様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１５　日現在

機関番号：　１３９０４
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２５０

研究課題名（和文）コドン縮約表現に基づくタンパク質－遺伝子モチーフ辞書システムの開発

研究課題名（英文）Development of the protein-gene motif dictionary system based on the codon reduced representation

研究代表者

加藤　博明（KATO, HIROAKI）

豊橋技術科学大学・工学（系）研究科（研究院）・講師

研究者番号：３０３０３７０４

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では、アミノ酸配列モチーフを原点として、その対応する遺伝子配列（コーディング領域）、およびそのタンパク質立体構造までを関連づけたタンパク質-遺伝子モチーフ辞書の構築を試みた。また、アミノ酸配列モチーフに対応するコドン重み行列をもとに、モチーフコドン縮約表現を提案するとともに、ゲノム配列の機能部位推定への応用を試みた。EF-handモチーフを例とした解析実験の結果、TNNC1中のイントロンによって隔てられた２ヶ所の機能部位を正しく検出することができた。

研究成果の概要（英文）：In the present work, the author has developed the protein-gene motif dictionary system which connected from an amino acid sequence motif to the corresponding gene sequence (coding region) and protein 3D structure. I have also proposed the motif codon reduced representation, and applied to the functional site prediction in the genomic sequence. As the result of the EF-hand calcium- binding motif, it is successfully identified two motif sites in TNNC1 which were interrupted by introns.

研究分野：　分子情報工学・分子生命情報学

キーワード：　分子構造情報処理　配列モチーフ　タンパク質　遺伝子　コドン縮約表現　コドン重み行列　三次元モチーフ　データベース

１．研究開始当初の背景

(1) 生物の遺伝情報は DNA の塩基配列として保存され、転写・編集した mRNA を通じて、最終的な発現系となるタンパク質アミノ酸配列へと翻訳される。また、その配列情報に従い固有の立体構造へと折りたたまれることで、生物学的機能を発現する。特にモチーフと呼ばれるタンパク質構造中に特定の配置で存在する局所構造特徴は、遺伝子配列の中でもよく保存されている部分であると考えられる。従って、タンパク質のモチーフ、あるいは広い意味での共通構造特徴はタンパク質の構造-機能解析だけでなく、遺伝情報解析においても極めて重要な問題の一つである。

(2) 一方、ポストゲノム計画の進展、並びに実験技術の進歩に伴い、生体高分子のデータは急速に増加しており、その構造データベースは生体高分子の構造と機能との関係解明など分子生物学上の新たな知識獲得のための基本要素としてその重要性はますます高まっている。そのため、これらのデータベースを有効に活用し、配列構造特徴の系統的な解析を行なうための方法論の確立、並びに有効なコンピュータツールの開発が切望されている。

(3) 筆者らはこれまでに、三次元分子構造特徴解析に基づく知識発見の視点から、アミノ酸配列レベルのモチーフデータベース PROSITE に登録されている配列パターンに注目し、これに対応する三次元部分構造情報を網羅的に集積した三次元モチーフ辞書の構築を試みた。また、グラフ論的な部分構造検索技法を基礎とした三次元モチーフ構造検索アルゴリズム、さらには質問構造の設定を要求しない複数タンパク質間の三次元共通構造特徴（新規モチーフ候補部位）の自動認識のためのシステムの開発を進めてきた。

２．研究の目的

(1) 本研究課題では、これらの成果をもとに、遺伝子-タンパク質の翻訳単位であるコドンに注目した、コドン縮約表現に基づく遺伝子領域の推定のための新たな方法を提案する。モチーフは、ゲノム配列中で、一つのエクソン領域に連続して保存されているだけでなく、図１に示すように、二つ以上のエクソン領域にまたがる、すなわち、イントロン領域を挟む状態のモチーフが複数存在すると考え、その情報を手掛かりとしたゲノムの機能部位解析への応用を試みる。

図１　ゲノム配列中で二つのエクソン領域にまたがる形で保存されるモチーフの概念図.

(2) 従来のグローバルな配列全体の比較だけではなく、機能に密接に関連する局所的な構造特徴であるモチーフ部位に注目した分子構造データマイニング手法の確立と、タンパク質-遺伝子の配列およびその対応する立体構造までを関連づけた新規モチーフ辞書（知識ベース）システムの構築を目指す。

３．研究の方法

(1) 遺伝子とタンパク質、すなわち、塩基配列とアミノ酸配列との関係はコドン表により定義されている。コドンとは、連続する三つの塩基（a, g, c, t の 4 種類から構成）のことであり、その組み合わせ（64 種類）から 20 種類のアミノ酸が生合成される。

コドン縮約表現は、ワイルドカード(例えば、r(プリン)、y(ピリミジン)、n(任意の塩基と対応)など、詳細は表１参照)を用いて、コドンとアミノ酸の関係を表現したものである。例えば、六つのコドンが対応するアミノ酸アルギニン(R)は二つの縮約コドン(cgn / agr)で表現される。本研究ではさらに、R を mgn として近似して表現することによって、コドンとアミノ酸との関係を１対１として表した。コドン縮約表現の近似に対応した拡張遺伝暗号表を図２に示す。

表１　塩基配列のワイルドカード表現.

code	nucleotide	code	nucleotide	code	nucleotide
a	a (Adenine)	r	a / g (puRine)	h	a / c / t
g	g (Guanine)	y	c / t (pYrimidine)	b	g / c / t
c	c (Cytosine)	m	a / c (aMino)	v	a / g / c
t	t (Tymine)	k	g / t (Keto)	d	a / g / t
		s	c / g (Strong)	n	a / g / c / t (aNy base)
		w	a / t (Weak)	—	(gap)

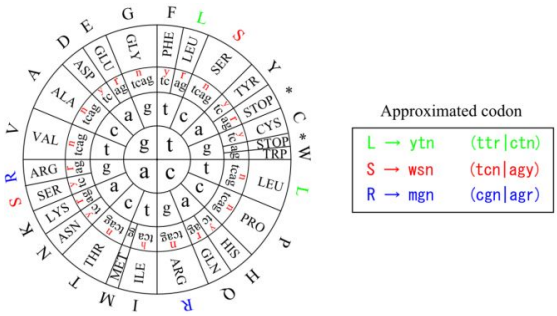


図２　コドン縮約表現と、その近似による拡張遺伝暗号表.

(2) 生成したコドン縮約配列をもとに、塩基配列データベースから対応するタンパク質のコーディング領域（遺伝子領域）の推定を行なうための方法について検討した。イントロン領域を挟むモチーフ部位を推定するために、ダイナミックプログラミング(DP)手法

に基づくローカルアラインメントプログラムを実装した。長大なイントロンの領域へはアフィンギャップペナルティにおける伸張ギャップを小さくすることにより対応し、コドン縮約表現へ対応するため、表２のように塩基の置換行列に対して拡張を行なった。ローカルアライメントによる機能部位検出のイメージ例を図３に示す。

表２　拡張した置換行列.

	a	g	c	t	r	y	m	k	s	w	h	b	v	d	n
a	10	-10	-10	-10	5	-10	5	-10	-10	5	2	-10	2	2	0
g	-10	10	-10	-10	5	-10	-10	5	5	-10	-10	2	2	2	0
c	-10	-10	10	-10	-10	5	5	-10	5	-10	2	2	2	-10	0
t	-10	-10	-10	10	-10	5	-10	5	-10	5	2	2	-10	2	0

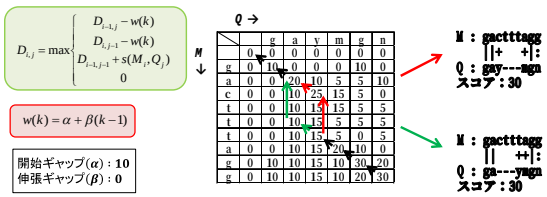


図３　ローカルアライメントによる機能部位検出.

(3) 次に、網羅的なタンパク質-遺伝子モチーフ辞書構築のため、NCBI RefSeq データベースから、ヒトをはじめとする７種類のモデル生物種ごとにタンパク質アミノ酸配列データを抽出し、コドン縮約表現をもとにそのコーディング領域を推定し、対応データセットを生成する。

アミノ酸配列モチーフ辞書 PROSITE で正規表現のパターンで定義されている配列モチーフを対象とし、アミノ酸配列データに対して配列パターン検索を行ない、配列モチーフのパターンと出現位置情報を取得する。そして、モチーフパターンの位置情報から対応塩基配列のモチーフ対応部位を特定・抽出する。この塩基配列の対応モチーフ部位について、コドン単位の塩基の出現頻度を計算し、重み行列を生成した。また、モチーフパターン中にギャップの幅（自由度）があるとき、重み行列はそれぞれの長さごとに分類し、各位置における塩基の出現頻度を求める。

(4) さらに、タンパク質の立体構造データベース PDB を参照し、対応する三次元モチーフの情報を関連付ける。ここで、立体構造は複数の鎖で構成される（複合体）があること、また、立体構造データにアミノ酸欠損や置換が存在する場合があることから、改めて、モチーフ配列をクエリに鎖単位のアミノ酸配列と文字列マッチングを行ない、該当部位が存在するかを確認・抽出する。

４．研究成果

(1) NCBI RefSeq データベースから抽出したヒト由来のゲノム塩基配列データセットに対して、GPCR の一種である匂い受容体に存在するモチーフを例として、タンパク質コーディング領域の推定実験を試みた。

匂い受容体に存在する MAYDRYVAIC モチーフを逆翻訳して近似した遺伝子配列で表すと、atg gcn tay gay mgn tay gtn gcn ath tgy となる（図４）。これをクエリとし、データセットに対して機能部位推定を行なった。その結果、パターンに完全に一致する機能部位だけでなく、類似する機能部位を検出することができた。これらのエントリのヘッダー情報を基に、RefSeq のアノテーション情報と照合した結果、いずれも匂い受容体のタンパク質のコーディング領域に保存されていることを確認した。特に、MAYDRYVAIC モチーフの第３残基目の Y と第７残基目の V は保存性が低いことが知られており、今回検出した機能部位もその保存性に対応した結果であることが確認できた。

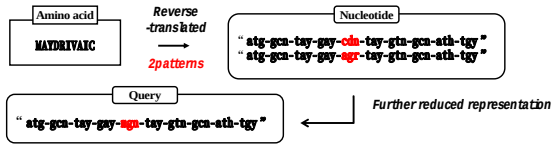


図４　モチーフパターン MAYDRYVAIC の逆翻訳と近似表現.

(2) コドン重み行列に対して閾値を与え、より出現頻度の高いものに注目した。このパターンをワイルドカードで表現したものを、モチーフコドン縮約表現と定義した。

これをクエリとし、データセットに対してゲノムの機能部位推定を行なった。モチーフパターンとして EF-hand モチーフ(PS00018)を使用した。ここで、EF-hand モチーフのモチーフコドン縮約表現は gay vhn ray vrn ray ggv hhv vtb dvy nwn vad gar htb となった。機能部位推定の結果から、このモチーフを保存していることが知られているエントリ TNNC1 (NG_008963.1) (9,951bp)に注目した。このエントリでは、２ヶ所の機能部位候補が検出された。この機能部位候補は、RefSeq のアノテーション情報と照合した結果、図５のように、エクソン-イントロン領域に保存されていることを確認した。

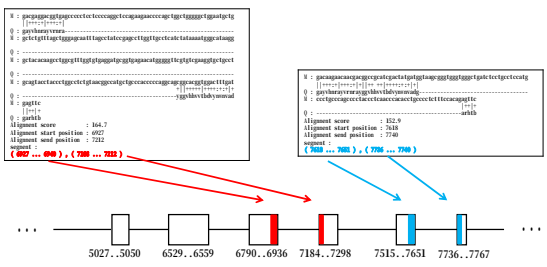


図５　troponin C type 1 (TNNC1) の EF-hand モチーフ保存領域.

(3) アミノ酸配列モチーフデータベース PROSITE の正規表現で定義された 1,309 パターンについて、RefSeq の７種類のモデル生物種ごとにタンパク質-遺伝子モチーフの検索と集積を行なった。ヒトデータセット（約４

万エントリ）の例では、907 パターンについてその対応モチーフ部位が 1 件以上ヒットし、それぞれコドン重み行列の生成を行ない、データベースを構築した。

構築したデータベースを管理・利用するためのユーザインターフェースを含めた検索システムを併せて開発した。ここで、遺伝子モチーフのコドン重み行列を可視化できるよう工夫した。これにより、どの位置でどの塩基が許容されやすいかが視覚的に確認できるようになった（図６）。

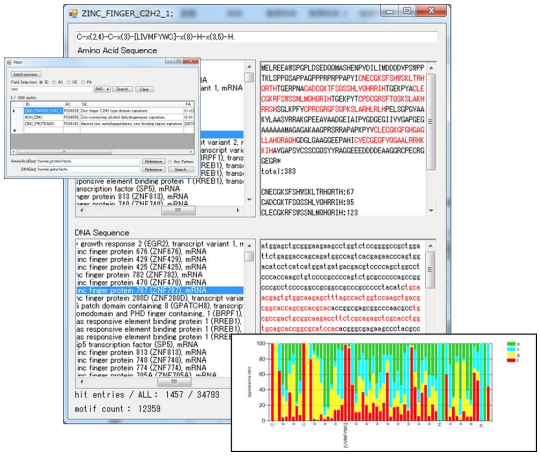


図６　タンパク質-遺伝子モチーフ辞書検索システムの実行画面例.

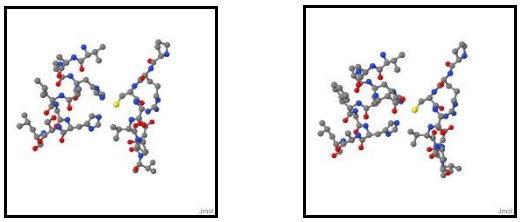
(4) 開発したシステムを用いて、例えば、Zinc-finger モチーフ(PROSITE-ID: PS00028)の特徴解析を試みた。ヒト由来のデータセットの場合、二つの明示的なシステインに挟まれた幅を有するギャップ領域 x(2,4)について、x(2)に対応するパターンの頻度が全体の約 95%と非常に高い値を示した。このシステインについて、配列全体ではコドンの第３位置において、シトシン(c)とチミン(t)、それぞれ約５割ずつの出現頻度であるのに対し、Zinc-finger モチーフ領域内ではシトシンの出現頻度が低く、特に 2 番目に出現するシステインのコドン出現頻度において、特徴的なパターンを見出すことができた。

(5) データベースに登録されている RefSeq エントリ全件に登録されたモチーフを対象にモチーフ部位抽出を行い、モチーフ立体構造ファイル(部分 PDB 形式)を生成し、データベースに追加した。

次に、モチーフ間の共起関係に注目した解析実験を試みた。ここでは、アミノ酸配列上でのモチーフの出現順序だけではなく、アミノ酸配列距離（モチーフ間の残基数）と、空間距離（モチーフ立体構造の重心座標間のユークリッド距離）を算出した。例として、分解酵素に関係する zinc_protease (PS00142) と cysteine_switch (PS00546)モチーフの二つを内包するヒトの MMP タンパク質ファミリーについて、表３にまとめた。その結果、配列距離は 106 残基から 295 残基と幅があるのに対して、空間距離はすべて約 12Å と構造的によく保存されていることを見出すことができた（図７）。

表３　zinc_protease と cysteine_switch モチーフの共起関係.

RefSeqID (タンパク質名)	PDBID	配列距離 (残基)	空間距離 (Å)
NM_002421.3 (MMP1)	1SU3A	108	12.0376
NM_002421.3 (MMP1)	1SU3B	108	12.0045
NM_002422.3 (MMP3)	1SLMA	116	12.0547
NM_004994.2 (MMP9)	1L6JA	295	12.0139



1SU3A (MMP1)　　1L6JA (MMP9)

図７　共起する二つのモチーフ部位の立体構造.

(6) 生体高分子の構造データは分子進化や統御メカニズムの解明だけでなく、医農薬品開発の標的としても極めて重要である。モチーフ構造の集積は、従来、アミノ酸配列のレベルで中心的に行われていた。遺伝子配列データに関しては、単純な TATA ボックスなど転写シグナルのごく限られたパターンのみで、その配列特徴を十分に表現し切れていないのが実情であった。本研究では、アミノ酸配列モチーフを原点として、その対応する遺伝子配列（コーディング領域）、およびそのタンパク質立体構造までを関連づけた新規のモチーフ辞書の構築を試みた。モチーフに注目したコドン出現頻度の解析や、モチーフをキーとしたゲノムの機能部位推定も含めたその手法の確立は生体高分子の構造情報解析における一つの重要な要素技術をなすものであり、その意義は極めて大きい。

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）

〔学会発表〕（計１１件）

[1] Akihiro Hirama, Hiroaki Kato, Development of the protein 3D fragment analysis system focused on ligand binding loop regions, 第 44 回構造活性相関シンポジウム, 2016 年 11 月 16 日～11 月 17 日, 京都大学芝蘭会館（京都府京都市）
[2] Yoshiyuki Sato, Hiroaki Kato, Development of the common fragment set extraction system for compound-protein relationship studies, 第 44 回構造活性相関シンポジウム, 2016 年 11 月 16 日～11 月 17 日, 京都大学芝蘭会館（京都府京都市）
[3] Takashi Yamamoto, Hiroaki Kato, Development of the protein motif

extraction system based on an Artificial Bee Colony algorithm, Chem-Bio Informatics Society (CBI) Annual Meeting 2016, 2016年10月25日～10月27日, タワーホール船堀（東京都江戸川区）
[4] Taku Matsubara, Hiroaki Kato, Development of the protein-gene sequence motif extraction system based on EM algorithm, Informatics in Biology, Medicine and Pharmacology 2016, 2016年9月29日～10月1日, 東京国際交流館プラザ平成（東京都江東区）
[5] Hiroaki Kato, Junki Yamamoto, Development of the protein-gene sequence motif analysis system based on the codon reduced representation, 2015 International Chemical Congress of Pacific Basin Societies, 2015年12月15日～12月20日, Sheraton Waikiki (Hawaii, USA)
[6] Souta Azuma, Hiroaki Kato, Development of the structural feature analysis system based on the motif combination pattern of proteins, Chem-Bio Informatics Society (CBI) Annual Meeting 2015, 2015年10月27日～10月29日, タワーホール船堀（東京都江戸川区）
[7] Takashi Kobayashi, Hiroaki Kato, Development of the DP-based genome sequence analysis system using the motif codon reduced representation,第43回構造活性相関シンポジウム, 2015年9月27日～9月29日, 新潟日報メディアシップ（新潟県新潟市）
[8] 三浦智大, 加藤博明, 回文フラグメントに注目したタンパク質アミノ酸配列特徴解析システムの開発, 第37回情報化学討論会, 2014年11月27日～11月28日, 豊橋商工会議所（愛知県豊橋市）
[9] 山本潤基, 加藤博明, コドン縮約表現に基づくタンパク質-遺伝子配列モチーフ解析システムの開発, 第42回構造活性相関シンポジウム, 2014年11月13日～11月14日, くまもと森都心プラザ（熊本県熊本市）
[10] 佐賀勇哉, 加藤博明, 三次元近傍情報に基づく分子の構造特徴解析システムの開発, 第42回構造活性相関シンポジウム, 2014年11月13日～11月14日, くまもと森都心プラザ（熊本県熊本市）
[11] Ryosuke Itoh, Hiroaki Kato, Construction of the protein 3D fragment library system based on glycine neighboring environment, Chem-Bio Informatics Society (CBI) Annual Meeting 2014, 2014年10月28日～10月30日, タワーホール船堀（東京都江戸川区）

〔その他〕
ホームページ等
http://www.mbi.cs.tut.ac.jp/

６．研究組織
(1)研究代表者
加藤　博明（KATO HIROAKI）
豊橋技術科学大学大学院・工学研究科・講師
研究者番号：30303704