BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 2. 117–128
Advance Access publication March 24, 2007
doi:10.1093/bib/bbm006

Forensic DNA and bioinformatics

Lucia Bianchi and Pietro Liò

Abstract

The field of forensic science is increasingly based on biomolecular data and many European countries are establishing forensic databases to store DNA profiles of crime scenes of known offenders and apply DNA testing. The field is boosted by statistical and technological advances such as DNA microarray sequencing, TFT biosensors, machine learning algorithms, in particular Bayesian networks, which provide an effective way of evidence organization and inference. The aim of this article is to discuss the state of art potentialities of bioinformatics in forensic DNA science. We also discuss how bioinformatics will address issues related to privacy rights such as those raised from large scale integration of crime, public health and population genetic susceptibility-to-diseases databases.

Keywords: *forensic science; DNA testing; CODIS; Bayesian networks; DNA microarray*

INTRODUCTION

Bioinformatics and forensic DNA are inherently interdisciplinary and draw their techniques from statistics and computer science bringing them to bear on problems in biology and law. Personal identification and relatedness to other individuals are the two major subjects of forensic DNA analysis. Typical contexts for forensic analysis are disputes on kinship; for example paternity disputes, suspected incest case, corpse identification, alimentary frauds (e.g. OGM, poisonous food, etc), semen detection on underwear for suspected infidelity, insurance company fraud investigations when the actual driver in a vehicle accident is in question, criminal matters, autopsies for human identification following accident investigations. Genetic tests have been widely used for forensic evidences and mass-fatality identification (terrorist attacks, airplane crash, tsunami disaster) [1,2]. Genetic testing results are integrated with information collected by multidisciplinary teams composed of medical examiners, forensic pathologists, anthropologists, forensic dentists, fingerprint specialists, radiologists and experts in search and recovery of physical evidence. Large scale tissue sampling and long-term DNA preservation under desiccation conditions with potential applications in mass fatalities has been recently described [2–4].

In several countries new rules could allow fingerprints and DNA samples to be taken from anyone they arrest, whether they are charged or not. This will be certainly facilitated by the introduction of three different key innovations, in data acquisition, such as thin film transistors (TFT) [5], in DNA sample identification, such as microarray re-sequencing and in statistical methodologies, such as Bayesian networks (BNs), which provide an effective way of evidence organization and inference.

The TFT, which can be seen as a combination of an intelligent version of liquid crystal display and wafer thin technology, will allow the DNA to be identified on the crime scene or in the police station [5,6]. Current genome sequencing projects employ high-throughput shotgun sequencing at large centers. Rapid DNA sequencing technology is nowadays based on microarrays; for example, the entire sequence of the mitochondrial genome (16 500 bases) can be re-sequenced in a single (three PCR) 48 h experiment allowing to detect variants over all the sequence and not just restricted to the hypervariable regions [7].

Corresponding author. Pietro Liò, Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, CB3 0FD, Cambridge, UK. E-mail: luciagbianchi@gmail.com

Lucia Bianchi has a degree in Law from the University of Firenze (Italy). After 2 years of legal practice she has attended courses on Forensic DNA analysis at the University of Bologna and courses on the English Legal system at the University of Cambridge (UK).

Pietro Liò is a lecturer at the University of Cambridge Computer Laboratory in England, where he undertakes research and teaching in the general area of Bioinformatics, Computational Biology, and System Biology.

Taking into account this scenario, we discuss links between bioinformatics and forensic statistics which is a discipline focusing on the experimental design of forensic examinations and evidences [8–10]. Bioinformatics will affect how forensic statistics will address hypothesis formulation on DNA samples, deciding on minimal population sample sizes when studying populations of similar units of evidence and determining the statistical significance of the outcome of tests. Particularly we aim at discussing the role of BNs which is becoming very useful in the study of the implications of forensic examinations on defendant and prosecution positions during crime investigation and criminal court proceedings.

THE BASICS OF FORENSIC DNA PROCEDURES

Personal identification relies on identifiable characteristics such as biological (DNA, blood, saliva, etc.), physiological (fingerprints, eye irises and retinas, hand palms and geometry and facial geometry), behavioral (dynamic signature, gait, keystroke dynamics, and lip motion) and on mixture of physiological and dynamical characteristics such as the voice.

DNA has become the most important personal identification characteristic because all genetic differences whether being expressed regions of DNA (genes) or some segments of DNA with no known coding function but whose pattern of inheritance can be monitored can be used as markers.

Any two humans are >99% identical in their DNA sequences, still have millions of genetic differences, making them different in their risk of getting certain diseases and response to environmental factors.

The most important sources of genetic variations are copy number variation (CNV) [10], large genomic regions that are absent from, or duplicated in different individuals, and SNP, single nucleotide polymorphisms i.e. single base difference among two different individuals of the same species. SNPs in humans occur in average every 1/2000 bases.

The human genome is also highly repetitious. Repetitions occur at most of sequence length scales, number and dispersion [11]. Examples of such repetitions are homo- and di-nucleotide repeats (microsatellites), and families of interspersed, mobile elements hundreds of base pairs long such as the ALU sequences. There are more than one million ALU sequences in the human genome, each 300 bases long, which are able to copy themselves in other parts of the genome, generating mutations.

Forensic DNA typing often requires the use of techniques that allow the detection of genetic variations among humans, usually short, repetitive loci. variable number of tandem repeats (VNTRs) polymorphism were used till few years ago. Such loci are composed of core units three, four or five nucleotides long and the number of repeated segments at a locus varies between individuals. One VNTR in humans is a 17 bp sequence of DNA repeated between 70 and 450 times in the genome. The total number of base pairs at this locus could vary from 1190 to 7650. VNTRs are identified by cutting genomic DNA with restriction enzymes such as HaeIII, HinfI or HindIII, separating the DNA fragments electrophoretically in a gel, and then detecting the variable fragments by the use of short DNA stretches that bind specifically to variable loci (probes). Nowadays the use of VNTR has been replaced by STR (short tandem repeats). CNV are supposed to be major determinants of human traits, and they may become useful in forensic science, particularly in the determination of the population substructures.

STR system and CODIS

The use of PCR allows to analyze DNA from samples as small as a single cell and, therefore, DNA typing analysis using STRs can be performed on a large variety of materials, such as cigarette ends, skeletal remains, urine, tissues on a gun muzzle and on bullets, dismembered and decayed body parts, paraffin embedded tumor tissue, dirt under fingernails, epithelia of an offender from the victim's neck after strangling, mummified newborns, blowflies preserved in ethanol, burned corpses, dentin, dried chewing gum, body parts after mass disasters, human feces and skeletonized flood victims.

At present, the most discriminative power in DNA identification is obtained by matching 13–17 of nuclear STR markers of a victim's profile (personal items, like toothbrushes and used shavers) to a direct antemortem sample of the victim or to family references: either or both biological parents of the victim. A system of 13 STRs constitutes the Combined DNA Index System (CODIS) which is used in USA and Canada [12], while most of European countries have their own systems and databases (see subsequently).

Although CODIS strictly represents the USA and Canadian felons and forensic samples database, sometimes it is used to match probability in mass disasters outside USA. For example in the Madrid

Table 1: Distribution of Codis STR on human chromosomes

Combined DNA Index System (CODIS)
D3S1358 3p21 (11449919)
vWA 12p12-pter (M25858)
FGA 4q28; located in the 3rd intron of human alpha fibrinogen gene (M64982)
D8S1179 8q24.1-24.2; (GO8710)
D21S11 21q21.1 (M84567)
D18S51 18q21.3 (X91254)
D5S818 5q21-q31 (G08446)
D13S317 13q22-q31 (G09017)
D7S820 7q (G08616)
D16S539 16q22-24 (G07925)
THO1 11p15-15.5; intron 1 of tyrosine hydroxylase gene (D00269)
TPOX 2p23-2pter; intron 10 of human thyroid peroxidase gene (M68651)
CSF1PO 5q33.3-34; c-fms proto-oncogene for CSF-1 receptor gene (X14720)
AMEL X Xp22.3-p22 (M86932)
AMEL Y Yp11 (M86933)

For each marker the table reports Chromosomal Location and GenBank accession. The Combined DNA Index System (CODIS) is the FBI's national databases of genetic identification codes.

terrorist attack case, the CODIS database was used to match probabilities of 220 body remains against 98 reference samples, including 67 samples from relatives, representing 40 family groups and 27 antemortem direct references.

In Table 1 we describe the Codis system of STR. For each marker chromosomal location and the GenBank accession are reported; for example, the table shows that the D3S1358 marker is on chromosome 3, has GenBank accession 11449919 from which we find that it has 18 repeats, of the form $TCTA(TCTG)_2(TCTA)_{15}$. The website http://www.cstl.nist.gov/div831/strbase/fbicore.htm contains information (frequencies) of the marker's alleles, i.e. the common variants.

When the genomic DNA is too scarce or too degraded for standard forensic analysis, the sequence of two hypervariable regions of the mitochondrial DNA is considered informative.

Multiple STR systems

The knowledge of the frequency of a certain STR allele in a population enables computing how often an allele combination appears in a given population. Because of their high variability, i.e. high numbers of rare alleles, classical VNTR loci alone often lead to much higher exclusion (or inclusion) probabilities than single modern STR systems alone, which often have quite common and widespread alleles. Thus the detection of an allele combination in only a single STR system in a biological stain seldom constitutes conclusive proof of identity. If, however, alleles in stain are observed not to be identical to those of a person's reference body fluid, in extreme cases even one STR profile can exclude the person from the suspicion of having left the stain.

Severely degraded DNA samples could contain only very short DNA template molecules (under 150 bp) making conventional STR typing (150–400 bp) unsuccessful. Damaged DNA templates (very old bones, hair shafts) and minute amounts of cells occasionally lead to the elimination of single or, in the worst case, all alleles, and occasionally one even obtains nonreproducible results. Multiplex PCR involves using several sets of PCR primers to the reaction and allows to target multiple locations throughout the genome. This is an ideal technique for DNA typing because the probability of identical alleles in two individuals decreases with the increase in the number of polymorphic loci examined. Currently, STR multiplex systems have a discrimination power (i.e. matching probability) greater than a combination of five classical single locus DNA fingerprints. For example, a third-generation multiplex PCR developed at the Forensic Science Service (FSS) in Birmingham (http://www.forensic.gov.uk) matches persons to a stain with a probability of 1:10 [13]. Figure 1 shows the key time developments of the forensic technology and the increased resolution of the DNA analyses.

The multiplexing technology can save time and money, but difficulties may arise when coamplifying several loci. Primers for one locus can complex with those of other loci and completely inhibit the amplification. This effect may be exhibited by dropout of a specific STR locus under certain conditions (e.g. sample mixtures). Finding the optimum PCR conditions, particularly the annealing temperature and the primer concentrations, can be challenging and time-consuming.

The procedure used when a match is found, consists in typing again the DNA by a scientist or technician who does not know which sample he/she is processing (bar code, no information about former typing result). If a new PCR analysis of the stored biological material confirms the match, fresh material is taken from the alleged suspect and analyzed in another laboratory. Only after a third confirmation of the PCR results in 13 of 14 STR systems, the match is communicated to the responsible authority.

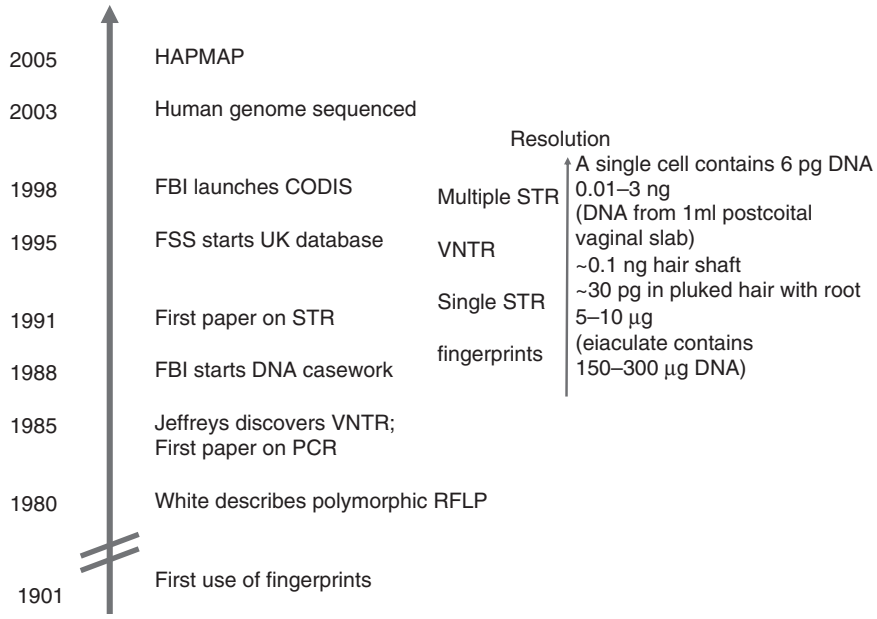

Figure 1: Time scales of major events in forensic DNA typing (left) and examples of resolution power of the different techniques (right).

Noteworthy, German courts generally consider five or six STRs to be sufficiently strong evidence of identity [14–18].

Gender determination

Forensic investigations may take full advantage of bioinformatics resources. For examples, given a tissues specimen, the length of the chromosomal ends and mitochondrial numbers may tell us about the age process while testing the amelogenin marker tells us information of the gender. Amelogenin is a matrix protein which comprises 90% of all the proteins in the tooth enamel. It regulates the initiation and growth of hydroxyapatite crystals during the mineralization of enamel and is involved in the development of cementum by directing cells that form cementum to the root surface of teeth. Using primers specific for intron 1 of the amelogenin gene, the X chromosome gives a 106 base pairs amplification product and the Y chromosome a 112 base pairs amplicon [19]. Therefore, samples from male sources (XY) will show two bands on an agarose gel, while females (XX) will show only one band. A region in the exon 6 is a hot spot of mutations, particularly amino acid insertions or deletions, in all mammals. In this region, numerous triplet repeats (PXQ) have been inserted recently and independently in five mammal lineages, while most of the hydrophobic exon 6 region probably had its origin in several rounds of triplet insertions, early in vertebrate evolution [14]. These differences may allow to use amelogenin in animal identification.

Using plant, bacteria, pollen and other bioinformatics data

Whenever crime scene investigation needs identification of bacteria, insects, and plants, genomic sequences can be resequenced using microarray [15–18, 20, 21] and analyzed using bioinformatics standard techniques. For example, practice in forensic entomology allows to determine postmortem intervals by analyzing the developmental status of certain hexapod species on corpses and has a role in toxicological analysis. The use of phylogenetic inference may lead to more precise taxonomic identification of the species, providing geographical information. Similarly, pollen and spores analysis, i.e. palynology, may provide information of a particular place and a certain time frame. Feline, canine and white-tailed deer DNA evidence has been presented in court, and follows the procedures for human DNA forensics.

Linkage disequilibrium and haplotyping

In order to avoid pitfalls in the inference process of the forensic evidences, it is important to discuss the patterns of occurrences of mutations in the human

genome. There are strong statistical associations between polymorphisms in the human genome, such that the presence of a particular variant at one site on a chromosome can predict or 'tag' the presence of a particular variant at another site. Linkage disequilibrium (LD), is the nonrandom pattern of association between alleles at different loci within a population. An association in inheritance between characters means that the parental character combinations appear among the progeny more often than the nonparental. The closer two or more markers are on a chromosome the greater the probability that they will be inherited together [22–24].

LD is generally low near telomeres, elevated near centromeres and correlated with chromosome length, particularly high in few regions, termed recombination hotspots which are enriched of retrotransposon-like elements. LD is low in regions containing genes involved in immune responses and neurophysiological processes, and high in regions containing genes involved in DNA and RNA metabolism, response to DNA damage and the cell cycle [13].

Variants that associate together are known as a 'haplotypes'. Therefore, a haplotype is a set of closely linked genetic markers present on one chromosome which tend to be inherited together. Intuitively, haplotypes (which can be regarded as a collection of ordered markers) may be more powerful than individual, unorganized markers [23].

Haplotype patterns reflect the fact that all modern humans originated in Africa more than 150 000 years ago. Some of the descendents of this group remained in Africa, whereas others migrated, eventually reaching all parts of the world. DNA events such as mutations and recombinations, natural selection and random drift which have caused population expansions and bottlenecks, founder effects, have influenced (generated or eliminated) the haplotype patterns in populations in different parts of the world.

While the reference sequence constructed by the Human Genome Project is informative of the vast majority of bases that are invariant across individuals, the HapMap project (http://www.hapmap.org) [13] focuses on DNA sequence differences among individuals. The Hapmap project consisted of compiling data on groups of individuals representative of four populations for more than a million single nucleotide polymorphisms, or SNPs. If a similar project will be carried out for CNV, identification of risk factors for common human diseases will be helpful in treatment or prevention and forensic information on human population will be complete.

Software for haplotype scoring, selection, visualization

There is a large variety of software useful for haplotype analysis. Most of this software comes with example data sets and manuals so it is easy to try different programs and make comparison on the basis of the specific needs and data sets. We describe a list of software relevant to haplotyping and linkage disequilibrium analysis that we have found particularly useful in forensic bioinformatics [25–33].

Haploview (www.broad.mit.edu/mpg/haploview) is designed to simplify and expedite the process of haplotype analysis by providing a common interface to several tasks relating to such analyses. Haploview currently supports the following functionalities. LD and haplotype block analysis, haplotype population frequency estimation, single SNP and haplotype association tests, permutation testing for association significance, implementation of Tagger (see subsequently), tag SNP selection algorithm. Haploview computes single locus and multimarker haplotype association tests, outputting the chi square and *P*-value for the allele frequencies in cases versus control. For family trios, all probands (affected individual with genotyped parents) are used to compute transmission disequilibrium test (TDT) values. Haploview can only interpret biallelic markers—markers with greater than two alleles (e.g. microsatellites) will not work correctly.

Haplofreq (http://www.cs.princeton.edu/haplofreq/) estimates the haplotype frequencies over a short genomic region given the genotypes with missing data. Haplofreq's approach incorporates a maximum likelihood model based on a simple random generative model which assumes that the genotypes are independently sampled from the population.

Haplofreq accepts as an input a set of gentoypes of the same length, and produces the haplotype distribution in the population, estimated from these haplotypes. The input format contains a genotype in each line. A genotype is described by a string of A,G,C,T,H and ?, where A,G,C,T correspond to homozygous sites, H corresponds to heterozygous

site and ? corresponds to missing data. Example:
AAGACCTT
GGAAHHHH
HHH???TT
GGGGTACC

Tagger (www.broad.mit.edu/mpg/tagger/) is a tool for the selection and evaluation of tag SNPs from genotype data such as that from the International HapMap Project. It combines the simplicity of pairwise tagging methods with the efficiency benefits of multimarker haplotype approaches. As input, users can upload genotype data in raw HapMap format or standard 'pedigree' linkage format. Alternatively, users can specify chromosomal landmarks to indicate genomic regions of interest within which tag SNPs are to be picked. This feature will be particularly useful for multiplex tag SNP design of candidate genes. Tagger has been implemented in the standalone program Haploview [25] (see above).

QTDT (http://bioinformatics.well.ox.ac.uk/project-ld.shtml) Linkage and association variance components analysis of quantitative traits.

GOLD (http://bioinformatics.well.ox.ac.uk/project-ld.shtml) Graphical Display of Linkage Disequilibrium: color-coding of LD-coefficient matrices. Distribution also includes some handy programs for calculation of LD coefficients [31].

GRR (http://bioinformatics.well.ox.ac.uk/GRR) Graphical Representation of Relationship errors. Simple representations of observed allele sharing in families to highlight erroneous coding of relationships amongst members [32].

SNPtagger (http://www.well.ox.ac.uk/~xiayi/haplotype) selective definition of haplotype tag SNPs: web-tool for picking minimal sets of non-redundant markers to capture information in input haplotypes.

PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/haplo.shtml) Whole genome association analysis toolset.

PHASE (http://www.stat.washington.edu/stephens/software.html) software for haplotype reconstruction, and recombination rate estimation from population data.

Snphap (http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt) program for estimating frequencies of haplotypes of large numbers of diallelic markers from unphased genotype data from unrelated subjects.

Arlequin (http://anthro.unige.ch/arlequin/software/) Multipurpose population genetics software implementing a large set of methods such as AMOVA using microsatellite data and dominant markers (RAPDs, AFLPs) [33].

Several 'R' functions (http://lib.stat.cmu.edu/R/CRAN/) are used for likelihood inference of trait associations with haplotypes and other covariates in generalized linear models. The functions accommodate uncertain haplotype phase and can handle missing genotypes at some SNPs. They need R release $\geq$2.0.1, and the following libraries: stats and survival.

Hapassoc: Likelihood inference of trait associations with SNP

Haplo.ccs: Estimates haplotype and covariate relative risks in case-control data by weighted logistic regression. Diplotype probabilities, which are estimated by the Expectation Maximization algorithm with progressive insertion of loci, are utilized as weights.

Haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates. It is a suite of routines for the analysis of indirectly measured haplotypes. The genetic markers are assumed to be codominant and it is possible to refer to the measurements of genetic markers as genotypes. The main functions in Haplo Stats are: haplo.em, haplo.glm and haplo.score.

Hapsim: Package for haplotype data simulation. Haplotypes are generated such that their allele frequencies and linkage disequilibrium coefficients match those estimated from an input data set.

Other repositories with interesting software and data sets are http://www.nslij-genetics.org/soft/,
http://linkage.rockefeller.edu/soft/,
http://www.animalgenome.org/soft,
www.hapmap.org/download/encode1.html,
http://www.broad.mit.edu/tools/data.html.

DATABASES

One of the most contentious issues in forensic use of DNA evidence is how to estimate the probability that two DNA profiles match by chance. In order to determine the probability that a particular genotype might occur at random in a population, extensive population data must be gathered to make an estimate of the frequency of each possible allele and genotype. A sample size much larger than 100 samples is needed to make reliable projections about a genotype's frequency in a large population [34]. Given allele frequencies, DNA profiles are screened automatically for matches between profiles

of person to person(s), person to scene(s) and scene to scene(s). Although the most used set of STR loci are spread on all the chromosomes (some chromosomes have just one), they have different mutation rate.

It is noteworthy to describe the information contained in the STR database (see for instance http://www.cstl.nist.gov/biotech/strbase/index.htm and http://www.str-base.org/index.php): the use of STRs; facts and sequence information on each STR system, descriptions of annotated sequence, population data, commonly used multiplex STR systems, their chromosomal locations, mutation rates for common loci, PCR primers and conditions, description of various technologies for analysis of STR alleles, addresses of scientists and organizations working in the field and a comprehensive reference listing of material on STRs used for DNA typing purposes.

A range of organizations are currently involved in developing and promoting DNA databases across the European Union (EU). For example: the European DNA Profiling Group (EDNAP) has existed since 1988 with the aim of establishing systematic procedures for data-sharing across the European community; the Standardization of DNA Profiling in the EU (STADNAP) exists to promote co-operation across the EU in order to use DNA profiling to detect 'mobile serial offenders'. The European Network of Forensic Science Institutes [35] (ENFSI, http://www.enfsi.org/) has similar ambitions to standardize forensic practices in support of policing across the whole of the EU. The EU itself provides funds (for example, to STADNAP) to ascertain best practices of data-sharing across criminal jurisdictions.

Differences exist amongst EU member and candidate states regarding the existence of a database (e.g. the Republic of Ireland and Portugal do not currently operate national databases); the conditions for including profiles (e.g. Belgium limits the inclusion of profiles to cases of serious offences against persons after a conviction is secured); profile removal (virtually all states, unlike the UK, remove profiles of the acquitted); the taking of samples with or without consent (e.g. France has strict rules for allowing DNA only to be taken with consent); and sample destruction (e.g. Germany specifies that all samples must be destroyed after profiling is completed, regardless of investigative and prosecutorial outcome).

In Italy collective efforts by large number of public (hospitals/universities/national criminal justice service) and private laboratories scattered around Italy brought to nine loci typed in about 2800 individuals and six loci typed in more than 1500 individuals [21] (see http://www.gefi-forensicdna.it.).

The ENFSI has undertaken an extensive study collecting STR-data from 24 European populations using the AMPFLSTR SGM Plus system, which has become one of the standard STR multiplexes to be used within Europe for the purpose of constructing national DNA criminal intelligence databases. This allele proportion (frequency) database—further referred to as the 'ENFSI DNA WG STR Population Database'—can be used to calculate match probabilities of DNA profiles from cosmopolitan Caucasian populations across all Europe, regardless of their specific country of origin.

Other examples are the ENFSI DNA WG STR Population Database (http://www.str-base.org/index.php) which uses 5699 samples from 24 European populations with SGM Plus kit loci. Noteworthy the Canadian Random Match Calculator for Profiler Plus and COfiler kits (http://www.csfs.ca/pplus/profiler.htm).

MASS DISASTER IDENTIFICATION

Pair-wise comparisons of DNA profiles in mass fatality incidents will require the bioinformatics capability to search (all-against-all) large STR and SNP data sets [36–38]. This task is performed using at least two different searching algorithms for autosomal markers: (i) looking for a perfect match: number of loci at which both alleles were found to match, a number which is expected to be equal to the number of loci analyzed among fragments of the same body or between a victim and a direct reference; (ii) allele sharing by kinship: number of loci at which at least one allele was found to match, a number which is expected to be equal to the number of loci analyzed for parent–child relationships.

The software should also have the capability to rank the significance of the DNA match. Specific software are Mass Disaster Kinship Analysis Program (MDKAP) and Mass Fatality Identification System (M-FYSis) which were widely used in the WTC disaster. They can align profiles derived from the remains to a reduced number of consensus profiles, assemble the overlapping partial

profiles and compute the likelihood ratios for each pair-wise comparison at various relationships, such as parent–child, sibling or half-sibling; apart from STR data; they also manage mtDNA and SNP data.

STATISTICAL INFERENCE: 'BEYOND THE REASONABLE DOUBT'

A key concept in forensic statistics is the Hardy–Weinberg equilibrium (HWE) [39,40] which assesses that in a large, randomly mating population, in which the evolutionary forces such as selection, migration and mutation are not acting, allele and genotypic frequencies do not change. Given a locus with two alleles, A, B, with frequencies p_A, $p_B = 1 - p_A$, and genotypes AA, AB, BB with genotypic frequencies P_{AA}, P_{AB}, P_{BB}, the relationship is $P_{AA} = p^2$, $p_{AB} = 2p_A(1 - p_A)$, and $P_{BB} = (1 - p_A)^2$, holds and the locus is said to be in HWE. One generation of random mating is sufficient to produce HWE. Classical statistical tests such as goodness-of-fit, exact Fisher, likelihood ratio and z-tests can be applied to HWE analysis [41–44]. The deviation from the HW is generally tested with the Pearson's chi-squared test, using the observed genotype frequencies obtained from the data and the expected genotype frequencies obtained using the HWE. Simulations show that one of the most powerful tests for HWE is the exact test, particularly when the number of alleles is large.

When there is a large number of alleles, this may result in data with many empty possible genotypes and low genotype counts, because there are often not enough individuals present in the sample to adequately represent all genotype classes. If this is the case, the asymptotic assumption of the chi-square distribution, will no longer hold, and it may be necessary to use Fisher's exact test.

In the common practice, if a bin contains very few bands, the FBI merges that bin into an adjacent bin of higher frequency. This merging of bins is believed to yield a more conservative estimate of the probability of a random match.

For each CODIS marker we can determine the genotype and the frequency of the different alleles; the probability (P) for a DNA profile is usually computed as the product of the probability for each individual locus, i.e. the profile probability = ($P1$) ($P2$) . . . (Pn). This is called the Product rule technique. This probability estimate is based on the assumption that the individual alleles are independent of each other, which is usually not the case, as discussed in the previous sections. If the probabilities of the individual alleles are not independent (i.e. if certain alleles are often associated), multiplying the individual allele frequencies may underestimate or overestimate the true probability of matching alleles in the chosen population and thereby mis-state the incriminating value of the evidence. Critics of the product rule technique contend that in some ethnic subpopulations the alleles identified by commonly used genetic probes are so extreme that the use of a broad-based comparison of populations is inappropriate. There is therefore a strong need to use large databases with information on ethnic haplotype frequencies. Current practice for estimating the probability of a genotype given the defendant's genotype is to use equations accounting for population substructure. See recommendation 4.2 from the National Research Council's 1996 report 'The Evaluation of Forensic DNA Evidence' (NRC Press, 1996).

Bayesian networks: a revolution?

All sort of heterogeneous information representing evidences in forensic science can be incorporated into BNs [45–49]. A BN is a graphical model, represented by a directed acyclic graph. BN describe the conditional dependence relationships between variables, i.e. joint probability distributions over all the variables in a graph. Nodes in the graph represent variables (they can be binary, multidimensional and continuous), and a directed link between node A and node B indicates that A is a parent of B i.e. B is conditionally dependent on A. A BN has at least these four components: (i) Priors which represent initial beliefs about nodes in the network; (ii) Conditional Probability Distribution (CPD) i.e the conditional probabilities between connected nodes; (iii) Posteriors which represent the computed beliefs after the evidence has been accounted for; (iv) Evidence i.e. the observations from extracted features.

The structure of the network can be based on expert knowledge, or learnt algorithmically if there is sufficient training data available. Bayesian inference is based on Bayes theorem: $P(A|B) = P(B|A)P(A)/P(B)$ where $P(A|B)$ is the conditional probability of A, given B, i.e the posterior probability depends on $Pr(B|A)$ which is the conditional probability of B given A, i.e the likelihood, on P(A) which is the prior probability i.e. the marginal probability of

A and on P(B) which is the prior or marginal probability of B and acts as a normalizing constant. Inference can be handled by marginalization—the summing out of all the irrelevant variables. In many cases BNs are quite slow and need powerful computers: the calculation of posterior takes exponential time, and the size of the network is exponential in the number of nodes, so it is often very inefficient. A number of more efficient methods and optimizations such as variable elimination and dynamic programming algorithms exist.

Nevertheless BN are powerful inference engines, as when the value of a node is unknown, the probability of it having a certain value conditional on the available evidence can be estimated using Maximum Likelihood. Maximum likelihood estimation trains the CPD to maximize the probability of assessing the evidence, given the distribution. Continuous variables can be dealt with in two ways: transform values into discrete one, that is, assigning a value to each bin in a histogram, or modeling the data with some continuous distributions such as for example Poisson and Gaussian. Given a graph where forensic evidences are the nodes connected by arrows, BNs can be used as a tool for lawyers to analyze evidence in judicial cases, and as an aid for constructing legal arguments. They can help determine to what extent the set of forensic evidences support the claims of the prosecution (defence) versus those of the defence (prosecution). For example, assessing the impact of a certain piece of forensic evidence on a given case involves (i) formalizing the respective claims of the prosecution and the defence, (ii) computing the probability that the evidence is found, given that the claim of the prosecution is true, and the probability that the evidence is found, given that the claim of the defence is true, and (iii) dividing the former probability by the latter to determine the likelihood ratio and compare it with the chi-square distribution or computing the *P*-value, i.e. the probability of obtaining by chance a similar result.

A widely used BN software is HUGIN, which is commercial but provides a free evaluation version available at http://www.hugin.dk. Peter Green from Bristol University (http://www.stats.bris.ac.uk/~peter/) has developed 'Grappa', which is a free suite of functions in R for calculating marginal and conditional probability distributions on collections of variables. It does a similar job to the Hugin, being 'programmable'; and with a text-based interface. Xmeta [47] is a Java BN focused on forensic inference. BNJ is an open-source suite of software tools for research and development using graphical models of probability (http://bndev.sourceforge.net). It is implemented in Java and distributed under the GNU General Public License (GPL).

Other software are described and can be downloaded from: http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html

Interesting cases in which Bayesian statistics has pointed to mistakes in judiciary cases ('Prosecutor's fallacy') are those of O.J. Simpson and Sally Clark. Press release by the Royal Statistical Society about the Sally Clark case can be found at
http://www.rss.org.uk/docs/RoyalStatistical%20Society.doc

http://www.colchsfc.ac.uk/maths/dna/discuss.htm

http://dna-view.com/profile.htm.

STAYING AWAY FROM FUNCTIONAL GENOMIC REGIONS IS NOT ENOUGH

Decisions on the use of genetic data have far-reaching consequences and reflect our society's basic values and priorities. Work still need to dispel popular myths about the infallibility of technologies. Not mentioning mistakes and different sensitivity in using the technologies or reporting results, particularly a good understanding of statistics (concepts of statistical tests and power). There is strong need of raising the discussion level on privacy rights, the nature of consent in relation to crime investigation, the confidentiality of genetic information and the proper form of database governance and maintenance. DNA profiles are different from fingerprints, which are useful only for identification. DNA can provide insights into many intimate aspects of a person and their families including susceptibility to particular diseases, legitimacy of birth, and perhaps predispositions to certain behaviors and sexual orientation. This increases the potential for genetic discrimination by government, insurers, employers, schools, banks and others. For example the FSS center has investigated the possibility of predicting physical characteristics of individuals for some time. It has created a 'Red Hair database' which claims to identify '84% of redheads', and now offer the police an 'ethnic inference service' which claims the capacity to discern—with unknown degrees

of certainly—ethnic origin from DNA profiles (see http://www.publications.parliament.uk/pa/ld199900/ldselect/ldsctech/115/115we20.htm).

A linkage is very often observed between heritable diseases and some repetitive DNA loci. Actually the observation of certain changes in repetitive DNA stretches allows to predict the probability to develop symptoms of a genetic disease. It is sometimes suggested that forensic DNA analysis uses STR loci composed of tri-, tetra- and pentameric core units, while most known repetitive DNA stretches that are linked to diseases have a dimeric substructure. This is not always true and should not be taken as a solution to genetic privacy concerns. Certainly a necessary-but-not-sufficient criterion to defend privacy rights is to forbid the inclusion of genetic susceptibilities information in crime databases. This is unlikely to be sufficient given the easiness to code software that access several databases. A possibility is to restrict the analysis to genomic regions which are informative for the identification but not for functional characters but such regions should be still characterized. Perhaps codon's third positions in some genes which are not under positive selection may be uninformative for diseases.

Bioinformatics institutions and associations should promote discussion and effective actions in cases of genetic information abuse. In other words, there is need to start sort of associations in forensic bioinformatics such as the CPSR [50] (Computer Professionals for Social Responsibility http://www.cpsr.org/) and the CCSR (Centre for Computing and Social Responsibility: http://www.ccsr.cms.dmu.ac.uk/) which are public interest alliances of computer scientists and others concerned about the impact of computer technology on society. They work to influence decisions regarding the development and use of computers.

As concerned citizens, bioinformaticians should direct public attention to critical choices concerning the applications of biocomputing and how those choices affect society, foster and support more public discussion of, and public responsibility for decisions involving the use of technology in systems critical to society not only in general but also on key cases. Moreover they should address the problem that <1% of court cases involving DNA are reviewed by experts working on behalf of the defendants. It is important that defendants and their attorneys are made aware of any commonly encountered problems that have occurred during the typing or the interpretation and comparative analysis of the DNA evidence associated with their case.

The main reason why Germany and France do not allow collection of 'nonintimate' biological samples and do not store STR profiles in a database is not the fear of misuse or mistyping but the view that any sampling of body tissue followed by storage of data violates the individual's privacy.

In the United Kingdom the collection of such samples is now allowed if the suspect's offense may lead to imprisonment. In other European countries which are establishing DNA databases, biological material of a suspect is allowed to be processed only if the alleged crime is severe enough to lead to a possible imprisonment for 1 year or more.

Remarkably, Professor Sir Alec Jeffreys [51–53], who developed the technique for the genetic markers (http://www.le.ac.uk/genetics/ajj/index.html) has stressed that the practice of storing the genetic profiles of suspects in the UK who have been cleared of any crime is highly discriminatory and measures should be taken to safeguard against particular individuals or groups being targeted. He has proposed the creation of a national database, storing the profiles of the entire UK population, which would be managed by an independent body.

Finally, the Council of Europe convention for the protection of human rights with regard to the application of biomedicine and the United Nations outline for an international declaration on genetic data should be the source of more decision power. They should play a more vital role in charting and homogeneizing the differing judicial and policy frameworks which exist throughout the EU for the use of DNA in support of criminal investigations.

Acknowledgment

The authors thank the organizers of the 'Forensic DNA course' held in September 2005 at the DNA Learning Center and The 'Fondazione Marino Golinelli' in Bologna. They also thank the EU F6 Bioinfogrid project.

References

1. Sherry ST, Sozer A, Walsh A. Epidemiology. DNA identifications after the 9/11 World Trade Center attack. *Science* 2005;**310**:1122–3.
2. Huffine E, Crews J, Kennedy B, *et al*. Mass identification of persons missing from the break-up of the former Yugoslavia: structure, function, and role of the International Commission on Missing Persons. *Croat Med J* 2001;**42**: 271–5.

3. Gill P. DNA as evidence — the technology of identification. *N Engl J Med* 2005;**352**:26.
4. Cash HD, Hoyle JW, Sutton AJ. Development under extreme conditions: forensic bioinformatics in the wake of the World Trade Center disaster. *Pac Symp Biocomput* 2003; **1**:638–53.
5. Estrela P, Stewart AG, Yan F, Migliorato P. Field effect detection of biomolecular interactions. *Electrochemica Acta* 2005;**50**:4995–5000.
6. Estrela P, Stewart AG, Migliorato P, Maeda H. Label-Free Detection of DNA Hybridization with Au/SiO2/Si Diodes and Poly-Si TFTs. *Technical Digest of 2004 IEDM - International Electron Devices Meeting*. San Francisco CA: IEEE, 1009–12.
7. Califano J. Interview with Amanda Parrish. Mitochondrial genome scan finds cancer mutations in saliva DNA samples. *Affymetrix Microarray Bull* 2005;**1**:17–9.
8. Foreman LA, Smith AFM, Evett IW. Bayesian analysis of deoxyribonucleic acid profiling data in forensic identification applications. *J R S S A* 1997;**160**:429–69.
9. Evett IW, Weir BS. *Interpreting DNA evidence*1998. Sunderland: Sinauer, 1998.
10. Redon R, Ishikawa S, Fitch KR, *et al*. Global variation in copy number in the human genome. *Nature* 2006;**444**: 444–54.
11. Piazza F, Lio' P. Statistical analysis of simple repeats in the human genome. *Physica A* 2005;**347**:472–88.
12. Mark B. DNA typing in forensic medicine and in criminal investigations: a current survey. *Naturwissenschaften* 1997;**84**: 181–8.
13. The International HapMap Consortium. The Hapmap project. *Nature* 2005;**437**:1299–320; http://www.hapmap.org
14. Mitchell RJ, Kreskas M, Baxter E, *et al*. An investigation of sequence deletions of amelogenin (AMELY), a Y-chromosome locus commonly used for gender determination. *Ann Hum Biol* 2006;**33**:227–40.
15. Zwick M, Mcafee F, Cutler DJ, *et al*. Microarray-based resequencing of multiple Bacillus anthracis isolates. *Genome Biology* **6**:R10doi:10.1186/gb-2004-6-1-r10.
16. Bang-Ce Y, Xiaohe C, Ye F, *et al*. Simultaneous genotyping of DRB1/3/4/5 Loci by oligonucleotide microarray. *J Mol Diagn* 2005;**7**:592–9.
17. Kemp JT, Davis RW, White RL, *et al*. A novel method for STR-based DNA profiling using microarrays. *J forensic Sci* 2005;**50**:1109–13.
18. Yancy HF, Mohla A, Farrell DE, Myers MJ. Evaluation of a rapid PCR-based method for the detection of animal material. *J Food Prot* 2005;**68**:2651–5.
19. Delgado S, Girondot M, Sire J. Molecular evolution of amelogenin in mammals. *J Mol Evol* 2005;**60**:12–30.
20. Rahimi M, Heng NC, Kieser JA, Tompkins GR. Genotypic comparison of bacteria recovered from human bite marks and teeth using arbitrarily primed PCR. *J Appl Microbiol* 2005;**99**:1265–70.
21. Graham EA, Tsokos M, Rutty GN. Can post-mortem blood be used for DNA profiling after peri-mortem blood transfusion? *Int J Legal Med* 2005;**10**:1–6.
22. Stephens JC, Schneider JA, Tanguay DA, *et al*. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001;**293**:489–93.
23. Lio' P, Morton N. Comparison of parametric and nonparametric methods to map oligogenes by linkage. *Proceedings of the National Academy of Sciences* 1997;**94**: 5344–8.
24. Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes*. Princeton, New Jersey: Princeton University Press.
25. de Bakker PIW, Yelensky R, Ipe'er I, *et al*. Efficiency and power in genetic association studies. *Nature Genetics* 2005; **37**:1217–23.
26. Dawid AP, Mortera J, Pascali VL, van Boxel D. Probabilistic expert systems for forensic inference from genetic markers. *Scand J Statist* 2002;**29**:577–95.
27. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of relationship errors. *Bioinformatics* 2001;**17**:742–3.
28. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;**21**:263–5.
29. Akey J, Jin Li, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Human Genet* 2001;**9**:291–300.
30. Wiltshire S, Morris AP, McCarthy MI, Cardon LR. How useful is the fine-scale mapping of complex trait linkage peaks? Evaluating the impact of additional microsatellite genotyping on the posterior probability of linkage. *Genet Epidemiol* 2005;**28**:1–10.
31. Pettersson F, Jonsson O, Cardon LR. GOLDsurfer: three dimensional display of linkage disequilibrium. *Bioinformatics* 2004;**20**:3241–3.
32. Brenner CH. Symbolic kinship program. *Genetics* 1997;**145**: 535–42.
33. Schneider S, Roessli D, Excoffier L., Arlequin: a software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Department of Anthropology, University of Geneva.
34. Presciuttini S, Cerri N, Turrina S, *et al*. Validation of a large Italian database of 15 STR loci. *Forensic Sci Int* 2006;**156**: 266–8.
35. Gill P, Foreman L, Buckelton JS, *et al*. Analysis of DNA databases across Europe compiled by the ENFSI group. *Forensic Sci Int* 2003;**131**:184–96.
36. Biesecker LG, Bailey-Wilson JE, Ballantyne J, *et al*. DNA Identifications After the 9/11 World Trade Center Attack. *Science* 2005;**310**:1122–3.
37. Primorac D, Anelinoviæ S, Definis-Gojanoviæ M, *et al*. Identification of war victims from mass graves in Croatia and Bosnia and Herzegovina through the use of DNA typing and standard forensic methods. *J Forensic Sci* 1996;**41**: 891–4.
38. Holland MM, Cave CA, Holland CA, Bille TW. Development of a quality, high throughput DNA analysis procedure for skeletal samples to assist with the identification of victims from the World Trade Center attacks. *Coatian Med J* 2003;**44**:264–72.
39. Balding DJ. Estimating products in forensic identification. *J Am Stat Assoc* 1995;**90**:839–44.
40. Balding DJ, Nichols RA. DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci Int* 1994;**64**:125–40.

41. Ewens WJ, Grant GR. *Statistical Methods in Bioinformatics. Statistics for Biology and Health*. New York: Springer, 2001.
42. Lang K. *Mathematical and Statistical Methods for Genetic Analysis. Statistics for Biology and Health*. 2nd edn. New York: Springer, 2001.
43. Shoemaker J, Painter I, Weir BS. A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics*;**149**:2079–88.
44. Weir BS. Genetic data analysis II. Sunderland, Massachussetts: Sinauer, 1996.
45. Taroni F, Aitken C, Garbolino P, Biedermann A. *Bayesian Networks and Probabilistic Inference in Forensic Science*. New York: Wiley, 2006.
46. Hepler AB. Improving forensic identification using Bayesian networks and relatedness estimation: allowing for population substructure. E-Book ProQuest Information and Learning (April 23, 2006)
47. Duval T, Jouga B, Roger L. XMeta: a Bayesian approach for computer forensics. *In: Work in Progress Session of the Annual Computer Security Applications Conference (ACSAC)*. Tucson, december 2004.
48. Aitken C, Taroni F, Garbolino P. A graphical model for the evaluation of cross-transfer evidence in DNA probes. *Theor Popul Biol* 2003;**63**.
49. Mortera J, Dawid AP, Lauritzen SL. Probabilistic expert systems for DNA mixture profiling. *Theor Popul Biol* 2003; **63**:191–205.
50. Kling R., Computerization and Controversy: Value Conflicts and Social Choices. 2nd edn London: Academic Press, 1991.
51. http://news.bbc.co.uk/2/hi/science/nature/3636050.stm
52. Jeffreys AJ, Wilson V, Thein SL. Hypervariable 'minisatellite' regions in human DNA. *Nature* 1985; **314**:67–73.
53. Jeffreys AJ, MacLeod A, Tamaki K, *et al*. Minisatellite repeat coding as a digital approach to DNA typing. *Nature* 1991; **354**:204–209.