氏名	Kryukov Kirill		
学位（専攻分野）	博士（理学）		
学位記番号	総研大甲第871号		
学位授与の日付	平成１７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	Development of new methods for evolutionary data analysis		
論文審査員	主査	教授	西川　建
		教授	舘野　義男
		教授	倉田　のり
		助教授	小出　剛
		教授	後藤　修（京都大学）

論文内容の要旨

My PhD study belongs to the field of computational biology and is focusing on development of new methods for molecular biology data analysis. My PhD paper includes three chapters, that are focusing on computational methods for different stages of biological study.

In the first chapter, titled: MISHIMA: a new method of multiple sequence alignment , I explore a possibility of applying advanced computational techniques to the problem of multiple molecular sequence alignment. Sequence alignment is one of the central tasks in molecular biology DNA or protein sequences must be aligned before any comparison can be done between them. Although alignment of two sequences already reveals valuable information about sequence relationship, some studies require multiple sequences aligned together. Such studies include phylogenetic analysis, identification of conserved genome elements and protein secondary structure prediction.

Common methods of multiple sequence alignment are usually based on pairwise sequence comparison all pairs of sequences are compared separately and then multiple alignment is constructed through the progressive alignment procedure. This method works well for aligning relatively short sequences, but takes too long time to align genomic sequences, and also when the number of sequences is large. These days the continuously increasing amount of available genomic sequences of various organisms requires some more efficient techniques for aligning such huge data.

The new method of multiple sequence alignment, that I was developing during the last year MISHIMA (a Method for Identifying Sequence History In terms of Multiple Alignment) is an attempt to reduce the computational requirement of alignment procedure of multiple genomic sequences. This is achieved through the heuristic approach to the quick extraction of potential homology information from the sequences. After that sequences are aligned using the Divide and Conquer approach: regions of homology shared by multiple sequences are used as a points of splitting sequences into parts, which are aligned independently from each other by conventional alignment method. The partial alignments are then assembled together to construct the final multiple alignment.

The homology extraction step is the key part of this method. It is based on the observation that the chance of every sequence motif (short sequence fragment) to represent a homology signal is related with the frequency of this motif occurrence in the sequence dataset. Sequence motifs that are rare, or oppositely very abundant in the sequence dataset, are unlikely to happen in the region of homology. On the other hand, the motifs that are occurring exactly once in each of the input sequences have a good chance to belong to the conserved element, thus revealing the probable homology shared by multiple sequences.

The heuristic method of homology extraction used in MISHIMA depends on counting the number of occurrences of every sequence motif of up to K nucleotides long in the sequence dataset. The number of all sequence motifs of length K is very large (it is proportional to K^4), so the important problem was to organize the information about motif frequencies. In MISHIMA method I use dictionary structure for storing the motif frequency data in efficient way, allowing information about motifs of up to 12 nucleotides long to be stored using about 0.5 GB of computer RAM.

MISHIMA alignment method was tested with several datasets, and compared with alternative methods. One of the datasets consisted of 10 complete mitochondrion genomic sequences of mammalian species. MISHIMA method could successfully construct the alignment for this dataset, taking about two minutes. ClustalW (most widely used multiple alignment software today) takes several hours to produce the alignment of the same data. Among the other test datasets was a set of 4 complete genomes of different strains of Streptococcus pyogenes, each about 2 MB long. MISHIMA method could align the dataset taking about 6 hours on Pentium 4 notebook machine with 1 GB of RAM. This test shows that this method can bring the possibility of large scale genomic multiple alignment experiment to the users of ordinary desktop or portable computers.

Second chapter of my work SMAP: Alignment with Reference Sequence is describing a technique for assisting a sequencing experiment. In a common whole genome shotgun-sequencing project a target species chromosome is divided into fragments, such as BAC (bacterial artificial chromosome), with length of several to one about hundred KB. These fragments are then sequenced, resulting in a number of sequence reads , usually less than 1 KB in length. These reads are assembled together to form contigs —a basic unit of resulting sequence. The location of each contig in the genome is not known at this stage.

The analysis of the set of contigs may be easier in case when a genome of a closely related species is already determined. In the process of sequencing genome of species A, genome of a closely related species B can be used as a reference, to supervise and assist the sequencing process. If A and B are close to each other most of the newly sequenced contigs will be found to be homologous to some part of the reference. This homology suggests their probable location in the A genome, that can be used to estimate the progress of sequencing process. Also this information can be used to assist the sequencing process, especially at the late stage of finishing the sequence. Comparison with reference sequence give the estimation of size and location of gaps —still unknown regions of target genome. Also reference sequence can help to assemble the contigs. In some cases the information about contig homology in reference sequence is enough to correctly assemble the continuous sequence of newly sequenced genome.

To implement this idea I developed SMAP —a software package for assisting a sequencing process with the help of the genomic DNA sequence of a closely related species. Its name came from the original idea —Sequence MAPping. BLAST local homology search tool is used for detecting homology between the original sequence fragments or contigs and the reference. SMAP then analyzes the result of BLAST search and performs the mapping and assembling of the set of contigs. SMAP was already applied in the process of chimpanzee chromosome 22 sequencing, when human chromosome 21 sequences were used as a reference.

Third part of my study Netview: Constructing and visually exploring phylogenetic networks is describing a new method for phylogenetic analysis. Phylogenetic relationship of a group of gene sequences is commonly represented as a tree. However a non-tree phylogenetic structure may be more appropriate in some cases. Such cases may result from recombination or horizontal gene transfer events.

Also a non-tree structure may appear because of ambiguity in the sequence data. In this study I proposed a method to explore such non-tree structures, based on contradictions between the aligned sequence data and a phylogenetic tree topology constructed by using the neighbor-joining method.

The Netview method of network construction is based on a comparison of a multiple sequence alignment data and a phylogenetic tree, based on that alignment. Every alignment position can be characterized by a certain relation with the tree —it can either support tree topology or contradict to it. Alignment sites that support tree topology don t require further analysis, but alignment positions that contradict with the tree represent the data that may need some additional explanation. Such sites show a conflict between the sequence data and the tree, so a more complex topology, such as network, may be needed to explain the data. Netview method counts different patterns of conflicting data and constructs a network by introducing an additional dimension to the tree.

I developed a program Netview implementing this method. Netview implements a graphical interface that lets user select a particular pattern of incompatibility between the aligned sequence data and the phylogenetic tree. The network is then re-constructed for selected pattern. The sequence data, which is shown for each case, also plays important role in interpreting the observed network structure. Also Netview has a convenient 3-dimensional network viewing tool, that is useful for navigating and exploring a phylogenetic structure. It is convenient to be able to change the size and projection angle to examine the network carefully.

論文の審査結果の要旨

申請論文は３章からなり、「進化的データ解析のための新手法の開発」として開発された３種類のコンピュータ解析ツール（MISHIMA, SMAP, Netview）について、それぞれ述べられている。これらのうち、第１の MISHIMA（Method for Identifying Sequence History In terms of Multiple Alignment の略）はゲノム規模の複数の DNA 塩基配列を一挙に多重整列させる、いわゆる多重配列アラインメントの新しい方法論の開発を目指したものであり、以下で詳述するように申請者の主要な研究に当たる。２番目の SMAP は、ゲノム配列決定のさい、近縁種ゲノムとのホモロジーを利用してアセンブリング過程を支援するための補助的ツール（既知の近縁種ゲノムを基準として実験的に得られた多数の配列断片を表示させるビューワー）であり、すでに実用化されている。３番目の Netview は、系統樹を３次元のネットワークとして立体的に表示させるビューワーである。遺伝的組換えなどによって生じた変異の系統関係は通常の系統樹では正確に表現できず、ネットワーク表現を必要とする。

多重配列アラインメント（MSA）は分子進化系統樹を作成するさいの重要なプロセスであり、すでに ClustalW、T-Coffee など、MSA 用のよく知られた専用プログラムが存在する。申請者の方法（MISHIMA）が、これらの従来法と比べて大きく異なる点は、従来法では先ず２本づつの配列のペアワイズ・アラインメントを行ない、その結果を組み合せて多重配列アラインメントに至るという段階を踏むのに対して、MISHIMA では与えられたすべての配列を同時に考慮し、一挙に多重整列させることを考える点である。そのために、最長 12 塩基（24 ビット）までの可能なすべての文字列に対し、それぞれの文字列の出現頻度を数えて「辞書」を作成する。M 本の入力配列に対して各々の配列に少なくとも１回づつ出現し、かつ出現回数の多くない文字列を探し、「モチーフ」とする。とくに各配列に一度づつしか出現しないものは「完全モチーフ」と呼ぶ。完全モチーフおよびモチーフどうしの組み合せ（前後関係が配列によって異なる組み合せは除く）を用いて配列のアンカー点を決める。アンカー点ごとに M 本の配列を一致させ、１つのアンカー点から次のアンカー点までの間の領域は、従来法の ClustalW を用いて多重整列させる。このように、アンカー配列の抽出とアンカー間領域の各個撃破的アラインメントの方法により、事実上、入力配列の長さに制約を課す必要がなくなり、ゲノム規模の多重配列アラインメントを可能にした。

応用例として、申請者は以下の３つの配列データセットに対して MISHIMA を適用した結果を示している。10 種類の動物のミトコンドリア完全長 DNA 配列（平均長 17 Kb、配列数 10 本）。ヒトおよびチンパンジー由来の Rh 血液型遺伝子（平均長 56 Kb,、８本）。*Streptococcus pyogenes* 菌の異型株４種の完全長ゲノム配列（平均長約 2Mb、４ 本）。卓上コンピュータ（PC）による計算時間は、それぞれ３分、10 分、約６時間であり、従来法（T-Coffee）と比べると１／10 以下に短縮できたという。これらの例のうち３番目のバクテリアゲノムの多重アラインメントは、従来法の適用限界（平均長×本数で決まる）を完全に越えており、新しい方法論によって達成された画期的な成果であると評価できる。また４種のゲノム配列は途中に数キロベースに及ぶ長大な挿入や欠失を随所に含み、アラインメントを行うのは決して容易ではない。事実、従来法（T-Coffee）を第１の配列データセット（ミトコンドリア）に適用して MISHIMA による結果と直接比較してみると、MISHIMA では１０本の配列のうちの１本に現れる長い（数百ベース長の）欠失部分が適切に処理されているが、T-Coffee を用いたときにはこの処理がうまくいかず、明らかなアランメントの乱れが見られた。このことは挿入や欠失に対してアンカー法が有効に作動していることを意味しており、アラインメントの質においても従来法より優れているといえる。

以上のように、申請者の開発した方法は、質的にも量的にも現在広く用いられている MSA 用の専用プログラムを凌駕するものであり、学位論文の水準を十分に満すものと判断した。