氏　　名	Noureen, Mehwish
学位(専攻分野)	博士(理学)
学 位 記 番 号	総研大甲第 2278 号
学位授与の日付	2021年9月 28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	An Algorithmic Approach for Identifying Rearrangements in Multiple Bacterial Genomes
論 文 審 査 委 員	主　査　黒川　顕 遺伝学専攻　教授 仁木　宏典 遺伝学専攻　教授 宮城島　進也 遺伝学専攻　教授 池尾　一穂 遺伝学専攻　准教授 内山　郁夫 遺伝学専攻　准教授

(Form 3)

Summary of Doctoral Thesis

Name in full Noureen, Mehwish

Title An Algorithmic Approach for Identifying Rearrangements in Multiple Bacterial Genomes

Each species has a unique genome structure that changes slowly with time. Genome of each species is subjected to both the local and global mutations during the evolution. Local mutations affect the genomes at a smaller scale and are more frequently occurring. On the other hand, genome rearrangements affect the large segments of the genome and occur less frequently. Increasing number of prokaryotic genomes and their comparison have revealed the presence of large number of genomic rearrangements. The dynamic nature of bacterial genome is the result of rearrangements, horizontal gene transfer and activity of the mobile genetic elements. Genome rearrangements not only change the orientation but also the order of the genes on the chromosomes. As genome rearrangements are rarer compared to the point mutations, therefore they can reveal the important events that occurred during the course of evolution.

Several approaches have been proposed to identify the genome rearrangements, however most of these approaches use only the pairwise comparison and consider the similar set of genes. I have developed an algorithmic approach to identify the genome rearrangements in multiple bacterial genomes considering highly conserved genes in a given set of genomes. Orthologous gene clusters were used to identify the gene order in each genome which was used as an input to identify the genome rearrangements. The obvious benefit of my approach is scalability: whole genome comparison is difficult for many genomes using previous approaches comparing two genomes. My method can handle hundreds of strains at the level of gene orders.

I have used *Helicobacter pylori* strains to demonstrate the use of my algorithm, as this bacterium has a very diverse genomic structure. Using my algorithm, the geographically region-specific rearrangements and those shared across continents were identified for 72 *H. pylori* strains in the public repository. Region specific breakpoints were overrepresented in Asia and Australia whereas all breakpoints were detected in Europe. Total 41 inversions were identified, 23 were shared whereas 18 were strain specific. Strain from Europe and East Asia shared as many as 11 inversions. Some inversions occurred more frequently and were found in strains from all geographical locations except for Africa. Two specific inversions were associated with disease states such as cancer. Three genomic loci were frequently involved in rearrangements (*rearrangement hotspots*)

in the analyzed strains. The pattern of inversions was most diverse in Japan probably because of the larger number of sampling. The North American region also had the diverse inversion pattern even though the number of samples was much smaller compared to Japan. This diversity occurred maybe because of human migration. Many inversions in *H. pylori* strains were shared across geographic regions, and only few were found to be geographically region-specific.

To identify the cause of rearrangements, the association of repeats, insertion sequences (IS) and genomic islands (GIs) were investigated. The correlation between the number of repeats and inversions was weak, suggesting that not only the occurrence of repeats but their relative position is also important for the homologous recombination. Shared inversions tend to possess more inverted repeats compared to the strain specific inversions. Beside this, world-wide inversion breakpoints had more IS elements compared to others. Moreover, GIs were mostly associated with region-specific and strain-specific breakpoints. Most of the shared inversions breakpoints possessed the similar genomic elements with a few exceptions. This suggests that these elements are well conserved irrespective of the different geographical region.

As some of the genome rearrangements were associated with the disease state, I performed the analysis on a larger dataset, 123 *Helicobacter pylori* genomes to find the association of the genomic features more specifically the genome rearrangements with the disease outcome. Comparative analysis of the strains revealed the presence of certain group-specific genes. Most of the identified inversions were shared and few were associated with the disease state. Weak association between the genomic features and disease state might be because of the fact that the disease outcome depends on several other factors such as environment, diet and host. Besides this, several genomes with no disease state information also makes it difficult to draw some conclusions.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　Noureen, Mehwish

Title
論文題目　An Algorithmic Approach for Identifying Rearrangements in Multiple Bacterial Genomes

ゲノム解析技術の発達により細菌のゲノム全配列を容易に解読できるようになった。DDBJ に代表される公共データベース等には、すでに数十万株の細菌のゲノム全配列が登録されている。これら細菌は、点突然変異だけにとどまらず、遺伝子の水平伝播や遺伝子重複など多様な方法で進化を繰り返している。中でも薬剤耐性やホストの免疫機構からの回避等の迅速な環境への適応は、点突然変異というよりはむしろ、遺伝子の水平伝播や相同組換えなど、ゲノムの大規模な再編成による新機能や新規遺伝子の急激な獲得により達成されていると考えられる。これら細菌における進化の履歴は、現存のゲノム配列そのものに蓄積されているため、ゲノム全配列を比較解析することで進化のダイナミクスを理解することが可能である。ゲノム再編成を同定するためにいくつかの解析手法が提案されているが、これらの解析手法の多くは一対一でゲノムを比較し、相同な遺伝子セットを抽出し、各株におけるゲノム再編成を定義するものである。

Noureen さんは、複数の細菌ゲノムにおいて保存性の高い遺伝子を同定し、ゲノム再編成の履歴を明らかにするアルゴリズムを開発した。一対一でゲノム全体を比較する従来の方法では、多数のゲノム配列を入力とした解析が困難であるため、ゲノム再編成の履歴を明らかにする事は容易ではなかった。本手法では、保存遺伝子の基準株における並び順と各株における並び順を比較し、再編成のイベントを最節約的に決定するため、多数の株におけるゲノム再編成の履歴を一度に解析する事を可能とした。この成果は Noureen さんが筆頭著者である論文として *BMC Bioinformatics* 誌に掲載されている。

続いて Noureen さんは、自ら開発したゲノム再編成の履歴を明らかにするアルゴリズムの有効性を示すために、種内において多様なゲノム構造を有している *Helicobacter pylori* 菌 72 株を対象とした解析を実施した。その結果、多様な地域から取得した 72 株の *H. pylori* のゲノム再編成のパターンを明らかにし、地域特異的なゲノム再編成、大陸間で共有されるゲノム再編成や、がんなどの疾患と関連する特異的な逆位などを同定した。*H. pylori* は多様なゲノム構造を有する一方で、多くの逆位は共通しており、地域特異的なものはごく僅かであることを示した。この成果は Noureen さんが筆頭著者である論文として *Microorganisms* 誌に掲載されている。さらに、がんなどの疾患とゲノム再編成との関係性を明らかにするため、123 株の *H. pylori* ゲノムを対象に同様の解析を行った。特定された逆位のほとんどは共通しており、胃炎など一部の疾患を除き、疾患と逆位との関連性を見出すことはできなかった。

Noureen さんは、疾患と細菌のゲノム進化との関連性を見出すためには、ゲノム再編成の履歴を明らかにする事が重要であると考え、それらを実現するための新規アルゴリズム

を開発し新たな比較ゲノム解析を実現させた。より大規模なデータによる解析を実現するために、国際コンソーシアムにも参加しており、ゲノム情報や疾患情報のみならず、今後は宿主の食事や住環境の情報など、多様なデータを統合した解析が可能となる。これらの成果は将来のゲノム科学の発展に寄与する学術上の優れた研究である。以上の理由により審査委員会は、本論文が学位の授与に値すると判断した。