様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２５　日現在

機関番号：　８２６２６
研究種目：　若手研究(A)
研究期間：　2014～2017
課題番号：　２６７０００３０

研究課題名（和文）Regulatory DNA conserved between Phyla

研究課題名（英文）Regulatory DNA conserved between Phyla

研究代表者

Ｆｒｉｔｈ　Ｍａｒｔｉｎ（Frith, Martin）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・上級主任研究員

研究者番号：４０４６２８３２

交付決定額（研究期間全体）：（直接経費）　　14,500,000円

研究成果の概要（和文）：ヒトDNAの配列再編成の調査は、NARで発表されました。簡略化された”ﾁｬｲﾙﾄﾞﾃｰﾌﾞﾙ”ｱﾙｺﾞﾘｽﾞﾑは、類似した配列を見つけるために、IEEE/ACM TCBBで発表されました。ｹﾞﾉﾑの変異を見つける方法は、BMC Medical Genomicsで発表されました。DNA隣接ﾍﾞｰｽ（細菌の系列差別を改善する）我々の方法は、BMC Bioinformaticsで発表されました。我々の分析法は、最近の長いDNAの読み取りに有用であることが証明されています。

研究成果の概要（英文）：This year, several direct and indirect research outcomes were finalized and published. A survey of sequence rearrangements in human DNA was published in NAR. A simplified "child table" algorithm, for finding similar sequences, was published in IEEE/ACM TCBB. A method for detecting genome variations was published in BMC Medical Genomics. Our method for considering DNA neighbor-base preferences, which improves bacterial strain discrimination, was published in BMC Bioinformatics. Our analysis methods are proving useful for recent "long" DNA-read data.

研究分野：　computational biology

キーワード：　genome

１．研究開始当初の背景

As this project began, more and more genomes of complex multicellular organisms were being sequenced and published. However, our understanding of the genetic information encoded therein was very limited, especially outside of protein-coding regions. It was well-established that deep evolutionary conservation indicates sequence with important biological function. Thus, a basic and interesting approach is to compare different genomes, to find conserved non-protein coding regions with no known function. Such regions may be: regulatory DNA (e.g. promoters, enhancers, silencers, insulators) that control the transcription of genes; non-protein-coding RNA genes such as micro-RNAs, or perhaps conserved domains in long noncoding RNAs; or even "unknown unknowns".

Although this approach was already well-known, a thorough analysis requires best-possible methods for comparing and aligning sequences. Specifically:

(1) We should determine the rates of insertion, deletion, and substitutions in the sequence data, and use them to find alignments in a statistically-optimal manner.

(2) We should evaluate the statistical significance (p-value) of any alignment. That is: we should evaluate the chance of such an alignment occurring between randomly shuffled sequences.

(3) We should thoroughly detect regions that encode proteins, or are descended from protein-coding sequence (i.e. pseudogenes): the latter requires frameshift-aware sequence comparison. Previous work in this field rarely used these procedures, and never all of them. On the other hand, statistically-tuned alignment criteria, and significance calculations, had been well developed for protein sequence analysis. It was a surprising scientific gap that analogous procedures were much less developed for analyzing nucleotide sequences. Previous work on comparing and aligning genomes was often surprisingly ad hoc, for example: reporting alignments that meet certain criteria, without considering the probability of randomly-shuffled sequences possessing such alignments by pure chance. Therefore, these criteria differed wildly in their strictness: sometimes allowing weak, chance, random alignments, and sometimes forbidding subtle alignments that are distinguishable from chance. Thus, there was much room to optimize comparison and alignment of genomic DNA sequences.

２．研究の目的

The initial aim was to optimize methods for comparing genetic sequences, with the long term goal of finding regions with deep evolutionary conservation but no previously-known function. However, optimizing these methods was also likely to bring other spin-off benefits, because fundamentally-similar sequence comparison is used to:

(1) Find chimeric and horizontally transferred DNA, such as NUMTs (insertions of mitochondrial DNA into a nuclear genome), or pathogenicity islands in microbial genomes.

(2) Compare long DNA reads to genomes, such as PacBio and (more recently) nanopore DNA reads, which have high error rates, thus requiring sensitive and statistically-optimized sequence comparison methods. Note that there is merely a quantitative, and not a qualitative, difference between a long DNA read and a chromosome sequence.

(3) Etc.

Thus, a more general aim was to know: where do the parts of a sequence come from, and how are the parts of sequences related to each other.

３．研究の方法

The research proceeded in several steps.

First, we developed a method to find statistically-significant DNA-to-protein alignments, allowing frameshifts, on a whole genome/proteome scale (S Sheetlin et al. Bioinformatics 2014, S Sheetlin et al. Bioinformatics 2016). Interestingly, we showed that many deeply-conserved non-protein coding regions of vertebrate genomes are actually recent pseudogenes: that is, the evolutionary ancestors of these sequences were conserved because they encoded proteins, but the present-day pseudogenic sequences may no longer be subject to conservation.

Then, we developed a method to find a most-probable division of two genomes into parts along with the most probable one-to-one alignments of these parts (M Frith & R Kawaguchi 2015). This is a new approach to finding and aligning orthologous regions between two genomes, and the first one that calculates alignment probabilities.

Next, we developed a useful software tool to determine the rates of insertion, deletion, and each kind of substitution between two sets of sequences, e.g. two genomes (M Hamada et al. Bioinformatics 2017). This turns out to be very broadly useful. It has been used to:
(1)
Compare whole genome sequences.

(2) Align various kinds of chemically-modified DNA reads to genomes, such as DNA reads from these types of experiment: CLIP-seq, PAR-CLIP, TimeLapse-seq, SLAM-seq, DMS-MaPseq.

(3) Compare DNA sequences with unusual at:gc composition, such as DNA from Plasmodium falciparum (malaria) which is about 80:20 at:gc.

These methods turned out to be extremely useful for comparing long, error-prone DNA reads (nanopore and PacBio) to a genome, especially for finding rearrangements and duplications (M Frith & S Khan 2018). Thus, our methods have been used by several research groups for analyzing long DNA reads.

４．研究成果

This year, several direct and indirect research outcomes were finalized and published.
A survey of sequence rearrangements in human DNA was published in NAR. A simplified " child table" algorithm, for finding similar sequences, was published in IEEE/ACM TCBB.
A method for detecting genome variations was published in BMC Medical Genomics.
Our method for considering DNA neighbor-base preferences, which improves bacterial strain discrimination, was published in BMC Bioinformatics. Our analysis methods are proving useful for recent "long" DNA-read data.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計４件）

〔学会発表〕（計　　件）

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：

番号：
取得年月日：
国内外の別：

〔その他〕

ホームページ等
https://sites.google.com/site/frithbioinfo/

６．研究組織
(1)研究代表者
フリス　マーティン
（Frith Martin）
国立研究開発法人
産業技術総合研究所・
情報・人間工学領域・
上級主任研究員
研究者番号：40462832

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）