氏　　名	増山　和花
学位（専攻分野）	博士（理学）
学位記番号	総研大甲第1248号
学位授与の日付	平成２１年３月２４日
学位授与の要件	生命科学研究科　遺伝学専攻 学位規則第６条第１項該当
学位論文題目	Evolutionary Analysis of Protein Domains in Mammals
論文審査委員	主　査　教授　五條堀　孝 教授　角谷　徹仁 准教授　酒井　則良 助教　高橋　文 室長　深海　薫（理化学研究所）

論文内容の要旨

Protein domains are considered as fundamental units of protein evolution. Many proteins consist of multiple domains, and domain combinations are closely linked with the function and evolutionary history of proteins. Domain loss and gene loss may lead to immediate loss of gene function. Domain combination of genes could also potentially attribute to the diversity of organisms. My purpose is to search human-, primate-, rodents-, and mammal-specific domain combinations created through domain gain and loss. I retrieved domain combinations of vertebrate proteins through genome data, and defined a repertoire of domain combinations of an organism as a set of combinations encoded in its genome. Pfam (http://pfam.sanger.ac.uk/) was used for the analysis as a domain database. I extracted domain information and functions of vertebrate proteins from GTOP (http://spock.genes.nig.ac.jp/~genome/gtop.html). In order to examine the phylogenetic distribution of domain combinations, I followed the nine-step procedures. At procedure 1, I enumerated 17,358 domain combinations from 37 Metazoa genome data set. At procedure 2, I selected domain combinations of eight species with steady genome data (human, chimpanzee, rhesus macaque, mouse, rat, dog, opossum, and chicken). Opossum and chicken were used as outgroup to placental mammals. At procedure 3, I removed domain combinations of outgroup species from the list of domain combinations left after procedure 2. At procedure 4, I selected lineage specific domain combinations (human-, primate-, rodents-, and mammal-specific). At procedure 5, I selected only the protein sequences existing in SWISS-PROT (http://au.expasy.org/sprot/). At procedure 6, I selected data with high scores by BLASTP as orthologue candidates. At procedure 7, I retained well-aligned sequences by using ClustalW multiple alignment. At procedure 8, I constructed phylogenetic trees to identify orthologues. At procedure 9, I eliminated domain combinations that were also found in more than one outgroup Metazoan species.

After following procedures 1 to 9, I selected 34 mammal-specific, 17 primate- specific, 10 human-specific, and 14 rodent-specific domain combinations. As the branch in the species tree becomes shorter, the frequency of the appearance of novel domains and novel domain combinations tends to decrease. This seems to be a major reason why the number of lineage-specific domain combinations decreases as we go down the species tree from mammalian ancestors via primate ancestors to humans.

I classified the remaining proteins into seven groups as follows; group 1: emergence of new domain, group 2: insertion of one (or more) domain with existing protein, group 3: deletion of one (or more) domain from existing protein, group 4: change of domain order, group 5: increase of domain copy number, group 6: decrease of domain copy number, and

group 5/6: case when it is not possible to distinguish increase or decrease of domain copy number. As a result, four lineage-specific domain combinations were decomposed into these seven groups: mammal-specific (14, 10, 1, 0, 7, 2 and 0), primate-specific (3, 4, 1, 0, 4, 5 and 0), human-specific (1, 0, 0, 0, 8, 0 and 1), rodent-specific (0, 0, 1, 0, 0, 0 and 13) in the order of groups 1, 2, 3, 4, 5, 6, and 5/6, respectively, in parentheses. There was no group 4 proteins. Mammal-specific group 1 proteins include various interleukins and macrophage scavenger receptor. This implies that the mammals upgraded the immunity system using these new proteins. Out of the seventeen proteins belonging to group 2 and group 3 for all four lineages, twelve in fact did not undergo deletion and insertion of domains that are very definitions of these two groups. These sequences have gradually changed their amino acid sequences via substitionts as the species evolved, and their regions came to be recognized as domains at certain points in evolution. To put it another way, even orthologous proteins that appear to have no domains actually have "pre-domain" sequences which can be detected with thresholds lower than the default value of the domain search program. This was revealed by the fact that, for these proteins, the taxonomic ranges of domain annotation vary depending on the thresholds fed into the domain search program. Other than the above-mentioned proteins, group 2 also includes signal transduction proteins, for example, regulator of G-protein signaling 3 protein, LIM kinase 2 and ubiquitin carboxyl terminal hydrolase. These proteins may have been involved in the evolution of lineage-specific signaling networks. Interestingly, lineage-specific domain combinations caused by insertions or deletions of domains always accompany alternative splicing products or paralogous proteins of the ancestral forms, which seem to alleviate the risk of impaired original functions, which could lead to lethality. It's interesting such backups are working as if they were a kind of insurance.

Aside from my main procedures 1 through 9, I also conducted secondary analyses and found several domain combinations that do not belong to the seven groups mentioned above. Examples of proteins with such domain combinations are: sex-determining region Y protein (SRY), which is conserved throughout therians but is absent in other vertebrates; uricase and L-gulonolactone oxidase, which are absent only in primates.

In this research, I focused on mammals and performed analysis. Compared to the numbers of lineage specific, new domain appearance including bacteria, appearance of nobel domain in the mammalian lineage is rare because of the short evolution time. In such a situation, however, I found 34 mammal-specific, 17 primate -specific, 10 human-specific, and 14 roden-specific domain combinations. As a result of detailed analysis, I was able to discover five cases where domain insertion occured.

論文の審査結果の要旨

タンパク質には、アミノ酸配列やそれをコードする遺伝子の塩基配列から機能ドメインという部分領域が予測され、それらの一部においては担当する機能が実験的に検証されている。通常、タンパク質は複数の異なる機能ドメインから構成されており、機能ドメインの獲得や消失などの変化がそのタンパク質に新規の機能を付与したり既存の機能を失くしてしまったりするものと考えられる。この機能ドメインや、それらの組み合わせの進化過程を理解することは、タンパク質やそれをコードする遺伝子の進化を解明する上で非常に重要である。

このため、増山さんは、タンパク質とそれをコードする遺伝子において、機能ドメインの組み合わせの進化過程が遺伝子機能の進化的変化にどう関連するかを解明することを目的として、完全ゲノムが分かっている哺乳類の遺伝子の塩基配列やタンパク質のアミノ酸配列のデータを比較解析することによって、系統特異的な機能ドメインの抽出を行うとともに機能ドメインの組み合せの出現過程を推定することを試みた。

具体的には、*Pfam* という機能ドメインのデータベースや遺伝研 DDBJ が公開している GTOP というデータベースを駆使して、そこで定義されまたアノテーションされた機能ドメインに基づいて、系統特異的ドメインを抽出した。基本的なアプローチとして相同性検索を用いたが、哺乳類という比較的に近縁な生物種に焦点を絞ったこともあって、機能ドメインの抽出は効率よく行われた。また、その抽出するプロセスにおいて、任意性や曖昧性をできるだけ排除するため、9つにわたる手順のステップを明確に設定して、配列の相同性に基づく機能ドメインの抽出を行った。

この結果、機能ドメインの出現過程は、獲得や消失を含む6つのグループに分類されることを明らかにした。また、機能ドメインが組み合わさることで出現したと思われるようなタンパク質も5つ発見し、哺乳類グループという近縁な生物種グループとはいえ、機能ドメインのダイナミックな進化過程が存在することを明らかにした。

増山さんのこの研究は、哺乳類における機能ドメインの全リストが完成したことを意味し、上記のような新しい知見の発見とともに、哺乳類のゲノムや遺伝子にコードされる現在わかっているすべてのタンパク質において、その機能ドメインの進化研究に、大きな知的基盤を提供したことになる。さらに、それぞれの遺伝子について、機能との関連についても詳細な記載がなされており、機能ドメインの進化的な起源や進化メカニズムを解明するための今後の発展の基礎を与えたものと考えられる。さらに、このような機能ドメインの進化過程が生物の多様性にどのように寄与したかなどの重要な問題にも大きな手がかりを与えるものである。

以上のことから、増山さんの本研究と学位論文は学位を授与するに十分に値すると判断した。