GMOtrack: Generator of Cost-effective GMO Testing Strategies

—

Appendix

Formal Problem Definition

Let us represent a biological sample, possibly containing GMOs, by a pair (A, X), where A is a known set of potentially present GMOs, and X is an unknown set of GMOs actually present in the sample. The set X can be any, possibly empty, subset of A. GMO traceability requires that all the GMOs present in the sample are identified, which corresponds to determining the unknown set $X \subset A$.

Set X is determined by testing the sample with assays. A T-assay is characterized by test set T of the GMOs it detects; T is a non-empty subset of set A of potentially present GMOs. We assume that we have at our disposal a suite of assays, one T-assay for each $T \in \mathcal{T}$, where $\mathcal{T}$ is a given collection of non-empty subsets of set A.

We can perform T-assays on the unknown subset X. A T-assay tells us whether X has some element in T or not. The outcome of a T-assay performed on X is the truth value of $X \cap T \neq \varnothing$, which we write as $X\,!\,T$ and read "X tests T-positive". Let us introduce $!\,T$ as a mapping from $\mathscr{P}A$ to $\mathbb{B} = \{\mathit{false}, \mathit{true}\}$, which assigns to each subset S of A the truth value $S\,!\,T$; in formal notation:

$$!\,T\colon \mathscr{P}A \to \mathbb{B} : S \mapsto S\,!\,T\,.$$

In standard mathematical notation, $\mathscr{P}A$ is the set of all subsets of A, also called the powerset of A. For a given set A of cardinality $|A|$ there are $2^{|A|}$ subsets of A.

The outcome opposite to X testing T-positive is $X \cap T = \varnothing$, denoted $X\neg!\,T$ and pronounced "X tests T-negative". The property "tests T-negative" is then the mapping

$$\neg!\,T\colon \mathscr{P}A \to \mathbb{B} : S \mapsto S\neg!\,T\,.$$

A remark is needed here. The properties $!\,T$ and $\neg!\,T$ consistently describe the possible outcomes of a T-assay only under the assumption that the set of GMOs present in the sample is in fact a subset of set A of potentially present GMOs. However, it is always possible that the sample being tested contains some bacterial or viral residue or an unexpected 'unofficial' GMO outside A, which by chance happens to be detected by some assays. Although the probability of this happening is low, the assays need to be divided into two types: screening assays $\mathcal{T}_S$ and event-specific assays $\mathcal{T}_E$. The suite of assays $\mathcal{T}$ is the disjoint union $\mathcal{T}_S \cup \mathcal{T}_E$.

Screening assays generally detect more than one GMO. For the reason expounded in the remark above, we do not consider that a screening T-assay reliably detects the presence of some GMO belonging to T in the sample if the sample tests T-positive, though it does reliably detect the *absence* of all GMOs in T when the sample tests T-negative.

Each event-specific assay corresponds to a one-element subset of A. We assume that our suite of assays $\mathcal{T}_E$ is complete in the sense that for each $a \in A$ there is an event-specific $\{a\}$-assay in $\mathcal{T}_E$. Unlike screening assays, event-specific assays reliably detect GMOs: given an event-specific $\{a\}$-assay, a sample tests $\{a\}$-positive if and only if the sample contains a.

To confirm the presence of a GMO, the corresponding event-specific assay must be applied. In principle, it is possible to determine any subset X of A of the GMOs present in the sample, by outcomes of event-specific assays only; performing the event-specific T-assays for all $T \in \mathcal{T}_E$, we find X as the union of all those (singleton test sets) T for which the sample tests T-positive:

$$X \;=\; \bigcup \{\, T \mid T \in \mathcal{T}_E,\, X\,!\,T \,\}\,.$$

The role of screening assays is that their use may reduce the number of event-specific assays needed to completely determine the subset X. For example, suppose that a sample undergoes a combination of three screening assays, with test sets T_1, T_2, and T_3, and it tests T_1-negative, T_2-positive, and T_3-negative; then it follows that X is a subset of $B = A \setminus (T_1 \cup T_3)$, so to finish testing, only the event-specific $\{a\}$-assays with $a \in B$ are needed. Moreover, instead of resorting at this point to event-specific assays, the sample may undergo another combination of screening assays, and so on. It could happen that all the screening assays performed on the sample have positive outcomes; in such a case the number of needed event-specific assays is not reduced.

The testing cost of a sample depends on the number of performed assays and on how these assays are parallelized: all the assays have the same cost, but if they are performed in parallel their cost is lower. If we split the testing into n phases and in phase i we perform a combination of assays $\mathcal{U}_i \subseteq \mathcal{T}$, the total testing cost is the sum of costs for all the phases:

$$cost = \sum_{i \in [1..n]} g(|\mathcal{U}_i|)\,.$$

Here g is a monotonically increasing function that maps the number of assays performed in parallel to their cost, while the cost per assay $g(k)/k$ is a monotonically decreasing function of k.

To determine the quality of a certain combination of assays, we must take into account *probabilities* of testing outcomes. In detail: If sample X undergoes the combination $\mathcal{U} \subseteq \mathcal{T}$ of assays, then the testing outcome is

$$X!\mathcal{U} = \begin{pmatrix} T_1 & T_2 & ... & T_{|\mathcal{U}|} \\ X\,!\,T_1 & X\,!\,T_2 & ... & X\,!\,T_{|\mathcal{U}|} \end{pmatrix}$$

and the probabilities just mentioned are probabilities $P(X!\mathcal{U})$. Since we assume GMO independence, the probability of a testing outcome can be computed as

$$P\big((\bigwedge_{i' \in I'} \neg!T_{i'}) \wedge (\bigwedge_{i \in I} !T_i)\big) = \sum (-1)^{|K|} q(\bigcup_{k \in I' \cup K} T_k),$$

where $i' \in I'$ are indexes of T-assays that test negative, $i \in I$ are indexes of T-assays that test positive and $q(\cup_{\mathcal{U}} T)$ is the probability that all $T \in \mathcal{U}$ test negative; $q(T)$ is computed as $q(T) = \prod_{a \in T}(1 - p(a))$.

GMO Traceability as an Optimization Problem

The GMO traceability optimization problem is to find a GMO testing strategy with the smallest expected cost. An instance of an optimization problem can be defined in a formal way as a tuple $(D, f, extr)$, where

- D is the solution space (on which f is defined);
- f is the objective function $f\colon D \to \mathbb{R}$;
- *extr* is the extreme (usually *min* or *max*).

Our solution space D is a set of solutions d; we call these solutions *testing strategies*. Every $d = (V, v_0, E, \mathcal{U}, P)$ is a tree, where V is a set of vertices, v_0 is the root of the tree, the set of edges E is a set of ordered pairs of vertices, $\mathcal{U}\colon V \to \mathscr{P}\mathcal{T}$ associates with each vertex V a combination of assays, and $P\colon E \to \mathbb{R}$ are probability weights of edges. This structure also has to satisfy the following conditions:

- edges e out of a non-terminal vertex v are in one-to-one correspondence with testing outcomes and are weighted by the probabilities $P(e)$;
- for every non-terminal vertex v, the sum of the probabilities $P(e)$ for all edges e out of v is 1.
- for every terminal vertex t, every GMO $a \in A$ is either confirmed by an event-specific assay from $\mathcal{T}_e$ or repudiated by any assay from $\mathcal{T}$.

The objective function f ($f : D \to \mathbb{R}$) is the expected total cost of a testing strategy. It is the sum of the cost of the first phase of testing $g(|d.v_0|)$ and the weighted (with probabilities P_i) sum of the costs of all the successive testing phases. It is a recursive function.

$$f(d) = g(|d.v_0|) + \sum_{i=1}^{2^{|d.v_0|}} P_i \cdot f(v_i)$$

extr is *min* since we are looking for the feasible solution that minimizes the expected cost f ($argmin_{d \in D} f(d)$). According to this definition, GMO traceability is a combinatorial optimization problem.

Example:

A sample testing strategy is here exemplified on data from Table 1, which represents the GMOs allowed on the European Union market in the year 1997. In the table, each line is one GMO: the first column are GMO names and the second are the crops. The following column represents the probabilities of the associated GMO to be present in a sample. The other columns represent assays: screening assays that detect specific genetic elements inserted into GM crop genomes are listed first, followed by event-specific assays that allow the identification of a unique GMO. An 'x' at the intersection of a column with a row means that the corresponding GMO has the corresponding genetic element. The data were preprocessed as described in Section Data acquisition.

The previously introduced formal notation is instantiated with data from Table 1 as follows. Set A is the known set of potentially present GMOs: A = {RRS, GT73, Bt176, MS1, RF1, RF2, HCN92}. $\mathcal{T}_E$ is a suite of all event-specific assay-sets: $\mathcal{T}_E$ = {eRRS, eGT73, eBt176, eMS1, eRF1, eRF2, eHCN92}, where the *eGMO* denotes a one element set with the GMO. $\mathcal{T}_S$ is a suite of screening assay-sets: $\mathcal{T}_S$ = {P-35S, P-TA29, P-nos, Cp4 EPSPS, BAR, Barstar, T-nos, T-35S, P-35S::BAR}. An example of a screening assay-set is P-35S = {RRS, Bt176, HCN92}, meaning that the P-35S-assay detects RRS, Bt176 and HCN92.

An example of a testing strategy (element of D) is shown on Figure 1. The root v_0 is {P-35S, T-nos}. There are four possible testing outcomes: both *negative*, both *positive*, and one *positive* one *negative*. Probabilities of outcomes are denoted as $p_1, .., p_4$.

The parallel cost function g is here approximated with a linear function $g(k) = m \cdot k + b$ for a positive integer number of assays k, while $g(0) = 0$. We compute the expected cost f for the example strategy d_1 on Figure 1 as follows. Note how, in a two-phase

strategy, the calculation of the total expected cost is simplified.

$$\begin{aligned} f(d_1) &= g(|v_0|) + \sum_{i\in[1,4]} p_i \cdot f(v_i) = g(|v_0|) + \sum_{i\in[1,4]} p_i \cdot g(|v_i|) \\ &= g(2) + p_1 \cdot g(1) + p_2 \cdot g(4) + p_3 \cdot g(4) + p_4 \cdot g(7) \\ &= (2 + p_1 + 4p_2 + 4p_3 + 7p_4)k + 2n \end{aligned}$$

From this calculation we can see that a good testing strategy has low probabilities associated with vertices with large sets of assays. It can also be proven that testing strategy d_1 is not optimal, since, regardless the outcome of the first phase, eGT37-assay needs to be performed. Therefore, it should have been included in the first phase. The optimal two-phase testing strategy employs assays P-TA29, CP4 EPSPS and T-35S in the screening phase.

The GMOtrack algorithm

The computation of the optimal GMO testing strategy by exploring the whole solution space is infeasible, since the space grows exponentially relative to the number of assays. We have therefore simplified the problem by reducing the solution space, limiting the testing strategies to have two phases: a screening phase and an identification phase. The screening phase applies screening assays only. The identification phase consists of event-specific assays only: the event-specific assays for GMOs that have not been repudiated in the first phase are performed. By adopting this simplification, we can not guarantee to find an optimal testing strategy, but rather to find an optimal two-phase testing strategy.

The expected-cost function f for two-phase testing strategies can be formulated as follows, where n is the number of possible outcomes of the screening phase ($n = 2^{|v_0|}$):

$$\begin{aligned} f(d) &= \mathit{screeningCost}(d) + \mathit{expected\ eventspecificCost} = \\ &= g(|v_0|) + \sum_{i\in[1..n]} p_i \cdot g(|v_i|) \end{aligned}$$

An exhaustive algorithm for finding the optimal two-phase testing strategy is presented in Algorithm 1. Its input is set A of GMOs a with their probabilities $P(a)$, a suite of screening assays $\mathcal{T}_s$, the maximum assays in the first phase constraint m and a parallel cost function g. It generates and evaluates all two-phases testing strategies with up to m assays in the first phase and returns the best one.

The $\mathit{XuniqueCombinations}(\mathcal{T}_S, m)$ function (line 2) returns all subsets of $\mathcal{T}_S$ of size up to m. Function $p(e, A)$ (line 5) returns the probability of outcome e on set A. Function $\mathit{possible}(e, A)$ (line 5) returns the number of GMOs that are not repudiated by outcome e. It is computed as those GMOs that are not in the union of assay-sets that tested negative: $\mathit{possible}(e, A) = |A \setminus \bigcup\{T \mid T \in e, \neg!\, T\}|$.

Algorithm 1 GMOTRACK$(A, \mathcal{T}_S, m, g)$

Input: set A of GMOs a with their probabilities $P(a)$,
suite of screening assay-sets $\mathcal{T}_s$,
maximum assays for first phase constraint m,
parallel cost function g.
Output: optimal two-phase testing strategy d.

```
1: d : d.assays ← {}, d.cost ← ∞
2: for all d_i.assays in XuniqueCombinations(T_S, m) do
3:    d_i.cost ← g(|d_i.assays|)
4:    for all e in d_i.outcomes do
5:       d_i.cost ← d_i.cost + p(e, A) · g(possible(e, A))
6:    end for
7:    if d_i.cost < d.cost then
8:       d ← d_i
9:    end if
10: end for
```

Note that even though this is a simplified (restricted) version of the GMO traceability optimization problem, the computational complexity is very high (exponential relative to the number of screening assays). If we take $m = 5$ for the small example in Table 1 (9 screening assays), *XuniqueCombinations* (line 2) returns 381 possible solutions. For each possible solution there are up to 32 possible outcomes. For every outcome, the probability and the number of possible GMOs need to be computed. For example, if we had 50 screening assays and $m = 5$, $\sum_{i=1}^{5}\binom{50}{i} = 2,369,935$ candidate solutions would need to be checked with up to 32 possible outcomes each. If we had 50 screening assays and $m = 8$, $\sum_{i=1}^{8}\binom{50}{i} = 655,023,685$ candidate solutions would need to be checked with up to 256 possible outcomes each. For these reasons, the *GMOtrack* algorithm is not scalable, but as shown in our experiments, sufficient for practical situations.