Published online 8 August 2013

Nucleic Acids Research, 2013, Vol. 41, No. 18 ***e173***
doi:10.1093/nar/gkt682

Targeted resequencing of HIV variants by microarray thermodynamics

Wahyu W. Hadiwikarta[1,2], Bieke Van Dorst[3], Karen Hollanders[1], Lieven Stuyver[3], Enrico Carlon[2] and Jef Hooyberghs[1,4,*]

[1]Flemish Institute for Technological Research, VITO, Boeretang 200, B-2400 Mol, Belgium, [2]Institute for Theoretical Physics, KULeuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium, [3]Janssen Diagnostics bvba, Turnhoutseweg 30, B-2340 Beerse, Belgium and [4]Theoretical Physics, Hasselt University, Campus Diepenbeek, Agoralaan - Building D, B-3590, Diepenbeek, Belgium

Received February 28, 2013; Revised July 10, 2013; Accepted July 14, 2013

ABSTRACT

Within a single infected individual, a virus population can have a high genomic variability. In the case of HIV, several mutations can be present even in a small genomic window of 20–30 nucleotides. For diagnostics purposes, it is often needed to resequence genomic subsets where crucial mutations are known to occur. In this article, we address this issue using DNA microarrays and inputs from hybridization thermodynamics. Hybridization signals from multiple probes are analysed, including strong signals from perfectly matching (PM) probes and a large amount of weaker cross-hybridization signals from mismatching (MM) probes. The latter are crucial in the data analysis. Seven coded clinical samples (HIV-1) are analyzed, and the microarray results are in full concordance with Sanger sequencing data. Moreover, the thermodynamic analysis of microarray signals resolves inherent ambiguities in Sanger data of mixed samples and provides additional clinically relevant information. These results show the reliability and added value of DNA microarrays for point-of-care diagnostic purposes.

INTRODUCTION

In human genetic research, targeted resequencing of genomic nucleic acid is applied extensively in population studies to search for associations between sequence variants and diseases. The technique is also applied in diagnostic or prognostic tests. In this context, one is often confronted with samples of mixed sequence variants, some possibly present in minority: e.g. biopsies from cancer tissue usually contain a mixture of cancerous and non-cancerous cells. Therefore, one needs to distinguish the presence of a specific mutation, possibly in low abundance (minority), in a majority of 'wild-type' sequences. Several techniques are in use to resolve the sequence composition in mixed sequence samples, like allele-specific PCR (1), melting curve analysis or sequencing (2).

In this article, we focus on the case of HIV-1/AIDS. To reduce the morbidity and mortality world-wide, there is a high need for simple point-of-care genotyping tests to screen for key resistance mutations. However, the development of a simple genotyping test to screen for key resistance mutations is a technical challenge, due to high genetic variability of HIV-1 (3). The variability is caused by the error-prone reverse transcriptase enzyme, which will introduce mutations during each replication cycle, combined with a short replication time. Within one single patient, different, but closely related, non-identical viral genomes can be present. This high variability makes the design of primers and probes for a simple genotyping assay difficult. A variety of high throughput genotyping assay for antiretroviral resistance testing are available in the market, which allow physicians to determine drug-resistance profiles (4–6). These genotyping assays are using capillary electrophoresis platforms that provide integrated systems for nucleotide sequence-based analysis and interpretation for drug-resistance mutations in the HIV-1 reverse transcriptase and protease. The development of these assays greatly advanced clinical care for HIV-1 patients, by allowing personalized disease management, using the most appropriate drugs and drug combinations available at any given point (6). These assays are requiring high-tech equipment and are performed by highly trained laboratory personnel limiting their practical use as point-of-care test.

In this article, a new method based on microarrays hybridization will be introduced to perform targeted resequencing on nucleic acid samples. The idea to use

*To whom correspondence should be addressed. Tel: +32 14 33 5277; Fax: +32 14 32 1183; Email: jef.hooyberghs@vito.be

hybridization for mutation detection is not new (7,8), and it has often been compared with other techniques (9,10). An advantage of hybridization is its simplicity and the possibility of miniaturization for point-of-care tests. An often reported disadvantage is its specificity: the possibility of cross-hybridization of not-perfectly matching sequences to a probe sequence complicates the data analysis. The situation complicates even further when the original sample contains two or more variants of a given sequence. As usually cross-hybridization is viewed as a limiting factor, efforts are often aimed at avoiding it, e.g. by introducing chemical agents in the nucleic acid probes (11,12). In this article, we show that if the probe-target affinities for cross-hybridization are quantified accurately, the measurements from multiple probes, which typically are not perfectly matching to the sample sequences, can be turned into a powerful targeted resequencing method. The analysis relies on estimates of the hybridization free energies of mismatching duplexes (cross-hybridization signals). The data are then checked against the isotherm expected from equilibrium thermodynamics (13–15).

MATERIALS AND METHODS

HIV samples

HIV-1 virus stocks were selected from the Janssen Diagnostics repository database, based on their known mutation profile in the region of codon 179 to codon 186 of the Reverse Transcriptase (RT) gene. This region was selected to cover key resistance mutations at position 179, 181 and 184. A mutation at position 179 or 181 causes resistance against non-nucleoside RT inhibitors (NNRTIs) (16–20), while nucleoside RT inhibitors (NRTIs) are selecting for a mutation at position 184 (21,22). An informed consent for research purposes is available for the HIV samples used.

Experimental protocol

The viral RNA extraction of virus stocks was carried out on an EasyMAG (bioMérieux, Boxtel, The Netherlands) according to the guidelines of the manufacturer, starting with 256 µl input material for plasma samples and 100 µl input material for virus stocks, both were eluted in 60 µl. A One-Step RT-PCR amplification (One-Step Superscript III HiFi, Invitrogen, CA, USA) was used to generate a 2.3-kb HIV-1 fragment (containing the gag-protease-reverse transcriptase (GPRT) region) using the 3-RT (5′-CATTGCTCTCCAATTACTGTGATATTTCTCAT G-3′) and 5-OUT (5′-GCCCCTAGGAAAAAGGGCTG TTGG-3′) primers. RNA input was 10 µl in a final volume of 35 µl. The complete amplification procedure was published by (23). The 2.3-kb HIV-1 outer fragment generated with the GPRT one-step PCR was used as template for the asymmetric amplification of the sequence around RT codon 184. Therefore, the HIV-1_Fw_184Cy3(5′-/Cy3/T AGAAAACAAAATCCAGAAATA-3′) and HIV-1_Rev_184 (5′-TGCCCTATTTCTAAGTCAGATCC-3′) primers were used, with the fluorescent labelled forward primer HIV-1_Fw_184Cy3 in excess to generate fluorescent labelled single-stranded DNA (ssDNA) fragments of 78 bp (containing RT codon 184). Forward primer HIV-1_Fw_184Cy3 and reversed primer HIV-1_Rev_184 were used at a concentration of 1 µM and 0.1 µM, respectively, with DNA input of 2 µl in a final volume of 100 µl. The microarray experiments were performed in an Agilent platform. Each experiment was performed, following the standard protocol discussed in (24). We considered hybridizing sequences of 25 nt. This is because in previous studies (13), sequences of this length were found to attain thermodynamic equilibrium after ~3h of hybridization (in the experiments, the hybridization time is of 17 h to ensure that equilibrium was reached).

The raw microarray data were subjected to a primary quality control using the Agilent Feature Extraction Software (Version 10.7). For all arrays, the spot centroids in the four corners of the microarray have been located properly and consequently the grid was placed correctly. The QC values for homogeneity and those for checking the hybridization and washing steps, all fall within the good range according to Agilent guidelines. In the present application, we work with single color, hence there is no need of color normalization. In addition, the analysis in extended previous hybridization experiments (13,15) showed that the experimental data closely follow the thermodynamics models in the full range of measured experimental intensities without additional normalization requirements.

Microarray design

Table 1 shows some sequences with high clinical frequency obtained with Sanger sequencing from a database provided by Janssen Diagnostics, of about 350 000 patients, on a region of 25 nt of the HIV-RT centred around codon 184. The sequence with the highest frequency occurs in only 25% of the patients. The other 75% of the cases contain mutations with respect to it. This makes the HIV an excellent test model to check the validity of the method presented in this article. Part of the samples in Table 1 consist of unique sequences. Other samples are mixed, i.e. HIV viruses with sequence differences within the 25-nt window considered coexist in a single patient. Further, the concept of unique and mixed sample has to be interpreted within the limited sensitivity of Sanger sequencing. This method can detect a mixture only if the relative abundance of the low abundance sequence is >20%. For instance, a mixed sequence sample with only 10% of low abundance sequence would be detected as unique sequence sample by the Sanger method (2).

To define a manageable and relevant diagnostic problem, each unique sample from the database was selected and ranked in decreasing order of clinical frequency. The goal was to diagnose the top 100 unique sequences and mixtures of two sequences thereof. This corresponds to the coverage of 82% of the whole database. The sequence no.143 shown in Table 1 is the 100th unique sequence relative to the most frequent one.

Table 1. Nucleotide sequences (from codon 179 to codon 186, written in 5′ to 3′ orientation) for different variants of the HIV-RT gene, as obtained from the analysis of 350 000 patients

Rank	Type	Sequence ...180...182...184...186.	Relative clinical Frequency
1	Unique	GTTATCTATCAATACATGGATGATT	1.000
2	Unique	GTTATCTATCAATAC**G**TGGATGATT	0.411
3	Unique	GTTATCTATCAATACATGGATGA**C**T	0.124
4	Unique	GT**C**ATCTATCAATACATGGATGATT	0.084
5	Unique	GTTATCT**G**TCAATACATGGATGATT	0.075
		⋮	
9	Unique	GTTATCTATCAATAC**G**TGGATGA**C**T	0.056
		⋮	
15	Mixed	GTTATCTATCAATAC**R**TGGATGATT	0.037
		⋮	
22	Unique	GT**C**ATCTATCAATA**T**ATGGATGA**C**T	0.027
		⋮	
143	Unique	GTTATCTATCAATACATGGATGA**CC**	0.002
		⋮	

The database, provided by Janssen Diagnostics, is obtained from Sanger sequencing, and only some selected sequences are shown here. The ranking follows the relative clinical frequency. Numbers above the first ranked sequence indicate the codons position. Nucleotides differing from those of the most frequent sequence are shown in bold. The Sanger sequencing method yields either *unique* sequences, i.e. with no ambiguities, or *mixed* sequences. In the Table, the mixed sequence with the highest clinical frequency is ranked no. 15. We use here the standard notation for degenerate bases (25); therefore, R means a purine (A or G). Note that this sequence is a mix between sequences ranked no. 1 and no. 2 in the Table. The sequence ranked no. 143 is the 100th unique sequence from the database and so is the last in the PM set.

This sequence has two mismatches with respect to the most frequent one.

The 15k custom Agilent array used in the experiment contains ~15 000 spots. The design of the probe sequences on the array was started from the 100 probes that are perfectly matching to the 100 unique sequences considered in the test (referred as the perfect match (PM) set in the design). The data analysis of the method discussed in this article is not solely based on the signal of perfectly matching spot, but also on cross-hybridizing signals. Previous experiments (15,25) showed that signals measured from spots with one or two mismatches with respect to the target sequence are still above the lower limit of detection. Therefore, probes with one or two mismatches with respect to those in the PM set were included in the microarray. The final design contains 2139 different probes replicated seven times to fill all the available spots of the 15k microarray.

Thermodynamic assessment

In the standard approach of hybridization-based targeted resequencing, the microarray contains probes that are perfectly matching to all possible variants of target sequences expected to be present in the biological sample. The brightest of all spots should indicate the sequence which is present in solution. Our approach is based on the analysis of the intensities of the brightest and also of many 'dimmer' spots which are expected to carry one or two mismatches with respect to the target (actually the large majority of signals comes from hybridizations to mismatching sequences). If a unique target sequence is present in solution, the fluorescence intensity I from different spots will be correlated according to equilibrium thermodynamics as

$$I = Ace^{-\Delta G/RT} \quad (1)$$

where A is a proportionality factor, c the target concentration in solution, ΔG the hybridization free energy as a sequence-dependent measure of the affinity between probe-target sequences, R the gas constant, and T the temperature. For convenience in the rest of the article, hybridization free energies are shifted by the corresponding perfect match value. This amounts to use $\Delta\Delta G \equiv \Delta G - \Delta G_{PM}$; therefore, for a PM hybridization $\Delta\Delta G = 0$. Note that as the free energies are shifted by a constant value, the same functional relationship as Equation (1), holds also for $\Delta\Delta G$. Equation (1) holds at sufficiently low concentrations, while at high concentrations saturation effects should be taken into account, as the intensity reaches a maximal value when all probes are hybridized [as predicted by the Langmuir model (24)]. These effects are not relevant in the range of concentrations investigated here. Equation (1) was thoroughly tested and validated in experiments on Agilent arrays using several different sequences (15,25).

Given the sequences of the target in solution and the surface-bound probes, the $\Delta\Delta G$ can be obtained from the nearest-neighbour model (26). In this model, the hybridization free energy is given by the sum of parameters that are dependent on pairs of neighbouring nucleotides. The computation of $\Delta\Delta G$ used in this article follows the same principles described in (15,24).

In a resequencing analysis, the target sequences are coming from a biological sample, and they are usually not known. One can, however, make a starting hypothesis about the sequence and compute the corresponding $\Delta\Delta G$. If the starting hypothesis agrees with the actual sequence in the sample, the measured intensities should be distributed according to Equation (1). Deviations from this law may have two causes: (i) The starting hypothesis is wrong; hence, the sequence in the sample is different from what originally assumed or (ii) the sample is a mixture, i.e. it contains the sequence of the original hypothesis together with other sequences. This concept is illustrated in Figure 1 that shows plots of I versus $-\Delta\Delta G$ in log-linear scale for which Equation (1) becomes a straight line with slope $1/RT$ (shown as dashed line). The two plots refer to the same experimental data. In one case (Figure 1a), the hypothesis matches the actual sequence in the sample. The data accurately follow Equation (1) over four orders of magnitude in the intensity scale. In the other case (Figure 1b), the wrong hypothesis leads to an incorrect computation of the $\Delta\Delta G$ and deviations from the expected thermodynamic behaviour. Here, the data are distributed into four distinct branches. The origin of this branching will be discussed in detail in the next section, which presents an algorithm used to infer the sequence composition from the analysis of the plots of the measured fluorescence

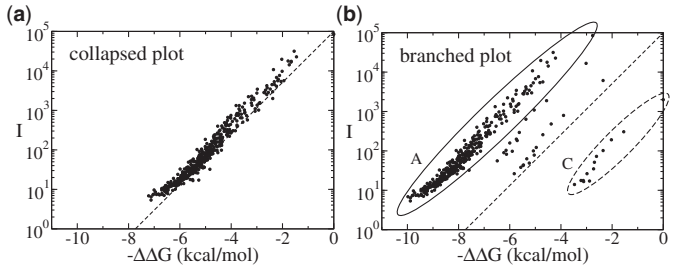

Figure 1. Two I versus $-\Delta\Delta G$ plots from the same set of experimental data for the hybridization of a synthetic oligomer target sequence 5′−AAGGGCCACGGATTACTCGTAATAA−3′ to a microarray containing a perfect match probe and many probes with one or two mismatches with respect to the target. The $\Delta\Delta G$ are calculated using the nearest neighbour model free energies by means that are described in (15,24), using two different hypotheses for the target sequence in solution. In (**a**) the hypothesis matches the actual target sequence and the data follows the Equation (1) that is shown as a dashed line. In (**b**) the hypothesis differs by one nucleotide with respect to the actual target sequence in solution. Four different branches appear in this plot. The origin of these branches is discussed in the text.

intensities I versus $-\Delta\Delta G$. We refer these plots as $I\Delta\Delta G$-plots.

Algorithm

Figure 2 shows a flowchart of the algorithm used for the targeted resequencing. The algorithm consists of a central loop that generates iteratively sequences $\ldots t_{n-1}, t_n, t_{n+1} \ldots$ (where n is the iteration step), which are successive *in silico* hypotheses about the composition of the sample. The loop is repeated until convergence is reached. Two types of convergence are obtained. In some cases, after a certain number of iterations, the algorithm shows a collapsed $I\Delta\Delta G$ plot similar to that seen in Figure 1a. This is a signature of the presence of a unique sequence in the sample. In this case, the algorithm ends at the block (b) of the flowchart and returns the output t_n as the sequence composition of the sample. In other cases, the algorithm converges to a two-cycle state i.e. $t_n = t_{n+2} = t_{n+4} = \ldots$ and $t_{n+1} = t_{n+3} = \ldots$, where $I\Delta\Delta G$ plots are always branched (similar to that seen in Figure 1b). This two-cycle state is a signature for a sample composed by a mixture of two sequences: t_n and t_{n+1}. In this case, the algorithm ends at the block (e) and returns the aforementioned two sequences as output.

An important part for generating new *in silico* hypotheses is the decision block (a) of the flowchart. At this point, the algorithm checks if the $I\Delta\Delta G$ plot is collapsed or branched. In the latter case, a new *in silico* hypothesis t_{n+1} is generated. To understand how this is done, the origin of the branching has to be elucidated in some detail.

The plot of Figure 1b is obtained by calculating $\Delta\Delta G$ using a wrong *in silico* hypothesis: Table 2 shows the actual sequence in solution used in the experiment to produce Figure 1a and the hypothesis sequence made for the calculation of Figure 1b. They differ by a single nucleotide at base position 13 (in the *in silico* hypothesis this is a G, while the actual target contains a T). Consider now

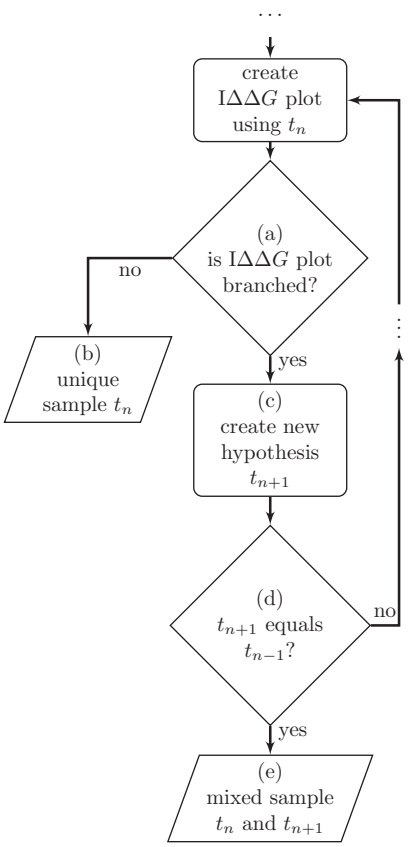

Figure 2. A flowchart showing the basics of the algorithm for targeted resequencing of HIV-RT. t_n is the hypothesis for the target sequence generated at the n-th iteration. The outputs are either a unique sequence [block (b)] or a mixed sequence composed by two sequences [block (e)] depending on the nature of the $I\Delta\Delta G$ plots.

all probe sequences in the microarray with an A at this position. The actual hybridization is a Watson-Crick AT pairing, while the *in silico* hypothesis estimates the $\Delta\Delta G$ as these were AG mismatches. This leads to overestimating the $-\Delta\Delta G$. The data points corresponding to the probes with nucleotide A at base position 13, are encircled with a solid line in Figure 1b. Conversely, for the probes with a nucleotide C the $-\Delta\Delta G$ are underestimated. The latter are encircled with a dashed line in Figure 1b. Thus, the splitting into four branches is due to the wrong estimates of free energy for each probe in the position where actual target and *in silico* hypothesis differ. It is important to notice here that the probes in the different branches systematically differ from each other by specific nucleotides at specific locations. This systematic sequence deviation will be used to decide whether the $I\Delta\Delta G$ plot is branched or not (see right panes in Figures 3 and 5 of the examples in the next section). The correct hypothesis can readily be constructed by selecting out the top left branch, determining the systematic sequence deviation (nucleotide position and type) in this probe subset, and implementing this nucleotide change in the previous hypothesis. This is precisely how a new *in silico* hypothesis denoted by t_{n+1} is generated in block (c) by the algorithm.

Table 2. Origin of branches appearing in the $I\Delta\Delta G$ plot

Targets	Actual AAGGGCCACGGA**T**TACTCGTAATAA hypothesis: AAGGGCCACGGA**G**TACTCGTAATAA	Actual	Hypothesis
Probes	----------**A**----------	PM	MM
	----------**G**----------	MM	MM
	----------**T**----------	MM	MM
	----------**C**----------	MM	PM
	01--------13--------25		

The top part of the graph shows two target sequences. The first sequence is the one that was used in the experiment that generates the data shown in Figure 1. The second sequence is a (wrong) hypothesis about sequence composition. These two sequences differ by a single nucleotide at base position 13. This base position is indicated by the numbers shown in the bottom. The microarray probes are grouped into four sets according to the type of nucleotide at the position where the two targets differ. The use of a wrong hypothesis leads to an error in the $\Delta\Delta G$ computation. For instance, probes with a nucleotide A at base position 13 are treated as having an AG mismatch (MM), while actually it has AT perfect match (PM).

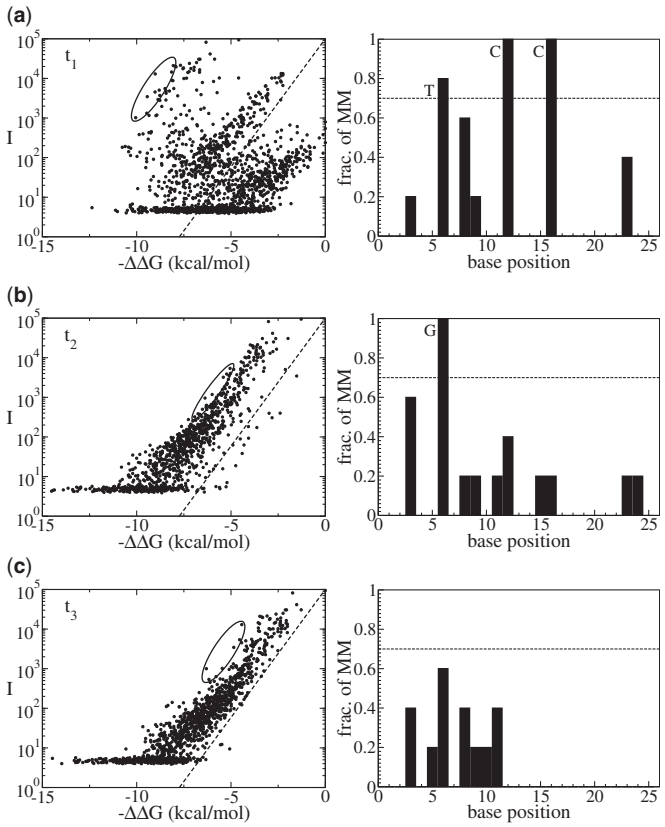

Figure 3. Analysis on sample no. 4; the first iteration tells that the hypothesis t_1 is not correct, as the $I\Delta\Delta G$ plot is branched (**a**). Analysis on the selected deviating points (encircled) reveals mismatching nucleotides per base positions that are commonly found in the selected probes. These nucleotides become the basis to generate a new hypothesis t_2. In the second iteration with the new hypothesis t_2, the $I\Delta\Delta G$ plot is apparently still branched (**b**); thus, the iteration continued by generating another new sequence t_3. The $I\Delta\Delta G$ plot with hypothesis t_3 turns out to be a collapsed plot; thus, the algorithm stops and conclude that the sample contains unique sequence (**c**).

In principle, for a unique sequence sample, one can start with any hypothesis and iterate until a collapsed $I\Delta\Delta G$ plot is found. However, when the sample is a mixture of two target sequences, the plots will always have branches, as over- and under-estimation of the data during the $\Delta\Delta G$ calculation will always occur. For each iteration, a new hypothesis is generated and when the t_{n+1} sequence is equal to the t_{n-1} sequence as evaluated in block (d), the algorithm is stopped and gives as output a mixed sample composed by sequences t_n and t_{n+1}.

RESULTS AND DISCUSSION

Decoding a set of coded samples

The algorithm was tested on seven coded clinical samples selected by Janssen Diagnostics out of a repository of samples for which the Sanger sequencing had been performed. The sequences composing these samples are shown in Table 3 where the left part of the Table shows the Sanger sequencing data provided as reference and the right part of the Table shows the sequences that were found by the algorithm. It can be seen that the proposed method decoded the samples successfully. To illustrate the working of algorithm, we discuss here in more details the sequences no. 3 and no. 4 of the seven sequences shown in Table 3, where the algorithm converges to a mixed and unique sequence, respectively. For convenience, in these two examples, the initial hypothesis t_1 corresponds to the sequence with highest clinical frequency (ranked no.1 in Table 1). We tested that the algorithm also converges to the same results when t_1 is one of the other 99 sequences of the PM set.

The first example is sample no. 4. The left pane of Figure 3a shows an $I\Delta\Delta G$ plot produced from the initial hypothesis t_1; in this first iteration the data are very scattered. First, a set of data points (encircled) is selected. These are points that deviate the most from the expected thermodynamic behavior (dashed line). The nucleotide composition of this set is analysed: the right pane Figure 3a shows a histogram of mismatches of this selected set with respect to the initial hypothesis t_1. In the algorithm, a threshold value of 70% is chosen as the minimum limit for the fraction of common mismatches per base positions between the selected probes and the current hypothesis (more details on the threshold selection and its influence on the output are presented in the Supplementary Section II). Based on this threshold, the selected probes are considered mismatching against the hypothesis t_1 at base positions no. 6, 12 and 16, by nucleotides T, C and C, respectively. This information is used to generate a new hypothesis t_2 by swapping nucleotides on hypothesis t_1 so it becomes complementary to the three nucleotides detected in the mismatching base positions (resulting sequence t_2 is given in Figure 4). The next iteration is the calculation of the $I\Delta\Delta G$ plot from t_2, which is the graph shown in Figure 3b. Although, compared with t_1, there is a better agreement with the expected thermodynamic behavior from Equation (1) (dashed line), the data are still scattered. The analysis of the most deviating set of points (encircled in Figure 3b) indicates that there is

Table 3. Seven resequenced clinical samples from Janssen diagnostics

No.	Sanger sequencing 179 180 181 182 183 184 185 186.	Proposed method 179 180 181 182 183 184 185 186.
1	GTT ATC TAT CAA TAC RTG GAT GAY T	GTT ATC TAT CAA TAC GTG GAT GAT T GTT ATC TAT CAA TAC ATG GAT GAC T
2	GTT ATC TGT CAA TAC ATG GAT GAT T	GTT ATC TGT CAA TAC ATG GAT GAT T GTT ATC TAT CAA TAC ATG GAT GAC T
3	GTT ATC TAT CAR TAC ATG GAT GAY T	GTT ATC TAT CAG TAC ATG GAT GAT T
4	GTT ATC TAT CAG TAC GTG GAT GAT T	GTT ATC TAT CAG TAC GTG GAT GAT T GTT ATC TAT CAA TAC GTG GAT GAT T
5	GTT ATC TAT CAA TAC RTR GAT GAT T	GTT ATC TAT CAA TAC ATA GAT GAT T GTT ATC TAT CAA TAC ATG GAT GAC T
6	GTT ATC TAT CAA TAC RTG GAT GAC T	GTT ATC TAT CAA TAC GTG GAT GAC T GTT ATC TAT CAA TAC GTG GAT GAT T
7	GTT ATY TAT CAA TAC RTG GAT GAT T	GTT ATT TAT CAA TAC ATG GAT GAT T

The numbers in the top row are indicating codon locations. Sequences on the left are the Sanger data whereas sequences on the right are the result from the method proposed in this article. The underlined nucleotides are mismatches against the sequence of highest clinical frequency including the degenerate bases i.e. R (purines: G or A) and Y (pyrimidines: C or T) (25). All sequences in this table are written in 5′ to 3′ orientation.

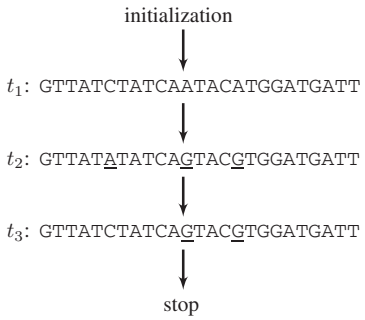

Figure 4. The hypothesis sequences from each iteration of the algorithm in the case of sample no. 4. It presents a conclusion that the sample contains a unique sequence t_3 after three iterations.

still a mismatch at base position no.6. The next *in silico* hypothesis t_3 (Figure 4) is then obtained. The analysis of the most deviating set of points that corresponds to $I\Delta\Delta G$ plot from t_3 (Figure 3c) does not provide a strong signature for any common mismatch. Therefore, the algorithm concludes that sample no. 4 contains a unique sequence t_3.

The second example to be discussed is sample no. 3. Figure 5 shows the $I\Delta\Delta G$ plots obtained during the iterations. In this case, the analysis of the most deviating branch of points generate a cycle in which sequences t_2 and t_4 are identical (Figures 6), indicating that the sample is actually a mixture of two sequences, differing in two nucleotides from each other. The difference with the case of sample no. 4 shown in Figure 3 is that in this case the $I\Delta\Delta G$ plot is always branched in all iterations forming a two-state cycle. The decision on whether a plot is branched or not is done by the algorithm based on the fraction of mismatches in the right panes, but note by visual inspection that in $I\Delta\Delta G$ plot of hypothesis t_2 (Figure 5b), the data points are quite close to the expected thermodynamic behavior of Equation (1) (dashed line). However, they are still more scattered

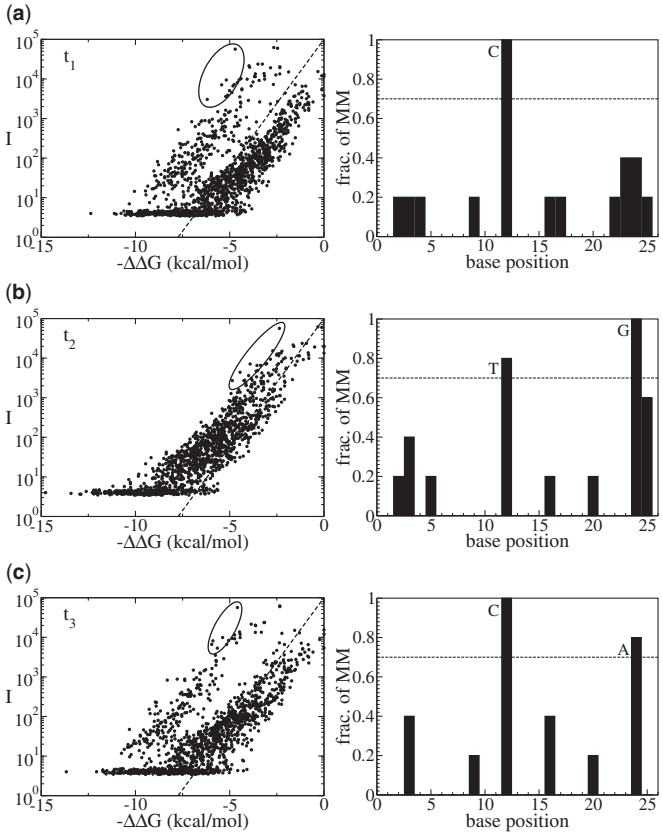

Figure 5. Analysis of sample no. 3, which is a mixed sample; in contrast to the analysis of unique samples such as shown in Figure 3, in the current case, the resulting $I\Delta\Delta G$ plots always have branches. Thus accordingly, fraction of mismatches higher than the threshold are always found in each iteration. In this Figure, results from analysing t_1 (**a**), t_2 (**b**) and t_3 (**c**) are shown. Notice that the next hypothesis generated from t_3, after implementing the nucleotide changes at position no. 12 and 24, is identical to t_2. These sequences are shown in Figure 6. Therefore, the algorithm converges into two-cycle state and concludes that the sample in this current case is a mixed sequences sample.

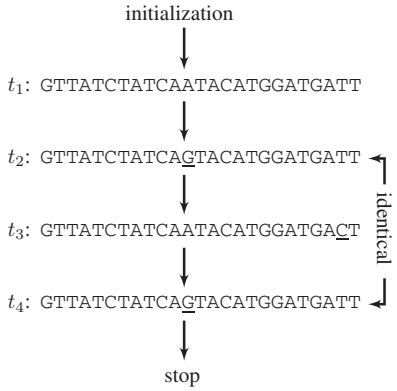

Figure 6. The hypothesis sequences from each iteration of the algorithm in the case of sample no. 3. It presents a conclusion that the sample contains mixed sequences as the generated hypothesis t_4 after t_3 is identical to previously generated hypothesis t_2. This is a two-cycle state between $t_2 = t_4 = \ldots$ and $t_3 = t_5 = \ldots$.

compared with the case in Figure 3c, which is a unique sequence sample.

We comment on the differences between the quality of the collapses between Figure 1 and Figures 3 and 5. Figure 1 reports calibration experiments using a single synthetic target sequence and probes carrying up to two mismatches (15). In Figures 3 and 5, corresponding to hybridization experiments on HIV samples, the probes can carry more mismatches. The intensity drops to a constant background level for $-\Delta\Delta G < -10$ kcal/mol. Low intensity data have not been considered in the analysis as the algorithm selects the most deviating branch of points in a range $I > 10^2$. The HIV data are typically more scattered than those from the synthetic oligo, which is probably due to more involved sample preparation where there is still room for further improvement.

The microarray probes were designed to detect the top 100 sequences with highest clinical frequency, and mixtures thereof. The seven samples presented in this article all fall within this scope, and our algorithm produces the correct sample composition. A test based on our algorithm to reject sequences that are outside a diagnostic scope is further discussed in the Supplementary Section I.

Resolving ambiguity from two degenerate bases

To discuss further about the result on mixed samples, we focused on degenerate nucleotides in the sequences due to uncertainties that are inherent in the Sanger method; as can be seen in the list of sequences on the left part of Table 3. The letters R and Y denote the degenerate bases A or G (purines) and C or T (pyrimidines), respectively (25). These uncertainties are caused by the presence of mixtures of two sequences in a given clinical sample. In the case of a single degenerate nucleotide, the sample is identified as a unique mixture. However, two different type of mixtures are possible in the case of two degenerate nucleotides in the same sequence. For instance a degenerate RR pair can mean either a AA/GG mixture or a AG/GA mixture. This type of ambiguity is present in the sequences no. 1, 3, 5 and 7 of Table 3. The analysis of the hybridization data from the microarray experiments allow to resolve this ambiguity because the hybridization free energies for each case are different.

The importance of this information lies in its clinical relevance, for example, in sample no. 5 where mutations RTR occur in codon 184. The presence of ATA as suggested by our method, gives an idea about the stage of resistance. This is interesting because during treatment with lamivudine, initially isoleucine mutants are present, which are subsequently replaced by valine variants (27). The isoleucine mutants are less fit than the valine mutants (28).

CONCLUSIONS

In this article, we used inputs from hybridization thermodynamics to perform targeted resequencing of a fragment of the HIV-1 RT gene. In the HIV example considered here, due to the high mutation rate of the virus, the fragment analysed can occur in more than a hundred different variants. In addition, one needs to distinguish between samples composed by a single sequence (unique) from samples in which two or more different sequences coexist (mixed). For clinical purposes, it is important to identify early enough the rising of a resistant strain. Our microarray analysis is based on a large number of measured intensities not just from perfectly matching probes but also mismatching probes. Although in general the effect of the hybridization of a very low abundance sequence in a mixed sample can be small for each individual intensities, the correlated effect on a large number of different probes can be detected even from a target sequence at relatively low abundance. The analysis of (14), in which artificial synthetic mixtures of two sequences were used, indicated that the detection limit of relative abundance is of about 1%. In the Supplementary Section III, we repeat that analysis for the HIV data to illustrate the potential of the thermodynamic approach to microarray data analysis.

Here, we presented an iterative algorithm that successively generates *in silico* hypotheses for the sample composition and checked them against thermodynamic models. The algorithm identified correctly the sequences in the sample and was tested on seven clinical samples from the Janssen Diagnostics database. We showed how hybridization thermodynamics can resolve some intrinsic ambiguities of the Sanger sequencing. The results show the reliability of DNA microarrays and in principle any hybridization-based technology.

A method based on hybridization has the advantage that it is a simple test that can be miniaturized to a fully automated lab-on-chip, which can be used as point-of-care test. Indeed hybridization-based method is promising, and it has been a focus of interest in the recent years on either its fundamentals or applications (29–33). Although DNA microarrays are considered nowadays a mature technology and these devices are providing high quality and reproducibile data, their potentials have not been fully exploited. A better understanding of the underlying

physico-chemical principles of hybridization is an issue of central interest and can lead to novel methods to improve the data analysis (34).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [35].

ACKNOWLEDGEMENTS

We thank Kim Steegen and David Nauwelaars for initial work on HIV resistance alignments and codon selection.

FUNDING

KULeuven [STRT1/09/042]; VITO [ZL39010200-401]. Funding for open access charge: VITO and Janssen Diagnostics.

Conflict of interest statement. None declared.

REFERENCES

1. Gaudet,M., Fara,A.-G., Beritognolo,I. and Sabatti,M. (2009) Allele-Specific PCR in SNP Genotyping. *Methods Mol. Biol.*, **578**, 415–424.
2. Tsiatis,A.C., Norris-Kirby,A., Rich,R.G., Hafez,M.J., Gocke,C.D., Eshleman,J.R. and Murphy,K.M. (2010) Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. *J. Mol. Diagn.*, **12**, 425–432.
3. Anastassopoulou,C.G. and Kostrikis,L.G. (2006) Global genetic variation of HIV-1 infection. *Curr. HIV Res.*, **4**, 365–373.
4. Hirsch,M.S., Guenthard,H.F., Schapiro,J.M., Brun-Vezinet,F., Clotet,B., Hammer,S.M., Johnson,V.A., Kuritzkes,D.R., Mellors,J.W., Pillay,D. *et al.* (2008) Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an international AIDS society-USA panel. *Clin. Infect. Dis.*, **47**, 266–285.
5. Van Houtte,M., Picchio,G., Van der Borght,K., Pattery,T., Lecocq,P. and Bacheler,L.T. (2009) A comparison of HIV-1 drug susceptibility as provided by conventional phenotyping and by a phenotype prediction tool based on viral genotype. *J. Med. Virol.*, **81**, 1702–1709.
6. Broder,S. (2010) The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. *Antiviral Res.*, **85**, 1–18.
7. Dramanac,R., Labat,I., Brukner,I. and Crkvenjakov,R. (1989) Sequencing of megabase plus DNA by hybridization: Theory of the method. *Genomics*, **4**, 114–128.
8. Ginot,F. (1997) Oligonucleotide micro-arrays for identification of unknown mutations: how far from reality? *Hum. Mutat.*, **10**, 1–10.
9. Shendure,J., Mitra,R.D., Varma,C. and Church,G.M. (2004) Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.*, **5**, 33 5–344.
10. Hurd,P.J. and Nelson,C.J. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief. Funct. Genomic. Proteomic.*, **8**, 174–183.
11. Lizardi,P.M. (2008) Next-generation sequencing-by-hybridization. *Nat. Biotechnol.*, **26**, 649–650.
12. Pihlak,A., Bauren,G., Hersoug,E., Lonnerberg,P., Metsis,A. and Linnarsson,S. (2008) Rapid genome sequencing with short universal tiling probes,' *Nat. Biotechnol.*, **26**, 676–684.
13. Hooyberghs,J., Baiesi,M., Ferrantini,A. and Carlon,E. (2010) Breakdown of thermodynamic equilibrium for DNA hybridization in microarrays, *Phys. Rev. E*, **81**, 012901.
14. Hooyberghs,J. and Carlon,E. (2010) Hybridisation thermodynamic parameters allow accurate detection of point mutations with DNA microarrays. *Biosens. Bioelectron.*, **26**, 1692–1695.
15. Hadiwikarta,W.W., Walter,J.C., Hooyberghs,J. and Carlon,E. (2012) Probing hybridization parameters from microarray experiments: nearest-neighbor model and beyond. *Nucleic Acids Res.*, **40**, e138.
16. Tombuyzer,L., Azijn,H., Rimsky,L.T., Vingerhoets,J., Lecocq,P., Kraus,G., Picchio,G. and de Bethune,M.P. (2009) Compilation and prevalence of mutations associated with resistance to non-nucleoside reverse transcriptase inhibitors. *Antivir. Ther.*, **14**, 103–109.
17. De Clercq,E. (1998) The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. *Antiviral Res.*, **38**, 153–179.
18. Das,K., Ding,J.P., Hsiou,Y., Clark,A.D., Moereels,H., Koymans,L., Andries,K., Pauwels,R., Janssen,P.A.J., Boyer,P.L. *et al.* (1996) Crystal structures of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant. *J. Mol. Biol.*, **264**, 1085–1100.
19. Ren,J. and Stammers,D.K. (2008) Structural basis for drug resistance mechanisms for non-nucleoside inhibitors of HIV reverse transcriptase. *Virus Res.*, **134**, 157–170.
20. Basson,A.E., Rhee,S.Y., Parry,C., de Oliveira,T., Pillay,D., Shafer,R. and Morris,L. (2011) Phenotypic resistance to etravirine in an HIV-1 subtype C background. *Antivir. Ther.*, **16**, A46–A46.
21. Sarafianos,S.G., Das,K., Clark,A.D., Ding,J.P., Boyer,P.L., Hughes,S.H. and Arnold,E. (1999) Lamivudine (3TC) resistance in HIV-1 reverse transcriptase involves steric hindrance with beta-branched amino acids,' *Proc. Natl Acad. Sci. USA*, **96**, 10027–10032.
22. Gao,H.Q., Boyer,P.L., Sarafianos,S.G., Arnold,E. and Hughes,S.H. (2000) The role of steric hindrance in 3TC resistance of human immunodeficiency virus type-1 reverse transcriptase. *J. Mol. Biol.*, **300**, 403–418.
23. Pattery,T., Verlinden,Y., De Wolf,H., Nauwelaers,D., Van Baelen,K., Van Houtte,M., Mc Kenna,P. and Villacian,J. (2012) Development and performance of conventional HIV-1 Phenotyping [Antivirogram (R)] and genotype based calculated phenotyping assay (virco (R) TYPE HIV-1) on protease and reverse transcriptase genes to evaluate drug resistance. *Intervirol.*, **55**, 138–146.
24. Hooyberghs,J., Van Hummelen,P. and Carlon,E. (2009) The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. *Nucleic Acids Res.*, **37**, e53.
25. 'Nomenclature Committee of the International Union of Biochemistry (NC-IUB). (1985) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Eur. J. Biochem.*, **150**, 1–5.
26. SantaLucia,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
27. Schuurman,R., Nijhuis,M., van Leeuwen,R., Schipper,P., de Jong,D., Collis,P., Danner,S.A., Mulder,J., Loveday,C., Christopherson,C. *et al.* (1995) Rapid changes in human immunodeficiency virus type 1 RNA load and appearance of drug-resistant virus populations in persons treated with lamivudine (3TC). *J. Infect. Dis.*, **171**, 1411–1419.
28. Larder,B.A., Kemp,S.D. and Harrigan,P.R. (1995) Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy. *Science*, **269**, 696–699.
29. Fuchs,J., Fiche,J.B., Buhot,A., Calemczuk,R. and Livache,T. (2010) Salt concentration effects on equilibrium melting curves from DNA microarrays. *Biophys. J.*, **99**, 1886–1895.
30. Burden,C.J. and Binder,H. (2010) Physico-chemical modelling of target depletion during hybridization on oligonulceotide microarrays. *Phys. Biol.*, **7**, 016004.
31. Irving,D., Gong,P. and Levicky,R. (2010) DNA surface hybridization: comparison of theory and experiment. *J. Phys. Chem. B*, **114**, 7631–7640.

32. Leinberger,D.M., Grimm,V., Rubtsova,M., Weile,J., Schröppel,K., Wichelhaus,T.A., Knabbe,C., Schmid,R.D. and Bachmann,T.T. (2010) Integrated detection of extended-spectrum-beta-lactam resistance by DNA microarray-based genotyping of TEM, SHV, and CTX-M genes,. *J. Clin. Microbiol.*, **48**, 460–471.
33. van Grinsven,B., Vanden Bon,N., Grieten,L., Murib,M., Janssens,S.D., Haenen,K., Schneider,E., Ingebrandt,S., Schöning,M.J., Vermeeren,V. *et al.* (2011) Rapid assessment of the stability of DNA duplexes by impedimetric real-time monitoring of chemically induced denaturation. *Lab. Chip*, **11**, 1656–1663.
34. Harrison,A., Binder,H., Buhot,A., Burden,C.J., Carlon,E., Gibas,C., Gamble,L.J., Halperin,A., Hooyberghs,J., Kreil,D.P. *et al.* (2013) Physico-chemical foundations underpinning microarray and next-generation sequencing experiments. *Nucleic Acids Res.*, **41**, 2779–2796.
35. Mullins,J.I., Heath,L., Hughes,J.P., Kicha,J., Styrchak,S., Wong,K.G., Rao,U., Hansen,A., Harris,K.S., Laurent,J.P. *et al.* (2011) Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside KP1461. *PLoS One*, **6**, e15135.