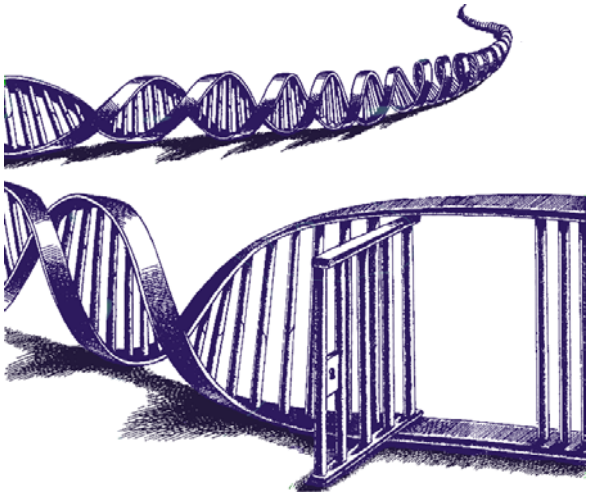


Charles H. Brenner, Ph.D.
consulting in forensic mathematics
6801 Thornhill Drive
Oakland, California, USA 94611
+1(510) 798-7139
E-MAIL: cbrenner@dna-view.com
WEB: www.dna-view.com

July 11, 2022

Mixture Solution™ overview

This is a very brief introduction to Mixture Solution,[1] an expert system for solving DNA mixtures. The complete documentation is the help file (see §3.2).

§1 presents some of the ideas, reasons, and discoveries behind the program. §2 is an overview tutorial of the key tasks of mixture analysis for casework and for automated validation. §3 is various information.

Table of contents

1 Design choices

1.1 Mixture Solution concept

Mixture Solution is an expert system for solving DNA mixtures. It consists of an expert exploratory component wrapped around a kernel that computes strength-of-evidence.

1.1.1 Fully continuous mixture model computation kernel

Mixture Solution's kernel is a program built on a fully continuous model[2] of DNA mixture data. The main point is to model stochastic variation – essential since dropout is an extremely important and ever-present complication in mixture analysis and, because dropout is the result of random sampling, it is a stochastic phenomenon. Degredation, dropin, and stutter fall out of the model naturally, with very little added complexity.

1.1.2 Expert system

The expert part systematically explores various mixture interpretations by repeatedly invoking the kernel (§1.1.1). It determines what hypotheses are most appropriate in terms both of contributors and number of contributors, separately for prosecution and for defense. I discovered (Brenner, How many contributors in a mixture?, 2019), and implemented in

[1] Mixture Solution is a module of the DNA·VIEW software package. A Kinship module is also part of DNA·VIEW but is not discussed here.
[2] The fully continuous category (also called "probabilistic genotyping") also includes STRmix, TrueAllele, Kongoh, and EuroForMix.

Mixture Solution the logic to answer mathematically how many contributors.[3] The expert system replaces human tedium and guesswork.

1.2 No MCMC

Markov Chain Monte Carlo is an excellent discovery, useful for many kinds of problems. For various reasons though, when I began the Mixture Solution project I did not want to follow the MCMC approach of TrueAllele and STRmix.

1.2.1 Black box

I'm distrustful, especially in a forensic situation, of using a mechanism whose behavior I don't fully understand and with pitfalls that I cannot fully control. For example, one can never be sure that an MCMC program will not overlook import aspects (technically referred to as "local maxima") of a mixture.

But initially I knew no alternative. Binary DNA mixture analysis methods treated the loci as independent bits of evidence each with an LR_i contribution to the overall LR. Even with the added complication of a continuous model rather than binary, independent per-locus computation isn't hard. But it's wrong, because although *genotypes* are independent between loci, *heights* are not. Imagine a two-locus mixture that looks to be two-person major-minor. If suspect S matches the major at one locus but the minor at the other, each locus by itself looks like strong evidence against him. But the two loci considered together provide much less evidence than the product of the two single-locus LRs – possibly even strong evidence in his favor. All researchers understood the problem but we had no answer except MCMC.

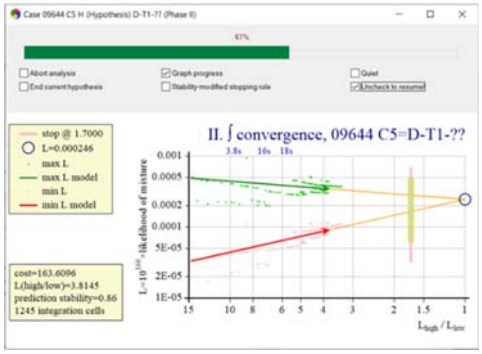
Jogging in the hills one afternoon, a new approach occurred to me. It is ***analytical*** rather than Monte Carlo, quite straightforward and correctly handles the dependence between loci – but had been universally overlooked. The next morning at first I doubted myself but fortunately I had written it down.

Limitation: This analytical method has trouble – results are less repeatable – with four or more unknowns when the mixture includes many alleles at a single locus (e.g. SE33).

1.2.2 Accuracy & precision

Another fuzzy MCMC issue is to find a stopping rule – a way to recognize when computing MCMC has gone on long enough and a sufficiently accurate result has been found. There are multiple published ideas, including Gelman-Rubin, but to my reading they signal when the computation is probably not going to change much more (is "reasonably" precise), but doesn't say anything about how much more.

Early implementation of the analytic method in Mixture Solution also did not have a good solution to the stopping problem. I was dissatisfied until I noticed an easy solution. The analytic idea can be implemented as numerical integration by rectangles. From there it is obvious to make two likelihood computations for a hypothesis simultaneously: one computation is a low-side likelihood estimate (red arrow) which gradually increases, the other a decreasing high-side estimate (green arrow), so the range of possible likelihoods is in between them and is squeezed down until a pre-specified accuracy target (blurry vertical bar) is achieved.



Progress bar and convergence graph

1.3 Direct, not deconvolution

Mixture Solution will, optionally, produce a deconvolution table as a tool for database searching. But deconvolution as a technique for computing likelihood ratios is simply a mistake. It prevents making full use of the data in various ways. For example when testing a non-contributor as a possible suspect, it often happens that a correct computation cannot exclude

[3] The concept that number of contributors can be computed (rather than guessed) was first discussed in (Brenner 1996) and later considered also in ISFG 2006 (Gill, Brenner, & Buckleton, 2006).

the suspect strongly (because a few alleles match and the correct computation includes the possibility of a quite minor contribution), but the deconvolution approach wrongly excludes strongly because of genotype combinations it overlooks.

Mixture Solution's alternative, a ***direct computation*** of the likelihood probability, is conceptually simpler, more flexible, and mathematically sound.

2 Mixture analysis

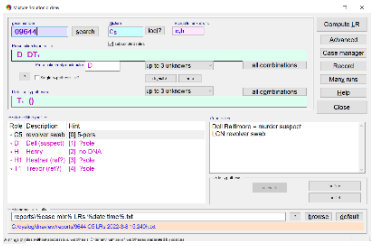
This section is an overview tutorial of the key tasks of mixture analysis for casework (§2.1.1) or automated validation (§2.1.2), and for importing mixture data (§2.2).

Mixture Solution analyses data from a *case*. A case is mainly a collection of mixture and reference *profiles*. Cases are denoted by *case numbers* like 2022-135 or 132499391 (a number – the hyphen doesn't matter). Each profile has a name or *role* like M, K11, or B007 – one uppercase letter followed by 0 or more digits.

- For a very quick first look, skip importing, go directly to §2.1 and use the pre-imported example case 128.
- Or, also for just learning, choose (and look at) one of the several tutorial case files provided in the **C:\dnaview\examples** subfolder and see §2.2.3. **Baltimore LCN mixture.xlsx** was interesting as a case (it shows the importance of considering differing numbers of contributors), and the file illustrates the Osiris format. **…\NPmix127.xlsx** is the simple case 127, equivalent to case 128 except the role naming is different. These cases are discussed in the help file.
- For real casework preparation see §2.2.

2.1 Mixture calculation

Full information starts with the **Mixture Solution** page of the help file.



Mixture Solution main form

2.1.1 Mixture analysis – basics

- The main form for mixture calculation is the **Mixture Solution** form. → From the **Casework** menu select **Mixture Solution**.[4]
- Fill in the important choices – case number, roles for the mixture and for prosecution and defense hypotheses. For other choices (output file and maximum numbers of unknown contributors) usually just press the two **default** buttons.
- If needed open the **Advanced** form, which has the parameters that, once determined, usually won't change from case to case.[5] It also has several default buttons and help buttons.
- Press **Compute LR**. A progress form monitors the computation (§1.2.2).

2.1.2 Mixture analysis by autopilot

A versatile "autopilot" mechanism allows computing many mixture analyses without user attention and with easy setup. See **Autopilot operation** under **Scripting and Validation** in the help file.

- A ***script*** occupies a sheet of an Excel workbook and defines a collection of ***tasks***, one task per row, one column per setting (case number, mixture, hypotheses, etc.). Each task can be a mixture analysis as in §2.1.1.
- The mechanism is efficient and user friendly.
    a. Automated script creation: Follow the basic procedure of §2.1.1 for the first task, except instead of **Compute LR**, touch **Record** then **Create script**. That creates an Excel script which contains task 1 as a model. The user doesn't need to memorize details like names of the settings.
    b. Abbreviation rule: No need to fill in the full details of a task, just the usually one or two settings that differ compared to the preceding task. The Excel sequence generator is often useful.

[4] Or select **Open case** or **Case Manager**.
[5] Like detection threshold, stopping rules, and *z* (§3.1.2)

i. Therefore modifying the script is also efficient.

c. Fast execution – if each task computes in seconds (see §3.4), 100 tasks take only minutes.

- Output from running the tasks includes a summary Excel spreadsheet, row per task. Graphs of speed and repeatability are already implemented and of course the user can create custom graphs.

2.2 Import case data

For details start from the **Import** page of the help file.

The importing steps:

1. Create one or several text or Excel files (typically from GeneMapper or Osiris) with profiles (genotypes; rfu is necessary for mixtures but may be omitted for references).
2. Decide on case number(s). Decide on role names for the mixtures and references. Manually add a new column with case numbers, roles, and (optional) brief descriptions.
3. Use **Import casework DNA profiles** (selection in the **File** dropdown) to import the data and create the case(s).

Sample info
22-135 M mask swab
22-135 K1 suspect

3 Various information

3.1 Preliminary

3.1.1 Installation

Download the installer and use the emailed password to install it in a folder with full user permission. For this document assume installation to **C:\dnaview**. The installer offers to put a start-up icon on the desktop.

A ***demo*** copy has a time restriction which can be extended on request, repeatedly and indefinitely for experimentation and evaluation, but only months at a time.

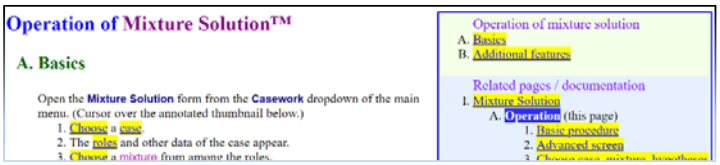
3.1.2 The stochastic ratio parameter z

Before live casework and serious exploration, evaluation, or validation are possible it will be necessary to estimate a laboratory-specific value for the stochastic ratio parameter z.[6] For initial exploration of the software though – for browsing and getting acquainted – it is fine to postpone that work and use the z value provided at software delivery.

3.2 Help file

A help file – 100+ pages, illustrated, cross-referenced and indexed – has much more information than this overview.

- **Help**, **Why?**, and **?** buttons on forms open a relevant help page.
- There's a **Glossary** page at the end with many definitions including some concept discussion. The **mixture** and **contributor** entries are interesting.
- Most pages have a navigation guide in the upper right corner with one or both of
  - index of this page (green)
  - a tree of links to nearby or to otherwise related pages (blue)



- There are many thumbnail images. Hold the cursor over the thumbnail to see the full-size image.

[6] Prepare some single-source samples in the laboratory from which the Mixture Solution "Stochastic model" tool estimates z. In effect z calibrates the rfu scale. It depends on the lab protocol – which multiplex, number of cycles, etc. – so it needs to be determined one time for each protocol. The **Stochastic** help pages are in **Research**.

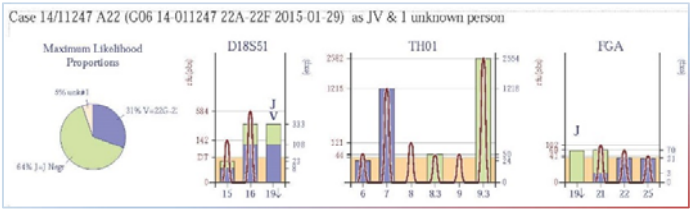
3.3 Program control console

On startup a console form appears. Use the menu bar to pop up other windows which will float independently. Nothing happens in the console space except some messages. There's no need to enlarge it.

3.4 More features

- Hardware: PC with 4 to 16 Gb, 1 core, 64 bit Windows
- Fast. Likelihood ratio calculation times ⟶
- Visual aids. Pictures explain, provide insight, expose data errors.

Number of persons	Time (typical)
2	2 seconds
3	6 seconds
4	1 minute



References

Brenner. (1996). Likelihood Ratios for Mixed Stains When the Number of Donors Cannot be Agreed. *International Journal of Legal Medicine*.

Brenner. (2019). *https://dna-view.com.* Retrieved 2022, from Forensic Mathematics: downloads/public/Determining the number of mixture contributors (B76).pptx

Gill, Brenner, & Buckleton, .. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International, 160*, 90-101.