ARCH. MED. SĄD. KRYMINOL., 2010, LX, 118-126 PRACE ORYGINALNE

Jarosław Berent

DNAStat wersja 2.1 – program do obsługi bazy danych profili genetycznych oraz do obliczeń biostatystycznych

DNAStat, version 2.1 – a computer program for processing genetic profile databases and biostatistical calculations

Z Katedry i Zakładu Medycyny Sądowej Uniwersytetu Medycznego w Łodzi
Kierownik: prof. dr hab. n. med. J. Berent

W pracy przedstawiono nową wersję programu DNAStat wersja 2.1 do obsługi bazy danych profili genetycznych oraz do obliczeń biostatystycznych. Rozpowszechnienie się badań DNA, wykorzystywanych dla potrzeb wymiaru sprawiedliwości, spowodowało konieczność opracowania odpowiednich programów komputerowych ułatwiających pracę biegłego genetyka. Programy takie muszą przede wszystkim rozwiązywać dwa problemy, tj. problem szeroko pojętej obsługi i archiwizacji danych oraz problem obliczeń biostatystycznych. Ponadto, z uwagi na coraz częstsze występowanie we współczesnym świecie zagrożeń terrorystycznych i klęsk żywiołowych, ważna jest możliwość analizy zgromadzonych danych pod kątem odnajdywania osób spokrewnionych. Takim programem jest właśnie DNAStat wersja 2.1. Program został opracowany w roku 2005 – wersja 1.0. W roku 2006 powstały wersja 1.1 i następnie 1.2. Wersje 1.1 i 1.2 usuwały jedynie kilka drobnych niedogodności z wersji 1.0, natomiast co do istoty nie różniły się wiele od pierwszej wersji. Wersja 2.0 powstała w roku 2007 – podstawowym udoskonaleniem programu w tej wersji było wprowadzenie możliwości obliczeń grupowych, których potencjalnym zastosowaniem jest identyfikacja osobnicza ofiar zamachów terrorystycznych lub katastrof masowych. W obecnej wersji 2.1 dodano możliwość obsługi programu – poza językiem polskim – także w języku angielskim.

Słowa kluczowe: biostatystyka, identyfikacja osobnicza, zamachy terrorystyczne, katastrofy masowe, badania ojcostwa, badania dowodów rzeczowych, bazy danych

This paper presents the new DNAStat version 2.1 for processing genetic profile databases and biostatistical calculations. The popularization of DNA studies employed in the judicial system has led to the necessity of developing appropriate computer programs. Such programs must, above all, address two critical problems, i.e. the broadly understood data processing and data storage, and biostatistical calculations. Moreover, in case of terrorist attacks and mass natural disasters, the ability to identify victims by searching related individuals is very important. DNAStat version 2.1 is an adequate program for such purposes. The DNAStat version 1.0 was launched in 2005. In 2006, the program was updated to 1.1 and 1.2 versions. There were, however, slight differences between those versions and the original one. The DNAStat version 2.0 was launched in 2007 and the major program improvement was an introduction of the group calculation options with the potential application to personal identification of mass disasters and terrorism victims. The last 2.1 version has the option of language selection – Polish or English, which will enhance the usage and application of the program also in other countries.

Key words: biostatistics, personal identification, terrorism, mass disasters, paternity testing, forensic cases, databases

WPROWADZENIE

Rozpowszechnienie się badań DNA, wykorzystywanych dla potrzeb wymiaru sprawiedliwości, spowodowało konieczność opracowania odpowiednich programów komputerowych ułatwiających pracę biegłego genetyka. Programy takie muszą przede wszystkim rozwiązywać dwa problemy, tj. problem szeroko pojętej obsługi i archiwizacji danych oraz problem obliczeń biostatystycznych. Ponadto z uwagi na coraz częstsze występowanie we współczesnym świecie zagrożeń terrorystycznych i klęsk żywiołowych, ważna jest możliwość analizy zgromadzonych danych pod kątem odnajdywania osób spokrewnionych. Takim programem jest właśnie DNAStat wersja 2.1. Program został opracowany przez prof. dr. hab. n. med. Jarosława Berenta, kierownika Katedry i Zakładu Medycyny Sądowej Uniwersytetu Medycznego w Łodzi przy wykorzystaniu obsługi informatycznej firmy Laser Systemy Informatyczne S.A. w Łodzi. Program powstał w ramach grantu na prace własne Uniwersytetu Medycznego w Łodzi nr 502-11-785(35).

WCZEŚNIEJSZE WERSJE PROGRAMU

Program został opracowany w roku 2005 – wersja 1.0 [4, 5]. W roku 2006 powstały wersja 1.1 i następnie 1.2 [6]. Wersje 1.1 i 1.2 usuwały jedynie kilka drobnych niedogodności z wersji 1.0, natomiast co do istoty nie różniły się wiele od pierwszej wersji. Wersja 2.0 powstała w roku 2007 – podstawowym udoskonaleniem programu w tej wersji było wprowadzenie możliwości obliczeń grupowych, których potencjalnym zastosowaniem jest identyfikacja osobnicza ofiar zamachów terrorystycznych lub katastrof masowych. W obecnej wersji 2.1 dodano możliwość obsługi programu – poza językiem polskim – także w języku angielskim.

INSTALACJA PROGRAMU

Plik instalacyjny programu DNAStat o nazwie DNAStat_setup.exe można uzyskać nieodpłatnie po zgłoszeniu e-mailowym do autora programu (J.Berent@eranet.pl). Po jego uruchomieniu cała instalacja następuje automatycznie i trwa około jednej minuty. Program zostaje zainstalowany do katalogu: C:\Program Files\DNAStat\, a na pulpicie umieszczona zostaje ikona o nazwie DNAStat 2.1.

INTRODUCTION

The popularization of DNA studies employed in the judicial system has led to the necessity of developing appropriate computer programs. Such programs must, above all, address two critical problems, i.e. the broadly understood data processing and data storage, and biostatistical calculations. Moreover, in case of terrorist attacks and mass natural disasters, the ability to identify victims by searching related individuals is very important. DNAStat version 2.1 is an adequate program for such purposes. The program has been elaborated by Professor Jaroslaw Berent, the Head of the Department of Forensic Medicine, Medical University of Lodz, with the cooperation of Laser Systemy Informatyczne S.A. in Lodz. The project was supported by Medical University of Lodz, grant no. 502-11-785(35).

PREVIOUS PROGRAM VERSIONS

The DNAStat version 1.0 was launched in 2005 [4,5]. In 2006, the program was updated to 1.1 and 1.2 versions [6]. There were, however, slight differences between those versions and the original one. The DNAStat version 2.0 was launched in 2007 and the major program improvement was an introduction of the group calculation options with the potential application to personal identification of mass disasters and terrorism victims. The last 2.1 version has the option of language selection – Polish or English, which will enhance the usage and application of the program also in other countries.

PROGRAM INSTALLATION

The DNAStat installation file named DNAStat_seyup.exe is freely available from its author (J.Berent@eranet.pl). After it is run, the installation starts automatically and lasts for about one minute. The program is installed to the directory: C:\Program Files\DNAStat\, and the “DNAStat 2.1“ icon is placed on a desktop.

Program można odinstalować przez aplet „Dodaj lub usuń programy" w panelu sterowania.

W katalogu C:\Program Files\DNAStat\ Databases\PL\ zostają automatycznie umieszczone dwa pliki baz danych: Baza.gdb i Pusta.gdb. Ta pierwsza zawiera już wprowadzone dane populacyjne dla 15 loci STR z zestawu multipleksowego Identifiler® dla n=250 alleli. Dane te pochodzą z publikacji: Jacewicz R., Berent J., Prośniak A., Gałecki P., Florkowski A., Szram S.: Population genetics of the Identifiler system in Poland. International Congress Series 2004, 1261, 229-232 [10]. Wprowadzone tam współczynniki mutacji pochodzą zaś z raportu: 2001 Paternity Testing Workshop of the English Speaking Working Group of the International Society for Forensic Genetics [13], przy czym współczynniki mutacji obliczono jako iloraz sumy niezgodności w układzie matka-dziecko i ojciec-dziecko przez całkowitą liczbę mejoz.

Natomiast baza o nazwie Pusta.gdb nie zawiera żadnych danych i stanowi miejsce, gdzie użytkownik może umieszczać swoje własne dane. Bazy te mogą być dowolnie kopiowane i mogą mieć dowolnie zmieniane nazwy. Również ich lokalizacja w komputerze może być dowolna, niekoniecznie w domyślnym miejscu, czyli katalogu C:\Program Files \DNAStat\Databases\PL\.

Podczas instalacji w katalogu C:\Program Files\DNAStat\ Examples\ PL\ zostaje umieszczonych siedem plików z przykładowymi danymi. Są to dwa pliki programu Microsoft® Office Excel: Import_1.xls i Import_2.xls. Pliki programu Excel zawierające genotypy, które użytkownik chciałby zaimportować do programu muszą mieć identyczną konstrukcję, tzn. w pierwszym wierszu muszą się znajdować opisy kolumn, a w kolejnych wierszach muszą się znajdować dane. Pierwsza kolumna o nazwie Numer zawiera numer sprawy (musi to być liczba), następne kolumny o nazwach układów zawierają genotypy (pierwsza kolumna nosi nazwę układu, np. D8S1179, a druga nazwę układu z rozszerzeniem „_2", np. D8S1179_2). W ostatniej kolumnie o nazwie uwagi może znajdować się dowolny tekst. Kolejne cztery pliki z tego folderu to pliki tekstowe Dane_1.txt, Dane_2.txt, Dane_3.txt i Dane_4.txt generowane przez sekwenator (zapis w standardzie CODIS). Zawierają one przykładowe dane, które mogą być automatycznie importowane przez program. Pliki te mają postać:

The program can be easily uninstalled by means of the "add/remove programs" applet in the control panel.

The C:\Program Files\DNAStat\Databases\ EN\ directory contains 2 database files named "Default_base.gdb" and "Empty_base.gdb". The first one already includes the population data of 250 alleles in a range of 15 STR loci contained in an Identifiler® kit. The source of the population data is the article: Jacewicz R., Berent J., Prosniak A., Galecki P., Florkowski A., Szram S.: Population genetics of the Identifiler marker in Poland. International Congress Series 2004, 1261, 229-232 [10]. Mutation ratios described in the article were taken from the Paternity Testing Workshop Report of the English Speaking Working Group of the International Society for Forensic Genetics launched in 2001 [13], and they were counted by dividing the sum of mother–child and father–child inconsistencies by the total meioses number.

The "Empty_base.gdb" file does not contain any data and thus can be used for inserting user data. Those databases can be optionally processed by changing name or location.

During the installation process, in the C:\ Program Files\DNAStat\ Examples\EN\ directory seven exampling files are placed. They are two Microsoft Excel files: "Import_1.xls" and "Import_2.xls". Those files, containing genotypes that the user wants to import to DNAStat program, have to be constructed identically, i.e. with a description in the first line and genetic data in the following lines. The first column named "Number" contains case number, the next columns contain genotypes (two columns per one marker, i.e. D8S1179 and D8S1179_2). The last column named Remarks may include any text. The other four files in this folder are sequencer generated CODIS files named "Data_1.txt", "Data_2.txt", "Data_3.txt" and "Data_4.txt". They contain data that may be automatically imported by the program. The files are in the following form:

Sample Info	Category	Peak 1	Peak 2
_207pl_ID	D8S1179	12	13
_207pl_ID	D21S11	31	32.2
_207pl_ID	D7S820	8	12
itd.			

Siódmy plik o nazwie Populacja.txt stanowi przykładowy plik z danymi populacyjnymi pięciu układów SNP, pochodzących z publikacji: Bąbol-Pokora K., Prośniak A., Jacewicz R., Berent J.: Pentapleks SNP – rozkład częstości alleli w populacji centralnej Polski. Arch. Med. Sąd. i Krym. 2006, 56(4), 228-231 [3]. Plik ten ma postać:

The seventh file named "Population.txt" is an example of population data for five SNP loci, which originates from the article: Babol-Pokora K., Prosniak A., Jacewicz R., Berent J.: [SNP pentaplex – the allele frequency database of central Poland population]. Arch. Med. Sadowej Kryminol. 2006, 56(4), 228-231 [3]. The file is in the following form:

```
*rs2294067/0,00000/160
C/0,48800
G/0,51200

*rs2070764/0,00000/160
T/0,62500
A/0,37500

*rs1063739/0,00000/160
A/0,48800
C/0,51200

*rs2282160/0,00000/160
G/0,51300
A/0,48700

*rs2277216/0,00000/160
C/0,79400
T/0,20600
```

Podczas instalacji w katalogu C:\Program Files\DNAStat\Help\PL\ zostaje umieszczony plik DNAStat_2.1_PL.pdf, który zawiera opis programu.

During the installation process, in the C:\ Program Files\DNAStat\Help\ EN\, a directory "DNAStat_2.1_EN.pdf" file with program description is placed.

ROZPOCZĘCIE PRACY Z PROGRAMEM

Po zainstalowaniu programu DNAStat należy wprowadzić własną bazę populacyjną albo – na początek – skorzystać z bazy instalowanej z programem Baza.gdb. Następnie należy wprowadzić genotypy i inne dane o badanych osobach albo – na początek – zaimportować jeden lub oba pliki zawierające genotypy badanych osób lub śladów Import 1.xls lub Import 2.xls. W tym momencie program jest gotowy do użycia, tzn. do przeszukiwania bazy danych lub do obliczeń biostatystycznych.

GETTING STARTED

After the DNAStat program is installed, new population database has to be inserted or "Default_base.gdb" can be used. Next, genotypes and other information have to be inserted or "Import_1.xls" or "Import_2.xls" files have to be imported. The program is then ready for genetic data processing and performing biostatistical calculations.

FUNKCJE PROGRAMU

Program DNAStat umożliwia tworzenie własnej bazy danych zawierającej: dane populacyjne o wykorzystywanych układach (nazwy alleli i ich częstości, współczynniki mutacji i wielkość populacji), dane o badanych osobach lub śladach (genotypy i różne informacje administracyjne) oraz dane o zleceniodawcach opinii (nazwa i adres). Wszystkie składniki tej bazy mogą być w dowolny sposób modyfikowane lub usuwane, jak również mogą być w każdym momencie dodawane nowe elementy. Tak utworzona baza danych jest zapisywana w postaci pojedynczego pliku *.gdb. Program DNAStat umożliwia korzystanie z wielu plików *.gdb zawierających różne bazy danych. Przełączanie pomiędzy poszczególnymi bazami następuje z poziomu programu.

Dane populacyjne, dotyczące wykorzystywanych układów, mogą być wprowadzane allel po allelu z klawiatury lub mogą być importowane automatycznie z pliku tekstowego *.txt przygotowanego np. w programie Microsoft® Notatnik lub EditPad. Plik taki ma postać: w pierwszej linii gwiazdka, nazwa układu łamane przez częstość mutacji, łamane przez wielkość bazy i w kolejnych liniach nazwa allela łamane przez jego częstość. Po liniach zawierających dane o pierwszym układzie następuje jedna linia wolna i w następnych liniach podane są dane o kolejnych układach. Dane wprowadzone do programu mogą też być eksportowane w formie takiego samego pliku. Zaimportowanie pliku z danymi usuwa wcześniej wprowadzone informacje o układach, nie naruszając bazy populacyjnej genotypów (osób). Taka opcja umożliwia szybkie i łatwe modyfikowanie posiadanej bazy np. o nowe układy lub allele oraz prowadzenie obliczeń dla różnych baz.

Genotypy badanych osób lub śladów mogą być również wprowadzane allel po allelu z klawiatury lub mogą być importowane automatycznie z plików. Program jest w stanie zaimportować pliki tekstowe *.txt generowane przez sekwenator lub pliki programu Microsoft® Office Excel *.xls.

Baza danych może być dowolnie przeszukiwana według takich pól, jak: numer sprawy, imię i nazwisko, data pobrania, itp. Możliwe jest również wyszukiwanie według genotypów, tzn. po wpisaniu (lub zaimportowaniu) interesującego nas genotypu program automatycznie wyszuka wszystkie osoby lub ślady z bazy, które posiadają identyczny genotyp. Ta ostatnia

PROGRAM FUNCTIONS

The DNAStat program enables the user to create a personal database that includes: population data concerning markers (allele names and frequencies, mutation ratios and population size), data concerning investigated individuals and samples (genotypes and administrative information) and information about ordering institutions (name and address). All the components of this database can be optionally modified or deleted, just as new elements can be added any time. The elaborated database is saved as a single *.gdb file. DNAStat allows for the use of many *.gdb files containing different databases, which can be optionally selected while running the program.

Population data of the investigated markers can be inserted manually (allele by allele) or imported automatically from sequencer generated files in the *.txt format for Microsoft® Notepad or EditPad. The form of such file is: asterisk in the first line, marker's name / mutation frequency / database size, and, in the following lines, allele name slash allele frequency. There is one blank line between the data concerning different markers. The inserted data can be exported in the same form. After the data file is imported, previous information concerning the markers is deleted without changing the population database. This allows for a quick and easy modification of the current base, e.g. by adding new alleles and markers, or making calculations for different bases.

Genotypes of investigated individuals and samples can be also inserted manually (allele by allele) or imported automatically from sequencer generated files in the *.txt format or in .xls format for Microsoft® Office Excel.

The database can be searched optionally according to: case number, name and surname, date of material collection, etc. It is possible to search the base via genotypes, i.e. after the genotype of interest is imported or inserted manually, the program will search automatically all individuals and samples sharing the same genotype. This works both for full profiles and for deficient ones, e.g. when only a partial genotype is imported, all samples sharing the same genotype in a range of investigated loci are indicated, while the remaining loci are omitted.

The same is true for searching only one allele (per locus). The program will search all samples having the compatible allele in the investigated locus, while the other allele will not be taken into

funkcja działa zarówno dla pełnych, jak i dla niepełnych genotypów, tzn. przy zadaniu genotypu przykładowo tylko w jednym układzie program wyszuka wszystkie osoby lub ślady, które mają taki genotyp w tym konkretnym układzie, pomijając informacje dla innych układów.

To samo dotyczy zadania informacji tylko o jednym allelu. Program wyszuka wówczas wszystkie osoby lub ślady, dla których jeden z alleli jest zgodny z zadanym, pomijając informacje o drugim allelu. Takie możliwości wyszukiwania mogą być przydatne dla zdegradowanych materiałów, gdzie pełny genotyp nie zawsze jest dostępny.

Program umożliwia także prowadzenie obliczeń biostatystycznych dla genotypów osób lub śladów wprowadzonych do bazy. Dla analizy śladów biologicznych program oblicza częstość profilu f oraz prawdopodobieństwo p(X|X), a przy analizie ojcostwa/macierzyństwa program oblicza szansę ojcostwa/macierzyństwa (ang. paternity/maternity index) i prawdopodobieństwo ojcostwa/macierzyństwa W (niem. Wahrscheinlichkeit) w układzie pełnej trójki, w układzie mężczyzna-dziecko (bez matki) i w układzie kobieta-dziecko (bez mężczyzny).

Program pozwala również na obliczenia grupowe, wykonując zadany rodzaj obliczeń dla wszystkich genotypów (osób) w bazie, czego potencjalnym zastosowaniem jest identyfikacja osobnicza ofiar zamachów terrorystycznych lub katastrof masowych. Możliwe są trzy rodzaje obliczeń. Pierwszy to poszukiwanie osób spokrewnionych w układzie ojciec-dziecko. Po wskazaniu wybranej osoby program przeprowadzi obliczenia szansy ojcostwa PI dla tej osoby w parze z kolejno wszystkimi pozostałymi osobami z bazy, a następnie uszereguje wyniki wg PI począwszy od największej do najmniejszej wartości. Analogiczne obliczenia są możliwe w układzie matka-dziecko i dla pełnej trójki matka-dziecko-ojciec, gdzie po wskazaniu obu rodziców program prowadzi obliczenia dla wszystkich pozostałych osób z bazy podstawiając je jako dziecko dla wybranej pary.

Wyniki wszystkich obliczeń mogą być eksportowane w formie plików *.xls, odczytywanych przez program Microsoft® Office Excel. Taka opcja umożliwia łatwe przeniesienie wyników dokonanych obliczeń do dowolnego edytora tekstów stosowanego w poszczególnych laboratoriach przy pisaniu opinii. Postępowanie takie zmniejsza możliwość popełnienia błędów poprzez wyeliminowanie ręcznego przepisywania wyników. Wyniki obliczeń mogą być także drukowane.

account. This option can be useful for analyzing degraded materials, with partial genetic profiles.

This program also allows for performing biostatistical calculations of the genotypes in the database. The program analyzes biological evidence by calculating the unconditional f and conditional p(X|X) profile frequency and it allows for analyzing paternity / maternity cases by calculating the paternity / maternity index PI / MI and probability of paternity W for full, motherless and fatherless cases.

The program also allows for group calculations, by applying the given calculation to all the genotypes (individuals) within the base, which can be potentially applied to personal identification of mass disasters and terrorism victims. Three calculation types are possible: the first is the searching of related individuals among father–child settings. The program will calculate the Paternity Indexes for the investigated individual paired with every individual existent in the base, and present the PI results from the highest value to the lowest one. The same is true for mother-child and mother-child-father settings, where both parents are given and every individual in the base is considered a child.

The results of all calculations can be exported in the .xls format for Microsoft® Office Excel. It allows for an easy transfer of the results into any text processor, which decreases the possibility of making mistakes, often caused by manual copying, and the results can be printed.

ANALIZA ŚLADÓW BIOLOGICZNYCH

Program DNAStat w analizie śladów biologicznych oblicza częstość profilu f oraz prawdopodobieństwo p(X|X), przy możliwości uwzględnienia współczynnika pochodzenia F_{ST} oraz zadania dolnego progu częstości alleli CP. F_{ST} – jest to współczynnik pochodzenia (ang. coancestry coefficient). Jest on definiowany dla całej populacji i określa, jakie jest prawdopodobieństwo, że dwa allele wzięte losowo od dwóch, również losowo, wybranych osób z populacji (jeden allel od jednej osoby i drugi od drugiej) są identyczne z pochodzenia (ang. identical by descent). Współczynnik ten jest wyrazem pewnej bliżej nieokreślonej liczby nieznanych wspólnych przodków w poprzednich pokoleniach. W typowych populacjach wynosi około 0.01, natomiast dla małych, odosobnionych populacji lub populacji trudno poddających się asymilacji może wynosić do 0.03 [1, 2]. CP – jest to dolny próg częstości alleli stosowany dla zapobieżenia przeszacowania częstości profili DNA wynikającego ze zbyt małych częstości allelicznych (ang. ceiling principle). Stosowanie progów zalecał I Raport NRC z roku 1992 (CP=0.1 dla interim ceiling principle albo CP=0.05 dla ceiling principle) [11]. Współcześnie nie zaleca się stosowania żadnych takich progów (CP=0) [12].

Częstość profilu f jest liczona najpierw dla każdego układu i dalej częstości genotypów w poszczególnych układach mnożone są przez siebie. Częstości genotypów obliczane są następująco:

– homozygoty:
$f = p * p + p * (1-p) * F_{ST}$, gdzie p – częstość allela

– heterozygoty:
$f = 2 * p_i * p_j$, gdzie p_i, p_j – częstość allela i, j

Drugim liczonym parametrem jest prawdopodobieństwo p(X|X). Jest to również iloczyn odpowiednich prawdopodobieństw w poszczególnych układach. Prawdopodobieństwa te liczymy następująco:

– homozygoty:
$$p(X|X) = [2*F_{ST}+(1-F_{ST})*p]*[3*F_{ST}+(1-F_{ST})*p] / / [(1+F_{ST})*(1+2*F_{ST})]$$

– heterozygoty:
$$p(X|X) = 2*[F_{ST}+(1-F_{ST})*p_i]*[F_{ST}+(1-F_{ST})*p_j] / / [(1+F_{ST})*(1+2*F_{ST})]$$

EVIDENTIAL SAMPLES ANALYSIS

The DNAStat 2.1 program allows for an analysis of biological evidence by calculating the unconditional f and conditional p(X|X) profile frequency, with the possibility of taking into account the coancestry coefficient FST, as well as setting the minimum allele frequency - CP. FST – the coancestry coefficient – it is defined for the whole population and describes the chance that two randomly chosen alleles of two randomly chosen individuals are identical by descent. This coefficient expresses a certain undetermined number of unknown common ancestors of the past generations. In standard populations, FST equals 0.01, while in small isolated populations or hardly assimilating ones, it can equal up to 0.03 [1,2]. The CP ceiling principle is the minimum allele frequency, which is used to prevent overestimation of the DNA profile frequencies caused by low allele frequencies. The 1st NRC report launched in 1992 recommended the use of CP (CP=0.1 for interim ceiling principle or CP=0.05 for ceiling principle) [11]. Nowadays, the CP usage is not recommended (CP=0) [12].

The profile frequency f is counted first for every marker and next, genotypes frequencies of particular markers are multiplied by themselves. Genotypes frequencies are counted in the following way:

– homozygote:
$f = p*p + p * (1-p) * F_{ST}$, p – allele frequency

– heterozygote:
$f = 2 * p_i * p_j$, p_i, p_j – i, j alleles frequencies

The second parameter is the conditional p(X|X). This is the product of appropriate probabilities for particular markers, counted in a following way:

– homozygote:
$$p(X|X) = [2*F_{ST}+(1-F_{ST})*p]*[3*F_{ST}+(1-F_{ST})*p] / / [(1+F_{ST})*(1+2*F_{ST})]$$

– heterozygote:

$$p(X|X) = 2*[F_{ST}+(1-F_{ST})*p_i]*[F_{ST}+(1-F_{ST})*p_j] / / [(1+F_{ST})*(1+2*F_{ST})]$$

Dla obu liczonych parametrów, tj. częstości i prawdopodobieństwa obliczenia prowadzimy albo dla faktycznych częstości alleli wynikających z danych w bazie populacyjnej, albo – gdy zadany próg CP jest różny od 0 – jeżeli częstość któregoś z alleli jest niższa od zadanego progu, to stosujemy zadany próg.

Częstość profilu f stosowana jest we wnioskowaniu wówczas, gdy znane jest pochodzenie osoby, do której należy analizowany ślad i istnieją bazy populacyjne dla osób o tym pochodzeniu. Np. podejrzewamy, że ślad należy do osoby z populacji polskiej i posiadamy bazy populacyjne dla takiej populacji.

Natomiast prawdopodobieństwo p(X|X) jest to prawdopodobieństwo, że losowo wybrana osoba inna niż osoba, od której pochodzi badany ślad, ma taki sam genotyp jak ten ślad. Stosowane jest, kiedy podejrzewamy, że osoba, do której należy ślad należy do pewnej subpopulacji, co do której nie istnieją bazy populacyjne, natomiast są odpowiednie bazy dla pełnej populacji. Np. podejrzewamy, że ślad należy do osoby z pewnego miasta, a nie są dostępne bazy populacyjne dla tego miasta, lecz tylko dla całego kraju.

ANALIZA OJCOSTWA

Program DNAStat podczas analizy ojcostwa oblicza szansę ojcostwa/macierzyństwa PI/MI (ang. paternity/maternity index) i prawdopodobieństwo ojcostwa W (niem. Wahrscheinlichkeit) w układzie pełnej trójki, w układzie mężczyzna-dziecko (bez matki) i w układzie kobieta-dziecko (bez mężczyzny), przy możliwości uwzględnienia częstości alleli zerowych null i prawdopodobieństwa a priori $p_{apriori}$. Współcześnie zaleca się stosowanie do obliczeń null=0 oraz $p_{apriori}=0.5$.

Obliczenia szansy ojcostwa PI prowadzone są według klasycznych zasad zaproponowanych przez Essen-Möllera [7] i podanych później wielokrotnie w piśmiennictwie, ostatnio np. przez Brennera [9] z uwzględnieniem częstości alleli zerowych. Przypadki mutacji traktowane są także według zasad zaproponowanych przez Brennera [8].

W przypadku niezgodności pomiędzy dzieckiem i pozwanym w postaci przeciwstawnych homozygot obliczenia są wykonywane w dwóch wariantach, w zależności od zadanej wcześniej wartości null. Jeżeli null>0, to wówczas stosowany jest wzór podany przez Brennera, a jeżeli null=0, to wówczas przypadek traktowany jest jako mutacja. Znajdowana jest wówczas

For both parameters, i.e. the frequency and probability, calculations are computed either for the actual allele frequencies resulting from the population database or CP values - when CP is different from 0 and the frequency of one of the alleles is lower than CP.

The profile frequency f is used for statistical calculations when the ethnic origin of the investigated individual is known and when there are population databases of people of the same ethnicity. For instance, to check if the evidence originates from the Polish population, the database of Poland has to be searched.

Conditional p(X|X) is the probability that randomly chosen individual, different from the individual from whom the investigated sample originates, shares the same genotype (with the sample). It is used when there is a possibility that the individual belongs to a certain subpopulation, which is not taken into account in any population database, however, there is a database for the population in a broader scale. For instance, this happens in a case when evidence comes from a resident of a certain town and there is no population database of that town, but there is a base of the whole country.

PATERNITY TESTING

DNAStat analyzes paternity cases by calculating the paternity / maternity index PI / MI and probability of paternity W for full, motherless and fatherless cases, with the possibility of taking into account the silent allele frequency and prior probability. It is recommended to take 0 for silent allele's values and 0.5 for prior probability values.

Paternity index calculation is made according to classical rules proposed by Essen-Möller [7] and repeatedly cited, recently by Brenner [9], taking into account silent alleles frequencies. Mutation events are treated according to the rules proposed by Brenner [8].

In case of an inconsistency manifested in the opposite homozygotes between the child and the alleged father, calculations are carried out in two variants, depending on setting the null allele value. If null>0, then the Brenner's formula is used, and if null=0, the case is treated as a mutation event. The minimal number of repetitive units between the child's and the alleged father's alleles is indicated and the Brenner's formula for mutation is used. In cases of other inconsistencies, the Brenner's formula for mutation is used.

najmniejsza ilość jednostek repetytywnych pomiędzy allelami dziecka i pozwanego i dla tej ilości jednostek stosowany jest wzór Brennera dla mutacji. W przypadkach pozostałych niezgodności stosowany jest każdorazowo wzór Brennera dla mutacji.

Obliczenia szansy macierzyństwa MI prowadzone są wg tych samych zasad, co obliczenia szansy ojcostwa PI.

Po obliczeniu w powyższy sposób szansy ojcostwa PI (lub macierzyństwa MI) dla każdego układu obliczana jest wartość całkowita jako iloczyn wartości cząstkowych. Z wartości całkowitej szansy ojcostwa PI (macierzyństwa MI) wyliczana jest następnie wartość prawdopodobieństwa ojcostwa W według wzoru:

$$W = 1 / [\, 1 + (((1 - p_{apriori}) / p_{apriori}) * (1 / PI)) \,]$$

Maternity Index is calculated according to the same rules as Paternity Index.

After paternity / maternity index is counted for every marker, it is multiplied, which results in a total PI. Next, the probability of paternity / maternity W is calculated according to the following formula:

$$W = 1 / [\, 1 + (((1 - p_{apriori}) / p_{apriori}) * (1 / PI)) \,]$$

PIŚMIENNICTWO / REFERENCES

1. Ayres K. L.: Measuring genetic correlations within and between loci with implications for disequilibrium mapping and forensic identification. Ph. D. Thesis, The University of Reading, Reading 1998, 181-204.
2. Ayres K. L.: Relatedness testing in subdivided populations. Forensic Sci. Int. 2000, 114, 107-115.
3. Bąbol-Pokora K., Prośniak A., Jacewicz R., Berent J.: Pentapleks SNP – rozkład częstości alleli w populacji centralnej Polski. Arch. Med. Sąd. i Krym. 2006, 56(4), 228-231.
4. Berent J.: DNAStat wersja 1.0 – program do obsługi bazy danych profili genetycznych oraz do obliczeń biostatystycznych. Arch. Med. Sąd. i Krym. 2006, 56(1), 15-18.
5. Berent J.: DNAStat wersja 1.0 – program do obsługi bazy danych profili genetycznych oraz do obliczeń biostatystycznych. Program komputerowy. Uniwersytet Medyczny w Łodzi, Łódź 2005.
6. Berent J.: DNAStat wersja 1.2 – program do obsługi bazy danych profili genetycznych oraz do obliczeń biostatystycznych. Arch. Med. Sąd. i Krym. 2007, 57(3), 322-325.
7. Essen-Möller E.: Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. Mitteilungen der Anthropologischen Gesellschaft in Wien 1938, 68, 2-53.
8. http://dna-view.com/mudisc.htm.
9. http://dna-view.com/patform.htm.
10. Jacewicz R., Berent J., Prośniak A., Gałecki P., Florkowski A., Szram S.: Population genetics of the Identifiler system in Poland. International Congress Series 2004, 1261, 229-232.
11. National Research Council Report. DNA Technology in Forensic Science. National Academy Press, Washington, D.C. 1992, 91-92.
12. National Research Council Report II. The Evaluation of Forensic DNA Evidence. National Academy Press, Washington, D.C. 1996, 96-97.
13. 2001 Paternity Testing Workshop of the English Speaking Working Group of the International Society for Forensic Genetics.

Adres do korespondencji / Address for correspondence:
Prof. Jarosław Berent
Katedra i Zakład Medycyny Sądowej
Uniwersytetu Medycznego w Łodzi
ul. Sędziowska 18a, 91-304 Łódź, Poland
J.Berent@eranet.pl