氏　　名	鈴木　留美子
学位（専攻分野）	博士（理学）
学位記番号	総研大甲第1387号
学位授与の日付	平成22年9月30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	Analyses of Highly Conserved Nucleotide Sequences within Protein Coding Regions of Eukaryotes
論文審査委員	主　査　教授　明石　裕 教授　角谷　徹仁 助教　鈴木　善幸 助教　髙橋　文 教授　池村　淑道　（長浜バイオ大学）

Nucleotide substitutions in the synonymous sites of codons do not alter amino acid sequences, therefore they are considered to be basically neutral. In some cases, however, synonymous sites accept selective constraints.

Requirement for translational efficiency or accuracy enhances the optimum codon usage and suppress the synonymous changes. Other than this, a certain region of a protein coding gene may function as exonic splicing signals, RNA editing targets, and RNA secondary structures that affect on gene expression. There is also possibility that messenger RNAs have interaction with non-protein coding transcripts. The existence of such functional regions would be detected from suppression of nucleotide substitution in the area.

Preceding studies have revealed many facts about codon biases and exonic splicing signals, however, other factors have not been extensively surveyed. The aim of this study is to explore unknown factors that affect on the nucleotide conservation in the coding regions in various taxa and to predict potential functionality of the conserved sequences.

For this purpose, I investigated invariant sequences in orthologous genes in seven taxa: mammals (*Homo sapiens, Macaca. mulatta, Mus musculus, Rattus norvegicus, Bos taurus, and Canis familiaris*), teleosts (*Tetraodon nigroviridis, Takifugu rubripes, Gasterosteus aculeatus, Oryzias latipes*), *Drosophilas* (*Drosophila melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae*), nematodes (*Caenorhabditis elegans, C. briggsae, C. remanei, C. japonica*), dicots (*Arabidopsis thaliana, A. lyrata, Vitis vinifera*), monocots (*Oryza sativa japonica, O. s. indica, Sorghum bicolor, Brachypodium distachyon*), and budding yeasts (*Saccharomyces cerevisiae, S. paradoxus, S. mikatae, S. bayanus*). Analysis on major codon ratio, GC content, and codon degeneracy revealed different characteristics of invariant sequences among the taxa. The major codon ratio decreases as the conservation length elongates in the four animal taxa (mammals, teleosts, *Drosophilas*, and nematodes), while GC content and codon degeneracy do not show notable sway. This result implies that selection toward optimum codons may not be the dominant factor in the above taxa.

To extract sequences whose conservation is stronger than others, I took a permutation approach. I permuted codons of each alignment and surveyed the length and frequency of invariant sequences in the permuted alignment. In the mammals, the result of permutation showed significant deviation from the number of invariant sequences in the original alignments ($p < 2.2E\text{-}16$) but deviation is subtle in budding yeasts. This result implies that the distribution of conserved sites is skewed in mammals, while the distribution is rather homogeneous in the budding yeasts. I extracted from the original alignments invariant sequences that have significantly low expectancy ($P < 0.01$) in comparison with the permutation results and defined them as significantly conserved coding sequences (SCCSs).

These analyses revealed different characteristics of conserved nucleotide sequences among the taxa. In mammals and teleosts, it's not likely that long SCCSs has been retained solely by amino acid constraint judging from the codon degeneracy and negative correlation between the conservation length and major codon ratio. The sequence characteristics and skewed distribution of conserved sites predicted from the permutation result suggest that SCCSs of the above tow taxa have rather preferable traits as functional nucleotide elements.

SCCSs are extracted from the alignments of orthologs in one taxon, but there are SCCSs that are observed in the similar regions of orthologs in multiple taxa. Although the sequence similarity varies case by case, it is notable that SCCSs are alignable between the distant taxa. This may happen if strong amino acid constraint is working on the similar regions of the genes. However, such constraint does not fully explain the SCCSs that exist out of protein domains, where amino acid constraint is not necessarily strong, and how synonymous sites in codons of moderate degeneracy are retained for a long period.

There are cases that specific RNA secondary structures exert some functions. I computationally predicted RNA secondary structures of SCCS regions using Vienna RNA package and detected five SCCSs that form secondary structures of significantly low folding free energy ($P < 0.05$). The corresponding regions of platypus and opossum orthologs showed sequence similarity but the structures are more stable in the placental mammals. Although the roles of these structures are unknown, strong conservation and significantly low free energy suggest the possibility that these regions have some functions.

As for mammals, I investigated exonic splicing signals and non-protein coding RNAs that overlaps with SCCSs or non-SCCS coding regions. No significant difference is observed in splicing signal density between SCCSs and non-SCCS coding regions, however, the component of non-protein coding RNAs overlapped with SCCSs show difference from those overlapped with non-SCCS regions. This result suggests that non-protein coding RNAs may have some association with SCCSs in mammals.

博士論文の審査結果の要旨

Detecting functional regions of genome sequences is a central challenge in bioinformatics. Nucleotide conservation among distantly related species has been used to identify candidates for functional DNA elements but most studies to date have examined only non-coding regions in a limited number of taxa. Rumiko Suzuki's doctoral dissertation research focused on detecting highly conserved regions within protein-coding regions and compared patterns among microbes, plants, insects, worms and vertebrates.

Predicted genes from over 30 genomes from seven taxa were analyzed. The taxa are budding yeasts (4 genomes), monocots (4), dicots (3), nematodes (4), Drosophila (6), teleosts (4), and mammals (6). Suzuki aligned orthologous genes within each taxon and employed a permutation approach to identify significantly conserved regions within protein-coding genes. Codon positions in each alignment were shuffled randomly to create sets of simulated genes that preserve the overall level of conservation of the data but randomize within-locus clustering of divergence. Simulated sequences were used to determine probabilites of perfect conservation of contiguous regions within taxa and to identify "significantly conserved coding sequences" (SCCS).

Interestingly, mammals and teleosts show strong excesses of SCCS's but other taxa show less deviation from expected numbers. This was surprising because natural selection should be more effective in taxa with large effective population sizes (such as yeast and Drosophila). Differences in levels of divergence within the taxa examined will need to be considered to establish differences in SCCS abundance among species.

Strong sequence conservation within protein-coding genes may reflect regional constraints related to protein structure and function. Translational selection (to optimize protein synthesis) can cause purifying selection at the synonymous positions of codons constrained at the amino acid level. Strong negative correlations between SCCS length and major codon usage (a measure of translational selection) are also unexpected findings of this research. Although the cause of these patterns are unclear, the trends suggest that translational selection is not a major contributor to sequence conservation of long SCCS regions. In addition, shared SCCS regions among distantly related taxa provide compelling evidence of their functional importance.

Functional annotations of genes that harbor SCCS regions are similar to those located near conserved non-coding regions. These include genes with important roles in transcriptional regulation. Overall, this research has addressed a relatively unexplored aspect of genome evolution, strong conservation of contiguous stretches of coding DNA, and has revealed a number of patterns that motivate further experimental study.