Determining the Identifiability of DNA Database Entries

Bradley Malin[1] and Latanya Sweeney[2]

[1]Department of Biological Sciences
[2]School of Computer Science and Heinz School of Public Policy
Carnegie Mellon University
Pittsburgh, Pennsylvania

CleanGene is a software program that helps determine the identifiability of sequenced DNA, independent of any explicit demographics or identifiers maintained with the DNA. The program computes the likelihood that the release of DNA database entries could be related to specific individuals that are the subjects of the data. The engine within CleanGene relies on publicly available health care data and on knowledge of particular diseases to help relate identified individuals to DNA entries. Over 20 diseases, ranging over ataxias, blood diseases, and sex-linked mutations are accounted for, with 98-100% of individuals found identifiable. We assume the genetic material is released in a linear sequencing format from an individual's genome. CleanGene and its related experiments are useful tools for any institution seeking to provide anonymous genetic material for research purposes.

INTRODUCTION

The genetic information of an individual is understood to be as, or more, personal than a fingerprint. Yet, having a database of only DNA entries, with no additional explicit demographic information or identifiers included, appears sufficiently anonymous. Associating autonomous DNA information to named persons seems impossible. After all, there does not exist a master registry against which the DNA information could be directly compared to reveal the associated person's identity. So how then, can persons be identified? This work demonstrates that inferences drawn from DNA information can be used to divulge the exact identities of the persons from whom the DNA originated.

The human genome is believed to harbor somewhere between 60,000 to 100,000 genes, which make up approximately 2% of the genome. National Center for Biotechnology Information (NCBI) statistics demonstrate that as of July 2000 almost 8500 human gene loci were established, hundreds of which have been characterized as being involved in genetic diseases[1,2]. Recent developments in molecular biology and sequencing have resulted in an exponential increase in the discovery of genetic loci and genes[3].

The rise in genetic data and resulting databases have a variety of uses in genetic and molecular biology research, including the ability to determine hotspots of mutation in genes and familial studies. Despite the considerable amount of basic and clinical research that may benefit from the availability of such data, care must be taken when such data are population based. The issue of genetic privacy is of utmost concern and must be addressed in terms of what information can be leaked before population based DNA data is shared.

BACKGROUND

To facilitate radical increases in human genetic research information and expedite scientific discovery, the National Center for Biotechnology Information established the Online Mendelian Inheritance in Man (OMIM) database in 1987[3]. This database contains the chromosomal location of known gene locus entries. It also includes important allelic variants that are known causes of clinical phenotypic abnormalities. The OMIM database has several search mechanisms, including one for providing the gene location given a known gene and another for providing the gene location given a known disease. Other major databases are specific to published mutations by gene and mutation type, such as the Human Gene Mutation Database, or by annotated sequence, such as the Database of Single Nucleotide Polymorphisms[4,5].

First Generation DNA Databases

There are currently more than 100 DNA databases available online over the World Wide Web (WWW)[6], most of which are publicly available. The online DNA databases are collections of known mutations in specific genes and loci. A common objective of these collections is to provide all known mutations of interest. Therefore, the only data these systems tend to include are known hotspots of mutation and other characterized mutations in genes. The individual who is the source of the mutation may be geographically located anywhere in the world. The means by which the information is provided to the database or otherwise validated is typically a published

1067—5027/00/$5.00

paper that appears in a respected publication. We term these databases first generation DNA databases.

First generation DNA databases appear sufficiently anonymous, however, there may exist individuals whose privacy may be compromised. For example, consider the Cystic Fibrosis Mutation Database maintained by the Cystic Fibrosis Genetic Analysis Consortium[7]. This database provides a mutation table of all published mutations in the CFTR gene, polymorphisms in the coding and non-coding regions of the gene, and publication references from which the mutations were submitted. This database lists all 918 known mutations from the United States, France, Italy, Soviet Union, and other places around the world. Databases such as the previous have been paramount in helping researchers determine hotspots for mutation and understand clinical phenotypes associated with specific mutations. While the database does not harbor specific data that could directly identify individuals, the database may be discredited in terms of privacy protection by someone performing at least one re-identification based on references within the publication that associate the reported piece of genetic information to its human source. The majority of first generation DNA databases have this level of protection and subsequently share the privacy risk.

Second Generation DNA Databases

We consider second generation DNA databases to be population based. In these cases, multiple submissions of DNA information corresponding to identical loci appear in the database. The submissions, which are partitioned into entries corresponding to each individual, include sequences from individuals within a geographically situated population. Data within second generation databases tend to have large amounts of sequences spanning many, or all, chromosomes. Maintaining such an abundance of additional genetic information can reveal sensitive information about the individual who was the source of the DNA, as well as, about blood relatives, who may not have been explicitly included in the database. As DNA sequencing speed continues to increase, the quantity and completeness of sequence entries in these databases are reasonably expected. A rise in the number of these databases has already appeared in university hospitals and other efforts, such as that by Genethon and deCode Genetics[8]. Collecting and sharing such information is useful to researchers and clinicians, however, because of geographical specification and additional inferences that can be drawn from the sequence, the privacy concerns for second generation databases are far more serious. The ways in which an attack could be launched on second generation DNA databases are more extensive than with first generation DNA databases. It is the records of second generation DNA databases that our research attempts to protect from breaches in privacy.

Computational Systems to Protect Privacy

Several computational systems presented that help render data anonymous have been designed. These include methods that locate and remove textual information in images[9], Scrub[10], which locates personally identifying information in unrestricted textual documents, and the Datafly[11] and Mu-Argus[12,13] systems, which attempt to render field-structured person-specific databases sufficiently anonymous. Before this work, no known system addressed the linear DNA information within genetic databases.

METHODS

Data used for this study includes hospital discharge data (called health data) for the state of Illinois for 1990 through 1997. In many states, such data are publicly available[14]. There are approximately 1.3 million hospital discharges per year in the health data, which reportedly corresponds to hospital compliance of 99+% for all discharges occurring in Illinois hospitals[15]. Diagnosis codes, procedure codes, patient demographics, and hospital identity are among the information stored for each visit. For convenience, we assume the second-generation DNA database under question results from patients that are in the hospital.

Correlating Disease Genes to ICD-9 Codes

Many diseases are known to have genetic influence. Some diseases depend on interactions of multiple gene products. Mutations in specific genes may not necessarily cause a certain disease, but many raise the risk of developing a disease, such as a mutation in the BRCT domain of BRCA1 and the acquisition of breast cancer[16]. However, there are diseases for which a single gene when mutated strongly correlates with (or has a casual relationship to) the development of a specific disease; we call these simple disease genes. The resulting diseases are diverse involving, for example, cancer, immunity, transporters, the nervous system, signaling, and metabolism. They may be autosomal or sex-linked, as well as dominant or recessive. Due to the growing number of simple disease genes, our first experiment was to relate a single gene detectable disease to ICD-9 diagnosis codes found in hospital discharge records. Over 20 such diseases were found to have a corresponding set of ICD-9 codes based on a first order search of specific disease names. Table 1 provides a sample listing. The preliminary search was not exhaustive due to the fact that some disease names are classified differently in ICD-9 codes than in its genetic counterpart; examples

include Menkes Syndrome, Machado-Joseph Disease, and diastrophic dysplasia.

Disease in Medical Release Data	Known Gene	Illness and Progression
Huntington's Chorea	HD	Imminent degeneration and death
Sickle Cell Anemia	HBB	Treatment available
Fragile X	FMR1	Imminent retardation
Refsum's Disease	PAHX	Treatment available
Phenylketo-nuria	PAH	Treatment available
Methemo-globinemia	HBB, HBA1, DIA1	Treatment available
Galactosemia	GALT	Treatment available
Amyotrophic Lateral Sclerosis (ALS)	SOD1	Imminent degeneration and death
Friedrich's Ataxia	Frataxin	Imminent degeneration and death

Table 1: Sample of ICD-9 classifications that match known gene counterparts.

Population-based Health Data Profiles

In preparation of using CleanGene, we generated population-based profiles from the health data, which we refer to as health data profiles. For each hospital visit in the health data that contained a diagnosis corresponding to that of a single disease gene, a profile was constructed consisting of {*date of birth, gender, ZIP, disease, hospital visit info*}, where *ZIP* was the patient's residential postal code. Profiles were then probabilistically merged based on census demographics for {*age, gender, ZIP*} so that values for *hospital visit info* from profiles that likely relate to the same person were combined. The set of resulting profiles contained the demographics for persons diagnosed with targeted diseases and who were geographically situated near the hospital collecting the second-generation DNA database.

It has been shown that by using data linkage algorithms, 80-100% of these kinds of health data profiles can be accurately re-identified using publicly available population registers[14,17]. Therefore, if we can match profiles of patients, who have been diagnosed with a simple genetic disease, to corresponding genetic sequences stored in a second-generation DNA database (maintained by a hospital), we could reveal the identity of the DNA contributors in almost all cases.

Computer Approach: Design and Implementation

CleanGene is a Java program that uses Java Database Connectivity (JDBC) to connect to a relational database. Given a set of population-based health data profiles and a second-generation DNA database, CleanGene identifies entries in the DNA database that are likely candidates for re-identification. It accomplishes this by employing knowledge-based algorithms to independently construct identifying profiles from the entries in the DNA database, which we call genetic profiles. These profiles are then deterministically linked to the health data profiles to pinpoint likely DNA candidates for re-identification.

Before we describe how CleanGene operates in detail, we must first describe the genetic data. Each data entry in the genetic database consists of linear sequences of the letters ACGT, each corresponding to a nucleotide of DNA. The genome in the database entry consists of both genes and noncoding sequences. The genes themselves are not continuous, many harboring noncoding regions (introns) between the coding regions (exons). However, the locations of the simple disease genes on their respective chromosomes used in this study are characterized and publicly available.

Each linear sequence of DNA in the database may be single or double stranded depending on the annotation strategy maintained by the data collector. Each of these sequences is considered a genetic profile to which additional information regarding {*Hospital identifier, gender, disease*} may be added.

Step 1. Gender. The first step in expanding the genetic profiles is to determine the gender of each DNA entry. Gender may be one of the fields associated with the DNA entry. If not, it has been demonstrated that the gene amelogenin can be used to categorize specimens as male or female. Amelogenin is an ideal gene due to the fact that there is a significant difference in size of the genetic locus in the X- and Y-chromosomes[18]. A straightforward search through the DNA entries establishes a gender determination in virtually all cases where the amelogenin gene is sequenced. This step is demonstrated in the top row of Diagram 2.

Step 2. Disease. The second step in expanding the genetic profiles is to identify instances of simple genetic diseases. DNA is double-stranded, but a gene only exists on one of the two stands. CleanGene ignores the noncoding strand during a particular search and proceeds in one of two search mechanisms depending on the genomic characterization of the gene in question. Some genes are completely sequenced, intronic regions as well as exons. Consensus sequences for disease genes are known and are used as the sequence to which all database entries are compared. The full consensus sequences are available via the WWW at sites such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/).

CleanGene utilizes a homologous search algorithm, similar to that of the Basic Local Alignment

Search Tool (BLAST), which is a set of similarity search programs designed to explore all of the available sequence DNA databases[19]. The sensitivity of CleanGene to distant sequence relationships is controlled by a range variable. CleanGene searches for global alignments with stringency controls in intronic sequences. Full sequenced gene loci are searched for (1) point mutations that change the protein of the gene product and (2) frameshift mutations for the addition or removal of nucleotides. This search strategy is depicted in the second row, left column of Diagram 2.

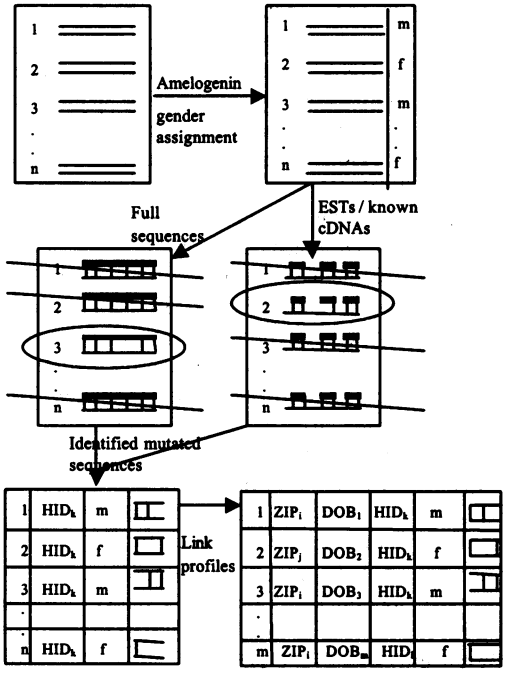

Diagram 2: Overview of CleanGene's operation. Circled entries represent examples of mutation. Horizontal lines represent DNA. Vertical lines symbolize positive alignment match of consensus and subject sequences. HID is an abbreviation for hospital identifier.

On the other hand, not all genes have a fully sequenced locus in the genome. For these genes, CleanGene searches expressed sequence tags (ESTs) and cDNA clone sequences. It aligns the smaller exon sequences from the consensus sequences with the database entries, which should be homologous to the consensus sequences. With genes containing introns, the alignment is strong for exon and EST/cDNA sequence. When intron overlapping occurs, the amount of homology observed significantly decreases. To prevent CleanGene from automatically returning such an entry as a mutated sequence, the program searches for runs of homology following or preceding the exon being compared to avoid erroneous mismatching that is explained by frameshift mutations. Nonhomolgous sequences unexplained by frameshifting or point mutations are discarded from homology comparison. The resulting alignments are searched for homology and mutation. This search strategy is depicted in the second row, right column of Diagram 2.

Step 3. Hospital and other inferences. If the second generation DNA database is collected by a hospital, then the identity of the hospital is implied and can be added to the genetic profiles. This is depicted in the third row, left column of Diagram 2. Additional information about geographical location, age and types of patients can often be inferred from supplemental knowledge about the data holder (e.g., Children's Hospital) or about particular diseases.

Step 4. Linking the profiles. Finally, the genetic profiles, with associated inferences, are directly linked to the health data profiles, thereby associating additional DNA-based inferences to identifiable data. This is depicted in the third row, right column of Diagram 2.

Disease	Total Hospital Visits	Number of Hospitals With Reported Cases	% of Unique Identifiable Individuals
Refsum's Disease	2	1	100%
Phenylketonuria	7	7	100%
Fragile-X	12	9	100%
Galactosemia	13	10	100%
Methemoglobinemia	20	11	100%
Friedrich's Ataxia	30	18	100%
Huntington's Chorea	136	53	100%
ALS	273	90	100%
Sickle Cell	7491	118	98%

Table 2. Selected genetic disease incidence in hospitals.

RESULTS

Table 2 reports the prevalence of some simple genetic diseases based on hospital profiles. We found that of the simple gene diseases profiled, the resulting patients were between 98-100% identifiable. Determining the number of distinct individuals and their identifiability is achieved using data linkage algorithms[17]. Intuitively, we can confirm these estimates by considering the following number of possible combinations:

2 genders * 365 days/year * 80 years = 58,400

Therefore, {*gender, date of birth, ZIP*} for any 5-digit ZIP having less than 58,400 individuals is likely to be unique. In the State of Illinois there are 1236 distinct ZIPs and 1212 (or 98%) of them have populations less than 58,400. These 1212 ZIPs account for 84% of the entire state's population[14,17]

DISCUSSION

CleanGene relies on health data profiles that need not necessarily be based on hospital visits. After all, some diseases have a greater occurrence of hospital visits than others. For some diseases there are no treatments available, such as Huntington's Chorea and Friedrich's Ataxia. Other diseases pertain more to diet treatments, which do not require hospital treatment, such as phenylketonuria. Still, some diseases such as sickle cell anemia, which are less deadly than some of the others diseases listed in Table 1, not only have a higher incidence of survival, but also has a variety of treatment options. Some genetic diseases are not explicitly annotated in health data by ICD-9 codes. On the other hand, there are many other sources of health data available (e.g., pharmacy records and ambulatory care data) and as DNA chips and arrays[20] become more prevalent, the presence of recordings in health data of mutated genes is likely to be far more common.

Genetic information disclosure is one of the most feared aspects of health related data in today's society[21]. The CleanGene program demonstrates that if genetic databases are to be shared or made available for research purposes, some necessary precautions must be taken beforehand. These precautionary measures may be in the form of full removal of the simple genetic genes. However, this seems to be an unfair tradeoff for researchers in population genetics and cytogenetics. A more research-friendly procedure may be to systematically remove specific hotspots of mutation in each simple disease gene. Regardless of the system enacted for anonymity maintenance, it is evident that second generation DNA databases must be safeguarded.

Acknowledgements

The authors thank the State of Illinois for use of their data. This work has been supported in part by the Howard Hughes Medical Institute.

References

1. Beroud C, et al. UMD (Universal Mutation Database); a generic software to build and analyze locus-specific databases. *Human Mutation*. 2000; 15(1): 86-94.
2. Brylawski B. OMIM statistics. Johns Hopkins University, Baltimore MD. July 13, 2000. http://www3.ncbi.nlm.nih.gov/Omim/Stats/mimstats.html
3. Hamosh A, et al. Online Mendelian Inheritance in Man (OMIM). *Human Mutation*. 2000;15:57-61.
4. Krawczak M, et al. The human gene mutation database. *Trends in Genetics*. 1997;13: 121-122.
5. Sherry ST. Use of molecular variation in the NCBI dbSNP database. *Human Mutation*. 2000; 15: 105-113.
6. Discala C, et al. DBcat: a catalog of 500 biological databases. *Nucleic Acids Research*. 2000; 28(1): 8-9.
7. Cystic Fibrosis Genetic Analysis Consortium. *Cystic Fibrosis Mutation Data Base*. 2/18/2000. http://www.genet.sickkids.on.ca/cftr
8. Greely HT. Icelands plan for genomics research: Facts and implications. *Jurimetrics*. 2000; 40: 153-191.
9. Wang JZ , Wiederhold G. System for efficient and secure distribution of medical images on the Internet Proc AMIA Symp. 1998;:907-11.
10. Sweeney L. Replacing personally-identifying information in medical records, the scrub system. In: Cimino, JJ, ed. Proceedings, *JAMIA*. Washington, DC: Hanley & Belfus, Inc., 1996:333-337.
11. Sweeney L. Guaranteeing anonymity when sharing medical data, the datafly system. Proceedings, *JAMIA*. Washington, DC: Hanley & Belfus, Inc., 1997.
12. Sweeney L. Three computation systems for disclosing medical data in the year 1999. *Medinfo*. 1998; 9 Pt 2: 1124-1129.
13. Hundepool A and Willenborg L. μ- and τ-argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality*. Bled: 1996.
14. Sweeney L. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics*. 1997, 25: 98-11.
15. "Data release overview," *State of Illinois Health Care Cost Containment*. Springfield: 1998.
16. Cortesi L, et al. Comparison between genotype and phenotype identifies a high-risk population carrying BRCA1 Mutations. *Genes, Chromosomes and Cancer*. 2000; 27 (2); 130-5
17. Latanya Sweeney's testimony in Southern Illinois vs. Department of Public Health, et al., No. 98-CH-5. Illinois, 1998.
18. Altschul SF, et al. Basic local alignment search tool. *Journal of Molecular Biology*. 1990 Oct 5; 215(3): 403-10.
19. Caenazzo L, et al. Prenatal sexing and sex determination in infants with ambiguous genitalia by polymerase chain reaction. Genetic Test. 1997-1998; 1(4): 289-291.
20. Granjeaud S, et al. Expression profiling: DNA arrays in many guises. *Bioessays*. 1999 Sep; 21(9): 781-90.
21. Gostin L and Hodge J. Genetics privacy and the law: An end to genetics exceptionalism. *Jurimetrics*. Fall 1999; 30: 21-58.