様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １６ 日現在

機関番号：６２６０３
研究種目：基盤研究(B)
研究期間：2012～2014
課題番号：２４３００１０９

研究課題名（和文）確率分割による個人ゲノム漏洩におけるリスク評価と秘匿の方法の確立

研究課題名（英文）Establishment of methodology for assessment of disclosure risk and confidentiality of individual genome by random partition

研究代表者
間野　修平（Mano, Shuhei）

統計数理研究所・大学共同利用機関等の部局等・准教授

研究者番号：２０３７２９４８

交付決定額（研究期間全体）：（直接経費）　　10,400,000円

研究成果の概要（和文）：個人ゲノムが漏洩し，保有者が特定されると，個人情報が漏洩します．統計科学において議論されてきた個票開示リスクと相似の問題で，標本一意の秘匿済み個人ゲノムが母集団一意になる確率を抑える方法が必要です．基礎となる統計モデルは確率分割です．本研究では，これらのリスク評価と秘匿の方法の確立を目的としました．統計モデルの検討として，ゲノム上の相関を考慮したモデルの数値計算の手法を提案し，交換可能な確率分割の一般的なクラスであるGibbs分割について，漸近的な性質を調べました．また，個人ゲノムの取得と利用に関わる方々との意見交換を行いながら，漏洩リスクの評価と秘匿の方法を検討しました．

研究成果の概要（英文）：If an individual genome is disclosed and the holder is identified, the person's sensitive information is disclosed. This problem is similar to the statistical disclosure control problem. We need methods to suppress the probability that a sample unique genome with some concealment becomes population unique. The fundamental statistical model is a random partition. In this research, the purpose was establishment of methodology for assessment of disclosure risk and confidentiality of individual genome. A computational method for a random partition with accounting correlation among points on a genome is proposed, and asymptotic properties of Gibbs-type partitions, which are in a general class of random partitions, were discussed. The methods for assessment of disclosure risk and confidentiality were discussed with people who obtain individual genome data and use them.

研究分野：統計科学

キーワード：統計数学　ゲノム

１．研究開始当初の背景

ゲノムを決定するために必要なコストは，時間的，金銭的に，急速に下がっています．各患者に重篤な副作用を避けながら奏功が確実な薬を投与すること，生活習慣病を発症する可能性の高い人の生活に介入し予防することなどが可能になりますので，その有用性から，将来的に多くの個人ゲノムが決定，保持されると考えられます．しかし，個人ゲノムの決定は良いことばかりではありません．ゲノムは，疾患へのかかりやすさ，民族的背景など，知られることを望まない個人情報を大量にもちますので，保有者が特定されると，それらの個人情報はすべて漏洩することになります．このような個人ゲノム漏洩のリスクは認識されていましたが，評価法がないまま事態が推移していました．

本研究では，上記の問題が統計科学において議論されてきた個票開示リスクと相似であることに着目しました（個票開示リスクについては竹村彰通編，「統計数理」51 巻 2 号,2003）．開示した調査票の提供者が特定され，提供者が知られることを望まない回答が漏洩することです．個票開示では，標本に 1 名（標本一意）の回答の組み合わせが母集団でも 1 名（母集団一意）になる確率が基本的なリスク評価の指標ですが，個人ゲノムの場合，一卵性双生児を除けば個人ゲノムは常に母集団一意ですので，個人ゲノムの情報をどの程度秘匿処理すれば，標本一意の秘匿済み個人ゲノムが母集団一意になる確率を抑えられるかが問題になります．

２．研究の目的

(1)　統計モデルの検討

個票開示リスクの基礎になる統計モデルは確率分割です．確率分割とは要素の分割に確率を与えたもので，あるサイズの標本を分類するとき，サイズごとの度数の分布です．調査表の回答の組み合わせは調査対象者の人数のサイズの分割になります．同様に，個人ゲノムを並べると，染色体の総数のサイズの分割が現れます．このことを，公開されている，米国に居住するヨーロッパ由来の住民について調べられた 120 本の染色体における酪酸分解酵素の遺伝子の DNA 配列のデータ（http://hapmap.ncbi.nlm.nih.gov）を用いて示します．下の図は，5 万塩基対ほどのうち，個人差のある塩基サイトだけを示していて，右の数字は各タイプの度数です．この例は，サイズ 94,8,7 が 1 つ，サイズ 2 が 3 つ，サイズ 1 が 5 つという分割です．

```
GGCGACATTCCGCTTCAGGCATTCCTATCTAAACAGACC 94
CACAGCGTGTTACCCGGAGCATGCTTAATCGGATCGGCT 8
GATGGTACTTCGTCCCAGAGTAGCTTCATTAGACAGATC 7
GGCGGCACTCCGCCCGGAGCATGATTAATTAGGCACACC 2
GGCGACACTTCGTCCCAGAGTAGCTTCATTAGACAGATC 2
GGCGACATTCCGCTTCAGGCATTCCTATTTAAACAGACC 2
GGCGGCACTCTATCCGGAGCATGCTCAATCGGATCGGCT 1
CACAGCGTGTTACCCGGAGCATGCTTAATCAGATCGGCT 1
CACAGCGTGTCGCTTCAGGCATTCCTATCTAAACAGACC 1
GGCGACACTCCGCCCGGAGCATGATTAATTAGGCACACC 1
CACAGCGTGTTACCCGGAGCATGCTCAATCGGATCGGCT 1
```

代表者は，共同研究者と，様々な歴史を想定したシミュレーションを行い，個人ゲノムから現れる分割の統計モデルとして Ewens 分割が適切であることを示していました．Ewens 分割は 1972 年に遺伝子の多様性のモデルとして提案されましたが，その当時は DNA 配列のデータは想定されておらず，遺伝子を染色体の上の点として扱っています．配偶子が作られるときに組み換えとよばれる現象がおきますので，DNA 配列を点として扱うことはできません．しかし，代表者らは，DNA 配列のデータについても，組み換えの効果を母数の尺度の変換とみなすことで，実効的に，Ewens 分割を適用できることを示しました．

本研究の目的の一つは，個人ゲノムから現れる分割の統計モデルをさらに検討することにありました．

(2) 漏洩リスクの評価と秘匿の方法の検討

母数は領域の長さに比例しますが，ゲノム全体（30 億塩基対）では十分に大きいため，すべての染色体が母集団一意です．特定の遺伝子を考えれば母数は小さくなり，母集団一意の染色体の数の期待値が減少します．この数に標本抽出率をかけたものが，標本一意かつ母集団一意の染色体の数の推定値で，リスク評価の指標になります．

ただし，ゲノムの多様性は，民族の歴史により異なるだけではなく，領域についても一様ではありません．例えば，移植の適合性判定に使われる組織適合性複合体は，生物学的

な自己認識に関わりますので，リスクが高いと想像されます．

本研究のもう一つの目的は，公開されている日本人の個人ゲノムのデータを用いて，個人ゲノム漏洩のリスクを評価し，秘匿の方法を確立することにありました．

３．研究の方法

(1)　統計モデルの検討

① ゲノム上の相関の検討

実用上，個人ゲノムから現れる分割の統計モデルとして Ewens 分割を用いることはできるのですが，組み換え，すなわち，ゲノム上の点の間の相関を陽に考慮したモデルとの比較も重要です．そこで，ゲノム上の相関を考慮したモデルの扱いについて検討しました．

② 一般化の検討

Ewens 分割は，確率分割の原点であり，ノンパラメトリック・ベイズにおいて事前分布として用いられる Dirichlet 過程からの標本であること，Poisson-Dirichlet 分布とよばれる測度値拡散過程の定常分布からの抽出公式として導かれることなど，確率論的に興味深い様々な性質をもっています．しかし，統計モデルとしての役割を考えると，母数は 1 つですし，十分な表現力のあるモデルとは言えません．より表現力のある統計モデルを求めて確率分割の一般化を検討することも重要です．

Ewens 分割に 1 つの母数を追加し，Ewens 分割を生成する中華料理店過程とよばれる推移の規則を拡張したものに，Pitman 分割があります．また，交換可能性（分割の各要素を置換しても不変）は自然な要請ですが，この 10 年間ほど，交換可能な確率分割の考察が盛んになっていますが，交換可能な確率分割のクラスに Gibbs 分割とよばれるものがあります．Gibbs 分割は各要素のサイズで定まる重みを要素の数について混合したものですが，ある条件の下で Pitman 分割に帰着します．Gibbs 分割は，これまでに検討された交換可能な確率分割のクラスとしておそらく最も一般的なものです．このような理由から，Gibbs 分割の性質を検討しました．

(2) 漏洩リスクの評価と秘匿の方法の検討

ゲノムの各領域について，確率分割の母数を最尤推定することで，リスク評価の指標がえられます．例えば，ゲノムの各領域について，標本サイズを与えると，母集団一意になる確率を求めることができます．この計算を計算機のソフトウェアに実装しました．

将来的に，大量の個人ゲノムが決定，保持されると考えられますが，ゲノムの多くの部分に有用性はありませんので，有用性に直結する最小限の情報のみを残し，他の情報を消去することで，個人情報を秘匿することになると予想されます．秘匿の方法は個人ゲノムの利用目的に密接に関わります．例えば，薬剤副作用を防ぐために秘匿処理済みの個人ゲノムを保持する場合は，重要な薬剤の副作用に関連するサイトは消去できません．そこで，個人ゲノムの取得と利用に関わる方々と意見交換を行い，利用目的をいくつか想定して，秘匿の方法を検討することにしました．

４．研究成果

(1)　統計モデルの検討

① ゲノム上の相関の検討

成果を雑誌論文⑥と学会発表⑫に公表しました．確率分割を，ある拡散過程の定常分布からの抽出公式として導かれること，すなわち，拡散過程が定める分布のモーメントであることに着目して，ゲノム上の相関を表現した拡散過程の双対な確率過程（状態空間が格子点）を考えました．一般に，モーメントの閉じた表示は得られないことを指摘し，数値計算の手法を提案しました．この手法では，マルコフ連鎖モンテカルロ法を用いますが，拡散過程ではなく，シミュレートの容易な双対な過程をシミュレートするところに工夫があります．

② 一般化の検討

成果を，Pitman 分割については雑誌論文⑤と学会発表⑧，⑨，⑭，⑮，⑯に，Gibbs 分割については雑誌論文①と学会発表③に公表しました．実用上，特に問題になるのは，確率分割は組み合わせ論と密接に関係しますので，標本のサイズが大きいと，巨大な数の計算が避けられないことです．そこで，サイズが大きいときの漸近的な性質に注目しました．特に，Ewens 分割における要素のサイズの極値の漸近分布は整数論と関係しているなど，古くからの興味の対象になっています（例えば Billingsley, Convergence of Probability Measures, 1999）．そこで，

Pitman 分割と，その一般化である Gibbs 分割について，解析的組み合わせ論の技術を用いて，要素のサイズの極値の漸近的な挙動を導きました．

(2) 漏洩リスクの評価と秘匿の方法の検討

公表された成果に，図書①（総説）と，Ewens 分割を用いて血縁者の間での DNA 配列の違いを評価した雑誌論文④があります．公開されている個人ゲノムのデータを用いた個人ゲノム漏洩のリスクの評価と秘匿の方法の検討に関しては，公表された成果はありませんが，公表次第，研究代表者のホームページに掲載します．

個人ゲノムの取得と利用に関わる方々との意見交換は以下の通りです．2013 年 12 月 20 日には，研究代表者の所属機関で研究集会「ゲノム多様性データの利用」を開催しました．2014 年 1 月には，ライフサイエンスデータベースの構築者と利用者向けの技術勉強会「バイオハッカソン」において日本人の個人ゲノム公開のルール策定を担当するグループと，2015 年 3 月には，東北メディカル・メガバンク機構の方々（学会発表①）と，意見交換を行いました．

研究分担者の角田達彦氏の成果は，大規模ゲノムデータ解析の観点からのもので，雑誌論文③，学会発表②，④，⑤，⑦，⑩，⑪，⑬があります．研究分担者の太田博樹氏の成果は，少数民族由来のゲノムデータの保護の観点からのもので，雑誌論文②，学会発表⑥，図書①があります．

５．主な発表論文等

〔雑誌論文〕（計 6 件）

① Mano S., Extreme sizes in the Gibbs-type exchangeable random partitions. Annals of the Institute of Statistical Mathematics, 査読有, 掲載確定.

② Takezawa Y., Kato K, Oota H. et al. (20 authors), Human genetic research, race, ethnicity and the labeling of populations: recommendations based on an interdisciplinary workshop in Japan. BMC Medical Ethics, 査読有, Vol.15, 2014, 33. doi: 10.1186/1472-6939-15-33.

③ 角田達彦，がん研究におけるゲノムビッグデータ解析と臨床応用，実験医学，査読無，Vol.32, 2014, 2046-2051.

④ Nishino J., Mano S., The number of candidate variants in exome sequencing for Mendelian disease under no genetic heterogeneity. Computational and Mathematical Methods for Medicine,査読有, Vol.2013, 2013, 179761. 13pages. doi: 10.1155/2013/179761.

⑤ Mano S., Asymptotics of Pitman random partition via combinatorics. ISM research memorandom 1177, 査読無, 2013, 1-30.

⑥ Mano S., Duality between the two-locus Wright-Fisher diffusion model and the ancestral process with recombination. Journal of Applied Probability, 査読有, Vol.50, 2013, 256-271. doi: 10.1239/jap/1363784437.

〔学会発表〕（計 16 件）

① 間野修平，Approximate Bayesian computation and its applications. インシリコ・メガバンク研究会，2015 年 3 月 13 日，東北大学.

② Tsunoda T., Medical Science Mathematics for Cancer Genome Analysis. The 73rd Annual Meeting of the Japanese cancer Association, September 26, 2014, Yokohama.

③ Mano S., Ordered sizes in the Gibbs-type exchangeable random partitions, 37th Conference on Stochastic Processes and their Applications. July 28-August 1, 2014, Buenos Aires (Argentina).

④ 角田達彦，全ゲノムビッグデータ解析によるオーダーメイド医療，富士通知創薬コンソーシアム，2014 年 4 月 7 日，東京.

⑤ Tsunoda T., Whole genome big data analysis for complex diseases. 3rd Global COE Workshop between BGI and University of Tokyo - Advances in Medical Genomics, Tokyo, 2014, March 20.

⑥ 松前ひろみ，間野修平，太田博樹，人類学の立場からパーソナルゲノムプロジェクトを考える，日本分子生物学会，2013 年 12 月 6 日，ポートアイランド（神戸市）.

⑦ 角田 達彦，全ゲノムとエクソームシークエンス解析．日本人類遺伝学会第 58 回大会，2013 年 11 月 23 日，江陽グランドホテル（仙台市）.

⑧ 間野修平，ピットマン分割の漸近論への解析的組み合わせ論による接近，統計関連学会連合大会，2013 年 9 月 10 日，大阪大学.

⑨ Mano S., Extremes of Pitman's random partition and their asymptotics, Asymptotic Statistics and Related Topics: Theories and Methodologies, September 4, 2013, University of Tokyo.

⑩ Tsunoda T., Whole genome approach is revolutionizing medicine. The 10th International Workshop on Advanced Genomics, May 21, 2013, Tokyo.

⑪ Tsunoda T., Whole genome sequencing and comprehensive mutation analysis of liver cancer. SNUCRI & SNUCH Cancer Symposium, May 3, 2013, Korea.

⑫ 間野修平，集団遺伝の確率モデルと統計的推測，日本遺伝学会大会，2012 年 10 月 26 日，九州大学.

⑬ Tsunoda T., Genomic medicine's milestones and future. International Conference on Bioinformatics 2012, October 2012, Thailand.

⑭ 間野修平，ピットマン分割における極値とその漸近的性質．研究集会「官庁統計データの公開における諸問題の研究と他分野への応用」，2012 年 9 月 24 日，統計数理研究所（東京都）.

⑮ 間野修平，ピットマン分割における極値とその漸近的性質，統計関連学会連合大会，2012 年 9 月 11 日，北海道大学.

⑯ Mano S., Extremes of random partition and the asymptotics, 研究集会 "Stochastic models and computational algorithms". 2012 年 7 月 19 日, 統計数理研究所（東京都）.

〔図書〕（計 1 件）

① 松前ひろみ，間野修平，太田博樹，ビッグデータ利用の最前線「個人ゲノムデータの利用と倫理的課題」（分担執筆），2014，7.

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

ホームページ等
http://www.ism.ac.jp/~smano

６．研究組織

(1)研究代表者
間野　修平（MANO, Shuhei）
統計数理研究所・数理・推論研究系・准教授
研究者番号：２０３７２９４８

(2)研究分担者
角田　達彦（TSUNODA, Tatsuhiko）
理化学研究所・統合生命医科学研究センター・医科学数理研究グループリーダー
研究者番号：１０２７３４６８

(3)連携研究者
太田　博樹（OOTA, Hiroki）
北里大学・医学部・准教授
研究者番号：４０４０１２２８