Studies on Security and Privacy of Genomic Data

Thesis submitted in partial fulfillment of the requirements
for the degree of
Master of Technology in Computer Technology

by

Vaishnavi R
Regn. No. - 160098 of 2021-2022
Exam Roll No. - M6TCT24022

under the supervision of

Dr. Mridul Sankar Barik
Assistant Professor

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University
2024

Certificate from the Supervisor

This is to certify that the work embodied in this thesis entitled **"Studies on Security and Privacy of Genomic Data"** has been satisfactorily completed by **Vaishnavi R** (Registration Number 160098 of 2021 − 2022; Class Roll No. 002110504027; Examination Roll No. $M6TCT24022$). It is a bona-fide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata for partial fulfilment of the requirements for the awarding of the **Master of Technology in Computer Technology** degree of the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021 − 24.

Dr. Mridul Sankar Barik,
Assistant Professor,
Department of Computer Science and Engineering,
Jadavpur University.
(Supervisor)

Forwarded By:

Prof. Nandini Mukhopadhyay,
Head,
Department of Computer Science and Engineering,
Jadavpur University.

Prof. Dipak Laha,
DEAN,
Faculty of Engineering & Technology,
Jadavpur University.

Department of Computer Science and Engineering
Faculty of Engineering And Technology
Jadavpur University, Kolkata - 700 032

Certificate of Approval

This is to certify that the thesis entitled **"Studies on Security and Privacy of Genomic Data"** is a bona-fide record of work carried out by **Vaishnavi R** (Registration Number 160098 of $2021 - 2022$; Class Roll No. 002110504027; Examination Roll No. $M6TCT24022$) in partial fulfilment of the requirements for the award of the degree of **Master of Technology in Computer Technology** in the **Department of Computer Science and Engineering, Jadavpur University**, during the period of June 2021 to June 2024. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose of which it has been submitted.

Examiners:

(Signature of The Examiner)

(Signature of The Supervisor)

Department of Computer Science and Engineering
Faculty of Engineering And Technology
Jadavpur University, Kolkata - 700 032

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that the thesis entitled **"Studies on Security and Privacy of Genomic Data"** contains literature survey and original research work by the undersigned candidate, as a part of his degree of **Master of Technology in Computer Technology** in the **Department of Computer Science and Engineering, Jadavpur University**. All information have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Vaishnavi R

Examination Roll No.: M6TCT24022

Registration No.: 160098 of 2021 – 2022

Thesis Title: Studies on Security and Privacy of Genomic Data

Signature of the Candidate:

ACKNOWLEDGEMENT

I am pleased to express my gratitude and regards towards my Project Guide **Dr. Mridul Sankar Barik**, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University, without whose valuable guidance, inspiration and attention towards me, pursuing my project would have been impossible.

Last but not the least, I express my regards towards my friends and family for bearing with me and for being a source of constant motivation during the entire term of the work.

Vaishnavi R
MTCT Final Year
Exam Roll No. - M6TCT24022
Regn. No. - 160098 of 2021 – 2022
Department of Computer Science and Engineering,
Jadavpur University.

Contents

Certificate from Supervisor **I**

Certificate of Approval **II**

Declaration of Originality **III**

Acknowledgement **IV**

1 Introduction **1**
1.1 Background . . . 1
1.2 Research Statement . . . 1
1.3 Contribution of the Thesis . . . 2
1.4 Outline of the Thesis . . . 2

2 Genomic Data Life Cycle **3**
2.1 Data Generation . . . 3
2.2 Data Storage . . . 4
2.2.1 Data Formats . . . 4
2.3 Data Transmission . . . 5
2.4 Data Processing . . . 6
2.5 Data Analysis . . . 7
2.6 Data Sharing . . . 9
2.7 Genomic Data and Precision Medicine . . . 10
2.8 Genomic Data Stakeholders . . . 11
2.9 International Scenario . . . 12
2.9.1 The Human Genome Project . . . 12
2.9.2 National Center for Bio-technology Information . . . 12
2.9.3 ISO Standards . . . 13
2.10 Indian Scenario . . . 13
2.11 Importance of Genomic data . . . 14
2.12 Usage of Genomic data . . . 14

3 Security and Privacy Threats to Genomic Data **16**
3.1 Security Threats in Genomic Data . . . 16
3.2 Privacy Threats to Human Genomic Data . . . 17

3.2.1 Re-identification Threats . . . 18
3.2.2 Phenotype Inference . . . 18
3.2.3 Other Threats . . . 19

4 State of the Art in Genomic Data Security **21**
4.1 Homomorphic Encryption . . . 21
4.1.1 Types of Homomorphic Encryption . . . 21
4.1.2 The Role of Homomorphic Encryption in Securing Genomic Data . . . 21
4.2 Blockchain . . . 22
4.2.1 Types of Blockchain Technology . . . 22
4.2.2 Blockchain in Genomics . . . 23
4.2.3 Use Case of Blockchain in Genomics . . . 23
4.2.4 Blockchain Operations in Genomics . . . 24

5 Modeling Privacy Threats to Genomic Data **26**
5.1 Threat Modeling Approaches . . . 27
5.1.1 STRIDE . . . 27
5.1.2 PASTA . . . 27
5.1.3 DREAD . . . 27
5.1.4 MITRE ATT&CK Framework . . . 28
5.1.5 LINDDUN . . . 29
5.2 The LINDDUN Methodology . . . 32
5.2.1 Mapping Privacy Threats to DFD . . . 33
5.2.2 Mapping Privacy Threat Categories . . . 35
5.2.3 Detailing Privacy Threats via Threat Tree Patterns . . . 38
5.2.4 Privacy Requirements Analysis . . . 51
5.2.5 From Privacy Requirements to Privacy Solutions . . . 52

6 Conclusion and Future Work **54**
6.1 Conclusion . . . 54
6.2 Future Work . . . 54

List of Figures

2.1 Genomic Data Lifecycle 3

5.1 LINDDUN Privacy Threat Types. Source: https://linddun.org/ 31
5.2 LINDDUN Privacy Threat Framework. 32
5.3 DFD of Genomic Data Processing Systems 33
5.4 Threat tree for linkability of an entity. 39
5.5 Threat Tree for Linkability of a Dataflow 40
5.6 Threat Tree for Linkability of a Datastore. 41
5.7 Threat tree for linkability of a Process. 41
5.8 Threat Tree for Identifiability of An Entity. 42
5.9 Threat Tree for Identifiability of a Dataflow. 43
5.10 Threat Tree for Identifiability of a Datastore. 44
5.11 Threat Tree for Identifiability of a Process. 44
5.12 Threat Tree for Non-Repudiation of Data Flow. 45
5.13 Threat Tree for Non-Repudiation of Data Store. 46
5.14 Threat Tree for Non-Repudiation of Process. 47
5.15 Threat tree for detectability of data flow 48
5.16 Threat Tree for Detectability of Data Store 49
5.17 Threat Tree for Detectability of Process. 49
5.18 Threat Tree for Information Disclosure. 50
5.19 Threat Tree for Content Unawareness. 50
5.20 Threat Tree for Consent and Policy Noncompliance. 51

List of Tables

5.1 STRIDE Threat Model 27
5.2 LINDDUN Methodology: Privacy Properties and the Corresponding Privacy Threats 34
5.3 DFD elements in the Bioinformatic tools 35
5.4 Mapping LINDDUN components (privacy threats) to DFD element types (E-Entity, DF-Data flow, DS-Data store, P-Process) 36
5.5 Determining privacy threats for DFD elements within the Bioinformatic application (From left to right: L-Linkability, I-Identifiability, N-Non Repudiation, D-Detectability, D-Information Disclosure, U-Content Unawareness, N-Consent/policy Noncompliance) 37

Abstract

The digital age has ushered in a heightened awareness of privacy concerns, yet systematic methods to address privacy threats in software-based systems are still lacking. To bridge this gap, we introduce a comprehensive framework tailored for modeling privacy threats with precision. Drawing inspiration from established models like STRIDE, our framework using Lindunn methodology employs an information flow-oriented approach to guide analysts in identifying and mapping privacy threats to system elements. We provide a detailed catalog of privacy-specific threat tree patterns to enhance threat analysis, facilitating a nuanced understanding of potential risks. Additionally, our framework simplifies the selection of privacy-enhancing technologies (PETs) by offering guidance on mapping these technologies to identified threats, streamlining the process of privacy requirements elicitation and fulfillment within software systems. Through these contributions, our framework aims to bolster privacy protection efforts and promote trust in digital environments.

Chapter 1

Introduction

1.1 Background

The digital age has brought about significant opportunities in healthcare, personalized medicine, and genomic research. Genomic data, which is the blueprint of an individual's genetic makeup, holds immense promise for understanding diseases, predicting predispositions, and tailoring treatments. However, the security and privacy of genomic data are crucial due to its unique, immutable, and highly revealing nature. The proliferation of genomic databases and bioinformatics tools has underscored the need for robust security measures to prevent unauthorized access, data breaches, and identity theft.

This thesis aims to establish a comprehensive framework for genomic data privacy using the Lindunn methodology. The Lindunn methodology, known for its robustness in data privacy analysis, offers a promising avenue for addressing the privacy concerns inherent in genomic data. The research aims to craft a structured framework encompassing privacy-preserving techniques, risk assessment protocols, and regulatory compliance measures tailored specifically to the genomic data landscape.

By harmonizing the principles of the Lindunn methodology with the unique challenges of genomic privacy, this thesis seeks to provide a systematic approach for safeguarding sensitive genetic information while facilitating legitimate research endeavors. Through a multidisciplinary lens spanning genomics, data privacy, and regulatory compliance, this thesis aims to contribute to the advancement of genomic privacy practices, fostering trust, transparency, and ethical stewardship in the handling of genomic data.

1.2 Research Statement

The aim of this thesis is: *To explore the suitability of existing techniques for modeling threats in gnomic data processing systems.* The concept of the Lindunn approach and the stages that make up the framework are provided in this thesis. We have organised the bioinformatic tools dataflow diagram and mapped privacy risks to DFD. Additionally, we have defined the Lindunn methodology's threat categories. Using the Lindunn threat tree pattern, we have created a threat tree for bioinformatic tools after specifying all these parameters.We have examined the tools' privacy needs using the threat tree pattern, and we have examined several solutions based on those requirements.

The same's outcomes are presented in this thesis. The results have also been thoroughly interpreted, outlining how the Lindunn technique will contribute to the improvement of genetic data security.

1.3 Contribution of the Thesis

This thesis contributes significantly to the fields of security, privacy, and threat modeling methodology, with a specific focus on the Lindunn methodology. Through a comprehensive analysis, it identifies and categorizes security and privacy threats within genomic data processing systems, shedding light on the unique challenges inherent in safeguarding genetic information. By exploring established threat modeling methodologies and applying them to genomic data, the research underscores the importance of a structured approach to threat analysis. Central to the thesis is the Lindunn methodology, which emerges as a powerful framework for privacy analysis. By defining its steps and integrating it with genomic privacy concerns, the research demonstrates its effectiveness in identifying and mitigating privacy threats. Moreover, the thesis provides practical applications and recommendations for enhancing security and privacy in genomic data processing systems, offering actionable insights for stakeholders in the field. In summary, this research significantly advances our understanding and practices in security, privacy, and threat modeling methodologies, particularly in the intricate realm of genomic data processing, ensuring the confidentiality and integrity of genetic information in an increasingly data-driven world.

1.4 Outline of the Thesis

The structure of this thesis begins with an abstract, followed by an Introduction in Chapter 1, which includes the research statement and contribution of the thesis. Chapter 2 briefly covers related materials to the thesis. Chapter 3 presents security and privacy threats to genomic data, while Chapter 4 covers the state of the art in genomic data security. In Chapter 5, the proposed work of this thesis, which is modelling privacy threats to genomic data using Lindunn methodology. Lastly, Chapter 7 presents the conclusion of the work completed in this thesis.

Chapter 2

Genomic Data Life Cycle

Genomic information, which may be processed from microbial, fungus, plant, and animal species, including humans, exists in different forms throughout its lifecycle, which may include the following phases: sample collection, sample preparation, data generation, and data analysis [27]. Fig 2.1 illustrates this lifecycle,



Figure 2.1: Genomic Data Lifecycle

2.1 Data Generation

Generating genomic data involves various techniques and technologies used to analyze and sequence an organism's entire genome or specific regions of its DNA. Here are some key methods and steps involved in genomic data generation:

- **DNA Extraction:** The process begins with extracting DNA from the biological sample (e.g., blood, tissue, saliva). Various methods like phenol-chloroform extraction, spin column-based extraction, or magnetic bead-based extraction are used.
- **Library Preparation:** After DNA extraction, the next step is library preparation. This involves fragmenting the DNA into smaller pieces and attaching specific adapters that facilitate sequencing. Library preparation methods can vary depending on the sequencing platform used.

Sequencing Technologies:

- Next-Generation Sequencing (NGS): NGS technologies like Illumina[18] sequencing use massively parallel sequencing to simultaneously sequence millions of DNA fragments. It involves steps such as amplification, sequencing by synthesis, and imaging.
- Third-Generation Sequencing: Technologies like PacBio [18] and Oxford Nanopore offer long-read sequencing, providing more contiguous sequences and helping to resolve complex genomic regions.

Data Acquisition and Analysis:

- Primary Data Generation: During sequencing, raw data (sequence reads) are generated and stored in FASTQ or other formats.
- Bioinformatics Analysis: Bioinformatics tools and algorithms are used to process, align, and assemble the sequence reads into a coherent genome. This includes base calling, quality control, read mapping, variant calling, and genome assembly.

Genome Assembly and Annotation:

- Genome Assembly: This process involves aligning and merging overlapping DNA sequences to reconstruct the complete genome.
- Genome Annotation: Identifying functional elements within the genome, such as genes, promoters, regulatory elements, and non-coding regions.

Quality Control and Validation: Checking the accuracy and quality of the assembled genome is crucial. Validation can be performed using various methods, including PCR, Sanger sequencing, or comparison with known reference genomes.

2.2 Data Storage

Genomic data, due to its size and complexity, requires specialized storage systems capable of handling large volumes of information. Here are some aspects and considerations regarding data storage for genomic data:

- High-Performance Computing (HPC): Genomic data often requires substantial computational resources for processing and analysis. HPC clusters or cloud-based platforms provide the necessary infrastructure for efficient data handling.
- Distributed Storage: Genomic data is typically stored across distributed systems to handle the massive amounts of data generated by sequencing technologies. Distributed file systems like Hadoop Distributed File System (HDFS) or object storage systems like Amazon S3, Google Cloud Storage, or Azure Blob Storage are commonly used.

2.2.1 Data Formats

Genomic data is often stored in specialized file formats like FASTQ (raw sequencing data), BAM (aligned sequencing data), and VCF (variant call format). These formats are optimized for storing and analyzing DNA sequencing data efficiently.

2.3 Data Transmission

Transferring genomic data involves moving large volumes of genetic information from one location to another, often between research institutions, sequencing facilities, or data repositories. Given the substantial size of genomic datasets, efficient and secure data transfer methods are essential. Here are some considerations and methods for genomic data transfer:

- **High-Speed Networks:** Utilizing high-speed networks, such as dedicated research networks or high-bandwidth internet connections, facilitates faster data transfer rates. These networks are specifically designed for large data transfers and offer higher throughput.
- **Data Compression:** Compressing genomic data before transfer reduces file sizes and speeds up transfer times. There are specialized compression algorithms and tools tailored for genomic data, such as CRAM (Compressed Random Accessible Memory) for sequence alignment/map format (SAM/BAM) files.
- **Aspera, Globus, and FTP:**
  - Aspera and Globus are high-speed data transfer solutions designed for moving large datasets efficiently over long distances. These platforms optimize data transfer and provide security features.
  - File Transfer Protocol (FTP) is commonly used for data transfer, although its speed may be limited for large genomic datasets compared to more specialized solutions.
- **Shipping Physical Storage Media:** For exceptionally large datasets or situations with limited network bandwidth, physically shipping hard drives or storage media (like tapes) containing the genomic data can be more practical. This method, though slower in transit time, can be faster than transferring data over the network.
- **Cloud-Based Data Transfer:** Cloud services provided by platforms like AWS (Amazon Web Services), Google Cloud, or Azure offer tools and services for transferring large volumes of data into their respective cloud environments. This can include direct data uploads, import/export services, or using specific cloud-based storage solutions.
- **Encryption and Security:** Encrypting data during transfer ensures its security and prevents unauthorized access or interception. Secure file transfer protocols (e.g., Secure File Transfer Protocol - SFTP, Secure Shell - SSH) and encryption methods (e.g., SSL/TLS) are commonly used to protect data during transit.
- **Quality of Service (QoS):** Prioritizing network traffic for data transfer can ensure that genomic data transfers receive sufficient bandwidth and priority, optimizing transfer times and reducing potential delays due to network congestion.
- **Monitoring and Error Handling:** Continuous monitoring of data transfer processes helps detect and address any errors or interruptions promptly. Implementing error-checking mechanisms and resuming transfers from the point of interruption can prevent data loss or corruption. Efficient and secure data transfer methods are crucial for ensuring the timely and reliable movement of large genomic datasets, supporting collaborative research efforts and enabling the utilization of genetic information for various scientific and medical purposes.

2.4 Data Processing

Analyzing genome data involves several steps and methods to extract meaningful information from the raw genetic information obtained through sequencing technologies. Here's a general overview of the key steps in genome data analysis:

- **Data Preprocessing:** Raw sequencing data often contains errors and artifacts that need to be corrected. Preprocessing steps may include quality control, adapter trimming, and filtering to remove low-quality reads and artifacts.
- **Genome Alignment:** Sequencing reads are aligned to a reference genome or assembled de novo to determine their genomic locations. This step helps identify variations, such as single nucleotide polymorphisms (SNPs) and structural variants.
- **Variant Calling:** Variants, including SNPs, insertions, deletions, and structural variations, are identified by comparing aligned sequencing reads to the reference genome. Variant calling algorithms detect differences between individual genomes and the reference.
- **Variant Annotation:** Annotating variants involves determining their functional effects and potential associations with phenotypic traits or diseases. This step may include predicting the impact of variants on protein structure and function, as well as assessing their evolutionary conservation and allele frequencies in population databases.
- **Population Genetics Analysis:** Population-level analyses examine genetic variation within and between populations to understand evolutionary processes, demographic history, and patterns of genetic diversity. Common analyses include estimating allele frequencies, measuring genetic relatedness, and detecting signatures of natural selection.
- **Genome-Wide Association Studies (GWAS):** GWAS identify genetic variants associated with specific traits or diseases by comparing genotype data from cases and controls. Statistical tests are used to assess the significance of genotype-phenotype associations while accounting for population structure and multiple testing corrections.
- **Functional Genomics:** - Functional genomic analyses aim to understand how genetic variants influence gene expression, protein function, and regulatory networks. Techniques such as RNA sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq), and assays for transposase-accessible chromatin using sequencing (ATAC-seq) provide insights into gene regulation and cellular processes.
- **Pathway and Network Analysis:** Pathway and network analysis integrate genomic data with biological knowledge to identify pathways, gene networks, and molecular mechanisms underlying complex traits and diseases. These analyses help prioritize candidate genes and pathways for further experimental validation.
- **Data Visualization and Interpretation:** Effective visualization tools and methods are essential for interpreting genomic data and communicating results. Visualization techniques include genome browsers, heatmaps, scatter plots, and network diagrams, which help researchers explore and interpret complex datasets.

- **Integration with other omics data:** - Integrating genomic data with other omics data such as transcriptomics, proteomics, and metabolomics can provide a more comprehensive understanding of biological systems and disease mechanisms.

Overall, genome data analysis is a multidisciplinary field that combines computational biology, statistics, and bioinformatics to extract biological insights from genomic information.

2.5 Data Analysis

Genomic data analysis is a complex process that transforms raw sequencing data into meaningful biological insights. This process involves several stages, each with specific methodologies, tools, and considerations for ensuring data quality, accuracy, and security. Here's a detailed breakdown of the genomic data analysis phase:

Initial Quality Control (QC)

- Quality Assessment: Tools like FastQC are used to evaluate the quality of raw sequencing data. Parameters such as read length, base quality scores, and GC content are analyzed to detect any potential issues with the sequencing data.
- Quality Filtering and Trimming: Low-quality reads and bases are removed or trimmed using tools like Trimmomatic or Cutadapt. This step helps to reduce errors in downstream analysis.

Sequence Alignment

- Reference Genome Alignment: Cleaned reads are aligned to a reference genome using aligners like Burrows-Wheeler Aligner (BWA), Bowtie2, or HISAT2. This process involves mapping the short reads to their corresponding positions in the reference genome.
- Output Formats: Aligned reads are stored in SAM (Sequence Alignment/Map) or BAM (Binary Alignment/Map) formats, which contain detailed information about the alignment, including read positions, quality scores, and alignment flags.

Post-Alignment Processing

- Sorting and Indexing: Aligned reads are sorted based on their genomic coordinates using tools like SAMtools. Indexing these sorted files allows for efficient access and visualization.
- Marking Duplicates: Tools like Picard are used to identify and mark PCR duplicates, which are reads that originate from the same DNA fragment. This helps to avoid biases in variant calling.
- Realignment and Recalibration: Realignment around indels (insertions/deletions) and base quality score recalibration are performed to correct alignment errors and improve variant calling accuracy. Tools like GATK (Genome Analysis Toolkit) are commonly used for these steps.

Variant Calling

- Variant Identification: Variant calling algorithms, such as those in GATK, FreeBayes, or VarScan, are used to identify genetic variants like single nucleotide polymorphisms (SNPs) and indels from the aligned data.

- Variant Filtration: Variants are filtered based on quality metrics to remove false positives. This step involves applying filters based on depth of coverage, quality scores, and other criteria to retain high-confidence variants.

Annotation

- Functional Annotation: Variants are annotated with functional information using tools like ANNOVAR or SnpEff. This step adds biological context to the variants, indicating their potential impact on gene function, regulatory elements, and protein structure.
- Clinical Annotation: For clinical applications, variants are annotated with known associations to diseases or phenotypes using databases like ClinVar, OMIM (Online Mendelian Inheritance in Man), and HGMD (Human Gene Mutation Database).

Data Integration and Multi-Omics Analysis

- Multi-Omics Integration: Genomic data can be integrated with other omics data (e.g., transcriptomics, proteomics, metabolomics) to provide a comprehensive view of the biological system. This integration helps in understanding the complex interactions between different molecular layers.
- Pathway and Network Analysis: Tools like Ingenuity Pathway Analysis (IPA) or Cytoscape are used to analyze the biological pathways and networks affected by the identified variants, providing insights into the molecular mechanisms underlying observed phenotypes.

Statistical and Computational Analysis

- Association Studies: Genome-wide association studies (GWAS) are conducted to identify genetic variants associated with traits or diseases. Statistical methods like linear regression, logistic regression, and mixed models are employed to test for associations.
- Machine Learning and AI: Machine learning algorithms are increasingly used to predict phenotypes, classify diseases, and identify potential biomarkers from genomic data. Techniques such as random forests, support vector machines, and neural networks can be applied to genomic datasets.

Visualization and Interpretation

- Data Visualization: Tools like Integrative Genomics Viewer (IGV) and UCSC Genome Browser are used to visualize genomic data, including read alignments, variant locations, and annotations. Visualization aids in interpreting complex data and identifying interesting patterns.
- Biological Interpretation: The results are interpreted in the context of existing biological knowledge. This step often involves collaboration with domain experts, including geneticists, biologists, and clinicians, to understand the implications of the findings.

Privacy-Preserving Computation

- Homomorphic Encryption: This method allows computations on encrypted data without decrypting it, ensuring data privacy throughout the analysis process.

- Differential Privacy: Techniques are applied to ensure that the output of any analysis does not reveal sensitive information about individual participants, protecting their privacy while allowing for useful data insights.

Ethical and Legal Considerations

- Informed Consent: Ensuring that participants have given informed consent for the use of their genomic data in research.
- Data Security and Compliance: Adhering to legal and regulatory requirements for data protection, such as GDPR, HIPAA, and other relevant frameworks. Ensuring that data access is restricted to authorized personnel and that data is encrypted both at rest and in transit.

Genomic data analysis involves a multi-step process that starts with quality control of raw data and progresses through alignment, variant calling, annotation, and integration with other data types. Advanced computational methods, including machine learning, are used to extract meaningful insights from the data. Throughout this process, maintaining data security and privacy, adhering to ethical standards, and complying with legal requirements are paramount. This ensures that the sensitive nature of genomic data is protected while enabling significant scientific discoveries and clinical advancements.

2.6 Data Sharing

Genomic data sharing refers to the practice of making genetic and genomic information available to researchers, clinicians, and other stakeholders for various purposes, such as scientific research, medical diagnostics, and treatment development. This data includes information about an individual's DNA sequence, genetic variations, and other molecular information.

There are several reasons why sharing genomic data is important:

- **Advancing Research:** Access to diverse genomic data sets facilitates research into the genetic basis of diseases, potential treatments, drug development, and understanding population genetics.
- **Precision Medicine:** Precision Medicine: Sharing genomic data enables the development of personalized or precision medicine by allowing clinicians to tailor treatments based on an individual's genetic makeup.
- **Data Integration:** Combining genomic data from different sources can lead to more comprehensive analyses, which may uncover new associations and patterns in genetic diseases or traits.
- **Improving Diagnostics:** Shared genomic data can enhance diagnostic capabilities, aiding in the identification of genetic markers for diseases or conditions.

However, genomic data sharing raises various ethical, legal, and privacy concerns:

- **Privacy and Confidentiality:** Genomic data contains sensitive and identifiable information. Ensuring privacy protection while sharing this data is crucial to prevent misuse or unauthorized access.

- **Consent and Ownership:** Clear consent mechanisms must be in place to ensure that individuals contributing their genetic information are aware of how their data will be used. Questions regarding data ownership and control also need addressing.
- **Data Security:** Genomic data needs robust security measures to prevent breaches or unauthorized access that could lead to misuse or exploitation.[5]

  Efforts have been made to establish guidelines, regulations, and platforms for responsible genomic data sharing. Initiatives like the Global Alliance for Genomics and Health (GA4GH) have developed frameworks and standards for sharing genomic and clinical data while addressing ethical and privacy concerns. Institutions and research consortia often have their own data sharing policies and platforms, ensuring compliance with legal and ethical standards.

  Overall, while genomic data sharing holds immense potential for scientific and medical advancements, it's crucial to balance the benefits with ethical considerations to protect individuals' privacy and ensure responsible data use.

2.7 Genomic Data and Precision Medicine

Genomic data and precision medicine are closely intertwined concepts that are revolutionizing healthcare and medical treatment. Precision medicine, also known as personalized medicine, is an approach to medical care that takes into account individual variability in genes, environment, and lifestyle for each patient. Genomic data plays a crucial role in enabling precision medicine by providing insights into an individual's genetic makeup.

Here are some key points regarding genomic data and precision medicine:

Genomic Data:

- Genomic data refers to the information encoded in an individual's DNA, including the sequence of their genes and variations in the genetic code. The Human Genome Project, completed in 2003, was a landmark effort to map and sequence all the genes in the human genome.
- Advances in DNA sequencing technologies, such as next-generation sequencing, have made it more affordable and accessible to generate genomic data for individuals.

Role of Genomic Data in Precision Medicine:

- Genomic data can identify genetic mutations or variations that may influence an individual's susceptibility to diseases, their response to treatments, and their risk of adverse drug reactions.
- This information can help healthcare providers tailor medical treatments and interventions to the specific genetic profile of each patient, increasing the effectiveness of healthcare and reducing potential side effects.

Applications of Genomic Data in Precision Medicine:

- **Cancer Treatment:** Genomic data can guide the selection of targeted therapies based on the genetic mutations present in a patient's tumor, leading to more effective cancer treatment.
- **Pharmacogenomics:** Genomic data can inform medication choices by identifying how an individual metabolizes drugs, helping to avoid adverse drug reactions and optimize drug dosages.

- **Rare Diseases:** Genomic sequencing can aid in diagnosing rare genetic disorders and provide insights into potential treatments.
- **Prevention:** Genomic data can identify individuals at higher risk for certain diseases, enabling early intervention and preventive measures.

Ethical and Privacy Considerations:

- The collection and sharing of genomic data raise important ethical and privacy concerns. Patients must have control over who accesses their genetic information and how it is used.
- Policies and regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, have been established to protect the privacy of patients' genetic data.

Data Sharing and Research:

- Sharing genomic data on a large scale can facilitate research into the genetics of various diseases, leading to a better understanding of their underlying causes and potential therapies.
- Initiatives like the Global Alliance for Genomics and Health (GA4GH) aim to establish data-sharing standards and guidelines to ensure responsible and secure sharing of genomic data

.

Future Outlook: The field of genomics and precision medicine continues to evolve rapidly, with ongoing research and technological advancements. As more data becomes available, precision medicine is expected to become increasingly effective in tailoring healthcare to individual patients.

In summary, genomic data plays a pivotal role in precision medicine by providing insights into an individual's genetic makeup, enabling more personalized and effective medical care. As our understanding of genomics advances, the potential for precision medicine to transform healthcare and improve patient outcomes continues to grow.

2.8 Genomic Data Stakeholders

Stakeholders in genomic data involve various individuals, organizations, and groups who have an interest or are impacted by the collection, storage, analysis, and application of genomic information. Here are some key stakeholders in genomic data:

- **Researchers and Scientists:** These individuals are involved in conducting studies, experiments, and analysis of genomic data. They aim to discover new information, patterns, and correlations within genetic sequences.
- **Healthcare Providers:** Clinicians, doctors, and genetic counselors use genomic data to diagnose diseases, predict susceptibilities, and provide personalized treatment plans based on an individual's genetic makeup.
- **Patients and Families:** Individuals whose genetic information is being collected and analyzed. They often seek insights into their health risks, hereditary conditions, and potential treatments.

- **Biotechnology and Pharmaceutical Companies:** These entities use genomic data for drug development, target identification, and personalized medicine. They may conduct research, clinical trials, and develop therapies based on genomic insights.
- **Ethical, Legal, and Social Implications (ELSI) Experts:** ELSI scholars and professionals address the ethical, legal, and social impacts of genomic research, ensuring that privacy, consent, and fairness are maintained while leveraging genomic data.
- **Regulatory Bodies and Governments:** Regulatory agencies set guidelines and regulations governing the collection, storage, and use of genomic data to ensure ethical standards, privacy protection, and data security.
- **Data Repositories and Biobanks:** These organizations store genomic data collected from various sources, providing access to researchers and ensuring data integrity and security.
- **Insurance Companies:** Insurers might be interested in genomic data to understand risks associated with certain diseases and conditions. However, ethical considerations regarding the use of this data in insurance practices exist.
- **Educational Institutions:** Universities and academic institutions train future researchers and professionals in genomics, providing education and conducting research in the field.
- **Community and Patient Advocacy Groups:** These groups advocate for patient rights, educate the public about genomic research, and ensure that diverse communities are represented and benefit from genomic advancements.

Understanding the perspectives and interests of these stakeholders is crucial for responsible and ethical use of genomic data while ensuring its benefits reach society at large. Balancing scientific advancements with ethical considerations and individual rights remains a key challenge in the field of genomics.

2.9 International Scenario

2.9.1 The Human Genome Project

The Human Genome Project (HGP)[1] is one of the greatest scientific feats in history. The project was a voyage of biological discovery led by an international group of researchers looking to comprehensively study all of the DNA (known as a genome) of a select set of organisms. Launched in October 1990 and completed in April 2003, the Human Genome Project's signature accomplishment – generating the first sequence of the human genome – provided fundamental information about the human blueprint, which has since accelerated the study of human biology and improved the practice of medicine.

2.9.2 National Center for Bio-technology Information

The National Center for Biotechnology Information[2] (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). It is approved

[1] https://www.genome.gov/human-genome-project
[2] https://www.ncbi.nlm.nih.gov/

and funded by the government of the United States. The NCBI is located in Bethesda, Maryland, and was founded in 1988 through legislation sponsored by US Congressman Claude Pepper.

The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for biomedical literature. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine. NCBI was directed by David Lipman, one of the original authors of the BLAST sequence alignment program and a widely respected figure in bioinformatics.

2.9.3 ISO Standards

ISO/IEC 23092-1:2020(en) Information technology — Genomic information representation — Part 1: Transport and storage of genomic information. This document specifies data formats for both transport and storage of genomic information, including the conversion process.

2.10 Indian Scenario

Indian Biological Data Center[3] is the first national repository for life science data in India. IBDC is mandated to archive all life science data generated from publicly funded research in India. IBDC is being developed in a modular nature wherein different sections would typically deal with particular types of life science data. Each IBDC section would have dedicated data submission and access schema. Further, besides archiving life science data, IBDC shall also develop highly curated data sets in order to facilitate knowledge discovery in various domains of life sciences. It would also provide infrastructure and expertise for biological data analysis.

According to a recent report[4], India lacks a privacy law and consent framework to ensure that submitted data remains confidential through robust de-identification practices, and to make certain that consent of the participants is obtained for the kind of research that can be conducted on such biological information.

Direct-to-consumer genetic testing services[5][6] in India are currently rising in number and becoming a hugely competitive market. However, passive consent guidelines and the lack of a privacy regime arising from a well-thought-out consent paradigm raise questions about the protection of genetic data in light of its collection and use.

The Digital Personal Data Protection Act[7], 2023

The Digital Personal Data Protection Act, 2023 (DPDP Act) is an act of the Parliament of India that aims to provide for the processing of digital personal data in a manner that recognises both the right of individuals to protect their personal data and the need to process such personal data for lawful purposes12. The Act regulates the governance of personal data collected by organisations, and aims at protecting the individual's privacy by empowering them with rights over the manner in which their data is processed3. The Act includes a negative-list regime for cross-border data transfers, among several other key changes.

[3] https://ibdc.rcb.res.in/

[4] https://ohrh.law.ox.ac.uk/the-need-to-revisit-indian-consent-paradigms-for-the-storage-and-use-of-digitalized-biological-data/

[5] https://mapmygenome.in/

[6] https://www.dnaforensics.in/

[7] https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf

2.11 Importance of Genomic data

Genomic data holds immense importance in various fields, particularly in biology, medicine, agriculture, and evolutionary studies. Here are some reasons why genomic data is valuable:

- **Understanding Genetic Basis:** Genomic data helps in comprehending the genetic makeup of organisms, including humans. It provides insights into DNA sequences, variations, mutations, and structural changes, aiding in understanding the genetic basis of traits, diseases, and evolutionary relationships.
- **Personalized Medicine:** Genomic data plays a pivotal role in personalized medicine by enabling tailored treatments and therapies based on an individual's genetic makeup. This approach helps in predicting disease risks, selecting suitable medications, and optimizing treatment plans.
- **Disease Research and Diagnosis:** Genomic data contributes significantly to disease research, facilitating the identification of genetic markers associated with various diseases. It aids in early detection, accurate diagnosis, and development of targeted therapies for conditions like cancer, genetic disorders, and infectious diseases.
- **Precision Agriculture:** In agriculture, genomic data assists in crop improvement, livestock breeding, and enhancing agricultural productivity. Understanding the genetic characteristics of plants and animals helps in developing resilient crops, disease-resistant breeds, and improving overall agricultural sustainability.
- **Evolutionary Studies:** Genomic data provides a wealth of information for understanding evolutionary relationships between species. By comparing genetic sequences across different organisms, researchers can reconstruct evolutionary histories and study how species have evolved and diversified over time.
- **Biotechnological Advancements:** Genomic data serves as a foundation for various biotechnological advancements. It aids in the development of genetically modified organisms (GMOs), gene editing techniques like CRISPR-Cas9, and synthetic biology applications.
- **Drug Development:** Genomic data contributes to drug discovery and development by identifying potential drug targets, understanding drug responses based on genetic variations among individuals, and aiding in the development of more effective and safer medications.
- **Forensic Science:** In forensic investigations, genomic data, particularly DNA profiling, plays a crucial role in identifying individuals, solving crimes, and establishing familial relationships in legal cases.

2.12 Usage of Genomic data

Genomic data has a wide range of applications across various fields. Here are thirteen common uses:

Medical Research: Genomic data is used extensively in medical research to understand the genetic basis of diseases, identify biomarkers, and develop targeted therapies.

Personalized Medicine: Genomic data can help tailor medical treatments to individual patients based on their genetic makeup, allowing for more effective and personalized healthcare.

Pharmacogenomics: Genomic data is used to study how an individual's genetic variations influence their response to drugs, enabling the development of personalized drug therapies with optimized efficacy and reduced side effects.

Cancer Genomics: Genomic data analysis is crucial in cancer research to identify genetic mutations associated with different types of cancer, predict disease progression, and develop targeted cancer therapies.

Population Genetics: Genomic data is used to study genetic variation within and between populations, providing insights into human evolution, migration patterns, and genetic diversity.

Forensic Genetics: Genomic data is utilized in forensic investigations to identify individuals, establish familial relationships, and solve crimes through techniques like DNA fingerprinting and DNA profiling.

Agricultural Genomics: Genomic data is employed in agriculture to improve crop yields, enhance disease resistance, and develop genetically modified organisms (GMOs) with desirable traits.

Livestock Breeding: Genomic data is used in animal breeding programs to select for desired traits such as milk production, meat quality, and disease resistance, leading to more productive and healthier livestock.

Conservation Genetics: Genomic data aids in the conservation of endangered species by assessing genetic diversity, identifying populations at risk, and designing effective conservation strategies.

Microbial Genomics: Genomic data is employed to study microbial communities, identify pathogens, track disease outbreaks, and develop strategies for disease control and prevention.

Bioinformatics: Genomic data analysis requires sophisticated computational tools and algorithms, leading to the development of the field of bioinformatics, which focuses on processing, analyzing, and interpreting large-scale genomic datasets.

Genomic Medicine: Genomic data is integrated into clinical practice to diagnose genetic disorders, predict disease risk, and guide treatment decisions, ushering in the era of genomic medicine.

Ethical, Legal, and Social Implications (ELSI): Genomic data raises important ethical, legal, and social issues regarding privacy, consent, discrimination, and equitable access to genetic testing and healthcare services, necessitating ongoing dialogue and regulation.

These are just a few examples of how genomic data is utilized across diverse fields, with ongoing advancements continually expanding its applications and impact.

Chapter 3

Security and Privacy Threats to Genomic Data

3.1 Security Threats in Genomic Data

Genomic data contains a wealth of sensitive information about an individual, including their susceptibility to certain diseases, ancestry, and other genetic traits. Consequently, there are several threats associated with handling and storing genomic data:

Privacy Risks: Genomic data can be used to identify individuals, posing a significant threat to privacy. Even if names are removed, it might be possible to re-identify individuals through their genetic information by cross-referencing with other publicly available datasets.

Data Breaches: Genomic databases are valuable targets for hackers. Breaches can lead to the exposure of sensitive genetic information, which could be exploited for identity theft, discrimination, or other malicious purposes.

Genetic Discrimination: Misuse of genetic information can lead to discrimination in various areas such as employment, insurance, and education. Employers or insurance companies might discriminate based on an individual's genetic predispositions to certain diseases or conditions

Informed Consent: Ensuring informed consent for the use of genetic data is crucial. Without proper consent mechanisms in place, there's a risk of unauthorized or unethical use of the data.

Accuracy and Interpretation: Misinterpretation of genetic data or errors in data analysis could lead to incorrect medical decisions or diagnoses, potentially causing harm to individuals.

Ethical Concerns: There are ethical dilemmas surrounding the ownership and sharing of genomic data. Determining who has access to this information and for what purposes raises significant ethical questions.

Long-Term Storage and Management: Genomic data needs to be stored securely for long periods due to its potential lifetime relevance. Maintaining the security of this data over extended periods requires robust storage and management protocols.

Genomic Surveillance and Profiling: As technology advances, concerns arise about the potential for mass genomic surveillance or profiling, which could be exploited for purposes of social control or discrimination.

Addressing these threats requires comprehensive data protection measures, including strong

encryption, access controls, strict regulations governing data use, and ethical guidelines to ensure responsible handling and sharing of genomic information. It's crucial to balance the potential benefits of genomic data with the protection of individual privacy and rights.

Security threats to human genomic data encompass various risks to the confidentiality, integrity, and availability of this sensitive information. These threats include: **Unauthorized Access:** Hackers or unauthorized individuals may attempt to gain access to genomic databases or storage systems, either through exploiting vulnerabilities in software or systems or through social engineering techniques. Unauthorized access could lead to the theft or manipulation of genomic data.

Data Breaches: Data breaches pose a significant threat to genomic data security. Breaches may occur due to system vulnerabilities, insider threats, or cyberattacks, resulting in the unauthorized disclosure or theft of genomic information. This can lead to privacy violations, identity theft, or misuse of the data for nefarious purposes.

Malware and Ransomware: Malicious software, such as malware or ransomware, can infect systems hosting genomic data, causing data loss, corruption, or unauthorized access. Ransomware attacks, in particular, can encrypt genomic data and demand payment for its release, disrupting research activities or clinical operations.

Data Tampering: Attackers may attempt to tamper with genomic data to manipulate research results, introduce errors or biases, or undermine the integrity of scientific studies. Data tampering can lead to erroneous conclusions, misdiagnoses, or compromised patient care in clinical settings.

Denial-of-Service (DoS) Attacks: DoS attacks aim to disrupt the availability of genomic data by overwhelming servers or network infrastructure with excessive traffic or requests. Such attacks can impede research activities, delay patient diagnoses, or disrupt healthcare services relying on genomic information.

Insider Threats: Insiders, including employees, contractors, or researchers with authorized access to genomic data, may misuse or abuse their privileges for personal gain or malicious purposes. Insider threats can involve unauthorized data access, theft, or unauthorized sharing of sensitive information.

Interception of Data in Transit: Genomic data transmitted over networks, such as the internet or internal networks, may be intercepted by attackers, potentially exposing sensitive information to unauthorized parties. Encryption and secure communication protocols are essential to mitigate this threat.

Third-Party Service Risks: Outsourcing genomic data storage or analysis to third-party service providers introduces additional security risks. Organizations must carefully vet and monitor third-party vendors to ensure they adhere to security best practices and compliance requirements.

Addressing security threats to human genomic data requires implementing robust security measures, such as encryption, access controls, intrusion detection systems, and regular security audits. Additionally, organizations must prioritize employee training, data governance practices, and compliance with relevant regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR), to safeguard genomic data against security breaches and unauthorized access.

3.2 Privacy Threats to Human Genomic Data

In this section, we examine an extensive range of privacy risks associated with human genomic data, drawing from findings documented in previous research [25].

3.2.1 Re-identification Threats

Re-identification stands out as a deeply researched privacy concern in the distribution and examination of human genomic data. In this form of attack, unauthorized parties scrutinize publicly available human genomes, which are typically safeguarded to conceal the identity information of their contributors (e.g., patients), aiming to unearth the identities of the individuals involved. The success of such an attack can inflict severe harm on these contributors, including discrimination and financial repercussions. This section delves into the vulnerabilities present in current privacy protection methods that enable such re-identification attacks.

Pseudo-anonymized data, commonly employed for safeguarding health information, entails the removal of explicit and quasi-identifying attributes like names and dates of birth. This practice aligns with legal mandates such as de-identification regulations under the U.S. Health Insurance Portability and Accountability Act (HIPAA) for conventional health records. However, genomic data cannot be rendered fully anonymous solely by eliminating identifiable details. There exists a persistent risk wherein adversaries may deduce the phenotype of a DNA-material donor, encompassing observable traits like eye, hair, or skin colors, from their genotypes or genetic makeup, leading to their identification. Despite ongoing advancements, techniques to mitigate this risk are not yet fully developed. The rapid progression in genomic research and technologies is steadily advancing towards addressing this challenge. Furthermore, re-identification can occur by scrutinizing the contextual information accompanying publicized DNA sequences [9] [11]. For instance, genomic variations on the Y chromosome have been linked to surnames for males, discernible through public genealogy databases. Other instances include the identification of participants in initiatives like the Personal Genome Project (PGP) via public demographic data and the retrieval of familial identities from data disclosed by projects like the 1000 Genome Project utilizing publicly available information such as death notices and other correlation attacks [21]. Notably, even cryptographically secure protocols have demonstrated vulnerabilities, leaking substantial information when applied to genomic data[10].

Attacks on Machine Learning Models Most attacks on genomic data use the entire dataset for the attack. Recently, the research paper [8] showed that even the machine learning models that are trained on the genomic data can reveal information about the people whose data was used for training the model as well as any arbitrary person given some background information. They study Warfarin dosage example in detail and show that the attack has significant advantage compared to a random guess. The authors further show that differential privacy is not useful for Warfarin dosage machine learning model.

3.2.2 Phenotype Inference

Phenotype inference involves the intricate process of extrapolating observable traits or characteristics of an organism from its genetic composition, often derived from DNA sequences. This process, fundamental to genetic analysis, encompasses the prediction of various phenotypic outcomes, including physical attributes, predispositions to diseases, and behavioral tendencies, based on the individual's genomic information. In essence, it delves into understanding how genetic variations manifest in observable traits and health conditions.

One of the foremost privacy concerns surrounding human genomic data pertains to the inference of sensitive phenotype information from DNA sequences. This threat revolves around the potential disclosure of deeply personal information, such as susceptibility to diseases or unique physical traits, which could lead to privacy breaches and discrimination against individuals. The following sections

delve into the nuanced landscape of this privacy threat, elucidating prior research findings and implications.

Aggregate Genomic Data: Beyond the well-documented risks of re-identification, the correlation between an individual's genomic data and publicly available information presents a distinct challenge. For instance, even a modest set of genetic markers, such as single nucleotide polymorphisms (SNPs), can be leveraged to discern an individual's participation in genomic studies [19]. This underscores the necessity of stringent protocols and agreements governing the dissemination of sensitive DNA raw data to mitigate the risk of unintended re-identification.

In 2008, Homer et al. [13] demonstrated a significant privacy vulnerability in genomic data. They showed that adversaries could determine a participant's group affiliation (e.g., case or reference) by comparing their DNA sample to frequencies of genetic variants in various study populations. This prompted the removal of publicly available aggregate genomic data from NIH and Wellcome Trust websites, with researchers now required to sign data use agreements. Subsequent research enhanced attack methods, making identification possible with fewer genetic variations. Wang et al. even showed the potential for recovering DNA sequences from aggregate statistical data. These findings underscored the urgent need for stronger privacy protections in genomic data sharing and analysis.

Quantification of Information Content: Sophisticated analyses have demonstrated the remarkable accuracy with which an individual's participation in genome-wide association studies (GWAS) and their corresponding phenotypes can be inferred from aggregate genomic statistics. Such findings raise profound concerns regarding participant privacy and the potential for unauthorized re-identification, necessitating heightened vigilance and safeguards in genomic data sharing practices [28].

Correlation of Genomic Data: The phenomenon of linkage disequilibrium, wherein genetic variants in close proximity exhibit non-random associations, introduces a novel dimension to privacy threats [12]. This correlation enables the inference of concealed genetic information from partially available genomic data, posing significant privacy risks, particularly in contexts where sensitive variants are deliberately withheld.

Kin Privacy Breach: Beyond individual privacy concerns, the genetic data of individuals who opt not to publicly disclose their information can still be inferred through the genetic relationships among their relatives. Through sophisticated haplotype reconstruction techniques, familial genetic data can be deduced from publicly available genomic datasets, accentuating the need for robust privacy frameworks encompassing broader familial implications [14].

These nuanced privacy threats underscore the intricate interplay between genetic information, privacy concerns, and data sharing practices. Addressing these challenges mandates a multifaceted approach, combining rigorous privacy protocols, ethical considerations, and technological innovations to safeguard the confidentiality and integrity of individuals' genomic data while facilitating advancements in genetic research and personalized medicine.

3.2.3 Other Threats

In addition to the aforementioned threats, there exist several other privacy concerns associated with genomics:

Unveiling Anonymous Paternity: The transmission of the Y chromosome from father to son facilitates the linkage of this chromosome with surnames in genealogy databases, allowing for ancestral tracing. This information has been exploited in instances where individuals conceived

through sperm donation successfully identified their biological fathers. For instance, a 15-year-old boy traced his genetic lineage by submitting a cheek swab to a genealogy service and conducting an online search. Similarly, adopted individuals have utilized genealogy databases, coupled with substantial manual effort, to locate their biological fathers[24]. Consequently, DNA testing has rendered the anonymity of sperm donors obsolete, potentially compromising their privacy [17].

Legal and Forensic Concerns: DNA collection for legal and forensic purposes, involving both criminals and victims, raises significant privacy implications. While forensic techniques continue to advance with evolving technology [15] [26], instances of DNA misuse, such as staging crime scenes, have already been documented, posing challenges for law enforcement agencies. Notably, individuals like Madonna have expressed concerns regarding the misuse of their DNA, going to the extent of hiring DNA sterilization teams to eliminate residual genetic material. Despite limited privacy risk assessments in the legal and forensic context, the potential future scenario where law enforcement agencies gain access to entire genome sequence databases warrants attention. This aspect is intertwined with the discussion on sperm donor paternity breaches, highlighting its relevance in legal contexts as well.

Chapter 4

State of the Art in Genomic Data Security

4.1 Homomorphic Encryption

Cryptography was originally developed as a technique for secure communication between multiple parties, where one party encrypts a message and sends it to another party, who then decrypts it. Homomorphic encryption is a cryptographic technique that allows computations to be performed on encrypted data without decrypting it. In other words, it enables operations to be conducted on data while it remains in an encrypted state. This is particularly important for preserving privacy and security in scenarios where sensitive information needs to be processed by third parties or in the cloud. [29]

4.1.1 Types of Homomorphic Encryption

There are three main types of homomorphic encryption:

- **Partially Homomorphic Encryption (PHE)**: A partially Homomorphic Encryption (PHE) scheme can perform a single operation, such as addition or multiplication, over encrypted data an arbitrary number of times.
- **Somewhat Homomorphic Encryption (SHE)**: Somewhat Homomorphic Encryption (SHE) allows for bounded computation of a set of operations. It supports both addition and multiplication operations on encrypted data, but not both simultaneously.
- **Fully Homomorphic Encryption (FHE)**: FHE enables arbitrary addition and multiplication over encrypted data. It Permits both addition and multiplication operations on encrypted data in any order, providing a more comprehensive and powerful solution.

4.1.2 The Role of Homomorphic Encryption in Securing Genomic Data

Homomorphic encryption is considered suitable for the security of genomic data due to its ability to perform computations on encrypted data without decrypting it. Genomic data is highly sensitive

and valuable, containing information about an individual's genetic makeup, potential health risks, and other personal details. Here are some reasons why homomorphic encryption is a good fit for securing genomic data:

Privacy Preservation: Homomorphic encryption allows computations to be performed on encrypted genomic data without revealing the actual genetic information. This ensures that sensitive data remains confidential and private, even when computations are being carried out.

Secure Data Sharing: Genomic research often involves collaboration among multiple parties, such as researchers, institutions, and healthcare providers. Homomorphic encryption enables secure data sharing and collaborative analysis without exposing the raw genomic information to unauthorized individuals.

Computation on Encrypted Data: Traditional encryption methods require decryption before performing computations, which can pose a significant security risk. Homomorphic encryption, on the other hand, enables computations directly on the encrypted data, ensuring that sensitive genomic information remains protected throughout the entire analysis process.

Protection against Insider Threats: Homomorphic encryption helps safeguard genomic data against insider threats, where individuals with authorized access may attempt to misuse or leak sensitive information. Even with access to the encrypted data, performing meaningful computations without the decryption key remains challenging.

Compliance with Regulations: Many countries and regions have strict regulations regarding the handling and sharing of sensitive health-related data, including genomic information. Homomorphic encryption provides a mechanism for complying with these regulations by ensuring data security and privacy.

Secure Cloud-Based Analysis: Genomic data is often stored and processed in cloud environments. Homomorphic encryption allows secure cloud-based analysis by enabling computations on the encrypted data, minimizing the risk of data exposure during processing.

4.2 Blockchain

Blockchain technology is a decentralized and distributed ledger system characterized by its secure and transparent transaction recording across a network of computers. It operates on a peer-to-peer basis, ensuring that each participant has an identical copy of the ledger, promoting decentralization and eliminating the need for a central authority. The immutability of blocks, achieved through cryptographic hashes, enhances security, while transparency is maintained as all participants can access the same information. Smart contracts automate and enforce agreements, reducing the need for intermediaries, and various consensus mechanisms ensure agreement among nodes. Applications span from cryptocurrencies like Bitcoin and Ethereum to supply chain management, smart contracts in legal and business processes, healthcare, genomics, and identity management.[3]

4.2.1 Types of Blockchain Technology

The three distinct types of blockchains are categorized based on node participation and network behavior:

Public Blockchain (public and permissionless): In a public blockchain, such as the Bitcoin blockchain, the network is open to anyone, and trust is not required to join. Any anonymous node can participate in the network, and transactions are publicly broadcasted to all nodes. Consensus mechanisms are open to any node in the network, allowing for decentralized validation of blocks.

Private Blockchain (private and permissioned): In a private blockchain, exemplified by platforms like MultiChain, nodes are pre-selected and vetted, ensuring trustworthiness upon joining the network. Access control mechanisms provide a higher level of privacy for blockchain transactions. This type of blockchain typically exhibits higher performance in terms of block confirmations due to the controlled and known set of participants.

Consortium Blockchain (public and permissioned): Consortium blockchains, represented by platforms like Hyperledger, strike a compromise between efficiency and trust. The ledger is managed by a pre-selected group of nodes, and nodes need permission to join the network. Validation of blocks and transactions occurs when a specific set of nodes reaches a consensus, following pre-established rules of the consensus mechanism. This type combines aspects of both public and private blockchains.

4.2.2 Blockchain in Genomics

Below are some of the points to understand why blockchain is useful in genomics:

Genomic Data Security: Genomics data is very sensitive and crucial, from a data security perspective, blockchain provides excellent data security and integrity. Security methods such as encryption are useful in combating data breaches, but they do not provide complete protection. Many systems of big organizations with the highest level of security are penetrated by hackers. However, blockchain technology helps organizations by providing better protection against data breaches. Blockchain uses hashing techniques to store data securely, which helps the company in securing data and also helps in data sharing.

Genomic Data Sharing: Using genomic data and the blockchain network, it is now possible to send anonymous genetic information around the world. The decentralization nature of blockchain allows easy and secure data exchange between organizations. Information can be stored in a special ledger in a blockchain database that keeps the information secure.

Immutability of genomic data: Blockchain provides immutability of genomic data for organizations, helping organizations protect information. Due to the decentralized structure of blockchain technology, genomic data cannot be modified, so any changes will be reflected on all nodes, so no one cheats here, and it can be said that genomic data sharing is very safe.

Efficiency: The organization uses blockchain technology for efficiency reasons, as blockchain eliminates any third-party interference between genomic data sharing and errors, making the system more efficient and faster. As a result, sharing data becomes easier, smoother, and faster.

Cost Reduction: Since blockchain does not require a third person, it reduces costs for organizations and gives trust to other partners. Before blockchain technology, organizations spends lots of money, as they have to hire a third person to maintain all the things which blockchain technology does.

4.2.3 Use Case of Blockchain in Genomics

Blockchain technology offers several potential use cases within the field of genomics, primarily focusing on enhancing data security, integrity, and accessibility. Here's a use case scenario for blockchain in genomics:

Secure and Immutable Storage of Genomic Data

Genomic data is highly sensitive and valuable, containing personal health information and genetic makeup. However, the current storage and sharing mechanisms often lack sufficient security

measures, leading to concerns about data breaches, unauthorized access, and data tampering.

Use Case: Implementing blockchain technology for secure and immutable storage of genomic data can address these concerns effectively. Here's how it can work:

- **Data Encryption and Storage:** Genomic data is encrypted and stored on the blockchain network. Each participant (such as research institutions, hospitals, or individuals) holds a private key, enabling access to their specific data while maintaining confidentiality.
- **Immutable Data Records:** Every genomic data transaction, including storage, access, and sharing, is recorded as a block on the blockchain. Once added, these blocks cannot be altered or deleted, ensuring data integrity and preventing unauthorized modifications.
- **Access Control and Permission Management:** Smart contracts are utilized to manage access control and permissions. Individuals can grant or revoke access to their genomic data selectively, specifying who can view, analyze, or utilize their data for research or medical purposes.
- **Traceability and Auditability:** Blockchain's transparent and decentralized nature enables comprehensive traceability and auditability of genomic data transactions. Every action performed on the data, including accesses, modifications, and sharing, can be tracked back to its source, promoting accountability and trust.
- **Data Interoperability and Collaboration:** Blockchain facilitates secure data sharing and interoperability among different stakeholders in the genomics ecosystem. Researchers, clinicians, and pharmaceutical companies can collaborate more effectively while ensuring data privacy and security.

Benefits

- **Enhanced Data Security:** Blockchain's cryptographic techniques and decentralized architecture provide robust security against data breaches and unauthorized access.
- **Data Integrity:** Immutable data records prevent tampering, ensuring the integrity and reliability of genomic data.
- **Patient Empowerment:** Individuals gain greater control over their genomic data, deciding how and with whom it's shared, promoting privacy and autonomy.
- **Streamlined Collaboration:** Blockchain fosters seamless collaboration and data exchange among researchers and institutions, accelerating genomic research and medical advancements.

By leveraging blockchain technology, the genomics industry can address critical challenges related to data security, integrity, and accessibility, paving the way for innovative research, personalized medicine, and improved healthcare outcomes.

4.2.4 Blockchain Operations in Genomics

In the realm of genomics, blockchain operations entail a fusion of genetic data management and decentralized ledger technology. These operations encompass securely recording genetic information onto the blockchain, ensuring data integrity, and facilitating secure sharing and access among

authorized parties. Initial steps involve collecting genetic data from various sources, such as sequencing platforms or healthcare providers, and encrypting it to preserve privacy. Subsequently, this encrypted data is stored as transactions on the blockchain, leveraging its immutable and tamper-resistant nature to safeguard against unauthorized alterations. Validation mechanisms, including consensus algorithms and cryptographic techniques, authenticate the accuracy and legitimacy of transactions. Smart contracts play a pivotal role in automating processes such as consent management, ensuring data access rights are enforced transparently and autonomously. Through decentralized networks, stakeholders can securely exchange genetic information, fostering collaboration while maintaining data sovereignty and privacy. Overall, blockchain operations in genomics herald a paradigm shift in data management, offering unprecedented security, transparency, and interoperability in the handling of sensitive genetic information.

Chapter 5

Modeling Privacy Threats to Genomic Data

Privacy, is the right to be free from interference or intrusion — the ability to be left alone without being watched. Information privacy refers to the right to have control over how your personal information and data are collected, stored, and used. Organizations often impose privacy policies to protect sensitive information and safeguard user identities. Organizations dealing with sensitive user information are required to be transparent about what forms of data they intend to collect, the purpose of the data collection, and where and with whom it is to be shared.

Cyber security risks are associated with cybersecurity incidences that arise from the loss of confidentiality, integrity or availability. Privacy attacks are actions or inactions that cause a perceived privacy harm that do not solely involve cyber security violations. In other words there are events which may cause substantial privacy risk, although they are not necessarily related to any cyber security violation. Example: Washington Health Identification

Privacy risk is a function of privacy threats, privacy vulnerabilities, and privacy consequences. Privacy vulnerabilities are weaknesses or loopholes in a system that causes privacy harms.

Solove's taxonomy is an ontology of privacy harms. It describes four types of privacy harms:

1. Information Collection: Surveillance, Interrogation
2. Information Processing: Aggregation, Identification, Insecurity, Secondary Use, Exclusion
3. Information Dissemination: Breach of Confidentiality, Disclosure, Exposure, Increased Accessibility, Blackmail, Appropriation, Distortion
4. Invasions: Intrusion, Decisional Interference

A system that is designed in concert with an effective privacy threat model could better protect itself from privacy attacks.

5.1 Threat Modeling Approaches

5.1.1 STRIDE

STRIDE [1] was developed by Loren Kohnfelder and Praerit Garg in 1999 to identify potential vulnerabilities and threats to Microsoft products. This software centric (developer focused) threat model uses the mnemonic STRIDE to categorize threats as follows:

Threat Category	**Property Violated**	**Threat Description**
Spoofing	Authentication	Illegal access and then use of another user's authentication information, such as username and password
Tampering	Integrity	The malicious modification of data
Repudiation	Accountability (Audit)	Associated with users who deny performing an action without other parties having any way to prove otherwise
Information Disclosure	Confidentiality	Exposure of information to individuals who are not authorized to access it
Denial of Service	Availability	Attacks that deny service to valid users
Elevation of Privileges	Authorization	Privileged access gained by an unprivileged user, who then has sufficient access to compromise or destroy the entire system

Table 5.1: STRIDE Threat Model

In the STRIDE methodology, data flow diagrams of the application's use cases are created by decomposing them to identify system entities, events, and boundaries. Threat categories are then used to apply a general set of known threats to the system and assess the system for mitigations to these threats.

5.1.2 PASTA

Process for Attack Simulation and Threat Analysis (PASTA) is a risk-based methodology to threat modeling that takes an attacker-centric approach in identifying threats to an application. This methodology follows the seven-stage process shown in the diagram at left.

As PASTA follows an attacker-centric approach, it uses "attack trees" to depict potential attacks on a system in a tree-form diagram. The tree root symbolizes the goal of the attack, and the leaves show ways to achieve that goal. In order to leverage a threat modeling methodology that uses attack trees, one must be familiar with an attacker mindset and capabilities.

PASTA is seven step methodology of threat modeling with the objective of identifying threat, enumerating them and assigning a score so that organizations can deploy appropriate

5.1.3 DREAD

The DREAD threat modeling methodology was developed by Microsoft. The DREAD model quantitatively assesses the severity of a cyber threat using a scaled rating system that assigns

numerical value to risk categories. It has 5 categories:

- Damage: Understand the potential damage a particular threat is capable of causing.
- Reproducibility: Identify how easy it is to replicate an attack
- Exploitability: Analyze the system's vulnerabilities to ascertain susceptibility to cyber attack.
- Affected Uers: Calculate how many users woould be affected by a cyber attack.
- Discoverability:Determine how easy it is to discover the vulnerable points in the system infrastructure.

The DREAD model enables analysts to rate, compare, and prioritize the severity of threats by assigning a given issue a rating between 0 and 10 in each of the above categories. The final rating, calculated as the average of these category ratings, indicates the overall severity of the risk.

5.1.4 MITRE ATT&CK Framework

The ATT&CK framework (Adversarial Tactics, Techniques, and Common Knowledge) was developed by MITRE Corporation, a private non-profit company in USA [2]. It is a globally accessible database which includes attack tactics, techniques, and mitigation measures for three different domains: enterprise, mobile, and industrial control systems (ICS). This knowledge base may act as a foundation for the development of other threat models, methodologies and tools.

ATT&CK contains 14 categories of adversary tactics and techniques, 12 of which are shared across all matrices. The 14 tactics are:

- Reconnaissance
- Resource Development
- Initial Access
- Execution
- Persistence
- Privilege Escalation
- Defense Evasion
- Credential Access
- Discovery
- Lateral Movement
- Collection
- Command and Control
- Exfiltration
- Impact

There are two major components, tactics and techniques. Tactics are the objectives of an attacker, and techniques are the method by which the attacker achieves the tactical goals. Tactics and techniques are arranged as rows and columns of MITRE ATT&CK matrix. There are 14 tactics and 218 major techniques listed in MITRE ATT&CK matrix. Primarily, there are three categories of MITRE ATT&CK. ATT&CK for enterprise focuses on adversarial behavior in the enterprise system also covering cloud environments, Windows OS, and Mac OS, ATT&CK for mobile focuses on aggressive conduct, while the ATT&CK for enterprise matrix includes pre-ATT&CK, which focuses on "pre-exploit" hostile behavior.

In the ATT&CK matrix, mitigations are security concepts and a collection of instruments that stop a group of methods or sub- techniques from being effectively applied to a system or in an organization. Analysis of these mitigations, for different techniques, offers deep understanding of attack and help to identify the suitable security controls from NIST security and privacy controls. With deep analysis of the use case, possible attacks are identified in MITRE ATT&CK matrix for smart industrial system of smart firefighting. Mitigations of these attacks are analyzed to find prevention strategies and how attacks are linked with mitigations.

5.1.5 LINDDUN

Deng et al. [7] introduced the LINDDUN privacy threat modeling approach for software based systems. LINDDUN is an acronym and it stands for **L**inkability, **I**dentifiability, **N**on-repudiation, **D**etectability, Information **D**isclosure, Content **U**nawareness, Policy and Consent **N**on-compliance. LINDDUN is a comprehensive approach that helps analysts to reason about privacy concerns in a systematic and structured way. It introduces 7 main privacy threat types[1] captured in the acronym LINDDUN.

LINDDUN Threat Types

LINDDUN has seven privacy threat types: Linking, Identifying, Non-repudiation, Detecting, Data Disclosure, Unawareness and Non-compliance.

Linking: Combining, associating, or linking data or actions to learn more about an individual or group, because the combined data provides more information and more insight. Linking can occur with identifiers such as usernames, IP addresses, phone numbers, etc. By linking (or combining) these types of data, a bigger picture of the individual is developed. Further, profiling and other advanced methods can be used to link data together. Linking can be used on individuals and groups.

Example Linking threats:

- A website combining actions from an anonymous user and based on those actions, the website manipulates offerings, reductions, etc.
- Combining a location data set, with a company's data set, to infer which users work at which companies.

Identifying: Identifying an individual (using available data, or actions). Identification occurs when it is not possible to identify a person, yet by data leakage, deduction, or inference, it becomes possible. Identified data allows for direct linking of data to an identity. Identifiable data allows for indirect linking of data to an identity.

[1] https://linddun.org/threat-types/

Example Identifying threats:

- A website forces users to register with name and address, this allows the website to link all (previous and future) actions on the website to the name and address.
- An online forum links registered chats and posts (which have a user associated with them), with anonymous chats and posts by combining cookie data.

Non-repudiation: The ability to link or attribute a claim (or action) to an individual. It relates to the ability to attribute a claim or action to an individual. This prevents the ability of an individual to deny a claim or action. Non-repudiation is similar to the Repudiation threat in STRIDE. Although for security purposes, it is helpful to be able to link individuals to actions, whereas, for privacy purposes, it is helpful to not be able to link individuals to actions. Evidence used for repudiation includes logs, signatures, and metadata. This evidence links to individuals.

Example non-repudiation threats:

- Having read receipts on messages, allows for confirmation that an individual has read a message.
- Health tracking data improves understanding of the personal health matters of an individual, which can be used against the same individual in insurance claims, or against legal accusations.

Detecting: Determining the involvement of an individual based on observation. It relates to inferring whether the mere existence of data proves something sensitive, rather than the contents of the data. Observed communications indicate communication between systems or people. Observed application side-effects indicate changes in a system from an application, such as memory usage, and temporary files. Evoked system responses indicate data through certain responses from a system.

Example detecting threats:

- A deleted application from a computer or smartphone leaves behind data, indicating that the application once existed, with perhaps even contents of the application still available.
- Attempts to check whether certain users, or certain Email addresses exist within an application.

Data Disclosure: Extreme or excessive collection, storage, processing, and/or sharing of personal data.It relates to excessively collecting, storing, or processing personal data. Data should only be collected when it is absolutely required. There are four types of excessive data processing: unnecessary data types, excessive data volume, unnecessary processing, and excessive exposure.

Example data disclosure threats:

- When a company collects data for its services within a web application, additional data is collected but not used. This additional data could be used for future sales purposes.
- Analytics tools on websites and web applications collect excessive amounts of data points.

Unawareness: Limited or insufficient notification, involvement, or empowerment for individuals regarding their (processing of) personal data. It relates to lack of transparency, lack of feedback and interminability. Lack of transparency means that a person is unaware of personal data collection and processing. Lack of feedback means that a person is not sufficiently informed of privacy impact. Lack of interminability means that a person cannot access or manage their personal data.

Example unawareness threats:

- Users of an application are not informed of how their data is used, what data is collected, etc.
- Users of an application do not have access to their personal data and are not able to change it.

Non-compliance: Non-compliance to best practice security and data management.It relates to not upholding best practice security and data management. This can include unlawful processing of personal data, lack of data lifecycle management, or a lack of cybersecurity.

Example non-compliance threats:

- Lack of security hardening of key components that make up an application. Collecting or using data clearly against GDPR, within the EU.



Figure 5.1: LINDDUN Privacy Threat Types. Source: https://linddun.org/

LINDDUN is a privacy threat modeling framework that provides support to systematically elicit and mitigate privacy threats in software architectures. It was inspired by STRIDE Microsoft's threat modeling approach for security. STRIDE was developed more than 20 years ago at Microsoft as part of their security development lifecycle. LINDDUN is currently one of the most mature privacy threat modeling approaches. While also other (security) threat modeling approaches exist all follow the same four high-level steps described as four questions (1) What are you building? (2) What can go wrong? (3) What are you going to do about it? (4) Did you do an acceptable job? LINDDUN provides both process and knowledge support, i.e. each of the steps of the method are com-plemented with privacy-specific knowledge to aid the analysts. The knowledge is structured according to the threat categories encompassed in the LINDDUN acronym: linkability, identifiability,

non-repudiation, detectability, disclosure of information, unawareness, non-compliance. LINDDUN consists of the following three steps

Model the system: Typically a data flow diagram (DFD) is used, which is a very simple representation of the system consisting of only 5 building blocks: processes (i.e. computational units), data stores (i.e.passive containers of information), external entities (i.e. users or third party services), data flows (i.e.communication between the different components), trust boundaries (i.e. logical or physical borders).

Elicit threats: To elicit threats, you need to systematically go over each of the DFD elements or interactions and determine which threats are applicable. LINDDUN provides knowledge support in the form of privacy threat trees, which describe the most common attack paths for each combination of LINDDUN threat category and DFD element type. Each identified threat needs to be documented.

Manage threats: In this solution-oriented step, the threats are prioritized according to their risk, and then suitable mitigation strategies and corresponding privacy enhancing solutions can be selected to resolve each elicited threat.

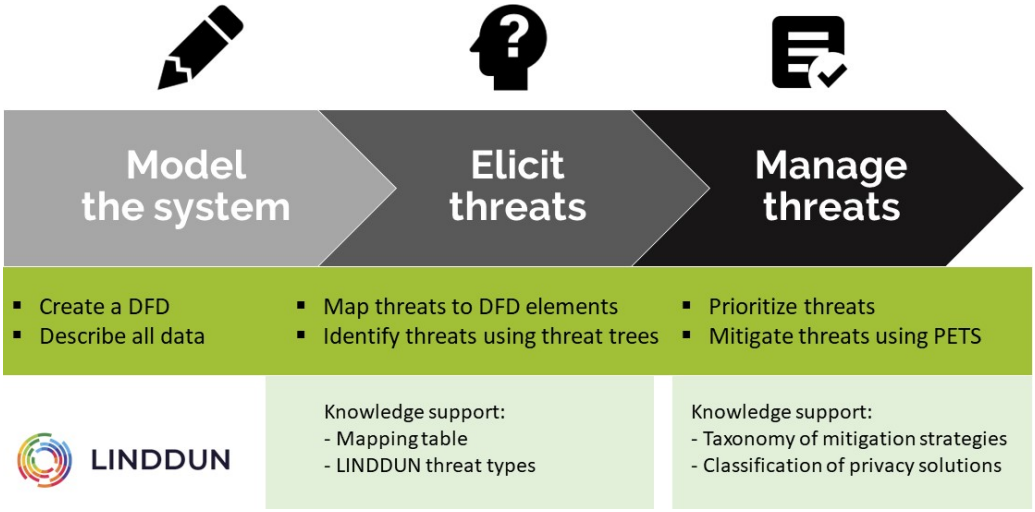


Figure 5.2: LINDDUN Privacy Threat Framework.

5.2 The LINDDUN Methodology

In this work, we propose a systemic approach for privacy threat modeling – the LINDDUN methodology to elicit the privacy requirements of genomic datas in tools and select privacy enhancing technologies accordingly. Each letter of "LINDDUN" stands for a privacy threat type obtained bynegating a privacy property. Privacy properties and threats types are briefly described in the following Sections respectively.

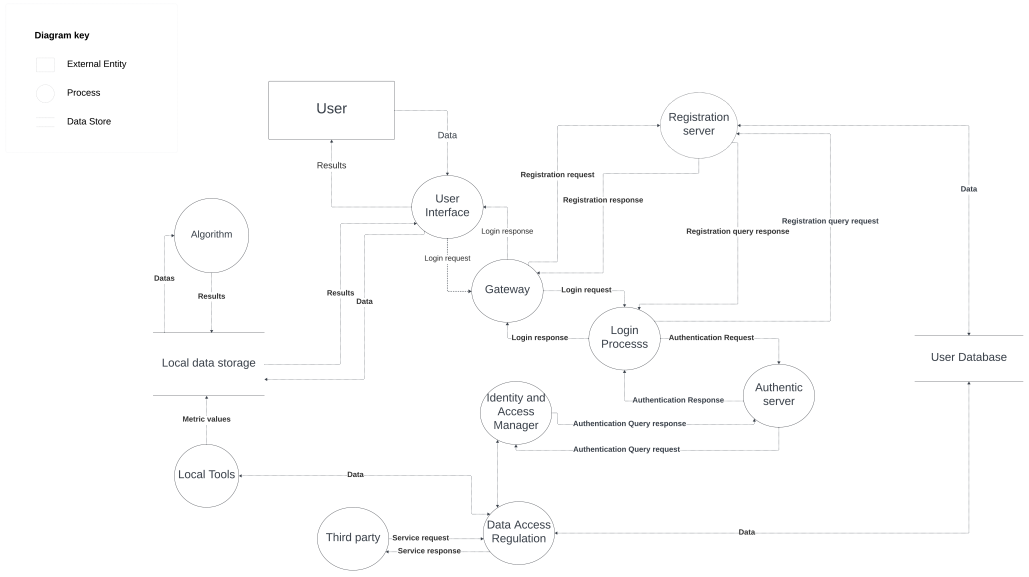


Figure 5.3: DFD of Genomic Data Processing Systems

From fig 5.3 The data flow diagram of bioinformatic tools is segmented into three distinct scenarios.

Scenario 1: Login Process Users input their data into the user interface, triggering a login request through the gateway. This request is directed to the login process, which authenticates via the authentication server. The authentication server verifies user identity through the identity and access manager. Upon authorization, the manager queries the data access regulation for access privileges stored in the user database. Access granted, the authentication process is confirmed back through the authentication server to the login process, ultimately providing access to the tools via the gateway.

Scenario 2: Registration New users initiate registration requests through the gateway, which are directed to the registration server. The registration server collects and stores user details in the user database for future login procedures.

Scenario 3: Tool Usage Following successful login, users can upload data to bioinformatics tools. Uploaded data is stored locally and analyzed using built-in algorithms. The results are then delivered to the user via the user interface. If a third party attempts access, the data access regulation verifies authentication against the user database, denying access if proper authentication is not obtained.

5.2.1 Mapping Privacy Threats to DFD

In this section, we present the privacy threat categories based on the above-mentioned privacy properties. We also discuss how to map these categories to the DFD elements.

Privacy Threat Categories

Privacy Properties	Privacy Threats
Unlinkability	Linkability
Anonymity and Pseudonymity	Identifiability
Plausible deniability	Non-repudiation
Undetectability and unobservability	Detectability
Confidentiality	Disclosure of information
Content awareness	Content unawareness
Policy and consent compilance	Policy and Consent Noncompilance

Table 5.2: LINDDUN Methodology: Privacy Properties and the Corresponding Privacy Threats

As shown in Table 5.2, the methodology considers seven types of threats. LINDDUN is the mnemonic acronym that we use.The following section describes LINDDUN components .

- **Linkability** of two or more items of interest (IOIs, e.g.,subjects, messages, actions, etc.) allows an attacker tosufficiently distinguish whether these IOIs are related ornot within the system.
- **Identifiability** refers to the ability of an attacker to adequately pinpoint the subject associated with an Identifier of Interest (IOI), such as the sender of a message. Typically, identifiability concerns a group of potential subjects known as the identifiability set. Essentially, identifiability represents a specific instance of linkability, particularly when both the subject and its attributes are involved. This concept poses a threat to both anonymity and pseudonymity, as it undermines efforts to conceal the identity of individuals or entities.
- **Non-repudiation**, unlike security concerns, poses a threat to privacy. It empowers attackers to accumulate evidence countering the repudiating party's claims, thereby establishing that a user possesses knowledge of, has performed, or has communicated something.
- **Detectability** of an Identifier of Interest (IOI) refers to an attacker's ability to adequately determine its existence. In the context of messages as IOIs, it implies that messages are distinguishable from random noise.
- **Information Disclosure** threats expose personal information to unauthorized individuals.
- **Content Unawareness** indicates that users are oblivious to information disclosed to the system. Users may inadvertently provide excessive information, facilitating an attacker's retrieval of their identity, or provide inaccurate information leading to erroneous decisions or actions.
- **Policy and consent Noncompliance** refers to instances where systems display privacy policies to users without necessarily adhering to them. Consequently, users' personal data may still be exposed despite the system's advertised policies.

5.2.2 Mapping Privacy Threat Categories

This section provides the guidelines to identify privacy threats of a genomic database. First, a Data Flow Diagram (DFD) is created in correspondence to the application's use case scenarios. Second, privacy threats are mapped to the DFD.

Creating Application DFD Based On Use Case Scenarios

DFD is chosen to represent a based on two reasons.First, DFD is proven to be sufficiently expressive in a number of case studies examined by the researchers. Second, DFD is also used by the SDL threat modeling process, hence by deploying the same modeling technique an interesting synergy can be created between the proposed framework and the SDL process.

Entity	User
Process	Registration server Authentic server Login process Identity and access manager Data access regulation Third party Gateway User Interface Algorithm Local tools
Data store	Local data storage User database
Data flow	Registration stream Authenticataion stream Login stream

Table 5.3: DFD elements in the Bioinformatic tools

Running Example: Bioinformatic Tools

In our running example Bioinformatic tools, Alice is a registered user of a bioinformatic tools. Each time Alice updates her data, she first connects to the tool's web portal. Accordingly, the portal communicates with the server, and eventually, the data information of Alice and all other data of a tool is stored in a database.

The DFD for the Bioinformatic tool application was already presented in Figure 5.3 of Section 5. Table 5.3 lists the DFD elements.

The creation of the DFD is an important part in the analysis. If the DFD was incorrect, the analysis results would be wrong as well. Since privacy focuses on the protection of user's information, it is important to consider where the information will be stored or passed by, as these are the crucial elements for building in privacy.

Mapping Privacy Threats to DFD

After the DFD elements are listed, we identify the privacy threat categories for each DFD element by following the mapping depicted in Table 5.4. Each intersection marked with the symbol × indicates a potential privacy threat at a corresponding DFD element in the system.

In essence, each DFD element is subject to certain privacy threats, and the nature of the potential privacy threat is determined by the DFD element type. For example, a data flow is subject to a number of privacy threats such as identifiability, linkability, detectability, non-repudiation, and information disclosure. The following sections will explain how privacy threats affect DFD elements. More threat scenarios corresponding to our running example will be discussed.

Threat categories	Entity	Data Flow	Data Store	Process
Linkability	x	x	x	x
Identifiability	x	x	x	x
Non-repudiation		x	x	x
Detectability		x	x	x
Information Disclosure		x	x	x
content Unawareness	x			
policy/consent Noncompliance		x	x	x

Table 5.4: Mapping LINDDUN components (privacy threats) to DFD element types (E-Entity, DF-Data flow, DS-Data store, P-Process)

The nature of linkability indicates that the threat affects DFD elements by pair. In other words, linkability of a DFD element refers to a pair (x1,x2), where x ∈(E,DF,DS,P) is the linkable IOI. Obviously, linkability at entity, from an attackers perspective means that within the system (comprising these and possibly other items), the attacker can sufficiently distinguish whether these entities are related or not. Similar description applies for that of data flow, data store, and process.

The **identifiability** threat affects all four DFD elements, such that each DFD element is made explicit as the attributes that identifiability (or its opposite property anonymity) re- lates to, by forming a pair with a subject. Essentially, iden- tifiability at each DFD element refers to a pair (x, y), where x ∈E is the identifiable subject, and y ∈E, DS, DF, P is the attribute identifiability relates to. For example, identifiability at entity refers to a pair (E, E), meaning to identify an entity within a set of entities. Identifiability at data flow refers to a pair (E, DF), meaning that a message is linkable to a potentially sending or receiving subject. Identifiability at data store refers to a pair (E, DS), meaning that a database entry is linkable to a potential data holder or subject. Iden- tifiability at process refers to a pair (E, P), meaning that a process is linkable to a potentially accessing subject.

Non-repudiation, opposite of plausible deniability, is a privacy threat that affects the DFD elements of data flow, data store and process. Non-repudiation might be appreciated for some system but undesirable for others. It depends on the system requirements. For e-commerce applications, non-repudiation is an important security property. Imagine a situation where a buyer signs for a

purchased item upon receipt, the vendor can later use the signed receipt as evidence that the user received the item. For other applications, such as off-the-record conversations, participants may desire plausible deniability for privacy protection such that there will be no record to demonstrate the communication event, the participants and the content. In this scenario, non- repudiation is a privacy threat. Even though entity is the only DFD element being able to (non-)repudiate, the non- repudiation privacy threat actually occurs at data flow, data store, and process. Similar to linkability and identifiability, non-repudiation at each DFD element refers to a pair (x, y), where x $\in$E is the non-repudiating subject, and y$\in$DS, DF, P is the attribute it relates to.

Detectability threats occur at data flow, data store, and process, meaning that the attacker can sufficiently distin- guish whether it exists or not. Though in some applications, techniques such as covert channel and steganography can be used to protect both messages (data flow) and communicating parties (entity), in this case the threat actually occurs at data flow instead of entity. In other words, the asset we want to protect against the detectability threat includes data flow, data store, and process.

Information disclosure threats affect data flow, data store, and process, referring to the exposure of information at these DFD elements to individuals who are not supposed to have access to it.

The **content unawareness** threat is related to entity, since the entity (data subject or data controller) is actually responsible to provide the necessary consents to process personal data and update or delete the expired information.

Policy and consent noncompliance is a threat that affects system as a whole, because each system component (including data flow, data store and process) is responsible to ensure that actions are taken in compliance with privacy policies and data subject's consents.

Running example: Bioinformatic Tools

Threat Target	L	I	N	D	D	U	N
User		x					
App Library	x						
User Interface	x	x			x		
Gateway	x	x	x	x	x	x	x
Registration Server	x						x
User Database	x						x
Login server	x	x	x	x	x	x	x
Authentication Server							
Identity and Access Manager							
Data Access Manager							
Third Party	x	x	x	x	x	x	x
Algorithm							
Local Data Storage							
Local Tools							

Table 5.5: Determining privacy threats for DFD elements within the Bioinformatic application (From left to right: L-Linkability, I-Identifiability, N-Non Repudiation, D-Detectability, D-Information Disclosure, U-Content Unawareness, N-Consent/policy Noncompliance)

Considering the Bioinformatic application, the list of generic privacy threats to the modeled

system is depicted in Table 5.5. This is obtained by gathering the elements from Table 5.3 and then determining the susceptible threats with Table 5.4.

Primarily, we assume that DFD elements within the trust boundary (marked as dashed line in Figure 5.1) are trustworthy. We trust the processes within the boundary, as well as all data flows in the trust boundary. Therefore, we will not discuss linkability, identifiability, and information disclosure threats on these elements. We however do not trust the user and its communication with the portal and we also want to protect the data store containing all the user's information.

Moreover, non-repudiation and detectability threats are considered irrelevant for social networks. Presumably, it depends on what privacy properties are required for a particular social network system. In case plausible deniability and undetectability would be desirable for a certain application, we should still consider these threats for each DFD element accordingly.

5.2.3 Detailing Privacy Threats via Threat Tree Patterns

This section provides a comprehensive collection of threat tree patterns, offering detailed insights into the privacy threats that a real-world system may face. Each intersection marked in Table 5.4 corresponds to a specific threat category, for which a corresponding threat tree pattern delineates the precise preconditions necessary for the threat's realization. These preconditions represent vulnerabilities that could be exploited in a privacy attack scenario.

Drawing from the latest advancements in privacy research, this catalog encompasses common attack patterns reflected in the threat trees, aiding application designers in considering privacy implications within their systems. However, it's important to note that the threat trees presented here represent the current best efforts and are subject to ongoing refinement to accommodate newly identified threats. Moreover, the catalog is intended to evolve continuously, incorporating insights from industrial validation experiments as they become available [7].

Linkability of Entity

Linkability of entity refers to an attacker can sufficiently distinguish whether two or more entities are related or not within the system. This implies that different pseudonyms can be linked to each other. The threat tree pattern is depicted in Figure 5.4. One precondition is that data flow or data store is not fully protected (e.g. unencrypted), which leads to the Information Disclosure threat of data flow and data store. The second precondition is that Personal Identifiable Information (PII) can be linked, e.g. based on user temporary ID, IP address, behavioral patterns such as time, frequency and location, session ID, identifier and biometrics, computer ID, communication content or any combination of these factors. The aforementioned data store refers to the identity management system's database or any other database which contains personal identifiers of users. Having accessed such a data store, the attacker could easily link different pseudonyms to the same user.

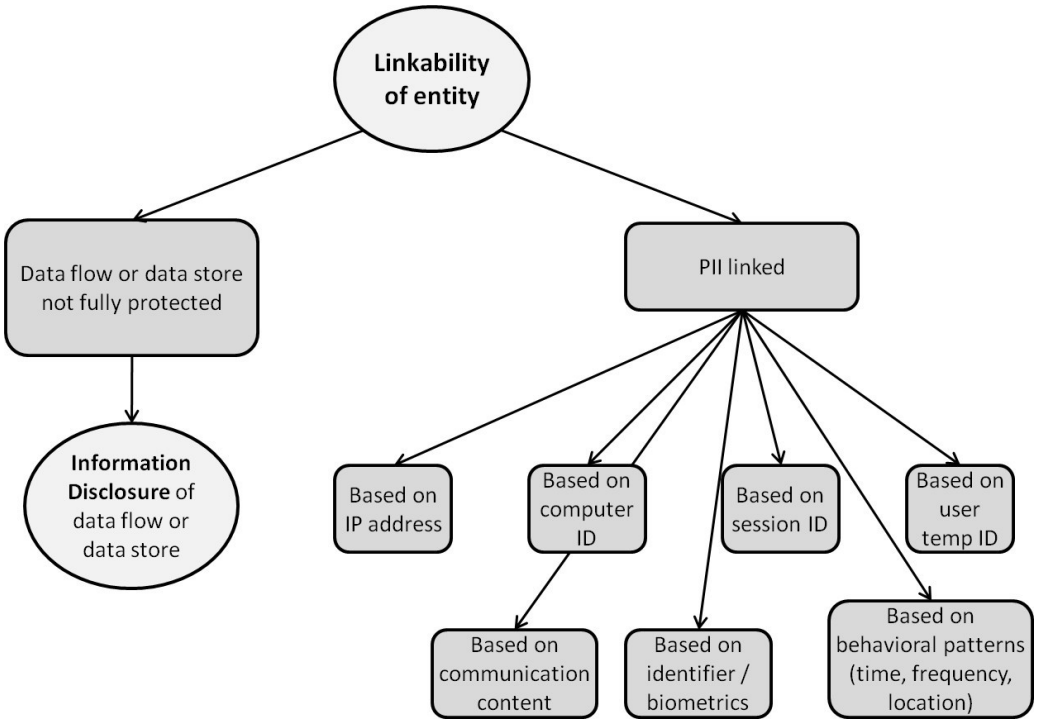


Figure 5.4: Threat tree for linkability of an entity.

Linkability of Data Flow

Linkability of data flow threat tree, as presented in Figure 5.5, suggests two preconditions. One precondition is that data flows are not fully protected (e.g. unencrypted), which leads to information disclosure of data flow; and communications are linkable due to little or insecure anonymity systems de- ployed. The other precondition is that communication can be linked. When no anonymous communication is deployed, basically the same preconditions apply as for the linkability of entity threat. Messages are linked to each other by user's identifiable information (e.g. based on user temporary ID, IP address, behavioral patterns such as time, frequency and location, session ID, identifier and biometrics, computer ID, communication content or any combination of these factors). Alternatively, when an insecure anonymity system is deployed, traffic analysis is possible to extract information out of patterns of traffic; passive attacks (e.g. longterm intersection attacks, traffic correlation and confirmation, fingerprinting, epistemic attacks (route selection), and predecessor attacks) and active attacks (e.g. N-1 attacks, Sybil attack, traffic watermarking, tagging attack, replay, and DoS attack) are possible to link entities together. An overview of these attacks can be found in [6].

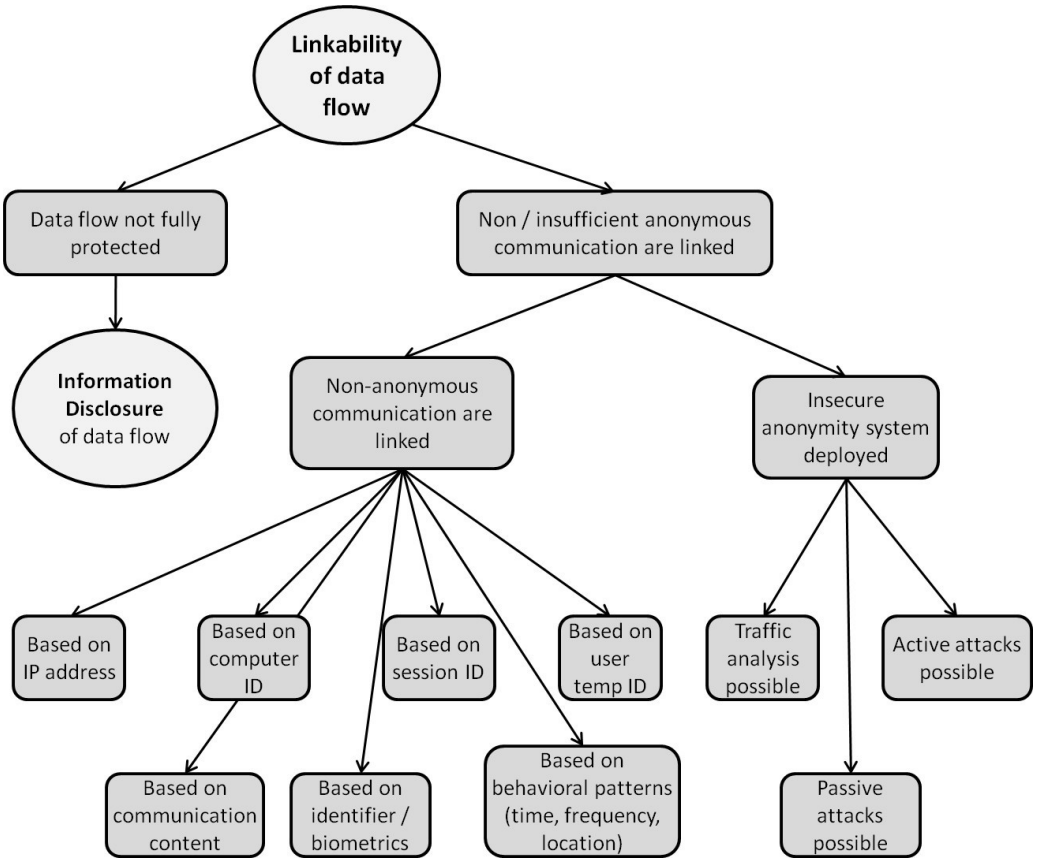


Figure 5.5: Threat Tree for Linkability of a Dataflow

Linkability of a data store Two preconditions correspond to the threat of linkability of a data store, as shown in Figure 5.6. First, there is insufficient access control of the data store leading to the information disclosure threat at a data store. Second, insufficient data anonymization is applied or strong data mining is possible in the data store, meaning that the stored information still contains sufficient references to the corresponding data subject, which makes it possible to link different data items to each other within the same database. Another possibility is that data can be linked from one database to another, and hence reidentification [16] is possible.

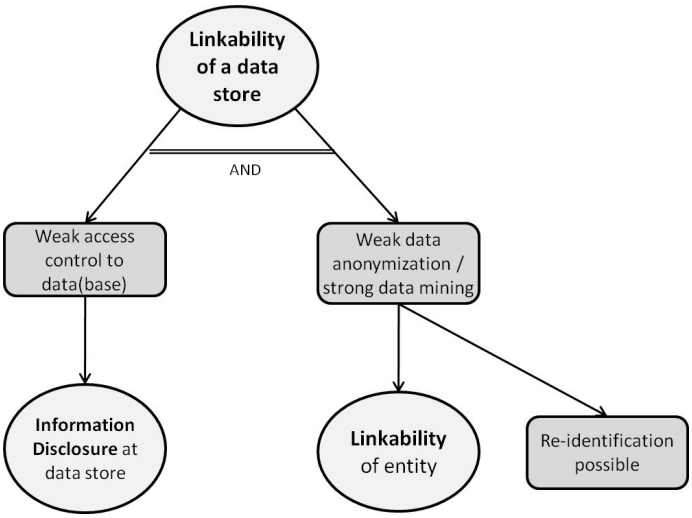


Figure 5.6: Threat Tree for Linkability of a Datastore.

Linkability of Process The threat tree of linkability of process suggests that the only way to prevent different actions being linked to the same subject is by gaining access to the process, as illustrated in Figure 5.7.

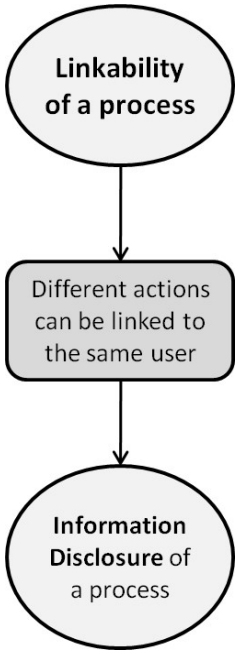


Figure 5.7: Threat tree for linkability of a Process.

Identifiability of Entity

Figure 5.8 shows the threat tree pattern for identifiability of entity. It gives an overview of the most common situations where an identifiability threat can occur at an entity. A first precondition is when the e-id is used as login (i.e., the user's actual identity is used), meanwhile the data flow between the user and login portal is not sufficiently protected (i.e., the threat of information disclosure of data flow), and thus the user's identity will be exposed. A second possibility occurs when a secret (e.g. a password) is used as login and the relationship between this secret and the user can be revealed. This can happen if the identity management database in not secure (e.g. passwords are stored in clear); if the communication channel is insecure and the communicated passwords are weak and can be connected to the user (e.g. using a birthdate as password); or if replay attacks are possible (e.g. a keylogger is installed, the communication can be eavesdropped or the person entering the secret can be observed). A third possible precondition for the threat is the use of a token as login which is weakly implemented or physically insecure. The final precondition is that biometrics is used as login, which means that biometrics is retrievable and can be linked to an entity. It is due to information disclosure of identity management database or data flow which contains biometrics, and linkability at data store.

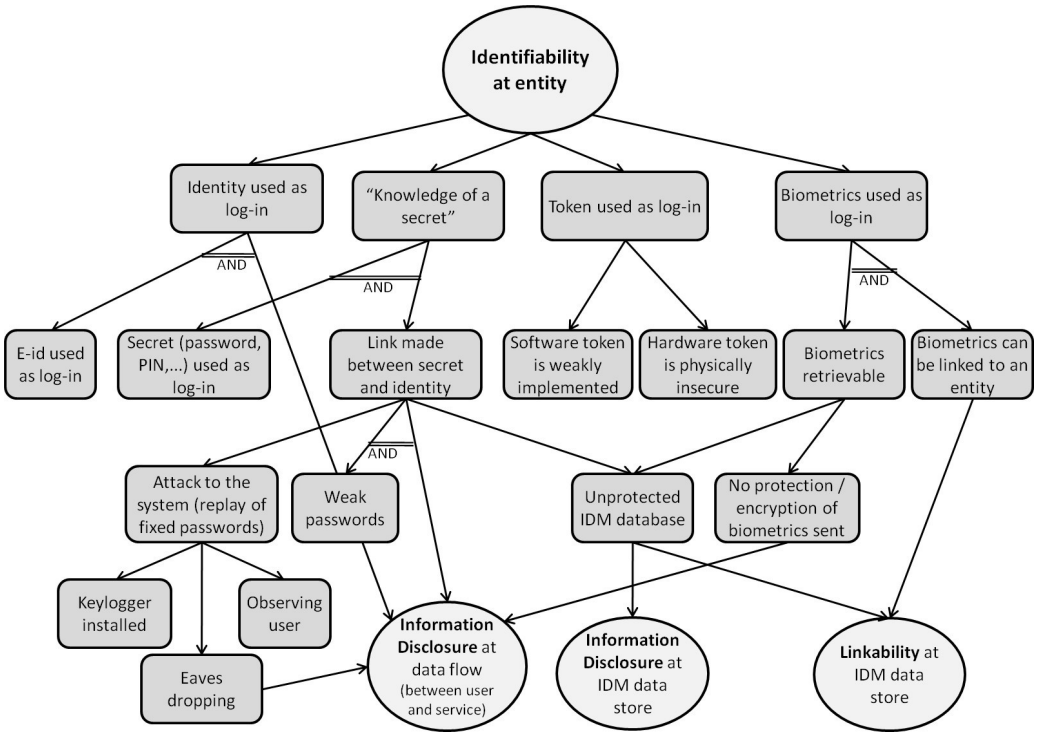


Figure 5.8: Threat Tree for Identifiability of An Entity.

Identifiability of Data Flow

The threat tree of identifiability of data flow is presented in Figure 5.9. Similar to the linkability of data flow threat tree, identifiability of data flow is possible when data flows are not fully protected,

which leads to information disclosure of data flow; or when the communication can be traced to an entity due to little or insecure anonymity system deployed. When no anonymity system is deployed, communication can be traced to an entity by means of identifiable information (e.g. based on user temporary ID, IP address, behavioral patterns such as time, frequency and location, session ID, identifier and biometrics, computer ID, communication content or any combination of these factors). Alternatively, when an insecure anonymity system is deployed, traffic analysis, passive attacks, and active attacks are possible to identify the particularly entity of interests.

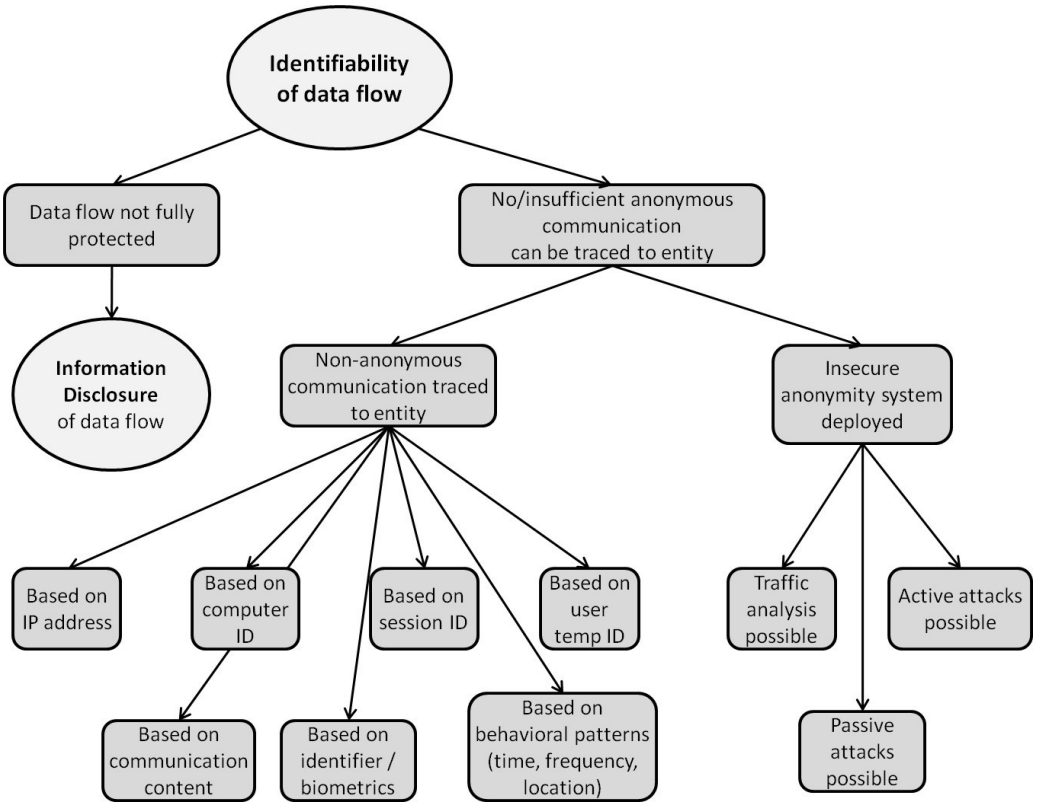


Figure 5.9: Threat Tree for Identifiability of a Dataflow.

Identifiability of Data Store

The preconditions for identifiability of data store, as presented in Figure 5.10, are similar to the preconditions of linkability at a data store. Either there is insufficient access control of the data store which refers to the information disclosure threat of a data store; or insufficient data anonymization or strong data mining techniques where applied in the data store, which means that the information stored still contains sufficient references to the corresponding person to reveal the person's identity, and hence it refers to the identifiability threat of entity or the data can be linked with other relevant information which can lead to reidentification of the data subject.

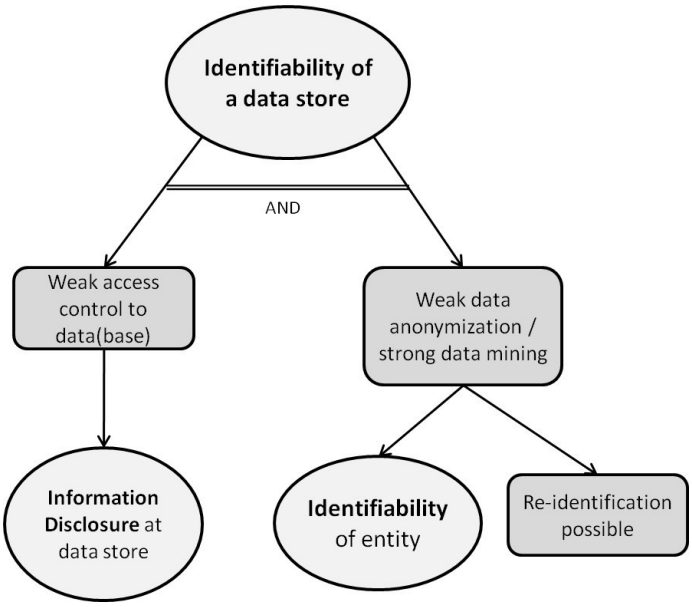


Figure 5.10: Threat Tree for Identifiability of a Datastore.

Identifiability of Process

The threat tree of identifiability of process is presented in Figure 5.11. The only condition for the identifiability threat at a process is when access to the process is not sufficiently secured, and thus it refers to the threat of information disclosure of a process

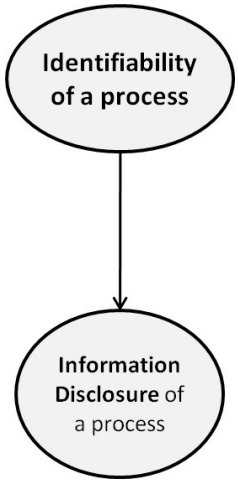


Figure 5.11: Threat Tree for Identifiability of a Process.

Non-repudiation of Data Flow

Four general preconditions can be applied to the threat tree of non-repudiation of data flow, as

presented in Figure 5.12. One condition is insufficient obfuscation for data sources or data flows, which means that the attacker can gain access to at least part of the data flow or data source. This can occur in a number of cases, for example, there is no automatic replay of broadcasts, such that the sender of a file is sufficiently distinguishable from those who are merely relaying it. Another example is when a complete decrypted log of all network connections to and from a user's computer is disclosed, resulting in the disclosure of the origin of data flow. The final examples are that there is insufficient protection against censors or insufficient obfuscation of data extensions, such that operators or users of the network are able to know where the data comes from.

The second precondition of this threat is that little or a weak deniable encryption technique is used to protect data flow. One possible attack path is to prove data is encrypted, either due to the encrypter proves the data is obviously an encryption or colluding users prove together that the data is encrypted. The second attack path is to prove data can be decrypted to a valid plain text, which can occur when the encrypter decrypts the file or colluding users can cooperate and show the decrypted message. The third attack path shows that all private keys are disclosed, and the last path suggests that cryptanalysis is possible to attack the used encryption scheme.

The third condition is that there are little or weak message authentication codes (MAC) used to ensure integrity of data flow content, such that an attacker can forge authentic looking messages and pretend that a certain data flow comes from a subject.

The final precondition indicates that there is little or a weak Off-the-Record Messaging (OTR) used, such that in a conversation it is not possible to provide both deniability for the conversation participants and confidentiality of conversations content at the same time. Possible attack paths include replaying of previous transferred messages, and the use of signatures to demonstrate communication events, participants and communication content.

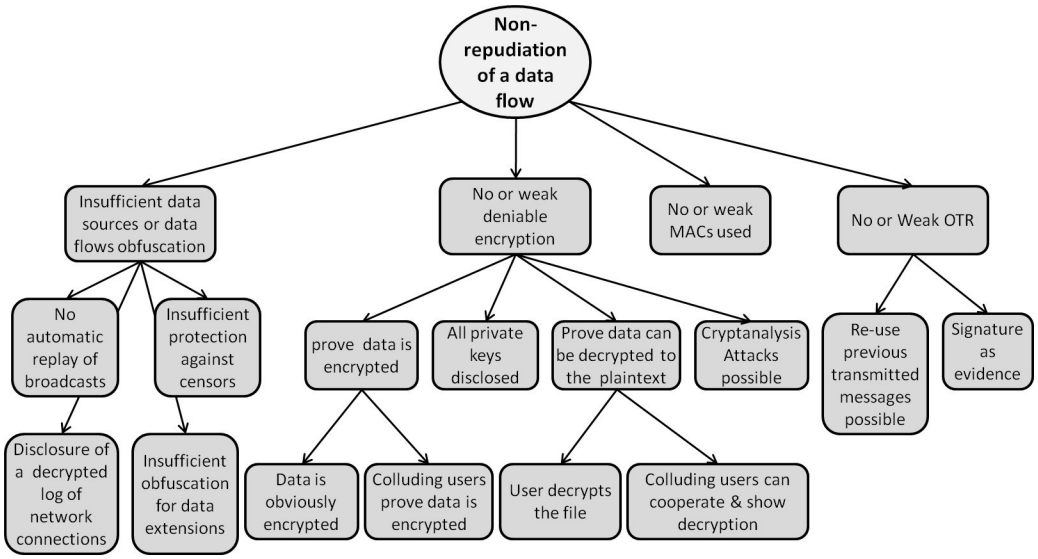


Figure 5.12: Threat Tree for Non-Repudiation of Data Flow.

Non-Repudiation of Data Store

The threat tree for non-repudiation of data store is depicted in Figure 5.13. Three preconditions

can apply to this threat, namely a weak access control to the database, which leads to the threat of information disclosure at the data store; little or a weak deniable encrypted is used to protect the data, such that data can be proven to be an encryption or can be decrypted to a valid plaintext; and subjects with deniability are not able to edit data in the database to cover their tracks, and it can be either impossible to remove or alter the user's own data or impossible to remove or alter someone elses data concerning the user himself.

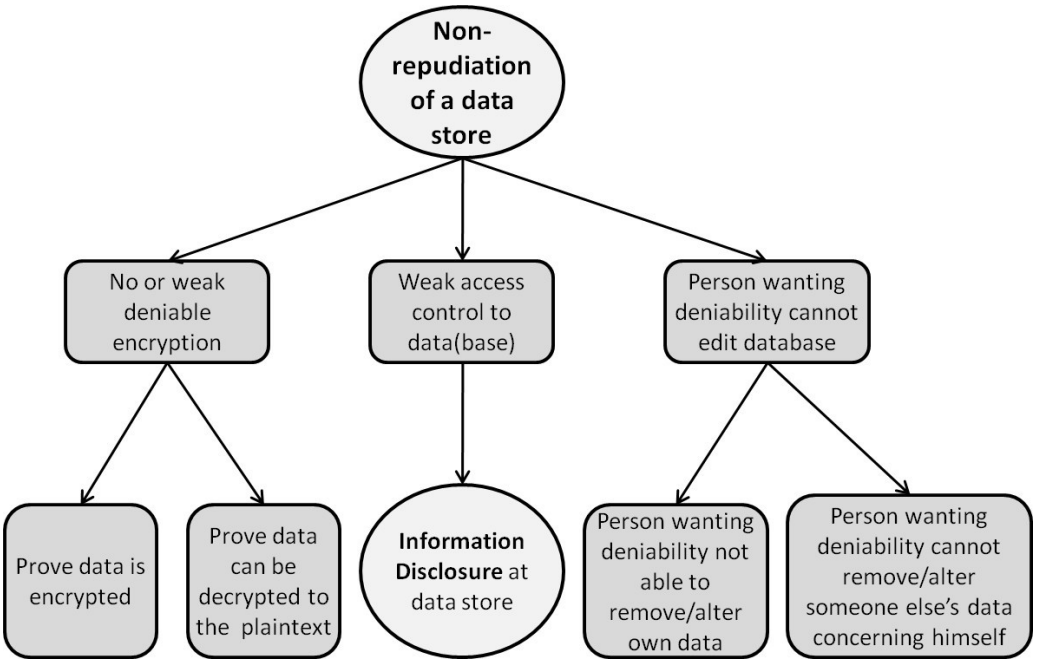


Figure 5.13: Threat Tree for Non-Repudiation of Data Store.

Non-Repudiation of Process Non-repudiation of process, as depicted in Figure 5.14 can be achieved in two ways: either the process looses its confidentiality and information disclosure attacks at the process are possible, or the process uses a secure log to create an overview of all actions, which can evidently be traced back to the user.

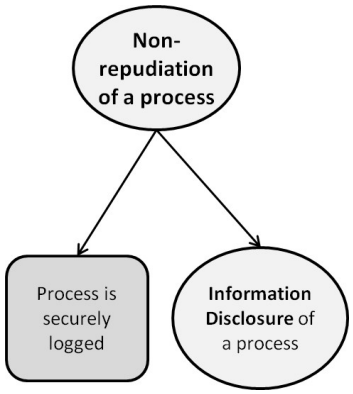


Figure 5.14: Threat Tree for Non-Repudiation of Process.

Detectability of Data flow The threat tree of detectability of data flow, as depicted in Figure 5.15 suggests five preconditions for the threat to occur. One condition is that the system lacks a covert channel. This can happen when the covert channel uses too much bandwidth from a legitimate channel, resulting in the detection of the covert communication. It can also be because the patterns or characteristics of the communications medium of the legitimate channel are controlled or examined by legitimate users, e.g. checking file opening and closing operations patterns or watching the timing of requests, such that covert communication is detected. The second condition is side channel analysis on timing information, power consumption, electromagnetic leaks, as an extra source of information which can be exploited to detect the communication. The third condition occurs when a weak information hiding techniques are used, which makes a number of steganalysis attacks possible. Another condition is when there is no or insufficient dummy traffic sent at some lower layer of communication network, such that messages fail to appear random for all parties except the sender and the recipient(s). Last but not least, the threat can occur because of a weak spread spectrum communication, resulting in deficiencies in the establishment of secure communications, resistance to natural interference and jamming, and detection prevention.

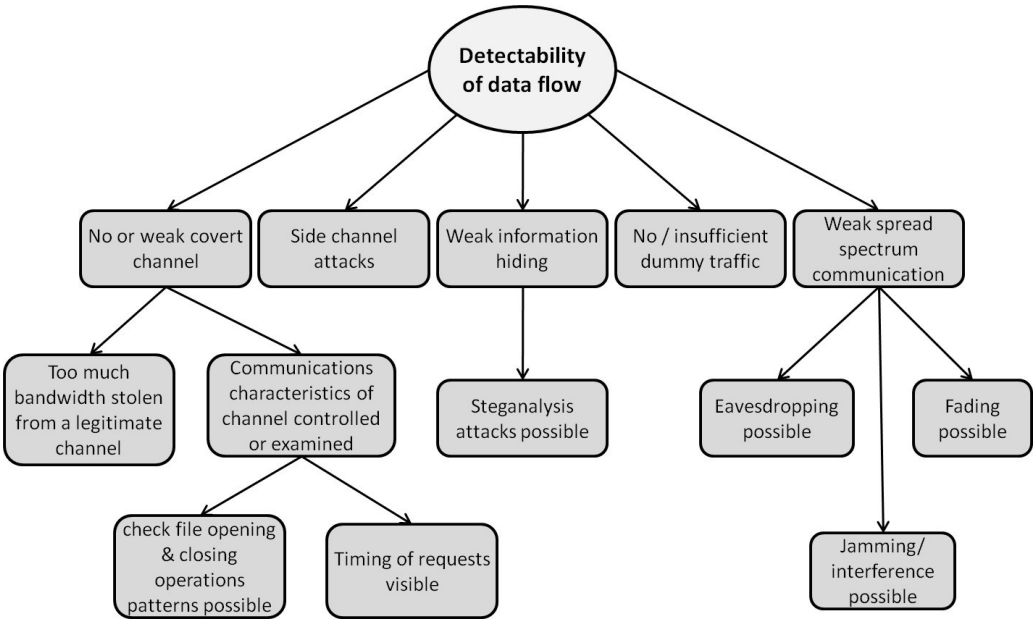


Figure 5.15: Threat tree for detectability of data flow .

Detectability of Data store As shown in Figure 5.16, detectability threats in a data store can occur if there is insufficient access control, which leads to the information disclosure threats for security, or if insufficient information hiding techniques are applied, such as information from a data store is revealed due to weak steganography algorithms employed.

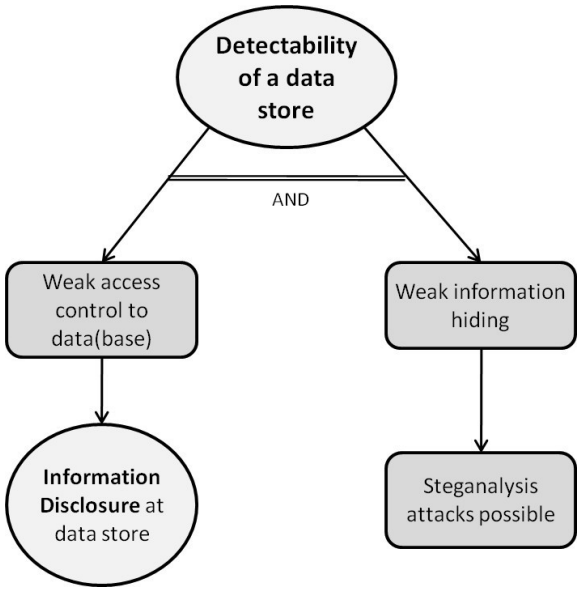


Figure 5.16: Threat Tree for Detectability of Data Store .

Detectability of Process Similar to the previously described threats related to a process, the detectability of process threat, depicted in Figure 5.17, also refers to the threat of information disclosure of process.

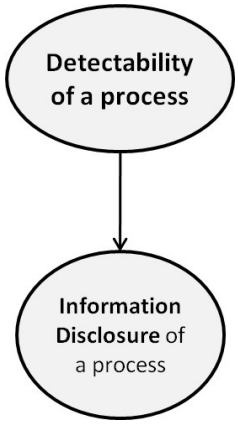


Figure 5.17: Threat Tree for Detectability of Process.

Information Disclosure of Data Flow, Data Store, and Process

The threat tree concerning information disclosure of data flow, data store, and process, depicted in Figure 5.18, refers to the security threat tree of information disclosure. This illustrates the fact that privacy properties are part of security properties, and privacy may depend on security.

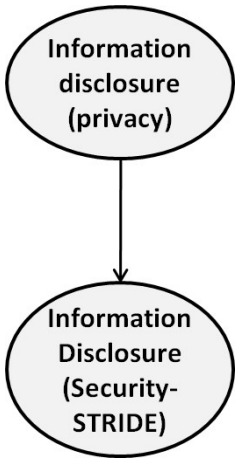


Figure 5.18: Threat Tree for Information Disclosure.

Content Unawareness of an Entity Content unawareness of an entity can occur in two situations: either the date subject provides more personal identifiable information than required, which has a negative influence on all the hard privacy objectives; or data subject does not keep information updated or does not remove the outdated information, and it can lead to wrong actions when decisions are based on this information.

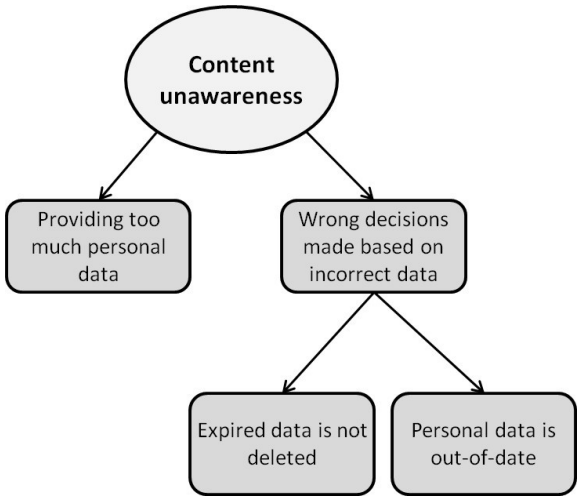


Figure 5.19: Threat Tree for Content Unawareness.

Consent and Policy Noncompliance of the System (Data Flow, Process and Data Store) The user's privacy can be violated when internal system rules do not correspond to privacy policies provided to the user. This can occur when an attacker tampers with the internal policies, which is actually a security threat; or when the policy rules are incorrectly managed or updated (according to the user's requests) by the system administrator.

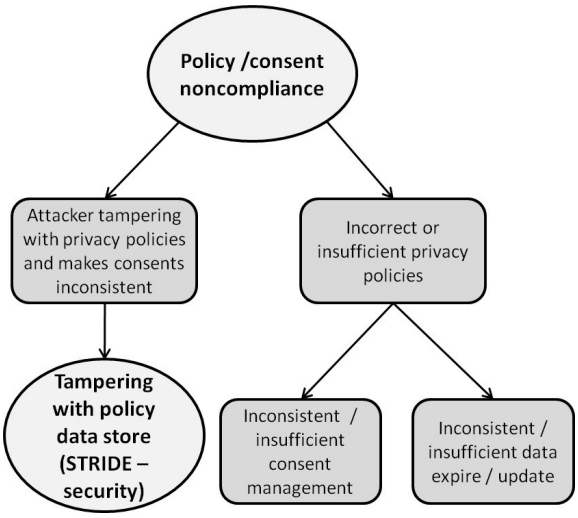


Figure 5.20: Threat Tree for Consent and Policy Noncompliance.

5.2.4 Privacy Requirements Analysis

Common privacy threats in the bioinformatic tools have been identified. However, the appropriate countermeasures to mitigate these threats are essential for the complete implementation of PE in this system. The following section presents common PET techniques to mitigate privacy threats in the bioinformatics.

Mylopoulos et al. [23] are the first to point out that complexity of an information system is determined partly by its functionality (i.e. what the system does) and partly by its nonfunctional requirements (also referred as constrains, goals, and quality attributes), and the nonfunctional requirements "play a crucial role during system development, serving as selection criteria for choosing among myriads of decisions". They proposed a comprehensive framework for representing and using non-functional requirements during the development process in a process oriented approach. The framework consists of five basic components – goals, link types, methods (i.e., goal decomposition methods, goal satisficing methods, and argumentation methods), correlation rules, and the labeling procedure – as the representation of non functional requirements in terms of interrelated goals. As suggested by Mylopoulos et al., "such goals can be refined through refinement methods and can be evaluated in order to determine the degree to which a set of non-functional requirements is supported by a particular design" [23]. This framework can serve as the foundation for the aforementioned privacy requirement analysis methodology.

The privacy guidelines provided by Microsoft describes some basic privacy concepts [88], such as different types of consents or data minimization concepts. Besides, a number of guidelines are presented for selected scenarios concerning the following principles: notice, choice, onward transfer, access, security, and data integrity. However, it only contains a flat list of the required and recommended guidelines and does not intend to describe a more structured approach. These guidelines can still be used as inspiration to determine possible threats, e.g. to extend our catalogue of threat trees.

In the documentation of SDL version 3.2 [6], privacy is also partially considered, but only at a

generic level. For example, the threat modeling process described in the forth stage of SDL only mentions that a design review with the privacy expert is necessary. SDL also presents ten general privacy guidelines. An example guideline indicates that it is important to collect the least sensitive form of data. At this stage, privacy is not yet well integrated in SDL.

Yu and Cysneiros [31] presented a framework using i , an agent oriented requirements modeling language, to deal with privacy requirements. This framework however focuses on reasoning about privacy and does not provide a structured methodology to examine the different privacy objectives.

Liu et al. [20] proposed a framework also using to deal with security and privacy requirements, which was inspired by the work of Yu and Cysneiros. They use four different analysis techniques to create a complete model: attacker analysis, dependency vulnerability analysis, counter measure analysis and access control analysis. Although this framework contains the attacker analysis technique which is similar to our idea of examining the possible privacy threats, the framework is again mainly meant to reason about privacy but lacks the necessary knowledge to empower the (non-expert) privacy analyst.

Miyazaki et al. [22] defined a computer-aided privacy requirements elicitation technique. This technique returns the appropriate requirements for the system to be compliant with the law, based on a questionnaire that the system engineer fills out. This is in contrast to our methodology, which helps the analyst eliciting the privacy requirements in accordance to the stakeholders' wishes.

5.2.5 From Privacy Requirements to Privacy Solutions

Several taxonomies have been proposed to create a link between privacy requirements and privacy solutions. This section gives an overview of some existing taxonomies which can be integrated in our threat modeling process to link the privacy requirements obtained by our methodology to the optimal privacy enhancing solutions.

The taxonomy of privacy goals, described by Anto´n et al. [4], is to analyze website privacy requirements. Privacy goals are divided into protection goals and (anti) vulnerability goals. The Code for Fair Information Practices is used to categorize the protection goals, namely notice and awareness, choice and consent, access and participation, integrity and security, and enforcement and redress. The (anti) vulnerability goals are classified according to the manner in which they violate the users privacy. The corresponding goals are monitoring, aggregation, storage, and information transfer. These goals are used to analyze and compare privacy policies and do not intend to link privacy requirements to general privacy solutions, instead, limited to website privacy requirements.

The Pris method presents a structured way to create systems which adhere to the specified privacy requirements. It consists of four different phases:

- Elicit privacyrelated goals,
- Analyze the impact of privacy goals on organizational processes,
- Model affected processes using privacy- process patterns,
- Identify the techniques that best support or implement the above processes.

The last step uses a table that classifies privacy implementation techniques in six categories, a) Administrative tools, b) Information tools, c) Anonymizer products, services and architectures, d) Pseudonymizer tools, e) Track and evidence erasers, and f) Encryption tools, and it can be an interesting source of inspiration to determine appropriate privacy solutions. Wuyts et al. [30] created a hierarchical taxonomy which categorizes privacy into objectives. The objectives are divided

in two branches. The proactive branch focuses on concealing the association between the data and the identity of the user before it is shared with the system, while the reactive branch focuses on guarding the relationship between the data and the identity when the data is already shared. Each objective corresponds to a number of strategies, which are a sub-classification of the objectives. These strategies can then be linked to their corresponding solutions. This paper only provides a sample solution for each category and does not provide a full overview of all existing solutions.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we have introduced a comprehensive framework designed to systematically model privacy threats within genomic data processing systems, elicit privacy requirements, and propose privacy-enhancing countermeasures. The primary contribution lies in the development of a structured methodology specifically tailored for modeling privacy threats. This methodology entails defining a comprehensive list of privacy threat types and establishing mappings to the various elements within the system model. Through this systematic approach, we aim to provide a clear and comprehensive understanding of the privacy risks inherent in the systems.

Furthermore, our work introduces an extensive catalogue of privacy-specific threat tree patterns, which serve as a supporting body of knowledge. These threat tree patterns offer a structured framework for analyzing and visualizing privacy threats, facilitating a more systematic and comprehensive threat assessment process. Additionally, our research provides a mechanism for mapping commonly known privacy-enhancing technologies (PETs) to the identified privacy threats and elicited privacy requirements. This integration enables stakeholders to identify suitable countermeasures to mitigate privacy risks effectively.

It is important to note that our framework is designed to be dynamic and adaptable to evolving privacy landscapes. As new threats continue to emerge and privacy technologies evolve, the threat tree patterns and catalog of PETs will be continuously updated and refined. By providing a flexible and evolving framework, we aim to empower stakeholders to effectively address the ever-changing privacy challenges within software-based systems, ultimately enhancing privacy protection and promoting trust in digital ecosystems.

6.2 Future Work

In our future endeavors, we are committed to extending the application of our proposed framework by undertaking larger case studies, with a notable focus on validating its efficacy within practical contexts. A significant aspect of this future work involves the validation of our framework within the intricate framework of a national e-health system.

This validation process encompasses the thorough application of our framework to analyze and

assess the privacy threats prevalent within a comprehensive national e-health infrastructure. Such systems, inherently complex due to their extensive data exchanges and management of sensitive healthcare information, pose multifaceted privacy challenges. By deploying our framework in this domain, we aspire to gauge its adaptability, scalability, and effectiveness in identifying and mitigating privacy risks within a real-world, high-stakes setting.

To accomplish this, several pivotal steps will be undertaken. Initially, a meticulous analysis of the national e-health system's architecture, data flows, stakeholders, and existing privacy protocols will be conducted. Subsequently, leveraging our framework, we will systematically model the specific privacy threats inherent in the system, categorizing them and mapping them to relevant system components. Based on these insights, we will then proceed to elicit privacy requirements, defining the necessary controls and policies to address identified risks effectively.

Furthermore, we will propose and evaluate privacy-enhancing countermeasures tailored to the system's needs, encompassing techniques such as encryption, access controls, and anonymization. The validation process will culminate in a comprehensive assessment of our framework's performance, gauging its ability to accurately identify threats, propose viable countermeasures, and enhance privacy protection within the national e-health system.

Bibliography

[1] Microsoft threat modeling tool. `https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool`. Accessed: 2023-10-01.

[2] The mitre attck framework. `https://attack.mitre.org/`. Accessed: 2023-10-01.

[3] M. Alghazwi, F. Turkmen, J. Van Der Velde, and D. Karastoyanova. Blockchain for genomics: A systematic literature review. *Distrib. Ledger Technol.*, 1(2), dec 2022.

[4] A. I. Antón, J. B. Earp, and A. Reese. Analyzing website privacy requirements using a privacy goal taxonomy. RE '02, page 23–31, USA, 2002. IEEE Computer Society.

[5] L. Bonomi, Y. Huang, and L. Ohno-Machado. Privacy challenges and research opportunities for genomic data sharing. *Nature genetics*, 52(7):646–654, 2020.

[6] G. Danezis, C. Diaz, and P. Syverson. Systems for anonymous communication. *CRC Handbook of Financial Cryptography and Security*, pages 341–390, 01 2010.

[7] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requir. Eng.*, 16(1):3–32, mar 2011.

[8] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*, pages 17–32, 2014.

[9] J. Gitschier. Inferential genotyping of y chromosomes in latter-day saints founders and comparison to utah samples in the hapmap project. *The American Journal of Human Genetics*, 84(2):251–258, 2009.

[10] M. T. Goodrich. The mastermind attack on genomic data. In *2009 30th IEEE Symposium on Security and Privacy*, pages 204–218. IEEE, 2009.

[11] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[12] E. Halperin and D. A. Stephan. Snp imputation in association studies. *Nature biotechnology*, 27(4):349–351, 2009.

[13] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[14] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1141–1152, 2013.

[15] M. Kayser and P. De Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179–192, 2011.

[16] Latanyasweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 05 2012.

[17] R. Lehmann-Haupt. Are sperm donors really anonymous anymore? dna testing makes them easy to trace. 2010.

[18] Z. Liang, B. Feng, L. Miao, W. Zhu, Y. Lin, M. Luo, S. Yang, L. Wang, J. Fu, H. Chen, and Z. Dong. Combined illumina and pacbio sequencing technology on transcriptome analysis reveals several key regulations during the early development of american shad (alosa sapidissima). *Aquaculture Reports*, 25:101264, 2022.

[19] Z. Lin, A. B. Owen, and R. B. Altman. Genomic research and human subject privacy. *Science*, 305(5681):183–183, 2004.

[20] L. Liu, E. Yu, and J. Mylopoulos. Security and privacy requirements analysis within a social setting. pages 151– 161, 10 2003.

[21] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37(3):179–192, 2004.

[22] S. Miyazaki, N. Mead, and J. Zhan. Computer-aided privacy requirements elicitation technique. pages 367–372, 12 2008.

[23] J. Mylopoulos, L. Chung, and N. BA. Representing and using non-functional requirements: A process-oriented approach. *Software Engineering, IEEE Transactions on*, 18:483–497, 07 1992.

[24] G. Naik. Family secrets: An adopted man's 26-year quest for his father. *Wall Street Journal*, 2, 2009.

[25] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang. Privacy in the genomic era. *ACM Comput. Surv.*, 48(1), aug 2015.

[26] A. J. Pakstis, W. C. Speed, R. Fang, F. C. Hyland, M. R. Furtado, J. R. Kidd, and K. K. Kidd. Snps for a universal individual identification panel. *Human genetics*, 127:315–324, 2010.

[27] R. Pulivarti. Cybersecurity of genomic data. Internal report 8432, National Institute of Standards and Technology, March 2023.

[28] E. E. Schadt, S. Woo, and K. Hao. Bayesian method to predict individual snp genotypes from gene expression data. *Nature genetics*, 44(5):603–608, 2012.

[29] A. Wood, K. Najarian, and D. Kahrobaei. Homomorphic encryption for machine learning in medicine and bioinformatics. *ACM Comput. Surv.*, 53(4), aug 2020.

[30] K. Wuyts, R. Scandariato, B. De Decker, and W. Joosen. Linking privacy solutions to developer goals. pages 847–852, 01 2009.

[31] E. Yu and L. Cysneiros. Designing for privacy and other competing requirements. 11 2002.