様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　25　年　5　月　27　日現在

機関番号：14301
研究種目：若手研究(B)
研究期間：2009～2011
課題番号：２１７００３２３
研究課題名（和文）　タンパク質立体構造における類似部分構造の大域的抽出と解析

研究課題名（英文）　Global extraction and analysis of protein tertiary structure patterns

研究代表者
林田　守広（HAYASHIDA MORIHIRO）
京都大学・化学研究所・助教
研究者番号：40402929

研究成果の概要（和文）：
タンパク質立体構造をラベル付きグラフとして表現したときに，実装に依存しないグラフ圧縮の手法を開発し，これを利用したグラフ間の類似度を提案した．また立体構造をＣα原子間の距離行列として表現したときに,行列を画像とみなし画像に対する文法を定義,できるだけ小さい文法を生成する手法を開発した．さらに最小の木文法を生成する手法も開発した．またドメイン情報を用いて，タンパク質間相互作用とタンパク質複合体を予測する手法を開発した．

研究成果の概要（英文）：
We developed a graph compression method that does not depend on implementation, and proposed a similarity measure between graphs. We regarded a distance matrix between Cα atoms of a protein as an image, and proposed a grammar-based image compression method. In addition, we proposed a grammar-based tree compression method. Furthermore, we developed prediction methods using protein domains for protein-protein interactions and protein complexes.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,000,000	300,000	1,300,000
2010 年度	700,000	210,000	910,000
2011 年度	700,000	210,000	910,000
総　計	2,400,000	720,000	3,120,000

研究分野：総合領域
科研費の分科・細目：情報学・生体生命情報学
キーワード：バイオインフォマティクス

１．研究開始当初の背景

タンパク質の機能を明らかにすることは生物学，医学などにとって重要である．多くのタンパク質はいくつかの特徴的な部分構造から構成されていると考えられ，これまでアミノ酸配列のアラインメントに基づくか，専門家の手によって見出されてきた．

２．研究の目的

本研究の目的は，タンパク質の詳細な内部構造を把握することで，タンパク質の機能推定に役立てることである．

３．研究の方法

(1) 多数のタンパク質立体構造から類似する部分構造を自動的に抽出するアルゴリズムを開発する．そのためにはまず，タ

ンパク質立体構造を扱いやすいデータ構造で表現することが必要である．その上で類似する部分構造を探索する．この目的のために，頂点にラベルの付いたグラフに対して効率良く類似部分を探索するアルゴリズムを開発した．

(2) 一般のグラフを生成するような最小のグラフ文法を効率良く見つけることは未だ困難であるので，辺にラベルの付いた根付き木構造について最小の文法を見つける整数計画法による手法を開発した．木文法の生成規則として，チョムスキー標準形を木構造へ応用した，縦または横に木を二分割する規則をもつ分割型木文法を用いた．

(3) タンパク質の立体構造をＣα原子間の距離行列として表現する．距離行列は画像とみなせ，画像圧縮を利用することでコルモゴロフ複雑性に基づき，立体構造間の類似度を提案した．また，画像に対する生成規則を定義し，できるだけ小さい文法を見つける近似アルゴリズムを開発した．

(4) ドメインはタンパク質内に存在する構造的または機能的なユニットであると考えられており，タンパク質間に相互作用があることは内部のドメイン間で相互作用があるとみなせる．この考えに基づき条件付き確率場を用いたタンパク質間相互作用予測手法を開発した．

(5) タンパク質の機能を明らかにする上で，アミノ酸残基単位での相互作用の理解は有用である．そこで条件付き確率場の一種で画像処理の分野で用いられている識別確率場をアミノ酸残基間の相互作用モデルとして導入した．

(6) タンパク質相互作用ネットワークから予測されたタンパク質複合体の検証手法に関する研究を行った．先行研究ではタンパク質ドメインの相互作用に着目し構造的な制約から，ある一つのドメインは高々一つのドメインと相互作用するという仮定のもと複合体中でのタンパク質間の相互作用数を最大化した．しかしこれは複合体を大きく二分する可能性があるので，できるだけ一つの大きくまとまった複合体を取ってくるように整数計画問題を改良した．またこれとグラフ理論での概念である極大成分などとの組合せ手法を提案した．

(7) ヒトなどの数種の生物種について，一つのタンパク質に含まれるドメインの種類と総数の分布はそれぞれ指数分布，べき乗則に従う分布となっている．そこで両方の分布が同時に現れるような，生物学的知識に基づいたタンパク質のドメイン獲得モデルを提案した．

４．研究成果

(1) ラベル付きグラフに対する圧縮アルゴリズムを検証するために，いくつかの主要な生物種の代謝ネットワーク間の類似度を計算した．代謝ネットワークには，各頂点に化合物がラベルとして付けられているので，純粋にネットワークの構造を比較するには適さない．そこで本研究では，化合物の構造情報に利用されるモルガンインデックスをラベルに用いた．モルガンインデックスは隣接するモルガンインデックスの足し合せをある条件が満たされるまで繰り返す．同じネットワークでも繰り返し回数が異なるとラベルが違ってくるため，繰り返し回数は固定する．このようにして求めた代謝ネットワーク間の類似度に対して，最短距離法によるクラスタリングを行い，一般に知られているような系統樹と矛盾のない結果が得られた．しかしいくつかの課題も残された．一つは開発したアルゴリズムが高速ではあるが不可逆であることである．つまり，得られたグラフ文法からもとのグラフをいつも再構成できるとは限らない．類似構造を抽出する観点からは必ずしも可逆である必要はないが，同じ規則によって縮約される部分グラフはある類似度以内であることが保証された方がよい．逆に全く同一の部分グラフのみを縮約する規則のみを生成するようなアルゴリズムは効率が悪くなり，多数の立体構造を扱うことが困難になるため，できるだけ効率性を失わずにアルゴリズムを改良することが今後の課題となる．

(2) 提案した分割型木文法圧縮手法は，部分問題として，指定した大きさの木文法が存在するかどうかを判定する整数計画問題を解く．最小の文法はこの大きさをいくつか試すことで得られる．整数計画問題を効率良く解くアルゴリズムが開発されてはきているが，人工的に作成した子頂点を多くもつ木に対しては 25 頂点で 7 時間程度を要する一方，子頂点の数が少ない木に対してはある程度の頂点数まで効率良く計算できた．さらに木構造であることが知られている，いくつかの糖鎖について，辺のラベルを根とは逆側の頂点の分子名として提案手法を適用し，実際の細胞内での糖鎖の形成と比較はしていないが，最小の生成文法を見出した．

(3) タンパク質立体構造間の類似度を求めるための画像圧縮手法をウェーブレット変換に似たアルゴリズムを改良することにより開発した．いくつかの立体構造のクラスに属するタンパク質群を用いた計算機による検証では，提案手法が他の圧縮手法に比べて良い分類精度を示した．

(4) ドメイン間の相互作用の指標として，ドメインに含まれるアミノ酸配列を用いて多重配列アラインメントを求め，ドメイン間に含まれる残基間での相互情報量のうち最大のものを利用した．開発した条件付き確率場を用いたタンパク質間相互作用予測手法において，このドメイン間の最大相互情報量を用いたモデルと用いないモデル，さらに主要な既存手法である EM 法と比較した．その結果，最大相互情報量を用いたモデルがテストデータに対する予測精度が最も優れていた．

(5) 画像に対する確率場にはしばしば二次元格子が用いられ，格子点には画素が対応する．一方残基間相互作用モデルでは，格子点をアミノ酸配列中の各残基のペアに対応させ，画素値の代わりに多重配列アラインメントから得られる残基間の相互情報量を利用した．これは相互作用するアミノ酸残基どうしは，相互作用を維持するために片方のアミノ酸残基が変異を受けて置換されればもう片方もこれに伴って変化するという考えに基づく．識別確率場の特徴として周囲の格子点の情報も条件付き確率に反映される点がある．いくつかのタンパク質のペアについて計算機実験による検証を行った結果，識別確率場を用いた方が周囲の格子点の情報を伴わないマルコフ確率場を用いるよりも残基間相互作用の予測精度が高かった．しかしながらまだ実用には耐えられないため改良が必要である．

(6) タンパク質相互作用ネットワークから予測されたタンパク質複合体の検証手法については，MCL と MCODE に対する計算機実験により，先行研究の手法よりも精度，再現率の両方で本研究による手法が上回った．

(7) 変異による新たなドメインの形成，遺伝子重複によるタンパク質の複製，遺伝子融合によるタンパク質配列の結合からは，数理的な解析により，ドメインの種類，総数について指数分布となることを証明した．さらにタンパク質内部でのドメインの複製を考慮することで，ドメインの種類について指数分布のまま，総数についてべき乗則に従う分布となることを解析と計算機シミュレーションから検証した．

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 7 件）

① Zhao, Y., Hayashida, M., Nacher, J. C., Nagamochi, H. and Akutsu, T., Protein complex prediction via improved verification methods using constrained domain-domain matching, International Journal of Bioinformatics Research and Applications, 査読有, vol. 8, 2012, 210-227
DOI: 10.1504/IJBRA.2012.048970

② Hayashida, M., Ruan, P. and Akutsu, T., A quadsection algorithm for grammar-based image compression, Integrated Computer-Aided Engineering, 査読有, vol. 19, 2012, 23-38
DOI: 10.3233/ICA-2012-0389

③ Hayashida, M. and Akutsu, T., Measuring the similarity of protein structures using image compression algorithms, IEICE Transactions on Information and Systems, 査読有, vol. E94-D, no. 12, 2011, 2468-2478
http://search.ieice.org/bin/summary.php?id=e94-d_12_2468

④ Hayashida, M., Kamada, M., Song, J. and Akutsu, T., Conditional random field approach to prediction of protein-protein interactions using domain information, BMC Systems Biology, 査読有, vol. 5, Suppl. 1, 2011, S8
DOI:10.1186/1752-0509-5-S1-S8

⑤ Zhao, Y., Hayashida, M. and Akutsu, T., Integer programming-based method for grammar-based tree compression and its application to pattern extraction of glycan tree structures, BMC Bioinformatics, 査読有, vol. 11, Suppl 11, 2010, S4
DOI:10.1186/1471-2105-11-S11-S4

⑥ Hayashida, M. and Akutsu, T., Comparing biological networks via graph compression, BMC Systems Biology, 査読有, vol. 4, Suppl. 2, 2010, S13
DOI:10.1186/1752-0509-4-S2-S13

⑦ Nacher, J. C., Hayashida, M. and Akutsu, T., The role of internal

duplication in the evolution of multi-domain proteins, BioSystems, 査読有, vol. 101, no. 2, 2010, 127-135
DOI:10.1016/j.biosystems.2010.05.005

〔学会発表〕(計6件)

① Zhao, Y., Hayashida, M., Nacher, J., Nagamochi, H. and Akutsu, T., Protein complex prediction via improved verification methods using constrained domain-domain matching, The 10th Asia-Pacific Bioinformatics Conference, 2012/1/19, メルボルン, オーストラリア
② Kamada, M., Hayashida, M., Song, J. and Akutsu, T., Discriminative random field approach to prediction of protein residue contacts, 2011 IEEE Conference on Systems Biology, 2011/9/3, 珠海, 中国
③ Zhao, Y., Hayashida, M. and Akutsu, T., Integer programming-based method for grammar-based tree compression and its application to pattern extraction of glycan tree structures, The 21st International Conference on Genome Informatics, 2010/12/16, 杭州, 中国
④ Hayashida, M., Kamada, M., Song, J. and Akutsu, T., Conditional random field approach to prediction of protein-protein interactions using mutual information between domains, The Fourth International Conference on Computational Systems Biology, 2010/9/9, 蘇州, 中国
⑤ Hayashida, M., Ruan, P. and Akutsu, T., A quadsection algorithm for grammar-based image compression, The 2010 International Conference on Advanced Science and Technology, 2010/6/25, 宮崎
⑥ Hayashida, M. and Akutsu, T., Comparing biological networks via graph compression, The Third International Symposium on Optimization and Systems Biology, 2009/9/20, 張家界, 中国

〔図書〕(計0件)

〔産業財産権〕
○出願状況(計0件)

○取得状況(計0件)

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
林田　守広(HAYASHIDA MORIHIRO)
京都大学・化学研究所・助教
研究者番号：40402929

(2)研究分担者

(3)連携研究者