様式 C-19

科学研究費補助金研究成果報告書

平成　21 年　6 月　8 日現在

研究種目：　基盤研究（B）
研究期間：　2005 年度　～　2008 年度
課題番号：　１７３１０１１６
研究課題名（和文）　タンパク質コード領域に特徴的な塩基配列パターンの網羅的な解析
研究課題名（英文）　Comprehensive analysis of characteristic sequence patterns among protein-coding regions

研究代表者
中村　保一　（NAKAMURA YASUKAZU）
財団法人かずさＤＮＡ研究所・植物ゲノム研究部・植物ゲノム情報研究室・特別客員研究員
研究者番号：　６０３７０９２０

研究成果の概要：

DNA 塩基配列データベースに登録された生物種のタンパク質遺伝子の配列を網羅的かつなるべく重複を省き集計するシステムを作成し運用した。このシステムにより集計した塩基配列の特徴を示す短い配列パターンを抽出し､配列の類似比較からだけでは見いだすことのできない、完全に未知なタンパク質遺伝子を発見するための基盤となる情報を集積すると同時に、タンパク質遺伝子発見に応用するための方法を検討した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2005 年度	2,900,000	0	2,900,000
2006 年度	500,000	0	500,000
2007 年度	500,000	150,000	650,000
2008 年度	500,000	150,000	650,000
総　計	4,400,000	300,000	4,700,000

研究分野：
科研費の分科・細目：
キーワード：　(1)　タンパク質遺伝子コード領域 (2)　遺伝子発見　(3)　　配列パターン解析
(4)　コドン使用傾向　(5)　ゲノム塩基配列解析

１．研究開始当初の背景

ほとんどのゲノムに、他生物種のゲノムに類似な遺伝子が存在しない、いわゆる「みなしご遺伝子」が存在する。研究開始当初、すでに数百種の生物種の全ゲノム塩基配列が決定されて来ているなかでも、新規に解読された塩基配列上の遺伝子領域の予測は、BLAST などによる、既知遺伝子をライブラリとした配列類似探索だけで発見できるものではなく、遺伝子の 1/4 から半数以上が、*ab initio* 遺伝子予測法を併用することによって予測する必要があった。*ab initio* 法による遺伝子コード領域の予測は、その生物種の既知のタンパク質遺伝子配列に特徴的に出現する数塩基程度の塩基配列パターンの現れ方をモデル化しておき、未知の塩基配列に沿ってモデルがよくあてはまる、高い「コード領域らしさ（コーディング・ポテンシャル)」が連続する領域をタンパク質遺伝子として検出するものである。

しかし、この「コード領域らしさ」の指標となる配列パターンは、生物種ごとに異なり、たとえ近縁の生物種であっても、ゲノム全体

を支配する GC 含量の異なりなどによりまったく違うパターンを示すことも少なくなく、そのため、ヒトや主要なモデル生物を対象とした遺伝子発見法は多数存在し利用可能である一方、それ以外の数多くの生物種からの遺伝子発見は配列パターンの収集や解析が行われておらず、遺伝子の発見に困難があった。一本以上の完全長タンパク質遺伝子注釈が国際 DNA データベースに記載されている生物種数は､本研究の計画時には 23,000 種（オルガネラゲノム、亜種ゲノムはそれぞれ別に集計）に達していたが、ほとんどの生物種で遺伝子領域推定のための効率的な配列パターンの収集と解析がなされていないという状況があった。

研究開始時点での類似の研究として、ゲノム上の数塩基までの配列パターンの出現頻度を自己組織化マッピング(SOM)により分類する解析が阿部らによりすすめられていたが、(Abe, T. *et al.*, 2003; *Genome Res.*, **13,** 693-702) が、全ゲノム塩基配列を対象とした解析が中心であるため、遺伝子発見への利用は未知数であった。本研究ではこうした自動的なゲノム全体に見いだされる固有のパターン探索ではなく、生物種ごとあるいは近縁生物分類群ごとにあらかじめ集計したコード領域から、積極的に共通なパターンを抽出し応用可能なモデルを構築することの可能性を考慮した。

２．研究の目的

本研究の目的は、DNA データバンクに塩基配列が存在する生物種について、網羅的にコード領域を収集し、生物種あるいは生物分類群ごとにコード領域に特徴づける数塩基までの配列パターンの効率的なモデル情報を集積することであった。

塩基配列決定技術の高速化・大容量化が加速度的に進むなかで、ますます推進されていくであろう多種類の生物種での概要（ドラフト）塩基配列決定や転写配列蓄積研究での配列解析に求められるのは、完全な遺伝子モデル構築技術とあわせ、短く不完全な断片配列からの効率的な新規コード領域の推定も重要になると考えられた。従って、本研究による生物種網羅的なコード領域とその解析結果の集積は、幅広い生物種由来の塩基配列からの新規な有用遺伝子発見に役立つものと考えられた。

本研究で得られた解析情報を利用することで、遺伝子コード領域推定、すなわち、ゲノムや転写産物由来の塩基配列から、タンパク質遺伝子である可能性の高い領域とフレームを推測することが幅広い生物種で可能になることが期待される。また、メタゲノム研究と呼ばれる、特定の環境由来の生物群からの直接塩基配列決定のように、生物種が混合された状態で、個々の塩基配列の由来生物種が未知の配列データからのコード領域推定による新規遺伝子の探索と同時に、コード領域の特定の生物種群への分類を行うための解析の根拠情報としても利用可能と考えられた。

３．研究の方法

本研究は、幅広い生物種におけるコード領域を特徴づける情報の収集と解析を目標とし、以下の方法に沿って研究を進めた。

まず、既知の塩基配列からの生物種ごとの網羅的なコード領域の集積を試みた。本研究の基盤は、コード領域を取得することが可能な全生物種から、そのコード領域を集積し、個々のゲノムを特徴づける配列パターンを収集することにある。その際、まずコード領域である根拠があり、かつ配列パターンの重複度の低い塩基配列を選抜する系の確立を行う必要があった。

既存の *ab initio* 遺伝子予測法の実行結果のばらつきは、方法論の問題と、既知の遺伝子を収集した学習セットの選択による問題の両方を含むと考えられる。生物種ごとの塩基配列のパターンの特徴を集計する際、同じ傾向の配列が重複することにより特定のパターンを過度に学習してしまうことが大きな問題となる。これを防ぐことを目的として、可能な限り重複を除去した遺伝子コード領域のデータセットを作成する方法論を検討し、実現可能なパイプラインの構築を行った。研究期間中、運用を継続しながら、タンパク質遺伝子コード領域を網羅的にかつ可能な限り重複の少ないデータセットとして収集するための、現実的に可能な方法の模索と改善を進めた。

次に、コード領域を特徴づける数塩基までの配列パターンの収集・モデル化を試みた。研究代表者が関連している研究プロジェクトにより、高精度な塩基配列が得られている、原核生物である複数のラン藻（シアノバクテリア）ならびに根粒菌（リゾビア）ゲノムと、高等真核生物の代表として植物のシロイヌナズナ、ミヤコグサの解析データに基づいた予測遺伝子セットについての検討を行い、配列特徴抽出法について考察した。配列データセットからの集計プログラムを用い、生物種ごとにコード領域に観察される単位を中心とした特徴的な配列パターンの出現頻度を集計しデータセット化した。

このデータセットを用いて生物種ごとのモデル作成と解析環境の作成を試みた。ここまでの研究で得られている、生物種ごとに遺伝子コード領域を特徴づけたデータセットとそこから構築したモデルを用い、未知の塩

基配列からのコード領域推定を行う判別プログラムの応用・公開を目標とした開発の検討を実施した。

４．研究成果

本研究計画実行の基盤となるのは、信頼性の高い配列情報を収集したタンパク質遺伝子をコードするデータセットの構築に他ならない。既知の塩基配列からの生物種ごとの網羅的なコード領域の集積を行うための方法論として、二つの方法の可能性について検討を行った。1) 国際 DNA データバンクからの網羅的な収集 2) ゲノムプロジェクト等の全ゲノム塩基配列報告生物種による提供データからの収集である。1) については、DDBJ/EMBL/GenBank の三極から構成される国際 DNA データベースに登録された塩基配列のエントリに、CDS フィーチャとして登録者によって構造注釈されたタンパク質コード領域が記載されている場合がある。本研究申請者が 1994 年より作成・公開している遺伝暗号（コドン）使用データベースでは、国際 DNA 配列データバンクのうち GenBank を用い、生物種により分類された division （例 hum: ヒト、pri: ヒト以外の霊長類、mam: 霊長類以外の哺乳類、などに分類される）フラットファイルから、注釈の存在するすべての完全長タンパク質遺伝子を切り出し、さらに生物種ごとに取りまとめた遺伝子ライブラリを作成し、それを用いてコドン使用頻度を集計し公開している（Nakamura, Y. *et al.*, 2000; *Nucleic Acids Res.* **28,** 292)。この集計では、配列登録者によって記載された遺伝子構造アノテーション情報を信頼して配列を収集し、部分配列と偽遺伝子配列を除いたすべての配列ごとに遺伝暗号頻度表を算出してきている。しかしながら、研究計画の段階でも予測していたが、国際 DNA データベースに登録された塩基配列注釈情報に基づき遺伝子を集計した場合、同一遺伝子塩基配列の重複が不可避的に集計されてしまう。完全な重複であればチェックサムを利用するなどの単純な情報処理によって重複を除去することは比較的容易であるが、不完全な類似ではそうした処理が困難である。塩基配列自体の重複登録も多く、同一と思われる配列領域に異なる複数の遺伝子構造が予測されている場合も多々あり、こうした重複を自動的に除去することは注釈の照合からだけでは困難であった。

研究計画当初に予定した方法は、まず信頼の置ける既知タンパク質遺伝子ライブラリに対する網羅的な類似配列検索を実行し、遺伝子コード領域と思われる領域を既知の塩基配列から可能な限り抽出し、特徴的な配列パターンの重複取得を避けるため、配列の類似によるクラスタ化を行い、クラスタごとの代表配列を選抜することで、生物種ごとに特異性の高い配列群からなるデータセットを作成する方法であった。同時に EST, HTG など配列注釈が存在しない登録ファイルからの遺伝子領域の抽出も計画した。

しかし、塩基配列情報の爆発的な増大に伴い、当初予定していたこの方法は高速な類似配列検索を行うための PC クラスタによっても実際的な時間内には実行することが不可能であることが判明し、現実的に可能な方法の検討をおこなった。国際 DNA データバンクの配列重複を回避し、ゲノムレベルでの網羅的かつ重複の少ない配列の提供を目指したデータコレクションとして網羅的なものに、米国 NCBI の The Reference Sequence (RefSeq) が存在する。ゲノム計画が実施された生物種を含む代表的な生物種については、この RefSeq データセットからのタンパク質遺伝子情報の抽出が現実的であり、RefSeq に登録されていない生物種については登録数の少なさから、重複が問題になることが皆無ではないが比較的少ないため、本研究のデータ収集の際に RefSeq に収集されているか否かを最初の分岐とし、RefSeq に登録されている生物種のデータセット(H21 年 5 月現在、約 8,400 種)についてはその配列を利用し、それ以外の生物種(約 3 万種)については国際 DNA データバンクの注釈情報からタンパク質遺伝子塩基配列を抽出するパイプラインを構築した。以上の処理を自動的に実行することで生物種網羅的な、実行可能なレベルで最大限クリーンな塩基配列データセットの選抜と集積を可能とし、これを継続して進めた。注意すべき点は、RefSeq に登録されている場合でも、NCBI の人的資源の限界から重複や注釈確認のキュレーションが充分でない場合が散見されることである。この問題の解消については、情報的なアプローチとして、配列注釈自動化のより良い方法の検討を続けるとともに、キュレーションの見地からは、国際協力体制による配列校正の枠組みを提案していくことが必要と考えられた。

次に、上記の系で得られた基盤配列データを用いて、それらコード領域を特徴づける数塩基までの配列パターンの収集とそのモデル化、解析環境の作成に関わる研究を行った。

まず、配列パターンの予備的な集計として研究代表者が所属する研究グループにより信頼性の高いコード配列を得ている原核生物であるラン藻ならびに根粒菌ゲノムと、高等真核生物のシロイヌナズナ、ミヤコグサのゲノム塩基配列から得られている解析データを用いて、遺伝子発見に応用可能な配列パターンである６塩基単位ならびに３塩基単位でのパターン抽出を行い、特徴分類のための検討データを作成した。これらの予備的研

究結果をふまえ、網羅的に収集したコード領域塩基配列から、生物種あるいは生物種群ごとに特徴的な配列パターンの確率モデルの作成を、既存の遺伝子発見ソフトウェアを用いて逐次試みた。遺伝子発見系については、当初の計画では、判定のブラックボックス化を回避するため、コード領域の判定にはニューラルネットワーク等の方法は応用せず、判別関数や決定木など、生物学者にも判定基準のパラメータやその根拠が明確に理解できることのできる手法を用いることを計画していたが、方法論の検討の結果、データの集計がきわめて大規模になることから、判別方法を完全自動化する必要があることと同時に、解析に供する際の実行速度や汎用性を考慮すると、の問題から、既存の隠れマルコフ連鎖法によるモデル作成を実施し、生物種網羅的にコード領域を推定するための基盤情報の準備を行った。

遺伝子領域のモデル作成と同時に、遺伝子非コード領域の集計と、そのネガティブデータを用いたモデルを作成し、本来の発見法と組み合わせることで遺伝子領域の効率的な発見をめざす方法論の検討を行った。この方法は予備的に高等真核生物のシロイヌナズナ、ミヤコグサのゲノム塩基配列から得られている解析データを用いて実施したが、結果的にはネガティブデータの組み込みでは、遺伝子領域の発見への貢献が得られなかった。このような結果が得られた原因としては、非遺伝子領域の配列パターンが存在しないか、もしくは遺伝子コード領域と同じ明瞭な数塩基単位のパターンを構成していないという理由が考えられた。非コード領域の利用については別の方法論をふまえた今後の展開が必要であると結論づけられた。

作成した収集、解析システムにより得られた基盤情報配列と配列パターン表については現在、遺伝暗号（コドン）使用データベースでのリリース作成に応用しており、同時に元データとしてファイル公開と、解析系の提供のためのウェブサイトを試作中である。

当初の計画では、本研究のより発展的な研究の展開として、一例として、塩基配列決定手法による問題に対する実際的な対応がある。高い完成度を前提としない決定配列、すなわち概要塩基配列決定や転写配列タグなど、低い重複度で解読された塩基配列には、人工的なフレームシフトやギャップが多数存在するが、こうした品質の低い配列にも適用可能な、配列の誤りに耐性のある遺伝子領域発見法の検討と提示があった。具体的には、ギャップを挟んだコード領域の対による遺伝子のスキャフォールド作成や、フレームの異なるコーディングポテンシャルの高い領域の連続を発見し、フレームシフトの可能性を予測する後処理の作成などを行い、配列解析リソースとしての応用をにらんだ展開が考えられたが、本研究終了時では応用的な課題のとりくみと解決には至っていない。情報基盤としてのタンパク質遺伝子配列セットの抽出と集積系は整備したので、今後の発展的な研究の推進を試みたい。また、データの公開と提供システム作成による、バイオインフォマティクス研究者への周知と利用を促進することも、今後の課題である。

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）

〔その他〕
データならびに解析環境公開 URL（試作）:
http://codon.kazusa.or.jp/

６．研究組織
(1)研究代表者
中村　保一（NAKAMURA YASUKAZU）
財団法人かずさＤＮＡ研究所・植物ゲノム研究部・植物ゲノム情報研究室・特別客員研究員
研究者番号：60370920

(2)研究分担者
なし

(3)連携研究者
なし