様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １１ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５７０１７６

研究課題名（和文）遺伝子細胞機能推定のための統合ネットワークの構築と解析

研究課題名（英文）Development and analyses of integrated networks for cellular function predictions

研究代表者
木下　賢吾（Kengo, Kinoshita）

東北大学・情報科学研究科・教授

研究者番号：６０３３２２９３

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：既に数千種を超える生物種でゲノムの全配列が明らかにされているが、ゲノムにコードされている遺伝子の約半数は機能未知のままである。すべての遺伝子の可能な機能を実験で検証するのは到底不可能であり、計算科学的な手法により機能を推定し実験で検証を行う手法が不可欠である。本申請では、マイクロアレイやRNA-seqデータから得られる共発現情報と大規模なタンパク質間相互作用データを利用して、統合ネットワークの構築と種間比較を行い、新規のクラスタリング手法の開発と細胞機能の推定法の開発を行った。

研究成果の概要（英文）：We have more than several thousands of whole genome sequences of many divergent species, but more than half of the genes on the genomes are still uncharacterized. It is almost impossible to determine all functions of all genes by experimental procedure, and thus some computational approaches to infer some possible functions of the uncharacterized genes. In this study, we have constructed integrated networks by using RNA-seq based co-expression networks and high throughput protein-protein interaction data, and developed a method to compare the integrated networks among the species. We also developed a new clustering method and applied it to infer the cellular function of uncharacterized genes.

研究分野：生命情報科学

キーワード：遺伝子共発現　蛋白質間相互作用　統合ネットワーク　種間比較　クラスター解析

１．研究開始当初の背景

ポストゲノム時代の遺伝子機能の推定は急務であり、既に多くの手法が試みられてきた。その中で最も信頼性の高い方法は配列の類似性検索による進化的類縁関係の利用であるが、似た配列を持つ遺伝子が無い場合や、有ったとしても類似性の低い場合は機能の推定がうまく行かないことが多い。これは、機能の類似性と進化的な類縁関係は間接的にしか相関していないことからくる原理的な限界である。実際、配列の類似性検索法の感度は、バイオインフォマティクス的に洗練された手法により飛躍的に向上したが、遺伝子の機能推定が大きく進歩したわけではない。これに対して、配列→構造→機能のパラダイムに従って、遺伝子産物であるタンパク質の立体構造情報を利用して機能を推定する手法が有効に働くことがある。立体構造情報は配列の類似性と異なり、様々な見方が出来るので扱いが難しいが、原子の空間配置のレベルで類似性が見られる場合には、機能推定の有効な手段となることも多い。しかしながら、立体構造情報、特に原子の空間配置のように化学的な機能と強く相関する情報を使った場合には、１分子で決まる分子機能が主な対象となり、細胞機能の推定は困難である。

一方、細胞機能はそもそも複数のタンパク質が相互作用することで実現される機能であるから、酵母ツーハイブリッド法などを利用して蛋白質の相互作用ネットワークを構築し、解析を行うというアプローチがある。このアプローチは一定の成果を上げつつあるが、解析するネットワークの信頼性の問題（偽陽性・偽陰性が多い）、細胞内局在や各タンパク質が発現するタイミングの情報に欠けているため、まだまだ実用にはほど遠いのが現状である。これに対して、近年 DNA マイクロアレイを利用して遺伝子の発現するタイミングの情報として、発現量情報が蓄積されてきた。DNA マイクロアレイのデータは初期段階では信頼性に問題があったが、近年 MAQC（MicroArray Quality Control project）といった大規模な精度検証実験によりその精度が飛躍的に向上している事が実証された。これらのデータを併せて利用すれば、タンパク質間相互作用ネットワークに欠けていた、発現する組織や発現のタイミングといった時間と空間の情報（時空間情報）を含んだ相互作用ネットワークを構築し、そのネットワークを解析から遺伝子のネットワーク上での役割、すなわち細胞機能の推定が可能になると期待される。

２．研究の目的

既に数千種を超える生物種でゲノムの全配列が明らかにされているが、ゲノムにコードされている遺伝子の約半数は機能未知のままである。すべての遺伝子の可能な機能を実験で検証するのは到底不可能であり、計算科学的な手法により機能を推定し実験で検証を行う手法が不可欠である。タンパク質の機能は、タンパク質一分子で決まる生化学的な機能である「分子機能」と、複数のタンパク質間相互作用で決まる生物学的な機能である「細胞機能」に分けることができるが、これまで計算手法による機能推定は主に分子機能を中心として行われてきた。これに対して本申請では、タンパク質の配列情報・立体構造情報・タンパク質間相互作用の実験データに加え、マイクロアレイから得られる共発現情報を利用して、統合ネットワークの構築と解析を行い、細胞機能の推定法の開発を行う。

３．研究の方法

DNA マイクロアレイのデータから共発現を計算して、PPI のデータと統合し、解析を行う。その際、データが多いことが重要であり、データの追加をまず行う。特に近年増えてきている RNA-seq のデータの拡充を行い利用する。解析としては種間の比較を行い、保存している機能モジュールの同定を行い機能推定につなげる。大規模 RNA-seq の課題と解決としては、１万サンプルに及ぶ RNA-seq のデータを解析するためには、通常よく使われる解析手法をそのまま適用することは難しい。そこで、出来るだけ精度を落とさず共発現情報のみを得られるように解析手法の検討を行った。具体的には、通常ゲノム配列にマッピングすることが多いショートリード情報を、転写産物にマッピングし、スプライスバリアントの解析は行わないことで計算時間の軽減を行うと共に、データ取得からマッピングまでをパイプライン化することで計算時間を可能な限り削減することができた。

最終的にできあがるネットワークは大規模なネットワークとなる。そこで、解析として、大規模なネットワークから意味のあるクラスターを抽出する手法を開発する。この方法では、従来は、完全に結合している部分グラフとして関連性の高い部分グラフを効率よく抽出する手法はあったが、完全に結合している部分グラフという、数学的には妥当な条件が生物学的には適用が困難であったことへの対応となっており、非常に一般性の高いネットワーククラスタリング手法となっている。具体的には、PageRank と呼ばれる、インターネットの情報を解析する際に用いられる重要度指標を利用し、重要度が高い遺伝子から順々に密に相互作用している部分ネットワークの抽出を行い、それら部分ネットワークを組み合わせて、より大きく関連性の高い部分ネットワークとして抽出を行った。抽出されたクラスターに関しては、含まれる遺伝子群の注釈が統計的に有意に濃縮されているか等の検討を行い、有意に多くの注釈がついているクラスターを生物学的に意味があるクラスターとしてリストアップし、個別に解析を進めることとした。

４．研究成果

本研究計画では、大きく分けて共発現データ

の解析の部分とタンパク質間相互作用ネットワークの解析を行った。
共発現データの解析に関しては、空間情報として、組織の違いに着目して解析を行った。まず、ことなる条件下での共発現パターンがどれぐらい似ているかを定量化するために、共発現パターンの比較法の開発を行った。次に、異なる種や組織での発現パターンを比較することで、保存している遺伝子群を見いだす手法へと発展させた。手法に関しては、生命医科学情報学連合大会で口頭発表に選ばれ、ポスター賞を受賞するなど、手法の独自性が高く評価された。

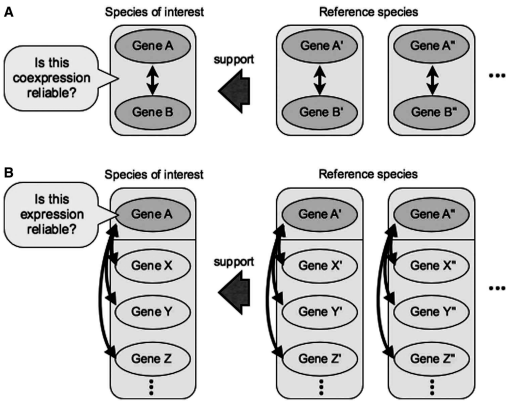

図１：共発現の種間比較の概念図

また、ヒトをターゲットとして、これまで使っていたのとは別のプラットフォーム（Affymetrix Human Gene 1.0 ST Array）での共発現データ（6865 サンプル）を ArrayExpress から取得し、mas5, RMA での規格化等の比較を行い、精度の見積を行った。また、従来はマイクロアレイデータを利用した共発現解析を行ってきたが、近年、RNA-seq による発現量情報が増えてきたことを受けて、RNA-seq での共発現の利用を開始した。具体的には、ヒト（11816 サンプル）、マウス（9518 サンプル）、ショウジョウバエ（1920 サンプル）の RNA-seq のデータを利用して共発現ネットワークの構築を行った。大規模な RNA-seq データを多サンプルに関して処理するために、発現量の推定に関して新たな手法の開発も行った。サンプル収集過程の途中経過であるが、ヒトサンプル（5626 サンプル）を利用して、アレイ（73,083 サンプル）の時と同様に機能予測を指標として性能を評価したところ、ヒトに関してはマイクロアレイで 2.60 であった partial AUC(/10,000)が、RNA-seq では 2.84 に向上するなど、大幅な性能向上をすることができた。結果は COXPRESdb の一部として公開を行った。また、共発現データの解析として、共発現ネットワークの種間解析を行った。その結果、その結果、種を超えて特徴的に保存している遺伝子群として 2717 遺伝子群を同定することができた。これらをさらに詳細にクラスター解析を行ったところ、タンパク質合成系や細胞周期に関わる基礎的な遺伝子群が予想通りあきらかになったのと同時に、予想とは異なり、免疫系などで共通の共発現パターンを有することを見いだした。また、その他のクラスターに関しても系統的な比較を行うことで、従来手法よりも種間で保存していることを加味したクラスターで有意に多くの GO-term が濃縮することもわかり、今回の計画で開発を行った手法の有効性を明らかにすることができた。以上の結果をまとめた論文の投稿を 3 月に行い、現在改訂中である。

以上の解析と平行して、共発現を生み出すメカニズムの解明を目指して、ChIP-seq データの解析も行った。特に、従来、定性的な解釈が主であった ChIP-seq の実験データに対して、酵母由来の DNA を内部標準として利用して、定量的な解析ができる手法の開発も行った（特許出願中）。

タンパク質間相互作用ネットワークの解析に関しては、タンパク質間相互作用ネットワークから機能モジュールを探すために、新規にクラスタリング手法の開発をおこなった。具体的には、これまではクリークと呼ばれる全てのノードが互いにエッジを持っている部分グラフを効率的に探す手法はあったが、クリークという条件は生物学的には厳しすぎることが問題であった。そこで我々は、PageRank と呼ばれるノードの重み付けに着目して、小さな部分グラフから大きな部分グラフを構成するアプローチで、クリークに近い大きな部分グラフを効率的に検出する手法の開発を行った。結果として、従来手法よりもより機能的に関連性の深い卵白室群を見いだすことができた。この手法をネットワーク解析の標準的なプログラムである Cytoscape で利用できるような実装（図２）を行うことができた。結果と手法を報告する論文を現在投稿準備中である。

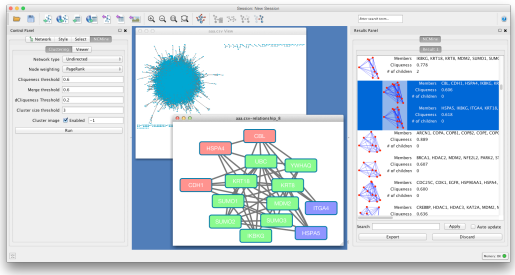
図２：Cytoscape での実施例

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５件）
1. Okamura Y, Aoki Y, Obayashi T, Tadaka S, Ito S1, Narise T, Kinoshita K, COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems, Nuc.

leic Acids Res, 43, D1, D82-D86, 2014.(査読有り)
2. Lensink M.F. et. Al (Kinoshita K, 57名 30 番目), Blind prediction of interfacial water positions in CAPRI. Proteins, 82, 620-632, 2014(査読有り)
3. Obayashi T, Okamura Y, Ito S, Tadaka S, Aoki Y, Shirota M, Kinoshita K. ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. 55, e6, Plant Cell Physiology, 2014(査読有り)
4. Shirota M and Kinoshita K. Analyses of the general rule on residue pair frequencies in local amino acid sequences of soluble ordered proteins, Protein Science, 22, 725-733, 2013(査読有り)
5. Obayashi T, Okamura Y, Ito S, Tadaka S, Motoike IN and Kinoshita K. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. Nucleic Acids Res, 41, D1014-20, 2013(査読有り)

〔学会発表〕(計 9件)

1. Tadaka S, Obayashi T, Kinoshita K, Detection of functional modules in protein networks by near-clique extraction, GIW Dec 15-18, 2014,Odaiba, Tokyo
2. Kinoshita K, Prediction of the biological and biochemical functions of uncharacterized genes, 2014 Bilateral Workshop between Tohoku University & National Tsing Hua University, 2014年11月21日、松島(宮城)
3. 岡村容伸, 大林武, 木下賢吾.「DEG.js : Streaming web-based differential expression genes analysis tool for RNA-seq」、生命情報科学若手の会 第6回研究会、2014年10月30日、理化学研究所 発生・再生科学総合研究センター(兵庫)
4. Okamura Y, Obayashi T, Kinoshita K, Classify RNA-seq runs as origin organs or other features by using machine learning, 22st Annual International Conference on Intelligent Systems for Molecular Biology, July 13-15, 2014, Boston (USA)
5. Tadaka S, Obayashi T, Kinoshita K, NCMine: a novel method for exploring clusters in biological networks″第36回日本分子生物学会年会, 2013年12月3日、神戸国際会議場(兵庫)
6. Okamura Y, Obayashi T, Kinoshita K, GO analysis of gene expression patterns comparison among organs and species, 生命医薬情報学連合大会,2013年10月29日、タワーホール船堀(東京)
7. Okamura Y, Obayashi T, Kinoshita K, Functional gene network prediction based on conservation of gene expression patterns, 21st Annual International Conference on Intelligent Systems for Molecular Biology, 2013年7月23日、Berlin, Germany
8. Murakami Y, Kanamori E, Sarmiento J, Liang S, Standley D.M, Shirota M, Kinoshita K, Tsuchiya Y, Nakamura H, An Automatic and Semi-automatic Approach for Predicting Protein-Protein Complex Structures, CAPRI Meeting 2013, 2013年4月17日, Utrecht, Portland
9. Tadaka S, Obayashi T, Kinoshita K. Identification of functional modules in protein network by near-clique detection, 情報処理学会第 33 回バイオ情報学研究会, 2013年3月21日、東北大学(宮城)

〔産業財産権〕

○出願状況(計 1件)

名称:免疫沈降用の内部標準分子および免疫沈降方法
発明者: 五十嵐和彦、落合恭子、中山啓子、木下賢吾、舟山亮、細金正樹、岡村容伸
権利者:国立大学法人東北大学
出願人: 国立大学法人東北大学
出願番号:特願2015-035734
出願年月日:平成27(2015)年2月25日
国内外の別:国際

〔その他〕

ホームページ等
http://coxpresdb.jp
http://atted.jp

６. 研究組織

(1)研究代表者
木下 賢吾(KINOSHITA, KENGO)
東北大学・大学院情報科学研究科・教授
研究者番号:60332293

(2)研究分担者
無し

(3)連携研究者
無し