Rational Design of Digital Assays

Pawel R. Debski,†‡ Kamil Gewartowski,‡ Magdalena Sulima,‡ Tomasz S. Kaminski,†‡ and Piotr Garstecki*†‡

†Institute of Physical Chemistry, Polish Academy of Sciences, Kasprzaka 44/52, 01-224 Warsaw, Poland
‡Curiosity Diagnostics Sp. z o.o. Kasprzaka 44/52, 01-224 Warsaw, Poland

Supporting Information

ABSTRACT: Optimum algorithm for digital assays treats chemical compartments as bits of probabilistic information and arranges these bits in a fractional positional system. Maximization of information gain reduces, by orders of magnitude, the number of partitions required to achieve the requested dynamic range and precision of the assay. The method simplifies the execution of digital analytical methods providing for more accessible use of absolute quantization in research and in diagnostics.

The standard “analogue” method to quantify concentration $C$ is to measure an amplitude $A$ of a physical quantity (e.g., absorbance of light) and to calculate an estimate of $C$ as a calibrated function $E(C) = f(A)$. In a fundamentally new concept, digital assays split the sample into a large number of partitions and estimate $C$ from the fraction of end-point signals. The concept dates back to 1915 when McCrady1 introduced the limiting-dilution assay and the most probable number method for quantization of bacteria. In spite of its common use over decades this method has never been optimized for the information gain as a function of the design of the assay.

In 1992 Sykes et al. described the method of quantization of initial targets present in the sample using limited dilution polymerase chain reaction (PCR)3−5 assays. Sykes’ design comprised 6 sets of 10 compartments. The consecutive sets were diluted 3-fold so the numbers of added DNA copies created a geometric sequence. Then, in 1999 Vogelstein and Kinzler proposed digital PCR assay using identical partitions of the sample and binary stochastic signals $s(C)$ adopting either a positive ($s = 1$) or negative ($s = 0$) value if the partition contains at least one target molecule ($m \geq 1$) or none ($m = 0$). Digital assays provide absolute, highly sensitive and precise5−9 estimates of $C$ and eliminate the need for calibration. As the method requires strong amplification of the presence of the analyte, its key applications are in quantitative PCR5−10 and ELISA.11

Digital methods have also important limitations. Their precision and dynamic range cannot be tuned independently. As a result the assays typically require massively large numbers of partitions: thousands to millions, elevating their cost and narrowing the range of applications. The problem stems from the coding scheme: as all compartments are identical, the maximum number $M$ of molecules that can be encoded is proportional to the number $N$ of partitions (Figure 1A). One solution is by combining classical digital schemes with different dynamic ranges.12,13 The solution relies on a simultaneously performed assays on multiple sets of identical compartments14−18 containing dilutions of the sample to increase the dynamic range. These approaches, although highly effective, have not been optimized for the information gain (the breath of the dynamic range and the tightness of the precision of the
estimate of $C$) against the design of the assay (the number, volume, and dilution factors of the compartments).

The method that we describe here minimizes the number of partitions and provides explicit formulas for the design of the assay for the requested dynamic range and precision of the estimate. This optimization is achieved via appropriate tuning of the expected number of molecules of the analyte (i.e., DNA molecules) in the partitions. The algorithms here described explicitly teach how to adjust the modification factor (i.e., the product of volume and dilution of the sample in each of the compartments) (see the Supporting Information, sections 1.1-1.4 for details).

The outcome of a digital assay (i.e., the collection of positive signals from a subset of partitions of the sample) codes the concentration of the analyte (i.e., the number of target molecules in the sample volume). The most natural way to code numbers is with equally valued bits; e.g., the number 157 can be represented by 157 dots. With some sophistication stemming from the Poisson statistics, this is the scheme of coding the number of molecules (i.e., concentration $C$) with the number of positive signals in a classic digital assay comprising identical partitions of the sample. As a result, the classical approach requires a large number of partitions to target large dynamic range of concentration of the analyte.

The requirement of large number of bits to express large values is readily remedied by positional systems. For example, in binary coding one needs only 8 digits to code the value of 157. In a chemical assay, the value of the bit (compartment) reflects the likelihood that it contains at least one molecule of analyte. The Boolean character of the signals suggests the binary positional system, yet the signals in a digital PCR assay are stochastic variables of $C$, the volume ($\nu$) and the dilution factor ($d = C_{\text{compartment}}/C$). Here, it is assumed that the efficiency of a single-molecule amplification is 100%, regardless of the volume and/or dilution ratio of any particular compartment. Therefore, the probability of obtaining a particular signal is based on Poisson distribution and equals

$$P_{\text{positive}} = p(m \geq 1|C) = 1 - e^{-E(m)} \quad \text{or} \quad P_{\text{negative}} = p(m = 0|C) = e^{-E(m)}$$

where $E(m) = dvC$ is the expected number of molecules in the compartment. As a result, the signal carries only a probabilistic information about the unknown concentration of the analyte. This information can be expressed with the Bayesian formalism as the probability $\rho(C)$ that a given value of $C$ has caused the recorded outcome. The lack of prior knowledge of $C$ can be encoded with an initially constant density of probability $\rho_0(C) = 1/C_\infty$ of finding any particular value of $C$ between $C = 0$ and an arbitrary upper bound $C_\infty$. Then the information about $C$ gained from a single signal ($s = 1$ or $s = 0$) is $\rho(C|s = 1) = (1 - e^{-dvC})/(C_\infty - (dv)^{-1})$ and $\rho(C|s = 0) = dv e^{-dvC}$ (Figure 2A).

Importantly, the probability distribution $\rho$ comprises all the information extracted from the signal. The density of this information is not evenly distributed along the axis of $C$. Shannon entropy $H_b = \sum_i p_i \log p_i$ where $i$ counts the two possible outcomes, measures the information density and achieves maximum at $C^* = \ln(2)/dv$. The signal $s$ provides most information about concentrations near $C^*$, for which $p(s = 1|C^*) = 1/2$ (Figure 2B), while it carries little information on much smaller or much larger concentrations.

In an assay, the information on $C$ combines the inputs from multiple partitions. The probability of observing a particular combination of signals $\{s_0, s_1, ..., s_N\}$, that we call a microstate $\mu$, is given by a product of individual probabilities (Supporting Information section 2.4): $\rho(\{s_0, s_1, ..., s_N\} = \mu(C) = \prod_{i=0}^{N} p(s_i|C)$. The full information on $C$ is then given by $\rho(C|\mu) = p(\mu(C))/\int p(\mu(C)) \, dC$.

For identical compartments, the probability of obtaining exactly $N_p$ positive and $N_n$ negative signals can be expressed with the Bernoulli distribution: $p(N_p, N_n|N_p + N_n = N) = \mu(C) = (N/N_p)(p(1|C))^N_p[p(0|C)]^{N_n}$. The information about $C$ gained from the outcome of a classic digital assay has the same functional form as the information from a single compartment (Figure 2B,C). The precision of the estimate from the classic assay is not constant within the dynamic range and obtaining reasonable estimates about high concentrations requires a very large number of partitions.

In order to provide a constant level of information on $C$ within the requested dynamic range, the compartments (i.e., their $C^*$ values) should be spread along the axis of $C$. As typically the precision of assays is parametrized by a fractional precision (expressed in the percentage of the estimated value), the compartments should be distributed uniformly along the logarithmic scale of $C$: the values of $C^*$ should follow a geometric sequence. What remains to be determined is the common ratio, the value of the first term, and the number of compartments for a given requested dynamic range and precision of the estimate of $C$.

To find the optimum positional system, we start with an arbitrary, known input concentration $C_{\text{input}}$ and center the assay at a compartment $d_{\text{center}} = 1/2(C_{\text{input}})$, We
supplement this central compartment with $\Delta N$ compartments on each “side”: $d_{i\Delta} = d_{0\Delta} x^i, i \in (-\Delta N,\Delta N)$, i.e., with larger and smaller compartments, all arranged into a geometric series with the common ratio $x$. The probability of reading a positive signal depends on the modulation factor of the compartment: $p_i = 1 - e^{-C_{\text{modulation}}}$. Compartments much smaller or much larger than the central one will have their probabilities either very close to zero or to unity. Thus, the relevant information will come from a finite set (an “active stripe”) of compartments around the central one (Figure 3A). For any fixed value of $x$, increasing the

number $2N + 1$ of partitions improves the $E(C)$, yet only to a limit: compartments with their $d_{i\Delta}$ values much larger or much smaller than $d_{0\Delta}$ progressively carry less and less information about $C^\ast$. Indeed, the standard deviation $\sigma$ of $E(C)$ falls with increasing $\Delta N$ to the limit $\lim_{\Delta N \to \infty} \sigma(x)$ that is a function of $x$ only (Supporting Information section 2.7). This limit can be closely approximated by a simple algebraic fit: $\lim_{\Delta N \to \infty} \sigma(x) = \alpha(1 - x)^{\beta}$. Substituting the requested precision $\sigma_{\text{max}}$ for $\lim_{\Delta N \to \infty} \sigma(x)$ and inverting this equation allows us to find the common ratio $x$ of the geometric sequence of compartments as an explicit function of the requested precision: $x = 1 - \alpha \cdot \sigma_{\text{max}}$ (Supporting Information 2.7).

Then, in order to find the optimum number $\Delta N_x$ of compartments in the active stripe for the calculated value of $x$ we find the lowest integer value of $\Delta N$ at which the derivative of $\sigma$ with respect to $\Delta N$ is less than $1/1000$. This arbitrary condition ensures that further addition of compartments to the active stripe does not appreciably increase the precision of the assay. We found that $\Delta N_x$ can also be closely approximated by a simple algebraic expression $\Delta N_x = \gamma \cdot (1 - x)^{\delta}$ (Supporting Information section 2.7). The two equations for $x$ and for $\Delta N_x$ define the active stripe as a function of the requested precision of the assay.

The assay must also be designed for the requested dynamic range of concentrations within which the precision is to be at least the requested value. As the equations for the active stripe are true for any value of $d_{0\Delta}$ (i.e., for any $C_{\text{input}}$) and since the geometric progression is self-similar, in order to guarantee the same precision within a required range $C \in (C^\ast,C^\ast)$, it suffices to span the assay, keeping the required $x$ and the required “margins” of compartments outside $(d_{0\Delta},d_{\Delta N})$, with $d_{\Delta N}^2 = \ln(2)/C^\ast$. Thus, the assay can be completely designed with the use of the following equations that use the requested dynamic range $\Omega = C^\ast/C^\ast$ and maximum allowed standard deviation $\sigma_{\text{max}}$ of the estimate of $C$ as an explicit input:

$$x = 1 - \alpha \cdot \sigma_{\text{max}}^\beta$$

$$\Delta N = [\Delta N_x] = [\gamma \cdot (1 - x)^{\delta}]$$

$$N = 2 \cdot \Delta N + \log(x/(1/\Omega))$$

$$d_{0\Delta} = \ln 2x^{-\Delta N}/C^\ast$$

with $\alpha$, $\beta$, $\gamma$, and $\delta$ being positive constants (Supporting Information sections 1.1–1.4 and 2.7).

The above analytical expressions are derived from the approximations and demonstrate the relation between the precision of the assessment and design of the active stripe. The model can be used over a wide range of requested precision $\sigma_{\text{max}} < 0.89$.

The assay can be designed in the following steps: (i) for a given $\sigma_{\text{max}}$ and $C^\ast$ calculate $N_1$, $\Delta N$, and $x_1$; (ii) using $x_1$ and $\Delta N$ calculate the volume $d_{0\Delta}$ of the first compartment in the sequence, and (iii) create the sequence of $N$ partitions of volume $d_{0\Delta} = d_{0\Delta}x^i$ for $i = 1 \ldots N - 1$.

This recipe provides a powerful analytical tool. The dynamic range can be tuned independently of precision (Figure S7A in the Supporting Information). For given, requested $\Omega$ and $\sigma_{\text{max}}$, the assays require orders of magnitude less compartments than the classic digital assays. For example, in classical assays in order to address $\Omega = 10^4$ or $\Omega = 10^6$ one needs $N = 2 \times 10^5$ or $2 \times 10^6$ compartments, respectively. If $\sigma_{\text{max}} = 50\%$ is satisfactory, the same $\Omega$ can be assessed with as few as 35 or 47 compartments. More precise estimates (e.g., $\sigma_{\text{max}} = 25\%$ or $\sigma_{\text{max}} = 10\%$) require 140 and 795 or 192 and 1120 test-volumes, respectively.

The method that we describe can be used to assess (i) a wide range with good precision (e.g., $\Omega = 10^5$ with $\sigma_{\text{max}} = 50\%$ with $N = 65$), (ii) a wide range with high precision (e.g., $\Omega = 10^6$, $\sigma_{\text{max}} = 10\%$, $N = 1615$) or (iii) a narrow range with high precision and again ultrasmall $N$ (e.g., $\Omega = 10^3$ offering $\sigma_{\text{max}} = 30\%$ can be realized with $N = 80$ and offering $\sigma_{\text{max}} = 10\%$ with $N = 631$).
We note that the logarithmic distance between the compartments ($x$) is typically much smaller than the standard deviation of $E(C)$: $(x^{-1} - 1) \ll \sigma$. One may thus expect that the precision of the estimate roots more in the number of compartments falling into a given active stripe than from the fine gradation of their $C^*$ values. As generation of fine gradations (e.g., $x = 0.99$) may be technically challenging, we consider assays built with larger common ratios (i.e., smaller values of $x$) with copies of each compartment, similarity to the multivolume approach of Ismagilov et al.~\cite{8–10} yet again, reducing the number of compartments needed to achieve the requested precision within the requested range of concentrations. Formally, each partition characterized by the dilution factor $d_i$ and volume $v_i$ can be prepared $N^i$ times, yielding $N_{\text{tot}} = \sum d_i^{-1} N^i$ partitions in the assay. Then, the microstate of the assay is no longer a set of binary values but a set of real numbers proportional to the number of positive compartments belonging to the $i$th family $d_i v_i$.

We derived equations that allow to freely exchange the value of $x$ into the number of copies of compartments $N^i$ (Supporting Information sections 1.2, 1.4, and 2.8). For example, $\Omega = 10^8$ and $\sigma_{\text{max}} = 10\%$, can be realized with $N^i = 1$, $x = 0.986$ and $N_{\text{tot}} = 1123$, or with $N^i = 10$, $x = 0.869$ and $N_{\text{tot}} = 1211$, or $N^i = 100$, $x = 0.2$ and $N_{\text{tot}} = 1400$. Again the assay can be completely designed from the input of $\Omega$, $\sigma_{\text{max}}$ and the maximum allowed value of $x$ (Supporting Information sections 1.2 and 1.4), with the limitation that $\sigma_{\text{max}} < 0.89$ and $x \in (0,1)$.

So far we used the microstates, i.e., the combination of signals from identifiable compartments. The classic digital assays use the sum of positive signals, as it does not require tracking the identity of the compartments during amplification. The method that we propose here can also be used with nonidentifiable compartments. The loss of information associated with the lack of knowledge from which compartment the signal comes from is small. This follows from the observation that the probability function that a given concentration caused the recorded sum of the signals comprises inputs from analogous functions for the microstates. This summed probability function is dominated by the most probable microstates (in the simple example shown in Figure 3B the most probable microstate contributes more than 60% to the stochastic outcome).

As a result, the precision of an assay operated on the sum of signals is only slightly smaller than the precision of the estimate based on the microstate. We approximate (Supporting Information section 2.9) that the standard deviation $\sigma_k$ of the estimate $E(C)$ based on the sum of signals is the following function of $\sigma_{\text{max}}$ as we used this parameter in the derivation of the assay based on the microstates: $\sigma_k \approx 1.023 \sigma_{\text{max}} + 0.021$ (linear part of the plot in Figure S8b in the Supporting Information, $\sigma_{\text{max}} \leq 0.55$). In other words, when designing an assay based on the sum of signals it is sufficient to replace the requested maximum standard deviation $\sigma_{\text{max}}$ in the equations with the expression $(0.978 \sigma_{\text{max}} - 0.02)$. We have also listed the guideline on how to construct the assay for protocols that do not track the identity of compartments in the Supporting Information sections 1.3 and 1.4.

The algorithms for the design of the assay use as an input the dynamic range within which the assay should return an estimate of concentration of the analyte with the requested precision. Depending on the method of analysis (identifiable and nonidentifiable compartments) and technical means for generation of the partitions of the sample (i.e., limitation in the minimum dilution factor), the assays can be divided into four groups, each having a different design protocol (Figure 4).

In short, the first decision is whether to execute the assay with tracking of the compartments during amplification or not. Tracking of the identity allows to extract full information from the outcome of the assay, yet tracking might be experimentally difficult. The second decision concerns the modulation factor. If the optimum value of the modulation factor is too high (i.e., too close to unity), one can input the maximum allowed value of the modulation factor to obtain the design of the assay with slightly higher number of compartments, yet one that can be easily realized in practice. The assays designed according to the method here described satisfy the requirements and compare well to the existing methods.\cite{8–10} Figure S5b–d shows a set of outcomes of the assays designed to deliver the precision $\sigma_{\text{max}} = 17\%$ over the dynamic range of $\Omega = 7650$, the same as in an exemplary assay comprising $N = 640$ compartments designed by Kreutz et al.\cite{10} (Figure S5a). The assays designed according the method here described deliver the requested precision uniformly over the whole dynamic range, while using less compartments ($N =$...
When allowed to use the same number of compartments $N = 640$ it is possible to improve the precision within the range to $\sigma_{\text{max}} = 11\%$ (Figure 5e,f).

We also tested the assays experimentally. In one experiment we designed a set of $N = 16$ (Figure 6), 32, 48, and 96 (Supporting Information section 4.1) compartment assays that cover the dynamic range $\Omega = 10^4$ with precision $\sigma_{\text{max}}$ varying from 70\% to 30\%.

Additionally we verified the assays with Monte Carlo simulations, including four assays designed to cover dynamic range $\Omega = 10^6$ and $\Omega = 10^9$ with precision $\sigma_{\text{max}} = 15\%$ and $\sigma_{\text{max}} = 50\%$ (Supporting Information section 5). All these results agree with the analytical predictions.
DISCUSSION

The proposed design of an assay is based on three assumptions. First, we use Bayesian approach to derive probability distribution of C. Second, we use the most probable microstate, i.e., the most probable output state of an assay for a given concentration to determine the precision of the assessment as the standard deviation of the probability distribution of C. Other approaches have also been tried, for example, Kreutz et al.\textsuperscript{10} determine the concentration using the most probable number method and precision of the assessment by the amount of Fisher information assuming the upper limit of precision given by the Cramer-Rao inequality. As the existence of such an ideal estimator cannot be rigorously assumed, we rested our method simply on the most probable precision of the estimate. The two methods are similar in that they both use a collection of compartments of different volumes to construct the assay. The differences include (i) the architecture of the solution and the assumptions accepted in the derivation, (ii) the actual number of compartments needed to realize the requested functionality (our model provides for a smaller number of compartments), and (iii) explicit mathematical formulation of the recipe for designing the assay. Finally, we assumed that the efficiency of the single-molecule amplification is 100% and does not depend on the volume and/or dilution ratio of a compartment. However, for real samples, the efficiency of amplification may depend on surface-to-volume ratio and on the extent of dilution of the sample. In such a case, running a series of control experiments is advised.

In conclusion, we have shown that it is possible to extract full information from stochastic realizations of digital assays. The optimized use of information\textsuperscript{20} allows to reduce the number of compartments needed to run the assay with respect to the multivolume and multidilution approaches.\textsuperscript{12,13} The algorithm that we described allows for rational and simple design of an assay with independently tuned dynamic range and precision, requiring orders of magnitude less partitions of the sample than the classic, single volume digital schemes.

It is further possible to construct assays that provide different required precision in different ranges of concentration. This could be useful whenever, e.g., at low and high concentrations precision can be sacrificed for the dynamic range, while in the intermediate range of concentration the exact estimate is clinically informative (Supporting Information section 3.2).

The drastic reduction of the number of partitions minimizes the technical requirements for running digital assays. These features may be important in bringing the reference-free, absolute, digital quantization of DNA to multiple new formats and in increasing the availability of digital PCR and digital immunoassays. The assays proposed here can offer same precision and sensitivity as the classic digital schemes. Because the assays use small number of partitions, these are likely to be large and thus not providing the advantage of increased specificity of PCR.\textsuperscript{12}

The dramatic reduction of the number of partitions qualitatively changes the technical requirements to run the assays and could perhaps lead to the development of an alternative technology for the RT-PCR to provide absolutely accurate estimates at requested precision and at the throughput of modern real-time devices. Finally, for the same reason for decreased technical requirements to execute the assays they may find use in quantitative identification of viral and microbial pathogens in the point-of-care applications.

ASSOCIATED CONTENT

Supporting Information

Practical guideline how to design a digital assay that provides the required dynamic range and precision of the assessment, derivation of the model, other possible applications of the algorithms, experimental verification of the algorithms, and numerical verification of the algorithms. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b00942.

AUTHOR INFORMATION

Corresponding Author

E-mail: garst@ichf.edu.pl.

Author Contributions

P.R.D. and P.G. developed the algorithms, analyzed the data, and wrote the manuscript; T.S.K. designed the experiments and, together with P.R.D. and P.G., and implemented the algorithms for the diagnostic assays; K.G. and M.S. executed the experiments.

Notes

The authors declare the following competing financial interest(s): P.R.D., T.S.K., and P.G. declare competing financial interest in possession of shares of Curiosity Diagnostics Sp. z o.o. that seeks commercial implementation of the method.

ACKNOWLEDGMENTS

The authors thank Curiosity Diagnostics Sp. z o.o. for financing the research. Curiosity Diagnostics Sp. z o.o. acknowledges the E‘8042 OPTIGENS Grant provided by the National Centre for Research and Development within the Eureka Initiative. P.G. acknowledges support within the European Research Council Starting Grant 279647 and within the Idea dla Polski subvention from the Foundation for Polish Science.

REFERENCES

One should remember, that the model provides the relation between a point in the "performance space" (range and precision) and a point in the "design space" (number and character of the partitions). The equations do this only for assays designed according to our prescription. These equations cannot be used to analyze the performance (range and precision) of assays designed according to different principles.