行政院國家科學委員會專題研究計畫　成果報告

基因體重組的研究及其軟體工具的發展和應用

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-009-141-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學生物科技學系(所)

計畫主持人：盧錦隆

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 30 日

行政院國家科學委員會專題研究計畫期末精簡報告

計畫名稱：基因體重組的研究及其軟體工具的發展和應用

計畫編號：NSC94-2213-E-009-141
執行期限：94 年 8 月 1 日至 95 年 7 月 31 日

主持人：盧錦隆 (Chin Lung Lu)
Email: cllu@mail.nctu.edu.tw
執行單位：國立交通大學生物科技系暨生物資訊研究所

一、中文摘要

隨著各種基因體序列(如 DNA、RNA 及蛋白質序列)的普及，基因體重組(Genome Rearrangement)的研究在計算生物(Computational Biology)及生物資訊(Bioinformatics)的領域上漸漸受到眾人的矚目。基因體重組研究的主要目的是藉由大型基因體的比較去衡量兩種生物體之間在演化上的差異。其中最為可行的大型基因體比較方法之一是比較兩種不同生物體之間其共有基因在基因體上次序的差異。不像插入(Insertions)、刪除(Deletions)、及取代(Substitutions)這些傳統的點突變，各種大型基因體突變(Large-Scale Mutations)例如翻轉(Reversals)、移位(Transpositions)、區塊互換(Block-Interchanges)、分裂(Fissions)、融合(Fusions)及易位(Translocations)等等已被提出並藉此以比較二個相關基因體之間共有基因次序的差異以決定出他們在演化上的距離。通常，基因體重組的研究都能夠被描述成一個找出最少的大型基因體突變來把一個基因體轉變成另一個基因體的組合數學問題。因此，這個計劃的主要目的是要去研究並解決在不同的大型基因體突變考量之下各種基因體重組的問題，特別是翻轉與區塊互換這二個大型基因突變。在計算理論方面，我們利用圖論的斷點圖的性質設計出有效率的求最佳解的演算法來解決有關於翻轉與區塊互換的基因體重組問題。在實際應用方面，我們對本計劃所設計出來的演算法去實做出他們的程式，並整合一些現有的程式去建構出基因體重組有關的網頁伺服器，來幫忙生物學家在各種大型基因體突變考量之下，去比較一些生物體的基因次序資料之間的差異，進而衡量出他們之間的演化差異。為了要讓這個系統也能處理序列資料，我們也提升系統的功能使之能夠自動化地偵測出所有輸入序列彼此之間所共有的同源或保守的區域。除此之外，這個系統也整合 NJ (Neighbor Joining) 和 UPGMA (Un-weighted Pair-Group Method using Arithmetic mean) 的程式使之也能夠利用系統所計算出來任兩生物體之間的距離重建出他們的演化樹。

關鍵詞：生物資訊；計算生物；基因體重組；翻轉突變；區塊互換突變

Abstract

With large amounts of various genomic data (DNA, RNA, and protein sequences) becoming available, the study of genome rearrangement has been drawing a lot of attentions in computational biology and bioinformatics. The main purposes of the genome rearrangement study are to measure the evolutionary difference between two organisms by conducting large-scale comparisons of their genomic data. One of the most promising ways to do this research is to compare the orders of the identical genes in two different genomes. Unlike from the traditional point mutations such as insertions, deletions, and substitutions, various large-scale mutations, such as reversals, transpositions, block-interchanges, fissions and fusions and translocations have been proposed to determine the evolutionary distance between two related genomes by comparing the gene orders. Usually, the genome rearrangement study can be modeled as a combinatorial problem of finding a series of large-scale mutations to transform one genome into another. Therefore, the main purposes of this proposal are to study and solve the genome rearrangement problem under reversals, and block-interchanges. In respect of computational theory, we have made use of breakpoint graphs to design efficient algorithm for solving the genome rearrangement problem under reversals and block-interchanges. In respect of practical applications, on the other hand, we have implemented our algorithm with incorporation of some existing programs to build a web server related to genome rearrangements that can help biologists measure the evolutionary differences among several organisms by comparing the differences of their gene-order data. To enhance the system's ability, we have also equipped it with the function of automatically identifying the homologous or conserved regions that are shared by all the input sequence data so that it can take sequence data as input. In addition, the system has further been integrated with the NJ (Neighbor Joining) and UPGMA (Un-weighted Pair-Group Method using Arithmetic mean) programs for reconstructing the evolutionary trees according to the computed distances for all pairs of input organisms.

Keywords: Bioinformatics; Computational Biology; Genome Rearrangements; Reversal Mutations; Block-Interchanges Mutations

二、計劃報告內容

In the late 1980s, J. Palmer and Herbon [PH88] discovered a novel kind of evolutionary changes, genome rearrangements, in plant organelles by comparing the mitochondrial genomes of *Brassica oleracea* (cabbage) and *Brassica campestris* (turnip). They found that these molecules are almost identical in gene sequences, but differ dramatically in gene order. With the parsimony (shortest) scenario of three reversal rearrangements (also known as inversions), the cabbage gene order can be transformed into the turnip gene order, where a reversal affects on a contiguous interval genes by inverting the gene order and signs. Since then, such genome rearrangement studies of comparing the gene orders have attracted a lot of attentions, because they can measure the evolutionary difference between two organisms by conducting the large-scale comparisons of their genomic data, especially of highly similar genomes. For highly similar genomes, their difference in sequence level is too small to be identified via the classical sequence comparison because the classical sequence comparison is based on only point mutations that act on individual characters.

In addition to reversals, the other large-scale mutations that have been considered in the study of genome rearrangement are transpositions, block-interchanges, fissions, fusions and translocations. The transpositions affect the chromosomes by moving a segment from one site to another in one chromosome. The block-interchanges affect the chromosomes by swapping two non-intersecting (but not necessarily adjacent) segments of any length on a chromosome [LLCT05]. The translocations affect the chromosomes by exchanging two segments between two different chromosomes. The fissions are a special case of translocations that affect on the chromosomes by splitting a chromosome into two chromosomes. The fusions are another special case of translocations that affect on the chromosomes by joining two chromosomes into one. The study of genome rearrangements can be modeled as a combinatorial problem of finding a series of large-scale mutations to transform one genome into another, where the considered genomes are usually denoted by a permutation of ordered (signed or unsigned) integers with each integer representing a identical gene in genomes and its sign (+ or -) indicating the transcriptional orientation.

In this proposal, we have studied the genome rearrangement problem under different large-scale mutations, especially such as reversals and block-interchanges. In respect of computational theory, we have applied the properties of breakpoint graphs for solving the genome rearrangement problem by considering both reversals and block-interchanges. If both reversals and block-interchanges are considered together, we adopted a strategy of unequal weight by using weight 1 for reversals and weight 2 for block-interchanges. This is mainly due to the following reasons. First, reversals have been favored as more frequent rearrangement operations when compared with block-interchanges. Second, a reversal affecting the chromosome removes at the most two breakpoints, whereas a block-interchange removes at the most four. Third, the rearrangement distance involving both reversals and block-interchanges can currently be computed in polynomial time only when the weight of reversals is 1 and the weight of block-interchanges is 2 (please see our paper [LLLT06] for details). In respect of practical applications, on the other hand, we have implemented the algorithm we have developed in this project and integrated them with some existing programs to build a web server related to genome rearrangements that can help biologists measure the evolutionary differences among several organisms by comparing the differences of their gene-order data. To enhance the system's ability, we have also equipped it with the function of automatically identifying the homologous or conserved regions that are shared by all the input sequence data so that it can take sequence data as input. In addition, the system has further been integrated with the NJ (Neighbor Joining) and UPGMA (Un-weighted Pair-Group Method using Arithmetic mean) programs for reconstructing the evolutionary trees according to the computed distances for all pairs of input organisms.

三、計畫成果自評

In this project, we have successfully developed a web server, called SPRING (http://algorithm.cs.nthu.edu.tw/tools/SPRING/), which is a tool for the analysis of genome rearrangement between two chromosomal genomes using reversals and/or block-interchanges. SPRING takes two or more chromosomes as its input and then computes a minimum series of reversals and/or block-interchanges between any two input chromosomes for transforming one chromosome into another. The input of SPRING can be either bacterial-size sequences or gene/landmark orders. If the input is a set of chromosomal sequences then the SPRING will automatically search for identical landmarks, which are homologous/conserved regions shared by all input sequences. In particular, SPRING also computes the breakpoint distance between any pair of two chromosomes, which can be used to compare with the rearrangement distance to confirm whether they are correlated or not. In addition, SPRING shows phylogenetic trees that are reconstructed based on the rearrangement and breakpoint distance matrixes. The research result of our SPRING (for details, please refer to our paper [LLLT06]) has been published in the journal of Nucleic Acids Research whose SCI impact factor is now 7.552.

四、參考文獻

[LLCT05] Lin, Y.C., Lu, C.L., Chang, H.-Y. and Tang, C.Y. (2005) An efficient algorithm for sorting by block-interchanges and its application to the evolution of vibrio species. *Journal of Computational Biology*, Vol. **12**, pp. 102-112. (*: corresponding author)

[LLLT06] Lin, Y.C., Lu, C.L.*, Liu, Y.-C. and Tang, C.Y. (2006), SPRING: A tool for the analysis of genome rearrangement using reversals and block-interchanges, *Nucleic Acids Research*, Vol. 34, pp. W696-W699. (*: corresponding author)

[PH88] Palmer J.D. and Herbon L.A. (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J. Mol Evol.*, Vol. 28, pp. 87-97.