The α-helical transmembrane domains and intrinsically disordered regions on the human proteins are coded for by the skews of their genes' nucleic acid composition with the "universal" assignment of the genetic code table

Genshiro Esumi
Pediatric Surgery, Hospital of the University of Occupational and Environmental Health, Kitakyushu, Japan

It is widely known that in living organisms, nucleic acid sequences of genes are translated into amino acid sequences according to the assignment of the genetic code table, resulting in the synthesis of proteins with various functions. It is also well known that the function of an individual protein is greatly influenced by its amino acid composition. On the other hand, there are few reports on how the gene's nucleic acid composition affects the protein's function.

In this study, using publicly available data from the UniProt and NCBI, the annotation of each human protein and its gene's cording sequence was matched by the RefSeq IDs. As a result, 25095 proteins were matched. First, I calculated each protein's fractions of α-helical transmembrane domains on their sequences and that of intrinsically disordered regions. Second, I made a scatter plot by each gene's GC content and TA skew. Third, I compared these plots and fractions. The plots show that proteins with higher fractions of α-helix transmembrane domains occupy the area with higher TA skew. On the other hand, the proteins with higher fractions of intrinsically disordered regions occupy both the lower TA skew area and partially higher GC content area.

Hydrophobic and hydrophilic amino acids cluster in the genetic code table, historically explained by "robustness to mutations." However, in the actual assignment of the genetic code table, codons containing T in the first or the second letters corresponded to amino acids that mainly constitute the alpha-helical transmembrane domains. In contrast, codons without T in the first and second letters corresponded to amino acids characteristic in the intrinsically disordered regions. The plot separation of the two types of proteins in this study was speculated to originate from the assignment of the "universal" genetic code table.

Chargaff's second parity rule (CSPR) says that even in the single DNA strand, both the number of Thymine (T) and Adenine (A) and the numbers of Guanine (G) and Cytosine (C) in a genome sequence are almost identical if the sequence is sufficiently long. On the other hand, the present study showed that the numbers of T and A differ in many protein genes and that their skew bias differentiates the α-helix transmembrane domains and the intrinsically disordered regions. The origin of CSPR has been a mystery in bioscience history. However, suppose the TA skews of the genes determine the functions of the proteins. In that case, CSPR might have been maintained to keep the proper proportion of protein functions in the proteome. This assumption could support the theory that CSPR is one of the backgrounds the genome must follow to keep functional proteomes.

The result of this study might indicate that all organisms might universally control the proportions of functional domains in their proteomes by using the universal genetic code table assignment and their non-random, precisely structured genome sequences.

Keywords: transmembrane domains, intrinsically disordered regions, nucleic acid composition, TA skew, GC content, genetic code table, Chargaff's second parity rule

E-mail: esumi@clnc.uoeh-u.ac.jp

ヒトタンパク質上のα-ヘリックス型膜貫通ドメインと天然変性領域は遺伝暗号表の配列の特性により遺伝子上の核酸組成の偏りとしてコードされている

江角 元史郎
産業医科大学病院 小児外科

生物において、遺伝子の核酸配列が遺伝暗号表の対応関係によりアミノ酸の配列に翻訳され、その結果として様々な機能を持ったタンパク質が合成されていることは汎く知られている。また、個々のタンパク質の機能はそのアミノ酸組成により大きく影響されることもよく知られている。一方で、タンパク質の機能に対しその遺伝子の核酸組成がどのような影響を及ぼしているかについてはほとんど報告がない。

今回、NCBI と UniProt の公開情報を用いて、ヒトの各タンパク質のアノテーションと遺伝子のコード領域の核酸組成を RefSeq ID で突合したところ、計 25095 個のタンパク質の情報が突合された。このデータセットにおける各タンパク質のアミノ酸シーケンスにおけるα-ヘリックス型膜貫通ドメインが占める割合、天然変性領域が占める割合をそれぞれ計算し、それらをそのタンパク質の遺伝子における核酸組成から計算した GC content および TA skew の値に基づいてプロットしたところ、αヘリックス型膜貫通ドメインの割合が高いタンパク質は TA skew が高い領域に集中し、一方で天然変性領域の割合が高いタンパク質は TA skew が低い領域全体と、GC content が高い一部一部の領域に集中していることが明らかとなった。

遺伝暗号表において、疎水性の高いアミノ酸、親水性の高いアミノ酸がまとまって存在していることは古くから知られているが、これは「変異に対する頑健性があるから」と説明されてきた。しかし、遺伝暗号表を確認すると、コドン 1 文字目、2 文字目に T が含まれるコドンがαヘリックス型膜貫通ドメインに多く含まれているアミノ酸と対応し、逆にコドン 1 文字目、2 文字目に T が含まれないコドンが天然変性領域に特徴的とされるアミノ酸と対応していた。以上より、2 種類のタンパク質がプロット上で分離する背景は、ヒトが用いている遺伝暗号表の配列に起源すると推測された。

ゲノム上の核酸配列について十分に長いシーケンスであればその配列に含まれる T と A、G と C の数はほぼ一致することは、シャルガフの第 2 パリティ則として報告されている。一方で、今回の結果より、各遺伝子においては大半の遺伝子の T と A の数は一致せず、実際はその偏りによってαヘリックス型膜貫通ドメインと天然変性領域がコードし分けられている可能性が示唆された。現在までシャルガフの第 2 経験則が保たれている理由は謎であるが、仮に遺伝子が TA skew で機能性領域の作り分けをしているという仮説が正しいとするならば、この第 2 パリティ則は、タンパク質全体における膜貫通ドメインと天然変性領域の占める割合といった、タンパク質機能のバランスをコントロールするために保たれている、という可能性が考えられた。

普遍遺伝暗号表を用いるすべての生物は、その遺伝子として DNA を用いるにあたり、機能性ドメインの作り分けを容易にするような配置の遺伝暗号表と、非ランダムかつ精密にコントロールされた核酸配列のゲノムを用いることで、ゲノム上にコードするタンパク質全体の機能性ドメインの存在量のバランスが指摘範囲に入るようにコントロールを行っていると推測された。

キーワード：膜貫通ドメイン、天然変性領域、核酸組成、TA skew、GC 含量、遺伝暗号表、シャルガフの第 2 パリティ則

背景

生物において、遺伝子の核酸配列が遺伝暗号表の対応関係によりアミノ酸の配列に翻訳され、その結果として様々な機能を持ったタンパク質が合成されていることは汎く知られている[1]。また、個々のタンパク質の機能はそのアミノ酸組成により大きく影響されることもよく知られている[2]。一方で、タンパク質の機能に対しその遺伝子の核酸組成がどのような影響を及ぼしているかについてはほとんど報告がない[3]。

本研究では、遺伝子の核酸組成とその遺伝子にコードされたタンパク質の関係を検討する過程で自分が明らかにした関係について報告する。

対象と方法

NCBI のデータベースに公開されている最新ヒトゲノム（T2T-CHM13v2.0）に対応したヒトの全タンパク質の CDS データ(n=65591)[4]と、UniProt のデータベースに公開されているヒトプロテオームの全タンパク質のアノテーション情報（n=81837）[5]について、RefSeq ID とタンパク質のアミノ酸残基数で突合を行ったところ、25095 個のタンパク質情報が突合された。

まず、このタンパク質のデータセット（n=25095）について、各タンパク質のアミノ酸シーケンスにおけるα-ヘリックス型膜貫通ドメイン（α-TMD）が占める残基数の割合（TMD fraction）、天然変性領域（intrinsically disordered region; IDR）が占める残基数の割合（IDR fraction）をそれぞれ計算した。

次に、同じタンパク質のデータセット（n=25095）について、各タンパク質の遺伝子における核酸組成から、GC content と TA skew を算出し、これを用いて二次元上にプロットを行った。

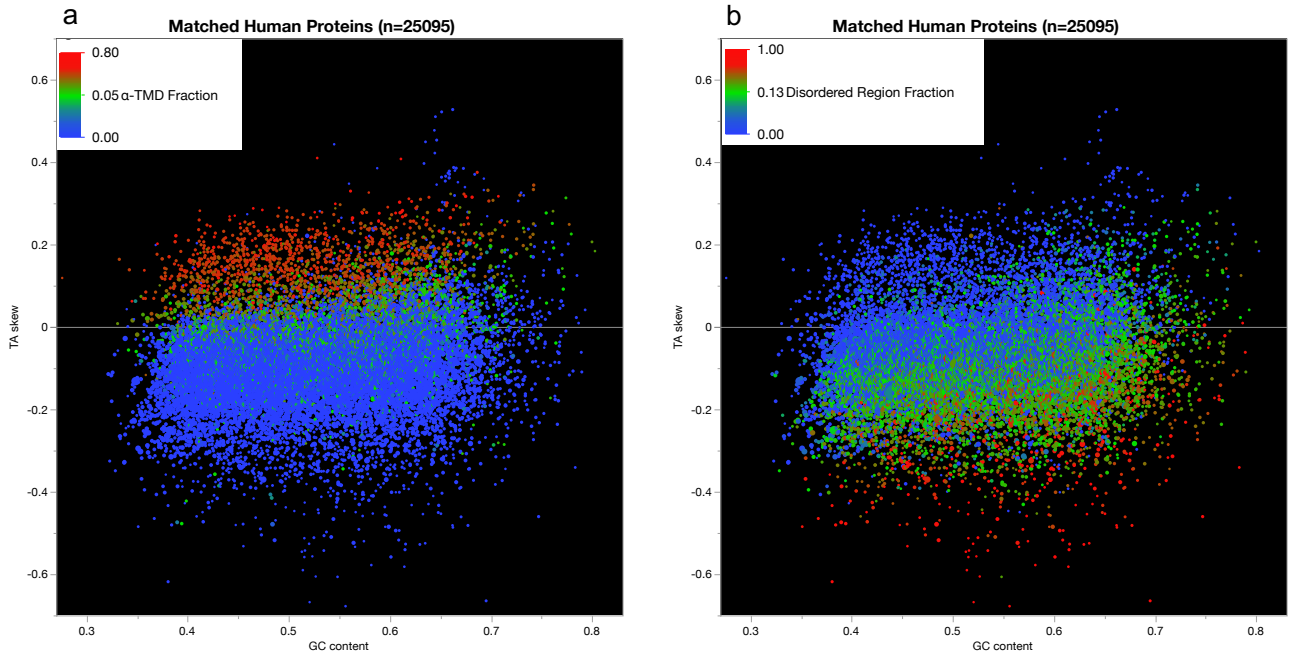
さらに、これらの 2 つの情報を合わせて、各タンパク質のプロット位置とα-TMD、IDR それぞれの fraction との関係を調べた。

本研究においては、データの突合、配列の集計・変換に Microsoft® Excel for Mac v16.69 (Microsoft Corporation, Redmond, WA, USA)を使用し、また、プロット・Figure の作成には JMP® 16.2.0 (SAS Institute Inc., Chicago, IL, USA)を使用した。

結果 1

各タンパク質のプロット（GC content、TA skew）位置と、それぞれの TMD fraction、IDR fraction を比較したところ、TMD fraction が高いタンパク質は TA skew が高い領域に集中し（Figure 1a）、一方で IDR fraction が高いタンパク質は TA skew が低い領域全体と、GC content が高い一部一部の領域に集中して分布していることが明らかとなった（Figure 1b）。

Figure 1



※Figure 1a,b においては、NCBI の CDS 情報と UniProt のアノテーション情報を RefSeqID を用いて突合した 25095 個のタンパク質データを利用した。各タンパク質の停止コドンを除いた CDS の核酸組成から計算した GC content、TA skew に基づいてプロット位置を決定し、各タンパク質の α-ヘリックス型膜貫通ドメイン（α-TMD）が全体の残基数に占める割合（Figure 1a）、天然変性領域（IDR）が全体の残基数に占める割合（Figure 1b）をもとにプロットの色を決定しプロットを行った。また各プロットのサイズは各タンパク質の残基数に合わせて調整を行った。

※Figure 1a では α-TMD の多いタンパク質が TA skew が高い領域を、Figure 2b では IDR が多いタンパク質が TA skew が低い領域全体と GC 含量が高い領域の一部を、それぞれ占拠することが判明した。

考察 1

遺伝暗号表において、疎水性の高いアミノ酸、親水性の高いアミノ酸がまとまって存在していることは古くから知られているが、これは「変異に対する頑健性があるから」と説明されてきた[6]。

しかし、今回突合したデータ（n=25095）において、全タンパク質上のアミノ酸残基数の総和の割合（アミノ酸組成）、α-TMD 部分のアミノ酸残基数の総和の割合（組成）、IDR のアミノ酸残基数の総和の割合（組成）の多寡をそれぞれ比較したところ、コドン 1 文字目、2 文字目に T（暗号表上の U）が含まれるコドンがα-TMD に多いアミノ酸の大半と対応し、逆にコドン 1 文字目、2 文字目に T（暗号表上の U）が含まれないコドンが IDR に多いアミノ酸の大半と対応していることが明らかとなった。（Figure 2a,b）

以上より、α-TMD と IDR の各ドメイン／領域を持ったタンパク質が核酸組成のプロット上で分離する背景は、遺伝暗号表の配列に起源すると推測された。

Figure 2

a

α-TMD

	U		A		C		G		
U	UUU	Phe	UAU	Tyr	UCU	Ser	UGU	Cys	U
	UUA	Leu	UAA	Stop	UCA	Ser	UGA	Stop	A
	UUC	Phe	UAC	Tyr	UCC	Ser	UGC	Cys	C
	UUG	Leu	UAG	Stop	UCG	Ser	UGG	Trp	G
A	AUU	Ile	AAU	Asn	ACU	Thr	AGU	Ser	U
	AUA	Ile	AAA	Lys	ACA	Thr	AGA	Arg	A
	AUC	Ile	AAC	Asn	ACC	Thr	AGC	Ser	C
	AUG	Met	AAG	Lys	ACG	Thr	AGG	Arg	G
C	CUU	Leu	CAU	His	CCU	Pro	CGU	Arg	U
	CUA	Leu	CAA	Gln	CCA	Pro	CGA	Arg	A
	CUC	Leu	CAC	His	CCC	Pro	CGC	Arg	C
	CUG	Leu	CAG	Gln	CCG	Pro	CGG	Arg	G
G	GUU	Val	GAU	Asp	GCU	Ala	GGU	Gly	U
	GUA	Val	GAA	Glu	GCA	Ala	GGA	Gly	A
	GUC	Val	GAC	Asp	GCC	Ala	GGC	Gly	C
	GUG	Val	GAG	Glu	GCG	Ala	GGG	Gly	G

b

IDR

	U		A		C		G		
U	UUU	Phe	UAU	Tyr	UCU	Ser	UGU	Cys	U
	UUA	Leu	UAA	Stop	UCA	Ser	UGA	Stop	A
	UUC	Phe	UAC	Tyr	UCC	Ser	UGC	Cys	C
	UUG	Leu	UAG	Stop	UCG	Ser	UGG	Trp	G
A	AUU	Ile	AAU	Asn	ACU	Thr	AGU	Ser	U
	AUA	Ile	AAA	Lys	ACA	Thr	AGA	Arg	A
	AUC	Ile	AAC	Asn	ACC	Thr	AGC	Ser	C
	AUG	Met	AAG	Lys	ACG	Thr	AGG	Arg	G
C	CUU	Leu	CAU	His	CCU	Pro	CGU	Arg	U
	CUA	Leu	CAA	Gln	CCA	Pro	CGA	Arg	A
	CUC	Leu	CAC	His	CCC	Pro	CGC	Arg	C
	CUG	Leu	CAG	Gln	CCG	Pro	CGG	Arg	G
G	GUU	Val	GAU	Asp	GCU	Ala	GGU	Gly	U
	GUA	Val	GAA	Glu	GCA	Ala	GGA	Gly	A
	GUC	Val	GAC	Asp	GCC	Ala	GGC	Gly	C
	GUG	Val	GAG	Glu	GCG	Ala	GGG	Gly	G

※今回突合したタンパク質データ（n=25095）において、全タンパク質上のアミノ酸残基数の総和の割合（アミノ酸組成）、α-ヘリックス型膜貫通ドメイン（α-TMD）部分のアミノ酸残基数の総和の割合（組成）、天然変性領域（IDR）部分のアミノ酸残基数の総和の割合（組成）について、全体組成よりもα-TMD において組成量が多いアミノ酸（Figure 2a）、全体組成よりも IDR において組成料が多いアミノ酸（Figure 2b）を、遺伝暗号表上に囲み表示した。

※遺伝暗号表上の U（Uracil）は DNA 配列上の T（Thymine）に対応する。

※一般に遺伝暗号表では 4 つの核酸を UCAG の順に並べて表示されるが、本論文では遺伝子上の TA skew、GC content に着目しているため、核酸の順番を UACG の順とした遺伝暗号表を掲載した。

考察 2

ゲノム上の核酸配列について十分に長いシーケンスであればその配列に含まれる T と A、G と C の数はほぼ一致することは、Chargaff's second parity rule（CSPR）として報告されている[7]。一方で、今回の結果より、各遺伝子においては大半の遺伝子の T と A の数は一致せず、実際はその偏りによってα-TMD と IDR がコードし分けられている可能性が示唆された。現在 CSPR が保たれている理由は謎とされている[8]が、仮に遺伝子が TA skew で機能性領域の作り分けをしているという仮説が正しいとするならば、CSPR は、タンパク質全体におけるα-TMD と IDR の占める割合をコントロールするために保たれている、という可能性が考えられた。

まとめ

今回の解析の結果、ヒトのタンパク質上のα TMD と IDR は、核酸組成の偏りとして遺伝子上にコードされていることが示された。

普遍遺伝暗号表を用いるすべての生物は、その遺伝子として DNA を用いるにあたり、機能性ドメインの作り分けを容易にするような配置の遺伝暗号表と、非ランダムかつ精密にコントロールされた核酸配列のゲノムを用いることで、ゲノム上にコードするタンパク質全体の機能性ドメインの存在量のバランスが指摘範囲に入るようにコントロールを行っていると推測された。

引用文献・サイト

1. B. Alberts 他著, 中村桂子, 松原謙一監訳,『細胞の分子生物学　第 6 版』(ニュートンプレス, 2017).
2. Carugo, O. (2008). Amino acid composition and protein dimension. *Protein Science*, *17*(12), 2187–2191. https://doi.org/10.1110/ps.037762.108
3. 江角　元史郎. (2022). 膜貫通ドメイン合成支援は遺伝暗号表配列の重要な機能である. *Jxiv.* https://doi.org/10.51094/jxiv.139
4. National Center for Biotechnology Information (NCBI). (2022). Genome assembly T2T-CHM13v2.0. *National Library of Medicine (NIH) website.* https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_009914755.1/
5. UniProt consortium. (2023). Proteomes · Homo sapiens (Human). *UniProt website.* https://www.uniprot.org/proteomes/UP000005640
6. Radványi, Á., & Kun, Á. (2021). The Mutational Robustness of the Genetic Code and Codon Usage in Environmental Context: A Non-Extremophilic Preference? *Life*, *11*(8), 773. https://doi.org/10.3390/life11080773
7. Rudner, R., Karkas, J. D., & Chargaff, E. (1968). Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proceedings of the National Academy of Sciences*, *60*(3), 921–922. https://doi.org/10.1073/pnas.60.3.921
8. Fariselli, P., Taccioli, C., Pagani, L., & Maritan, A. (2021). DNA sequence symmetries from randomness: the origin of the Chargaff's second parity rule. *Briefings in Bioinformatics*, *22*(2), 2172–2181. https://doi.org/10.1093/bib/bbaa041