

平成 22 年 6 月 29 日現在

研究種目：若手研究（スタートアップ）

研究期間：2008～2009

課題番号：20810034

研究課題名（和文） 確率的ダイナミカルシステムとしての生物構造予測とその高精度化

研究課題名（英文）

研究代表者

鍋木 崇史（KABURAGI, Takashi）

早稲田大学・理工学術院・助手

研究者番号：10468861

研究成果の概要（和文）：

この研究はタンパク質のうち様々な分野から注目されている膜タンパク質の構造を予測するアルゴリズムを構築するものである。膜タンパク質構造予測問題は生命科学のみならず、創薬の分野などへの効果も期待されている。膜タンパク質は生体膜に存在し、細胞間情報伝達や物質の輸送など生命活動の基本的機能を果たしている。ゲノム中の 20～30% は膜タンパク質をコードしているといわれており、その構造解明は創薬を含めいくつかの分野で重要と考えられている。しかし、膜タンパク質は結晶化が困難であり、立体構造が解明されているタンパク質のうち膜タンパク質は 1% 程度とまだ少ない。このような背景から機械学習による膜タンパク質構造予測アルゴリズムの開発が注目されている。

本研究では機械学習手法の 1 つ、隠れマルコフモデル (Hidden Markov Model, HMM) を用いる。HMM は時系列性のあるデータのモデル化に広く用いられており、手書き文字認識、音声認識、署名照合などの幅広い分野に応用されている。HMM では観測されるデータの背後に「状態」が存在すると仮定する手法である。本研究では膜タンパク質の特徴的な構造をグラフ構造として表現し、膜貫通領域の数と膜貫通領域を予測するアルゴリズムを構築することを目指した。

具体的には膜タンパク質構造の予測問題に対し、HMM の Bayes 的アプローチをモンテカルロ実装したアルゴリズムを提案し、国際会議にて報告した。その際、新たな課題も見つかった。その対策としてモデルを Dirichlet Process 先験分布を用いて拡張することを提案し、初期的な実験を行った。

また、遺伝子はタンパク質を符号化し、一方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興味以外に、いくつかの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。この問題に対しても機械学習的手法により多くの成果がある。

遺伝子発現制御ネットワーク予測問題に対しては、グラフ構造の事前知識としてスケールフリー性を考慮したアルゴリズム提案し国際会議にて報告し、論文化した。さらに、データの非正規性・非線形性を考慮する必要が確認されたため、2 つのアプローチを試みた。データの非正規性には Box-Cox 変換を用いたデータ正規化アルゴリズムを、非正規性には無限中間素子ニューラルネットワークを用いたアルゴリズムを提案し、ごく初期的な実験結果を得つつある。

研究成果の概要（英文）：

交付決定額

(金額単位：円)

	直接経費	間接経費	合 計
2008年度	900,000	270,000	1,170,000
2009年度	360,000	108,000	468,000
年度			
年度			
年度			
総 計	1,260,000	378,000	1,638,000

研究分野：

科研費の分科・細目：

キーワード：

1. 研究開始当初の背景

バイオインフォマティクスにおけるいくつかの予測問題に対して機械学習的手法によるアプローチは多くの成果を生み出している。生物データには不確実性が含まれる事が多いため、確率・統計的手法は自然な枠組みといえよう。従って対象とする量の確率分布を計算することになるが、現実の生物データでは当該確率分布を解析的に表示する事は困難、あるいは不可能の場合がままある。この研究の目的はバイオインフォマティクスにおける2つの予測問題をBayes的枠組みから定式化し、対象確率分布計算をモンテカルロ実装して予測精度の向上を目指す事である。

2. 研究の目的

本研究では2つの研究目標、

【A】タンパク質構造予測

【B】遺伝子制御ネットワーク推定

に着目して研究を推進した。

これらの予測問題には少なくとも2つの共通点がある：

1. データの背後にある逐次性

遺伝子制御ネットワークにおいては遺伝子発現量がダイナミックに動くという意味で、時間が自然な形で逐次性を与えている。タンパク質の折りたたみは時間に関する逐次性を内包しているが、例えばX線回折像には時間に関するダイナミクスはない。一方しかし、タンパク質を1次元の鎖と考えれば、空間に関する逐次性を備えていると考えられ、それは例えばHMM (Hidden Markov Model)で定式化される。

2. データは不確実性を含むことが多い

これらの予測問題に取り組むためには逐次性をいかにとらえるか、と不確実性をどのような枠組みでとらえるかが本質的である。

3. 研究の方法

本研究を進めるにはいくつかの条件が必要であった：

1. 対象を的確にとらえたモデル構築

2. 健全な予測アルゴリズムの構築と予測確率分布の計算手法の実装

3. ふさわしいデータ

本研究では前述の通り2つの研究目標、

【A】タンパク質構造予測

【B】遺伝子制御ネットワーク推定

のそれぞれに対し、以下の2つのアプローチで達成することを目標とした。

<アプローチ1> 確率的ダイナミカルシステムとしてのモデル化とその改良

<アプローチ2> Bayes的枠組みでの高精度モンテカルロ法の実装

4. 研究成果

各年度で得られた本研究の成果は以下のとおりである。

平成20年度

平成20年度は特に<アプローチ1>に注力をして研究を進めた。

【A】については、現在の基本アルゴリズムを拡張し、ベイジアンモンテカルロHMMを用いた。また、モデル構造の改良には一般化HMMなどを導入した。(論文2.学会4.6.7.)

【B】にはZipf則を考慮した基本モデルの構築ならびにその改良を行った。(論文1.学会5.)

平成 21 年度

平成 21 年度は平成 20 年度で得られた結果を用いて、＜アプローチ 2＞を特に注力をして研究を進めた。

【A】への適用を予定しているベイジアンモンテカルロ HMM を Dirichlet Process 先験分布に用いる混合モデルの提案・実装を試みた。(学会 1.)

【B】に対してはデータの非正規性・非線形性を考慮する必要が確認され、対応策を試みた。(学会 2.3.)

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)

〔雑誌論文〕(計 2 件)

1. "Monte Carlo-based Mouse Nuclear Receptor Superfamily Gene Regulatory Network Prediction: Stochastic Dynamical System on Graph with Zipf Prior", Y. Kitamura, T. Kimiada, J. Maruyama, T. Kaburagi, T. Matsumoto, and K. Wada, IPSJ Transactions on Bioinformatics (accepted), 2010.

2. "Backward Smoothing Approach to Transmembrane Protein Structure Prediction with Stochastic Dynamical Systems", T. Kaburagi, and T. Matsumoto, J. of Comp. Intelligence in Bioinformatics Vol. 1(No. 1) pp. 13-33, Jun. 2008.

〔学会発表〕(計 7 件)

1. "A Novel Gene Ontology Prediction Algorithm Using Infinite Mixtures of Hidden Markov and Binary Models with a Dirichlet Process Prior", T. Kaburagi, N. Tagoto, T. Matsumoto ISMB 2010 (accepted)

2. "A Gene Regulatory Network Prediction Algorithm Using a Gaussian Bayesian Network Model with a Box-Cox Transformation" H. Miyachika, J. Maruyama, T. Kaburagi, Y. Nakada, T. Matsumoto, T. Kimiada, K. Wada, ISMB2010 (accepted).

3. "A non-parametric Bayesian algorithm for predicting gene regulatory networks with a Gaussian process", T. Kikuchi, Y. Nakada, T. Kaburagi, T. Matsumoto, ISMB2010 (accepted).

4. "A Bayesian Monte Carlo Hidden Markov Model Approach to Transmembrane Protein Structure Prediction", T. Kaburagi and T. Matsumoto, 17th Annual International Conference on Intelligent Systems for Molecular Biology & 8th European Conference in Computational Biology, Jun.

2009.

5. "Monte Carlo-Based Bayesian Prediction of Gene Regulatory Networks with Zipf Distribution: Mouse Nuclear Receptor Superfamily", H. Miyachika, Y. Kitamura, T. Kimiada, J. Maruyama, T. Kaburagi, T. Matsumoto, K. Wada, 17th Annual International Conference on Intelligent Systems for Molecular Biology & 8th European Conference in Computational Biology, Jun. 2009.

6. "状態維持確率にポアソン分布を用いた一般化隠れマルコフモデルによる膜タンパク質構造予測", T. Kaburagi and T. Matsumoto, 第 14 回バイオ情報学研究会, Sep. 2008

7. "An Approach to Predict Transmembrane Protein Structure with Stochastic Dynamical Systems Using Backward Smoothing Scheme", T. Kaburagi and T. Matsumoto, 16th Annual International Conference Intelligent Systems for Molecular Biology, Jul. 2008.

〔図書〕(計 件)

〔産業財産権〕

○出願状況(計 件)

名称:

発明者:

権利者:

種類:

番号:

出願年月日:

国内外の別:

○取得状況(計 件)

名称:

発明者:

権利者:

種類:

番号:

取得年月日:

国内外の別:

〔その他〕

セミナー

1. "Transmembrane Secondary Structure Prediction: A Hidden Markov Model Approach and its Extensions", T. Kaburagi, 医薬基盤研究所 バイオインフォマティクスプロジェクト, July, 2009.

2. "Developing Transmembrane Protein Structure Prediction Algorithms: Monte

Carlo HMM ”, T. Kaburagi and T. Matsumoto, TWIns Joint Symposium, Dec. 2009.

3. “Predicting transmembrane structures with stochastic dynamical systems: An approach using generalized hidden Markov model”, T. Kaburagi, Japan-Slovenia Seminar on Nonlinear Science, (Invited) Nov. 2008

4. “二次元物理・化学量出力を持つ隠れマルコフモデルによる膜タンパク質構造予測とその拡張”, T. Kaburagi, 生命情報工学研究センター, (Invited) Jun, 2008.

6. 研究組織

(1)研究代表者

鍋木 崇史 (KABURAGI, Takashi)

早稲田大学・理工学術院・助手

研究者番号：10468861

(2)研究分担者

()

研究者番号：

(3)連携研究者

()

研究者番号：