

科学研究費助成事業 研究成果報告書

平成 29 年 5 月 31 日現在

機関番号：82401

研究種目：若手研究(B)

研究期間：2013～2016

課題番号：25870197

研究課題名(和文) ネットワーク解析と比較ゲノムを応用した薬剤標的分子を予測する数理モデルの構築

研究課題名(英文) Computational framework to infer potential novel drug-targets based on network analysis and comparative genomic analysis

研究代表者

長谷 武志 (Hase, Takeshi)

国立研究開発法人理化学研究所・統合生命医科学研究センター・客員研究員

研究者番号：70569285

交付決定額(研究期間全体)：(直接経費) 3,300,000円

研究成果の概要(和文)：タンパク質間相互作用ネットワーク(PIN)は、遺伝子の機能や薬剤標的分子の探索のための、有用なリソースである。本課題では、主に2つの研究に取り組んだ。人のPINの情報を使用し、薬剤標的分子を予測する計算手法の開発を行った。この計算手法を使用し、アルツハイマー病、関節リウマチ、そして、いくつかのがんに対する、新規の薬剤標的候補分子を予測した。多様な生物種間における、PINの構造の、類似性と差異について調査を行った。さらに、比較ゲノム解析、および、遺伝子重複に基づくネットワーク成長モデルを使用したコンピューターシミュレーションを行い、このような類似性と差異を説明できる進化メカニズムを探索した。

研究成果の概要(英文)：Protein-protein interaction networks are useful resources for investigating gene functions and potential drug targets. In this project, we mainly conducted two studies. By using genome-wide data of protein-protein interactions, we developed a computational framework to infer potential drug-target genes. Then, we used the framework and inferred potential novel therapeutic targets for Alzheimer's disease, rheumatoid arthritis, and several cancerous diseases. We investigated differences and similarities in the structure of protein-protein interaction networks among various species. In order to investigate potential evolutionary mechanisms that can explain the differences and similarities, we conducted comparative genome analyses and computer simulations that use network growth model based on gene duplications.

研究分野：情報学

キーワード：生命情報 システムバイオロジー ネットワークバイオロジー 機械学習 タンパク質ネットワーク
比較ゲノム 薬剤標的分子 ネットワーク進化

1. 研究開始当初の背景

現在、Yeast Two-hybrid 法等のハイスループットな実験手法によって、大量のタンパク質間相互作用のデータが蓄積されている。網羅的なタンパク質間相互作用ネットワークは、多数の遺伝子の関係をシステムとして理解するための、有用な研究資源として利用されている。

今まで、我々は、タンパク質間相互作用ネットワークの構造の解析、および、その構造と薬剤標的分子との関係性について研究を行ってきた。その結果、中程度の相互作用数をもつタンパク質はお互いに密に結合し合っているが、相互作用数の多いタンパク質間の相互作用は抑制されていることを発見した(引用文献)。さらに、薬剤標的分子をヒトのタンパク質間相互作用ネットワークにマップしたところ、薬剤標的となる蛋白質は、相互作用数が中程度から低いところには殆ど存在しないことが判った。このことは、相互作用数が多いタンパク質を薬剤の標的としてしまった場合、その副作用が大きすぎるために標的としては適当ではないことを示唆する。実際、抗癌剤の標的タンパク質は一般的に副作用がその他の薬のものより大きい傾向があるが、抗癌剤の標的タンパク質の相互作用数はその他の薬剤の標的タンパク質よりも有意に大きい(引用文献)。また、中程度の相互作用数の遺伝子群中に、synthetic lethality に関係する遺伝子が最も高確率で存在することも発見した。以上の研究は、タンパク質間相互作用の情報は、薬剤標的分子の予測や副作用などについて、システムの視点からの理解を助ける重要なツールとなることを示唆している。

また、我々は、生物の進化をシステムとして理解するためにも、タンパク質間相互作用ネットワークを使用してきた。具体的には、我々は、タンパク質間相互作用ネットワークには、局所的に相互作用が集中した特徴的な構造が存在し、この構造はランダムに遺伝子重複や相同遺伝子間の相互作用の保存を行う進化モデルでは説明できないことを発見した。一方、我々が構築した、相同遺伝子間の相互作用が保存される割合が共有する結合相手の数に依存する過程を組み込んだ進化モデルにより、この構造が自然に説明できることを示した(引用文献)。以上から、遺伝子および相互作用の進化と、それらのネットワーク上の位置には関係性があることが示唆される。

さらに、我々は、マラリア原虫の相互作用ネットワークの構造が、マラリア原虫以外(酵母、線虫、ショウジョウバエ、ヒト)と大きく異なることを発見した。55種の真核生物のゲノムの比較およびシミュレーションにより、この違いは、遺伝子重複のし易さの相互作用数に対する依存性が、マラリア

とその他の生物で大きく異なることによって生じることが示された(引用文献)。以上の結果と、マラリア原虫と他の生物と生息している環境が大きく違うことから、環境がネットワーク構造とその進化に影響を与えている可能性が示唆される。

2. 研究の目的

本研究の目的は、主に、パブリックドメインのタンパク質間相互作用情報を利用して、(1)それぞれの遺伝子が、薬剤ターゲット候補となるかどうかを予測する計算手法を構築すること、(2)それぞれの病気に対して、drug repositioning の対象となる可能性のある薬剤を推測する計算手法を構築すること、(3)原核生物のタンパク質間相互作用ネットワークの進化メカニズムの解明、および、(4)生物種間における、タンパク質間相互作用ネットワークの構造の違いを明らかにすること、である。

3. 研究の方法

上記の目的に必要な情報を集めるために、公共のデータベースから、様々な生物種のゲノムワイドなタンパク質間相互作用ネットワークに関する情報を取得する。

これらのネットワークから有用な情報を抽出することを目的として、各種ネットワーク統計量の計算、および、サブネットワーク分割などを行う。抽出した情報を基にして、薬剤標的分子を予測するための classifier モデルを構築する。

原核生物を含む、様々な種のタンパク質間相互作用ネットワークの解析を行う。これらの生物種間での、ネットワーク構造の比較を行う。遺伝子重複に基づく進化モデルを基盤とした、シミュレーションを行い、このモデルが、生物種間での、ネットワーク構造の類似性と差異を説明できるかどうか検証する。さらに、比較ゲノム解析を行うことで、シミュレーションに使用した進化モデルの妥当性について、検証を行う。

これらの解析のために必要なプログラムを作成する。

4. 研究成果

(1)ゲノムワイドなタンパク質間相互作用ネットワークは複雑で巨大であり、このようなネットワークから、有用な情報を取り出すことは困難である。この問題を解決するために、我々は、巨大なネットワークを複数のシンプルで小さいサブネットワークに分割する方法を開発した。さらに、この方法を、人のタンパク質間相互作用ネットワークに適用することにより、薬剤標的分子を有意に多く含む、複数のサブネットワークの同定に成功した。これらのサブネットワークについて詳細な調査を行い、抗がん剤の標的分子の半分以上を含む、2つのサブネットワークを発

見した。また、pathway enrichment analysis を行い、この2つのサブネットワークは、がんの成長や生存に重要な役割を果たすパスウェイ(vascular endothelial growth factor signaling pathway など)と強い関係があることを発見した。このようなサブネットワークを利用することにより、新しい薬剤標的分子の探索を、効率良く進めることが可能になると期待される。この研究成果は、2014年11月に *Artificial Life and Robotics* 誌に掲載された(雑誌論文5)。

(2) タンパク質間相互作用ネットワークを含む、3種の分子間相互作用ネットワークの情報を用いて、薬剤標的分子を予測する computational workflow の構築を行った(図1)。

この computational workflow は、以下の3ステップで構成される。(i) ネットワーク解析による特徴量抽出、(ii) 抽出した特徴量から、予測に有用な特徴量を選択するステップ、(iii) 選択した特徴量を用いた予測モデル(classifier モデル)の構築、であ

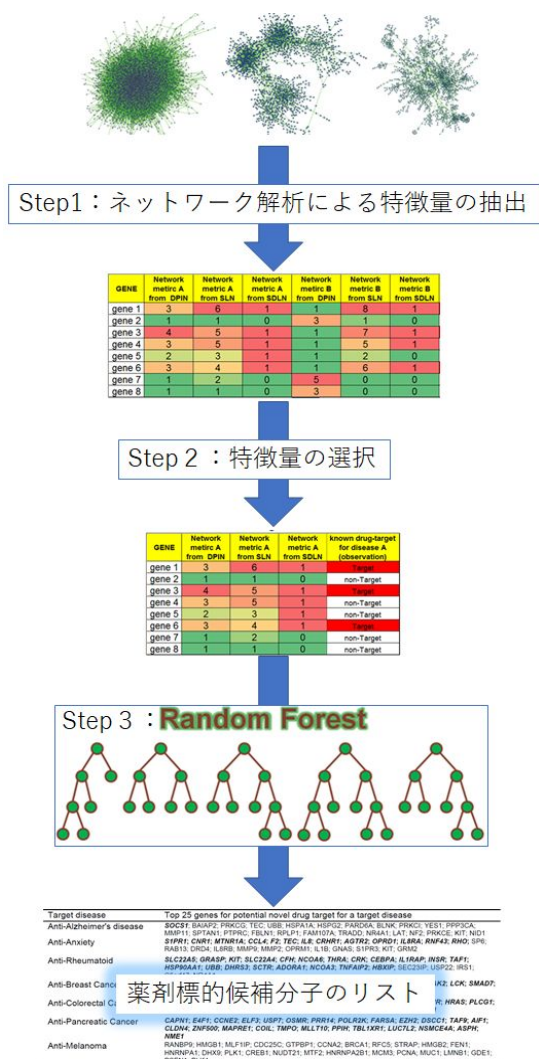


図1. 薬剤標的分子を予測するための computational workflow (学会発表より改変)

具体的には、(i)では、それぞれのネットワークにおいて、25種のネットワーク統計量を計算した。これらの25種のネットワーク統計量は、それぞれのタンパク質に対して計算される。この解析の結果、それぞれのタンパク質に対して、75個のネットワーク統計量が計算される。(ii)では、計算した75個の特徴量から、Relief アルゴリズムを使用して、予測に有用な特徴量を選択する。この解析の結果、25個の特徴量を選択した。(iii)では、選択した25個の特徴量をインプットデータとして、薬剤標的分子を予測するための classifier モデルを構築する。このモデルは、それぞれのタンパク質を、2つのクラス、“薬剤標的候補分子”と、“非薬剤標的分子”のどちらに属するかについて予測を行う。classifier モデルの構築には、ランダムフォレストを用いた。10 fold-cross validation を行い、classifier モデルの予測精度の検証を行った。その結果、構築した classifier モデルは、高い予測精度を示した (Area under the ROC curve : 0.948-0.990)。

構築した classifier モデルを、ネットワーク上にある全てのタンパク質に対して適用し、幾つかの種類のがん、アルツハイマー病、そしてリウマチなどに対して、“薬剤標的分子候補”となりうるタンパク質をリストアップした。アルツハイマー病に対する新規薬剤ターゲット候補として推測できたもののうち、特に、SOCS1 は JAK/STAT シグナリングの抑制を通じてサイトカイン反応を調整し、中枢神経系の炎症を制御することが知られている(引用文献)。したがって、SOCS1 は、アルツハイマーの病態をコントロールする鍵となる因子である可能性がある。

ところで、あるタンパク質 A が、病気 1 の薬剤標的分子候補であると予測された場合を考える。この時、タンパク質 A が、ほかの病気 2 に対する薬剤 R の標的分子であるとする。この場合、薬剤 R はタンパク質 A を通じて、病気 1 の疾患メカニズムを制御できる可能性があり、病気 1 に対して repositionable な薬剤候補であると推測される。このように考え、上記の classifier モデルでリストアップした薬剤標的分子候補を、薬剤ターゲットデータベース上にマップし、repositionable な薬剤を予測した。この方法により、特に、白血病の治療薬である Nilotinib は、アルツハイマー病に対して repositionable であると推測された。Nilotinib とアルツハイマー病の関連性については、Lonskaya らが、Nilotinib と bosutinib のコンビネーションが、Ab1 の抑制を通じて中枢神経系の免疫システムを調整し、アミロイドを取り除き、神経の炎症を低減させることを、報告している(引用文献)。我々の予測結果とこの報告から、Nilotinib は、初期のアルツハイマー病に対して、repositionable な薬剤候補である可能性が示唆される。

以上の結果から、我々が開発した計算手法は、薬剤標的分子候補および repositionable な薬剤を絞り込み、効率よく創薬を進めるための有用なツールとなりうると期待される。この研究成果について、国際学会 1st International Symposium on BioComplexity で口頭発表を行った(学会発表)。

(3) 最近、ネットワークのダイナミクスを制御するために重要な役割を果たすノード(driver)を同定する解析手法である、controllability 解析、が注目を集めつつある。しかしながら、タンパク質間相互作用ネットワークにおける、driver と、薬剤標的分子との関係性はよくわかっていなかった。そこで、我々は、人のタンパク質間相互作用ネットワークに対して controllability 解析を行い、driver と薬剤標的分子との関連性を調査した。その結果、driver の中でも、相互作用の多い driver 群は、有意にがんの薬剤標的分子とオーバーラップしていることが判った。この研究成果について、国際学会、20th International conference of Artificial Life and Robotics において、口頭発表を行った(学会発表)。

(4) メチシリン耐性ブドウ球菌(MRSA)は、多くの抗生物質に対する耐性を獲得した、多剤耐性菌である。MRSA による感染症を、効果的に治療する新薬の開発のために、新しい薬剤標的分子の開発が望まれている。

最近、網羅的な、MRSA のタンパク質間相互作用ネットワークの情報が利用可能となり、この情報は、MRSA に対する薬剤標的分子の探索に利用されつつある。また、人の遺伝子と相同性のない MRSA のタンパク質は、薬剤標的分子として使用した場合に、副作用を及ぼす危険性が低いと考えられる。

そこで、我々は、MRSA のタンパク質間相互作用ネットワークの情報、MRSA の遺伝子と人の遺伝子の相同性の情報、および、MRSA のタンパク質の特徴量を入力データとした、薬剤標的分子を推定する予測モデルの開発を行った。この予測モデルは、代表的な 10 種の機械学習アルゴリズム(サポートベクターマシン、ランダムフォレスト、一般化線形モデル、partial least square、Naïve Bayes、regression tree、neural network、multi-layer perceptron、self-organizing maps、k-nearest neighbors)の予測結果を組み合わせて、最終的な予測結果を生成する。この予測モデルにより、3つの新規の薬剤標的分子候補(YP_041656、YP_040643、YP_040661)を推定した。この研究成果について、国際シンポジウム、8th Asian Young Researchers Conference on Computational and Omics Biology(学会発表)にて発表を行った。

現在、この研究結果について、論文

を準備中である。

(5) 原核生物には寄生性の種が多く、多様な宿主の環境に対応するために、タンパク質間相互作用ネットワークの構造を変化させるため、ネットワーク構造に多様性があると考えられる。しかし、モデル生物である酵母以外の生物種、特に原核生物については、ネットワークの構造は詳しく調べられていない。

そこで、我々は、特に原核生物に着目し、多様な生物種のタンパク質間相互作用ネットワークについて詳細な比較解析を行い、生物種間での構造の差異について調べた。その結果、特に、真核生物では、相互作用数の多い蛋白質と少ない蛋白質間の相互作用(非同類結合)が多数存在するという性質を持つものが多かった。一方、原核生物では、非同類結合性を持つものの割合は、真核生物よりも少なかった。

さらに、代表的な原核生物6種のタンパク質間相互作用ネットワーク構造の比較と、遺伝子重複に基づく進化モデルを用いたシミュレーション解析により、相互作用数の少ない(多い)遺伝子が重複されればされるほど、PINの非同類結合性の程度は強く(弱く)なることを示した。また、原核生物の33種について比較ゲノム解析を行い、このシミュレーションの結果が妥当なものであることを確認した。これらの結果は、以前の我々の真核生物に対する研究結果と一致する(引用文献)。このことから、原核生物においても、真核生物と同様に、ネットワークの構造の差異は、遺伝子重複のし易さの相互作用数に対する依存性の違いにより生じていることを示唆する。

この研究結果について、現在、論文を準備中である。

(6) ハイスループットな実験により得られたタンパク質間相互作用データには、多数の偽陽性が含まれる。この問題を解決するために、科学文献より抽出した、タンパク質間相互作用に関する情報が利用されつつある。例えば、研究者が科学文献を読んで情報を抽出することにより、パーキンソン病などに関連する、生体内分子間の相互作用のデータセットが構築されている。このようなデータセットは、それぞれの相互作用が、小規模な実験から得られたもののため、信頼性が高いと考えられている。しかし、このようなデータベースを構築するには、研究者が大量の文献を読む必要があり、多大な時間とコストが必要である(例えば、パーキンソン病のデータベースを構築するためには、研究者が429本の論文を詳細に読んで、相互作用に関する情報を抽出している(引用文献))。さらに、出版された文献の数は膨大であり(例えば、パーキンソン病の文献数は80,000を超える)、研究者がこれらの全ての文献を読んで、網羅

的なタンパク質間相互作用のデータベースを構築することは、非常に困難である。

計算機によるテキストマイニングは、上記の問題を解決できる可能性がある。現在、急速な機械学習技術の進歩により、テキストマイニングのための有用な計算手法が利用可能である。これらのツールの中で、Word2Vec 法は、最も有望である(引用文献)。Word2Vec 法は、巨大なコーパスを解析し、コーパス内の各単語に対して、ベクトルを生成する(一般には、ベクトル空間の次元を 100~1,000 次元として設定する)。その後、コーパス内において同様の文脈で使用される 2 つの単語が、ベクトル空間においてお互いに近い場所に位置するように、ベクトル空間を最適化する。例えば、V[“ミルク”](V[“ミルク”]は“ミルク”のベクトルを表す)は、料理に関するコーパスから生成されたベクトル空間において、V[“魚”]よりも V[“チーズ”]に類似している可能性がある。

この Word2Vec 法は、大量の科学文献をもとに生成されたコーパスから、タンパク質間相互作用ネットワークに関する情報を抽出するための、有用なツールとなり得る。例えば、あるタンパク質 A のベクトルが他のタンパク質 B のものと類似している場合、2 つのタンパク質 A と B は、科学文献において同様の文脈で使用されている可能性が高い。そのような 2 つのタンパク質は、互いに相互作用し、共通の疾患メカニズムや、同じシグナル伝達経路と関連している可能性がある。そこで、我々は、2 つのタンパク質のベクトルが類似している場合、これらの 2 つのタンパク質は相互作用すると仮定し、word2vec をベースとした、特定の疾患(例えば、統合失調症、パーキンソン病およびアルツハイマー病)に関連するタンパク質間相互作用の情報を抽出する、計算手法を構築した。この計算手法は、以下の 4 ステップで構成される。

ステップ(i): NCBI の pubmed データベースから、対象とする疾患と関連する論文の抽象ストラクトをダウンロードする。

ステップ(ii): ダウンロードした抽象ストラクトをフォーマットし、コーパスを生成する。

ステップ(iii): word2vec アルゴリズムをコーパスに適用して、コーパス内の各単語(この場合は単語はタンパク質)のベクトルを計算する。

ステップ(iv): 2 つのタンパク質のベクトルの類似性を計算し、類似性が高いタンパク質のペアをリストアップする。これらのペアにおいて、一方のタンパク質が他方のタンパク質と相互作用すると仮定し、タンパク質間相互作用データベースを構築する。

この計算手法を使用することにより、パーキンソン病、および、アルツハイマー病に関連する、タンパク質間相互作用のデータベースを作成した。このデータベースを、代表的なパブリックドメインのデータベースと比較した。その結果、データベース間において、多数の相互作用がオーバーラップしていることが判った。このことは、我々の手法により抽出されたタンパク質間相互作用の情報は、信頼性が高いことを示唆している。この研究成果は、国際学会、2nd International Symposium on BioComplexity において口頭発表を行った(学会発表)。この研究成果について、現在論文を準備中である。

<引用文献>

- Hase T et al. PLoS Comput Biol 2009 5(10):e1000550
Hase T et al: PLoS ONE 2008 3(2):e1667
Hase T, Niimura Y, Tanaka H BMC Evol Biol 2010 10:358.
Baker BJ, Akhtar LN, Benveniste EN Trends Immunol 2009 30:392-400.
Lonskaya I et al. Neuroscience 2015 304:316-327
Fujita KA et al. Molecular Neurobiolog 2014 49:88-102
Mikolov et al. arXiv 2013 arXiv:1301.378

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1件)

Hase T, Kikuchi K, Ghosh S, Kitano H, and Tanaka H. Identification of drug-target modules in the human protein-protein interaction network. Artificial Life and Robotics, 2014, 19: 406-413. doi:10.1007/s10015-014-0178-5 [査読有]

[学会発表](計8件)

Takeshi Hase, Samik Ghosh, Ayako Yachie, and Hiroaki Kitano. A neural network based text mining approach for inference of protein-protein interaction networks. 2nd International Symposium on BioComplexity 2017 年 1 月 20 日, B-Con PLAZA (別府市・大分県)
長谷 武志. 生体分子間相互作用ネットワークの情報をを用いたドラッグリポジショニングの可能性について. Chem-Bio Informatics Society(CBI) Annual Meeting 2016, 2016 年 10 月 26 日, タワーホール船堀(江戸川区・東京都)

Takeshi Hase, Kaito Kikuchi, Samik Ghosh, Hiroaki Kitano, and Hiroshi Tanaka. A machine-learning approach that leverages information from genome-wide molecular interaction networks towards identification of novel drug-target. International Conference on Systems Biology, 2016年9月17日, Barcelona (Spain)

長谷 武志, 新村 芳人, 多様な生物種におけるタンパク質間相互作用ネットワーク構造の差異とその進化. 日本進化学会第18回大会, 2016年8月25日, 東京工業大学大岡山キャンパス(目黒区・東京都)

Takeshi Hase, Kaito Kikuchi, Samik Ghosh, Hiroshi Tanaka, and Hiroaki Kitano. A computational approach to prioritize drug-target genes in the human protein interaction network. 1st International Symposium on BioComplexity, 2016年1月21日, B-Con PLAZA(別府市・大分県)(in The Proceedings of The 21st International Symposium on Artificial Life and Robotics 2016, pp.871-876, https://isarob.org/symposium/index.php?main_page=index)

Takeshi Hase, Kaito Kikuchi, Samik Ghosh, Hiroaki Kitano, and Hiroshi Tanaka. Controllability of protein-protein interaction networks and their relationships with drug-targets, essential genes, and degree connectivities. International Symposium on Artificial Life and Robotics AROB 20th 2015, 2015年1月21日, B-Con PLAZA(別府市・大分県)(in The proceedings of The 20th International Symposium on Artificial Life and Robotics 2015, pp.813-818, https://isarob.org/symposium/index.php?main_page=index)

Kaito Kikuchi, **Takeshi Hase**, Samik Ghosh, and Hiroaki Kitano. A network guided approach towards identification of novel drug targets in MRSA. 8th Asian Young Researchers Conference on Computational and Omics Biology, 2015年1月19日, HsinChu (Taiwan)

Takeshi Hase, Kaito Kikuchi, Samik Ghosh, Hiroaki Kitano, and Hiroshi Tanaka. Identification of drug-target modules in the human protein-protein interaction network. The International Symposium on Artificial Life and Robotics 19th 2014, 2014年1月22日, B-Con PLAZA(別府市・大分県)

〔図書〕(計0件)

〔産業財産権〕

出願状況(計0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0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〕

ホームページ等

<http://researchmap.jp/read0099633/?lang=japanese>

6. 研究組織

(1) 研究代表者

長谷 武志 (HASE, Takeshi)

国立研究開発法人理化学研究所・統合生命医科学研究センター・客員研究員

研究者番号: 70569285

(2) 研究分担者

()

研究者番号:

(3) 連携研究者

()

研究者番号:

(4) 研究協力者

()