Published online 13 March 2012

Nucleic Acids Research, 2012, Vol. 40, No. 11 ***e87***
doi:10.1093/nar/gks230

Sequence dependence of isothermal DNA amplification via EXPAR

Jifeng Qian[1], Tanya M. Ferguson[1], Deepali N. Shinde[1], Alissa J. Ramírez-Borrero[1], Arend Hintze[2], Christoph Adami[1,2] and Angelika Niemz[1,*]

[1]Keck Graduate Institute, Claremont, 535 Watson Drive, Claremont, CA 91711, USA and [2]Michigan State University, East Lansing, MI 48824, USA

Received November 23, 2011; Revised February 21, 2012; Accepted February 22, 2012

ABSTRACT

Isothermal nucleic acid amplification is becoming increasingly important for molecular diagnostics. Therefore, new computational tools are needed to facilitate assay design. In the isothermal EXPonential Amplification Reaction (EXPAR), template sequences with similar thermodynamic characteristics perform very differently. To understand what causes this variability, we characterized the performance of 384 template sequences, and used this data to develop two computational methods to predict EXPAR template performance based on sequence: a position weight matrix approach with support vector machine classifier, and RELIEF attribute evaluation with Naïve Bayes classification. The methods identified well and poorly performing EXPAR templates with 67–70% sensitivity and 77–80% specificity. We combined these methods into a computational tool that can accelerate new assay design by ruling out likely poor performers. Furthermore, our data suggest that variability in template performance is linked to specific sequence motifs. Cytidine, a pyrimidine base, is over-represented in certain positions of well-performing templates. Guanosine and adenosine, both purine bases, are over-represented in similar regions of poorly performing templates, frequently as GA or AG dimers. Since polymerases have a higher affinity for purine oligonucleotides, polymerase binding to GA-rich regions of a single-stranded DNA template may promote non-specific amplification in EXPAR and other nucleic acid amplification reactions.

INTRODUCTION

The number of reported isothermal nucleic-acid amplification methods is rapidly growing (1–14). Isothermal nucleic-acid amplification permits less complex and less expensive instrumentation resulting in a significant advantage for low cost point-of-care diagnostic applications (14). Unfortunately, assay design for isothermal DNA amplification is, in most cases, more complex than for the Polymerase Chain Reaction (PCR), and new computational design tools are needed to facilitate assay development. Existing computational design tools for PCR (15,16) and isothermal methods (6,17) focus mainly on the thermodynamics of nucleic-acid hybridization, calculated through nearest neighbor interaction energies (18,19). Other considerations include the length of the primer (20); GC content (21,22); the propensity of the primer to form internal secondary structure (23); and the propensity of a pair of primers to form primer dimers, especially with overlapping 3′ ends (24).

Assay performance may be predicted more accurately by also considering interactions between the polymerase and its template. DNA polymerases interact with the DNA-phosphate backbone through electrostatic interactions and hydrogen bonding (25–27), and with the nucleoside bases of the DNA template through hydrophobic interactions (25). Studies have shown an increase in the affinity between DNA polymerase and homo-oligonucleotides in the order $d(pC)n < d(pT)n < d(pG)n \approx d(pA)n$, which may be a reflection of the relative hydrophobicity of the nucleotides $C < T < G \approx A$ (28). The rate at which DNA polymerases incorporate nucleotides into the elongating strand also depends on the sequence of the DNA template (29). These observations support the view that polymerases interact with DNA in a sequence-dependent manner. Capturing protein–DNA interactions requires computational

*To whom correspondence should be addressed. Tel: +1 909 607 9854; Fax: +1 909 607 9826; Email: aniemz@kgi.edu

methods that go beyond standard thermodynamic characterization of DNA–DNA interactions. Here, we use two complementary methods to quantify the sequence dependence of the polymerase–DNA template interaction: a position weight matrix (PWM) approach and a Naïve Bayes machine learning technique. The PWM approach is one of the leading methods used to identify DNA sequences recognized by certain proteins, such as transcription factor binding sites (30,31), splice sites, and translational start sites (32). The standard PWM approach relies on the assumption that each position in the DNA sequence contributes independently to protein binding (33), which may lead to erroneous classification of sequences with correlations between nucleotides within the binding region. However, recent improvements to the PWM approach mitigate this concern (34). As an alternate approach, machine-learning algorithms such as the Naïve Bayes method classify data based on conditional probability. Naïve Bayes machine learning is widely used in text (35) and graph classification (36). In contrast to the standard PWM approach, this method can take into account DNA motifs consisting of multiple bases.

The studies described in this report use the Exponential Amplification Reaction (EXPAR, Figure 1), an isothermal amplification method, which efficiently amplifies short oligonucleotides at 55°C (11,12,37). The short trigger oligonucleotides that initiate EXPAR, called trigger X, can be enzymatically generated from specific sites within the targeted genomic DNA (12), and therefore represent the analyte. Exponential amplification of trigger X is facilitated by an EXPAR amplification template oligonucleotide provided in excess in the reaction. The EXPAR template contains two copies of the trigger reverse complement X′, separated by the reverse complement of a nicking enzyme recognition site plus a required post-cut site spacer. Trigger X primes the template and is extended by a polymerase, which generates a double-stranded 5′-GAGTC-3′ on the top strand that is recognized by the nicking enzyme Nt.BstNBI. The nicking enzyme nicks the top strand four bases to the 3′ end of its recognition sequence. This creates another copy of the oligonucleotide trigger X that melts off or is displaced from the amplification template. The polymerase elongates the recessed 3′-hydroxyl, created by the departing trigger, and the process repeats. Newly formed triggers then prime other amplification templates, creating true chain (exponential) reactions. During EXPAR, templates with similar thermodynamic characteristics often exhibit very different trigger amplification rates. This suggests that efficient EXPAR amplification depends in part on the template sequence.

Most nucleic-acid amplification reactions exhibit non-specific amplification in the absence of the targeted sequence, which limits the attainable assay sensitivity. Non-specific amplification is often due to mis-priming (38) or primer–dimer formation (39). Furthermore, thermophilic polymerases can carry out *ab initio* DNA synthesis in the absence of templating or priming DNA strands (40–44). This *ab initio* DNA synthesis is accelerated in the presence of restriction endonucleases (45) or nicking endonucleases (46,47). We have observed a late-stage amplification phenomenon under the EXPAR

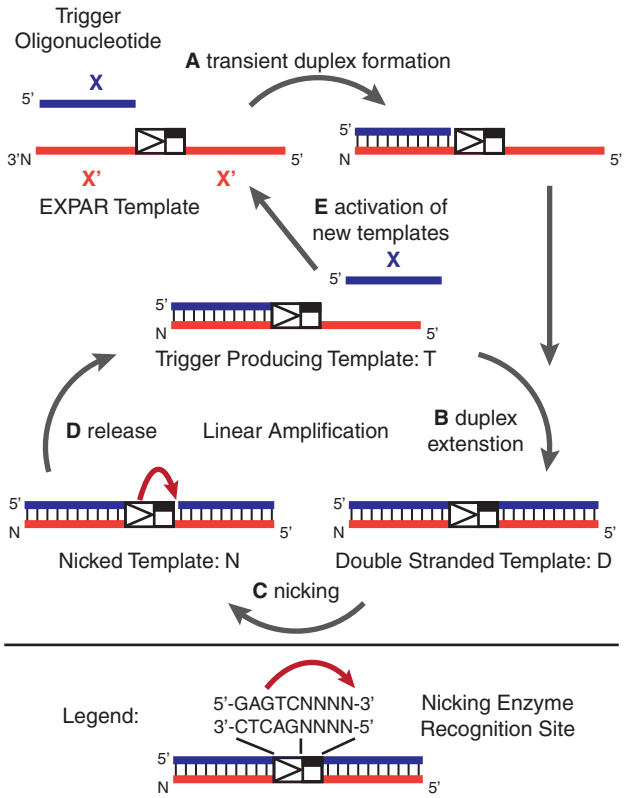

Figure 1. Overview of the EXPAR: (**A**) Trigger X transiently binds to the complementary recognition sequence at the 3′-end of the amplification template; (**B**) the trigger sequence is extended by the DNA polymerase, forming the double-stranded nicking enzyme recognition site 5′-GAGTCNNNN-3′ on the top strand; (**C**) the top strand is cleaved through the nicking endonuclease Nt.BstNBI; (**D**) at the temperature of the reaction (55°C), the newly formed trigger is released from the amplification template. The trigger-producing form of the amplification template re-enters the linear amplification cycle, and new trigger oligonucleotides are generated through duplex extension, nicking, and release; (**E**) the newly formed trigger oligonucleotides activate additional template sequences, giving rise to exponential amplification of the trigger X.

reaction conditions consistent with *ab initio*, untemplated and unprimed DNA synthesis (47). However, EXPAR also exhibits early phase non-specific background amplification, which is observed only in the presence of an EXPAR template and which generates the trigger sequence specific for the template present in the reaction (47). Although contamination with other oligonucleotides or primer-dimer formation may play some role in this process and cannot be ruled out with absolute certainty, control experiments indicate that these are not the main causes of the observed phenomenon (47). This early-phase non-specific background amplification may involve a novel and unconventional polymerase activity, which becomes noticeable in EXPAR due to the positive feedback loop present in the reaction. Elucidating which types of sequences or sequence motifs are over-represented in templates particularly prone to early-phase non-specific background amplification may provide clues about the underlying reaction mechanism.

We report the experimental characterization of over 300 different EXPAR templates that were generated in

accordance with a set of design rules based on thermodynamic parameters and secondary structure analysis. Within these EXPAR template sequences, we observed considerable variation in the efficiency of specific amplification in the presence of trigger X, and in the propensity for early phase non-specific background amplification in the absence of trigger X. We have applied a PWM approach and a Naïve Bayes machine learning algorithm to identify sequence motifs that influence template dependent specific versus non-specific amplification, and to enable EXPAR template classification based on sequence. We have combined these different approaches into a computational tool called EXPAR Template Sequence Analysis tool (ETSeq), for EXPAR template selection based on thermodynamic and sequence dependence criteria, and have tested the performance of this tool on *de novo* designed sequences. ETSeq facilitates EXPAR assay design by increasing the likelihood that a chosen template sequence will perform well in the assay. Furthermore, the results may provide a better fundamental understanding of the underlying reaction mechanism for the observed non-specific amplification phenomenon.

MATERIALS AND METHODS

EXPAR template design—thermodynamic criteria

Out of the 384 template sequences used in this study (Supplementary Table S1), sequences #1–64 had previously been characterized in our laboratory, whereas sequences #65–384 were newly designed. Design of sequences #2–11 was based on sequence #1 (Figure 2) described in an earlier publication (11) with a single nucleotide base change each in the trigger complement positions 1–10. For all remaining 373 sequences, we randomly assigned A, T, C, or G with equal probability to each of the 14 variable positions that make up the trigger complement and nicking enzyme post-cut site (Figure 2; positions 1–14, with 1–10 replicated at 21–30). An additional 'T' was added to each template at the 3′ end of the nicking enzyme recognition site (position 20), because we earlier hypothesized that the polymerase may append an A to newly generated triggers through non-templated adenylation. Although we found this not to be the case under the current EXPAR conditions, based on mass spectrometric analysis of the generated trigger, we kept the extra T in the template sequences for consistency.

Template sequences newly designed for this study were selected to have a calculated template/trigger duplex melting temperature (T_M) between 35 and 55°C; a calculated T_M for template/template self-hybridization less than 25°C; and fewer than 10 predicted secondary structure bonds for the template–template interaction. In EXPAR, the trigger-template melting temperature is at or below the reaction temperature (11). Therefore, the incoming trigger binds transiently to the template, but the interaction becomes stabilized upon polymerase extension. The newly formed trigger, however, readily dissociates from the template, which facilitates rapid amplification. In previous studies (data not shown), we found that if the trigger-template melting temperature is lower than 35°C, trigger-template binding is too weak to initiate the reaction. If the trigger-template T_M is much higher than 55°C, newly formed triggers will not readily dissociate from the template, which impedes the linear amplification cycle. In addition, templates with extensive secondary structure tend to amplify slowly or not at all. These design rules based on thermodynamic parameters and secondary structure analysis provide some level of consistency between the different templates. We calculated the T_M values for template/trigger and template/template hybridization and determined the number of bonds involved in template–template self-hybridization via the Zuker–Turner algorithm, through the UNAFold application (48), (see Supplementary Methods section of Supplementary Data). Out of the 64 previously characterized sequences, nine did not fit our thermodynamic acceptance criteria. Two sequences had a trigger template T_M around 29°C, i.e. lower than the 35°C cutoff. Both sequences did not amplify, which supports the validity of our selected trigger template T_M cutoff value. Seven sequences did not satisfy the template self-hybridization criteria, and had a template–template T_M higher than 25°C. All sequences with template–template T_M larger than 45°C did not amplify, and the remaining sequences had poor or intermediate performance. This observation again supports the validity of our chosen selection criteria. Out of the nine sequences that did not fit our thermodynamic acceptance criteria, only four sequences with a template–template T_M between 26 and 39°C remained in our data set after data analysis using further exclusion criteria.

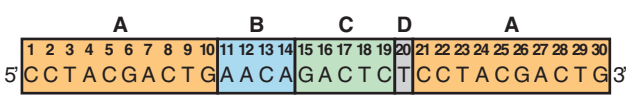

Figure 2. Example template sequence #1 (11) consisting of (**A**) the trigger reverse complement X′, which is replicated at positions 1–10 and 21–30, (**B**) the nicking enzyme post-cut site, (**C**) the reverse complement of the nicking enzyme recognition site, (**D**) an additional T at the end of the trigger-binding site.

EXPAR performance screening

All EXPAR template sequences used in this study were ordered from Integrated DNA Technologies, Inc. (Coralville, IA) in standard uncapped desalted form for cost reasons. Although EXPAR reactions using 3′-amine capped and HPLC purified amplification templates perform more consistently, we previously demonstrated that acceleration of non-specific background amplification using uncapped versus capped templates was small and not statistically significant (47), therefore use of uncapped templates is not expected to skew the results of this study. Corresponding trigger sequences were obtained from Eurofins MWG Operon (Huntsville, AL). EXPAR was performed in a 30 µl reaction mixture containing 0.2 units/µl Nt.BstNBI nicking enzyme (New England Biolabs, Ipswich, MA), 0.03 U/µl *Bst* DNA Polymerase (NEB), 0.24 mM of each dNTP (Fermentas, Glen Burnie, MD), 3 mM $MgCl_2$ (Sigma-Aldrich, St. Louis, MO), 1× Sybr Green I (Invitrogen), 20 mM

Tris–HCl pH 7.9, 15 mM ammonium sulfate, 30 mM KCl, 0.005% Triton X-100, and 50 nM EXPAR template oligonucleotide. Positive trigger-containing reactions included 1 pM trigger oligonucleotide, which was left out of the No Trigger Control (NTC) reactions. The EXPAR mastermix and template/trigger dilutions were prepared manually. We then used a Beckman Coulter Biomek FX Liquid Handling System to combine the EXPAR mastermix with the template only or template-trigger mixes, and to transfer the final reaction mixture into a 96-well PCR plate, with random distribution to mitigate confounding effects. All reaction components were kept on cold blocks during the manual and automated steps of reaction setup. The final plate was then heated at a constant temperature of 55°C for 50 min inside a Bio-Rad Opticon real time thermocycler, and the fluorescence was monitored using 488 nm excitation. For each template sequence, we acquired six replicates respectively for the positive and negative reactions, with two sets of three replicates acquired in separate experiments performed on two different days. For some templates with ambiguous data, three additional replicates were acquired in a third experiment.

Data analysis and classification

To analyze the EXPAR real-time fluorescence amplification curves, we developed a MATLAB program called EXPAR Data Analysis Tool (EDAT), which performs a nonlinear least-squares curve fit of the real-time fluorescence data to a sigmoidal curve (see Supplementary Methods section of Supplementary Data). The sigmoidal fit enables mathematical parameterization of the data, which is required for machine learning-based classification. For each curve, EDAT determines and records the time required for the fitted data to reach 10 and 90% of the maximum plateau. These time points were designated as P10 and P90 for positive, trigger containing reactions, and N10 and N90 for negative no trigger controls. The time difference between positive, specific amplification and negative, non-specific amplification, was calculated as Diff = N10–P90. The program also records the quality of data fitting as the normalized residual, and we checked each curve fit manually to ensure that the experimental data was suitably fitted by the program. Curves with no amplification, very late amplification, and with non-sigmoidal amplification behavior (Supplementary Figure S1) were considered to be 'unfitable', and were excluded from further analysis. Curves that amplified in a sigmoidal fashion to nearly the final plateau, and curves wherein the final plateu either drifted up or down in intensity were partially fitted as described under Supplementary Data. For these curves, we only considered the 10% intensity to be reliably determined. From the data set of evaluated curves, we removed outliers, defined as individual values within a set of P90 or N10 replicates for a given template sequence more than 1.5 times the interquartile range below the first quartile or above the third quartile of the replicate set. Only template sequences with at least three measurable P90 and N10 values were used for further analysis. For these template sequences, we calculated the mean and standard deviation for P90, N10, and Diff over all plates. The standard deviation for Diff was calculated from the variances of P90 and N10, based on standard error propagation. We further excluded any templates whose Diff or P90 standard deviation was larger than 30% of the mean. Following these exclusion criteria, we obtained data of suitable quality for 307 template sequences. Out of the 77 omitted template sequences, 49 were omitted because of late amplification, no amplification or non-sigmoidal behavior. These templates could have been assigned to the 'poorly performing' class, but as we could not derive reliable P90 and Diff values, we could not use these templates in our analysis. Another 17 templates were omitted because the final plateau drifted either up or down for a significant number of the replicates. Although we partially fitted this data by cutting off the sloping section of the plateau, we did not consider the data reliable for determining P90. We repeated the analysis with this data included, but observed no significant differences in the results. Only 11 template sequences were omitted purely due to a large standard deviation, which in most cases was caused by significant plate-to-plate differences. Such experimental variations can occur, especially for a large screening study such as this.

The 307 sequences used for further analysis were divided into three types: 102 Class I templates (well performing), 102 Class II templates (poorly performing) and 103 Class III templates (with intermediate performance). Class I templates were defined as having a Diff value larger than one minute and a Diff $> 0.85 \cdot (\text{P90}) - 15$ min. Class II templates had a Diff value smaller than negative 1.3 min and a Diff $< 0.85 \cdot (\text{P90}) - 15$ min. The remaining templates were classified as Class III.

PWM

Using a 10-fold cross-validation approach, the combined set of 204 Class I and II templates, called here the total data set, was randomly divided into 10 parts, of which one part (20 or 21 sequences) was left out as a test set. The remaining 183 or 184 sequences were used as the training set. From the training set, we generated a PWM based on templates whose P90 was in the lowest 30% of the P90 value range. A second PWM was generated using templates in the training set whose Diff was in the highest 40% of the Diff value range. These two PWMs were then applied to calculate scores for all template sequences within the training set. Further details of this method are provided in the Supplementary Methods section of Supplementary Data. Using a support vector machine (SVM) algorithm (49) with sequential minimal optimization, we obtained a linear classification boundary in the two-dimensional space of P90 and Diff scores that optimally separates the two classes. The two PWMs were then applied to the template sequences in the test set to calculate their P90 and Diff scores. Based on these scores and the classification boundary, the test set sequences were classified as either Class I or Class II. This procedure was performed 10 times, each time using a different 10th part of the total data set as test set. The

entire procedure was repeated using a shuffled data set, in which the 204 sequences were randomly assigned one of the 204 performance metrics (P90 and Diff values).

Naïve Bayes classification approach

In order to predict EXPAR template performance using the Naïve Bayes machine learning approach, we generated a position motif count matrix based on sequence motifs present in the 204 Class I and Class II EXPAR templates. As before, Class III templates were excluded from the analysis. Position motifs represent subsequences composed of all possible combinations of the four bases, up to four nucleotides long, starting at defined positions within the EXPAR template. For example, motif 'ATG-7' represents the 3mer subsequence 'ATG' starting at the 7th base pair of the templates. Since the trigger complement sequence in positions 1–10 is repeated in positions 21–30, and since positions 15–20 are conserved, only motifs starting in positions 1–14 are truly unique. Motifs up to four bases in length starting at positions 15–17 fall entirely within the conserved sequence region, while certain motifs starting at positions 18–20 are correlated with and occur at the same frequency as 1–3mer motifs starting at position 1; therefore, they encode no additional information.

Out of the 4760 possible position motifs $[14 \cdot (4 + 4^2 + 4^3 + 4^4) = 4760]$, we only considered in our analysis the 957 motifs that occurred at least three times in the entire data set. To capture which motifs occur in which template, we generated the motif count matrix $A(i,j)$, where i is the template index ($1 < i < 204$), and j is the motif index ($1 < j < 957$). If template i contains the position motif j, then $A(i,j) = 1$, otherwise $A(i,j) = 0$. This motif count matrix was used as input for the machine learning procedure, which was executed using the open source software suite WEKA (50). Just as for the PWM approach, we used a 10-fold cross-validation approach, wherein the whole data set of 204 sequences was randomly partitioned into 10 subsets. Each subset was used as test set once, while the other nine subsets were used as training data set. In the training process, important features were selected from the training set using RELIEF attribute evaluation (51). Then, Naïve Bayes machine learning (52) was applied to obtain classifiers for the selected features. The features and classifiers were then used to predict which sequences within the test set fall into Class I versus Class II. This procedure was performed 10 times, each time using a different tenth part of the total data set as test set. The entire procedure was repeated using a shuffled data set, in which the 204 sequences were randomly assigned one of the 204 performance metrics (P90 and Diff values).

Combined computational tool

We have combined the methods for EXPAR template performance prediction based on thermodynamic and sequence-related criteria into a python tool with graphical user interface, called EXPAR Template Sequence analysis tool (ETSeq), which can be downloaded from https://github.com/expartools/ETSeq/wiki. This tool requires as input a set of user defined EXPAR template sequences to be analyzed. The program UNAFold (48) is used to predict parameters related to template trigger binding and template self-hybridization. ETSeq allows users to change thermodynamic selection criteria from the default values if needed. The tool performs sequence dependence classification using both the PWM and Naive Bayes approaches. For the PWM approach, we have used all 204 Class I and Class II templates to generate PWMs for P90 and Diff, and one SVM-based boundary line for classification. The PWMs and boundary line were then embedded in the tool. All submitted sequences that according to their calculated P90 and Diff scores fall below the boundary line are predicted to be Class I (well performing). All other sequences are categorized as Class II (poorly performing). For the Naïve Bayes approach, we provided the model with selected significant position motifs identified in this study. All 204 Class I and Class II templates were then used to generate one Naïve Bayes prediction model using a module from the python-based software suite Orange (http://orange.biolab.si) (53). We used the Orange module since the WEKA tool could not be readily incorporated into this program. For each submitted sequence, if the calculated probability for the sequence to belong to Class I is higher than the probability to belong to Class II, then that sequence is predicted to be Class I, and vice versa. The tool generates an output excel file containing one sheet for just the sequences predicted to be well-performing based on both sequence dependence prediction methods, one sheet containing all sequences that have passed the thermodynamic selection criteria, and a third sheet containing all submitted sequences. Each sheet lists the predicted trigger-template T_M, template–template T_M, the predicted number of bonds for template self-hybridization, the P90 score, Diff score, and the classification results for the PWM and Naïve Bayes methods.

Performance verification of the computational tool

Using the program ETSeq, we analyzed 100 000 newly designed EXPAR template sequences, wherein the variable positions 1–14 (Figure 2) were randomly populated with A, T, C and G with equal probability. Based on the previously discussed thermodynamic selection criteria, 45 134 sequences were excluded. The remaining sequences were then classified as 'well performing' or 'poorly performing', using both the PWM and Naïve Bayes approaches. We selected 23 sequences classified as well performing and seven sequences classified as poorly performing by both methods. These 30 sequences were then experimentally characterized as described under EXPAR performance screening, and under data analysis and classification.

RESULTS AND DISCUSSION

There is significant variability in the performance for different EXPAR templates. Well-performing template sequences (Figure 3A) typically show amplification of the 1 pM containing positive reaction in less than

25 min, along with significant temporal separation between the positive reaction (P) containing 1 pM trigger, and the negative reaction (N) containing no trigger. Poorly performing template sequences (Figure 3B) show slow amplification of the 1 pM containing positive reaction, and/or no significant temporal separation between the positive reaction (P) and the negative reaction (N). The amplification kinetics of positive, trigger containing reactions vary significantly even among sequences with similar hybridization thermodynamics (Supplementary Figure S3). Based on previous studies (47), non-specific background amplification in EXPAR is not primarily caused by oligonucleotide contamination or 'primer-dimer' type extension, but may involve interactions between the polymerase and single-stranded EXPAR template sequence.

Template performance characterization

In order to investigate the cause of this variability, we experimentally characterized the performance of 384 EXPAR template sequences (Supplementary Table S1), each 30 nucleotides long (Figure 2), consisting of a 10-base trigger complement region at the 3′ and 5′ ends, separated by the five-base reverse complement of the nicking enzyme recognition site and four-base spacer. In EXPAR with real time SYBR based detection (Figures 3 and 4A), activation of the amplification template is monitored. The sigmoidal curve signifies conversion from single-stranded to partially or fully double-stranded templates (Figure 1, template forms D, N and T). Once all template oligonucleotides have become activated, amplification occurs at maximum efficiency, yet the curve reaches a final plateau. Conversely, at the bottom of the sigmoidal curve, amplification has just started and very little trigger is present. In contrast, in real-time PCR using SYBR detection, the increase in double-stranded amplicon concentration is monitored.

To characterize efficient amplification, we used the time at which the amplification curve reaches 90% of the final plateau, labeled P90 for the positive and N90 for the negative reaction (Figure 4A). In order to characterize the initiation of an amplification reaction, we used the time at which the amplification curve reaches 10% of the final plateau, labeled P10 for the positive and N10 for the negative reactions. These values were derived from a normalized sigmoidal curve obtained from the experimental real-time data via non-linear least squares curve fitting. In most cases, we obtained excellent agreement between experimental and fitted data. The normalized residuals between the experimental and fitted data shown in Figure 4A were 0.03 and 0.04 for the positive and negative curves, respectively. For 95% of our data, curve fitting resulted in normalized residual less than 0.16. Curves that did not amplify, amplified late, or were non-sigmoidal (Supplementary Figure S1A–C) were characterized as 'unfitable' and were omitted from further analysis. Curves that amplified to nearly the final plateau or that did not exhibit a flat plateau (Supplementary Figure S1, D and E) were considered partially 'fitable', and only the N10 of these curves was considered reliable enough for further analysis. For positive reactions, the time of efficient amplification (P90) is the most important criterion, whereas for the negative reactions, the start of amplification (N10) is more significant. An important performance parameter in EXPAR is the temporal separation between positive, specific and negative, non-specific amplification, characterized as Diff, the time difference between P90 and N10

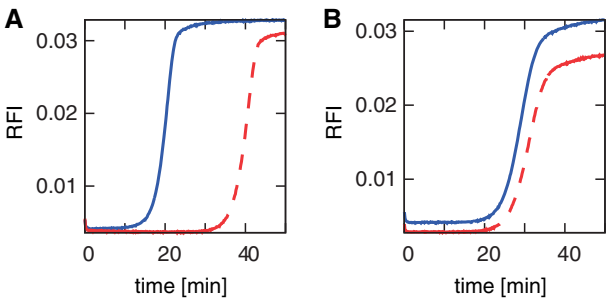

Figure 3. EXPAR amplification curves with real time fluorescence monitoring. (**A**) Example of a good performer (sequence # 356) with significant separation between sample containing 1 pM trigger (positive (P): blue solid line) and no trigger (negative (N): red dashed line). (**B**) Example of a poor performer (sequence # 85) with negligible separation between P (blue solid line) and N (red dashed line). Template sequences corresponding to these sequence numbers are listed in Supplementary Table S1.

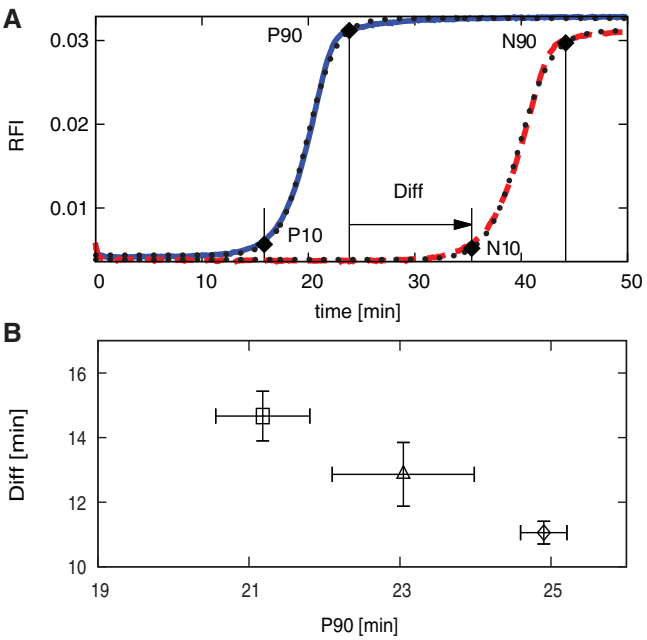

Figure 4. Data evaluation of real time amplification curves. (**A**) Example amplification curves for template sequence # 356 (positive (P): blue solid line, negative (N): red dashed line, fitted sigmoidal curves: dotted), with lines for P10, P90, N10, and N90 indicating the times at which the normalized sigmoidal curves for the positive and negative reactions reach 10 and 90% of the final plateau. The difference in time between P90 and N10 is indicated as Diff. (**B**) Means and standard errors for P90 and Diff obtained for template sequence #356. Values for three replicates from plate 1 (square), three replicates from plate 2 (diamond), and six replicates overall (triangle). The template sequence corresponding to sequence #356 is listed in Supplementary Table S1.

(Diff = N10 – P90). A plot of Diff versus P90 provides a suitable graphical representation of a template's performance. For each template sequence, we acquired two sets of three replicates in separate experiments performed on two different days. As anticipated, the intra-assay variability was smaller than the inter-assay variability (Figure 4B); however, we used the mean values over all replicates of each template for further evaluation.

Following the exclusion criteria described in the 'Materials and methods' section, we obtained suitable data for 307 template sequences, which were used for further analysis. The performance of these 307 template sequences (Figure 5) varied significantly. In the graph of Diff versus P90 (Figure 5A), well-performing templates appear at low P90 and positive Diff values, while poorly performing templates appear at high P90 and negative Diff values. In poorly performing templates with minimal or no separation between positive and negative amplification curves, the Diff value can become negative, since Diff measures the separation between P90 and N10, not between the inflection points of the positive and negative amplification reactions. For example, the sequence shown in Figure 3B has a Diff of negative 13 min. Conversely, for template sequences that show good temporal separation between specific and non-specific amplification, Diff is smaller than the actual separation between the inflection points of the positive and negative curves. Using Diff for performance characterization also penalizes templates with shallow amplification kinetics. Figure 5A and B contains the same information, but in a different graphical representation. In the graph of N10 versus P90 (Figure 5B), well-performing templates appear above the diagonal at lower P90 values, while poorly performing templates appear below the diagonal.

The observed variability in P90, Diff and N10 cannot be adequately explained based on the thermodynamics of template-trigger hybridization or template self-dimerization. We found that the template-trigger melting temperature T_M is the only parameter that shows any appreciable correlation with P90, Diff and N10, and even here the correlation is very low (Supplementary Figure S3), meaning template sequences with the same trigger-template T_M vary considerably in performance. Templates with a template-trigger $T_M < 40°C$ tend to amplify slower and have more negative Diff values, and for future assay design we have therefore increased the lower T_M cutoff from 35°C to 40°C. However, many well-performing templates also have a trigger-template $T_M < 40°C$, and T_M overall is not a good predictor of template performance.

Template sequence performance prediction

We categorized as well-performing Class I templates a set of 102 template sequences that amplify rapidly with good separation between specific and non-specific amplification (Figure 5A). We categorized as poorly performing Class II templates a set of 102 template sequences that amplify slowly and have minimal or no separation between specific and non-specific amplification. Some templates

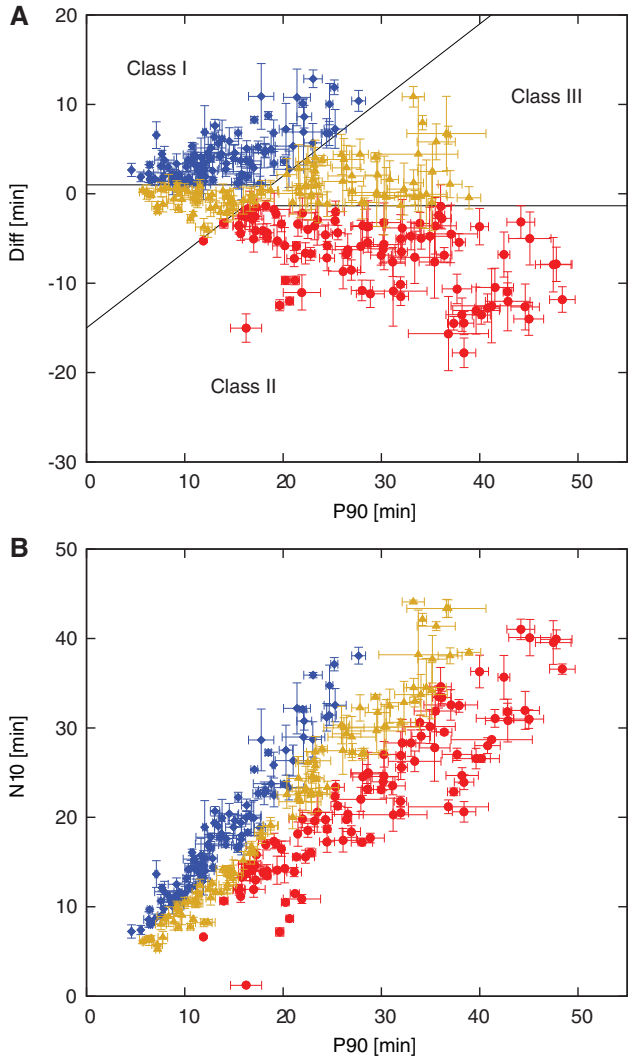

Figure 5. Results of data analysis and classification: mean values and standard errors of (**A**) Diff plotted versus P90 and (**B**) N10 plotted versus P90 for 307 template sequences that were classified into 102 Class I templates (good performers, blue diamond), 102 Class II templates (poor performers, red circle) and 103 Class III templates (intermediate performance, orange triangle).

with P90 larger than ~20 min were still considered to be good performers as long as the Diff is large, since it is possible to accelerate the amplification by changing the reaction conditions. Some templates with P90 smaller than ~20 min but very negative Diff were still considered to be poor performers, due to early onset of non-specific amplification. The remaining templates with intermediate performance are categorized as Class III. This Class includes templates that amplified rapidly but had no separation between specific and non-specific amplification, and templates that amplified slowly, but had good temporal separation between specific and non-specific amplification. To develop computational tools that can predict the performance of EXPAR template sequences and identify relevant motifs, we only considered the combined set of 204 Class I and Class II sequences. We excluded the Class III sequences with intermediate

performance while training the model, since these sequences likely combine characteristics of Class I and II sequences, cannot be categorized in a clear and experimentally meaningful manner, and may mask otherwise observable effects.

In order to predict the performance of EXPAR templates based on their sequence, we used two different computational methods: a PWM approach and Naïve Bayes machine learning (Figure 6).

We used the test set, which was never used during the training period, to measure the performance of each classifier method. The model performance therefore should reflect the performance of each classifier method when applied to new sequences. The procedure was repeated 10 times, each time using a different part as test set. Through this cross-validation approach, we can determine if results are dependent on which sequences were selected as test versus training set. Each sequence within the set of 204 Class I and Class II sequences was classified in this process. However, a slightly different model was obtained and applied for scoring during each of the 10 cycles during cross-validation, and the results presented in the following sections capture the aggregate performance of all 10 cycles.

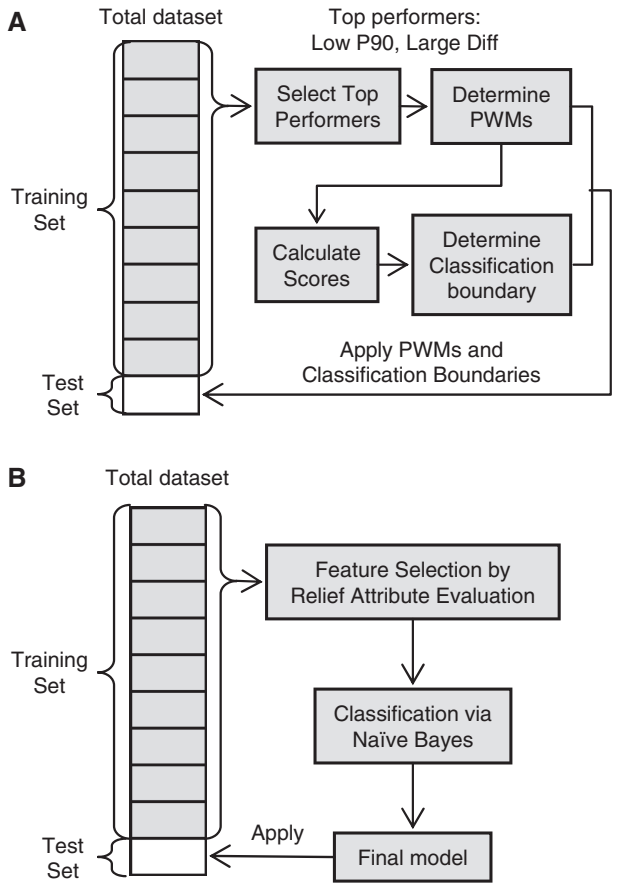

Figure 6. Procedure overview for EXPAR template classification using (**A**) a PWM approach, and (**B**) a Naïve Bayes classification method. In each case, 90% of the total data set consisting of Class I and II templates was used as training set to derive a model that was then applied to classify the templates in the remaining 10% used as test set. The process is repeated 10 times, each time using a different part of the total data set as test set.

PWM with SVM classification

Using the PWM approach, we scored the sequences based on two separate performance criteria: speed of amplification (P90) and separation between specific and non-specific amplification (Diff). Of all possible models based on combinations of PWMs derived from sequences with low versus high P90 and large positive versus large negative Diff, the best predictive power was obtained from the two PWMs that captured low P90 and large positive Diff, i.e. features present in templates with fast amplification and good separation between specific and non-specific amplification. Using these two PWMs, we obtained scores for all sequences in the training set (Figure 6A), and from these scores, we derived a linear classification boundary via a SVM algorithm (49) that optimally separates the well-performing templates from the poorly performing templates. We then applied the PWMs and classification boundary derived from the training set to categorize the test set sequences into either Class I or Class II. To check the predictive power of this approach, we also ran the same procedure using the set of 204 templates with shuffled values for P90 and Diff. If the real data set contains no information correlating template sequence with template performance, then the predictive power of the Classifier model derived from the real and shuffled data should be very similar. Note that a machine learning approach can classify even random data, as any finite data set will contain structure that can be used for discrimination. Thus, the baseline expectation for classification performance on shuffled data is not the unbiased expectation (50%). Rather, classification performance on shuffled data provides the baseline to determine whether classification on the real data set captures functional as opposed to random characteristics.

The classification performance using the PWM approach for real versus shuffled data is illustrated in Figure 7 and summarized in Table 1. Using the real data, Class I and II sequences are clearly segregated based on their PWM score, and the classification boundaries for the 10 iterations as part of cross-validation are very similar (Figure 7A). The respective PWMs for different iterations are also very similar, which supports the hypothesis that the results are largely independent of which part of the data was selected as test set versus training set. With the real data, 67.7% of the well-performing Class I templates were correctly classified as Class I (67.7% sensitivity), and 80.4% of the poorly performing Class II templates were correctly classified as Class II (80.4% specificity). Based on a Fisher's exact test, this classification is significant at a level $P = 4.4 \times 10^{-12}$ when compared to the null hypothesis of unbiased classification (50%). The method had a positive predictive value (PPV) of 78%, defined as the percentage ratio of correctly predicted Class I versus all predicted Class I sequences. The negative predictive value (NPV) was 78%, defined as the percentage ratio of correctly predicted Class II to all predicted Class II sequences. With the shuffled data,

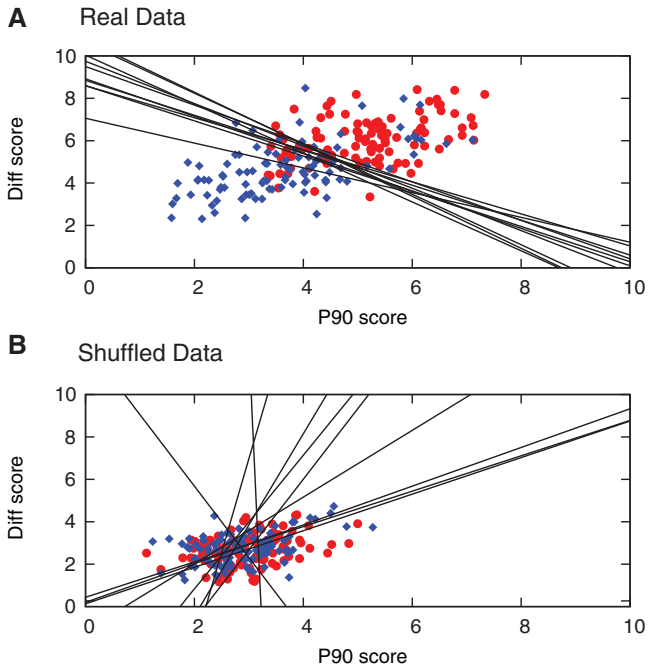

Figure 7. Categorization of Class I templates (good performers, blue diamond), and Class II templates (poor performers, red circle) using the PWM approach. Plot of PWM scores for Diff and P90 obtained for all test set sequences after 10 iterations based on (**A**) real data, and (**B**) shuffled data, along with the classification boundaries (black lines) derived in each cycle. Lower scores signify closer similarity with the characteristics on which the PWMs are based (in this case, low P90 and large positive Diff), therefore Class I templates appear in the bottom left, and Class II templates appear in the top right of the graph.

Table 1. Confusion matrix using a PWM based classification approach

		Actual Class I	Actual Class II
Real data		102	102
Predicted Class I	89	69	20
Predicted Class II	115	33	82
Shuffled data		102	102
Predicted Class I	100	59	41
Predicted Class II	104	43	61

Class I: well-performing templates. Class II: poorly performing templates. Values represent the number of template sequences in each set.

a significantly lower proportion of templates were correctly classified (Table 1). The association is still significant ($P = 0.0172$, Fisher's exact test), which implies that the PWM approach is able to extract signatures even of random data, but the significance is 10 orders of magnitude less than for the real data. As shown in Figure 7B, Class I and II sequences are not well-segregated using shuffled data based on their PWM scores for P90 and Diff. In addition, the scores across all templates are more similar for shuffled than for real data, and the classification boundaries for shuffled data are randomly orientated. The difference in results obtained using real versus shuffled data supports the conclusion that the PWM approach was able to capture specific information linking the performance of an EXPAR template to its sequence.

RELIEF attribute evaluation with Naïve Bayes classification

The Naïve Bayes machine learning approach requires as input a training set of categorized sequences, in contrast to the PWM approach that required scoring by P90 and Diff as separate parameters. In addition, machine learning methods can analyze the significance of multi-base sequence motifs. In our case, we developed a model that captures the occurrence of sequence motifs up to four bases long at specified positions within the EXPAR template. We limited the analysis to motifs no longer than four bases, since the probability of longer motifs occurring at sufficiently high frequency in our data set was too low. Less than one-quarter of the theoretically possible 4760 position-motifs up to four bases long occurred at least three times in our data set and was considered in the analysis. To capture all possible position motifs up to four bases long would require us to experimentally characterize more than 1000 template sequences, which although desirable is expensive and laborious. Therefore, with our currently limited data, this approach cannot fully characterize the template's sequence-dependent performance, but we anticipate that the model can be refined in the future by incorporating additional sequence data into the data set.

For the machine learning approach, we first applied a feature selection method to identify significant position motifs that are over-represented in either Class I or Class II template sequences, and to determine the relative weights of these motifs. We then used a classifier method to derive a model that can categorize templates as Class I and II based on these significant position motifs. We tested many possible combinations of feature selection and classifier methods available within the Weka program suite (50). Best results were obtained by using feature selection through RELIEF attribute evaluation, coupled with Naïve Bayes classification. Which feature selection and classifier method works best in any particular application of machine learning depends on the topography of the search landscape intrinsic to the data set. Although some guiding principles exist, selecting the most appropriate feature selection and classification method is mainly empirical. Again, the same process was performed using shuffled values for P90 and Diff to test the predictive power of the model.

The classification performance using the machine learning approach for real versus shuffled data is illustrated in Figure 8 and summarized in Table 2. Using the real data, 70.6% of the well-performing Class I templates were correctly classified as Class I (70.6% sensitivity), and 77.5% of the poorly performing Class II templates were correctly classified as Class II (77.5% specificity). The Naïve Bayes method performed reasonably well despite the limited data set. This classification is significant at the level $P = 6 \times 10^{-12}$ (Fisher's exact test),

thus the predictive power of this approach was similar to that of the PWM method, with a slightly higher sensitivity, and a slightly lower specificity. The method had a PPV of 76% and an NPV of 72%. This implies that the detected features are robust and have clear predictive value regardless of the specific classification method used. Again, significantly fewer templates were correctly classified when using shuffled data (Figure 8B, Table 2), consistent with the null hypothesis of unbiased classification ($P = 0.89$ Fisher's exact test). This finding validates that the site specific motifs are meaningful and are linked to the EXPAR templates' performance in a functional manner.

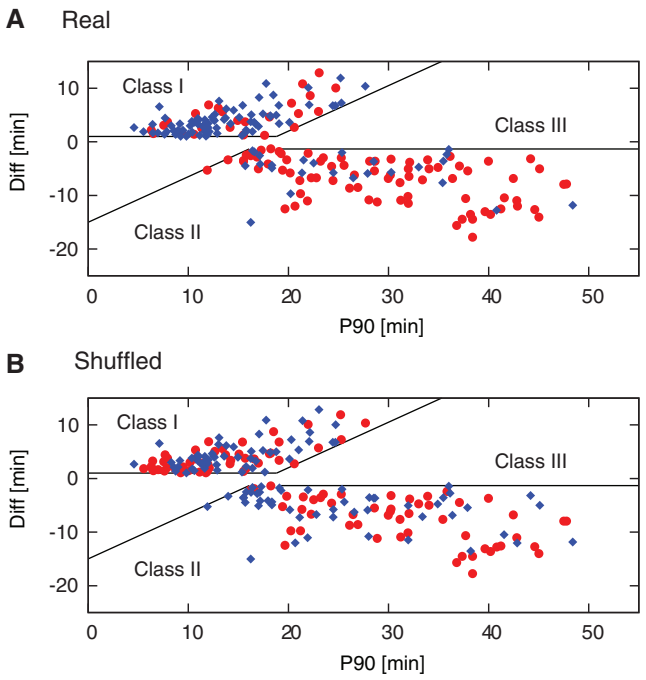

Figure 8. EXPAR template performance predicted using the Naïve Bayes machine learning approach. Plot of experimentally determined P90 versus Diff values for Class I and Class II templates, with color coding to indicate the predicted classification into Class I templates (good performers, blue diamond), and Class II templates (poor performers, red circles) for all test-set sequences after 10 iterations based on (**A**) real data, and (**B**) shuffled data.

Table 2. Confusion matrix using Naïve Bayes machine learning classification

		Actual Class I	Actual Class II
Real data		102	102
Predicted Class I	95	72	23
Predicted Class II	109	30	79
Shuffled data		102	102
Predicted Class I	108	55	53
Predicted Class II	96	47	49

Class I: well-performing templates. Class II: poorly performing templates. Values represent the number of template sequences in each set.

Computational tool for performance prediction

We have combined these different methods into one computational tool called ETSeq, which is open to the public and can be downloaded at https://github.com/expartools/ETSeq/wiki, to facilitate EXPAR template selection for new assay design based on thermodynamic and sequence dependence criteria. For a set of given input sequences, this tool determines thermodynamic parameters related to template-trigger binding and template self-hybridization, analyzes the sequences using the PWM and Naïve Bayes approaches, and generates a list of templates considered to be 'well performing' or 'poorly performing' by both methods, while template sequences with discordant classification are considered to be ambiguous.

We have applied this tool back to the full training set of 204 Class I and II sequences, and also to the 103 Class III template sequences with intermediate performance that were omitted in training the model (Table 3). Of the 102 well-performing Class I sequences, 81.4% were predicted to be Class I by both methods, 10.8% were predicted to be Class II, and 7.8% had discrepant classification results for the two methods, and were therefore considered ambiguous. The sensitivity for this combined tool therefore appears to be higher, with the caveat that the tool was used on its training set without cross validation. Of the 102 poorly performing Class II sequences, 75.5% were predicted to be Class II by both methods, 4.5% were predicted to be Class I, and 19.6% were ambiguous. The specificity for this combined tool therefore is slightly lower than for each individual method, possibly due to the larger percentage of sequences that gave ambiguous results. Out of the 103 Class III templates with intermediate performance, approximately 40% each were classified as either Class I or Class II, and the rest gave ambiguous results. As expected, the models cannot readily classify templates with intermediate performance, which likely have a mixture of features found in well and poorly performing template sequences. However, Class III sequences predicted to be Class I tend to be closer to the boundary between the Class III and Class I regions, while Class III sequences predicted to be Class II tend to be closer to the boundary between the Class III and Class II regions.

In practice, this tool would be used to narrow down a list of possible EXPAR template sequences for a given application to a set of templates that are likely to perform well, which would then be characterized

Table 3. Confusion matrix for the combined computational tool applied to the original data set

		Actual Class I	Actual Class III	Actual Class II
		102	103	102
Predicted Class I	128	83	40	5
Ambiguous	50	8	22	20
Predicted Class II	129	11	41	77

Class I: well-performing templates. Class II: poorly performing templates. Class III: templates with intermediate performance. Values represent the number of template sequences in each set.

experimentally. To test the value of our tool for such a scenario, we have experimentally characterized a small set of 30 EXPAR template sequences, of which 23 were predicted to be 'well performing' and seven were predicted to be 'poorly performing' by both methods. Due to cost and time reasons, we limited the size of this follow-on study to only 30 sequences, and we are aware that such a small study only provides rough preliminary performance estimates with large confidence intervals. Based on experimental performances, these 30 sequences were classified into Class I ('well performing'), Class II ('poor performing') or Class III ('intermediate performance'), as shown in Table 4, using the same criteria as for the 307 template sequences studied earlier.

Out of the actual Class I ('well performing') sequences, 100% were correctly predicted to be Class I, while out of the actual Class II ('poorly performing') sequences, 67% were correctly predicted to be Class II. Of the 23 templates predicted to be Class I ('well performing'), 12 were indeed class I templates, eight were Class III templates with intermediate performance, and only three were in the poorly performing Class II. Most of the templates with intermediate performance had acceptable Diff values, but amplified slower than our P90 cut-off. Although not optimal, these templates would still be acceptable in many cases. The PPV of the computational tool is therefore ~52% for predicting well-performing templates, and ~88% for predicting templates with good and intermediate performance, as opposed to ~33 and ~66% expected for randomly selected sequences, respectively. In training the model, we set the classification boundaries such that ~33% each out of the 307 sequences used were considered well-performing Class I templates, poorly performing Class II templates, and intermediate Class III templates. In the absence of a sequence based algorithm, one would expect a similar distribution for a randomly picked set of sequences. The tool can therefore enrich a set of predicted Class I ('well performing') templates with templates that have good and intermediate performance, while ruling out poor performers. Out of the seven templates predicted to be poorly performing, six were indeed in Class II, the model therefore appears to have a promising NPV.

This computational tool has significant practical value for new assay development, beyond the predictive power using only standard thermodynamic approaches. For example, we have recently developed another computational tool (Klaue, Qian *et al.* manuscript in preparation) that designs EXPAR templates to identify specific pathogens through the fingerprinting–two-stage EXPAR reaction (12). However, using only standard thermodynamic exclusion criteria for template design, we obtain many more possible EXPAR templates for conserved and non-cross reactive fingerprinting sites than can be experimentally screened. Using this tool, we can significantly narrow down the list of template sequences to be tested experimentally. Assay development therefore becomes more systematic and efficient even if the predictive power of the model is not 100%.

Given the limitations of our current data set, it is unreasonable to expect the current computational tool to perfectly classify well and poorly performing templates. For example, for the Naïve Bayes approach, less than one-quarter of the theoretically possible 4760 position-motifs were present with enough replicates in the data. To improve the predictive power of this computational tool, we will continue to experimentally characterize more sequences so we can re-train the model on a larger data set, thereby improving the classification accuracy. Such improvements will be made available to the public through continual upgrades to ETSeq.

Table 4. Confusion matrix for the combined computational tool applied to a new set of sequences

		Actual Class I	Actual Class III	Actual Class II
		12	9	9
Predicted Class I	23	12	8	3
Predicted Class II	7	0	1	6

Class I: well-performing templates; Class II: poorly performing templates; Class III: templates with intermediate performance. Values represent the number of template sequences in each set.

Sequence motif analysis

In addition to facilitating assay design, another motivation for this study was to better understand which sequence motifs give rise to efficient specific EXPAR amplification in the presence of trigger, and which motifs facilitate non-specific background amplification. Using the PWM approach, we have determined position-dependent nucleotide frequency maps (54,55) for templates that exhibit fast versus slow amplification (Figure 9A and C) and for templates with good versus poor separation between specific and non-specific amplification (Figure 9B and D). We further calculated the entropy (randomness) (56) of each variable position within Class I templates only, Class II templates only, and Class I and II templates combined (Figure 9E).

Using the Naïve Bayes machine learning approach, we have identified significant position motifs related to the characteristics of well-performing Class I templates (Figure 10A) that represent rapid amplification and good separation between specific and non-specific amplification, and significant position motifs related to poorly performing Class II templates (Figure 10B), which represent slow amplification and poor separation between specific and non-specific amplification, i.e. relatively facile non-specific amplification. Some specific positions within the template appear to be more informative than others (Figures 9E, 10C and D). Positions 7–10, which are repeated in positions 27–30, contain the most information, with some variability based on the parameter evaluated and approach used, with smaller contributions by other positions in the template.

Within the top 25 significant position motifs identified using the Naïve Bayes method (Figure 10; Table 5), well-performing templates have an over-representation of cytidine (C) and to some degree thymidine (T), while G and A are significantly under-represented. In contrast,

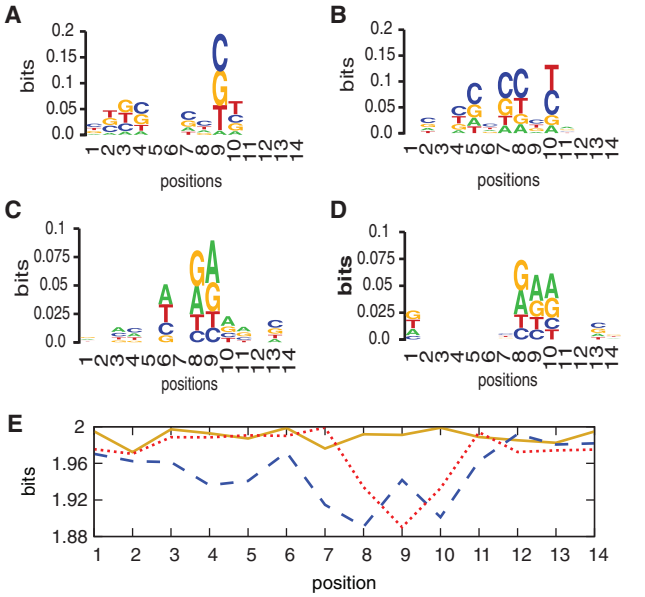

Figure 9. Position-dependent nucleotide frequency maps (sequence logos) (**A**) low P90 (fast amplification) (**B**) large positive Diff (good temporal separation between specific and non-specific amplification) (**C**) high P90 (slow amplification), and (**D**) large negative Diff (poor temporal separation between specific and non-specific amplification). **A–D** were generated by an online graphical sequence representation tool (54,55) using the template sequences which were used to generate PWM. At each position, the size of the overall stack indicates the conservation of that position, while the size of each symbol represents the frequency of that nucleic acid at that position. (**E**) Shannon entropy (56) for each variable position in Class I templates (blue dashed line), Class II templates (red dotted line) and Class I and Class II combined (orange solid line).

in poorly performing Class II templates A and G are significantly over-represented within the top 25 motifs, while C and T are significantly under-represented. A similar trend can be observed for the nucleotide composition of all variable positions within all Class I and Class II templates, but the effect is much more pronounced within the significant position motifs. For well-performing Class I templates, the GC content is 63% within the top 25 motifs and 55% within all variable positions. For poorly performing Class II templates, the GC content is 45% within the top 25 motifs and 50% within all variable positions. The GC content of the overall template sequences only weakly correlates with the template performance metrics P90, Diff and N10 (Supplementary Figure S3). In predicting template performance, the type of base and the specific position of the base appear to be more significant than overall GC content.

The Naïve Bayes machine learning approach further identified several significant multi-base motifs, some of which occur multiple times in several positions within the template. For example, the top 25 motifs for Class I include CC three times. The top 25 motifs for Class II include AG five times, and GA three times. The single-base position motif A appears six times at different positions, yet the single-base position motif G occurs only once. Therefore, G adjacent to A (i.e. GA or AG) appears to be correlated with poor template performance, rather than a single G by itself. Some shorter motifs further overlap and/or are contained within longer motifs, such as CC-8, CT-9, and CCT-8 within the top 25 Class I motifs. These results indicate that multi-base motifs are important in understanding the performance characteristics of different EXPAR templates. Multi-base motifs cannot be identified using the standard PWM approach as implemented here.

We have considered the hypothesis that poor performance and non-specific background amplification is caused by primer-dimer type template self-priming, since the most significant positions 7–10, repeated in positions 27–30, appear near the 3′ end of the template. However, poorly performing Class II templates do not have a higher amount of 3′ self-complementarity than well-performing Class I templates. The template sequences used in this study were selected to have a self-hybridization T_M less than 25°C, and thus are unlikely to interact efficiently at 55°C. In addition, poorly performing templates contain GA rich motifs near the 3′ end that do not readily hybridize with themselves, unlike GT-rich motifs that can form wobble pairs. Motifs found in poorly performing templates are not expected to self-dimerize more effectively than motifs found in well-performing templates. Although primer-dimer type template self-priming cannot be ruled out as a contributing factor, it is unlikely to be the predominant cause of poor performance and non-specific background amplification.

We previously observed that pre-incubating the polymerase with the EXPAR template significantly accelerates non-specific background amplification (47). This and other observations described in (47) led to the hypothesis that non-specific background amplification observed in EXPAR involves binding of the polymerase to the single-stranded template, which is present at nanomolar concentrations in the reaction. In contrast, we found that the nicking enzyme is not critical for initiating non-specific amplification (47), even though it is required to propagate the reaction once initiated.

Interestingly, G and A, which are over-represented in certain positions of poorly performing templates, are both bicylic purine nucleotide bases that are larger and more hydrophobic than the monocylic pyrimidine bases C and T, which occur more frequently in well-performing template sequences. Cytidine, which is significantly over-represented in well-performing templates, has the lowest hydrophobicity of all four nucleotides (28). Furthermore, it has been shown that the affinity between a DNA polymerase and homo-oligonucleotides increases in the order $d(pC)n < d(pT)n < d(pG)n \approx d(pA)n$ (28), which suggests that the polymerase present in the reaction may interact more strongly with certain regions of the single-stranded template of poorly performing sequences, compared to well-performing template sequences.

We previously hypothesized that during non-specific EXPAR background amplification, the polymerase synthesizes DNA complementary to the single-stranded template, but that this reaction is primed in an unconventional manner (47). Such templated but unprimed DNA amplification may also occur in other isothermal DNA

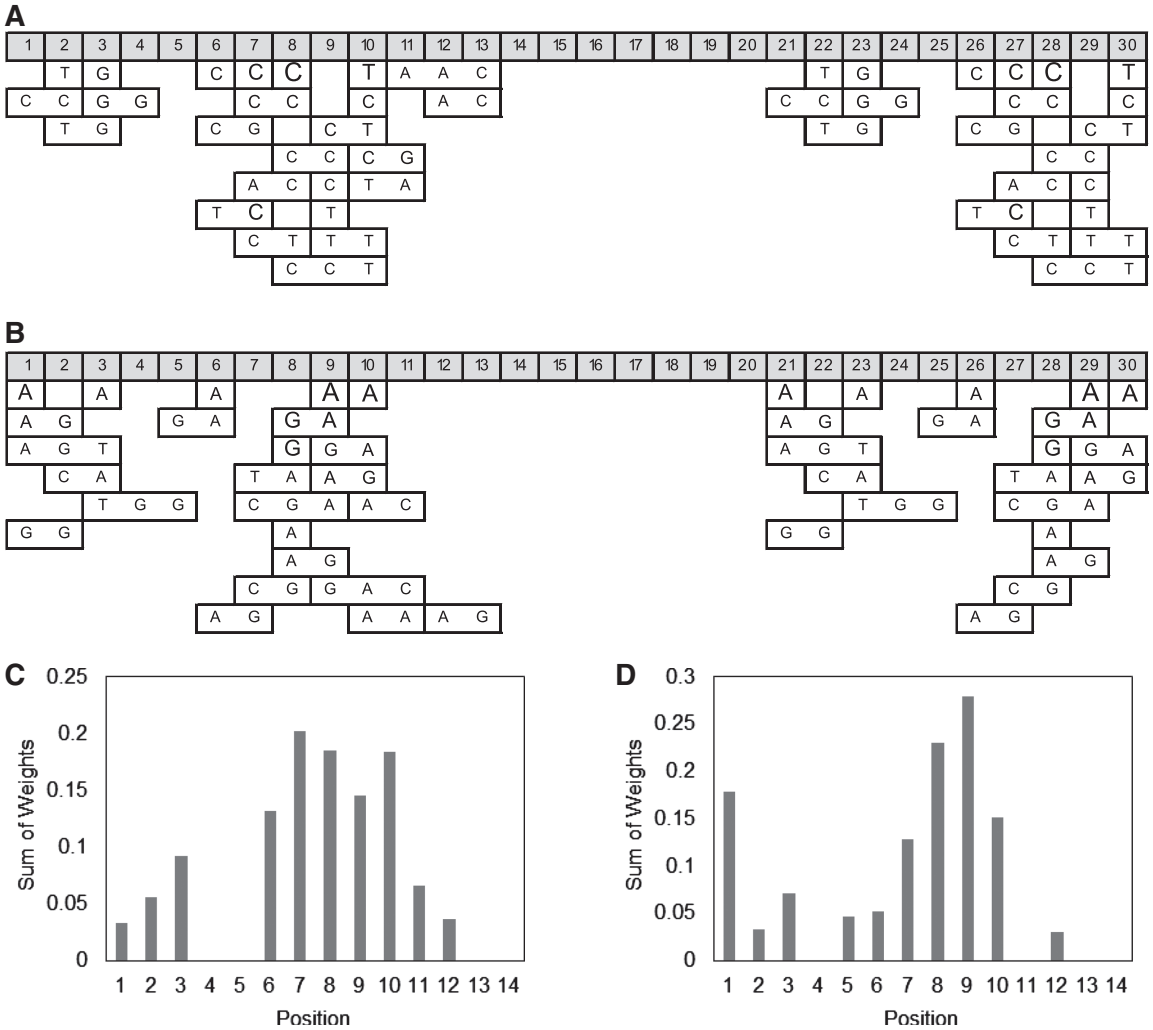

Figure 10. Graphical representation of significant position motifs identified via Naïve Bayes machine learning. The 25 highest-ranking position motifs of (**A**) well-performing Class I templates and (**B**) poorly performing Class II templates. The font size for significant motifs shown in A and B is correlated to each motif's relative weight, i.e. relative importance. Sum of weights calculated for each variable position using the top 25 motifs identified by RELIEF attribute evaluation within (**C**) well-performing Class I templates and (**D**) poorly performing Class II templates.

Table 5. Base composition of position sequence motifs and of all variable positions by Class

	Nucleotide	A (%)	G (%)	T (%)	C (%)
Class I	Top 25 motifs	12	14	26	49
	Variable positions	20	25	25	30
Class II	Top 25 motifs	46	35	8	10
	Variable positions	27	27	23	23

Class I (well-performing) templates; Class II (poorly performing templates). Top 25 motifs: highest ranking sequence motifs identified via Naïve Bayes. Variable positions: positions 1–14 within the template sequences.

amplification reactions, but unless this rare event leads to efficient amplification it will go unnoticed. However, due to the positive feedback loop present in EXPAR, generating even small numbers of short DNA sequences complementary to a template of general structure X′rX′ can trigger the reaction, resulting in exponential amplification of trigger X. The precise mechanism of the priming process is unclear, but could involve priming from a single-dNTP bound within the post-insertion site of the polymerase. Polymerases are known to interact specifically with the base present within the post-insertion site that contains the recessed 3′ hydroxyl group to which the incoming nucleotide is added (27,57). It has been shown that a single nucleotide can serve as a primer for DNA extension by thermophilic polymerases, albeit with much lower K_M and v_{max} values than longer primers (25,58). The higher affinity of the polymerase to certain GA rich regions in the template might allow for a nucleotide bound in the post-insertion site to be used as primer analog.

CONCLUSIONS

We systematically characterized the amplification performance of over 300 randomly designed EXPAR

template sequences. The results show that templates with similar thermodynamic characteristics related to trigger-template binding and template self-hybridization can perform quite differently in the reaction. Simple thermodynamic rules are of limited value in predicting the performance of a new template sequence. Therefore, the design of new assays often turns into laborious trial and error. By applying two different computational models, a standard PWM approach and a Naïve Bayes machine learning algorithm using position motifs, we were able to differentiate with ~67–70% sensitivity and ~77–80% specificity between EXPAR templates that perform well or poorly in the reaction. We then combined these methods into one computational too, and have characterized the performance of a small set of sequences designed *de novo* using this tool. This computational tool can significantly facilitate new assay design by enriching a set of EXPAR templates designed for a particular application with well-performing sequences, and by ruling out templates with poor performance. We will continually improve the predictive power of this computational tool by adding new experimental data to the set used for training the model.

The results of this study further indicate which sequence characteristics favor slow versus fast and specific versus non-specific amplification, related to most influential positions within the template, base composition at these positions, and occurrence of certain multi-base motifs. Stronger interactions between the polymerase and purine-rich regions of a single-stranded DNA template may facilitate templated DNA polymerization not primed in the conventional manner. We are currently conducting enzyme kinetic experiments and in silico-molecular modeling to obtain a clearer understanding of the reaction mechanism involved, and are combining the tool described herein with other computational tools for EXPAR assay design, starting from a genomic target sequence.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table S1, Supplementary Figures S1–S3, Supplementary Methods and Supplementary Reference [59].

ACKNOWLEDGEMENTS

We thank Dr Steven Youra and Dr Craig Adams for helpful discussions concerning the manuscript, and Dr Ali Nadim for developing the MatLab code behind the EDAT tool, plus a write up on this code.

FUNDING

The National Institutes of Health (award R01AI076247 and an ARRA supplement to this award); National Science Foundation's Frontiers in Integrative Biological Research (FIBR-0527023); NSF's BEACON Center for the Study of Evolution in Action (contract No. DBI-0939454). Funding for open access charge: National Institutes of Health (R01AI076247).

Conflict of interest statement. None declared.

REFERENCES

1. Hofmann,W.P., Dries,V., Herrmann,E., Gartner,B., Zeuzem,S. and Sarrazin,C. (2005) Comparison of transcription mediated amplification (TMA) and reverse transcription polymerase chain reaction (RT-PCR) for detection of hepatitis C virus RNA in liver tissue. *J. Clin. Virol.*, **32**, 289–293.
2. Gracias,K.S. and McKillip,J.L. (2007) Nucleic acid sequence-based amplification (NASBA) in molecular bacteriology: a procedural guide. *J. Rapid Methods Autom. Microbiol.*, **15**, 295–309.
3. Jeong,Y.J., Park,K. and Kim,D.E. (2009) Isothermal DNA amplification in vitro: the helicase-dependent amplification system. *Cell Mol. Life Sci.*, **66**, 3325–3336.
4. Lutz,S., Weber,P., Focke,M., Faltin,B., Hoffmann,J., Muller,C., Mark,D., Roth,G., Munday,P., Armes,N. *et al.* (2010) Microfluidic lab-on-a-foil for nucleic acid analysis based on isothermal recombinase polymerase amplification (RPA). *Lab Chip.*, **10**, 887–893.
5. Piepenburg,O., Williams,C.H., Stemple,D.L. and Armes,N.A. (2006) DNA detection using recombination proteins. *PLoS Biol.*, **4**, e204.
6. Mori,Y. and Notomi,T. (2009) Loop-mediated isothermal amplification (LAMP): a rapid, accurate, and cost-effective diagnostic method for infectious diseases. *J. Infect. Chemother.*, **15**, 62–69.
7. Mitani,Y., Lezhava,A., Sakurai,A., Horikawa,A., Nagakura,M., Hayashizaki,Y. and Ishikawa,T. (2009) Rapid and cost-effective SNP detection method: application of SmartAmp2 to pharmacogenomics research. *Pharmacogenomics*, **10**, 1187–1197.
8. Hellyer,T.J. and Nadeau,J.G. (2004) Strand displacement amplification: a versatile tool for molecular diagnostics. *Expert Rev. Mol. Diagn.*, **4**, 251–261.
9. McHugh,T.D., Pope,C.F., Ling,C.L., Patel,S., Billington,O.J., Gosling,R.D., Lipman,M.C. and Gillespie,S.H. (2004) Prospective evaluation of BDProbeTec strand displacement amplification (SDA) system for diagnosis of tuberculosis in non-respiratory and respiratory samples. *J. Med. Microbiol.*, **53**, 1215–1219.
10. Jung,C., Chung,J.W., Kim,U.O., Kim,M.H. and Park,H.G. (2010) Isothermal target and signaling probe amplification method, based on a combination of an isothermal chain amplification technique and a fluorescence resonance energy transfer cycling probe technology. *Anal. Chem.*, **82**, 5937–5943.
11. Van Ness,J., Van Ness,L.K. and Galas,D.J. (2003) Isothermal reactions for the amplification of oligonucleotides. *Proc. Natl Acad. Sci. USA*, **100**, 4504–4509.
12. Tan,E., Erwin,B., Dames,S., Voelkerding,K. and Niemz,A. (2007) Isothermal DNA amplification with gold nanosphere-based visual colorimetric readout for herpes simplex 2 virus detection. *Clin. Chem.*, **53**, 2017–2020.
13. Connolly,A.R. and Trau,M. (2010) Isothermal detection of DNA by beacon-assisted detection amplification. *Angew. Chem. Int. Ed. Engl.*, **49**, 2720–2723.
14. Niemz,A., Ferguson,T.M. and Boyle,D.S. (2011) Point-of-care nucleic acid testing for infectious diseases. *Trends Biotechnol.*, **29**, 240–250.
15. Abd-Elsalam,K.A. (2003) Bioinformatic tools and guideline for PCR primer design. *Afr. J. Biotechnol.*, **2**, 91–95.
16. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
17. Kimura,Y., de Hoon,M.J., Aoki,S., Ishizu,Y., Kawai,Y., Kogo,Y., Daub,C.O., Lezhava,A., Arner,E. and Hayashizaki,Y. (2011) Optimization of turn-back primers in isothermal amplification. *Nucleic Acids Res.*, **39**, e59.

18. Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
19. SantaLucia,J., Allawi,H.T. and Seneviratne,P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.
20. Wu,D.Y., Qian,J., Wallace,R.B., Pal,B.K. and Ugozzoli,L. (1991) The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol.*, **10**, 233–238.
21. Dieffenbach,C.W., Lowe,T.M.J. and Dveksler,G.S. (1993) General concepts for PCR primer design. *PCR Methods Appl.*, **3**, S30–S37.
22. Spencer,W.J., Rhoads,R.E. and Rychlik,W. (1991) Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.*, **19**, 698.
23. Marky,L.A., Frank,R., Breslauer,K.J. and Blocker,H. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
24. Sheffield,V.C., Cox,D.R., Lerman,L.S. and Myers,R.M. (1989) Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes. *Proc. Natl Acad. Sci. USA*, **86**, 232–236.
25. Nevinsky,G.A., Nemudraya,A.V., Levina,A.S. and Khomov,V.V. (1989) The algorithm of estimation of the Km values for primers of various structure and length in the polymerization reaction catalyzed by Klenow fragment of DNA-polymerase-I from Escherichia coli. *FEBS Lett.*, **258**, 166–170.
26. Beese,L.S., Derbyshire,V. and Steitz,T.A. (1993) Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science*, **260**, 352–355.
27. Kiefer,J.R., Mao,C., Braman,J.C. and Beese,L.S. (1998) Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. *Nature*, **391**, 304–307.
28. Kolocheva,T.I., Nevinsky,G.A., Volchkova,V.A., Levina,A.S., Khomov,V.V. and Lavrik,O.I. (1989) DNA-polymerase-I (Klenow fragment)—role of the structure and length of a template in enzyme recognition. *FEBS Lett.*, **248**, 97–100.
29. Pavlov,A.R., Pavlova,N.V., Kozyavkin,S.A. and Slesarev,A.I. (2004) Recent developments in the optimization of thermostable DNA polymerases for efficient applications. *Trends Biotechnol.*, **22**, 253–260.
30. Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.
31. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
32. Salzberg,S.L. (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, **13**, 365–376.
33. Ben-Gal,I., Shani,A., Gohr,A., Grau,J., Arviv,S., Shmilovici,A., Posch,S. and Grosse,I. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, **21**, 2657–2666.
34. Siddharthan,R. (2010) Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *Plos One*, **5**, e9722.
35. Chen,J.N., Huang,H.K., Tian,S.F. and Qu,Y.L. (2009) Feature selection for text classification with Naïve Bayes. *Expert Syst. Appl.*, **36**, 5432–5435.
36. Liu,Y.F., Guo,J.M. and Lee,J.D. (2011) Halftone image classification using LMS algorithm and naïve Bayes. *IEEE Trans. Image Process.*, **20**, 2837–2847.
37. Tan,E., Wong,J., Nguyen,D., Zhang,Y., Erwin,B., Van Ness,L.K., Baker,S.M., Galas,D.J. and Niemz,A. (2005) Isothermal DNA amplification coupled with DNA nanosphere-based colorimetric detection. *Anal. Chem.*, **77**, 7984–7992.
38. Andreson,R., Mols,T. and Remm,M. (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Res.*, **36**, e66.
39. Brownie,J., Shawcross,S., Theaker,J., Whitcombe,D., Ferrie,R., Newton,C. and Little,S. (1997) The elimination of primer-dimer accumulation in PCR. *Nucleic Acids Res.*, **25**, 3235–3241.
40. Hanaki,K., Odawara,T., Muramatsu,T., Kuchino,Y., Masuda,M., Yamamoto,K., Nozaki,C., Mizuno,K. and Yoshikura,H. (1997) Primer/template-independent synthesis of poly d(A-T) by Taq polymerase. *Biochem. Biophys. Res. Commun.*, **238**, 113–118.
41. Hanaki,K., Odawara,T., Nakajima,N., Shimizu,Y.K., Nozaki,C., Mizuno,K., Muramatsu,T., Kuchino,Y. and Yoshikura,H. (1998) Two different reactions involved in the primer/template-independent polymerization of dATP and dTTP by Taq DNA polymerase. *Biochem. Biophys. Res. Commun.*, **244**, 210–219.
42. Ogata,N. and Miura,T. (1997) Genetic information 'created' by archaebacterial DNA polymerase. *Biochem. J.*, **324**, 667–671.
43. Ogata,N. and Miura,T. (1998) Creation of genetic information by DNA polymerase of the thermophilic bacterium Thermus thermophilus. *Nucleic Acids Res.*, **26**, 4657–4661.
44. Ogata,N. and Miura,T. (1998) Creation of genetic information by DNA polymerase of the archaeon Thermococcus litoralis: influences of temperature and ionic strength. *Nucleic Acids Res.*, **26**, 4652–4656.
45. Liang,X.G., Jensen,K. and Frank-Kamenetskii,M.D. (2004) Very efficient template/primer-independent DNA synthesis by thermophilic DNA polymerase in the presence of a thermophilic restriction endonuclease. *Biochem.*, **43**, 13459–13466.
46. Zyrina,N.V., Zheleznaya,L.A., Dvoretsky,E.V., Vasiliev,V.D., Chernov,A. and Matvienko,N.I. (2007) N.BspD6I DNA nickase strongly stimulates template-independent synthesis of non-palindromic repetitive DNA by Bst DNA polymerase. *Biol. Chem.*, **388**, 367–372.
47. Tan,E., Erwin,B., Dames,S., Ferguson,T., Buechel,M., Irvine,B., Voelkerding,K. and Niemz,A. (2008) Specific versus nonspecific isothermal DNA amplification through thermophilic polymerase and nicking enzyme activities. *Biochemistry*, **47**, 9987–9999.
48. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.
49. Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, MA.
50. Frank,E., Hall,M., Trigg,L., Holmes,G. and Witten,I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.
51. Sikonja,M. and Kononenko,I. (2003) Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.*, **53**, 23–69.
52. Cheeseman,P. and Stutz,J. (1996) Bayesian classification (AutoClass): theory and results. In: Fayyad,U., Piatelsky-Shapiro,G., Smyth,P. and Uthurusamy,R. (eds), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, pp. 153–180.
53. Curk,T., Demsar,J., Xu,Q.K., Leban,G., Petrovic,U., Bratko,I., Shaulsky,G. and Zupan,B. (2005) Microarray data mining with visual programming. *Bioinformatics*, **21**, 396–398.
54. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
55. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos:a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
56. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
57. Johnson,S.J., Taylor,J.S. and Beese,L.S. (2003) Processive DNA synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. *Proc. Natl Acad. Sci. USA*, **100**, 3895–3900.
58. Bukhrashvili,I.S., Chinchaladze,D.Z., Lavrik,O.I., Levina,A.S., Nevinsky,G.A. and Prangishvili,D.A. (1989) Comparison of initiating abilities of primers of different length in polymerization reactions catalyzed by DNA-polymerases from thermoacidophilic archaebacteria. *Biochim. Biophys. Acta*, **1008**, 102–107.
59. Coleman,T.F. and Li,Y. (1994) On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Math. Program.*, **67**, 189–224.