ARCH. MED. SĄD. KRYM., 2006, LVI, 15-18 PRACE ORYGINALNE

Jarosław Berent

DNAStat wersja 1.0 – program do obsługi bazy danych profili genetycznych oraz do obliczeń biostatystycznych[1]

DNAStat, version 1.0 – a software package for processing a genetic profile database and for biostatistical calculations

Z Zakładu Orzecznictwa Sądowo-Lekarskiego i Ubezpieczeniowego
Katedry Medycyny Sądowej Uniwersytetu Medycznego w Łodzi
Kierownik: dr hab. n. med. Jarosław Berent

Rozpowszechnienie się badań DNA wykorzystywanych dla potrzeb wymiaru sprawiedliwości spowodowało konieczność opracowania odpowiednich programów komputerowych. Programy takie muszą rozwiązywać dwa istotne problemy, tj. problem szeroko pojętej obsługi i archiwizacji danych oraz problem obliczeń biostatystycznych. W niniejszej pracy omówiony jest program DNAStat wersja 1.0. Program ten umożliwia: 1) tworzenie i obsługę własnej bazy danych; 2) analizę śladów biologicznych przez obliczenie częstości profilu f oraz prawdopodobieństwo p(X|X), przy możliwości uwzględnienia współczynnika pochodzenia oraz zadania dolnego progu częstości alleli; 3) analizę ojcostwa przez obliczenie szansy ojcostwa PI i prawdopodobieństwa ojcostwa W w układzie pełnej trójki i w układzie pozwany-dziecko (bez matki), przy możliwości uwzględnienia częstości alleli zerowych i prawdopodobieństwa a priori.

The application of DNA studies to the administration of justice has led to the necessity of developing appropriate computer programs. Such programs must address two critical problems, i.e. the broadly understood data processing and archivization, and biostatistical calculations. This paper discusses DNAStat 1.0, a program that enables its user to 1) create and process an individual database; 2) analyze biological evidence by calculating the unconditional f and conditional p(X|X) profile frequency, with the possibility of taking into account the inbreeding (coancestry) coefficient, as well as setting the minimum allele frequency; 3) analyze paternity cases by calculating the paternity index PI and probability of paternity W for full and motherless trios, with the possibility of taking into account the silent allele frequency and prior probability.

Słowa kluczowe: biostatystyka, badania ojcostwa, badania dowodów rzeczowych, bazy danych
Keywords: biostatistics, 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 rozwiązywać dwa istotne problemy, tj. problem szeroko pojętej obsługi i archiwizacji danych oraz problem obliczeń biostatystycznych. W niniejszej pracy omówiony jest program DNAStat wersja 1.0 [1]. Program został opracowany przez dr. hab. n. med. Jarosława Berenta, kierownika Zakładu Orzecznictwa Sądowo-Lekarskiego i Ubezpieczeniowego Katedry Medycyny Sądowej Uniwersytetu Medycznego w Łodzi przy wykorzystaniu obsługi informatycznej firmy Laser Systemy Informatyczne S.A. w Łodzi.

[1] Praca powstała w ramach grantu Uniwersytetu Medycznego w Łodzi nr 502-11-785(35).

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.

Genotypy badanych osób lub śladów mogą być 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® 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 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 program oblicza szansę ojcostwa (ang. paternity index) i prawdopodobieństwo ojcostwa W (niem. Wahrscheinlichkeit) w układzie pełnej trójki i w układzie pozwany-dziecko (bez matki). Wyniki obliczeń mogą być drukowane.

ANALIZA ŚLADÓW BIOLOGICZNYCH

Program DNAStat w analizie śladów biologicznych oblicza częstość profilu f (ang. unconditional profile frequency) oraz prawdopodobieństwo p(X|X) (ang. conditional profile frequency), 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 [2, 3]. 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) [4]. Współcześnie nie zaleca się stosowania żadnych takich progów (CP=0) [5] lub ewentualnie zastosowanie progu opartego o wielkość populacji i wynoszącego CP=5/n, gdzie n – liczba alleli w populacji [5, 6]. Ten ostatni wzór można zinterpretować w taki sposób, że dla zapewnienia wiarygodnych wyników obliczeń statystycznych w analizowanej populacji musi być co najmniej 5 alleli danego rodzaju, a jeżeli jest mniej to ich liczbę podnosi się do 5.

Częstość profilu f jest liczona najpierw dla każdego układu i dalej częstości genotypów w poszczególnych układach mnożone są przez siebie (ang. product rule). 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:

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

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 takich osób.

Natomiast prawdopodobieństwo p(X|X) jest to prawdopodobieństwo, że losowo wybrana osoba inna niż ta, 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 PI (ang. paternity index) i prawdopodobieństwo ojcostwa W (niem. Wahrscheinlichkeit) w układzie pełnej trójki i w układzie pozwany-dziecko (bez matki), 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 [8] z uwzględnieniem częstości alleli zerowych. Przypadki mutacji traktowane są także według zasad zaproponowanych przez Brennera [9].

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 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.

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

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

INSTALACJA PROGRAMU

Dysk instalacyjny programu DNAStat zawiera jeden plik o nazwie DNAStat – Install.exe. Po jego uruchomieniu cała instalacja następuje automatycznie i trwa około jednej minuty. Program zostaje zainstalowany do katalogu: C:\Program Files\Laser\DNAStat\, a na pulpicie umieszczona zostaje ikona o nazwie DNAStat 1.0. Program można odinstalować przez aplet Dodaj lub usuń programy w Panelu sterowania.

W katalogu C:\Program Files\Laser\DNAStat\, w folderze o nazwie Database 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 [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 [11], 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\Laser\DNAStat\Database\.

Podczas instalacji w katalogu C:\Program Files\Laser\DNAStat\, w folderze o nazwie Przykładowe dane zostaje umieszczonych sześć plików z przykładowymi danymi. Są to dwa pliki programu Microsoft® Excel: Import 1.xls i Import 2.xls. Pliki programu Microsoft® 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 folderu Przykładowe dane 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:

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

Podczas instalacji w katalogu C:\Program Files\Laser\DNAStat\, w folderze o nazwie Instrukcja zostaje umieszczony plik DNAStat 1.0.pdf, który zawiera instrukcję instalacji i użytkowania programu.

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.

PIŚMIENNICTWO

1. 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.

2. 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, pp. 181-204.

3. Ayres K. L.: Relatedness testing in subdivided populations. Forensic Sci. Int. 2000, 114, 107-115.

4. National Research Council Report. DNA Technology in Forensic Science. National Academy Press, Washington, D.C. 1992, pp. 91-92.

5. National Research Council Report II. The Evaluation of Forensic DNA Evidence. National Academy Press, Washington, D.C. 1996, pp. 96-97.

6. Budowle B., Giusti A. M., Waye J. S., Baechtel F. S., Fourney R. M., Adams D. E., Presley L. A., Deadman H. A., Monson K. L.: Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. Am. J. Hum. Genet. 1991, 48, 841-855.

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/patform.htm.

9. http://dna-view.com/mudisc.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. 2001 Paternity Testing Workshop of the English Speaking Working Group of the International Society for Forensic Genetics.

Adres do korespondencji:
Dr hab. n. med. Jarosław Berent
Kierownik Zakładu Orzecznictwa
Sądowo-Lekarskiego i Ubezpieczeniowego
Katedry Medycyny Sądowej
Uniwersytetu Medycznego w Łodzi
ul. Sędziowska 18a
91-304 Łódź
J.Berent@eranet.pl