一般口演
# 一般口演11
# セキュリティとプライバシー保護2
2017年11月21日(火) 16:00 〜 17:30 H会場 (10F 会議室1008)

[2-H-3-OP11-5] 塩基配列データの改ざんの可能性とその防止策

作佐部 太也[1], 木村 通男[2] （1.藤田保健衛生大学 医療科学部 臨床工学科, 2.浜松医科大学附属病院 医療情報部）

【目的】塩基配列データについての 解析の手法やツールの進歩により、より多くの情報が得られるようになったが、一方では、改ざんなどの不正をより高度にするリスクも高まったと考えられる。本研究では塩基配列データの意図的な改ざんの可能性とその防止について検討する。
【方法】塩基配列データについてのソフトウェアやデータ形式、公共データベースなどについての調査を行った。
【結果】 NGSシミュレータと呼ばれるソフトウェアがあり、リアリティのある塩基配列データを合成することができることがわかった。特に遺伝子変異を指定して埋め込むなど高度なプログラミングができる機能を持つものがあり意図的な捏造が可能であることがわった。一方、塩基配列データに関するデータ形式やデータベースには改ざん防止などの対策がほとんど取られていないことがわかった。
【結論】改ざん防止のためには学術団体だけでなく機器メーカなどを含め 社会的な基盤整備が必要であることがわかった。手法としては医療情報の分野で既に確立した技術となっている電子署名が有効であると考えられる。また、一度改ざんされた塩基配列データが公共データベースに登録されると、そのデータは他の研究で再利用されるため汚染が拡大する可能性もあり、公共データベースの運営組織との国際的な取り組みも必要である。

塩基配列データの改ざんの可能性とその防止策

作佐部 太也[*1]、木村 通男[*2]

*1 藤田保健衛生大学 医療科学部 臨床工学科、*2 浜松医科大学附属病院 医療情報部

Possibility of tampering with nucleotide sequence data and precautionary measure

Takaya Sakusabe[*1] , Michio Kimura [*2]

*1 Fujita Health University, School of Health Sciences, Faculty of Clinical Engineering
*2 Hamamatsu University Hospital, Department of Medical Informatics

[PURPOSE] Progress in analysis method and tool of nucleotide sequence data brings more information, but it is considered that fraud such as tampering raised the risk. [METHOD] We investigated software, data formats, public databases on nucleotide sequence data. [RESULT] There are software called NGS simulator, and it is possible to synthesize reality base sequence data. Particularly, it is possible to intentionally fabricate sequence data which contains a specified gene mutation by programming. The data formats and the public databases does not have preventing methods for tampering. [CONCLUSIONS] In order to prevent tampering, efforts not only academic groups but also equipment manufacturers is necessary. As a method, electronic signature technology which is already established in the medical informatics is effective. Once sequence data that has been tampered with is registered in the public database, pollution may be expanded because that data is reused in other research. International efforts with the public database management organization are also necessary.

Keywords: NGS, Scientific misconduct, Simulation, Electronic signature

1. 目的

今日、次世代シーケンサー（NGS）により生成された大量の塩基配列データ(NGS データ)の取得が可能になり、それに応じて高度な解析ソフトウェアがオープンソースとして配布されている。加えて応じて高速・大容量な計算機の低価格化もあり、ゲノム解析研究の裾野は広まりつつある。このことは研究の発展という観点からは良好な状況と考えられるが、研究における不正の防止という観点からはリスクを孕んだ状況でもある。特にデータに関する不正の防止について検討の余地があると考えられる。これは、は画像処理ツールの普及が研究における不正を少なからず助長したという指摘 [1]からも、類似の現象が NGS データにおいても起こることは推測しうることである。不正データが生成された場合、それが単一の研究に用いられるだけであれば、不正が指摘され研究が否認されることにより、その影響は局限される。しかし、塩基配列データは公共データベースに登録され他の研究者により流用されるため、ひとたび不正データが登録されれば、その影響圏を把握することすら困難である。

本研究では、NGS データに関わるデータの不正の可能性と防止について、調査および提案を行う。

2. 方法

不正データの生成のためのツールとなりうる計算機リソースについて検討を行った。とりわけ塩基配列データを人為的に生成しうる NGS シミュレータについて調査を行った。

先行研究により塩基配列データに暗号技術を適用することで改竄検知自体は可能であることが示唆されており [2]、公共データベースに改竄検知のためのフレームワークをどのように組込むことが可能であるかについて調査を行った。

3. 結果

3.1 NGS データ処理の障壁の低下

不正な NGS データの生成を意図する者が十分な分子生物学的な知識を持つのは当然であり、一方、そのような者が実際に改竄を実施しようとする場合に最初に障壁となると想定されるのが計算機関連の機材の調達やスキルの習得である。そして今日、そのハードルが低くなってきているということである。

NGS データのサイズは数ギガバイトから数十ギガバイトであり、以前であればこれは特殊なハード／ソフトを必要とする巨大なデータであった。しかし、今日の PC のメインメモリは、コンシューマ向けのデスクトップ PC でも 64GB、ノート型 PC ですら 32GB のものもある。また、補助記憶装置としても SSD の大容量、低価格化、更には PC 向けの OS が仮想記憶を備え、アドレス空間も 64-bit 化したことなどもあり、テキストエディタで直接に編集することすら不可能ではなくなっている。

また、NGS データを処理する主要なプログラムの多くはオープンソースとして配布されている。それらの多くは UNIX 系統の OS 上で動作し、その操作には GUI ではなくコマンドラインを用いるものが多い。プログラムの使用方法についての情報が書籍やインターネット上で掲載されているが [3]、特にインターネットから場合、コマンドラインであればコピー＆ペーストにより簡単に実行させることができる。加えて、医学生物学系の研究者が Apple 社製の PC を好むことは頻繁に言及されるが、現在、それらのオペレーティングシステム（OS）は UNIX 系統であり、プログラムのインストールから実行についてのスキルの障壁は低いものとなっている。

3.2 NGS シミュレータ

NGS データを用いた研究のうち、疾患などに関する臨床研究においては、基準となる塩基配列（リファレンス）に対する SNP や indel などの小さな変異の探索し変異データを取得する。ヒトゲノムの変異についは位置、遺伝子との関係などがデータベース化されており、臨床研究では疾患などの現象と遺伝子変異との対応付けが解析対象となる。

NGS データは塩基配列の断片（リード）の集合である。あるリードが染色体のどの部分から得られたものであるかというのは、配列解析処理（パイプライン）の一環であるアライメント処理によって推定されるものである。アライメント処理は統計解析に基づいた確率的な処理である。このため、あるリードに小さな改変を加えた場合、そのリードが改変後にも同じ位置にアライメントされるとは限らない。すなわち、配列データを改変できたとしても、変異データが望み通りにできるとは限らない。この問題に対する解決策となる可能性があるのが NGS シミュレータである。

NGS シミュレータは、NGS を用いずに NGS データを生成するソフトウェアである。NGS シミュレータについての調査研究 [4]によると、変異データに基づいて NGS データを生成できる機能をもつ NGS シミュレータがある。本来であれば、この機能はパイプラインの挙動を詳細に検証するため用いられるのだが、意図的な改ざんを行う上でも有効なツールになり得ると考えられる。

NGS シミュレータの一つで NGS メーカの illumina 社が配布している EAGLE について実際にその動作の調査を行った。EAGLE は、リレファレンス配列への変異の適用、塩基配列の分割、NGS での塩基配列の読み取り、という現実の現象に沿って、シミュレーションを行う。その為、出力される NGS データは実際の NGS より出力されたものと形式的には全く区別できない。変異情報としては、通常のゲノム解析の結果として用いられる VCF 形式が適用される。例えば図 1 のように、実際のゲノムから得られた変異データに対して、編集を加えて改変し、NSG シミュレータに入力することで、意図的に変異を僅かに改変した配列データを容易に生成できる可能性がある。

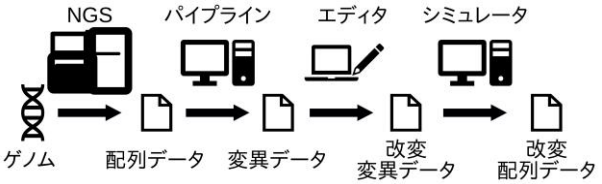

図 1 データ改ざんのフロー

3.3 公共データベース

塩基配列情報の解析に基づく研究では研究成果の公表の際に、取得した NGS データの公共データベースへの登録、公開が義務付けられることがある。そして、登録された NGS データは当該研究の証拠としてだけではなく、以後の別の研究において参照され再利用されることになる。公共データベースについては、国際的に協調、集約の動向があり、現在その中心となっているのは、INSDC (The International Nucleotide Sequence Database Collaboration) の活動として米国の NCBI（The National Center for Biotechnology Information）、欧州の EMBL-EBI (The European Bioinformatics Institute)、日本の DDBJ（ DNA Data Bank of Japan）が協調して蓄積、管理している SRA（Sequence Read Archive)である [5]。

SRA への登録の際には NGS データに加えて、研究や実験に関する情報はメタ・データが付加される。

SRA の内部では、NGS データは NGS によって異なるデータ形式の違いを吸収するために独自のフォーマット（SRA 形式）で保存される。SRA から NGS データを取得する際には、解析システムに入力できる FastQ 形式など変換するためのツールや Web システムが用意されている。

SRA 形式は、データ形式記述(schema)が内包された自己記述的なポータブルなデータベース・ファイル形式であり、vertical database と呼ばれている。schema としてフィールド名やデータ型の情報があるため、フィールド名の衝突に注意さえすれば、新たなフィールドの追加などが容易に行える非常に柔軟な形式であるため、電子署名を格納するためフィールドを追加することは技術的には問題がない。

4. 考察

情報技術の側面からだけ言えば、偽造あるいは改ざんした不正 NGS データを生成することは可能であることが分かった。ただし、特定の研究上の結論を肯定（あるいは否定）するだけでなく、分子生物学的および遺伝統計学的に矛盾のない人工的な NGS データが生成できるかについては検討が必要である。加えて、NGS シミュレータを使いこなすためには分子生物学的な知識に加えて、NGS の仕組みについての理解が必要であった。したがって、実際に有意な不正 NGS データを生成することは現時点では容易なことではないと考えられる。しかし、不可能ではない以上は事前に対策をすべきである。

一般にデータの偽造あるいは改ざんを検知するために、PKI にもとづく電子署名を用いることが有効である。電子署名を NGS データに適用する場合、電子署名の生成および埋め込みは NGS 内部で行われなければならない。そして、不正 NGS データの拡散を防ぐためには、公共データベースが NGS データを受け入れる際には、電子署名を検証するようにしなければならない。このような社会的なフレームワークの構築には、学会・研究者、メーカ・ベンダそして行政の連携が必要である。その点では、医療情報の分野では、すでに PKI などを活用した情報基盤の構築について多くの実績がある。そこで、医療情報関係者が NGS データの運用について関心を持ち、協力することが重要であると考えられる。

参考文献

1) 榎木英介. 生命科学の研究倫理 なぜ不正が絶えないのか？. KEIO SFC JOURNAL 2015；15-1：340-362.
2) 作佐部太也, 大内雄矢, 澤智博, 渡辺浩, 中島直樹, 木村通男. 証拠性のある医学研究 － 次世代シーケンサーからのデータの証拠性確保における暗号技術の利用についての評価と提案. 医療情報学 2016；36(Suppl.)：720-721.
3) 清水厚志, 坊農秀雅. 細胞工学別冊 次世代シーケンサーDRY 解析教本. 学研メディカル秀潤社, 2015.
4) Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data. NATURE REVIEWS GENETICS 2016；17：459-469.
5) Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Research 2011；39 Database issue：D19-D21.