# 論文の内容の要旨

論文題目　Comparative Analysis of Genomic Landscapes

（ゲノムランドスケープの比較解析）

氏　　名　　芦田　広樹

**Motivation**

To date, the number of sequenced genomes of non-human species is more than 3,800 and the cost per nucleotide to sequence DNA has dropped 100,000-fold between 1999 and 2009. The growth rate of our sequencing capabilities has far exceeded that of Moore's Law. The drastic acceleration in biological enquiry enabled by the current high-throughput technology is just beginning. At the current rate of technological progress, DNA sequencing is soon likely to become a commodity for all the studies in biology.

However, what is rapidly growing in even faster rate is production of comprehensive catalogues of genetic features that are mapped on to the primary sequence. For instance, the number of genome annotation tracks in the UCSC genome browser has increased exponentially over the past few years and now tracks for human genome hg19 alone exceeds 150, with around 1000 data tables. The types of data include histone modifications, SNPs, structural variation sites, CpG methylation, splicing sites, non-coding RNA and many more. The next important step is to determine how these genomic landscapes are associated with each other, both globally and locally, and to start piecing together the puzzle in order to grasp the whole picture of the genome system. Our goal in this thesis is to develop a method for comparing genomic landscapes according to their shapes and extracting regions that show high correlations.

**Method**

Although new data continue to arrive at a prodigious rate and thorough investigation of each measurement is done individually, not much work has been done to provide an overview and bring together the different views of the landscapes. The general idea of our approach is to align genomic landscape data (collections of real-valued observations made at sequential positions along a chromosome) based on their topology. This will allow us to detect regions with similar shapes, which can lead to finding functionally interrelated regions. We overcame the size problem for genome-wide data by converting the data into series of symbols and then carrying out sequence alignment. We also decomposed the oscillation of the landscape data into different frequency bands before analysis, since the real genomic landscape is a mixture of embedded and confounded biological processes working at different scales of the cell nucleus. Our approach has five phases: (i) Wavelet transformation, (ii) Dimensionality (data) reduction, (iii) Symbolic representation, (iv) Local alignment and (v) Filtering. The dimensionality reduction feature of our process makes approximating large datasets like genomic landscape feasible.

**Result**

To verify the usefulness and generality of our method, we applied our approach to well investigated landscapes from the human genome, including several histone modifications. Furthermore, by applying our method, we made the novel biological finding that DNA replication timing and the density of Alu insertion are highly correlated genome-wide.

**Conclusion**

We have developed an ultra fast method for comparing the genome-wide data of genomic landscapes. To our knowledge, this is the first method to align the landscapes according to their topology at multiple resolutions. Our approach is robust to position distortion and copes with the high dimensionality of genomic data. We have processed vast numbers of human genomic landscape data in order to find links between previously untested factors. The information discovered through our approach should facilitate further exploration of genomic landscapes and how they affect each other within a living cell nucleus.