),ACG(52),TGC(53),CAC(54),CCC(55),GGG(59)	86.67	0.31
H3K9ac	71.04	72.68	ATT(1),...,CCC(62)	72.86	ATT(1),...,TCC(58), AGG(59),GTC(60), CCC(62), GGA(63)	72.90	0.22
H3K14ac	68.64	69.98	AAT(1),...,GCC(62)	70.23	AAT(1),...,GGG(58), GGC(60),CCC(61), GCC(62)	70.36	0.38
H4ac	67.65	69.26	AAT(1),...,CGG(64)	69.26	AAT(1),...,GGA(60), GGG(61),CCG(63), CGG(64)	69.32	0.06
H3K4me1	66.21	67.13	TTT(1),...,GCG(62)	67.28	TTT(1),...,CGT(58), CGG(64)	67.56	0.44
H3K4me2	66.09	68.23	ATT(1),...,CGG(64)	68.23	ATT(1),...,CCC(60), CCG(61),GCG(63), CGG(64)	68.28	0.05
H3K4me3	62.37	65.79	ATT(1),...,CCG(63)	65.87	ATT(1),...,GGC(59), GCC(60),GCG(61), CGC(62), CGG(64)	65.92	0.14
H3K36me3	71.74	73.80	CCA(1),...,CGG(64)	73.80	CCA(1),...,GGC(60), GCC(61),CGT(62), CCG(63), CGG(64)	73.80	0.00
H3K79me3	78.25	79.56	ATA(1),...,CGA(61)	79.77	ATA(1),...,ACG(57), GCT(59),CGA(61), TCG(63), CGT(64)	79.94	0.38

表 5.4: Pham による予測性能と全属性, ステップ 1, ステップ 2

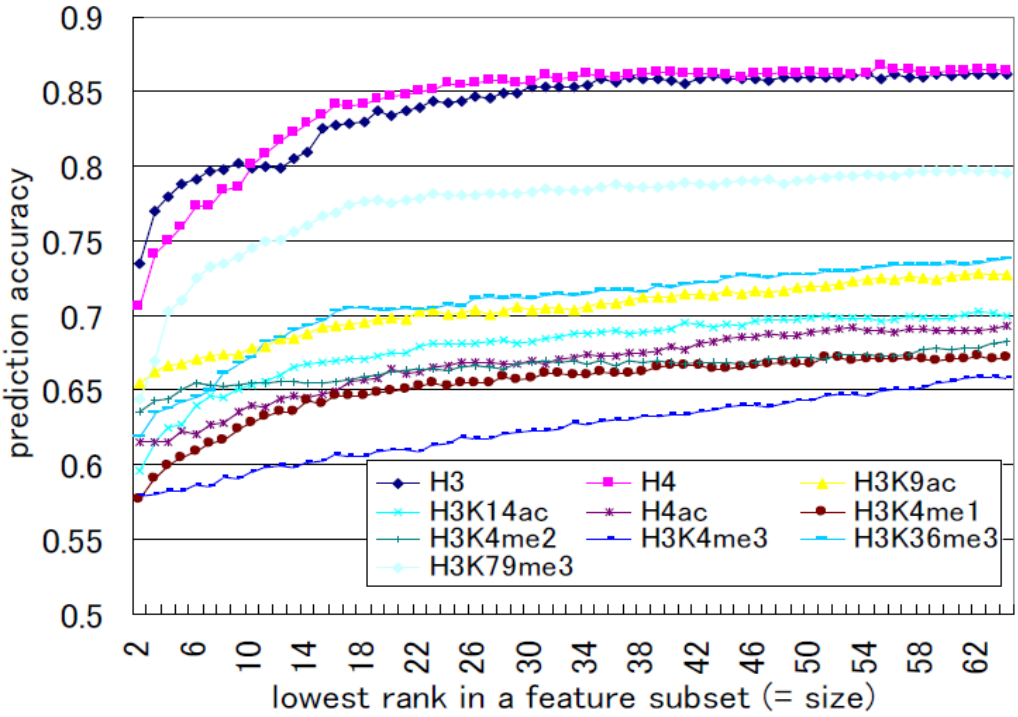


図 5.4: ランキングに沿った属性選択の効果

Dataset \ Rank	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
H3	CGA	GTG	GAG	TGT	AGT	TCG	AAC	GGG	GTA	CTA	ACT	TAC	TAG	ACG	CGT
H4	CGT	AGC	ACG	TGC	CAC	CCC	TAG	GCC	AGT	GGG	GTG	GCT	CGA	TCG	GGC
H3K9ac	TAG	CAG	ACT	CGA	AGT	GGT	ACC	GGG	TCC	AGG	GTC	CCT	CCC	GGA	GAC
H3K14ac	TGC	CTG	AGG	GAC	GTC	CGT	CCT	ACG	GGG	TCG	GGC	CCC	GCC	CGG	CCG
H4ac	CTA	AGG	TCG	CTG	GTC	TCC	GCC	CCT	CAG	CCC	GGA	GGG	GAC	CCG	CGG
H3K4me1	GAC	GTC	AGG	CCC	CGA	ACG	TCG	GGG	CGT	CGC	GGC	GCC	GCG	CCG	CGG
H3K4me2	GAC	CAG	AGC	TCC	TCG	GCC	CCT	GGG	ACG	CGT	CCC	CCG	CGC	GCG	CGG
H3K4me3	GTC	CGA	GGG	CCT	CGT	GAC	TCG	CCC	ACG	GGC	GCC	GCG	CGC	CCG	CGG
H3K36me3	CAG	CTG	CCT	GAC	CCC	CGA	AGG	GTC	ACG	TCG	GGC	GCC	CGT	CCG	CGG
H3K79me3	CTA	GGG	GAG	CGG	GAC	CTC	AGC	AGG	ACG	GCT	GTC	CGA	CCT	TCG	CGT

図 5.5: ステップ 2 で最高の属性の部分集合の属性

5.3.4 ラッパ法で他の属性選択との比較

ラッパ法で人気のある手法と本研究の手法を比較するために、Weka を用いた属性選択の広範囲の実験を行った。この実験で、我々は 6 つの判別器 (BayesNet, NaiveBayes, SMO, J48,AdaBoostM1, RandomForest) と 5 つの探索方法 (BestFirst, GeneticSearch, GreedyStepwise,LinearForwardSelection, RankSearch) を組み合わせてテストした。そのあとに各属性の部分集合の予測性能が図 5.2 と図 5.3 として同様の方法で計算した。その結果、表 5.6 から 5.11 までを示せた。これらの表の場合、各属性の部分集合の予測性能は属性の部分集合のサイズを () に示されている。いくつかの組合せは、属性の部分集合が生成されないけれども（これらの表では、-とする.）上の判別器と探索手法で生成される属性の大半は本研究の手法の表 4 での比較のなかでより低い予測性能を実現している。

Dataset/Rank	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.7899(6)	0.8507(29)	0.7899(6)	0.8258(20)	0.8607(56)
H4	0.8453(20)	0.8475(34)	0.8453(20)	0.8399(19)	0.8463(24)
H3K9ac	0.6984(23)	0.7141(41)	0.6984(23)	0.6919(21)	0.7163(51)
H3K14ac	0.6767(20)	0.6921(44)	0.6767(20)	0.6680(18)	0.6794(26)
H4ac	0.6511(17)	0.6716(34)	0.6511(17)	0.6576(20)	0.6892(53)
H3K4me1	0.6448(19)	0.6556(33)	0.6448(19)	0.6425(15)	0.6657(42)
H3K4me2	0.6475(7)	0.6600(30)	0.6475(7)	0.6464(5)	0.6491(7)
H3K4me3	0.5997(10)	0.6354(35)	0.5997(10)	0.6013(10)	0.6090(17)
H3K36me3	0.7098(28)	0.7182(38)	0.7098(28)	0.7024(19)	0.7352(58)
H3K79me3	0.7653(20)	0.7846(45)	0.7653(20)	0.7768(26)	0.7913(43)

表 5.5: BayesNet classifier

BayesNetclassifier での属性選択手法の比較

表 5.6 では、BayesNet 判別器を用いた。属性選択の手法では、Genetic search と Rank search の値が高い。Genetic search では、H4、H3K14ac、H3K4me2、H3K4me3 で最も予測率が高く。Rank Search は、H3、H3K9ac、H4ac、H3K4me1、H3K36me3、H3K79me3 が最も高い。選択した属性数では、BestFirst と GreedyStepwise が同じ属性数である。LinerFowardSelection と合わせて 3 つの選択手法はほぼ同じ属性数を選択している。Genetic Search と Rank Selection の選択する属性数は、総じて多いが、一番予測率が高い選択法は、一番属性数が多いという傾向がみられる。更に図 5.2 と比べてみると H3 の場合、Gini 係数が 0.5 以上の属性は 6 個である。選択方法では、BestFirst、GreedyStepwise が属性数 6 となっている。次に LinearForwardSelection は属性数 20 であるが、Gini 係数は、0.1 程度である。Genetic search は、29 個で、Rank search は、56 個であるが、それぞれの Gini 係数はほとんど 0 に近い。H4 についてみると、Gini 係数が 0.5 以上は、11 個程度である。属性数が一番少ない選択方法は、LinearForwardSelection の 11 個で、一番多い属性数は 34 個で GeneticSearch である。図 5.2 では、属性数 20 個程度のところで勾配の傾きが変化している。H3 と H4 は、予測率は 10 データの中で最も高いデータであるが、H3 は、RankSearch で 56 個の属性数であるが、H4 は、GeneticSearch で 34 個と選択される属性数がかなり少ないことが特徴的である。H3K9ac は、図 5.2 では、重要な属性がかなり少数である。Gini

係数が 0.5 の値で 3 個である。最も予測率の高い RankSearch での属性数は 51 となり数に開きがある。H3K14ac は、図 5.2 では 10 データのうちで最もなだらかな曲線になっている。GeneticSearch で選択された 44 個の属性が最も予測率が高いが 0.6921 である。H4ac も図 5.2 では、少数の属性が寄与している急こう配になっている。しかし Rank Search では、属性数 53 個となっている。H3K4me1 は、図 5.2 では、なだらかな勾配でかつ裾野の部分が厚い。Rank Search では、属性数 42 個となっている。H3K4me2 は、図 5.2 では Gini 係数 0.5 で 6 個位の属性数があり、裾野が厚い。Genetic　Search で属性数 30 個となっている。H3K4me3 は、図 5.2 では Gini 係数 0.5 で 7,8 個位の属性数があるが、裾野が薄い。Genetic Search で属性数 35 個となっている。H3K36me3 は、図 5.2 で Gini 係数 0.7 のところまで急に落ち、そこから Gini 係数 0.4 のところまで第 2 の勾配、そしてそこから第 3 の勾配となっている。裾野は厚い部類に入る。RankSearch では、属性数 58 である。H3K36me3 は、図 5.2 では、Gini 係数 0.4 のところまで急に落ち、そこから 4 つほど階段状の勾配があり、Gini 係数 0.1 のところで 0 に漸近している。Rank Search で属性数 43 個が最も予測率が高い。Rank Search は、図 5.2 で少数の属性が重要としているデータでも選択する属性数はかなり多い傾向がある。逆に Genetic　Search が Rank Search よりも優位のとき、Rank Search が選択する属性数はかなり少ない傾向がある。

Dataset/Rank	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.8394(26)	0.8445(28)	0.8174(14)	0.8386(21)	0.7909(8)
H4	0.8568(30)	0.8554(36)	0.8568(30)	0.8395(18)	0.8463(24)
H3K9ac	0.7002(22)	0.7100(36)	0.6945(18)	0.7018(24)	0.7151(50)
H3K14ac	0.6663(18)	0.6888(35)	0.6620(15)	0.6765(24)	0.6709(16)
H4ac	0.6655(22)	0.6774(33)	0.6655(22)	0.6628(26)	0.6926(64)
H3K4me1	0.6391(19)	0.6568(31)	0.6390(16)	0.6427(18)	0.6639(41)
H3K4me2	0.6443(6)	0.6577(23)	0.6450(6)	0.6422(4)	0.6491(7)
H3K4me3	0.6231(32)	0.6378(34)	0.6054(10)	0.6092(15)	0.6579(64)
H3K36me3	0.7123(28)	0.7151(33)	0.7123(28)	0.6992(24)	0.7126(30)
H3K79me3	0.7890(43)	0.7850(36)	0.7727(18)	0.7829(34)	0.7926(46)

表 5.6: NaiveBayes classifier

NaiveBayes classifier での属性選択手法の比較

表 5.7 では、NaiveBayes classifier での結果を示している。H3 では、Genetic Search が最も予測率が高く属性数は 28 である。ここでは Rank Search が最も低く属性数は 8 である。H4 では、BestFirst と GreedyStepwise が同じ予測率になり属性数は 30 である。これは、珍しい例である。H3K9ac では、Rank Search が最も高く属性数は 50 である。H3K14ac は、Genetic Search が最も予測率が高く属性数は 35 である H4ac は、Rank Search が最も高いが、全属性の 64 である。しかし、他の選択方法、たとえば Genetic Search では属性数 33、BestFirst と GreedyStepwise はともに属性数は 22、LinearForwardSelection は属性数は 26 となっている。H3K4me1 では、Rank Search が最も高く属性数は 41 である。H3K4me2 では、Genetic Search が最も高く属性数は 23 である。このデータに関しては半分以下の属性数になっている。H3K4me3 では、Rank Search が最も高く属性数は 64 と全属性である。しかし、GreedyStepwise は、属性数 10 と差が大きい例である。H3K36me3 は、Genetic Search が最も高く属性数は 33 である。このデータでは、他の選択方法も属性数は半分程度である。H3K79me3 では、、Genetic Search が最も高く属性数は 46 である。NaiveBayes classifier でも最も予測率が高い選択方法の属性数は、他の選択方法よりも属性数が一番高い選択方法である同程度の予測率ではあるが、属性数に差が大きいデータが存在する。

Dataset/Rank	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.8382(24)	0.8581(41)	0.8228(12)	0.8563(41)	0.8539(44)
H4	0.8527(47)	0.8613(48)	0.8494(31)	0.8478(24)	0.8625(62)
H3K9ac	0.7072(40)	0.7187(40)	0.6954(20)	0.6994(23)	0.7142(44)
H3K14ac	0.6786(26)	0.6892(41)	-	-	0.6984(58)
H4ac	0.6669(25)	0.6827(47)	0.6667(23)	0.6684(27)	0.6880(52)
H3K4me1	0.6484(32)	0.6614(39)	0.6457(19)	0.6459(29)	0.6717(61)
H3K4me2	0.6623(14)	0.6688(37)	0.6623(14)	0.6606(22)	0.6810(62)
H3K4me3	0.6271(27)	0.6460(43)	0.6134(16)	0.6042(13)	0.6573(59)
H3K36me3	0.7058(26)	0.7208(44)	0.7036(23)	0.7238(45)	0.7368(63)
H3K79me3	0.7825(35)	0.7953(44)	0.7816(29)	0.7880(45)	0.7961(61)

表 5.7: SVM(SMO) classifier

support vector machine(SMO)classifier での属性選択手法の比較

表 5.8 では、support vector machine(SMO)classifier での結果を示している。H3 では、Genetic Search が最も予測率が高く属性数は 41 である。H4 では、Rank Search が最も高く属性数は 62 である。。H3K9ac では、Genetic Search が最も予測率が高く属性数は 40 である。H3K14ac は、Rank Search が最も高く属性数は 58 である。ここで GreedyStepwise と LinearForwardSelection は値が出なかった。原因は不明である。H4ac は Rank Search が最も高く属性数は 52 である。他の選択方法の BestFirst、GreedyStepwise、LinearForwardSelection は属性数 25 前後である。H3K4me1 では、Rank Search が最も高く属性数は 61 である。H3K4me2 では、Rank Search が最も高く属性数は 62 である。他の選択方法の BestFirst、GreedyStepwise は属性数 14 である。H3K4me3 では、Rank Search が最も高く属性数は 59 と全属性である。H3K36me3 は、Rank Search が最も高く属性数は 63 である。H3K79me3 では、、Rank Search が最も高く属性数は 61 である。support vector machine(SMO)classifier でも最も予測率が高い選択方法の属性数は、他の選択方法よりも属性数が一番高い選択方法である。

Dataset/Rank	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.8072(8)	0.8505(36)	0.8016(6)	0.8072(8)	0.7909(8)
H4	0.8231(10)	0.8482(26)	0.8231(10)	0.8132(9)	0.8373(21)
H3K9ac	0.6698(4)	0.7000(28)	0.6698(4)	0.6698(4)	0.6659(4)
H3K14ac	0.6373(5)	0.6509(12)	0.6373(5)	0.6275(4)	0.6424(10)
H4ac	0.6190(4)	0.6368(12)	0.6190(4)	0.6190(4)	0.6210(3)
H3K4me1	0.6205(7)	0.6276(10)	0.6111(5)	0.6137(6)	0.5993(4)
H3K4me2	0.6475(5)	0.6363(5)	0.6418(3)	0.6475(5)	0.6491(7)
H3K4me3	0.5858(4)	0.5832(7)	0.5858(4)	0.5858(4)	0.5804(4)
H3K36me3	0.6549(5)	0.6712(7)	0.6549(5)	0.6549(5)	0.6788(10)
H3K79me3	0.7127(6)	0.7851(33)	0.7302(7)	0.7127(6)	0.7226(7)

表 5.8: J48 classifier

J48 classifier での属性選択手法の比較

表 5.9 では、J48 classifier での結果を示している。H3 では、Genetic Search が最も予測率が高く属性数は 36 である。H4 では、Genetic Search が最も予測率が高く属性数は 26 である。H3K9ac では、Genetic Search が最も予測率が高く属性数は 28 である。H3K14ac は、Genetic Search が最も予測率が高く属性数は 12 である。H4ac は、Genetic Search が最も予測率が高く属性数は 12 である。H3K4me1 では、Genetic Search が最も予測率が高く属性数は 10 である。H3K4me2 では、Rank Search が最も高く属性数は 7 である。H3K4me3 では、BestFirst、GreedyStepwise、LinearForwardSelection がともに最も高く属性数は 4 と全属性である。H3K36me3 は、Rank Search が最も高く属性数は 10 である。H3K79me3 では、、Rank Search が最も高く属性数は 33 である。決定木である J48 であるが、他の判別器よりも少ない属性を選択される傾向がある。J48 classifier でも最も予測率が高い選択方法の属性数は、他の選択方法よりも属性数が一番高い選択方法である。

Dataset/Rank	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.7873(5)	0.8204(21)	0.7873(5)	0.7873(5)	0.7909(8)
H4	0.8110(10)	0.8527(32)	0.8110(10)	0.8134(9)	0.8286(16)
H3K9ac	0.6784(8)	0.6909(27)	0.6784(8)	0.6784(8)	0.6709(5)
H3K14ac	0.6569(10)	0.6726(29)	0.6569(10)	0.6569(10)	0.6737(21)
H4ac	0.6289(6)	0.6666(29)	0.6289(6)	0.6289(6)	0.6709(39)
H3K4me1	0.6308(10)	0.6485(30)	0.6308(10)	0.6259(9)	0.6590(29)
H3K4me2	0.6471(5)	0.6611(21)	0.6471(5)	0.6471(5)	0.6526(10)
H3K4me3	0.5918(6)	0.6188(29)	0.5918(6)	0.5918(6)	0.6169(26)
H3K36me3	0.6765(8)	0.7016(33)	0.6765(8)	0.6614(6)	0.7118(27)
H3K79me3	0.7500(11)	0.7706(32)	0.7500(11)	0.7500(11)	0.7642(19)

表 5.9: AdaBoostM1 classifier

AdaBoostM1 classifier での属性選択手法の比較

表 5.10 では、AdaBoostM1 classifier での結果を示している。H3 では、Genetic Search が最も予測率が高く属性数は 21 である。H4 では、Genetic Search が最も予測率が高く属性数は 32 である。H3K9ac では、Genetic Search が予測率が高く属性数は 27 である。H3K14ac

は、Rank Search が最も予測率が高く属性数は 21 である。H4ac は、Rank Search が最も予測率が高く属性数は 39 である。H3K4me1 では、Rank Search が最も予測率が高く属性数は 29 である。H3K4me2 では、Genetic Search が最も高く属性数は 21 である。H3K4me3 では、Genetic Search が最も高く属性数は 29 と全属性である。H3K36me3 は、Rank Search が最も高く属性数は 27 である。H3K79me3 では、Genetic Search が最も高く属性数は 32 である。AdaBoostM1 classifier では、H3K14ac と H3K36me3 は、最も高い予測率の属性数より高い属性数をもつデータである。

Dataset/Rank	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.7296(2)	0.8501(33)	0.7296(2)	0.7296(2)	0.8553(48)
H4	0.8434(21)	0.8508(32)	0.6967(2)	0.8492(21)	0.8551(32)
H3K9ac	0.6544(2)	0.7017(36)	0.6544(2)	0.6544(2)	0.7058(35)
H3K14ac	0.5993(2)	0.6861(39)	0.5993(2)	0.5993(2)	0.6950(49)
H4ac	0.6148(2)	0.6754(33)	0.6148(2)	0.6148(2)	0.6926(63)
H3K4me1	0.5878(2)	0.6578(38)	0.5878(2)	0.5878(2)	0.6674(47)
H3K4me2	0.6390(2)	0.6721(35)	0.6390(2)	0.6390(2)	0.6742(47)
H3K4me3	0.5785(2)	0.6357(37)	0.5785(2)	0.5785(2)	0.6577(60)
H3K36me3	0.6187(2)	0.7132(35)	0.6187(2)	0.6187(2)	0.7218(37)
H3K79me3	0.6659(2)	0.7852(40)	0.6659(2)	0.6659(2)	0.7903(41)

表 5.10: RandomForest classifier

RandomForest classifier での属性選択手法の比較

表 5.11 では、RandomForest classifier での結果を示している。H3 では、Rank Search が最も予測率が高く属性数は 48 である。H4 では、Rank Search が最も予測率が高く属性数は 32 である。H3K9ac では、Rank Search が予測率が高く属性数は 35 である。H3K14ac は、Rank Search が最も予測率が高く属性数は 49 である。H4ac は、Rank Search が最も予測率が高く属性数は 63 である。H3K4me1 では、Rank Search が最も予測率が高く属性数は 47 である。H3K4me2 では、Rank Search が最も高く属性数は 47 である。H3K4me3 では、Rank Search が最も高く属性数は 60 と全属性である。H3K36me3 は、Rank Search が最も高く属性数は 37 である。H3K79me3 では、Rank Search が最も高く属性数は 41 であ

る。RandomForest classifier では、Rank Search が全ての選択方法で最も予測率が高い。

5 つの属性選択手法の比較のまとめ

5 つの選択手法の比較として、Genetic Search または Rank Search が最も予測率が高い属性を選択している。

選択された属性数の最も多い選択手法が予測率が最も高い。決定木 J48 は、属性選択では他の判別器よりも少ない属性を選択している。RandomForest classifier では、Rank Search が全てのデータで最も高い予測率を示している。図 5.2 では、寄与度のグラフは、急勾配のグラフ、なだらかなグラフに分かれる。また裾野の厚いグラフ、裾野の薄いグラフと分かれる。

- 急勾配で裾野が薄いグラフ　　H3,H3K9ac,H4ac,H3K4me2,H3K79me3
- 急勾配で裾野が厚いグラフ H3K4me3
- なだらかなグラフで裾野が厚いグラフ H3K14ac,H3K4me1,H3K36me3
- なだらかグラフで裾野が薄いグラフ H4,

5.3.5　長いウインドウサイズ (k=4) の効果

上の実験から我々はウインドウサイズが k=4 のみの場合を考える。本研究の手法は、大きな属性数に対して有用なので、同じ実験で k=4 の場合を実行した。属性数は 256 に増加しているが、表 5.12 で示したように以下のように、本研究の手法は効果的である。H3 の場合、先行研究の予測率は 85.88 %、本研究の SVM での全属性の予測率は 86.43 %、step.1 での属性部分集合の数は 243 で、step.1 での予測率は 86.47 %、近傍探索した結果の予測率は 86.47 %で全属性の 86.43 %より 0.04 %上昇した。H4 の場合、先行研究の予測率は 87.14 %、本研究の SVM での全属性の予測率は 87.04 %、step.1 での属性部分集合の数は 244 で、step.1 での予測率は 87.24 %、近傍探索した結果の予測率は 87.32 %で全属性の 87.04 %より 0.29 %上昇した。H3K9ac の場合、先行研究の予測率は 73.64 %、本研究の SVM での全属性の予測率は 74.98 %、step.1 での属性部分集合の数は 250 で、step.1 での予測率は 75.07 %、近傍探索した結果の予測率は 75.08 %で全属性の 74.98 %より 0.10 %上昇した。H3K14ac の場合、先行研究の予測率は 71.28 %、本研究の SVM での全属性の予測率は 73.28

%、step.1 での属性部分集合の数は 256 で、step.1 での予測率は 73.28 %、近傍探索した結果の予測率は 73.28 %で予測率は変化なかった。H4ac の場合、先行研究の予測率は 69.93 %、本研究の SVM での全属性の予測率は 72.06 %、step.1 での属性部分集合の数は 256 で、step.1 での予測率は 72.06 %、近傍探索した結果の予測率は 72.06 %で予測率は変化なかった。H3K4me1 の場合、先行研究の予測率は 68.29 %、本研究の SVM での全属性の予測率は 69.53 %、step.1 での属性部分集合の数は 251 で、step.1 での予測率は 69.64 %、近傍探索した結果の予測率は 69.71 %で全属性の 69.53 %より 0.18 %上昇した。H3K4me2 の場合、先行研究の予測率は 67.05 %、本研究の SVM での全属性の予測率は 68.89 %、step.1 での属性部分集合の数は 255 で、step.1 での予測率は 68.97 %、近傍探索した結果の予測率は 68.89 %で全属性の 69.53 %より 0.08 %上昇した。H3K4me3 の場合、先行研究の予測率は 65.09 %、本研究の SVM での全属性の予測率は 68.38 %、step.1 での属性部分集合の数は 254 で、step.1 での予測率は 68.57 %、近傍探索した結果の予測率は 68.38 %で全属性の 69.53 %より 0.19 %上昇した。H3K36me3 の場合、先行研究の予測率は 73.37 %、本研究の SVM での全属性の予測率は 75.09 %、step.1 での属性部分集合の数は 256 で、step.1 での予測率は 75.19 %、近傍探索した結果の予測率は 75.09 %で全属性の 69.53 %より 0.09 %上昇した。H3K79me3 の場合、先行研究の予測率は 79.91 %、本研究の SVM での全属性の予測率は 80.39 %、step.1 での属性部分集合の数は 244 で、step.1 での予測率は 80.58 %、近傍探索した結果の予測率は 80.39 %で全属性の 69.53 %より 0.19 %上昇した。 4 塩基での場合も近傍探索が作用し、 9 データで予測率が上がったの近傍探索の有効性が示せた。

5.4 属性選択手法の計算量による比較

属性数が n 個の場合、探索する属性の組合せは以下の通りである。

3 塩基の場合　4*4*4=64 の属性数

探索空間 $2^{64} = 18446744073709551616$ 個の組合せ数

4 塩基の場合　4*4*4*4=256 の属性数

探索空間 $2^{256} = 1.157920892373161954235709850086 9e + 77$

$=1.157920892373161954235709850086 9 * 10^{77}$ 個の組合せ数

5 塩基の場合　$4^5 = 1024$ の属性数

探索空間　$2^{1024} = 1.797693134862315907729305190789e + 308$ 個の組合せ数

6 塩基の場合　$4^6 = 4096$ の属性数

Dataset	予測性能(Pham)	全属性	step1での属性選択なしでの予測性能	step1での最高の予測性能をもつ予測	step2での最高の予測性能をもつ属性部分集合	step2での最高の属性部分集合の予測	step1とstep2との改良の合計
H3	85.88	86.43	AAAA(1),...,ACGC(243)	86.47	AAAA(1),...,GGGT(239),GACG,GGGC,GCGT,ACGC	86.47	0.04
H4	87.14	87.04	TATA(1),...,AGGG(244)	87.24	TATA(1),...,CCTA(240),CGCC,GGCG,AGGG,GGCC	87.32	0.29
H3K9ac	73.64	74.98	AATT(1),...,GACC(250)	75.07	AATT(1),...,ACCC(246),CGGG,GGAC,CGGA,GACG,TCCG,CCGA	75.08	0.10
H3K14ac	71.28	73.28	TATA(1),...,GGGG(256)	73.28	TATA(1),...,CCGG(252),CGGC,CGGG,GGGG	73.28	0.00
H4ac	69.93	72.06	AATT(1),...,CCCC(256)	72.06	AATT(1),...,GCGG(252),GCGC,CGGG,GGGG,CCCC	72.06	0.00
H3K4me1	68.29	69.53	TATA(1),...,CGGC(251)	69.64	TATA(1),...,CGCC(247),GCGG,CCGC,CGGC,CCGG,GGGG	69.71	0.18
H3K4me2	67.05	68.89	ATTT(1),...,CGGG(255)	68.89	ATTT(1),...,CCCG(251),GGGG,GCGC,CGGG,CGCG	68.97	0.08
H3K4me3	65.09	68.38	AATT(1),...,GCGC(254)	68.46	AATT(1),...,CCGG(250),CCCG,CCCC,CGGG,CGCG	68.57	0.19
H3K36me3	73.37	75.09	TATA(1),...,CGGG(256)	75.09	TATA(1),...,CGGC(252),CCGG,CCCC	75.19	0.09
H3K79me3	79.91	80.39	TATA(1),...,CGGG(244)	80.53	TATA(1),...,ACCC(240),CGGT,CCCC,CGGG,GGCG,TCCG,GACG	80.58	0.19

表 5.11: Pham による予測性能と全属性, ステップ 1, ステップ 2(k=4)

探索空間　$2^{4096} = 1.044388881413152506691752710716 6e + 1233$ 個の組合せ数

膨大な数値になる。属性選択のアルゴリズムの中で予測率が最も高い遺伝的アルゴリズム (genetic search) と Rank search であるが、時間計算量を比較すると遺伝的アルゴリズムは時間計算量が大きくかかるため sliding window の window サイズが大きくなると実用的な時間ではなくなる。Rank search は、filter 法であるが、スピードは速いが、n 通りの組合せしかためしていない。近傍探索は、時間計算量は、Genetic search より実用的であり、Rank search よりも広い探索空間を探索しているためより予測率の高い属性の部分集合を選択する可能性がある。また、計算機実験では、予測率も他の 2 つと同程度に高いため有用である。

5.5　まとめ

本研究では、ヌクレオソームの占有率、アセチレン化、メチル化の予測の正確さの改良を目指した。第 1 のステップで、属性の重要性は、randomForest の学習と予測を通して計算される。属性は MeanDecreaseGini の値を計算してランク付けする。属性のランキングの結果は、属性のなかにはデータのグループの中で選択的に重要なものがあることを示している。例 : アセチレンについての 3 データ。次に第 2 のステップで属性選択を行った。ステップ 1 は、ランキングに沿った最高の選択を探索する。すべての属性と比較して、より

良い予測性能をもつ属性の部分集合はこの属性選択によって発見される。属性の大部分は小さい重要性しかないことが分かる。しかし、SVM はできるだけそれらを有効化しようとする。最終的にはステップ 2 で、ステップ 1 の最高の属性の部分集合がテストされる。その結果、性能は再び改良される。ステップ 2 で最高の属性の部分集合が発見されたら、ステップ 2 では、補足的に実行されるかもしれない。同様の実験が k=4 の場合でも実行される。本研究の手法は依然として 256 の属性に対して効果的である。本研究の手法に対して、Weka を使って属性選択の比較実験を行った。6 つの判別器と 5 つの探索手法の全ての組合せがテストされ、本研究の手法は全て上回った。

第 6 章

寄与度からの知見

RandomForest では、決定木の分岐に評価基準として MeanDecreaseGini 係数を用いる。判別分析では、判別に寄与する数値 (variable importance) が算出され、定量評価できる。先行研究では、実際の多方面のデータに対しての応用例がある。この章では、さらに寄与度と SOM、クラスタリング、相関性などの解析と比較して関連を調べる。本研究は 2008 年に発表したものだが、現在 2011 年度までにも RandomForest の統計学の理論研究がなされており、改良を必要とする部分もある。

6.1 先行研究

RandomForest の variable importance を用いた研究は、初期には寄与度を提示するのみの研究が多かった。機械学習においては、学習アルゴリズムの研究が活発に行われているが、生物学や医学などの分野では、予測率よりも説明能力が求められることが多い。

一例として、医学的な実データでの例をとり説明する。[44] では、頭痛の診断補助装置を自己組織化マップ（Self Organizing Map(SOM)）を用いて作成している。この研究では、、頭痛の主訴をもつ患者 208 人に対して、100 項目の頭痛に関する定性的な問診をし、その各々の項目を特徴ベクトルとし、Random Forest 法の Mean Decrease Gini 係数を用いて重み付けを行い、SOM を作成している。結果として診断に関する多次元の情報が、SOM により 2 次元に有意な形で投影されている。また、その判別には主に 10 程度の問診内容が寄与していることが示唆された。結論としては、自己組織化マップを用いて、頭痛の診断補助装置を作成している。10 程度のある・なしの定性的問診情報のみで、6 割程度現場の医師と同じ機能性頭痛の判別を行うことが可能であることが示唆された、とされている。この博士論文の評価を引用する。

「本論文は自己組織化マップ (Self Organizing Map:SOM) を用いて頭痛の診断支援を行

うシステムの構築と評価について述べている。診断支援における人工知能（AI）の応用にはShortliff の MYCIN 以来長い歴史があり、ニューラルネットワークを用いた研究も多数知られている。また頭痛は鑑別診断が問診を主体にして進められる代表的症状であることから、AI による診断支援の対象としてしばしば用いられてきた。今回は Mean Decrease GiniIndex(MIGI) を採用した Random Forest(RF) 法で 100 の質問項目に重みづけをし、SOM を作成している。その結果として正診率は約 70 であることを示した。頭痛の診断自身がこうした問診項目のみで 100 %正しく診断できるわけではなく、また現在の診断が最終的な病態生理診断と言えるかどうかも疑問である中で、こうしたシステムに高い正診率だけが求められるわけではない。むしろここで重要なのは診断理由が論理的に示されること（Black box を作らないこと）、可視化すること、ルール変更による改善が容易なこと、入力情報の信頼性を高める工夫があることなどが上げられる。本手法ではとくに可視化することでその状況を把握するだけでなく、相互の疾患間の関係を理解するのに有用であり、またどの因子が特に影響があるのかを逆に解析して新知見を得ることができることが示された。このことは今後他領域における応用にも期待が持てるものであり、AI を用いた診断支援として高く評価できるものであって、学位審査にあたり価値ある論文であると認める。」

このように、予測率ばかりではなく、その診断の説明力が重視されている例もある。

RandomForest の先行研究として 2 種類がある。実際に現実のデータに適用した先行研究と、variable importance の bias とその改良についての理論研究である。先行研究では、RandomForest の説明能力が注目されている。実際に適用されている分野は、医学、土木など多方面にわたっている。[28] では、車の衝突事故での運転者の回避行動を分析に RandomForest を用いている。ドライバーの視認性の有無、ドライバーの身体障害、ドライバーの気晴らし行動などの要因に対しての判別分析を行い、その重要度を variable importance でランク付けしている。[29] では、渡り鳥の冬の行動の分析に適用している。2005 年から 2006 年ベーリング海での衛星発信機で、95 羽の渡り鳥に発信機を付け冬期での行動を観察した。渡り鳥が移動する環境要因として日照時間、場所、海氷密接度、生息地を要因として RandomForest を用い判別分析を行い、要因の重要度を variable importance によってランク付けしている。結果として個体差が大きいこと、場所、日付、及び海氷密接度も順に重要な要因であることが分かった。[30] では、歯科の分野について適用し、歯を損失した臨床的な要因に RandomForest 相対的な重要性のランク付けしている。[31] では、間接リュウマチにおける自己申告うつ病患者に対して、うつ病を発症する臨床的に有用な予測因子を特定しランク付けに使われている。要因として、人口統計、臨床及び治療データ、家計

所得、雇用と労働の状態、併存疾患、痛み、疲労、局所疼痛スケール (PRS)、症状の強度 (SI)、健康状態質問票をしている。結論として痛みの程度と疲労が自己申告うつ病患者の予測因子となっている。[31] は、細菌を用いて突然変異を検出する Ames 試験の QSAR（定量的構造活性相関、化学物質の構造と生物学的（薬学的あるいは毒性学的）な活性との間になりたつ量的関係のこと。) に対してその局所的な部分に対して RandomForest を用いて要因をランク付けしている。[32] では、関節リューマチのデータに対してロジステック回帰分析と RandomForest を持ちた解析を行った。その結果、ロジステック回帰分析では、高いランクに位置付けられていないが重要な遺伝子である個人差を示す遺伝子である 1 塩基多型 (SNIP) と遺伝子を randomFortest では特定している。[33] では、同じアンサンブル学習のアルゴリズムである boosting と RandomForest の折衷的なアルゴリズムを提案しており、創薬のデータに対して解析をしている。また、属性のランキングとクラスタリングを使った、診断ツールを提案している。[34] では、抗レトロウイルス療法を開始する HIV-1 患者に CD4T 細胞の回復を予測する実験を行い、木構造ベースの解析として CARTs(分類と回帰)、RandomForest(RFs)、ロジステック回帰 (LR) の 3 種類の比較を行った。さらに分割表を用いた実験をも行っている。RandomForest と分割は、予測因子のの潜在的な重要性を示しているが CARTs と LR は、変数の組合せについての情報を示唆している。3 種類の木構造ベースの解析では、ともに CD3-DR-CD56+CD16 という因子を共通して重要としている。結論として、木構造ベースの解析は、このフローサイトメトリー (flow cytometry) の実験で、潜在的な予測因子の発見についての情報が提供できており、単変量分析では発見できない関連を発見しているとされる。[35] では、直腸癌を切除したあとのエピジェネティクス現象を示すメチル化の分析を行っている。これは癌の再発に関しての有益な情報を提供する。実験では、直腸癌の直腸間全切除した 325 人の患者からメチル化された部分の定量評価をし、RandomForest を用いてクラスタリングを行った。結果として、属性の重要度 variable importance によって、患者を 4 つのグループにクラスタリングされ同定された。患者の 73 %を示すクラスタには再発のリスクが高い傾向をしめす徴候があった。このグループは放射線治療を受けた患者よりも高い局所再発の可能性を示した。

2008 年位から randomForest の寄与度 (variable importance) に関する統計的な理論研究が出始めてきた。Strobl らは、Gini index についての variable　imporatnce は、生成する決定木のノード数やサンプルサイズなどで bias が生じること、説明変数に相関する属性がある場合なども bias が生じることを示し、これを修正するために permutation importance、conditional variable importance などを提案した [39],[40],[41],[42],[43]。R のパッケージとし

てはRandomForestがあるが、更にStroblらの研究成果を実装したparty、RandomForestを基に関連する全属性を表示するBoruta[45]などがある。

6.2 寄与度とSOMの関連

SOMは、ニューラルネットワークの１つであり、可視化に優れている。先に説明した決定木とSOMを組み合わせることにより、可視化と定量評価が行える。SOMの統計的応用として

・データ可視化

・連想、想起などの情報処理

・大規模データの要約

・非線形モデルの作成

があげられる。SOMのソフトはViscovery SOMine5.2を用いた。(http://www.mindware-jp.com/somine/index.html)

同社のサイトには、SOMの利点として下記のことがあげられている。

1. パターンと支配変数の識別
2. 複雑な従属性 (関係性) の識別
3. 変数選択と重要度
4. マップの最適化
5. ポジショニング (ラベリング)
6. クラスタリング

この章では、前章で行ったデータ解析で予測率の高いデータH3、H4と予測率率が低いデータH3K4me3を例にとり説明をする。全データの予測率での順位は、

$H4(86.67\%) > H3(86.45\%) > H3K79me3(79.94\%) > H3K36me3(73.80\%) > H3K9ac(72.90\%) > H3K14ac(70.36\%) > H4ac(69.32\%) > H3K4me2(68.28\%) > H3K4me1(67.56\%) > H3K4me3(65.92\%)$ である。

図6.1はH4、図6.2はH3、図6.3はH3K79me3のSOMによる解析結果である。彩色されているが赤い方が数値が高く、青いほど低い。SOM表示でのLABELを見ると赤い色と青い色に明確に分かれている場合と全体にに散布している場合がある。H3,H4は、明確に分かれておりSOMによるクラスタリングが有効である場合である。H3K4me3は、SOMによるクラスタリングが有効に作用していない場合である。H4について、上位のランキング

の特に ATA、TAT は LABEL と反対の色相をしている。これは負の相関を示している。つまり、LABEL の赤い色は、メチル化されており、その部分が不活性化を示している。ATA と TAT は活性化している部位といえる。逆に LABEL と同じ色相は AGG、GCC、CGT、CCT が部分的に色相が類似している。相関係数との関係をみると、相関係数の高い属性は、類似した色相を示す。H3 の場合、LABEL の色相は明瞭である。右半分と左半分で、負例と正例で判別されている。寄与度のランキング上位でも負例の方に高い頻度を示す属性が多い。

結論としては、RandomForest のランキングの上位の属性と LABEL は負の属性を示している。これは、LABEL との相関係数を比較で示されている。予測率が高いデータ程 SOM でのクラスタリングでも有用な属性つまり相関関係が明瞭な数値を取る。

H4 の寄与度と SOM の比較

H4 の予測率は 10 データのうちでも最も高く 86.64 % である。図 6.3 の寄与度のグラフを見ると急こう配からなだらかな勾配となっているが裾野が薄い。次の H3 のグラフでも同様であるが、予測率が高いデータのグラフのなだらかな勾配になった時の裾野が薄い。この部分は、判別分析において正例と負例の境界領域だと推測される。寄与度が 0.5 以上の属性は、8 属性ある。逆に 0.2 以下の属性は、34 属性ある。相関係数が 0.5 以上の属性は 13 組ある。TCT と TTC (0.6301)、CTT と TTC (0.6208)、AGA と AAG (0.6181)、CTT と TCT (0.6103)、GAA と AAG (0.6101)、GAA と AGA (0.6055)、TAA と AAT (0.5706)、TTA と ATT (0.5543)、TAT と ATA (0.5543)、CTC と TCT (0.5448)、CCT と TCC (0.5262)、GAG と AGA (0.5113)、図 6.1 の SOM でのクラスタリングで、LABEL の表示では、右上と左下に頻度の高い赤い部分が現れている。頻度の低い部分は青い部分である。クラスは 2 クラスである。黒い線が 2 クラスの境界線である。右下の部分が頻度が低い部分がある。この部分で頻度が高い属性は、AGG(53)、CCG(21)、GCC(62)、GGC(60) である。これらの属性の属性数の最大値は、順番に AGG が 12、CCG が 11、GCC が 10、GGC が 11 と低い。左上で頻度の高い属性は、AAA(2)、AAT(36)、ATA(4)、ATT(33)、TAT(3) である。右上で頻度の高い属性は、AGG(53)、TGA(15)、GAA(28)、GAC(43)、GGA(17)、GAT(11) である。左下で頻度の高い属性は、ATC(7)、TTC(30)、TCT(31)、CTC(51) である。LABEL において右上の左下の属性の集合は、メチル化され不活性を示す属性の集合である。左上で頻度の高い属性は、負例でメチル化されていない活性化している属性の集合である。RandomForest のランキングでも上位の属性が多い。右上は、不活性化されているが、ラ

ンキングで上位のものがTGA、GGA、GATとある。左下も不活性化されている属性であるが、ランキング上位はATCがある。まとめとしては、予測率の高い寄与度のグラフは、急勾配からなだらかな裾野の部分が薄いグラフが予測率が高いと推測される。ランキング上位には、正例で重要な属性と負例で重要な属性が混じっている。

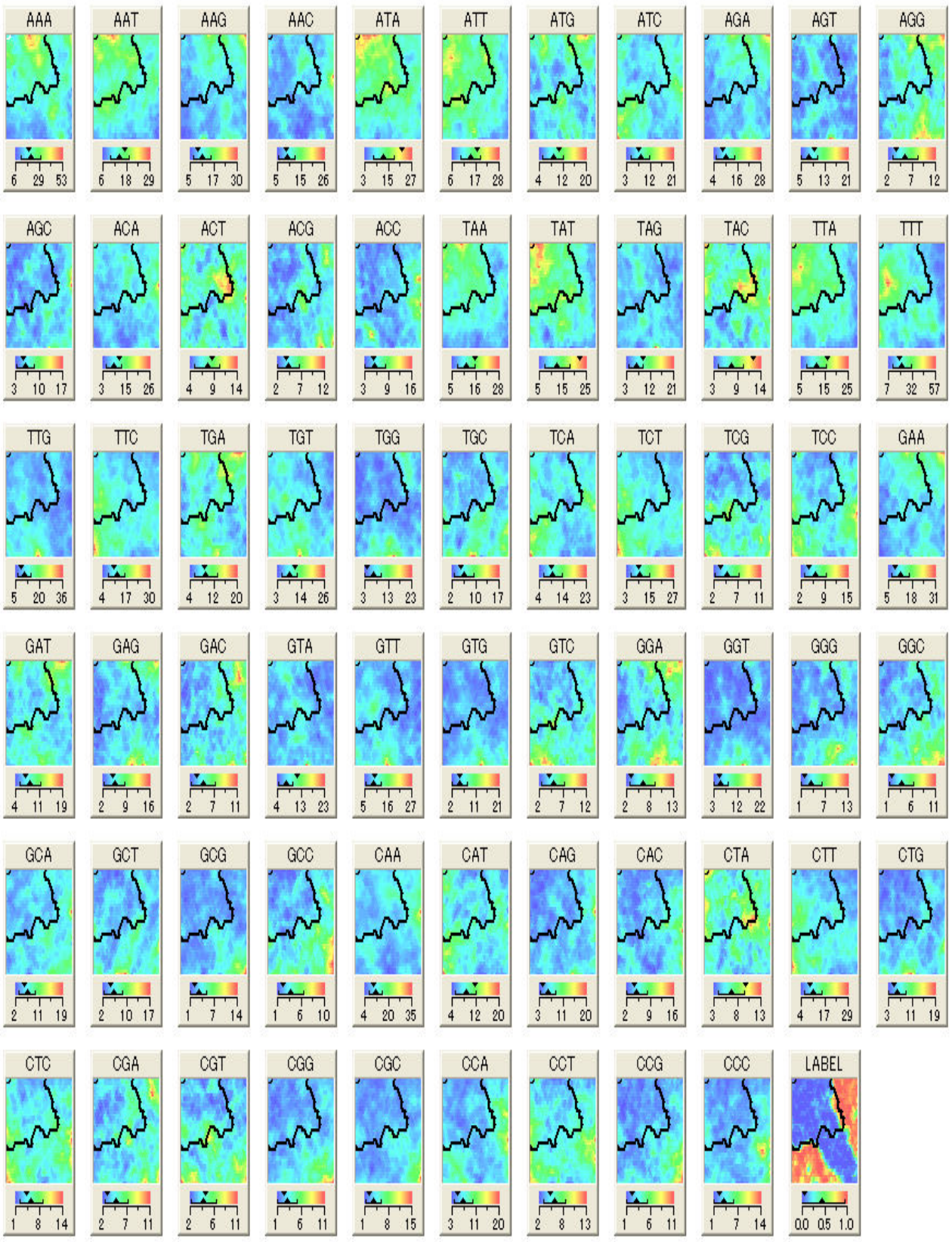


図 6.1: H4 の SOM 表示

RFの寄与度

TTT	263.976	1	GGT	51.41878	0.194786
AAA	239.0813	0.905693	ATT	50.98936	0.193159
TAT	233.403	0.884183	CAT	50.33339	0.190674
ATA	224.7352	0.851347	AAT	50.13733	0.189931
CCA	204.6619	0.775305	ACA	48.4919	0.183698
TAA	160.8464	0.609322	GTC	47.81805	0.181145
ATC	141.634	0.536541	AAC	47.43477	0.179694
TGG	137.0678	0.519243	ATG	46.65185	0.176728
CAA	130.2156	0.493286	GTT	46.61299	0.17658
TTA	126.0297	0.477429	GCA	46.47442	0.176055
GAT	117.2072	0.444007	GAC	44.4637	0.168438
TTG	109.9524	0.416524	GAG	43.72756	0.16565
GCG	105.5663	0.399909	CCT	43.7032	0.165557
TCA	101.5153	0.384563	TGC	43.62825	0.165274
TGA	92.66913	0.351051	ACT	43.34414	0.164197
CGC	89.39406	0.338645	AGC	42.81947	0.16221
GGA	70.62932	0.26756	TCG	42.64064	0.161532
TCC	70.53683	0.267209	ACG	42.28057	0.160168
CAG	63.25153	0.239611	CTC	42.00756	0.159134
CTT	59.62013	0.225854	TAG	41.58379	0.157529
CCG	59.44996	0.22521	AGG	41.56476	0.157457
AAG	56.82367	0.215261	CCC	41.52716	0.157314
CTG	56.33918	0.213425	CGT	41.16442	0.15594
GTA	55.90299	0.211773	AGT	40.89853	0.154933
AGA	55.63098	0.210743	CGA	40.41557	0.153103
TGT	54.81781	0.207662	CAC	40.30006	0.152666
CGG	54.49368	0.206434	CTA	40.06741	0.151784
GAA	53.78869	0.203764	GGC	39.90794	0.15118
ACC	53.30441	0.201929	GCT	39.87078	0.151039
TTC	53.18167	0.201464	GCC	39.12456	0.148213
TCT	52.26912	0.198007	GTG	38.45976	0.145694
TAC	51.67835	0.195769	GGG	35.43484	0.134235

相関係数

属性 1	属性 2	相関係数
TCT	TTC	0.6301
TTC	TCT	0.6301
CTT	TTC	0.6208
TTC	CTT	0.6208
AGA	AAG	0.6181
AAG	AGA	0.6181
CTT	TCT	0.6103
TCT	CTT	0.6103
GAA	AAG	0.6101
AAG	GAA	0.6101
GAA	AGA	0.6055
AGA	GAA	0.6055
TAA	AAT	0.5706
AAT	TAA	0.5706
TTA	ATT	0.5543
ATT	TTA	0.5543
TAT	ATA	0.5473
ATA	TAT	0.5473
CTC	TCT	0.5448
TCT	CTC	0.5448
CCT	TCC	0.5262
TCC	CCT	0.5262
GAG	AGA	0.5113
AGA	GAG	0.5113
GGA	AGG	0.5039
AGG	GGA	0.5039
GAT	TGA	0.4691
TGA	GAT	0.4691
CCA	ACC	0.4565
ACC	CCA	0.4565
TAA	ATA	0.4533

相関係数

属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	CAA	0.2038	LABEL	GCC	0.0389
LABEL	CCA	0.1848	LABEL	AAG	0.0281
LABEL	ACT	0.1817	LABEL	TCG	0.0179
LABEL	AAC	0.1509	LABEL	GGC	0.0115
LABEL	GAT	0.1441	LABEL	CCC	0.0094
LABEL	CGA	0.1277	LABEL	CCT	0.0057
LABEL	TGA	0.1256	LABEL	GGT	0.0039
LABEL	TGG	0.1234	LABEL	AGG	0.0007
LABEL	GAC	0.122	LABEL	CCG	-0.0025
LABEL	ACC	0.1199	LABEL	CGG	-0.0031
LABEL	CTG	0.1138	LABEL	GGG	-0.0157
LABEL	GGA	0.1069	LABEL	ATT	-0.0206
LABEL	TAC	0.1062	LABEL	CAT	-0.022
LABEL	CAG	0.0998	LABEL	CGC	-0.0222
LABEL	GAA	0.0951	LABEL	AAT	-0.0263
LABEL	CTA	0.093	LABEL	GCG	-0.0467
LABEL	ACA	0.0806	LABEL	AGT	-0.0488
LABEL	ACG	0.076	LABEL	GTG	-0.0518
LABEL	CAC	0.0739	LABEL	CGT	-0.0522
LABEL	AGA	0.0681	LABEL	CTT	-0.0558
LABEL	TCA	0.0629	LABEL	TCT	-0.0624
LABEL	AGC	0.0586	LABEL	TTC	-0.0638
LABEL	ATC	0.0585	LABEL	GTT	-0.0707
LABEL	GCA	0.0548	LABEL	TAG	-0.0743
LABEL	TGC	0.0507	LABEL	AAA	-0.088
LABEL	TCC	0.05	LABEL	TTA	-0.1102
LABEL	GCT	0.0488	LABEL	TGT	-0.1171
LABEL	GTC	0.0485	LABEL	GTA	-0.1316
LABEL	GAG	0.0484	LABEL	ATA	-0.1529
LABEL	TTG	0.047	LABEL	TAA	-0.1541
LABEL	ATG	0.0457	LABEL	TAT	-0.1607
LABEL	CTC	0.0405	LABEL	TTT	-0.2116

図 6.2: H4 の寄与度 (VI)、相関係数、LABEL との相関係数

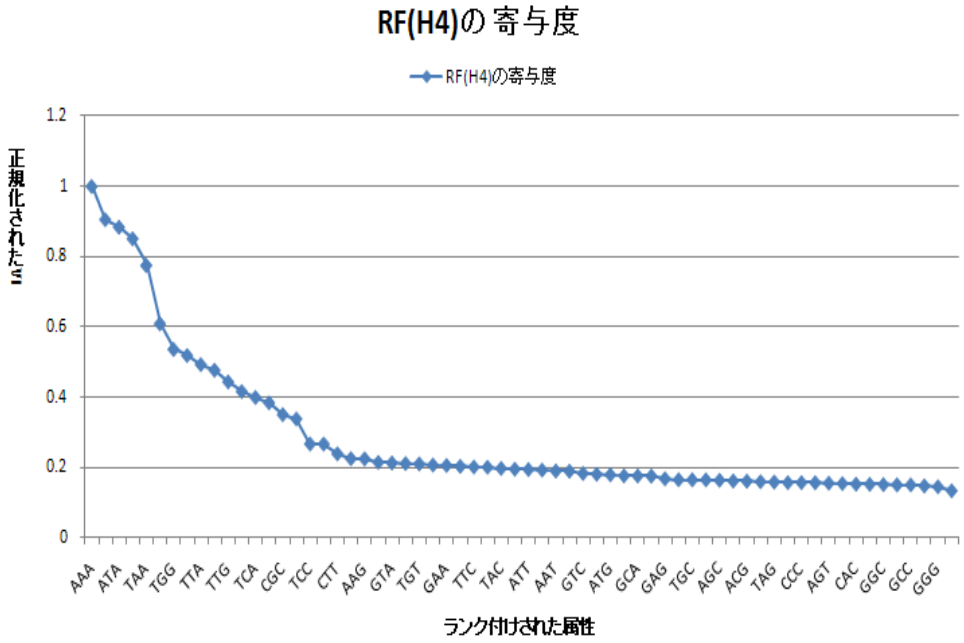


図 6.3: H4 の寄与度

H3 の寄与度と SOM の比較

H3 の予測率は 10 データのうちで 2 番目に高く 86.19 %である。図 6. の寄与度のグラフを見ると急こう配からなだらかな勾配となっているが裾野が薄い。H4 のグラフでも同様であるが、予測率が高いデータのグラフのなだらかな勾配になった時の裾野が薄い。この部分は、判別分析において正例と負例の境界領域だと推測される。相関係数をみると、0.5 以上のものが H4 よりも多い。() 内は相関係数。

GAA と AAG (0.705) 、AGA と AAG (0.6832) 、GAA と AGA (0.6827) 、CTT と TCC (0.6976) 、CTT と TTC (0.6785) 、TCT と TTC (0.6698) 、CCA と ACC (0.6048) 、GGT と TGG (0.6001) 、CAA と AAC (0.574) 、GTT と TTG (0.5601) 、TAA と AAT (0.5589) 、GAG と AGA (0.5573) 、TAT と ATA (0.5497) 、CTC と TCT (0.5466) 、CCT と TCC (0.5446) 、TTA と ATT (0.5435) 、図 6.4 の SOM のグラフでは、右半分と左半分で色相が分かれている。2 クラスの実線もその境界線に沿っている。LABEL の図において左半分のメチル化されていない不活性化されている部分の属性は、AAG(25)、AGA(34)、AGC(36)、ACA(46)、ACG(58)、TGA(30)、TTC(37)、GAA(32)、GAT(23)、GAC(50)、GGA(24)、GCC(13)、CCA(8)、CCT(21)、CGA(45)、() 内はランキング順位である。逆に右半分の

活性化されている部分の属性は、AAT(7)、ATA(5)、TAA(3)、TAT(6)、TTT(1) である。右半分の活性化されている部分の属性はランキング上位の属性が多いが、左半分の不活性化されている部分の属性もランキングの上位にあるものがあるので、H3 の場合でも正例、負例の重要な属性が混じっている。

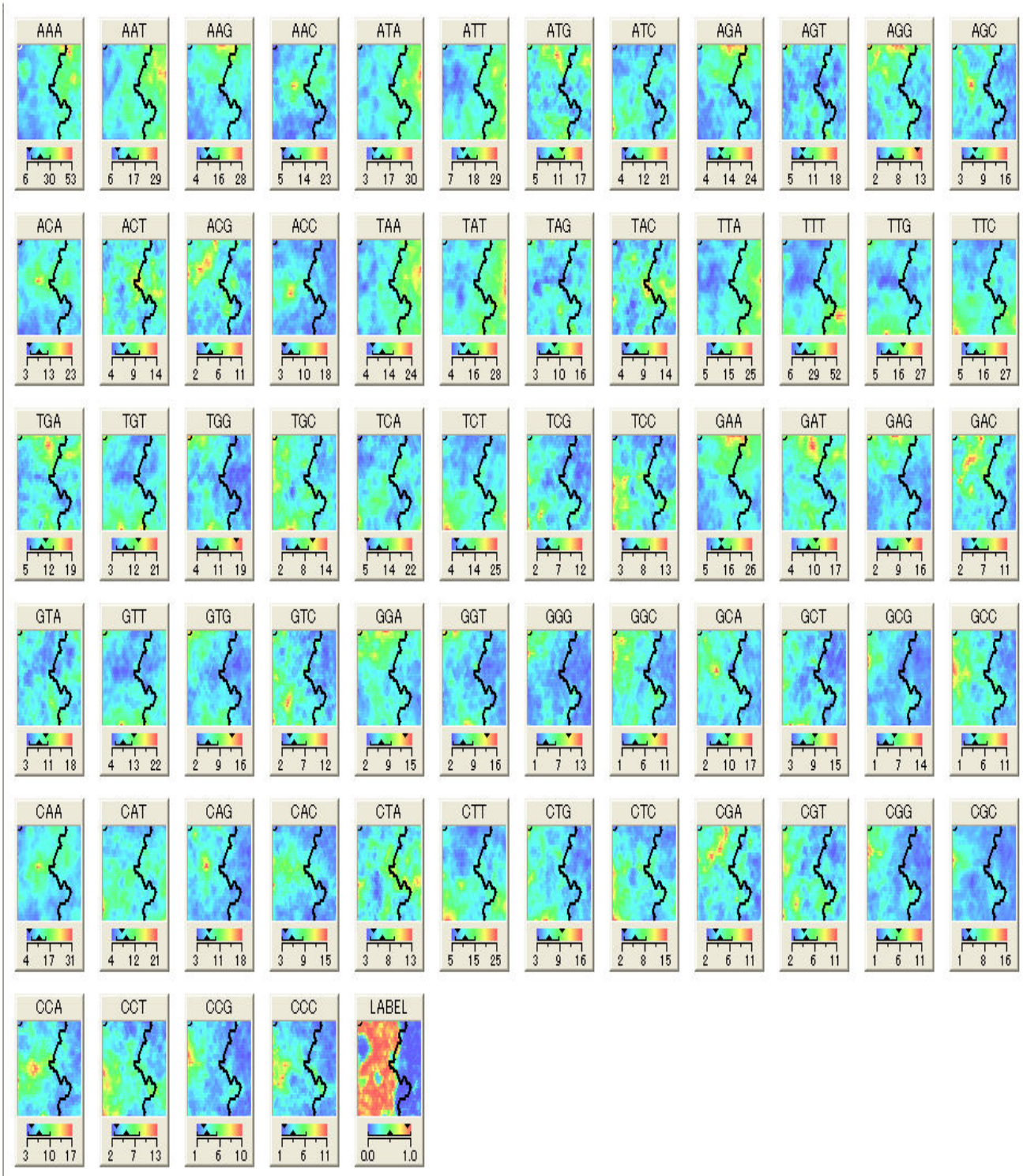


図 6.4: H3 の SOM 表示

RFの寄与度

TTT	311.4368	1	CAA	52.90703	0.16988
AAA	305.9853	0.982496	AGA	52.53166	0.168675
TAA	293.784	0.943318	TCC	51.87378	0.166563
TTA	281.8441	0.90498	AGC	51.4077	0.165066
ATA	204.1833	0.655617	TTC	50.963	0.163638
TAT	169.6143	0.544619	CCC	50.46207	0.16203
AAT	138.9121	0.446036	GGT	50.161	0.161063
CCA	126.6322	0.406606	TCT	49.19733	0.157969
ATT	116.9933	0.375657	AGG	49.05985	0.157527
TGG	103.6879	0.332934	ACC	48.89757	0.157006
CAG	91.94483	0.295228	CTC	48.00605	0.154144
CTG	89.50783	0.287403	CAT	47.41496	0.152246
GCC	79.61648	0.255643	CGA	46.94874	0.150749
GGC	77.2308	0.247982	ACA	46.30274	0.148675
CGC	74.5646	0.239421	GGG	46.21079	0.148379
GCG	73.18387	0.234988	GTG	46.17113	0.148252
TGC	72.86627	0.233968	GAG	46.16076	0.148219
TTG	72.40253	0.232479	GAC	46.0973	0.148015
GCA	69.07127	0.221783	TGT	46.0377	0.147824
ATC	65.8092	0.211308	ATG	46.00569	0.147721
CCT	60.56116	0.194457	TCG	44.6495	0.143366
CGG	60.34142	0.193752	GTT	44.08424	0.141551
GAT	59.37766	0.190657	CAC	43.97569	0.141203
GGA	59.23912	0.190212	ACT	42.88331	0.137695
AAG	58.76568	0.188692	AGT	42.70572	0.137125
TCA	57.73196	0.185373	AAC	42.06115	0.135055
CCG	56.41992	0.18116	TAC	41.73572	0.13401
GCT	55.9902	0.17978	ACG	41.07941	0.131903
GTC	55.43377	0.177994	TAG	41.05484	0.131824
TGA	54.05054	0.173552	GTA	40.80815	0.131032
CTT	53.57272	0.172018	CTA	39.8361	0.127911
GAA	53.30172	0.171148	CGT	38.94543	0.125051

相関係数

属性 1	属性 2	相関係数
GAA	AAG	0.705
AAG	GAA	0.705
AGA	AAG	0.6832
AAG	AGA	0.6832
GAA	AGA	0.6827
AGA	GAA	0.6827
CTT	TCT	0.6796
TCT	CTT	0.6796
CTT	TTC	0.6785
TTC	CTT	0.6785
TCT	TTC	0.6698
TTC	TCT	0.6698
CCA	ACC	0.6048
ACC	CCA	0.6048
GGT	TGG	0.6001
TGG	GGT	0.6001
CAA	AAC	0.574
AAC	CAA	0.574
GTT	TTG	0.5601
TTG	GTT	0.5601
TAA	AAT	0.5589
AAT	TAA	0.5589
GAG	AGA	0.5537
AGA	GAG	0.5537
TAT	ATA	0.5497
ATA	TAT	0.5497
CTC	TCT	0.5466
TCT	CTC	0.5466
CCT	TCC	0.5446
TCC	CCT	0.5446
TTA	ATT	0.5435
ATT	TTA	0.5435
GGA	AGG	0.5427
AGG	GGA	0.5427
GTT	TGT	0.5269
TGT	GTT	0.5269
ACA	AAC	0.5222
AAC	ACA	0.5222
TCA	ATC	0.5143
ATC	TCA	0.5143
GAT	TGA	0.5099
TGA	GAT	0.5099
CAA	ACA	0.5092
ACA	CAA	0.5092
TGT	TTG	0.4957

LABELとの相関係数

属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	ATC	0.198	LABEL	ACA	0.0245
LABEL	TCA	0.1747	LABEL	CGA	0.0196
LABEL	CCA	0.1704	LABEL	TTC	0.012
LABEL	CAG	0.1505	LABEL	AGG	0.0081
LABEL	TGG	0.1487	LABEL	TAG	0.008
LABEL	GAT	0.1474	LABEL	TGT	0.0058
LABEL	CTG	0.1321	LABEL	GTT	0.0032
LABEL	TGA	0.131	LABEL	GCC	0.0021
LABEL	CAT	0.1306	LABEL	CTA	0.0012
LABEL	TTG	0.1285	LABEL	GTG	0.001
LABEL	CAA	0.114	LABEL	TAC	0
LABEL	TCC	0.1028	LABEL	GTA	-0.0148
LABEL	GTC	0.0971	LABEL	ATT	-0.0314
LABEL	CCT	0.0951	LABEL	AAT	-0.032
LABEL	ACT	0.0908	LABEL	CCC	-0.0349
LABEL	ATG	0.0791	LABEL	GGC	-0.0394
LABEL	AGT	0.0788	LABEL	CGT	-0.0476
LABEL	GAC	0.076	LABEL	ACG	-0.0549
LABEL	CTC	0.0747	LABEL	AGA	-0.0575
LABEL	ACC	0.0659	LABEL	GGG	-0.069
LABEL	TGC	0.0659	LABEL	GAA	-0.0699
LABEL	GGA	0.0589	LABEL	AAG	-0.0925
LABEL	GCA	0.0508	LABEL	TAT	-0.1001
LABEL	AGC	0.0463	LABEL	ATA	-0.1017
LABEL	TCT	0.0432	LABEL	TTA	-0.15
LABEL	CTT	0.0407	LABEL	CCG	-0.1612
LABEL	AAC	0.0387	LABEL	CGC	-0.1616
LABEL	GCT	0.036	LABEL	TAA	-0.1646
LABEL	TCG	0.031	LABEL	CGG	-0.1681
LABEL	GGT	0.0264	LABEL	GCG	-0.172
LABEL	CAC	0.0262	LABEL	TTT	-0.2413
LABEL	GAG	0.0251	LABEL	AAA	-0.2632

図 6.5: H3 の寄与度 (VI)、相関係数、LABEL との相関係数

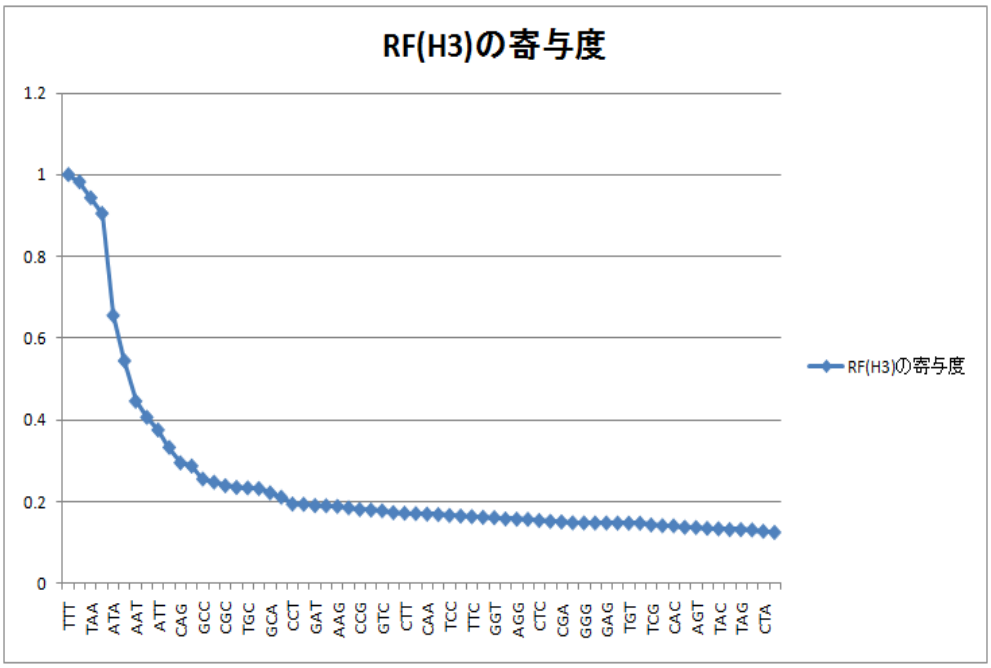


図 6.6: H3 の寄与度

H3K79me3 の寄与度と SOM の比較

H3K79me3 の予測率は 10 データのうちで 3 番目に高く 79.56 %である。図 6.9 のグラフの特徴的な点は、2 属性ほど寄与度が 1 に近い属性があり、次に緩やかな勾配が 2 段ある点である。2 段目の勾配は寄与度 0.2 に近く比較的裾野は薄いタイプである。相関係数に着目すると、高い相関を示す属性が複数ある。

GAA と AAG (0.7119) 、GAA と AGA (0.7083) 、AGA と AAG (0.7046) 、TCT と TTC (0.6292) 、CTT と TCT (0,6232) 、CTT と TTC (0.622) 、GAG と AGA (0.6038) 、CCA と ACC (0.5907) 、TTA と ATT (0.5856) 、TAT と ATA (0.5783) 、GGA と AGG (0.5767) 、CAA と AAC (0.5393) 、GAT と TGA (0.5308) 、TTA と TAT (0.5265) 、CCT と TCC (0.5161) 、TAA と ATT (0.5142) 、となっている。図 6.7 の SOM のクラスタリングの図では、正例の赤い部分と負例の青い部分に明瞭に分かれているが 2 クラスの境界線は、図を上下に 2 分している。正例に属する属性は、ATG、AGA、ACG、TGA、GAA、GGA、CTA、CGA、CCT である。負例に属する属性は、AAA、AAT、ATA、ATT、TAA、TAT、TTA、CTA である。

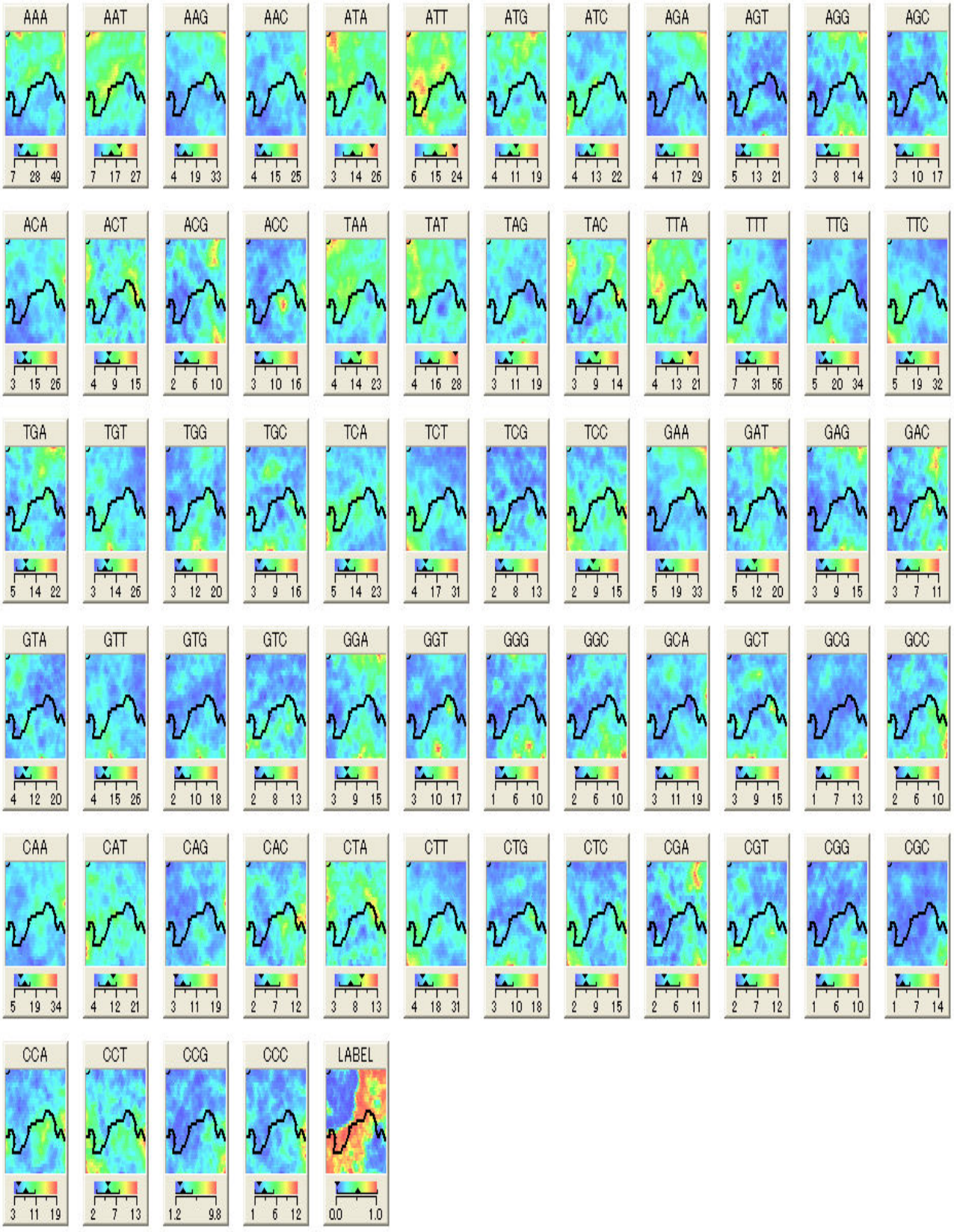


図 6.7: H3K79me3 の SOM 表示

RFの寄与度		
ATA	443.7917	1
TAT	430.9104	0.970975
ATC	253.1508	0.570427
CAA	247.3213	0.557291
CGC	234.7542	0.528974
TTG	229.5167	0.517172
TTT	219.7638	0.495196
TGA	215.097	0.48468
GCG	209.8933	0.472955
TCA	207.8618	0.468377
AAA	206.8259	0.466043
CCA	205.5231	0.463107
TGG	198.8833	0.448146
GAT	195.1829	0.439808
GAA	181.4495	0.408862
TTC	178.3807	0.401947
TCT	175.8172	0.396171
AGA	170.5105	0.384213
TAA	169.6988	0.382384
TTA	163.7568	0.368995
TAC	158.8202	0.357871
CTT	153.6762	0.34628
AAG	151.9397	0.342367
GTA	148.2156	0.333976
ATT	141.0893	0.317918
AAT	140.3256	0.316197
ACA	128.2135	0.288905
GTT	126.4189	0.284861
GGA	124.974	0.281605
TGT	123.6164	0.278546
GCA	120.0073	0.270414
AAC	119.8447	0.270047

相関係数		
属性 1	属性 2	相関係数
GAA	AAG	0.7119
AAG	GAA	0.7119
GAA	AGA	0.7083
AGA	GAA	0.7083
AGA	AAG	0.7046
AAG	AGA	0.7046
TCT	TTC	0.6292
TTC	TCT	0.6292
CTT	TCT	0.6232
TCT	CTT	0.6232
CTT	TTC	0.622
TTC	CTT	0.622
GAG	AGA	0.6038
AGA	GAG	0.6038
CCA	ACC	0.5907
ACC	CCA	0.5907
TTA	ATT	0.5856
ATT	TTA	0.5856
TAT	ATA	0.5783
ATA	TAT	0.5783
GGA	AGG	0.5767
AGG	GGA	0.5767
CAA	AAC	0.5359
AAC	CAA	0.5359
GAT	TGA	0.5308
TGA	GAT	0.5308
TTA	TAT	0.5265
TAT	TTA	0.5265
CCT	TCC	0.5161
TCC	CCT	0.5161
TAA	AAT	0.5142

相関係数					
属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	CAA	0.2754	LABEL	CTC	0.0078
LABEL	GAA	0.2717	LABEL	TCG	0.0036
LABEL	GAT	0.2588	LABEL	GCC	-0.0032
LABEL	TGA	0.2559	LABEL	CGG	-0.0066
LABEL	AGA	0.2503	LABEL	AAA	-0.0095
LABEL	GGA	0.2334	LABEL	GCT	-0.013
LABEL	AAG	0.2081	LABEL	ACT	-0.0194
LABEL	CCA	0.188	LABEL	CAC	-0.0195
LABEL	GAC	0.185	LABEL	CCC	-0.0232
LABEL	GAG	0.1752	LABEL	GTT	-0.0466
LABEL	TGG	0.1699	LABEL	GTG	-0.05
LABEL	CGA	0.1616	LABEL	CTA	-0.0502
LABEL	CAG	0.1569	LABEL	TTC	-0.0561
LABEL	AAC	0.1559	LABEL	CCT	-0.0604
LABEL	ATC	0.1492	LABEL	AGT	-0.0613
LABEL	AGG	0.1406	LABEL	ATT	-0.0648
LABEL	TCA	0.1217	LABEL	TCT	-0.0823
LABEL	ACC	0.0952	LABEL	TGC	-0.084
LABEL	ATG	0.0792	LABEL	CGT	-0.0988
LABEL	TCC	0.0788	LABEL	CGC	-0.112
LABEL	ACG	0.0681	LABEL	CAT	-0.1128
LABEL	AGC	0.0673	LABEL	GCG	-0.1173
LABEL	CTG	0.0561	LABEL	TAG	-0.1403
LABEL	TTG	0.0545	LABEL	CTT	-0.1464
LABEL	GGT	0.0527	LABEL	TAC	-0.1481
LABEL	ACA	0.0413	LABEL	TGT	-0.2083
LABEL	GTC	0.0222	LABEL	TAA	-0.2297
LABEL	GGC	0.0198	LABEL	TTA	-0.2406
LABEL	GCA	0.0159	LABEL	GTA	-0.2528
LABEL	GGG	0.0104	LABEL	TTT	-0.2662
LABEL	CCG	0.0099	LABEL	ATA	-0.339
LABEL	AAT	0.0097	LABEL	TAT	-0.354

図 6.8: H3K79me3 の寄与度 (VI)、相関係数、LABEL との相関係数

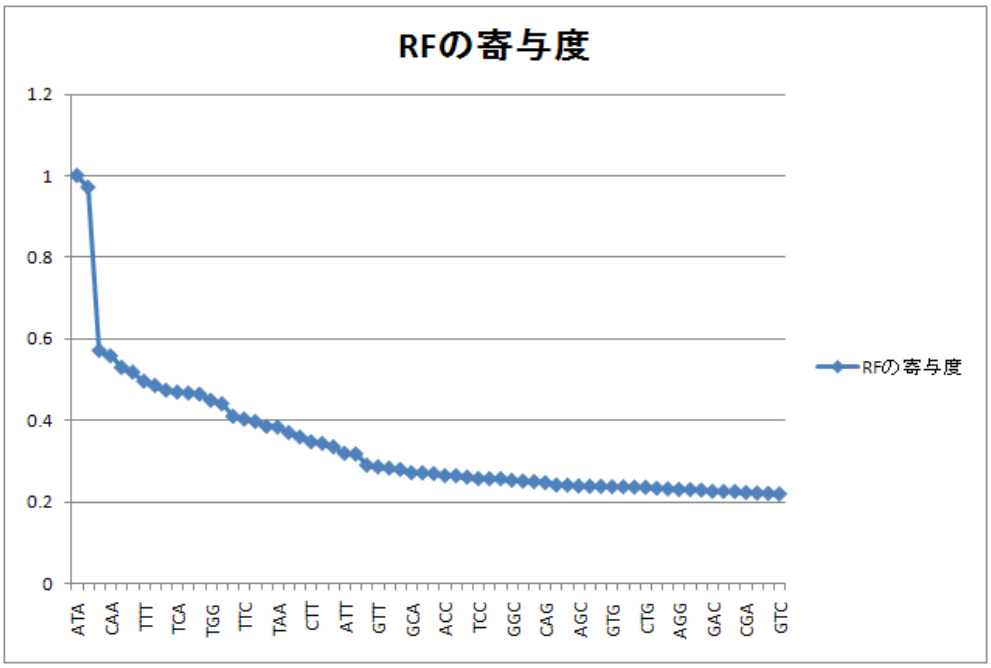


図 6.9: H3K79me3 の寄与度

H3K36me3 の寄与度と SOM の比較

H3K36me3 の予測率は 10 データのうちで 4 番目に高く 73.80 %である。図 6.12 の寄与度のグラフの特徴は、全体的になだらかで裾野が厚いことである。裾野の部分は寄与度が 0.6 から 0.4 の範囲にある。相関係数をみると 0.5 以上の属性が多くある。

GAA と AAG (0.6929) 、AGA と AAG (0.6784) 、GAA と AGA (0.6754) 、CCA と ACC (0.6454) 、CTT と TCT (0.6045) 、CTT と TTC (0.6045) 、TAA と AAT (0.591) 、TTA と ATT (0.5895) 、CAA と AAC (0.5768) 、TCT と TTC (0.5738) 、GGT と TGG (0.5706) 、TAA と ATA (0.5501) 、CCT と TCC (0.5385) 、GAT と TGA (0.5341) 、TAT と ATA (0.5326) 、CAA と ACA (0.5241) 、である。図 6.10 の SOM の図では、赤い色の正例の属性と青い負例の属性、２クラスの境界線が正例の部分を横断している。正例では、AAA、AATATG、AGG、GCA、CCA が特に明瞭である。負例は、ATA、TAT が特に顕著である。

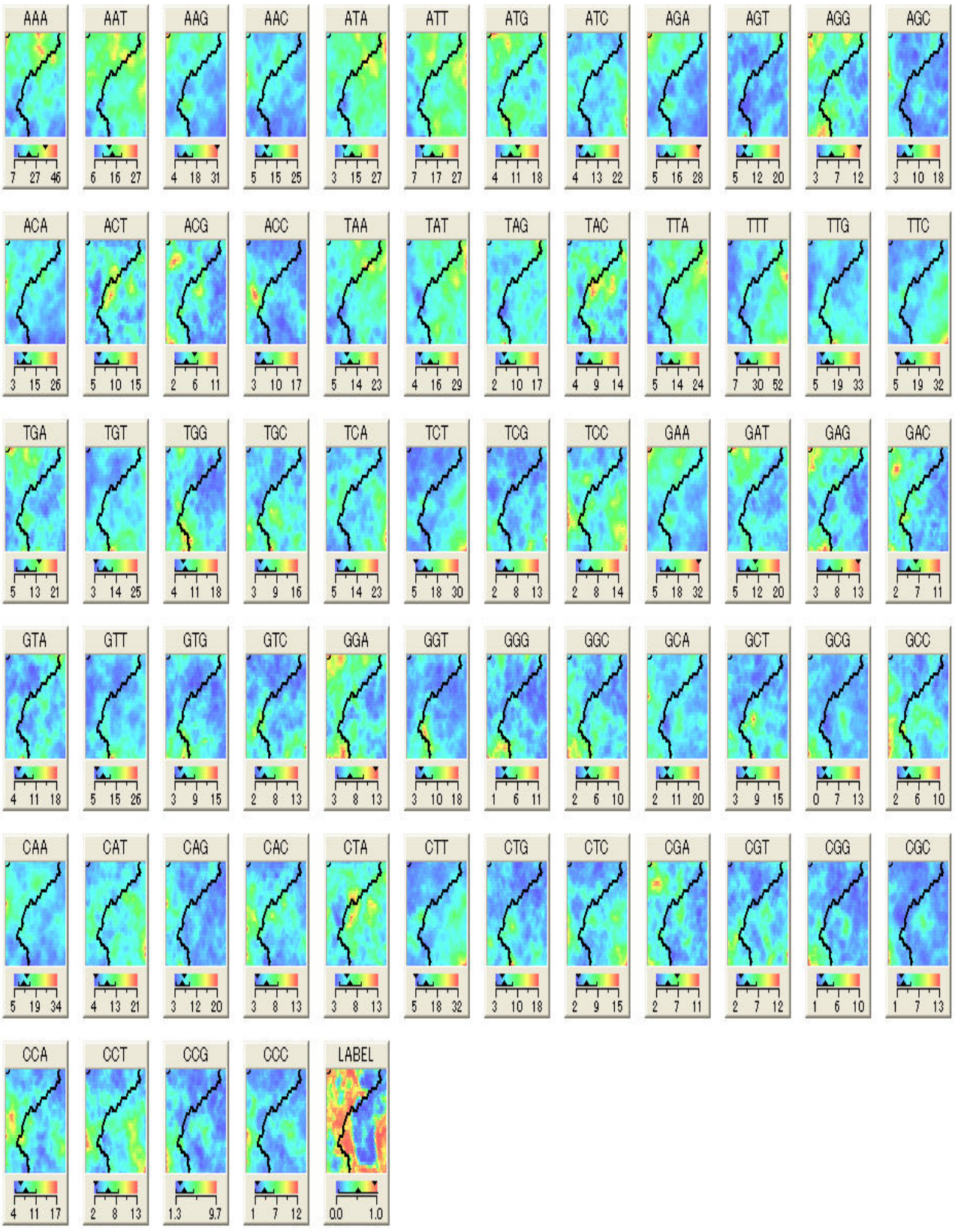


図 6.10: H3K36me3 の SOM 表示

RFの寄与度		
CCA	325.2978	1
TGG	312.289	0.96001
ATA	300.3533	0.923318
TAT	288.0474	0.885488
CAA	251.8409	0.774186
AAA	245.5457	0.754834
TTT	241.3317	0.741879
TTG	228.5074	0.702456
ATC	224.3046	0.689536
GAA	217.7165	0.669284
GAT	215.28	0.661794
TGA	211.5334	0.650276
TCA	209.3119	0.643447
TTC	200.1808	0.615377
AGA	198.1873	0.609249
AAT	198.1641	0.609177
TAA	192.2052	0.590859
AAG	190.6174	0.585978
TCT	189.0793	0.58125
ATT	188.5602	0.579654
CTT	187.3218	0.575847
TTA	186.4064	0.573033
GGT	185.9556	0.571647
ACC	184.469	0.567077
CGC	183.854	0.565187
GTA	180.7161	0.55554
TAC	178.8372	0.549765
GCG	175.7284	0.540208
ACA	168.6301	0.518387
TGT	168.1483	0.516906
GTT	162.3068	0.498948
AGT	160.7488	0.494159

相関係数		
属性 1	属性 2	相関係数
GAA	AAG	0.6929
AAG	GAA	0.6929
AGA	AAG	0.6784
AAG	AGA	0.6784
GAA	AGA	0.6754
AGA	GAA	0.6754
CCA	ACC	0.6475
ACC	CCA	0.6475
CTT	TCT	0.6045
TCT	CTT	0.6045
CTT	TTC	0.5933
TTC	CTT	0.5933
TAA	AAT	0.591
AAT	TAA	0.591
TTA	ATT	0.5895
ATT	TTA	0.5895
CAA	AAC	0.5768
AAC	CAA	0.5768
TCT	TTC	0.5738
TTC	TCT	0.5738
GGT	TGG	0.5706
TGG	GGT	0.5706
TAA	ATA	0.5501
ATA	TAA	0.5501
CCT	TCC	0.5385
TCC	CCT	0.5385
GAT	TGA	0.5341
TGA	GAT	0.5341
TAT	ATA	0.5326
ATA	TAT	0.5326
CAA	ACA	0.5241

相関係数					
属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	CAA	0.1466	LABEL	TAC	-0.0179
LABEL	CCA	0.1288	LABEL	GAG	-0.0192
LABEL	ACC	0.1287	LABEL	CTT	-0.02
LABEL	TGA	0.1233	LABEL	CCT	-0.0251
LABEL	GAT	0.1222	LABEL	CTG	-0.0274
LABEL	AAC	0.104	LABEL	CCG	-0.0277
LABEL	AAT	0.1026	LABEL	TTA	-0.0278
LABEL	ATC	0.1008	LABEL	TAA	-0.0378
LABEL	TGG	0.0983	LABEL	GTG	-0.039
LABEL	TCA	0.0908	LABEL	CTC	-0.04
LABEL	GAA	0.0862	LABEL	AAA	-0.0441
LABEL	ATT	0.0811	LABEL	CCC	-0.0522
LABEL	GAC	0.0752	LABEL	AGC	-0.0546
LABEL	AGA	0.0703	LABEL	TGT	-0.0569
LABEL	ATG	0.0699	LABEL	CGT	-0.0613
LABEL	TTG	0.0675	LABEL	GCC	-0.0618
LABEL	GGA	0.0398	LABEL	TCG	-0.0618
LABEL	TTC	0.0309	LABEL	AGT	-0.0648
LABEL	ACA	0.0261	LABEL	AGG	-0.0703
LABEL	TCC	0.0228	LABEL	GCA	-0.0717
LABEL	CGA	0.0209	LABEL	CGG	-0.0739
LABEL	AAG	0.0156	LABEL	GGC	-0.0758
LABEL	TCT	0.0154	LABEL	TAG	-0.0821
LABEL	GGT	0.0142	LABEL	ATA	-0.0909
LABEL	CAT	0.0047	LABEL	TTT	-0.0947
LABEL	GTT	0.0021	LABEL	GCT	-0.0978
LABEL	CAG	0.0016	LABEL	TAT	-0.1027
LABEL	CTA	0.0007	LABEL	TGC	-0.1096
LABEL	ACT	-0.0014	LABEL	GGG	-0.1122
LABEL	ACG	-0.007	LABEL	GTA	-0.1191
LABEL	CAC	-0.0098	LABEL	CGC	-0.1255
LABEL	GTC	-0.0118	LABEL	GCG	-0.1402

図 6.11: H3K36me3 の寄与度 (VI)、相関係数、LABEL との相関係数

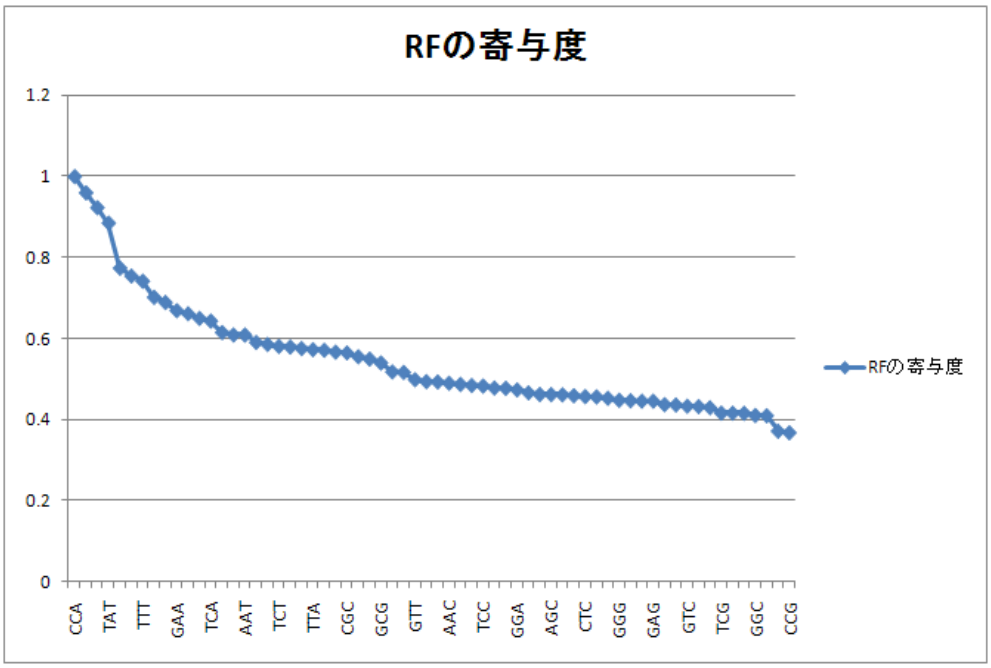


図 6.12: H3K36me3 の寄与度

H3K9ac の寄与度と SOM の比較

H3K9ac の予測率は 10 データのうちで 5 番目に高く 72.68 %である。図 6.15 の寄与度のグラフの特徴は、2 属性が寄与度 1 に近く、急勾配からなだらか勾配の部分は、寄与度 0.4 から 0.2 の範囲にある。このグラフの特徴は寄与度の高い属性が少なく、そして、寄与度が比較的少ないが、その裾野が横に長い点である。

TTC と TCT (0.6902) 、TCT と CTT (0.6863) 、TTC と CTT (0.6774) 、TCT と CTC (0.5809) 、AAG と GAA (0.5752) 、AAG と AGA (0.5727) 、AGA と GAA (0.5668) 、TGG と GGT (0.5613) 、AAT と TAA (0.557) 、ATC と TCA (0.5453) 、ATA と TAT (0.5334) 、TCC と CCT (0.5272) 、ATA と TAA (0.4989) 、AGA と GAG (0.4864) 、TCA と CAT (0.4766) 、AGG と GGA (0.4722) 、となっている。図 6.13 の SOM の図では、正例と負例の属性は分かれている。2 クラスの境界線もほぼ正例と負例を分割している。正例に属している属性は、AGG、GAC、GTC、GGA、CGA である。負例に属している属性は、ATA、ATG、TAA、TAC、TTA、TGA である。

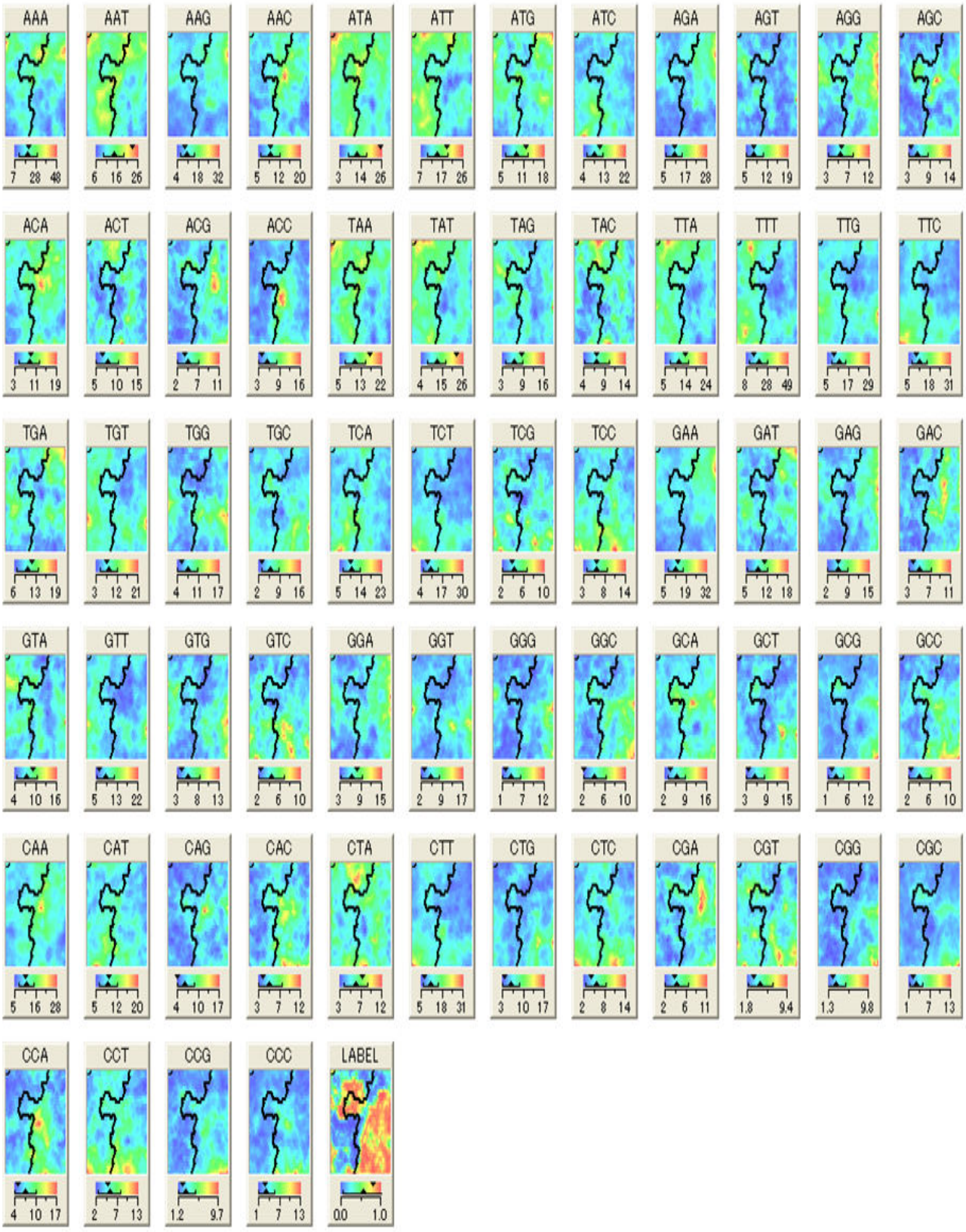


図 6.13: H3K9ac の SOM 表示

RFでの寄与度		
ATT	416.2088	1
AAT	397.4384	0.954901
GCG	203.5539	0.489067
CGC	202.7459	0.487125
TTA	199.2506	0.478728
TAA	176.3764	0.423769
GGC	167.3869	0.40217
TTT	161.1426	0.387168
AAA	159.5532	0.383349
ATC	155.6334	0.373931
TCA	152.5141	0.366437
ATA	151.2461	0.36339
TGA	151.136	0.363125
TTG	150.3472	0.36123
TAT	148.1047	0.355842
GAT	146.6283	0.352295
AGC	146.5562	0.352122
CAA	142.5438	0.342481
GCC	142.2971	0.341889
CGG	141.8337	0.340775
GCT	137.6655	0.330761
TGC	137.4579	0.330262
CAT	137.3216	0.329934
GCA	134.8448	0.323984
ATG	133.8937	0.321698
GAA	132.2596	0.317772
TTC	132.1632	0.317541
CCA	129.2773	0.310607
CTT	128.1367	0.307866
GTT	127.8576	0.307196
AAG	127.8152	0.307094
AGA	127.7178	0.30686

相関係数		
属性 1	属性 2	相関係数
TTC	TCT	0.6902
TCT	TTC	0.6902
TCT	CTT	0.6863
CTT	TCT	0.6863
TTC	CTT	0.6774
CTT	TTC	0.6774
TCT	CTC	0.5809
CTC	TCT	0.5809
AAG	GAA	0.5752
GAA	AAG	0.5752
AAG	AGA	0.5727
AGA	AAG	0.5727
AGA	GAA	0.5668
GAA	AGA	0.5668
TGG	GGT	0.5613
GGT	TGG	0.5613
AAT	TAA	0.557
TAA	AAT	0.557
ATC	TCA	0.5453
TCA	ATC	0.5453
ATA	TAT	0.5334
TAT	ATA	0.5334
TCC	CCT	0.5272
CCT	TCC	0.5272
ATA	TAA	0.4989
TAA	ATA	0.4989
AGA	GAG	0.4864
GAG	AGA	0.4864
TCA	CAT	0.4766
CAT	TCA	0.4766
AGG	GGA	0.4722

相関係数					
属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	CGC	0.0969	LABEL	CAC	0.0204
LABEL	GCG	0.0948	LABEL	CCT	0.019
LABEL	TAC	0.0936	LABEL	TGC	0.0177
LABEL	ACG	0.0858	LABEL	TAA	0.0133
LABEL	CTA	0.0853	LABEL	GTT	0.0124
LABEL	GTA	0.0756	LABEL	GCA	0.0115
LABEL	AGT	0.0733	LABEL	GGG	0.0114
LABEL	GCT	0.0691	LABEL	GCC	0.0087
LABEL	GTG	0.0667	LABEL	CTG	0.0027
LABEL	CGA	0.063	LABEL	GAA	-0.0013
LABEL	TAG	0.0627	LABEL	GTC	-0.0046
LABEL	AAG	0.0585	LABEL	CAG	-0.0073
LABEL	CGT	0.0574	LABEL	TTA	-0.0091
LABEL	AGC	0.0546	LABEL	CCC	-0.0112
LABEL	ACT	0.0533	LABEL	AAC	-0.0131
LABEL	GAG	0.0532	LABEL	TTC	-0.0225
LABEL	CGG	0.051	LABEL	GGA	-0.024
LABEL	AAA	0.0471	LABEL	GGT	-0.0258
LABEL	TCG	0.0452	LABEL	ACC	-0.0375
LABEL	TGT	0.0441	LABEL	ATG	-0.0439
LABEL	ATA	0.0387	LABEL	TGA	-0.0735
LABEL	CCG	0.0369	LABEL	TCC	-0.0783
LABEL	CTC	0.033	LABEL	GAT	-0.0829
LABEL	AGG	0.0327	LABEL	TTG	-0.0842
LABEL	TTT	0.0295	LABEL	TGG	-0.0864
LABEL	CTT	0.029	LABEL	CAT	-0.1022
LABEL	ACA	0.028	LABEL	CAA	-0.1178
LABEL	AGA	0.0254	LABEL	TCA	-0.1234
LABEL	TAT	0.0234	LABEL	AAT	-0.1258
LABEL	TCT	0.0214	LABEL	ATC	-0.1282
LABEL	GGC	0.0208	LABEL	CCA	-0.1432
LABEL	GAC	0.0207	LABEL	ATT	-0.1527

図 6.14: H3K9ac の寄与度 (VI)、相関係数、LABEL との相関係数

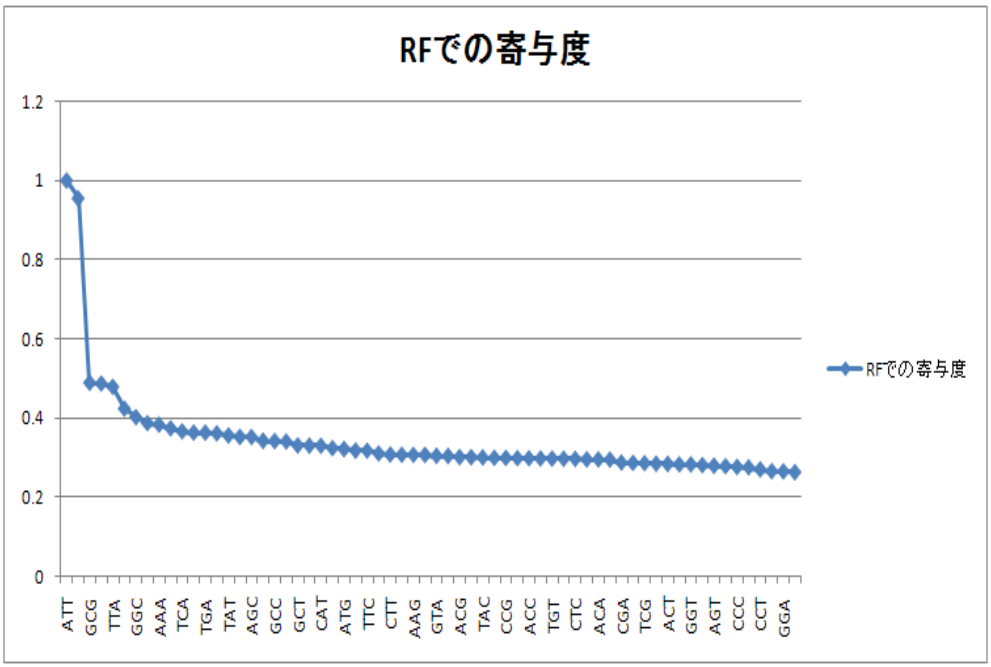


図 6.15: H3K9ac の寄与度

H3K14ac の寄与度と SOM の比較

H3K4ac の予測率は 10 データのうちで 6 番目に高く 69.98 %である。図 6.18 の寄与度のグラフの特徴は、なだらかな部分つまり裾野が厚いことである。寄与度が 0.6 から 0.4 の範囲内に、44 属性が属している。相関係数は、
GAA と AAG (0.6634)、AGA と AAG (0.6579)、GAA と AGA (0.6512)、TTA と ATT (0.6204)、CCA と ACC (0.6147)、TAT と ATA (0.6114)、TAA と AAT (0.5797)、TTA と TAT (0.55) 、CTT と TCT (0.5444)、CTT と TTC (0.5377)、TAA と ATA (0.5374)、GGT と TGG (0.5348)、TCT と TTC (0.5319)、GAG と AGA (0.5265)、GGA と AGG (0.523) 、CAA と AAC (0.5229)、図 6.16 の SOM では、正例を示す赤い部分が大半を占めており、負例を示す部分が少ない。負例の部分では、ATT、ATG、GTA が属している。正例では、ATA、CTA があげれらる。

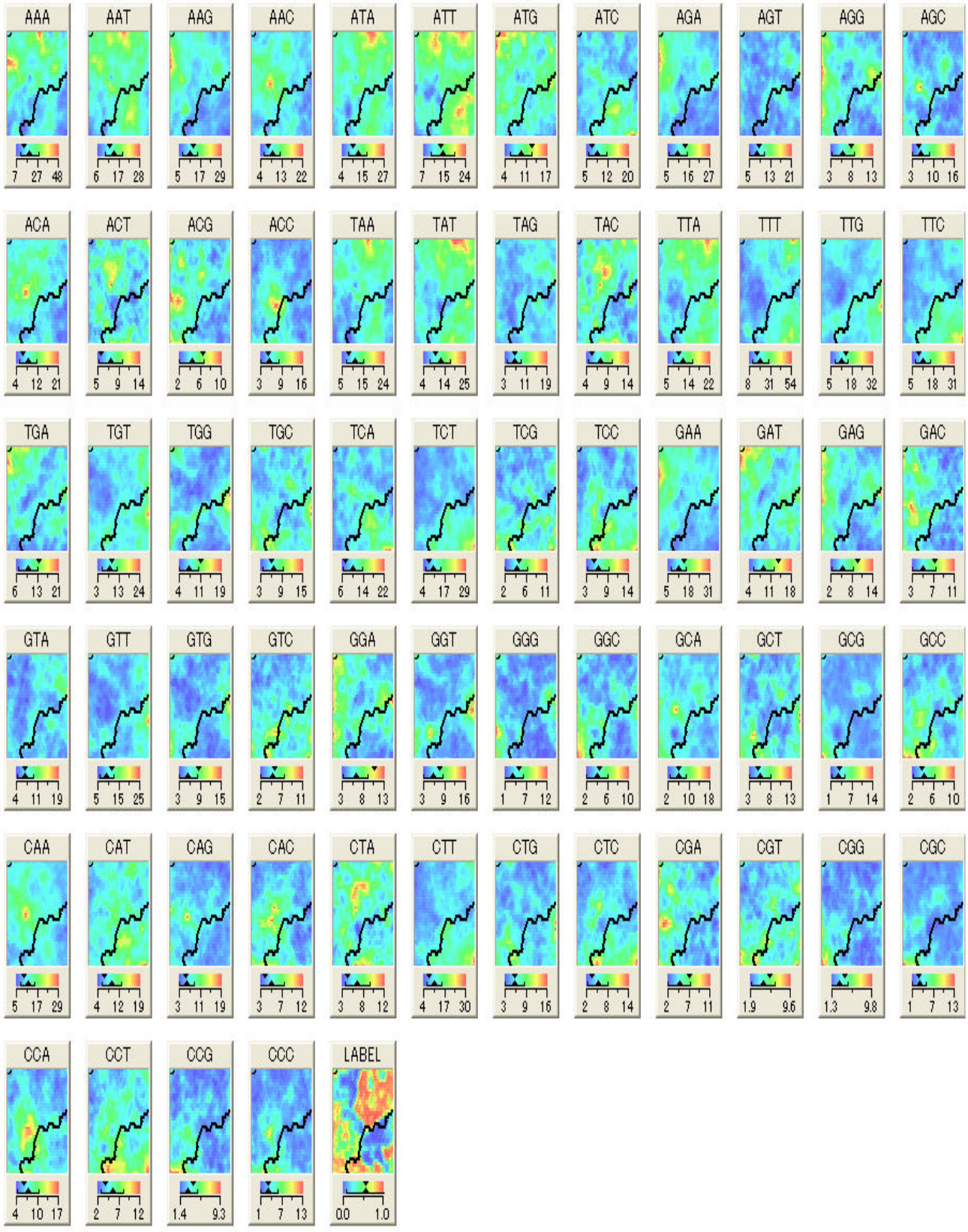


図 6.16: H3K14ac の SOM 表示

RFの寄与度		
AAT	288.9311	1
ATT	265.4506	0.918733
CAA	241.0859	0.834406
TTG	227.6648	0.787955
TGA	225.05	0.778905
CCA	223.018	0.771873
TCA	222.2487	0.76921
ATC	220.1801	0.762051
AAA	216.8368	0.750479
TTT	211.7438	0.732852
GAT	209.8048	0.726141
TGG	205.7409	0.712076
ATA	203.4191	0.70404
TAT	197.8229	0.684672
GAA	179.5229	0.621335
CAT	177.1933	0.613272
TTC	175.9651	0.609021
TAA	174.8981	0.605328
AAG	173.1493	0.599275
ATG	171.2897	0.592839
GTA	171.2417	0.592673
TTA	169.9904	0.588342
AGA	168.4909	0.583153
GTT	168.3082	0.58252
CTT	166.0171	0.574591
TCT	164.8501	0.570552
TGT	162.8551	0.563647
AAC	162.5057	0.562438
TAC	162.2704	0.561623
AGT	161.589	0.559265
GGT	161.4501	0.558784
ACA	158.5601	0.548782

相関係数		
属性1	属性2	相関係数
GAA	AAG	0.6634
AAG	GAA	0.6634
AGA	AAG	0.6579
AAG	AGA	0.6579
GAA	AGA	0.6512
AGA	GAA	0.6512
TTA	ATT	0.6204
ATT	TTA	0.6204
CCA	ACC	0.6147
ACC	CCA	0.6147
TAT	ATA	0.6114
ATA	TAT	0.6114
TAA	AAT	0.5797
AAT	TAA	0.5797
TTA	TAT	0.55
TAT	TTA	0.55
CTT	TCT	0.5444
TCT	CTT	0.5444
CTT	TTC	0.5377
TTC	CTT	0.5377
TAA	ATA	0.5374
ATA	TAA	0.5374
GGT	TGG	0.5348
TGG	GGT	0.5348
TCT	TTC	0.5319
TTC	TCT	0.5319
GAG	AGA	0.5265
AGA	GAG	0.5265
GGA	AGG	0.523
AGG	GGA	0.523
CAA	AAC	0.5229

相関係数					
属性1	属性2	相関係数	属性1	属性2	相関係数
LABEL	GCG	0.1375	LABEL	CAC	0.0146
LABEL	TGT	0.1369	LABEL	GAG	0.0121
LABEL	GTA	0.136	LABEL	GGT	0.0119
LABEL	CGT	0.1288	LABEL	AGG	0.0073
LABEL	TTT	0.1187	LABEL	TTA	-0.0035
LABEL	CGC	0.1135	LABEL	CTA	-0.008
LABEL	CTT	0.1093	LABEL	CGA	-0.0119
LABEL	TCG	0.1006	LABEL	TCC	-0.012
LABEL	TAG	0.1	LABEL	TTG	-0.0157
LABEL	TCT	0.0992	LABEL	CAT	-0.0159
LABEL	GTG	0.093	LABEL	CAG	-0.0194
LABEL	CGG	0.083	LABEL	TAC	-0.0224
LABEL	AGT	0.0821	LABEL	AAA	-0.0242
LABEL	GCT	0.0798	LABEL	ACT	-0.0285
LABEL	GTC	0.0719	LABEL	AAG	-0.0518
LABEL	TGC	0.0712	LABEL	GAC	-0.0579
LABEL	GTT	0.0664	LABEL	GGA	-0.0606
LABEL	TTC	0.0621	LABEL	TGG	-0.0638
LABEL	CTC	0.0593	LABEL	ACA	-0.0738
LABEL	CCG	0.0565	LABEL	TCA	-0.08
LABEL	AGC	0.0564	LABEL	ACC	-0.0806
LABEL	TAT	0.056	LABEL	AGA	-0.0858
LABEL	GGC	0.0558	LABEL	ATC	-0.0867
LABEL	CCT	0.0521	LABEL	ATT	-0.1032
LABEL	GGG	0.0495	LABEL	AAC	-0.1099
LABEL	ATA	0.0479	LABEL	ATG	-0.1196
LABEL	GCA	0.0459	LABEL	CCA	-0.1245
LABEL	GCC	0.0383	LABEL	GAA	-0.1309
LABEL	TAA	0.0266	LABEL	AAT	-0.1383
LABEL	CTG	0.0247	LABEL	GAT	-0.1433
LABEL	CCC	0.0228	LABEL	TGA	-0.1637
LABEL	ACG	0.0163	LABEL	CAA	-0.1789

図 6.17: H3K14ac の寄与度 (VI)、相関係数、LABEL との相関係数

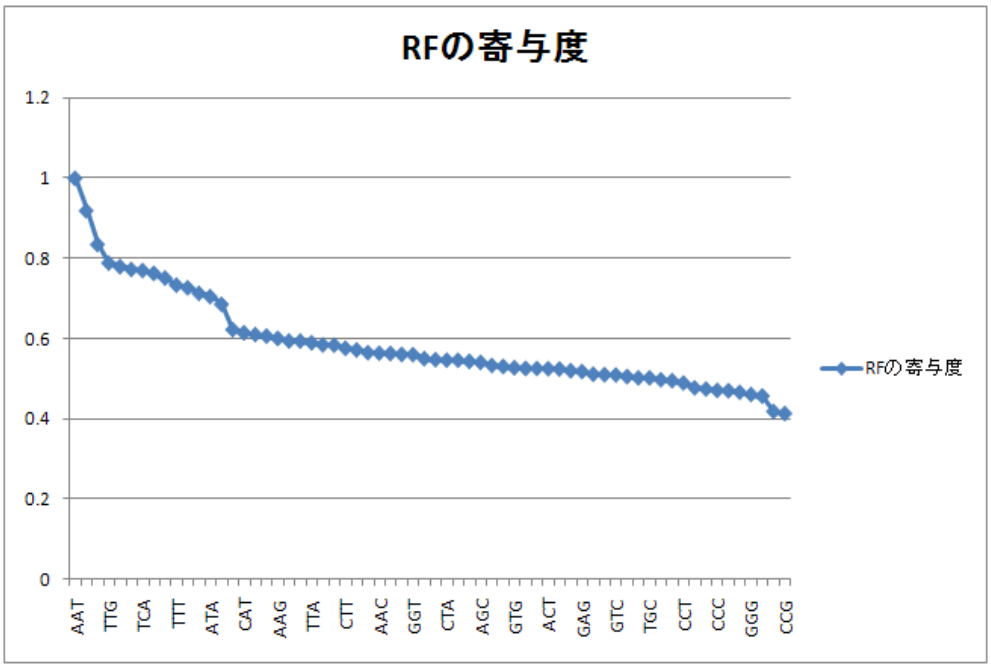


図 6.18: H3K14ac の寄与度

H4ac の寄与度と SOM の比較

H4ac の予測率は 10 データのうちで 7 番目に高く 69.26 %である。図 6.2 の寄与度のグラフの特徴は、横に長い裾野の長さである。寄与度は、0.4 から 0.3 までの範囲に入っている。相関係数は、0.5 以上は比較的少ない。

CTT と TCT (0.671)、CTT と TTC (0.6623)、TCT と TTC (0.6553)、TAT と ATA (0.6147)、TAA と AAT (0.5936)、CTC と TCT (0.5664)、CCT と TCC (0.5492)、TCA と ATC (0.5332)、TTA と ATT (0.5114)、GGT と TGG (0.5094)、TAA と ATA (0.5078)、AGA と AAG (0.5024)、である。図 6.19 の SOM の図では、LABEL は赤い正例の部分が図の大半を占めている。正例では、TAC、CTA があげられる。負例では、AAA、ATT があげられる。

図 6.19: H4ac の SOM 表示

RFの寄与度

ATT	430.1257	1
AAT	406.8147	0.945804
TTT	217.4483	0.505546
AAA	209.9822	0.488188
CAA	206.6947	0.480545
TTA	197.8625	0.460011
TAT	197.4562	0.459066
TGA	196.7159	0.457345
ATC	192.3805	0.447266
CCA	191.7656	0.445836
TAA	191.6185	0.445494
TTG	191.3347	0.444834
ATA	190.5608	0.443035
CGC	184.0592	0.42792
TCA	184.0286	0.427848
GCT	179.5191	0.417364
GCG	177.3555	0.412334
GAT	177.3078	0.412223
AGC	176.4735	0.410284
GAG	172.4527	0.400936
TGG	172.2257	0.400408
ATG	172.1989	0.400346
CTT	171.9069	0.399667
TCT	171.6458	0.39906
GTA	171.2637	0.398171
CAT	171.1614	0.397933
AAG	171.0198	0.397604
AGT	170.8692	0.397254
GCA	170.8522	0.397215
GTT	170.6082	0.396647
TTC	169.8639	0.394917
AGA	169.4662	0.393992

相関係数

属性 1	属性 2	相関係数
CTT	TCT	0.671
TCT	CTT	0.671
CTT	TTC	0.6623
TTC	CTT	0.6623
TCT	TTC	0.6553
TTC	TCT	0.6553
TAT	ATA	0.6147
ATA	TAT	0.6147
TAA	AAT	0.5936
AAT	TAA	0.5936
CTC	TCT	0.5664
TCT	CTC	0.5664
CCT	TCC	0.5492
TCC	CCT	0.5492
TCA	ATC	0.5332
ATC	TCA	0.5332
TTA	ATT	0.5114
ATT	TTA	0.5114
GGT	TGG	0.5094
TGG	GGT	0.5094
TAA	ATA	0.5078
ATA	TAA	0.5078
AGA	AAG	0.5024
AAG	AGA	0.5024
GAA	AAG	0.4999
AAG	GAA	0.4999
GAA	AGA	0.4979
AGA	GAA	0.4979
CAT	TCA	0.4869
TCA	CAT	0.4869
CGT	TCG	0.4842

相関係数

属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	GCT	0.1754	LABEL	TAC	0.0512
LABEL	TGC	0.1597	LABEL	GAC	0.0502
LABEL	CCT	0.1505	LABEL	TCC	0.0492
LABEL	CTG	0.1456	LABEL	TAG	0.0488
LABEL	CGC	0.1427	LABEL	CAG	0.0463
LABEL	GCC	0.1393	LABEL	GTT	0.0409
LABEL	GCG	0.1296	LABEL	AGG	0.0393
LABEL	CTC	0.1264	LABEL	ACC	0.0342
LABEL	GTG	0.1264	LABEL	TTG	0.0073
LABEL	CGT	0.1188	LABEL	TTC	0.0034
LABEL	GTC	0.1177	LABEL	ATG	-0.014
LABEL	CCG	0.1097	LABEL	GGA	-0.0148
LABEL	GGC	0.1091	LABEL	AAG	-0.0152
LABEL	AGC	0.1053	LABEL	CCA	-0.0343
LABEL	TCG	0.0965	LABEL	AGA	-0.0403
LABEL	CTT	0.0939	LABEL	ACA	-0.0467
LABEL	CCC	0.0907	LABEL	TGA	-0.0535
LABEL	TCT	0.0851	LABEL	TTT	-0.0557
LABEL	GGT	0.0838	LABEL	AAC	-0.0628
LABEL	TGT	0.0829	LABEL	GAA	-0.0697
LABEL	ACG	0.0825	LABEL	CAT	-0.0725
LABEL	CGG	0.0815	LABEL	GAT	-0.0736
LABEL	CTA	0.0773	LABEL	TAT	-0.08
LABEL	CGA	0.0771	LABEL	ATC	-0.0814
LABEL	CAC	0.0738	LABEL	TCA	-0.1015
LABEL	GCA	0.0724	LABEL	AAA	-0.1185
LABEL	GAG	0.0701	LABEL	TTA	-0.1196
LABEL	GGG	0.0617	LABEL	CAA	-0.142
LABEL	GTA	0.0566	LABEL	ATA	-0.1457
LABEL	TGG	0.0533	LABEL	TAA	-0.1817
LABEL	ACT	0.052	LABEL	ATT	-0.2346
LABEL	AGT	0.0515	LABEL	AAT	-0.2736

図 6.20: H4ac の寄与度 (VI)、相関係数、LABEL との相関係数

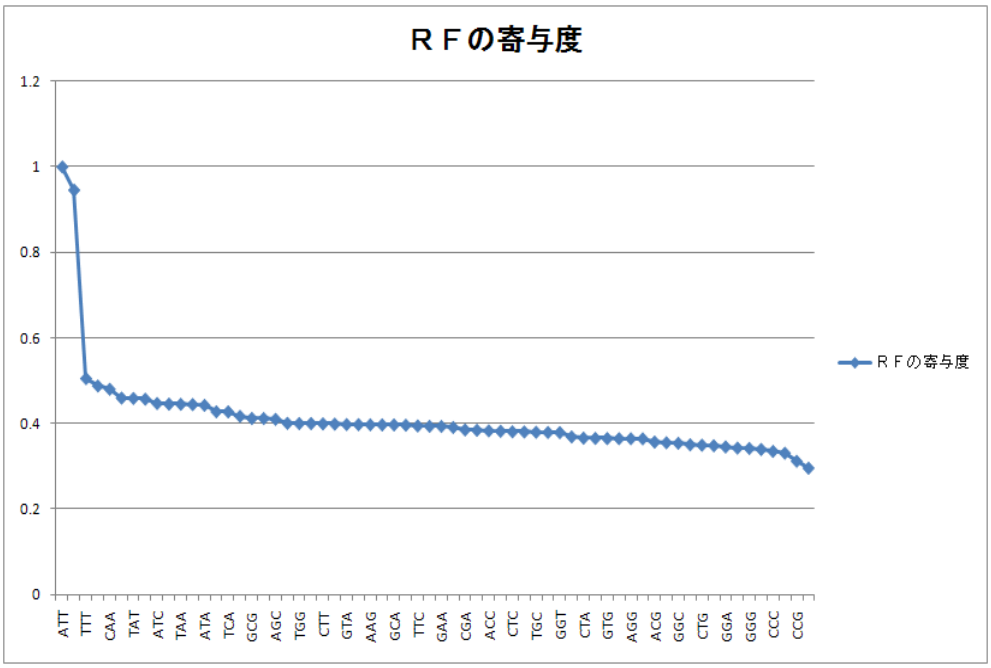


図 6.21: H4ac の寄与度

H3K4me2 の寄与度と SOM の比較

H3K4me2 の予測率は 10 データのうちで 8 番目に高く 68.23 %である。図 6.14 の寄与度のグラフでは、裾野は厚くかつ横に長い。寄与度は 0.5 から 0.4 の範囲に入っている。TCT と CTT (0.7051)、TTC と TCT (0.6995)、TTC と CTT (0.6949)、AAG と AGA (0.6433)、AAG と GAA (0.6399)、AGA と GAA (0.6393)、TCT と CTC (0.5918)、TGG と GGT (0.5602)、AAT と TAA (0.549) 、ATA と TAT (0.5454) 、TCC と CCT (0.5315) 、AGA と GAG (0.5285) 、ACC と CCA (0.5248) 、ATC と TCA (0.5208) 、ATT と TTA (0.5029) 、である。図 6.22 の SOM のグラフでは、赤い正例の部分が大半を占めている。負例の部分はわずかだが、個々の属性をみると AAA、AAT、ATA、ATT、TAA、TAA が属しいている。正例は、GCC、CCA、CCT が属している。LABEL では、正例の領域は多いが、個々の属性をみると赤い頻度の高い部分が少ない。これは、寄与度のグラフと照らし合わせると正例部分は、裾野の部分に対応していると推測される。

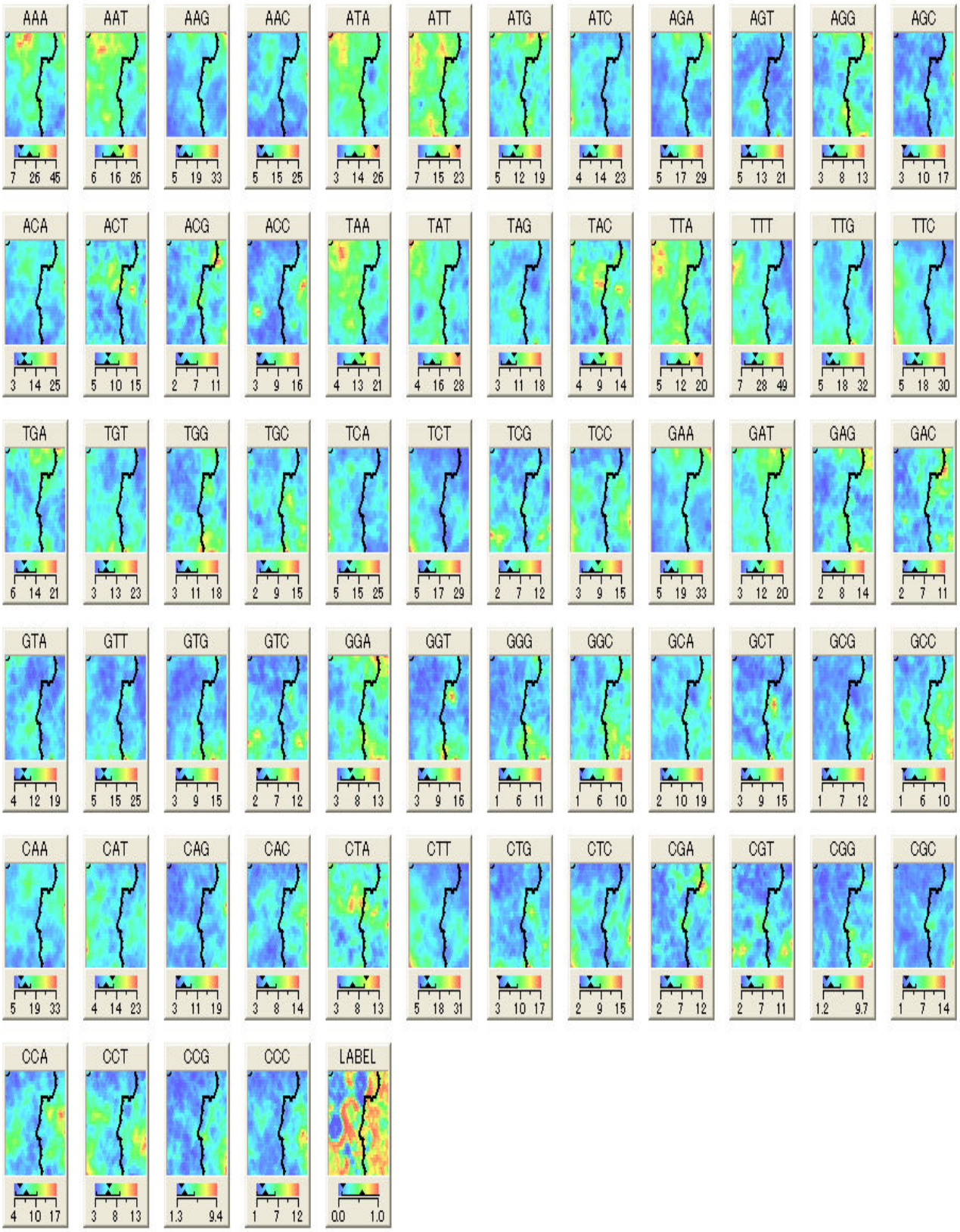


図 6.22: H3K4me2 の SOM 表示

RFでの寄与度		
ATT	327.1605	1
AAT	294.8887	0.901358
TTA	263.7496	0.806178
TTT	236.999	0.724412
TAT	233.9282	0.715026
TAA	210.5522	0.643575
ATA	203.2961	0.621396
AAA	195.8734	0.598707
TGA	158.5021	0.484478
CAT	157.1003	0.480193
TCA	155.9265	0.476606
TTG	155.2569	0.474559
AAG	153.9669	0.470616
TGG	153.1085	0.467992
GAT	152.9306	0.467448
TTC	152.4965	0.466121
AGA	152.2388	0.465334
CAA	151.502	0.463082
CTT	151.424	0.462843
TCT	151.1443	0.461988
GAA	151.1217	0.461919
ATG	150.1927	0.45908
ATC	150.0107	0.458523
GCA	148.8773	0.455059
ACA	148.3718	0.453514
GTC	148.1158	0.452731
TGT	146.5269	0.447875
GTT	146.2337	0.446978
CCA	145.8647	0.445851
GAG	145.5081	0.444761
AGT	144.2946	0.441051
AAC	143.9907	0.440123

相関係数		
属性 1	属性 2	相関係数
TCT	CTT	0.7051
CTT	TCT	0.7051
TTC	TCT	0.6995
TCT	TTC	0.6995
TTC	CTT	0.6949
CTT	TTC	0.6949
AAG	AGA	0.6433
AGA	AAG	0.6433
AAG	GAA	0.6399
GAA	AAG	0.6399
AGA	GAA	0.6393
GAA	AGA	0.6393
TCT	CTC	0.5918
CTC	TCT	0.5918
TGG	GGT	0.5602
GGT	TGG	0.5602
AAT	TAA	0.549
TAA	AAT	0.549
ATA	TAT	0.5454
TAT	ATA	0.5454
TCC	CCT	0.5315
CCT	TCC	0.5315
AGA	GAG	0.5285
GAG	AGA	0.5285
ACC	CCA	0.5248
CCA	ACC	0.5248
ATC	TCA	0.5208
TCA	ATC	0.5208
ATT	TTA	0.5029
TTA	ATT	0.5029
ATA	TAA	0.4971

相関係数					
属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	GTC	0.1139	LABEL	TGC	0.0414
LABEL	TCG	0.0963	LABEL	CCC	0.0392
LABEL	GGC	0.0951	LABEL	GTA	0.0365
LABEL	CGT	0.0901	LABEL	TGT	0.024
LABEL	GGG	0.0841	LABEL	AAG	0.0238
LABEL	GCA	0.0812	LABEL	TAG	0.0234
LABEL	CGA	0.0803	LABEL	TCC	0.0228
LABEL	GTG	0.0799	LABEL	TTC	0.0154
LABEL	AGC	0.0796	LABEL	GTT	0.0141
LABEL	GCG	0.0793	LABEL	AAC	0.004
LABEL	CAG	0.079	LABEL	TTG	0.004
LABEL	CGG	0.0725	LABEL	TCA	0.0015
LABEL	TGG	0.0719	LABEL	AGA	0.0008
LABEL	AGG	0.0713	LABEL	ATC	-0.0019
LABEL	CAC	0.0683	LABEL	CAA	-0.004
LABEL	GGT	0.0674	LABEL	ACA	-0.0169
LABEL	CCG	0.0657	LABEL	GAA	-0.0172
LABEL	ACC	0.0648	LABEL	CAT	-0.0215
LABEL	GAG	0.0645	LABEL	ACT	-0.0227
LABEL	GCC	0.0637	LABEL	GAT	-0.0236
LABEL	ACG	0.063	LABEL	ATG	-0.0324
LABEL	CGC	0.0617	LABEL	CTA	-0.0329
LABEL	CTC	0.0613	LABEL	TAC	-0.0399
LABEL	GGA	0.0572	LABEL	TGA	-0.0534
LABEL	GAC	0.0564	LABEL	AAA	-0.0553
LABEL	AGT	0.0547	LABEL	TTT	-0.0803
LABEL	GCT	0.0545	LABEL	TAA	-0.1261
LABEL	CCT	0.0485	LABEL	ATA	-0.1272
LABEL	CTG	0.0483	LABEL	TAT	-0.1416
LABEL	TCT	0.0463	LABEL	AAT	-0.1607
LABEL	CTT	0.0441	LABEL	TTA	-0.1666
LABEL	CCA	0.0418	LABEL	ATT	-0.1903

図 6.23: H3K4me2 の寄与度 (VI)、相関係数、LABEL との相関係数

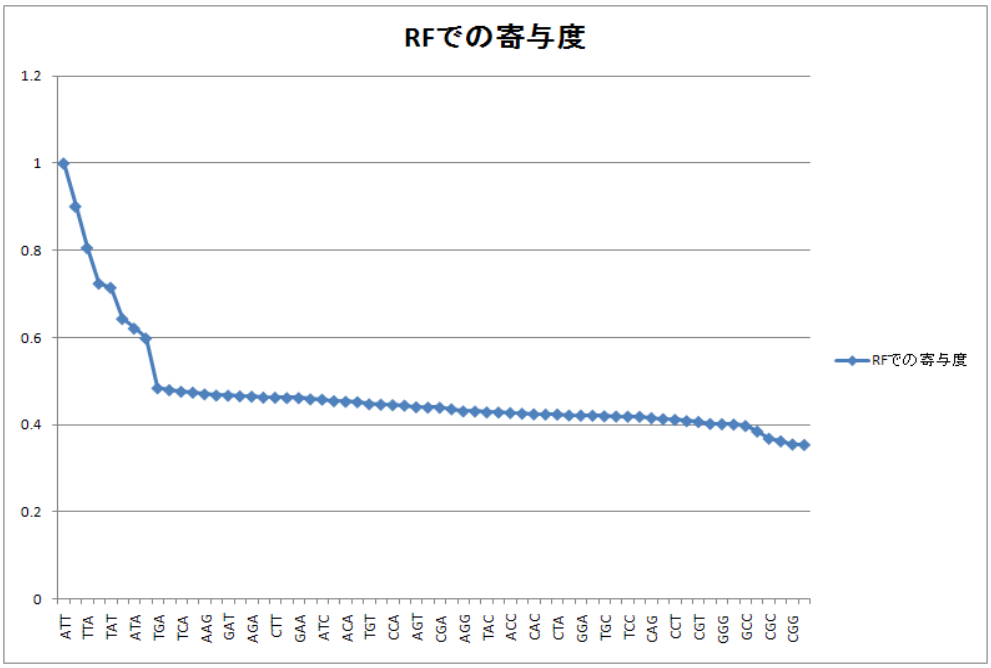


図 6.24: H3K4me2 の寄与度

H3K4me1 の寄与度と SOM の比較

H3K4me1 の予測率は 10 データのうちで 9 番目で 67.13 %である。図 6.27 の寄与度のグラフでは、裾野の部分が厚い。相関係数では、CTT と TCT (0.667)、CTT と TTC (0.6666)、TCT と TTC (0.6663)、TAT と ATA (0.5937)、TAA と AAT (0.5845)、CTC と TCT (0.5505)、CCT と TCC (0.5437)、AGA と AAG (0.5377)、GAA と AAG (0.5307)、CCA と ACC (0.5254)、TTA と ATT (0.525)、GAA と AGA (0.5169)、TCA と ATC (0.5155)、TAA と ATA (0.5002)、である。図 6.25 では、LABEL では、赤い頻度の高い部分が大半を占めいている。負例の青い部分は、AAT、ATA、ATT、TAA、TAT、TTA である。正例での赤い部分では、TAC、CCA である。

図 6.25: H3K4me1 の SOM 表示

RF寄与度		
TTT	290.0573	1
TAT	261.4651	0.901426
ATA	249.6921	0.860837
CCA	245.3697	0.845935
AAA	239.0204	0.824045
TGG	224.7817	0.774956
CAA	209.9321	0.723761
TTA	199.3005	0.687107
TTG	195.5113	0.674044
GAT	184.3479	0.635557
ATC	180.995	0.623997
TGA	180.7979	0.623318
TAA	180.4663	0.622175
CTT	178.4712	0.615296
ATT	177.5747	0.612206
TCA	176.0058	0.606797
AAT	175.7782	0.606012
GAA	175.0312	0.603437
AGA	173.639	0.598637
AAG	172.5791	0.594983
TTC	171.1328	0.589997
ATG	170.7273	0.588599
TGT	167.205	0.576455
TCT	166.3123	0.573377
CAT	164.8919	0.56848
ACA	162.6655	0.560805
TAC	158.2089	0.54544
AAC	156.66	0.5401
GGA	154.9113	0.534071
GTA	154.3119	0.532005
AGT	153.8164	0.530297
GGT	153.8005	0.530242

相関係数		
属性1	属性2	相関係数
CTT	TCT	0.667
TCT	CTT	0.667
CTT	TTC	0.6666
TTC	CTT	0.6666
TCT	TTC	0.6663
TTC	TCT	0.6663
TAT	ATA	0.5937
ATA	TAT	0.5937
TAA	AAT	0.5845
AAT	TAA	0.5845
CTC	TCT	0.5505
TCT	CTC	0.5505
CCT	TCC	0.5437
TCC	CCT	0.5437
AGA	AAG	0.5377
AAG	AGA	0.5377
GAA	AAG	0.5307
AAG	GAA	0.5307
CCA	ACC	0.5254
ACC	CCA	0.5254
TTA	ATT	0.525
ATT	TTA	0.525
GAA	AGA	0.5169
AGA	GAA	0.5169
TCA	ATC	0.5155
ATC	TCA	0.5155
TAA	ATA	0.5002
ATA	TAA	0.5002
GGT	TGG	0.4949
TGG	GGT	0.4949
TTA	TAT	0.494

相関係数					
属性1	属性2	相関係数	属性1	属性2	相関係数
LABEL	CCA	0.1678	LABEL	CCG	0.0126
LABEL	ATC	0.165	LABEL	GCA	0.0091
LABEL	TTG	0.1477	LABEL	CAC	0.0068
LABEL	TCA	0.146	LABEL	AGT	0.0025
LABEL	TCC	0.1422	LABEL	AAT	-0.0001
LABEL	TTC	0.1401	LABEL	GTA	-0.0005
LABEL	CAT	0.1298	LABEL	ATT	-0.0008
LABEL	TCT	0.1293	LABEL	GTG	-0.0013
LABEL	TGG	0.121	LABEL	GCT	-0.0116
LABEL	GTC	0.1169	LABEL	TAG	-0.0126
LABEL	CTT	0.0984	LABEL	ATG	-0.0223
LABEL	CTC	0.0772	LABEL	AGC	-0.0227
LABEL	TCG	0.0745	LABEL	AAC	-0.0312
LABEL	ACC	0.0731	LABEL	GCG	-0.0368
LABEL	CGT	0.0652	LABEL	CGC	-0.0386
LABEL	CTG	0.0621	LABEL	CGA	-0.0412
LABEL	GGT	0.0599	LABEL	AGG	-0.0432
LABEL	CCT	0.0583	LABEL	GAG	-0.0437
LABEL	GTT	0.0508	LABEL	ACG	-0.0501
LABEL	CAA	0.0437	LABEL	TTT	-0.0606
LABEL	CAG	0.0419	LABEL	ACT	-0.0632
LABEL	TGT	0.041	LABEL	GAA	-0.0923
LABEL	GCC	0.0398	LABEL	ACA	-0.1067
LABEL	GGA	0.0382	LABEL	CTA	-0.1097
LABEL	TGC	0.0365	LABEL	TAA	-0.1128
LABEL	GAT	0.0355	LABEL	AGA	-0.1186
LABEL	TGA	0.0333	LABEL	TAC	-0.1338
LABEL	GGG	0.0239	LABEL	AAG	-0.1464
LABEL	CGG	0.021	LABEL	TTA	-0.1563
LABEL	GGC	0.0203	LABEL	ATA	-0.1586
LABEL	CCC	0.0178	LABEL	TAT	-0.1661
LABEL	GAC	0.0166	LABEL	AAA	-0.1804

図 6.26: H3K4me2 の寄与度 (VI)、相関係数、LABEL との相関係数

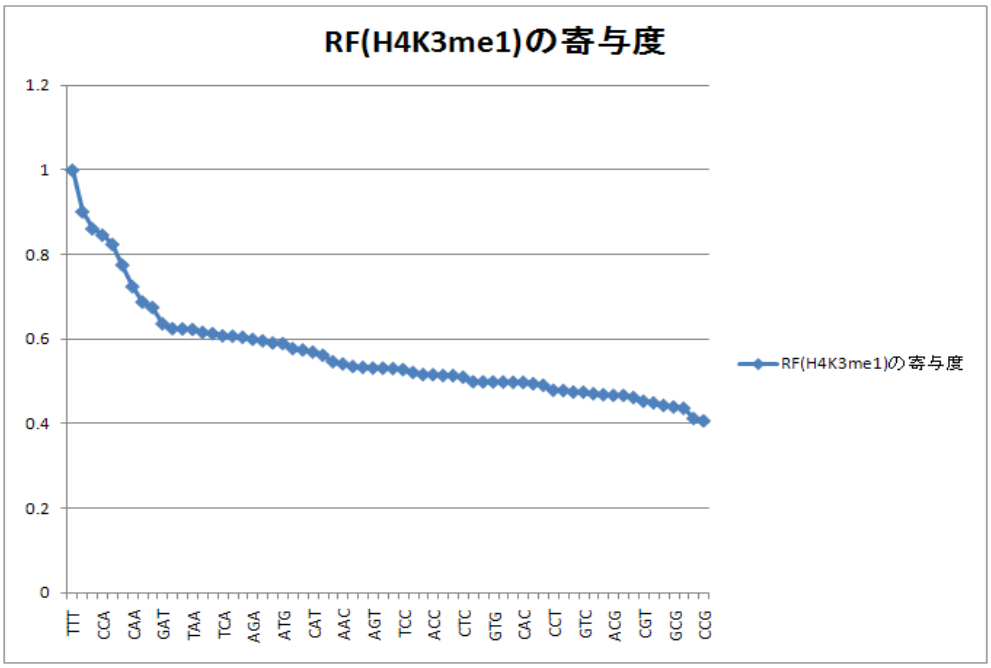


図 6.27: H3K4me1 の寄与度

H3K4me3 の寄与度と SOM の比較

H3K4me3 の予測率は 10 データのうちで 10 番目で 65.29 %である。H3K4me3 の図 6.9 のグラフの特徴は、急勾配の部分が短く、裾野の勾配がなだらかな部分が厚いことである。予測率は 10 データの中で最も低く、65.87 %である。相関係数は、CCA と ACC (0.6775)、GAA と AAG (0.676)、AGA と AAG (0.6711)、GAA と AGA (0.6519)、TTA と ATT (0.645)、TAT と ATA (0.6034)、TTA と TAT (0.5676)、GAG と AGA (0.5645)、TAA と AAT (0.5596)、TAA と ATA (0.5547)、CAA と AAC (0.5496)、GCA と AGC (0.5493)、GGA と AGG (0.5279)、CCT と TCC (0.5102)、GAT と TGA (0.5062)、CGA と ACG (0.5047)、である。図 6.28 の SOM では、LABEL では、ほぼ全域に赤い部分に占めらている。負例の部分は、AAA、AAT、ATA、TAA である。正例の部分は、CTA があげられる。これは、正例と負例の部分の境界領域の属性が多い事を示している。特徴ベクトルは sliding window で作成しているため、3-gram での頻度が均等に近いとも言える。

図 6.28: H3K4me3 の SOM 表示

RFの寄与度

AAT	266.8243	1
ATT	264.1867	0.990115
AAA	217.2471	0.814195
TTT	208.1982	0.780282
GAT	188.689	0.707166
TGA	187.4493	0.70252
TAT	187.1067	0.701236
ATA	183.5593	0.687941
ATC	183.463	0.68758
CAA	183.3192	0.687041
TTG	180.559	0.676696
TCT	177.6938	0.665958
TCA	177.6217	0.665688
ATG	176.2945	0.660714
GAA	176.1432	0.660147
TTA	176.1363	0.660121
TAA	174.7599	0.654962
CAT	174.6886	0.654695
AAG	173.8622	0.651598
AGA	173.2218	0.649198
CTT	170.6261	0.63947
TTC	169.4362	0.63501
TGG	166.3313	0.623374
CCA	165.4131	0.619933
TGT	165.1554	0.618967
GTA	163.7311	0.613629
GTT	163.5248	0.612856
ACA	161.612	0.605687
AAC	160.7112	0.602311
TAC	159.542	0.597929
AGC	159.163	0.596509
ACT	158.7378	0.594915

相関係数

属性 1	属性 2	相関係数
CCA	ACC	0.6775
ACC	CCA	0.6775
GAA	AAG	0.676
AAG	GAA	0.676
AGA	AAG	0.6711
AAG	AGA	0.6711
GAA	AGA	0.6519
AGA	GAA	0.6519
TTA	ATT	0.645
ATT	TTA	0.645
TAT	ATA	0.6034
ATA	TAT	0.6034
TTA	TAT	0.5676
TAT	TTA	0.5676
GAG	AGA	0.5645
AGA	GAG	0.5645
TAA	AAT	0.5596
AAT	TAA	0.5596
TAA	ATA	0.5547
ATA	TAA	0.5547
CAA	AAC	0.5496
AAC	CAA	0.5496
GCA	AGC	0.5493
AGC	GCA	0.5493
GGA	AGG	0.5279
AGG	GGA	0.5279
CCT	TCC	0.5102
TCC	CCT	0.5102
GAT	TGA	0.5062
TGA	GAT	0.5062
CGA	ACG	0.5047
ACG	CGA	0.5047
ACA	AAC	0.5028
AAC	ACA	0.5028
CAA	ACA	0.5016
ACA	CAA	0.5016
CCA	GCC	0.4989

LABELとの相関係数

属性 1	属性 2	相関係数	属性 1	属性 2	相関係数
LABEL	AGG	0.1287	LABEL	CCA	0.0262
LABEL	AGC	0.1167	LABEL	CGT	0.0253
LABEL	GCG	0.1142	LABEL	GTA	0.0117
LABEL	GGC	0.1029	LABEL	GGT	0.0111
LABEL	AAG	0.1024	LABEL	CTC	0.0107
LABEL	GAG	0.1018	LABEL	AGT	0.0067
LABEL	GCA	0.1016	LABEL	TGG	0.0055
LABEL	GAC	0.0964	LABEL	CCT	0.001
LABEL	GGA	0.0952	LABEL	TAG	-0.0157
LABEL	GGG	0.0911	LABEL	ACT	-0.0205
LABEL	ACG	0.0905	LABEL	ATG	-0.0243
LABEL	CGA	0.0897	LABEL	ATA	-0.0284
LABEL	CGG	0.0897	LABEL	TCC	-0.0309
LABEL	CAG	0.0871	LABEL	CTA	-0.0347
LABEL	GCC	0.0758	LABEL	TGT	-0.0374
LABEL	CGC	0.0744	LABEL	GAT	-0.0395
LABEL	CAC	0.0717	LABEL	CAT	-0.0432
LABEL	AGA	0.0704	LABEL	TAA	-0.0448
LABEL	ACA	0.0645	LABEL	TAC	-0.045
LABEL	GTG	0.0516	LABEL	CTT	-0.0485
LABEL	AAC	0.0487	LABEL	TGA	-0.0569
LABEL	CCG	0.0485	LABEL	TCA	-0.0588
LABEL	GCT	0.0476	LABEL	TCT	-0.0642
LABEL	GAA	0.0424	LABEL	ATC	-0.0661
LABEL	GTC	0.0416	LABEL	GTT	-0.0883
LABEL	AAA	0.0373	LABEL	TTC	-0.0956
LABEL	TGC	0.0347	LABEL	TTG	-0.0959
LABEL	TCG	0.0336	LABEL	TAT	-0.1071
LABEL	CCC	0.0336	LABEL	TTT	-0.1206
LABEL	ACC	0.0324	LABEL	AAT	-0.1236
LABEL	CTG	0.0317	LABEL	TTA	-0.1704
LABEL	CAA	0.0264	LABEL	ATT	-0.2063

図 6.29: H3K4me3 の寄与度 (VI)、相関係数、LABEL との相関係数

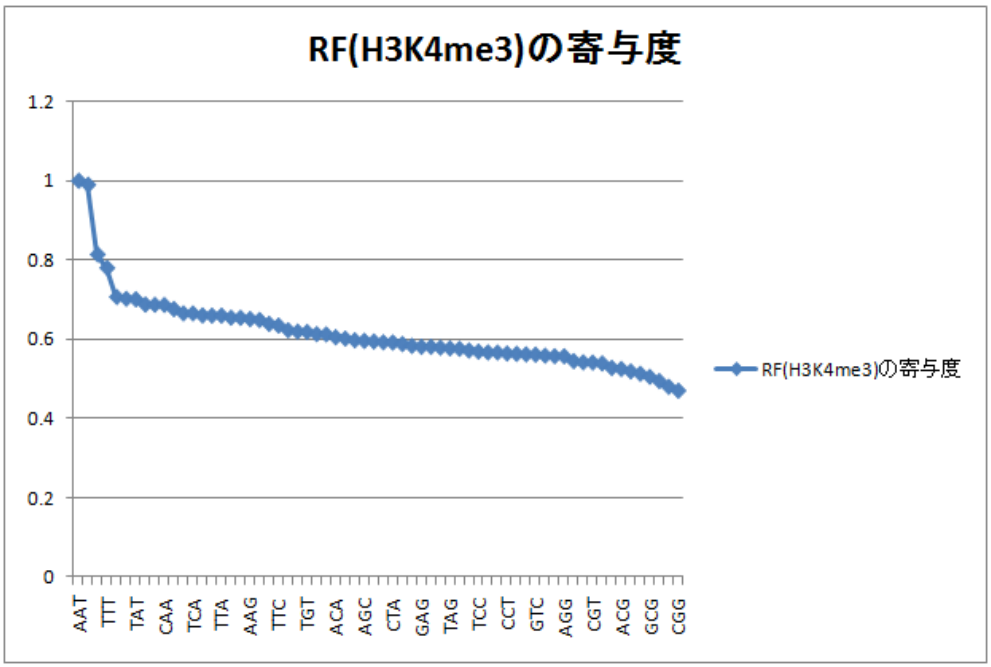


図 6.30: H3K4me3 の寄与度

まとめ

寄与度と予測率の関係であるが、寄与度のグラフから、グラフの裾野が薄く、寄与している属性が明確な場合予測率が高い傾向がある。これは、正例と負例の境界領域の部分が少ないことが推測される。逆に、予測率が低く場合は、寄与度のグラフの裾野の部分が厚い傾向が見られる。これは、境界領域に属する属性が多いということと、属性の頻度が均等に近いためと推測される。SOM でのクラスタリングでの図では、その結果を裏付けている。H3、H4 などの予測率が高いデータの LABEL は、明確に色相が 2 部されている。逆に、H3K4me2 や H3K4me3 など予測率が低いデータの LABEL は、全体に一様にひろがっている。ランキングの順位には、正例で重要な属性と負例で重要な属性が混じっている。相関係数をみると、どのデータでも相関係数が高い属性の組みがある。AGA と AAG、GAA と AAG、CTT と TCT がその例である。先行研究では、寄与度と予測率に言及している論文はないが、寄与度のグラフから予測率の性能、性能を左右する属性の定量評価され特定できれば、寄与度からの知見として興味深いと考えられる。

第 7章

位置特異な情報を用いた特徴ベクトルでの予測と属性部分集合選択とその近傍探索

7.1 背景

機械学習によるエピジェネティクス関連領域の予測の先行研究としては、Pham らによるSVM を用いた研究がある [15]。彼らはRBF カーネルを用いて予測を行う一方で、別途polynomial kernel で学習した際の重みを用いて特徴のランキングを行うことにより、特徴ベクトルの属性の重要性を解析している。さらに、Tran らによる研究では、Conditional Random Field を用いて予測を行い、SVM との比較を行っている [16]。また、先のような研究では、しばしば高次元データになるため属性選択に関する研究も活発に行われている。新島ら [18] は、化合物とタンパク質の相互作用や活性の予測ばかりではなく、それらに関与する属性を抽出する数理的手法としてカーネル空間での化合物・タンパク質の活性空間を表現し、その空間で特徴抽出する手法を提案している。

7.2 目的

本研究では、エピジェネティクス現象を解析した配列データを対象として、位置情報の特徴ベクトルを用い その化学反応に対して機械学習による判別分析を行った。その際、属性選択の評価指標に randomForest で生成される Mean Decrease Gini index を用い、そのランキングに従い属性部分集合探索を行った。また、DEWS2008 での頻度データでの実験との比較も行った。実験結果からモチーフ抽出した配列の組合せによる位置をカウントした特徴ベクトルが高い予測率を示すことがわかった。

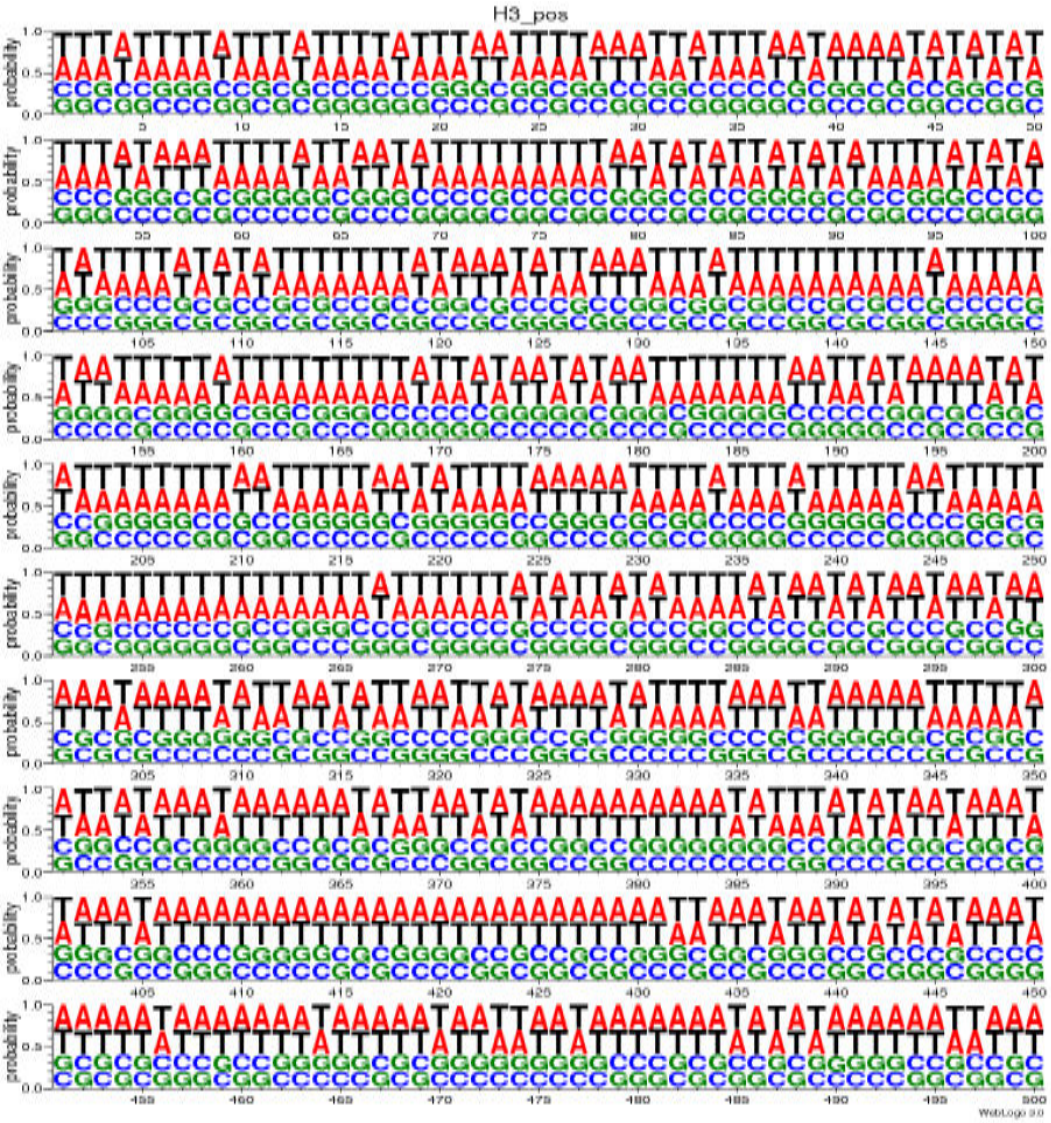


図 7.1: 正例 (H3) の塩基の位置毎の頻度 (Weblogo による出力) 横軸は塩基の位置　縦軸は頻度のパーセント表示

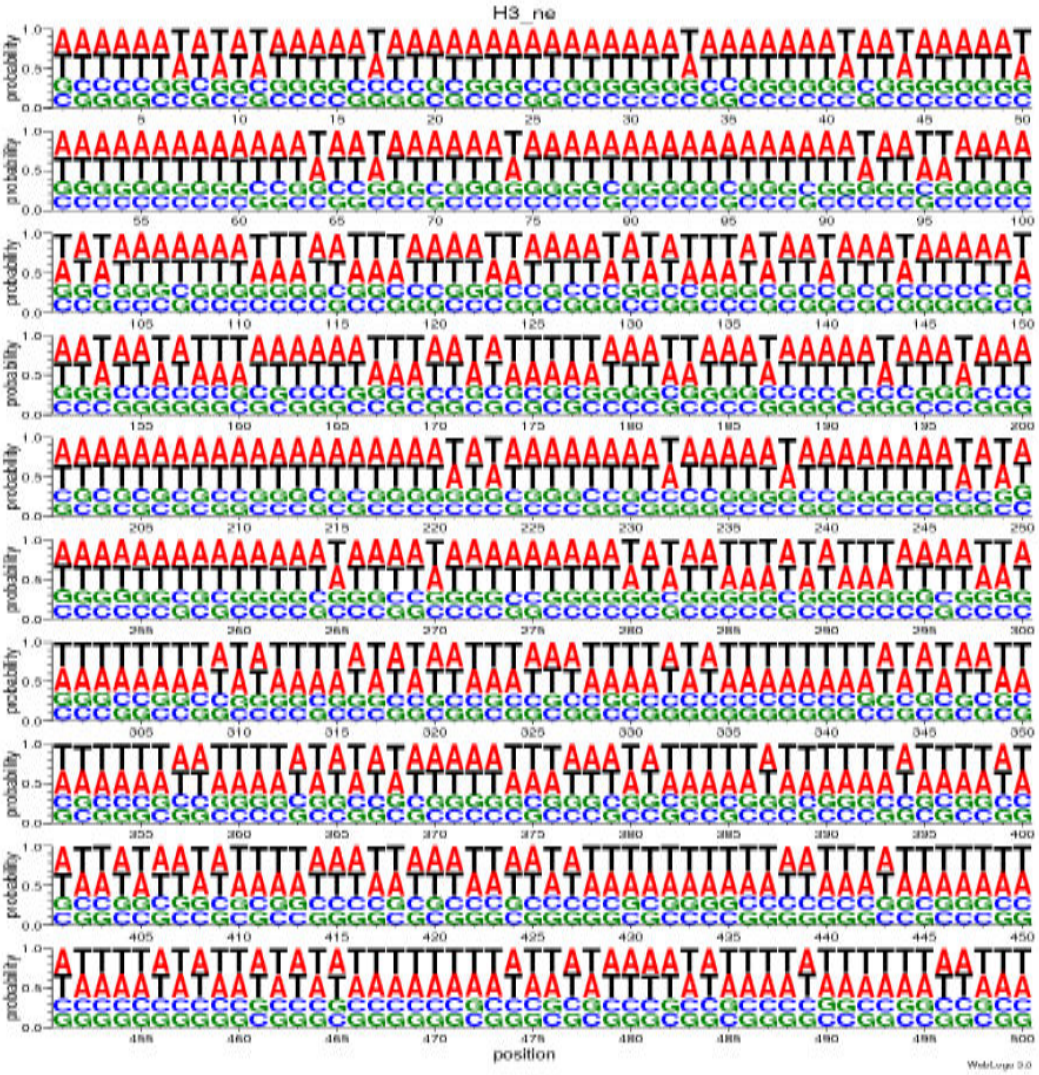


図 7.2: 負例 (H3) の塩基の位置毎の頻度 (Weblogo による出力) 横軸は塩基の位置　縦軸は頻度のパーセント表示

7.3 提案手法

本研究では、Pokholok らのデータを用い位置情報に着目した特徴ベクトルを用いる。この特徴ベクトルを randomForest というアンサンブル学習を用いて MeanDecrease Gini index を用いて判別に重要な属性を求める。次に全属性から属性選択をし、機械学習を行う。機械学習には、support vector machine を用いる。第 5 章でおこなった頻度データとの結果との比較も行う。2.1. Randomforest Random Forest は、Breiman により提案されたアンサンブル学習の 1 つである。RandomForest は、リサンプリングに bootstrap を用い、サブデータを作成し、それぞれのサブデータセットの決定木を組み合わせる方法である。RandomForest のアルゴリズムは、

1) 与えられたデータセットから数組のブートストラップデータを作成する。

2) 各々のブートストラップサンプルデータを用いて枝刈りされていない最大の決定・回帰木を作成する。分岐のノードとしては、ランダムサンプリングされた変数の中で最も良いものを選択する。

3) すべての結果を組み合わせて (回帰の問題では平均、分類は多数決)、新しい予測・分類器を構築する。

本研究では、R のパッケージである randomForest を用いた。

support vector machine は、R のパッケージである kernlab を用いた。カーネルは、RBF カーネルでパラメータ μ=0.01 である。2.3. 属性選択、属性部分集合選択属性選択のアルゴリズムは次のように行った。

1)randomForest の学習で、Mean Dcrease Gini index を計算する。

2) 1) で得られた ranking に従って属性数を減らし SVM によって学習を行う。

探索の方向は、後ろ向き探索で、探索の戦略としては、全属性で最高の予測率を出した属性の近傍の組合せを考え、再度学習を行う。

7.4 実験結果

計算機は、大学内の PC クラスターを用いた。CPU は AMD Opteron Dual Processor Model 250(2.4GHz) × 32、メモリ:4GB、OS は、SuSE Linux Enterprise Server 8, SCore 5.8 である。3.1. データセット Pokholok[6] が発表しているクロマチンデータは、出芽酵母

のゲノム DNA 上の異なる部分領域 (41282 箇所) に対して 14 種類の実験データを提供している。データは DNA 上の 1 点という形で（ 染色体の先頭からの位置）という形式で表現されている。Pham らは、この 14 種類のデータのうち 10 種類（ 表 1 では 5 種類のみ表示） に対して、指定された 1 点を中心として 200, 500, 1000 という長さの部分配列をとり、固定長 k=3〜11（ 塩基）の k-gram の出現頻度を sliding window でカウントしたものを特徴ベクトルとし、これを用いて予測実験を行っている。実験結果は、公開されている。(http://www.jaist.ac.jp/ tran/nucleosome/) 本研究では、この提供されている長さ 500 の部分配列を実験に用いた。ラベルは、連続量で表現されている Pokholok らのデータを正規化し、1.2 以上の場合は正例のラベル、0.8 以下の場合は負例のラベルを付与した。3.2. 位置を考慮した特徴ベクトル

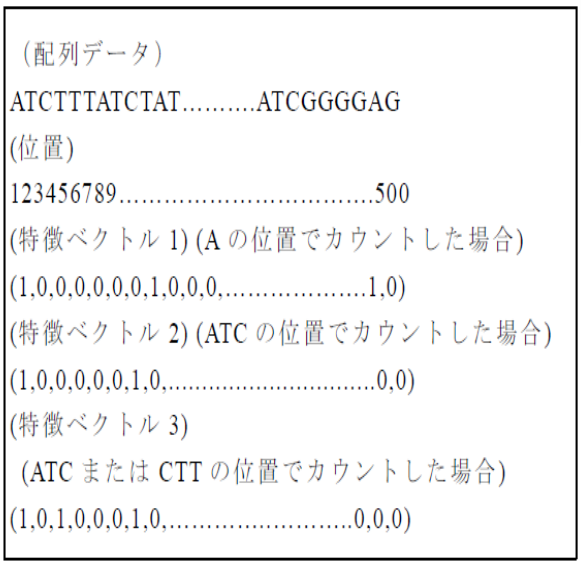


図 7.3: 位置を考慮した特徴ベクトル

本研究では、長さ 500 のクロマチンデータを次のように、その位置でカウントした。特徴ベクトル 1: 1 塩基｛A,T,G,C｝ をそれぞれの位置でカウント特徴ベクトル 2: 3 つの塩基の組合せ｛AAA,AAT,…,CCC｝の 64 種類の組合せをそれぞれの位置でカウント特徴ベクトル 3: Gini 係数で ranking した上位からの組合せを位置でカウント (例)AAA,TTT…が上位であれば AAAorTTT で位置をカウント位置毎の塩基の頻度を図 6.1、図 6.2 では Weblogo を用いて表示した。特徴ベクトル 2 は、3 塩基の先頭の位置で 1 とカウントしたため最後の 499,500 の位置は 0 となる。特徴ベクトル 3 では、上位の 3 塩基の組合せを用いた。図

6.3 で具体的な特徴ベクトルの例を示す。3.3. 実験結果（feature ranking）randomForest での Gini index の値によって位置データをランキングした。図 6.4 では、特徴ベクトル 1 の 1 塩基の属性のランキングを表し、図 6.5 では、特徴ベクトル 2 の先の研究で行った頻度データで H3 の場合の順位の高い属性を選び、その位置を示した。図 6.5 では、特徴ベクトル 3 の実験結果である。

3.4. 実験結果 (属性部分集合選択) 属性部分集合選択の計算機実験では、

・ 頻度データ

・ 位置データ

を用いて比較する。表 3 では、H3 の位置データの randomForest での予測率の上位 10 属性をあげており、頻度データは、Gini index の上位 10 属性をあげている。表 4 は、1 塩基での randomForest での予測率である。実験では 5 つのクロマチンデータを用いたが、最も予測率の高い結果を示す H3 の場合のみ詳述する。

7.5 まとめ

図 4 は、1 塩基で位置毎にカウントした特徴ベクトルでの Gini index であり判別分析に対する寄与を表している。表 5 は、各 1 塩基で判別に寄与している位置を表している。図 5 は、先の頻度データで判別の寄与が高い 3 塩基 TTT、AAA、TAA、TTA、ATA の位置毎の Gini index を表している。表 6 は、その 5 つの塩基が、位置での学習での判別での寄与の順位を表している。表 4,5 では、頻度データで上位の予測率を示す属性のと位置データでの上位の予測率のデータは、同じものが多いことを示している。今後の課題としては、

・頻度データでの判別分析での寄与率の高い属性の組合せの位置での特徴ベクトルが高い予測率を示した。より精度の高いモチーフ抽出の手法の利用があげられる。今回塩基の長さが長さ 1 と長さ 3 の場合のみ考えているが、モチーフ抽出により長さを固定してない塩基の抽出を行う予定である。配列の曖昧さはモチーフ抽出の過程で解消されると考えている。

・複数の塩基の共起が考慮されていない。今後の課題とする。

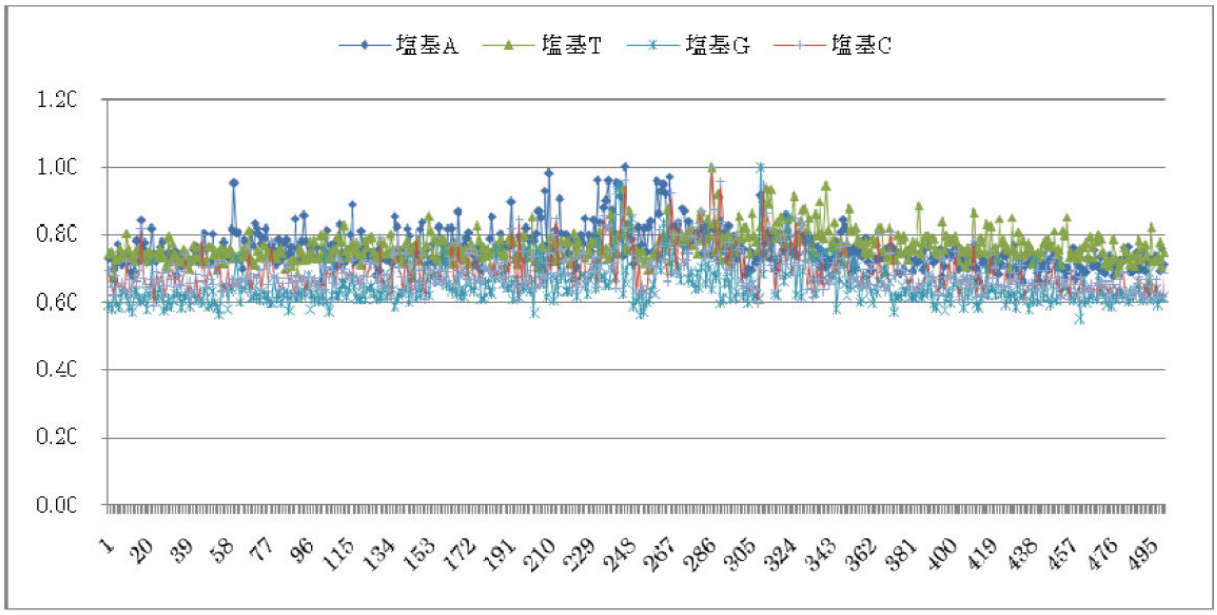


図 7.4: 位置ごとに 1 塩基をカウントした属性の正規化した Gini index

位置ごとの属性の順位（Gini 係数)

A		T		G		C	
順位	位置	順位	位置	順位	位置	順位	位置
1	245	1	286	1	309	1	286
0.982065	209	0.944559	340	0.949009	242	0.959978	245
0.970615	266	0.937709	312	0.931382	241	0.955079	290
0.960901	232	0.935196	244	0.857159	322	0.946513	310
0.959826	237	0.933726	314	0.84852	248	0.924445	267

表 7.1: 位置ごとの属性の順位（Gini 係数)

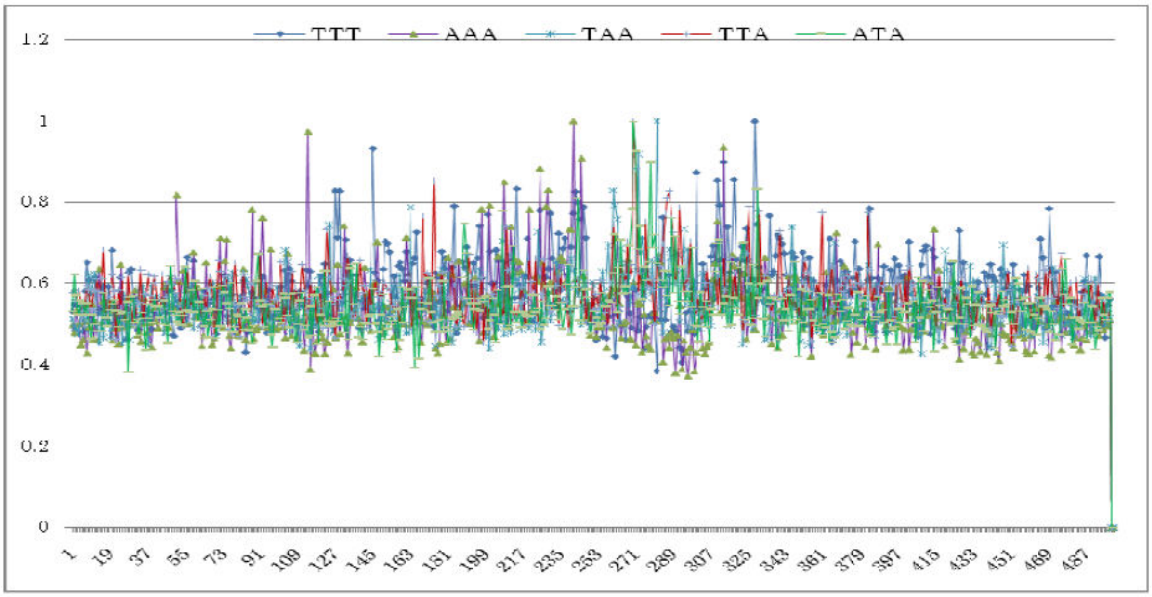


図 7.5: 位置ごとに 3 塩基をカウントした属性の正規化した Gini index

位置ごとの属性の順位（Gini 係数)

TTT		AAA		TAA		TAA		ATA	
順位	位置	順位	位置	順位	位置	順位	位置	順位	位置
1	328	1	241	1	281	1	270	1	270
0.932129	145	0.973186	114	0.918874	272	0.885142	271	0.927609	271
0.899915	313	0.935867	313	0.828508	260	0.852877	174	0.90029	278
0.874886	300	0.909056	245	0.792661	261	0.825822	287	0.832291	329
0.855109	318	0.88551	225	0.78813	163	0.8109	286	0.81059	243

表 7.2: 位置ごとの属性の順位（Gini 係数)

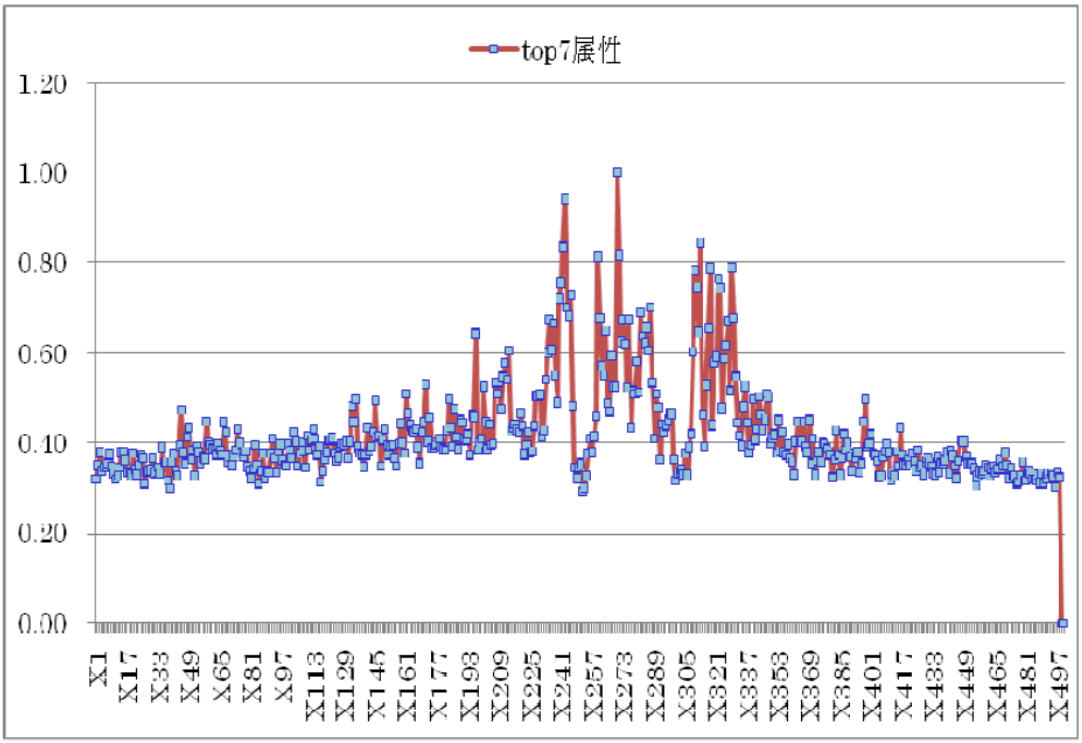


図 7.6: H3 の属性部分集合で最も高い予測率の位置毎の正規化した Gini index
TTT,AAA,TAA,TTA,ATA,TAT,ATT

第 8 章

機械学習ベンチマークデータでの予備実験

8.1 目的

本予備実験の目的は、n 個の属性の全探索空間 $2^n - 1$ の中で RandomForest の MeanDecreaseGini 係数によりランキングされた属性の n 個の組合せ以外で、予測率が高い属性の組合せの探索をすることである。

8.2 提案手法

以下の手法により計算機実験を行った。

1. データは、UCI の機械学習ベンチマークテストデータを用いた。RandomForest により MeanDecreaseGini 係数によるランキングを行った。(統計ソフト R を使用)
2. UCI の機械学習ベンチマークテストデータについて support vector machine による全探索をおこなった。(データマイニングツール Weka polynomial kernel 使用)
3. 1. と 2. の属性の組み合わせによる予測率を比較した。

8.3 計算機実験及び実験結果

データセットは UCI バークレイの機械学習ベンチマークテストデータを使用した。
データの詳細は以下の通りである。
ピマ族の糖尿病の判定　Pima Indians Diabetes Data Set

データセットの特性:多変数

データ数:768

分野:Life

属性のデータ型:整数, 実数

属性数:8

データ提供日:1990-05-09

関連する処理:Classification

欠損値:あり

web での利用数:39314

属性の情報: 計算機実験の表では、属性を a と置いている。a1 とは、下の属性の順番通りである。

1. 妊娠の回数　Number of times pregnant (a1 と置く)
2. 経口グルコーステストの結果　Plasma glucose concentration a 2 hours in an oral glucose tolerance test (a2 と置く)
3. 血圧　Diastolic blood pressure (mm Hg) (a3 と置く)
4. 皮膚の厚さ Triceps skin fold thickness (mm) (a4 と置く)
5. 2 時間後のインシュリンの摂取量 2-Hour serum insulin (mu U/ml) (a4 と置く)
6. 体重と身長の比率 Body mass index (weight in kg/($heightinm$)^2) (a6 と置く)
7. 糖尿病の家系調査の結果 Diabetes pedigree function (a7 と置く)
8. 年齢 Age (years) (a8 と置く)
9. クラス　Class variable (0 or 1)

8.4　実験結果

表 8.1 は属性数 1 , 2 の予測率を示している。属性数 1 では、a2 つまり Plasma glucose concentration a2 hours in an oral glucose tolerance test が予測率でもっとも高く 74.6094 %を示している。これは、RandomForest のランキングで a2 が最も寄与度が高く一致する。属性数 2 では、$\{a2, a6\}$ の組合せがもっとも高く 75.651 %を示す。これは、RandomForest のランキングでの 1 位と 2 位である。その他の結果をみると a2 が含まれている組合せが予測率が高いことを示している。

$\{a1, a2\}$ が 75.1302 %

$\{a2, a3\}$ が 74.4792 %
$\{a2, a5\}$ が 74.6094 %
$\{a2, a7\}$ が 74.6094 %
$\{a2, a8\}$ が 74.4792 %
となっており、a2 が含まれている場合予測率は高くなっている。

表 8.2 では、属性数 3 の予測率を示している。図 8.1 はグラフである。予測率が高くなっている山が 2 つある。RandomForest の属性ランキングの組合せは $\{a2, a5, a8\}$ である。
図 8.1 の最初の山の属性の組合せ
$\{a1, a2, a3\}$ は 75.3906% $\{a1, a2, a4\}$ は 75.0000%
$\{a1, a2, a5\}$ は 74.8689% $\{a1, a2, a6\}$ は 76.3021%
$\{a1, a2, a7\}$ は 76.1719% $\{a1, a2, a8\}$ は 75.2604%
が第 1 群の組合せの集合である。
第 2 群の組合せの集合は、
$\{a2, a3, a4\}$ は 74.6094% $\{a2, a3, a5\}$ は 74.7396%
$\{a2, a3, a6\}$ は 74.6094% $\{a2, a3, a7\}$ は 75.0000%
$\{a2, a3, a8\}$ は 75.0000% $\{a2, a4, a5\}$ は 74.6094%
$\{a2, a4, a6\}$ は 75.1302% $\{a2, a4, a7\}$ は 74.4792%
$\{a2, a4, a8\}$ は 73.4375% $\{a2, a5, a6\}$ は 75.5208%
$\{a2, a5, a7\}$ は 75.0000% $\{a2, a5, a8\}$ は 77.8648%
$\{a2, a6, a7\}$ は 76.3201% $\{a2, a6, a8\}$ は 76.6927%
である。第 1 群、第 2 群ともに $a2$ の属性が予測率の高い要因となっている。

表 8.3 は、属性数 4 の予測率を示している。図 8.1 のグラフでは、山が 2 つと 1 つのピークがみられる。RandomForest での属性の組合せは $\{a2, a6, a7, a8\}$ 第 1 群の山の属性の組合せは、
$\{a1, a2, a3, a4\}$ は 74.4702%
$\{a1, a2, a3, a5\}$ は 75.3902%
$\{a1, a2, a3, a6\}$ は 76.5625%
$\{a1, a2, a3, a7\}$ は 75.9115%
$\{a1, a2, a3, a8\}$ は 74.4792%
$\{a1, a2, a4, a5\}$ は 74.4792%
$\{a1, a2, a4, a6\}$ は 75.5208%

表 8.1: 属性数 1,2 の予測率

最高順位	予測率	属性 1	
	65.1042	a1	
○	74.6094	a2	
	65.1042	a3	
	65.1042	a4	
	65.1042	a5	
	65.1042	a6	
	65.1042	a7	
	65.1042	a8	
	予測率	属性 1	属性 2
	75.1302	a1	a2
	65.1042	a1	a3
	65.1042	a1	a4
	65.1042	a1	a5
	66.4063	a1	a6
	65.1042	a1	a7
	65.1042	a1	a8
	74.4792	a2	a3
	65.1042	a2	a4
	74.6094	a2	a5
○	75.651	a2	a6
	74.6094	a2	a7
	74.4792	a2	a8
	65.1042	a3	a4
	65.1042	a3	a5
	65.1042	a3	a6
	65.1042	a3	a7
	65.1042	a3	a8
	65.1042	a4	a5
	65.1042	a4	a6
	65.1042	a4	a7
	65.1042	a4	a8
	65.1042	a5	a6
	65.1042	a5	a7
	65.1042	a5	a8
	65.1042	a6	a7
	64.3229	a6	a8
	65.1042	a7	a8

表 8.2: 属性数 3 の予測率

順位	予測率	属性 1	属性 2	属性 3
	75.3906	a1	a2	a3
	75	a1	a2	a4
	74.8698	a1	a2	a5
	76.3021	a1	a2	a6
	76.1719	a1	a2	a7
	75.2604	a1	a2	a8
	65.1042	a1	a3	a4
	65.1042	a1	a3	a5
	66.5365	a1	a3	a6
	65.625	a1	a3	a7
	65.1042	a1	a3	a8
	65.1042	a1	a4	a5
	66.5365	a1	a4	a6
	65.7552	a1	a4	a7
	65.1042	a1	a4	a8
	67.4479	a1	a5	a6
	65.3646	a1	a5	a7
	65.2344	a1	a5	a8
	68.099	a1	a6	a7
	66.7969	a1	a6	a8
	65.2344	a1	a7	a8
	74.6094	a2	a3	a4
	74.7396	a2	a3	a5
	74.7396	a2	a3	a6
	75	a2	a3	a7
	75	a2	a3	a8
	74.6094	a2	a4	a5
	75.1302	a2	a4	a6
	74.4792	a2	a4	a7
	73.4375	a2	a4	a8
	75.5208	a2	a5	a6
	75	a2	a5	a7
◎	77.8646	a2	a5	a8
	76.3021	a2	a6	a7
○	76.6927	a2	a6	a8
	65.2344	a2	a7	a8
	65.1042	a3	a4	a5
	65.1042	a3	a4	a6
	65.1042	a3	a4	a7
	65.1042	a3	a4	a8
	65.1042	a3	a5	a6
	65.1042	a3	a5	a7
	64.974	a3	a5	a8
	65.2344	a3	a6	a7
	64.974	a3	a6	a8
	64.8438	a3	a7	a8
	65.1042	a4	a5	a6
	65.1042	a4	a5	a7
	64.974	a4	a6	a8
	64.974	a4	a6	a7
	65.2344	a4	a6	a8
	64.322	a4	a7	a8
	64.7135	a5	a6	a7
	66.276	a5	a6	a8
	64.4531	a5	a7	a8
	66.0156	a6	a7	a8

$\{a1, a2, a4, a7\}$ は 75.3908%
$\{a1, a2, a4, a8\}$ は 74.4792%
$\{a1, a2, a5, a6\}$ は 75.9112%
$\{a1, a2, a5, a7\}$ は 75.9115%
$\{a1, a2, a5, a8\}$ は 74.7396%
$\{a1, a2, a6, a7\}$ は 77.3438%
$\{a1, a2, a6, a8\}$ は 76.3021%
$\{a1, a2, a7, a8\}$ は 76.1719%

ピーク $\{a1, a4, a5, a6\}$ は 77.8646 %
第２群の山の属性の組合せは、
$\{a2, a3, a4, a5\}$ は 74.4792%
$\{a2, a3, a4, a6\}$ は 75.3906%
$\{a2, a3, a4, a7\}$ は 74.0886%
$\{a2, a3, a4, a8\}$ は 73.5677%
$\{a2, a3, a5, a6\}$ は 75.9115%
$\{a2, a3, a5, a7\}$ は 74.8698%
$\{a2, a3, a5, a8\}$ は 74.349%
$\{a2, a3, a6, a7\}$ は 76.4323%
$\{a2, a3, a6, a8\}$ は 76.8229%
$\{a2, a3, a7, a8\}$ は 74.6094%
$\{a2, a4, a5, a6\}$ は 75.3906%
$\{a2, a4, a5, a7\}$ は 74.349%
$\{a2, a4, a5, a8\}$ は 73.4375%
$\{a2, a4, a6, a7\}$ は 75.7813%
$\{a2, a4, a6, a8\}$ は 76.4323%
$\{a2, a4, a7, a8\}$ は 74.0885%
$\{a2, a5, a6, a7\}$ は 76.3021%
$\{a2, a5, a6, a8\}$ は 76.4323%
$\{a2, a5, a7, a8\}$ は 74.349%
$\{a2, a6, a7, a8\}$ は 77.3438%

となっている。$\{a1, a2, a6, a7\}$ は 77.3438 %と $\{a1, a4, a5, a6\}$ は 77.8646 %が RandomForest の属性組合せの予測率を同じか、それ以上である。

表 8.4 は属性数 5 の予測率である。図 8.2 ではグラフを示している。予測率が高くなる山が 2 つある。RandomForest での属性の組合せは $\{a2, a3, a6, a7, a8\}$

第 1 群は、$\{a1, a2, a3, a4, a5\}$ は 74.7396%

$\{a1, a2, a3, a4, a6\}$ は 76.0417%

$\{a1, a2, a3, a4, a7\}$ は 75.7813%

$\{a1, a2, a3, a4, a8\}$ は 74.8698%

$\{a1, a2, a3, a5, a6\}$ は 76.5625%

$\{a1, a2, a3, a5, a7\}$ は 76.3021%

$\{a1, a2, a3, a6, a8\}$ は 74.8698%

$\{a1, a2, a3, a7, a8\}$ は 76.3021%

$\{a1, a2, a4, a5, a6\}$ は 75.651%

$\{a1, a2, a4, a5, a7\}$ は 75.3906%

$\{a1, a2, a4, a5, a8\}$ は 74.2188%

$\{a1, a2, a4, a6, a7\}$ は 76.6927%

$\{a1, a2, a4, a6, a8\}$ は 76.0417%

$\{a1, a2, a4, a7, a8\}$ は 75.7813%

$\{a1, a2, a5, a6, a7\}$ は 76.8229%

$\{a1, a2, a5, a6, a8\}$ は 75.9115%

$\{a1, a2, a5, a7, a8\}$ は 75.651%

$\{a1, a2, a6, a7, a8\}$ は 77.0833%

第 2 群は、$\{a2, a3, a4, a5, a6\}$ は 75.7813%

$\{a2, a3, a4, a5, a7\}$ は 74.349%

$\{a2, a3, a4, a5, a8\}$ は 73.5677%

$\{a2, a3, a4, a6, a7\}$ は 74.0885%

$\{a2, a3, a4, a6, a8\}$ は 76.4323%

$\{a2, a3, a4, a7, a8\}$ は 74.0885%

$\{a2, a3, a5, a6, a7\}$ は 75.7813%

$\{a2, a3, a5, a6, a8\}$ は 76.8229%

$\{a2, a3, a5, a7, a8\}$ は 74.4792%

表 8.3: 属性数 4 の予測率

順位	予測率	属性 1	属性 2	属性 3	属性 4
	74.4792	a1	a2	a3	a4
	75.3906	a1	a2	a3	a5
	76.5625	a1	a2	a3	a6
	75.9115	a1	a2	a3	a7
	74.4792	a1	a2	a3	a8
	74.4792	a1	a2	a4	a5
	75.5208	a1	a2	a4	a6
	75.3906	a1	a2	a4	a7
	74.4792	a1	a2	a4	a8
	75.9115	a1	a2	a5	a6
	75.9115	a1	a2	a5	a7
	74.7396	a1	a2	a5	a8
◎	77.3438	a1	a2	a6	a7
	76.3021	a1	a2	a6	a8
	76.1719	a1	a2	a7	a8
	65.1042	a1	a3	a4	a5
	66.276	a1	a3	a4	a6
	65.7552	a1	a3	a4	a7
	65.1042	a1	a3	a4	a8
	67.8385	a1	a3	a5	a6
	65.1042	a1	a3	a5	a7
	65.3646	a1	a3	a5	a8
	67.8385	a1	a3	a6	a7
	67.0573	a1	a3	a6	a8
	64.974	a1	a3	a7	a8
◎	77.8646	a1	a4	a5	a6
	65.2344	a1	a4	a5	a7
	65.8854	a1	a4	a5	a8
	67.9688	a1	a4	a6	a7
	66.7969	a1	a4	a6	a8
	65.7552	a1	a4	a7	a8
	67.4479	a1	a5	a6	a7
	66.9271	a1	a5	a6	a8
	66.0156	a1	a5	a7	a8
	67.0573	a1	a6	a7	a8
	74.4792	a2	a3	a4	a5
	75.3906	a2	a3	a4	a6
	74.0885	a2	a3	a4	a7
	73.5677	a2	a3	a4	a8
	75.9115	a2	a3	a5	a6
	74.8698	a2	a3	a5	a7
	74.349	a2	a3	a5	a8
	76.4323	a2	a3	a6	a7
	76.8229	a2	a3	a6	a8
	74.6094	a2	a3	a7	a8
	75.3906	a2	a4	a5	a6
	74.349	a2	a4	a5	a7
	73.4375	a2	a4	a5	a8
	75.7813	a2	a4	a6	a7
	76.4323	a2	a4	a6	a8
	74.0885	a2	a4	a7	a8
	76.3021	a2	a5	a6	a7
	76.4323	a2	a5	a6	a8
	74.349	a2	a5	a7	a8
○	77.3438	a2	a6	a7	a8
	64.974	a3	a4	a5	a6
	65.1042	a3	a4	a5	a7
	64.974	a3	a4	a5	a8
	65.1042	a3	a4	a6	a7
	65.1042	a3	a4	a6	a8
	64.1927	a3	a4	a7	a8
	64.7135	a3	a5	a6	a7
	66.0156	a3	a5	a6	a8
	64.0625	a3	a5	a7	a8
	65.625	a3	a6	a7	a8
	64.7135	a4	a5	a6	a7
	66.1458	a4	a5	a6	a8
	64.1927	a4	a5	a7	a8
	66.1458	a4	a6	a7	a8
	66.5365	a5	a6	a7	a8

$\{a2, a3, a6, a7, a8\}$ は 76.9531%
$\{a2, a4, a5, a6, a7\}$ は 76.0417%
$\{a2, a4, a5, a6, a8\}$ は 76.1719%
$\{a2, a4, a5, a7, a8\}$ は 74.0885%
$\{a2, a4, a6, a7, a8\}$ は 76.6927%
$\{a2, a5, a6, a7, a8\}$ は 77.2135%
となっている。
$\{a1, a2, a6, a7, a8\}$ は 77.0833%と $\{a2, a5, a6, a7, a8\}$ は 77.2135% は RandomForest のランキングの属性組合せの予測率を上回っている。ランキング上位の a3 よりも a1、a5 が含まれている方が予測率が高くなる現象が起きている。

表 8.5 は属性数 6 の予測率である。図 8.2 ではグラフを示している。予測率が高くなる山が 2 つある。RandomForest での属性の組合せは $\{a1, a2, a3, a6, a7, a8\}$
第 1 群は、$\{a1, a2, a3, a4, a5, a7\}$ は 75.9115%
$\{a1, a2, a3, a4, a5, a8\}$ は 74.349%
$\{a1, a2, a3, a4, a6, a7\}$ は 74.3448%
$\{a1, a2, a3, a4, a6, a8\}$ は 74.349%
$\{a1, a2, a3, a4, a7, a8\}$ は 76.4323%
$\{a1, a2, a3, a5, a6, a7\}$ は 77.7344%
$\{a1, a2, a3, a5, a6, a8\}$ は 76.9531%
$\{a1, a2, a3, a5, a7, a8\}$ は 76.0417%
$\{a1, a2, a3, a6, a7, a8\}$ は 77.474%
$\{a1, a2, a4, a5, a6, a7\}$ は 76.8229%
$\{a1, a2, a4, a5, a6, a8\}$ は 76.0417%
$\{a1, a2, a4, a5, a7, a8\}$ は 75.7813%
$\{a1, a2, a4, a6, a7, a8\}$ は 77.2135%
$\{a1, a2, a5, a6, a7, a8\}$ は 77.9948%

第 2 群は $\{a2, a3, a4, a5, a6, a7\}$ は 75.651%
$\{a2, a3, a4, a5, a6, a8\}$ は 76.8229%
$\{a2, a3, a4, a5, a7, a8\}$ は 74.6094%
$\{a2, a3, a4, a6, a7, a8\}$ は 76.5625%

表 8.4: 属性数 5 の予測率

順位	予測率	属性 1	属性 2	属性 3	属性 4	属性 5
	74.7396	a1	a2	a3	a4	a5
	76.0417	a1	a2	a3	a4	a6
	75.7813	a1	a2	a3	a4	a7
	74.8698	a1	a2	a3	a4	a8
	76.5625	a1	a2	a3	a5	a6
	76.3021	a1	a2	a3	a5	a7
	74.8698	a1	a2	a3	a5	a8
	77.3438	a1	a2	a3	a6	a7
	76.8229	a1	a2	a3	a6	a8
	76.3021	a1	a2	a3	a7	a8
	75.651	a1	a2	a4	a5	a6
	75.3906	a1	a2	a4	a5	a7
	74.2188	a1	a2	a4	a5	a8
	76.6927	a1	a2	a4	a6	a7
	76.0417	a1	a2	a4	a6	a8
	75.7813	a1	a2	a4	a7	a8
	76.8229	a1	a2	a5	a6	a7
	75.9115	a1	a2	a5	a6	a8
	75.651	a1	a2	a5	a7	a8
◎	77.0833	a1	a2	a6	a7	a8
	67.1875	a1	a3	a4	a5	a6
	64.8438	a1	a3	a4	a5	a7
	64.7135	a1	a3	a4	a5	a8
	69.2708	a1	a3	a4	a6	a7
	66.276	a1	a3	a4	a6	a8
	65.4948	a1	a3	a4	a7	a8
	69.0104	a1	a3	a5	a6	a7
	67.4479	a1	a3	a5	a6	a8
	66.0156	a1	a3	a5	a7	a8
	68.6198	a1	a3	a6	a7	a8
	67.8385	a1	a4	a5	a6	a7
	67.0573	a1	a4	a5	a6	a8
	66.0156	a1	a4	a5	a7	a8
	67.3177	a1	a4	a6	a7	a8
	68.099	a1	a5	a6	a7	a8
	75.7813	a2	a3	a4	a5	a6
	74.349	a2	a3	a4	a5	a7
	73.5677	a2	a3	a4	a5	a8
	75.2604	a2	a3	a4	a6	a7
	76.4323	a2	a3	a4	a6	a8
	74.0885	a2	a3	a4	a7	a8
	75.7813	a2	a3	a5	a6	a7
	76.8229	a2	a3	a5	a6	a8
	74.4792	a2	a3	a5	a7	a8
○	76.9531	a2	a3	a6	a7	a8
	76.0417	a2	a4	a5	a6	a7
	76.1719	a2	a4	a5	a6	a8
	74.0885	a2	a4	a5	a7	a8
	76.6927	a2	a4	a6	a7	a8
◎	77.2135	a2	a5	a6	a7	a8
	64.8438	a3	a4	a5	a6	a7
	65.8854	a3	a4	a5	a6	a8
	64.3229	a3	a4	a5	a7	a8
	65.3646	a3	a4	a6	a7	a8
	66.5365	a3	a5	a6	a7	a8
	66.1458	a4	a5	a6	a7	a8

$\{a2, a3, a5, a6, a7, a8\}$ は 77.474%
$\{a2, a4, a5, a6, a7, a8\}$ は 76.8229%
となる。
RandomForest の属性ランキングの予測率を越えている属性の組合せは、
$\{a1, a2, a5, a6, a7, a8\}$ は 77.9948%と
$\{a1, a2, a3, a5, a6, a7\}$ は 77.7344%と
$\{a2, a3, a5, a6, a7, a8\}$ は 77.474%である。ランキングの属性 a3 の代わりに a5 が入っている組合せが予測率が高くなっている。この組合せは全組合わせで最高の予測率となっている。

図の 8.5 が属性数 7 の場合である。これは Leave-one-out を行った場合と同じである。a2 の属性がない場合、最も予測率が下がる。RandomForest のランキングでは、a4 を抜いた場合が最も予測率が高い。

8.5 まとめ

予測率が 77 %台が最も高い予測率であるが、属性数が 4 以上から 77 %を出している。特に a2 の属性が加わることで予測率が急激にあがることでわかる。少数の属性が予測に寄与している例である。
Leave-one-out の場合、$a2 > a6 = a7 > a1 > a3 = a8 > a5 > a4$ の順番である。
この組合せでの予測率を列挙する。
○ $\{a2\}$ 74.6094 %
○ $\{a2, a6\}$ 75.651 %
$\{a2, a7\}$ 74.6094 %
◎ $\{a1, a2, a6, a7\}$ 77.3438 %
$\{a1, a2, a3, a6, a7\}$ 76.8229 %
○ $\{a1, a2, a3, a6, a7, a8\}$ 77.474 %
○ $\{a1, a2, a3, a5, a6, a7, a8\}$ 77.6042 %

○印は、Randomforest のランキングと一致している属性の組合せを表す。◎は、RandomForest のランキングを越えた予測率の属性の組合せを表す。Randomforest の寄与度の順位と同じ数が出てきているが、予測率最高の組合せ $\{a1, a2, a5, a6, a7, a8\}$ は出てきていない。よって Leave-one-out では予測率が最高の属性の組合せを選択しない場合が起こり

表 8.5: 属性数６，７，８の予測率

順位	予測率	属性１	属性２	属性３	属性４	属性５	属性６
	65.1042	a1	a2	a3	a4	a5	a6
	75.9115	a1	a2	a3	a4	a5	a7
	74.349	a1	a2	a3	a4	a5	a8
	77.3438	a1	a2	a3	a4	a6	a7
	74.349	a1	a2	a3	a4	a6	a8
	76.4323	a1	a2	a3	a4	a7	a8
◎	77.7344	a1	a2	a3	a5	a6	a7
	76.9531	a1	a2	a3	a5	a6	a8
	76.0417	a1	a2	a3	a5	a7	a8
○	77.474	a1	a2	a3	a6	a7	a8
	76.8229	a1	a2	a4	a5	a6	a7
	76.0417	a1	a2	a4	a5	a6	a8
	75.7813	a1	a2	a4	a5	a7	a8
	77.2135	a1	a2	a4	a6	a7	a8
◎	77.9948	a1	a2	a5	a6	a7	a8
	67.9688	a1	a3	a4	a5	a6	a7
	67.4479	a1	a3	a4	a5	a6	a8
	65.8854	a1	a3	a4	a5	a7	a8
	67.5781	a1	a3	a4	a6	a7	a8
	67.5781	a1	a3	a5	a6	a7	a8
	66.7969	a1	a4	a5	a6	a7	a8
	75.651	a2	a3	a4	a5	a6	a7
	76.8229	a2	a3	a4	a5	a6	a8
	74.6094	a2	a3	a4	a5	a7	a8
	76.5625	a2	a3	a4	a6	a7	a8
◎	77.474	a2	a3	a5	a6	a7	a8
	76.8229	a2	a4	a5	a6	a7	a8
	66.1458	a3	a4	a5	a6	a7	a8

順位	予測率	属性１	属性２	属性３	属性４	属性５	属性６	属性７
	76.9531	a2	a3	a4	a5	a6	a7	a8
	67.1875	a1	a3	a4	a5	a6	a7	a8
	77.3438	a1	a2	a4	a5	a6	a7	a8
○	77.6042	a1	a2	a3	a5	a6	a7	a8
	77.474	a1	a2	a3	a4	a6	a7	a8
	76.4323	a1	a2	a3	a4	a5	a7	a8
	76.4323	a1	a2	a3	a4	a5	a6	a8
	77.3438	a1	a2	a3	a4	a5	a6	a7

順位	予測率	属性１	属性２	属性３	属性４	属性５	属性６	属性７	属性８
○	77.8646	a1	a2	a3	a4	a5	a6	a7	a8

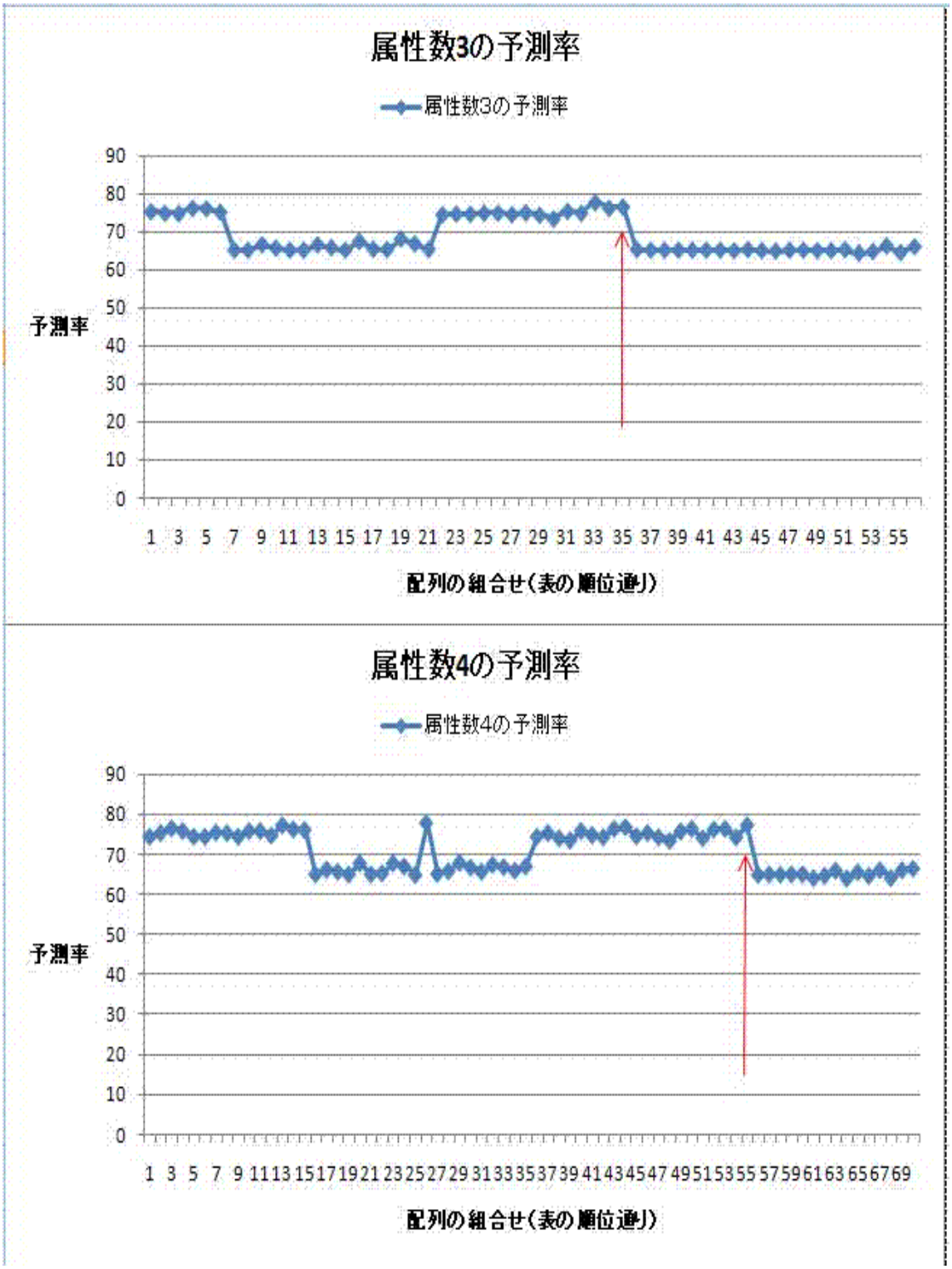


図 8.1: 属性数 3,4 の予測率のグラフ

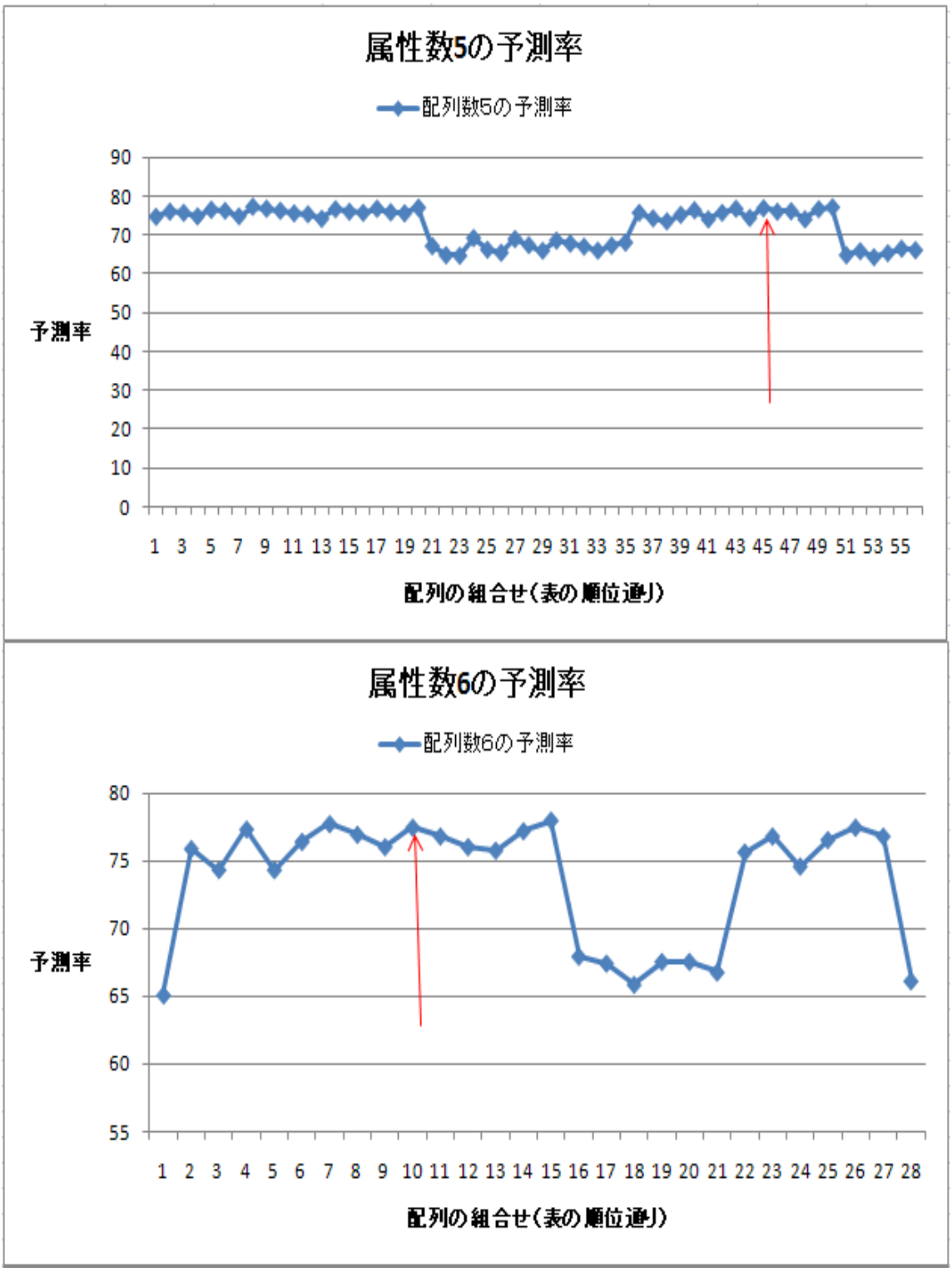


図 8.2: 属性数 5,6 の予測率のグラフ

うる。

計算機実験が示すことは、予測率が高い属性の組合せは密集しており、MeanDecreaseGini係数による選択される属性の組合せ以外の近傍を探索することが有効であることが分かる。特徴ベクトルが sliding window で表されている場合、同じ辞書式順序で配列されるため同様の傾向を示すことが予想される。データ依存の結果ではあるが、近傍探索が有効であることは示されている。

第 9 章

まとめ

9.1 結論

本研究において得られた新たな知見としては以下のことが挙げられる。第 5 章での主論文では、エピジェネティクス現象を示す酵母菌の配列データをデータセットとして活性化・不活性を示す判別分析の計算機実験を行った。先行研究 [14] では、上位 3 位までの sliding window で作成した特徴ベクトルの属性を infromative feature として列挙していたが、本研究では RandomForest の variable importance を用いて寄与度を全て計算して、定量評価をしランキングを行った。その結果、寄与度のグラフと予測率には関連があることが分かった。寄与度のグラフは以下の 3 種類に分かれる。属性数が 0 から variable importance の高い属性を順次追加していく前向き手法で、

1. 最初は急勾配で、次になだらかな勾配のグラフが続くが、裾野が薄いグラフ
2. 最初は急勾配で、次になだらかな勾配のグラフが続くが、裾野が厚いグラフ
3. なだらかな勾配が続くが裾野が厚いグラフ

があった。予測率が高いグラフは 1 番目の裾野の薄いグラフである。これは、正例と負例との境界領域と推測される。その知見は、第 6 章での SOM での可視化でも裏付けられる。予測率が比較的低い寄与度のグラフは第 2 , 3 番目のグラフに見られる。裾野が厚いまたは、裾野が薄いが横に長い場合も予測率が比較的低くなっている。これは、正例と負例の境界領域にある属性が多いことを示している。また特に sliding window での頻度の特徴ベクトルでは、3-gram での頻度が均等であることが予測率が低い原因とも考えられる。

第 6 章では、その他に属性間の相関係数を調べた。その結果複数のデータにおいて相関係数の高い属性が見つかった。

第5章では、variable importanceに基づいて属性部分集合選択を行った。探索の戦略として第8章の予備実験に基づいた近傍探索 (Local Search) を提案した。これは、

1. variable importanceでのランキングでSVMで判別分析を行う。その際、前向きの探索方向で、予測率が最高の属性の組合わせを探す。

2. 1の予測率が最高の属性の組合せの前後3近傍を含めた7属性の全ての組合わせについてSVMでの判別分析を行う。

比較研究としてBayesNet、NaiveBayes、SMO、J48、AdaBoostM1、RandomForestの判別解析を用い、探索方法してBestFirst、GeneticSearch、 GreedyStepwise、LinearForwardSelection、RankSearchを用い計算機実験を行った。その結果rank searchのほかにGAを利用した手法が探索の速度と選択された属性の予測率が高かった。一般的に、探索の戦略としてヒューリステック手法が用いられるが、データ数が巨大な場合、近傍探索が実用的である。属性部分集合選択の計算機実験では、提案手法が全データに対してわずかではあるが上回っている。そして、近傍探索により予測率が先行研究よりもよい属性の部分集合が求められた。また、機械学習のパラメータによってn個のランキングの予測率も大きく変わるアルゴリズムがあることが計算機実験により分かったがパラメータのチューニングに関しては今後の研究課題とする。

第7章では、位置特異的な情報による特徴ベクトルを作成し、判別解析を行った。データセットは第5章の同様のデータセットを用いた。3塩基のトリプレット単位で考えた場合、RandomForestのランキングで6割程度の予測率だが、ランキング上位を組合わせると8割近くの数字になる。

第8章では、Randomforestのランキングの属性組合せと全探索空間$2^n - 1$での予測率の分布を調べた。その結果、Randomforestのランキングの属性組合せ以外にも、全探索空間$2^n - 1$で予測率が上回る組合せが数組みあることが分かった。また、Leave-one-outでも属性の組合せでは、全探索した場合の最高予測率は探索できなかった。この生物学的意味は、離れたところで相互作用を引き起こしている部位を意味していると考えられる。

9.2 今後の課題

今後の課題としては、

・機械学習のパラメータと属性選択の手法との関連が挙げられる。パラメータの設定によって予測率の値が大きく変化する。両者の関係はまだあまり研究されていない。

・属性選択の戦略の計算量を比較することである。これは、アルゴリズムの分野のテーマになるが、まだ戦略についてはあまり提案が少ない。まだアルゴリズム研究の提案または、計算の高速化において研究の余地があるものと考えている。

・属性間の相関性を考慮した研究が提案されているが、現在の randomForest を用いた機械学習との比較実験を行う [19][20]。

・遺伝子配列には固有の性質がある。突然変異により欠損、置換などの揺らぎがある。本研究では、この揺らぎには考慮されていない。

・ベイズ統計で提案されている生成モデルの次元削減の手法 (LDA,latent dirichit allocation）との比較を予備審査で指摘されたが今回はできなかったため、今後の課題としたい。

・本研究は 2008 年に発表したものである。この頃から variable importance に関する統計学的性質の理論的研究が行われている。主に bias に関する研究とその問題を修正した提案などがなされているが、まだあまり少ないので、今後の課題とする。

謝辞

本研究を行なうに当たり, 御指導を賜わった佐藤賢二教授に深謝致します。

また、中森先生、池田先生、国藤先生には様々なご配慮を頂きありがとうございました。

博士審査をおこなって頂いたHo先生、ダム先生に深く感謝いたします。Ho研究室の方々には、大変お世話になりました。

修士の指導教官であり、博士課程では、副テーマにおいて熱心に指導を賜りました田島敬史先生には深く感謝いたします。

また、中森研の上野國光さんには、輪講や様々な有益な議論をして頂きました。今回の論文でもその議論がアイデアの一つとなっています。

現在、奉職しております法政大学理工学部経営システム工学科の先生方には、様々な配慮を賜り深く感謝いたします。

参考文献

[1] マウント デービッド W., 岡崎 康司, 坊農 秀雅, “バイオインフォマティクス ゲノム配列から機能解析へ 第2版”, メディカル・サイエンス・インターナショナル,2005.

[2] T.A.Brown, “ゲノム 第3版—新しい生命情報システムへのアプローチ ”, メディカル・サイエンス・インターナショナル,2007.

[3] 元田浩, 津本周作, 山口高平, 沼尾正行, “データマイニングの基礎”, オーム社.

[4] Huan Liu, Hiroshi Motoda, “Computational Methods of Feature Selection (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)”. *Chapman and Hall/CRC* ,2007.

[5] Trevor Hastie,Robert Tibshirani,Jerome Friedman, “The Elements of Statistical Learning:Data Mining, Inference, and Prediction.Second Edition(Springer Series in Statistics) ”. *Springer-Verlag*.

[6] D.K. Pokholok et al., “Genome-wide Map ofNucleosome Acetylation and Methylation”, *Cell*,Vol.122, pp.517-527.

[7] Eran Segal et al., “A genomic code for nucleosome positioning”, *Nature*,2006.

[8] 杉本知之, 下川敏雄, 後藤昌司, “樹木構造接近法と最近の発展”, 計算機統計学　第 *18* 巻・第 *2* 号：*2005 p.123-p.164*.

[9] 金森 敬文, 畑埜 晃平, 渡辺 治, 渡辺 治, “ブースティング - 学習アルゴリズムの設計技法 (知能情報科学シリーズ)”, 森北出版株式会社,2006.

[10] 赤穂 昭太郎, “カーネル多変量解析—非線形データ解析の新しい展開 (シリーズ確率と情報の科学)”, 岩波書店,2008.

[11] C. M. ビショップ, 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇, “パターン認識と機械学習 上 - ベイズ理論による統計的予測”, シュプリンガー・ジャパン株式会社,2007.

[12] C. M. ビショップ, 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇, “パターン認識と機械学習 下 - ベイズ理論による統計的予測”, シュプリンガー・ジャパン株式会社,2008.

[13] Yvan Saeys,Inaki Inza,and Pedoro Larranaga, “A review of feature selection techniques in bioinfomatics” *Bioinfomatics*, Vol.23, No.19,pp.2507-2517(2007).

[14] T. H. Pham, D. H. Tran, T. B. Ho, K. Satou and G. Valiente, “Qualitatively Predicting Acetylation and Methylation Areas in DNA sequences” *Genome Informatics*, Vol.16, No.2,pp.3-11(2005).

[15] D. H. Tran, T. H. Pham, K. Satou and T. B. Ho, “Conditional Random Fields for Predicting and Analyzing Histone Occupancy, Acetylation and Methylation Areas in DNA Sequences, Applications of Evolutionary Computing”, *Lecture Notes in Computer Science*, Vol.3907,pp.221-230(2006).

[16] Zenglin Xu, Rong Jin, Jieping Ye, Michael R. Lyu, and Irwin King. “Non-Monotonic Feature Selection”, it Proceedings of the 26th International Conference on Machine Learning (ICML2009), Montreal, Quebec,, 2009.

[17] 新島 聡, 奥野 恭史, “化合物-タンパク質活性空間における特徴選択”, *IBIS2009*,2009.

[18] 押村 光雄, “注目のエピジェネティクスがわかる—ゲノムの修飾・構造変換と生命の多様性、疾患との関わり (わかる実験医学シリーズ—基本 &トピックス)”, 羊土社 ,2004.

[19] Hai Nguyen,Katrin Franke and Slobodan Petrovic, “Optimizing a Class of Feature Selection Measures”. *NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML)*, 2009.

[20] Hanchun Peng,Fuhui Long and Chris Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependancy,Max-Relevance, and Min-Redundancy”. *IEEE*

TRANSACTION ON FATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol 27,No 8, 2008.

[21] Ian B.Dodd, Mille A.Micheelsen,Kim Sneppen, and GenevieveThom, "Theoretical Analysis of Epigenetic Cell Memory by Nucleosome Modification". *Cell*,Vol.129,pp.813-822,2007.

[22] John A. Lee, Michel Verleysen, "Nonlinear dimensionality reduction", *Springer-Verlag*,2008.

[23] George A. F. Seber, "Multivariate Observations (Wiley Series in Probability and Statistics)", *Wiley*,1984.

[24] 矢部 博史, "工学基礎 最適化とその応用 (新・工科系の数学)", 数理工学社,2006.

[25] H. Liu et al, "Evolving Feature Selection", *IEEE Intelligent Systems*,Vol.20, No.6, 2005,pp. 64-76.

[26] L. Breimanr, "Random Forests, ", *Machine Learning*,Vol.45, No.1, 2001, pp. 5-32.

[27] Harb R, Yan X, Radwan E, Su X, "Exploring precrash maneuvers using classification trees and random forests., ", *Accid Anal Prev*,Vol.41, No.1, 2009, pp. 98-107.

[28] Oppel S, Powell AN, Dickson DL, "Exploring precrash maneuvers using classification trees and random forests., ", *J Anim Ecol*,Vol.78, No.3, 2009, pp. 524-531.

[29] Fan J, Nunn ME, Su X, "Multivariate Exponential Survival Trees And Their Application to Tooth Prognosis.,", *Comput Stat Data Anal*,Vol.53, No.4, 2009, pp. 1110-1121.

[30] Fan J, Nunn ME, Su X, "Multivariate Exponential Survival Trees And Their Application to Tooth Prognosis.,", *Comput Stat Data Anal*,Vol.53, No.4, 2009, pp. 1110-1121.

[31] Tang R, Sinnwell JP, Li J, Rider DN, de Andrade M, Biernacka JM, "Identification of genes and haplotypes that predict rheumatoid arthritis using random forests., ", *BMC Proc*,Vol.3, Supppl.7, 2009.

[32] Culp M, Johnson K, Michailidis G, "The ensemble bridge algorithm: a new modeling tool for drug discovery problems., ", *J Chem Inf Model.*,Vol.50, No.2, 2010.

[33] Culp M, Johnson K, Michailidis G, "The ensemble bridge algorithm: a new modeling tool for drug discovery problems., ", *J Chem Inf Model.*,Vol.50, No.2, 2010.

[34] de Maat MF, van de Velde CJ, Benard A, Putter H, Morreau H, van Krieken JH, Meershoek Klein-Kranenbarg E, de Graaf EJ, Tollenaar RA, Hoon DS, "Identification of a quantitative MINT locus methylation profile predicting local regional recurrence of rectal cancer., ", *Clin Cancer Res*,Vol.16, No.10, 2010, pp.2811-2818.

[35] Julia Lasserre, "Hybrid of Generative and Discriminative Methods for Machine Learning, ", *PhD thesis*,March 2008, University of Cambridge.

[36] Julia Lasserre, "Principled Hybrids of Generative and Discriminative Models, ", *IEEE CVPR* ,Vol1.No.1,2006,pp.87-94.

[37] Julia Lasserre, "Generative or discriminative? getting the best of both worlds, ", *BAYESIAN STATISTICS*,Vol8.No.1,2007,pp.3-24.

[38] C. Strobl, A.L. Boulesteix and T. Augustin , "Unbiased split selection for classification trees based on the Gini index. ", *Computational Statistics Data Analysis* ,Vo52.No.1,2007,pp.483-501.

[39] C. Strobl, A.L. Boulesteix and T. Augustin , "Danger: High Power!-Exploring the statistical properties of a test for random forest variable importance. ", *COMPSTAT 2008 - Proceedings in Computational Statistics* ,Vo II ,2008,pp.59-66,Physica Verlag, Heidelberg.

[40] C. Strobl, A.L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis , "Conditional variable importance for Random Forests. ", *BMC Bioinformatics* ,Vo.9,No.37,2008.

[41] C. Strobl, "Statistical Issues in Machine Learning - Towards Reliable Split Selection and Variable Importance Measures, ", *PhD thesis* ,2008,Ludwig-Maximilians-University (LMU) Munich.

[42] C. Strobl, J. Malley and G. Tutz, "An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests.", *Psychological Methods* ,Vo.14,No.4,p. 323-348,2009.

[43] C. Strobl, T. Hothorn and A. Zeileis, “Party on! A new, conditional variable importance measure for random forests available in the party package. ”, *The R Journal* ,2009.

[44] 佐藤 寿彦, “Diagnostic decision support system for headache using SOM (SOM を用いた頭痛診断補助装置)”, *PhD thesis* 社,March 2010, 千葉大学.

[45] Miron B. Kursa, Witold R. Rudnicki, “Feature Selection with the Boruta Package. ”, *Journal of Statistical Software* ,Vo.36,No.11,p.1-13,2010.

[46] 小西貞則, 越智 義道, 大森 裕浩, “計算統計学の方法—ブートストラップ・EM アルゴリズム・MCMC ”, 朝倉書店.,2008.

[47] 金明哲, “R とブートストラップ”, *ESTRELA*.,No.156,p.58-p.63,2007.

[48] 汪金芳・田栗正章, “計算統計 I ：統計科学のフロンティア 11”, 岩波書店.,2003.

[49] O. Hobert, “Gene Regulation by Transcription Factors and MicroRNAs,. ”, *Science* ,Vo.319,No.5871,p.1785-1786,2008.

本研究に関する発表論文

◎ジャーナル（査読あり）

Higashihara.M., Rebolledo-Mendez.J.D., Yamada.Y., Satou.K.: “Application of a Feature Selection Method to Nucleosome Data: Accuracy Improvement and Comparison with Other Methods”, WSEAS Transactions on Biology and Biomedicine, Vol.5, Issue 5, pp.95-104,2008.5.

◎国際会議（査読あり）

Higashihara.M., Rebolledo-Mendez.J.D., Yamada.Y., Satou.K.: “Ranking and Selection of Features for Improved Prediction of Nucleosome Occupancy and Modification”, Proc. of the 9th WSEAS International Conference on MATHEMATICS & COMPUTERS IN BIOLOGY & CHEMISTRY (MCBC ’08), pp.188-193,2008. 6.

◎国内会議（査読なし）

東原正智, “位置情報を考慮したエピジェネティクス関連領域の予測と属性部分集合選択に関する研究”, 電子情報通信学会第 21 回データ工学ワークショップ, 第 8 回日本データベース学会年次大会 (DEIM ’10), 電子情報通信学会第 21 回データ工学ワークショップ (DEIM 2010) 論文集.

東原正智, 佐藤賢二., “RandomForest を用いたエピジェネティクス関連領域の予測と属性選択”, 電子情報通信学会第 19 回データ工学ワークショップ, 第 6 回日本データベース学会年次大会 (DEWS ’08), 電子情報通信学会第 19 回データ工学ワークショップ (DEWS 2008) 論文集.

○本研究以外に発表した論文

◎国内会議（査読あり）

東原正智, 田島敬史., “問合せ処理に適した XML 圧縮形式に関する研究”, 電子情報通信学会第 15 回データ工学ワークショップ, 第 2 回日本データベース学会年次大会 (DEWS ’04), 電子情報通信学会第 15 回データ工学ワークショップ (DEWS 2004) 論文集.

◎国内会議（査読なし）

東原正智, 田島敬史., “XBW 変換による圧縮 XML データ上の構造問合せの性能評価”, 第 3 回データ工学と情報マネジメントに関するフォーラム, 第 9 回日本データベース学会年次大会 (DEIM ’11).

○第 1 著者以外

◎ジャーナル (査読あり)

Rebolledo-Mendez.J.D., Higashihara.M., Yamada.Y., Satou.K.: “Characterization and Clustering of GO Terms by Feature Importance Vectors Obtained from Microarray Data”, WSEAS Transactions on Biology and Biomedicine, Vol.5, Issue , pp.163-172,2008.7.

Yamada.Y., Miyata.Y., Higashihara.M., Satou.K.: “Comparison of Cluster Identification Methods for Selection of GO Terms related to Gene Clusters”, WSEAS Transactions on Biology and Biomedicine, Vol.5, Issue 3, pp.54-63,2008.3.

◎国際会議 (査読あり)

Rebolledo-Mendez.J.D., Higashihara.M., Yamada.Y., Satou.K.: “Finding Hidden Relationship among Biological Concepts in Gene Ontology”, Proc. of the WSEAS International Conference on BIOMEDICAL ELECTRONICS and BIOMEDICAL INFORMATICS (BEBI’08), pp.237-241,2008. 8.

Yamada.Y., Miyata.Y., Higashihara.M., Satou.K.: “Estimation of Identification Methods of Gene Clusters Using GO Term Annotations from a Hierarchical Cluster Tree”, Proc. of the 9th WSEAS International Conference on MATHEMATICS & COMPUTERS IN BIOLOGY & CHEMISTRY (MCBC ’08), pp.194-198,2008. 6.