科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１８　日現在

機関番号：　３２６０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００２７４

研究課題名（和文）無限要素を持つノンパラメトリックベイズモデルによる生命情報予測アルゴリズム研究

研究課題名（英文）Nonparametirc Bayes-based infinite mixture model algorithms for Bioinformatics

研究代表者
鏑木　崇史（Kaburagi, Takashi）

青山学院大学・理工学部・助教

研究者番号：１０４６８８６１

交付決定額（研究期間全体）：（直接経費）　　2,200,000円、（間接経費）　　660,000円

研究成果の概要（和文）：本研究の目的は、無限要素を持つノンパラメトリックベイズモデルを適用して、タンパク質機能予測問題と遺伝子制御ネットワーク構造予測問題へ適用することを目標とした。タンパク質機能予測には隠れマルコフモデル(HMM)とベイジアンネットワーク(BN)の複合数理モデルを無限混合拡張したアルゴリズムを、遺伝子発現ネットワークのモデル設計にノンパラメトリックモデルとダイナミックベイジアンネットワークを用いて構築されたベイズ的枠組みを,リバーシブルジャンプマルコフ連鎖モンテカルロ法（RJMCMC 法）を用いることで推定する手法をそれぞれ提案した。

研究成果の概要（英文）：We proposed a non-parametric Bayesan models to two bioinformatics applications: 1) automatic protein function prediction and 2) gene expression network inference. For automatic protein function prediction,we proposed a novel method to predict protein functions, called PreGO. PreGO is an algorithm based on an infinite mixture of hidden Markov models. Given an unannotated protein sequence, PreGO predicts the probability of existence of Gene Ontology terms. For time-varying network inference for gene expression data, we adopted a nonparametric Bayesian regression method to predict interactions between the genes. This method is expected to achieve more flexible regression capability in time-varying network. To obtain stronger robustness to noisy data, we employed the T-Process. The basic algorithm employed reversible jump Markov Chain Monte Carlo for inference of whole network structures. The method can handle (i) change point detection and (ii) network structure inference simultaneously.

研究分野：　総合領域

科研費の分科・細目：　情報学・感性情報学・ソフトコンピューティング

キーワード：　バイオインフォマティクス　ベイズ学習　隠れマルコフモデル　ベイジアンネットワーク　ノンパラメトリックベイズモデル　タンパク質機能予測　遺伝子発現データ　時系列データ解析

１．研究開始当初の背景

本研究の目的は、無限要素を持つノンパラメトリックベイズモデルを適用して、網羅的タンパク質機能予測問題と遺伝子制御ネットワーク構造予測問題の精度向上を目指す事である。

タンパク質機能や遺伝子制御ネットワーク構造などの生命情報には非線形性・不確定性を含む事が多い。その複雑な生命情報の特徴を的確に捉えるためには、利用する数理モデルが極力柔軟性を備えていることが望ましい。本研究では、既存のモデルよりかなり柔軟性を持つことで、近年機械学習の分野で注目されつつあるノンパラメトリックベイズモデルに着目した。ノンパラメトリックベイズモデルは背景構造が未知の複雑なデータから自動的に適切な学習を行う事ができるため、非線形性・不確定性を含むデータについても的確にモデル化でき、網羅的タンパク質機能予測問題と遺伝子制御ネットワーク構造予測問題の精度向上に繋がることが期待できる。

本研究では生命情報予測問題のうち、生命科学と医学の視点から重要とされる次の２つに注目する。

【A】　網羅的タンパク質機能予測

ポストゲノム時代の現在、次世代シーケンサーなどの登場により、多くのアミノ酸配列を短時間で求めることができるようになった。しかしながら、タンパク質の機能までが判明しているものはきわめて少なく、アミノ酸配列データベース TrEMBL の 5%に満たない。このような背景のもと、信頼性が高い網羅的タンパク質機能予測アルゴリズムの開発は急務と言われている。

【B】　遺伝子制御ネットワーク予測

遺伝子はタンパク質を符号化し、タンパク質は遺伝子発現量を制御する。したがって一般に遺伝子制御ネットワークはフィードバックを含むダイナミックなネットワークと考えられる。その構造の解明は、構造そのものに対する生物学的興味以外に、困難な疾患の原因解明とその治療への足がかりとして重要な役割を演じている。

これらの予測問題には少なくとも２点の共通点がある：

(1)データの背後にある複雑性と非線形性

(2)データが含む確率的不確定性

上記の２点を考慮すると「非線形かつ複雑なデータをモデル化できる柔軟なモデルの使用」と、「不確定性をとらえる枠組みの検討」がきわめて重要である。

２．研究の目的

本研究では上述の柔軟なモデルと不確定性を考慮して、各課題に対し無限要素を持つノンパラメトリックベイズモデルを適用することを提案する。具体的には以下のような目標を設定した。

【A】に対し提案するアルゴリズでは、アミノ酸配列を用いてタンパク質の機能を予測することを目標とした。タンパク質の機能としては木構造で体系的にまとめられた遺伝子オントロジー(GO)を利用する。機能既知のタンパク質ではアミノ酸配列と機能情報の両方が利用可能である。一方、機能未知のタンパク質では機能情報が与えられず、アミノ酸配列のみが利用できる。アミノ酸配列はアミノ酸間の位置関係を時系列とみなす状態数無限の隠れマルコフモデル(HMM)でモデル化され、機能情報は各機能間の依存を表現するためベイジアンネットワーク(BN)でモデル化される。この HMM と BN を組合せた複合数理モデルが無限個混合されている混合モデルを提案する。アミノ酸配列を用いてそれぞれの複合モデルの尤度を求め、無限個の重み付き平均を求めることにより、機能の有無を確率として求める。本研究ではディリクレ過程事前分布を仮定することにより、無限の要素を持つ提案モデルでも計算が可能となる。

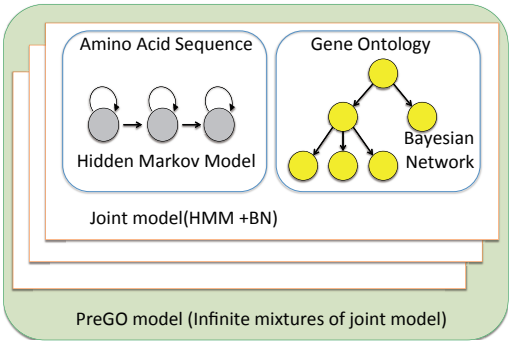

図１　本研究で用いた HMM-BN 複合数理モデル "PreGO"モデル

【B】に対しては、ネットワークを構成する各ノードが遺伝子に対応し、ノード間をつなぐネットワークが遺伝子の依存関係を示すと考えた。申請者がこれまでに提案してきたモデルでは各遺伝子間の関係性が線形であることを仮定してきた。しかし実際の遺伝子発現データは必ずしも単純な線形結合で表現されるような単純なものではなく、非線形性を持つと考えるのが自然である。ここでは非線形性を考慮する目的で無限中間素子ニューラルネットワークを 用いたベイジアンネットワークモデルを利用する。提案するモデルでは Gaussian Process 事前分布を仮定することで、通常必要のあるニューラルネットワークの重みパラメタを積分消去（周辺化）できるため、より柔軟に依存関係を表現できることが期待される。実装にあたっては高精度な Monte Carlo 法を用いた。

３．研究の方法

【A】に対しては隠れマルコフモデル(HMM)とベイジアンネットワーク(BN)の複合数理モデルを無限混合拡張したものを、【B】に対

しては無限中間素子ニューラルネットワークを用いたベイジアンネットワークを用いる。
本研究はそれぞれのモデルに対し、基本アルゴリズムの構築とその精度向上をベイズ的枠組みでの改良などを通して目指すものである。

４．研究成果

【A】 網羅的タンパク質機能予測研究成果

ベイジアンネットワークは,不完全な事象の連鎖において,それらの相互作用を確率的に扱うことを目的としたモデルの一種であり,確率変数をノードで表し,各ノード間の因果関係や相関関係といった依存関係をリンクで表す.また,リンクは因果関係を示す 方向を向き,リンクの向きにグラフを辿った時にそのグラフが循環しないという非循環有向グラフで表される.
本研究では、数理モデルのパラメタと、観測されたデータ、ここではアミノ酸配列yと機能情報xが与えられたときに、どの程度適合しているかの指標（尤度）が必要になる。HMM のパラメタをΦ、ベイジアンネットワークのパラメタをΨとしたとき、尤度は次のように定義される：

$$P(y|\Phi) = \sum_{z} P(y,z|\Phi)$$

$$= \sum_{z} P(y|z,\Phi)P(z|\Phi)$$

$$P(x|\Psi) = \prod_{i=1}^{N} P(x_i|\mathrm{Pa}(x_i),\Psi_i)$$

ここでzは隠れ状態列、Nは機能の総数、$Pa()$は変数x_iに対する親ノードのすべてである。
本研究では、この複合数理モデルに重みp_kをつけた線形和で示される混合モデルとして扱う。つまり

$$P(x,y|\Phi,\Psi) = \sum_{k=1}^{K} p_k P(x|\Psi_k)P(y|\Phi_k)$$

$$\sum_{k=1}^{K} p_k = 1$$

さらに、タンパク質データはいくつの集合でモデル化するのが適切か不明なため、$K \to \infty$とするのが適切である。本研究では、Dirichlet Process 事前分布による無限混合モデルを採用した。Dirichlet Process 事前分布には幾つかの表現方法があるが、本研究では Stick-Breaking 表現を用いた。
パラメタΦ、Ψ、p_kの学習には最大事後確率(MAP)期待値最大化法（EM）を用いた。
提案したアルゴリズムの予測精度を比較するため、実験を行った。データセットはタンパク質配列数 378、機能数 153 であり、10 分割交差検定を行った。
ここで、比較手法として Baseline を定義する。この方法では、学習データにおける機能の出現頻度のみを用いたものであり、国際的なタンパク質機能予測コンテストでも比較用に持ちられた方法である。性能の指標としては ROC 曲線(Receiver Operating Characteristic)を用いて比較する。（図２）この方法では、縦軸に True Positive Rate(真陽性)と横軸に False Positive Rate(偽陽性)としてプロットしたものである。ROC 曲線下の面積（Area under the curve，AUC）はアルゴリズムの性能の良さを表す。0 から 1 までの値をとり，完全な分類が可能なときの面積は 1 で，ランダムな分類の場合は 0.5 になる。図３は１０回交差検証を行った手法間の AUC を表している。混合数を固定したときは混合数 30 がもっとも性能が高かった。一方、混合数を無限とした場合、真の混合数を与える事なく、固定の最高値とほぼ同等の性能を得る事ができた。

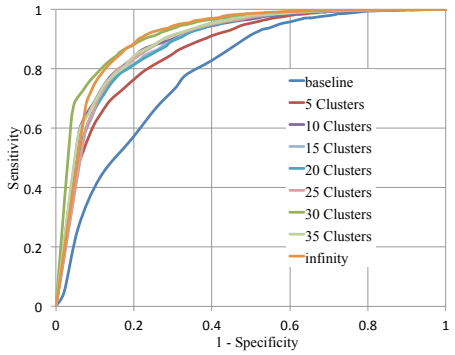

図２　ベースライン手法と提案手法の性能比較（ROC）

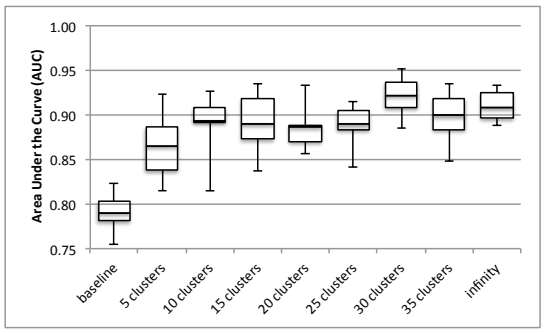

図３　ベースライン手法と提案手法の性能比較（AUC）
各種法のプロットは下から最低値、第一四分位、中央値、第三四分位、最高値を示す

【B】 遺伝子制御ネットワーク予測研究成果

本研究では,ノンパラメトリックモデルを用いることで,遺伝子発現データから遺伝子発現ネットワークを 推定することを主目的とする.更に,遺伝子発現生成メカニズムに内在すると言われる非線形性をノンパラメ トリックモデルを用いることで捉えるのみならず,遺伝子発現ネットワーク構造の突発的な変化に柔軟に対応 するアルゴリズムの構築を目指す.こうした変化は,例えば,変態生物の形態変化に伴う遺伝子間の相互作用 の強さの変化に対応する形で現れると考えられる.提案手法においては,変化前後のネットワー

ク構造を独立 なものと捉え,これらをまとめて推定対象とする.具体的には,遺伝子発現ネットワークのモデル設計にノンパラメトリックモデルとダイナミックベイジアンネットワークを用いて構築されたベイズ的枠組みを,リバーシブルジャンプマルコフ連鎖モンテカルロ法 (RJMCMC 法) を用いることで推定する.提案手法を,人 工的に作成したデータセットに対して適用しその有効性を確かめた後,キイロショウジョウバエ (Drosophila melanogaster) の筋肉の発達を司る遺伝子群のデータセットに適用した.

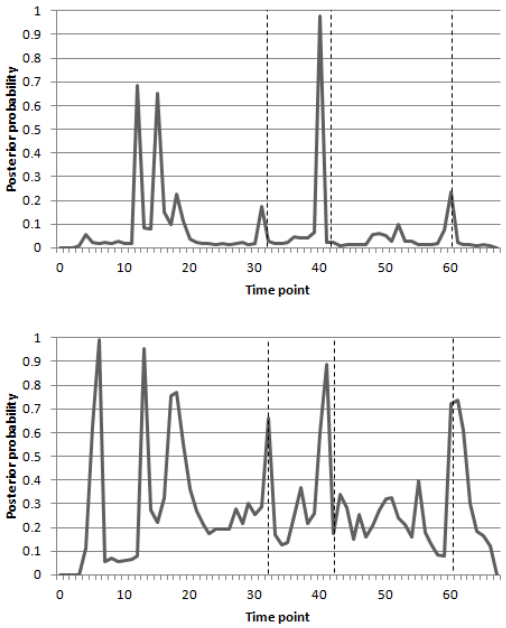

図４　構造変化点位置推定結果.上段 : 提案手法による推定結果.下段 :Dondelingerらの先行研究による推定結果. 縦点線は形態変化が起きる時刻を示している.

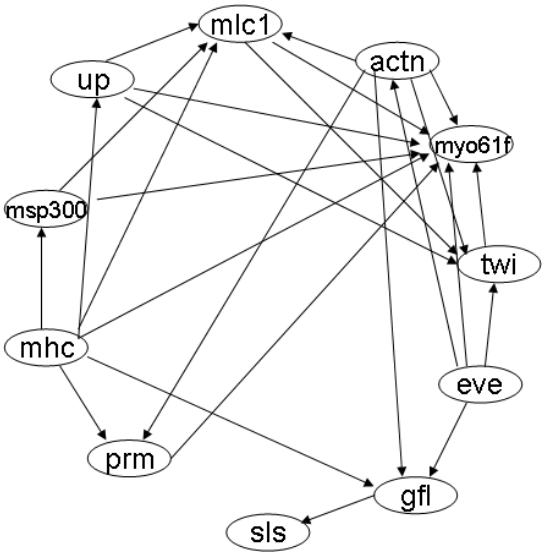

図５　embyo フェーズにおけるネットワーク構造推定結果 (閾値 = 0.25).

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕(計　６　件)

① PreGO: A Protein Function Prediction Algorithm Based on an Infinite Mixture of Hidden Markov and Bayesian Network Models
T. Kaburagi, Y. Koizumi, K. Oota, T. Matsumoto
International Conference on Bioinformatics and Computational Biology 2013 (2013 年 3 月)

② Nonparametric Bayes-based heterogeneous Drosophila melanogaster gene regulatory network inference: T-Process regression.
H. Miyashita, T. Nakamura, Y. Ida, T. Kaburagi, T. Matsumoto
The 12th International Association of Science and Technology for Development (IASTED) International Conference on Artificial Intelligence and Applications (2013 年 2 月)

③ ノンパラメトリックベイジアン T 過程アルゴリズム を用いた時間的構造変化を考慮した遺伝子発現ネットワーク推定
宮下弘樹, 中村拓磨, 井田安俊, 鏑木崇史, 松本隆
情報処理学会, 第 91 回数理モデル化と問題解決・第 32 回バイオ情報学合同研究会 (2012 年 12 月)

④ Protein Function Prediction Algorithm Based on Infinite State Hidden Markov Model and Bayesian Network Model,
T. Kaburagi, Y. Koizumi, G. Kobayashi, T. Matsumoto,
International Conference Intelligent Systems for Molecular Biology 2012 (2012 年 7 月).

⑤ Protein Function Prediction Algorithm Based on Infinite State Hidden Markov Model and Bayesian Network Model,
T. Kaburagi, Y. Koizumi, G. Kobayashi, T. Matsumoto,
International Conference Intelligent Systems for Molecular Biology 2012 (2012 年 7 月).

⑥ Infinite Mixture Model Approach for Protein Function Prediction Algorithm Utilizing Hidden Markov Model and Bayesian Network Model with Dirichlet Process Prior
T. Kaburagi, Y. Koizumi, G. Kobayashi, K. Oota, Y. Nakada, T. Matsumoto,
International Conference Intelligent Systems for Molecular Biology 2011 (2011

年７月).

６．研究組織

(1)研究代表者

鏑木　崇史（KABURAGI, Takashi）
青山学院大学・理工学部・助教
研究者番号：10468861