Landeskriminalamt Mecklenburg-Vorpommern

POLIZEI Mecklenburg-Vorpommern

EFQM Recognised for excellence 3 star - 2016

COMPLETELY AUTOMATED INTERPRETATION OF REFERENCE SAMPLES

Volker Weirich, MD
State Office of Criminal Investigation Mecklenburg-Vorpommern
19067 Rampe, Germany

Contact
Volker.Weirich@LKA–MV.de

Introduction

Inspired by full-continuous models a completely automated interpretation workflow for reference samples was developed and introduced in our lab.
It is based on a quite simple model that takes into account DNA amount and degree of degradation as well as backward and forward stutters.
Current implementation can handle CE results of any number of replicates, even from different autosomal or Y-chromosomal STR kits.

Workflow

Basically, there are few major steps after generating .fsa or .hid files from a Genetic Analyzer. In our lab we use GeneMapper ID-X (GMIDX; Thermo Fisher Scientific) for allele calling. Hence, five steps are needed:

I. Import files into GeneMapper ID-X
II. Size and allele calling using Analysis Methods without filtering (➔ next figure: remarks to allele calling software)
III. Export of results as .csv or .txt files
IV. Automated continuous interpretation using Statistefix v.4.0 alpha
V. Automated generation of database records including QR codes for National DNA Database

- Labs using GMIDX should use v1.6, because there is a new "Export Table With Stutter" option, which allows data export for continuous interpretation while using Analysis Methods with usual filters for human interpretation.
  In earlier GMIDX versions, separate projects have to be created.
- OSIRIS as a free alternative allows automatization of step I. – III. too; for detailed information visit ncbi.nlm.nih.gov/osiris.

Steps I. and II. can easily be performed in the command line interface (CLI) of GMIDX.
Step III. is an additional line in CLI (➔ next figure: see GMIDX's AdminGuide for detailed information; customizable arguments are shown in *italics*; typically used in batch file).

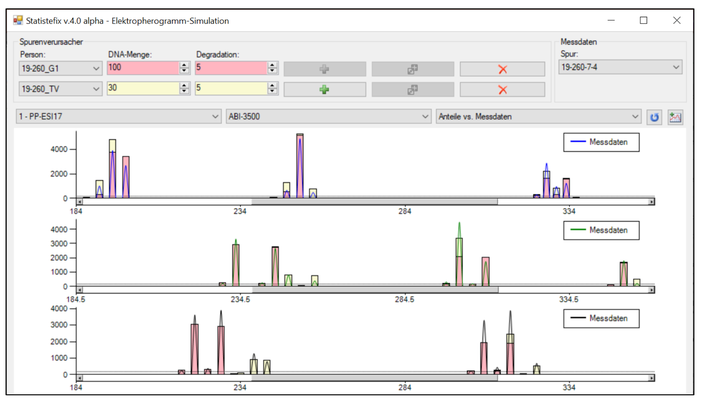
Steps I. and II.:

```
C:\> GeneMapper.exe -commandline -option h
     -username "UserName" -password "Password"
     -project "ProjectName" -folder "Folder"
     -analysismethod "AnalysisMethod"
     -sizestandard "SizeStandard"
```

Step III.:

```
C:\> GeneMapper.exe -commandline -option h
     -username "UserName" -password "Password"
     -project "ProjectName"
     -exportgenotypetable "FileName"
```

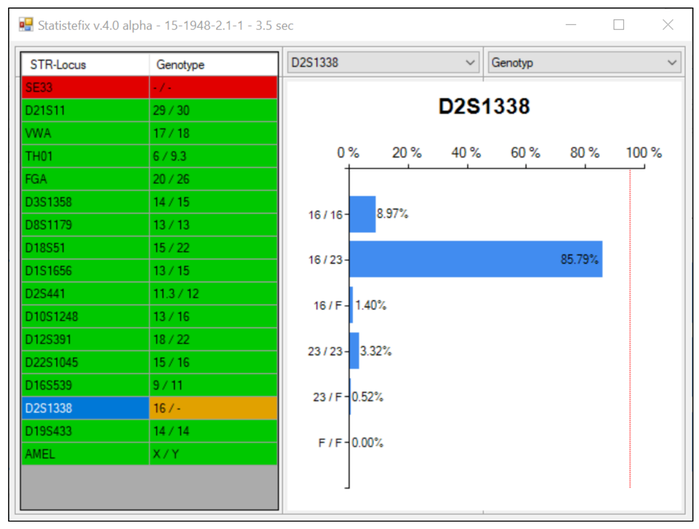
Step IV. represents core functionality:
In general, observed peak heights are compared to expected peak heights using a model with DNA amount, degree of degradation, and – of course – possible genotypes as parameters. (➔ next figure: easy-to-understand visualization of the main principle; screenshot-section only).



Genotype weights are calculated assuming a log-normal distribution for ratio of observed and expected peak heights (➔ next figure: observed probability distribution for observed / expected ratio of peak heights for PowerPlex ESI 17 Pro and ESX 17; Promega Corporation).

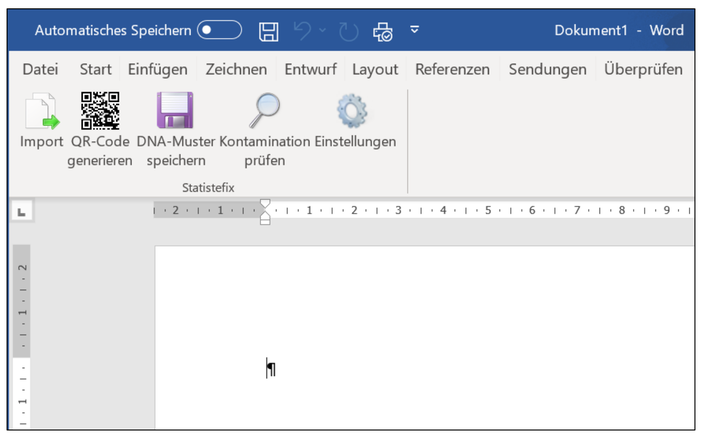
In order to avoid discordant results caused by primer binding site mutations, most forensic science services in Germany use more than one STR kit to amplify forensic samples to confirm results. Therefore, handling results from different STR kits was included from early stage on conceptually.

An easily-to-understand visual output is available as well as result tables and / or XML files for serial analyzes (➔ next figure).



Even if current tests use 95%, threshold level (dotted line) is customizable too.

Step V. is a user-friendly interface to Statistefix output files as an AddIn for Word 2013 and newer versions (Microsoft Corporation; ➔ next figure).



Proof of concept

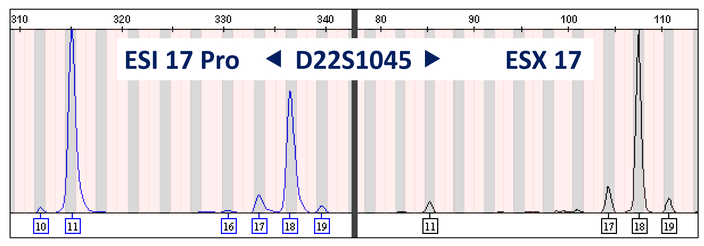
All data collected from reference samples between July 2018 and June 2019 were analyzed automatically. These automatically generated genotypes were compared to classical output from interpretation by experienced DNA experts. STR kits used for routine: ESI 17 Pro + ESX 17; Promega. Additionally, few samples were amplified with NGM-Detect; Thermo Fisher Scientific.

In a first series, the majority of discordant results were caused by rare off-ladder alleles. Therefore, bin sets provided by kit suppliers were filled with virtual bins using a VBA macro in Excel (Microsoft Corporation; ➔ next figure: ESI 17 Pro bins for D2S441, original [v2.1] and completed with virtual bins).

The same CE data were analyzed with modified bin sets for a second test.
Results from second test series are shown in ➔ next table.

	Amount	Share
Persons	555	
Samples	1,117	
.hid files	1,275	
Alleles (17 loci)	18,870	100%
Concordant	18,856	99.926%
Discordant	14	0.074%
Inconclusive	7 (6 samples)	0.037%
Erroneous	7 (4 samples)	0.037%

In a second series, most discordant results were caused by incorrect or failed size calling. After solving the incorrect size calling only four discordant alleles (0.021%) remained. Three discordant allele calls were caused by extreme locus imbalances, probably due to primer binding site mutations (➔ next figure: example). A fourth was caused by a huge bleed-through.



Still, some issues need to be solved. Therefore, drafts automatically generated by the current workflow require confirmation by DNA experts.
For prospective usage, it is strongly recommended to check size calling first.

Current tests and prospects

Given those useful results even in cases of poor-quality reference samples, currently first test series of extended usage for crime scene samples are performed on lab's backlog data, without EPG inspection by an expert.
Basically, the model is extended by one variable only: the number of contributors. Above-mentioned parameters (DNA amount, degradation, possible genotypes) are handled individually.
In one analysis 5.834 .hid files of 3.167 crime scene samples were analyzed using Statistefix v.4.0 alpha. After 30 hours for 901 samples complete profiles (17 loci) were deduced automatically, from both single stains and mixtures. These profiles were submitted to the National DNA Database and searched for matches; causing a hit ratio of roughly 30%. Even DNA-profiles (≥10 complete loci) generated from poor-quality stains provided valid hits in the National DNA Database.
Furthermore, a mass screening of over 1,000 men based on Y-chromosomal STRs was analyzed with this approach successfully, identifying a relative of the true offender, who was not included in the screening himself.
In conclusion, even if the above-mentioned probabilistic interpretation is still to improve it is a very useful tool in our lab already.

Acknowledgements

The author thanks Charles Brenner (➔ dna-view.com), Øyvind Bleka (➔ euroformix.com), and the STRmix crew (➔ strmix.com; alphabetical order, no ranking!) for valuable discussions.
Additionally, thanks to George Riley from NCBI for supporting me with OSIRIS.
Last but not least, thanks to members of our biostatistics project group from German police labs.