個人ゲノムデータの利用とプライバシー保護

筑波大学 システム情報系/理化学研究所 革新知能統合研究センター 佐久間 淳

近年の分子生物学の発展により、大規模 SNP 解析のコストは劇的に減少し、各個人の全 SNP が十数万円程度のコストで解析可能になった。臨床的要因(例えば血圧や家族の罹患歴)や遺伝的要因(例えばSNP)と疾患の関連性も徐々に明らかになりつつあり、個人の体質に合わせた効果的な予防医療・先制医療が広く普及すると予想されている。近い将来、研究者のみならず、医療業務従事者や患者自身が個人ゲノムに触れる機会をもつことになると予想される。講演では、個人ゲノムの医療応用と医学研究の観点から、二つの個人ゲノム利用のプライバシー上のリスクと、プライバシー保護手法を紹介する。

(1) 疾患リスクからの個人ゲノム推定の抑制:
個人ゲノムは様々な疾患を生じさせる原因になっていると考えられ、個人ゲノムを調べることでそれらの疾患の発症リスクを評価することができる。疾患リスクをロジスティック回帰でモデル化した場合、疾患を発症する確率についての対数オッズ(疾患リスク)は、個人ゲノムを属性に持つ離散値の特徴量ベクトルと有限精度の回帰係数の内積値で評価される。このとき、回帰係数の有効桁数が大きい場合には、対数オッズから個別の個人ゲノムの値が特定あるいは高い確率で推測されるリスクがある。講演者らの研究グループでは、複数の生活習慣病に関する疾患の発症リスクモデルにおいて、実際の対数オッズから個別の個人ゲノムが特定される確率を評価した。また、特定される確率を指定した値以下に抑えるための、リスク値の区分化方法およびその最適な区分の設計方法を紹介する[1]。

(2) ゲノム疫学における検定統計量からの個人ゲノム推定の抑制:
Homer らは、ゲノムワイド相関解析(GWAS)における症例対象研究において、症例群および対照群における対立遺伝子の頻度表を公開したときに、その研究に参加した被験者の SNP がわかれば、その被験者が症例群と対照群のどちらに属するかを統計的に推測可能であることを示した[2]。また、同様に求めた検定統計量からも、類似の推測が可能であることが報告されている。この報告を受け、NIH は公開情報としていた GWAS に関連する頻度表を非公開とすることを決定した。講演では、このような統計量からの差分プライバシーにもとづき、GWAS に用いられる(多重)カイ二乗検定を安全に行いつつ、第一種の過誤と第二種の過誤を適切に制御する方法を紹介する[3]。

参考文献

[1] Kosuke Kusano, Ichiro Takeuchi, Jun Sakuma: Privacy-preserving and Optimal Interval Release for Disease Susceptibility.Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (AsiaCCS 2017), pp.532-545, 2017.
[2] Homer et al.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, PLoS genetics, vol. 4, no. 8, e1000167, 2008, Public Library of Science.
[3] Kazuya Kakizaki, Kazuto Fukuchi, Jun Sakuma: Differentially Private Chi-squared Test by Unit Circle Mechanism, Proceedings of the 34th International Conference on Machine Learning (ICML2017), to appear.