Problems of Forensic Sciences 2006, LXVIII, 362–367 ISSN 1230-7483

EFFECT OF RARE ALLELES ON DNA PROFILE FREQUENCY IN SGM PLUS STUDIES*

Ireneusz SOŁTYSZEWSKI[1], Magdalena SPÓLNICKA[2], Magdalena KONARZEWSKA[2], Jarosław BERENT[3]

[1] *Department of Criminalistics and Forensic Medicine, University of Varmia and Mazury, Olsztyn*
[2] *Department of Biology, Central Police Forensic Laboratory, Warsaw*
[3] *Department of Forensic Medicine, Medical University of Łódz, Łódź*

Abstract

To calculate the DNA-profile frequency, it is essential to know the allele distribution frequency in a given population. The study and determination of the allele frequencies of ten STR loci (SGM plus) in the Polish population served as a basis for assessing the effect of rare alleles on the statistical interpretation of DNA analysis results and their relation to the National Research Council Report recommendations. The statistical analysis we carried out showed that low-frequency alleles appear rarely in the multiplex STR profile, which means that their effect on the final frequency profile is not great. Frequency profiles calculated for the SGM Plus system with correction for rare alleles, and without correction, are never greater than ca. 1.9 10^{-11}, which shows the great evidence value that the system provides in crime investigation.

Key words

Profile frequency; Rare alleles; STR; Population genetics; AmpF*l*STR SGM Plus.

Received 6 November 2006; accepted 20 December 2006

1. Introduction

Studies using STR markers are currently the most popular method applied to determining blood-relations and the origin of organic/biological traces. Multiplex systems make it possible to determine gender on the basis of amelogenin (AMG) locus analysis, as well as genotyping of 7 to 15 STR loci, depending on the system used. A characteristic feature of these systems is their great sensitivity, the amount of information provided, reliability, repeatability and the fact that analysis is reduced to only one study process [2]. These advantages have also led to the utilization of STR loci in national DNA bases [e.g. 6]. In connection with the above, the examination and determination of STR locus allele frequencies in the population is of great significance, as is demonstration that these data, calculated on the basis of sample population studies, can be a tool serving to determine probability values describing the chance of a DNA profile occurring randomly [1, 3]. Owing to the possibility of overestimating this value, in the 1st Report of the National Research Council in 1992, use of the so-called interim ceiling principle, or ceiling principle was recommended. In the first case (interim ceiling principle) use is

* Population studies were carried out by staff of the Department of Biology, Central Criminalistics Laboratory, Headquarters of the Polish Police and Department of Forensic Medicine, Medical University in Warsaw as part of project 128-270/C-T00/99. The project was financed by the Polish Government Research Committee. The head of the project was Halina Dąbrowska M.Sc. and the scientific co-ordinator was Prof. Andrzej Płócienniczak.

recommended of an allele frequency of at least 0.1, and in the ceiling principle, 0.05. In the 2nd NRC Report published in 1996, the use of any such thresholds was no longer recommended [4, 5].

In connection with the above, the authors have attempted to assess the real influence of rare alleles on the evidential value of expert reports in relation to a big population sample originating from all over Poland. The aim of the study was to investigate, for a very large data base of genetic profiles, how often low-frequency alleles (i.e. lower than 0.1, 0.05, and 0.01) appear in multiplex profiles, and how these alleles affect the final frequency of the genetic profile. Furthermore, the influence of increasing rare-allele frequency values to the thresholds given (0.1, 0.05, and 0.01) on the final frequency of the whole profile was examined.

2. Materials and methods

In this study we used a database of 2233 genetic profiles obtained using the SGM Plus system. Biological material in the form of oral smears was taken from non-related, anonymous, Caucasian volunteers. In order to obtain a sample representative of the Polish population, donors were sought from the entire national territory.

All profiles gathered in the database were subjected to tests in order to find out how often alleles of a frequency less than 0.10, 0.05 and 0.01 appear in these profiles, and the results were assembled in a table (Table I). In the next step, the frequencies of the individual DNA profiles were calculated. First they were calculated without any corrections, using the formula $f = p \quad p$ for homozygotes, $f = 2 \quad p_i \quad p_j$ for heterozygotes, and then with such a correction that, in the case of the appearance of a rare allele, its frequency was increased to a required threshold (i.e. up to 0.10, 0.05 or 0.01). Finally, the quotient of one frequency (with correction) divided by the other (without correction) was calculated; in other words, we checked by how many times the introduction of the correction increases the profile's frequency, or by how many times the evidence-value of the result is decreased by the introduction of such a correction.

The calculations were carried out using an author-developed computer program written in Delphi, and the statistical analysis, using STATGRAPHICS® Plus 5.0 Professional Edition.

3. Results

In the first stage of the analysis, we tested how many alleles in the entire multiplex profile of 10 loci, i.e. 20 alleles, had frequencies below a given threshold. It was ascertained that at a threshold of 0.10, 4.46574 alleles on average had a frequency below that threshold ($SD = 1.81066$, minimum = 0, maximum = 11). For a threshold of 0.05, an average of 1.17278 alleles had a frequency below it ($SD = 1.17278$, minimum = 0, maximum = 6), and for a threshold of 0.01, only 0.241379 alleles on average had a lower frequency than this threshold ($SD = 0.485865$, minimum = 0, maximum = 3). Detailed calculation results are presented in Table I.

TABLE I. ANALYSIS OF THE NUMBER OF ALLELES IN A PROFILE HAVING A FREQUENCY LESS THAN 0.10, 0.05 AND 0.01 RESPECTIVELY

Number of alleles	Number of profiles which have alleles with frequency		
	$f < 0.10$	$f < 0.05$	$f < 0.01$
0	16	488	1749
1	84	734	433
2	214	601	47
3	358	291	4
4	478	87	–
5	468	26	–
6	326	6	–
7	181	–	–
8	72	–	–
9	26	–	–
10	9	–	–
11	1	–	–

Next, frequencies of all profiles in the database were calculated without any corrections, as well as with a correction increasing the frequency of rare alleles to the given threshold, i.e. to 0.10, 0.05 and 0.01 respectively. The results of the calculations are presented below on scatterplots. In Figure 1, frequencies of profiles are calculated without corrections in relation to frequencies with correction to 0.10, in Figure 2 to 0.05, and in Figure 3 to 0.01.

Analysis of the results revealed that profile frequencies lie between $2.2 \quad 10^{-20}$ and $1.9 \quad 10^{-11}$. Specific values are shown in Table II.

TABLE II. MINIMAL AND MAXIMAL VALUES OF GENETIC PROFILE FREQUENCIES

Frequency correction	Min frequency	Max frequency
No correction	$2.2\ 10^{-20}$	$1.9\ 10^{-11}$
0.10	$2.2\ 10^{-16}$	$1.9\ 10^{-11}$
0.05	$1.0\ 10^{-17}$	$1.9\ 10^{-11}$
0.01	$1.5\ 10^{-19}$	$1.9\ 10^{-11}$

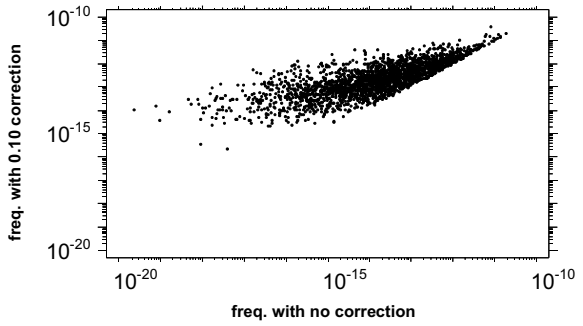

Fig. 1. Comparison of genetic profile frequencies for SGM Plus. Horizontal axis: profile-frequency values calculated without correction; vertical axis: profile frequencies calculated with replacement of frequencies of alleles rarer than 0.10 by 0.10.

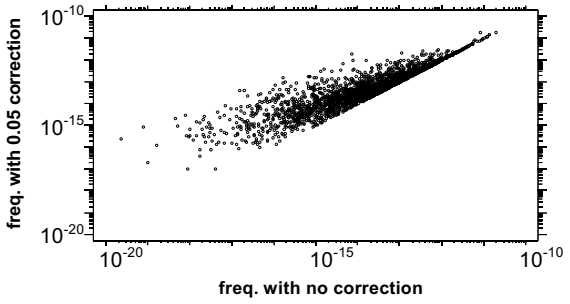

Fig. 2. Comparison of genetic profile frequencies for SGM Plus. Horizontal axis: profile-frequency values calculated without correction; vertical axis: profile frequencies calculated with replacement of frequencies of alleles rarer than 0.05 by 0.05.

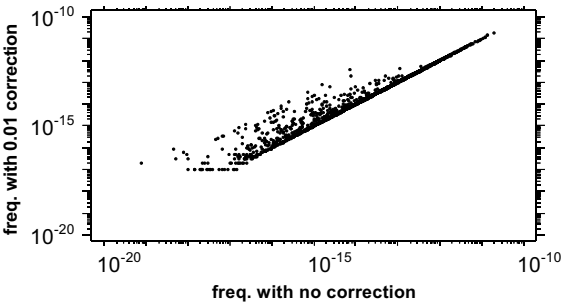

Fig. 3. Comparison of genetic profile frequencies for SGM Plus. Horizontal axis: profile-frequency values calculated without correction; vertical axis: profile frequencies calculated with replacement of frequencies of alleles rarer than 0.01 by 0.01.

Additionally, we calculated by how many times the introduction of specific corrections to allele frequencies increases the final value of the profile frequency; that is, we tested how introduction of particular corrections lowers the evidence value of an expert opinion. The results of these calculations are presented in Table III.

4. Discussion

The performed studies showed that alleles of low frequency appear in the multiplex profile rather rarely: even at a very high threshold of 0.10, in only somewhat over 4 cases out of 20, on average. In other words, in nearly 16 cases out of 20, allele frequencies are greater than 0.10 (this very high threshold was taken from the interim ceiling principle). But, if we take the threshold $f = 0.05$, then such alleles appear on average only in slightly more than 1 allele in 20 (this threshold was from the ceiling principle). However, with a threshold of $f = 0.01$, an allele of low frequency appears only in an average of 0.25 alleles out of a possible 20.

Thus, rare alleles presented a problem in earlier times, when few loci were studied. Today by contrast, in multiplex studies, they appear in the total profile so rarely that

TABLE III. ANALYSIS OF QUOTIENTS OF PROFILE FREQUENCY CALCULATED BY APPLICATION OF SPECIFIC CORRECTIONS TO ALLELE FREQUENCY DIVIDED BY PROFILE FREQUENCY CALCULATED WITHOUT ANY CORRECTION

Frequency threshold	Min. quotient	Max. quotient	Average quotient	Number of samples where quotient					
				1–10	10–100	100–1000	1000–10,000	10,000–100,000	> 100,000
0.10	1	481 211	766.61	1056	776	275	101	23	2
0.05	1	10 702	36.18	1813	312	98	8	2	–
0.01	1	258	2.18	2175	56	2	–	–	–

their influence on the final result is slight (since the vast majority of alleles are those of great frequency).

Profile frequencies calculated without any corrections occur in the range from ca. 10^{-20} to ca. 10^{-11}, and with correction for $f < 0.10$, from ca. 10^{-16} to ca. 10^{-11}. This means that the introduction of such a correction limits the lowest profile frequencies, in certain cases even by ca. 4 decimal places. Correction, of course, only limits the lowest profile frequencies (thus the graph shows a "cloud" dispersing to the left), while higher profile frequencies are not limited. The highest profile frequencies are ca. 10^{-11}, that is, these genotypes appear in the study population with a frequency not higher than ca. 10^{-11}. Thus, this profile occurs most often once in 10^{11} persons, or once in one-hundred billion persons. This means that the evidential value of examinations based on SGM Plus is enormous and fully sufficient for criminal investigations.

The issue of "limiting" allele frequency by raising the actual frequency, before calculating the profile frequency, to some established value (as proposed in the early stages of DNA research), no longer has any great practical significance. Profiles of very low frequencies occur relatively rarely, as seen in the scatterplot of Figure 1: on the left there are fewer points than on the right, i.e. the graph "densifies" from left to right.

Similarly, an analysis of the scatterplots shown in Figure 2 and 3 leads to the same conclusions. These two graphs are very similar to Figure 1, only "denser", which means that, the smaller the correction, the more rarely it must be applied, and the smaller is its effect on the final result. This is best seen on the last scatterplot, Figure 3 (where $f < 0.01$). Most of the points lie on the line dividing it in half, and only a few are distant from it.

Analysis of the specific frequency quotients shows that the quotient, of course, reaches its greatest value where the correction for $f < 0.10$ was applied. Then, the maximum quotient is as high as 481, 211. When the corrections are smaller, the quotients also diminish to a maximum of 10, 702 and 258, respectively. But such great values are sporadic, since average values are, respectively, 766.61, 36.18 and 2.18. This means that calculating the profile frequency with the correction $f < 0.10$ causes on average a 766.61 times increase in profile frequency, while corrections $f < 0.05$ and $f < 0.01$ cause smaller changes: the average profile frequency increase is then only 36.18 and 2.18 times respectively.

Summarising, years ago rare alleles could pose a problem, since only few loci were studied. Then, the appearance of such an allele as one among only a few others could significantly affect the final result, markedly lowering the final value of the profile frequency. Today, in multiplex studies many loci are examined and rare alleles appear in the total profile so infrequently that their influence on the final result is slight, since the decisive majority of all alleles are those of high frequency.

5. Conclusions

Alleles with low frequencies appear in the multiplex STR profile rarely, which causes their effect on the final profile frequency result to be small. Profile frequencies calculated for the SGM Plus system, with correction for rare allele frequencies as well as without correction, are never greater than ca. $1.9 \quad 10^{-11}$, which indicates the enormous evidence value provided by this system in criminal investigations.

References

1. Buckleton J. S., Triggs C .M., Walsh S. I., Forensic DNA evidence interpretation, CRC Press 2005.
2. Budowle B., Niezgoda S. B., Charkabort R., CODIS STR loci data from 41 sample populations, *Journal of Forensic Science* 2001, 46, 453–489.
3. Budowle B., Monson K. L., Chakraborty R., Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci, *International Journal of Legal Medicine* 1996, 108, 173–176.
4. National Research Council Report: DNA, technology in forensic science, National Academy Press, Washington 1992.
5. National Research Council Report II: The evaluation of forensic DNA evidence, National Academy Press, Washington 1996.
6. Parson W., Steinlechner M., Efficient DNA database laboratory strategy for high throughput STR typing of reference samples, *Forensic Science International* 2001, 22, 1–6.

Corresponding author
Jarosław Berent
Zakład Medycyny Sądowej
Uniwersytetu Medycznego w Łodzi
ul. Sędziowska 18 a
PL 91-304 Łódź
e-mail: J.Berent@eranet.pl

WPŁYW RZADKICH ALLELI NA CZĘSTOŚĆ PROFILU DNA PRZY BADANIU ZESTAWEM SGM PLUS*

1. Wstęp

Badania z użyciem markerów typu STR są aktualnie najpopularniejszą metodą stosowaną do ustalenia pokrewieństwa i w badaniach śladów biologicznych. Systemy multipleksowe umożliwiają identyfikację płci na podstawie analizy lokus amelogeniny (AMG) i genotypowanie, w zależności od zestawu, od 7 do 15 *loci* typu STR. Charakterystyczną cechą tych systemów jest ich wysoka czułość, duża informatywność, wiarygodność i powtarzalność wyników oraz ograniczenie czasu analizy do jednego procesu badawczego [2]. Powyższe zalety spowodowały wykorzystanie *loci* STR również w krajowych bazach DNA [np. 6]. W związku z powyższym istotne znaczenie ma zbadanie i ustalenie częstości występowania alleli *loci* STR w populacji oraz wykazanie, że dane te, wyliczone na podstawie przebadanej próby populacyjnej, mogą stanowić narzędzie służące do określenia wartości prawdopodobieństwa opisującego losową szansę wystąpienia profilu DNA [1, 3]. Z uwagi na możliwość przeszacowania tej wartości, w pierwszym raporcie opracowanym przez National Research Council w 1992 r. zalecano stosowanie tzw. *interim ceiling principle* lub *ceiling principle*. W pierwszym przypadku zalecano stosowanie częstości alleli o wartości co najmniej 0,1, a w drugim 0,05. W drugim raporcie NRC opublikowanym w 1996 r. nie zaleca się już stosowania takich progów [4, 5]. W związku z powyższym autorzy podjęli próbę oszacowania faktycznego wpływu rzadkich alleli na wartość dowodową ekspertyzy w odniesieniu do dużej próbki populacyjnej pochodzącej z terenu całej Polski.

Celem pracy była analiza częstości pojawiania się alleli uznawanych za rzadkie (o częstościach niższych niż 0,1, 0,05 i 0,01) w profilach genetycznych wchodzących w skład dużej bazy danych oraz ocena wpływu, jaki tego typu allele mają na częstość całego profilu genetycznego. Badano również wpływ, jaki ma podniesienie wartości częstości rzadkich alleli do zadanych wartości progowych (0,1, 0,05 i 0,01) na ostateczną częstość całego profilu.

* Badania populacyjne opisane w niniejszej pracy przeprowadzone zostały przez pracowników Zakładu Biologii Centralnego Laboratorium Kryminalistycznego Komendy Głównej Policji w Polsce oraz Zakładu Medycyny Sądowej Akademii Medycznej w Warszawie i wykonano je w ramach projektu 128-270/C-T00/99. Projekt sfinansowany został przez Komitet Badań Naukowych w Polsce. Kierownikiem projektu była mgr Halina Dąbrowska, a koordynatorem naukowym prof. Andrzej Płócienniczak.

2. Materiał i metody

W pracy wykorzystano bazę danych liczącą 2233 profile genetyczne, które oznaczono przy wykorzystaniu zestawu SGM Plus. Materiał biologiczny w postaci wymazów z jamy ustnej został pobrany na zasadzie dobrowolności od niespokrewnionych anonimowych dawców rasy kaukaskiej. W celu uzyskania reprezentatywnej dla polskiej populacji próbki, kompletowanie dawców odbywało się z uwzględnieniem całego obszaru Polski, a miejsce urodzenia przyjęto za kryterium przynależności terytorialnej.

Dla wszystkich profili z bazy sprawdzono, jak często pojawiają się w niej allele o częstości mniejszej niż 0,10, 0,05 i 0,01, a wyniki przedstawiono tabelarycznie (tabela I). Następnie obliczono częstości poszczególnych profili DNA. Najpierw obliczono je bez żadnych poprawek, stosując wzory dla homozygot $f = p \quad p$, dla heterozygot $f = 2 \quad p_i \quad p_j$, a następnie z taką poprawką, że w przypadku pojawienia się rzadkiego allela jego częstość zwiększano do zadanego progu (tj. kolejno 0,10, 0,5 i 0,1). Na koniec obliczono iloraz jednej częstości (z poprawką) przez drugą (bez poprawki), czyli, innymi słowy, sprawdzano, ile razy zwiększy częstość profilu wprowadzenie poprawki albo ile razy zmniejszy się wartość dowodowa wyniku przez wprowadzenie takiej poprawki.

Obliczenia przeprowadzono, stosując własnoręcznie napisane programy komputerowe w języku Delphi, zaś do analizy statystycznej pakiet statystyczny STATGRAPHICS® Plus 5.0 Professional Edition.

3. Wyniki

W pierwszym etapie analizy sprawdzono, ile alleli w całym multipleksowym profilu składającym się z 10 *loci*, czyli 20 alleli, ma częstości mniejsze od zadanego progu. Stwierdzono, że przy progu wynoszącym 0,10 (średnio 4,46574) allela ma częstość mniejszą od tego progu (*SD* = 1,81066, minimum = 0, maximum = 11). Przy progu wynoszącym 0,05 średnio 1,17278 allela ma częstość mniejszą od niego (*SD* = 1,17278, minimum = 0, maximum = 6), a dla progu wynoszącego 0,01 już tylko średnio 0,241379 allela ma częstość mniejszą od tego progu (*SD* = 0,485865, minimum = 0, maximum = 3). Szczegółowe wyniki Następnie obliczono częstości wszystkich profili z bazy danych bez jakichkolwiek poprawek oraz z poprawką zwiększającą częstość rzadkich alleli do zadanego progu, tj. odpowiednio do 0,10, 0,05 i 0,01. Wyniki obliczeń przedstawiono poniżej na wykresach

rozproszenia. Na rycinie 1 porównano wartości częstości profili obliczone bez poprawek w stosunku do częstości z poprawką dla 0,10, na rycinie 2 dla 0,05 a na rycinie 3 dla 0,01.

Analizując wyniki stwierdzono, że częstości profili zawierają się w zakresach pomiędzy $2{,}2 \quad 10^{-20}$ a $1{,}9 \quad 10^{-11}$. Szczegółowe wartości podano w tabeli II.

Dodatkowo obliczono jeszcze, ile razy wprowadzenie poszczególnych poprawek do częstości allelicznych zwiększa końcową wartość częstości profilu, czyli innymi słowy, sprawdzono, jak wprowadzenie poszczególnych poprawek zmniejsza wartość dowodową ekspertyzy. Wyniki tej części obliczeń przedstawiono w tabeli III.

4. Dyskusja

Przeprowadzone badania wykazały, że allele o niskiej częstości pojawiają się w multipleksowym profilu stosunkowo rzadko. Nawet przy bardzo wysokim progu równym 0,10 tylko średnio w nieco ponad 4 przypadkach na 20, czyli w blisko 16 przypadkach na 20, częstości alleliczne są większe niż 0,10 (ten bardzo wysoki próg wzięto z *interim ceiling principle*). Przy zastosowaniu progu $f = 0{,}05$ takie allele pojawiają się średnio tylko w ponad 1 przypadku na 20 (taki próg wynikał z *ceiling principle*). W przypadku progu $f = 0{,}01$ allele o niskiej częstości pojawiają się już tylko w średnio 0,25 allela na 20 możliwych.

Rzadkie allele mogły zatem stanowić problem w czasach, gdy badano mało *loci*. Natomiast obecnie przy badaniach multipleksowych pojawiają się one w całym profilu tak rzadko, że ich wpływ na końcowy wynik będzie niewielki (zdecydowana większość alleli to allele o dużej częstości).

Częstości profili obliczone bez żadnych poprawek zawierają się w zakresie od około 10^{-20} do około 10^{-11}, a z poprawką dla $f < 0.10$ od około 10^{-16} do około 10^{-11}. Oznacza to, że wprowadzenie takiej poprawki ogranicza najmniejsze częstości profili w niektórych przypadkach nawet o około 4 rzędy wielkości. Poprawka ogranicza oczywiście tylko najmniejsze częstości profili (stąd wykres tworzy „chmurę" rozpraszającą się w lewo), natomiast największe częstości profili nie są ograniczane. Najwyższe częstości profili wynoszą około 10^{-11}, czyli dane genotypy pojawiają się w badanej populacji z częstością nie wyższą niż około 10^{-11}. Taki profil występuje najczęściej raz na 10^{11} osób, czyli raz na sto miliardów. Wartość dowodowa badania zestawem SGM Plus jest więc miarodajna i wystarczająca do badań kryminalistycznych.

Kwestia „ograniczania" częstości alleli przez podwyższanie faktycznych częstości alleli przed obliczeniami częstości profilu do jakiejś założonej wartości (proponowana w początkowej fazie badań DNA) nie ma już większego praktycznego znaczenia. Profile o bardzo małych częstościach zdarzają się stosunkowo rzadko (co widać na rycinie 1 – po lewej stronie jest mniej punktów niż po prawej, wykres się „zagęszcza" przy przesuwaniu się od lewej do prawej). Również analiza rycin 2 i 3 prowadzi do takich samych wniosków. Oba te wykresy mają bardzo podobny charakter jak wykres przedstawiony na rycinie 1, tyle tylko, że są „bardziej zwarte", co oznacza, że im mniejsza poprawka, tym rzadziej trzeba ją stosować i tym samym mniejszy jest jej wpływ na końcowy wynik. Najlepiej to widać na rycinie 3 (dla $f < 0{,}01$). Większość punktów na nim leży na linii dzielącej go na połowę, a tylko nieliczne oddalają się od tej linii.

Analiza ilorazów poszczególnych częstości wskazuje, że iloraz ten największe wartości osiąga oczywiście tam, gdzie stosowano poprawkę dla $f < 0{,}10$. Wówczas maksymalny iloraz wynosi nawet 481 211. Gdy poprawki są mniejsze, to ilorazy też maleją do – odpowiednio – maksymalnie 10 702 i 258. Jednak tak duże wartości są sporadyczne, bowiem wartości średnie wynoszą odpowiednio 766,61, 36,18 i 2,18. Oznacza to, że prowadzenie obliczeń częstości profilu z poprawką $f < 0{,}10$ spowoduje średnio 766,61-krotny wzrost częstości profilu. Wprowadzenie zaś poprawek $f < 0{,}05$ i $f < 0{,}01$ spowoduje jeszcze mniejsze zmiany, bowiem średnio odpowiednio tylko 36,18-krotne i 2,18-krotne.

5. Podsumowanie

Rzadkie allele mogły stanowić problem przed laty, gdy badano mało *loci*. Wówczas pojawienie się takiego allela jako jednego spośród kilku innych mogło znacząco wpływać na wynik końcowy, wyraźnie obniżając końcową wartość częstości profilu. Natomiast obecnie, przy badaniach multipleksowych, badanych jest wiele *loci*, a rzadkie allele pojawiają się w całym profilu tak rzadko, że ich wpływ na końcowy wynik będzie niewielki (zdecydowana większość alleli jest allelami o dużej częstości).