Nucleic Acids Research, 1998, Vol. 26, No. 23 5525–5532

Interdependence between DNA template secondary structure and priming efficiencies of short primers

Lev Lvovsky, Ilya Ioshikhes[+], Mugasimangalam C. Raja[§], Dina Zevin-Sonkin, Irina A. Sobolev, Arthur Liberzon, J. Shwartzburd and Levy E. Ulanovsky[§,*]

Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

Received April 29, 1998; Revised September 3, 1998; Accepted September 28, 1998

ABSTRACT

Here we analyze the effect of DNA folding on the performance of short primers and describe a simple technique for assessing hitherto uncertain values of thermodynamic parameters that determine the folding of single-stranded DNA into secondary structure. An 8mer with two degenerate positions is extended simultaneously at several complementary sites on a known template (M13mp18) using one, two or three (but never all four) of the possible dNTPs. The length of the extension is site specific because it is limited by the first occurrence in the downstream template sequence of a base whose complementary dNTP is not present. The relative priming efficiencies of different sites are then ranked by comparing their band brightnesses on a gel. The priming efficiency of a short primer (unlike conventional long primers) depends dramatically on the secondary structure of the template at and around the priming site. We calculated the secondary structure and its effect on priming using a simple model with relatively few parameters which were then optimized to achieve the best match between the predictions and the actual rankings of the sites in terms of priming efficiency. This work introduces an efficient and conceptually novel approach that in the future can make use of more data to optimize a larger set of DNA folding parameters in a more refined model. The model we used, however crude it may be, significantly improved the prediction of priming efficiencies of 8mer primers and appreciably raised the success rate of our DNA sequencing technique (from 67 to 91% with a significance of $P < 7 \times 10^{-5}$), which uses such primers.

INTRODUCTION

Whether the dramatic variation in priming efficiency of the same short primer at different priming sites in the same template can be explained by the local secondary structure differences is a fascinating question. One purpose of this paper was to investigate the folding of the template around the priming sites to verify this hypothesis. Our ultimate goal was to predict the sequencing efficiencies of 8mer primers by computing the local secondary structure stability and its interference with priming. One problem we encountered was that, compared with the extent of knowledge about RNA, relatively little is known about the values of thermodynamic parameters for the folding of DNA. The DNA parameters that have been characterized are those for the base pairing and stacking of dinucleotides (1–4) and the destabilization effects of some specific mismatches (5–10). The values of other parameters, such as the destabilization effects of other mismatches and of different loop sizes, are unclear (to the best of our knowledge).

We discovered that ranking the priming efficiencies of 8mer priming sites provides a technique that dramatically reduces the number of experiments and the amount of effort required to determine the values of DNA folding parameters. In this technique, DNA folding parameters are optimized by minimizing the difference between the predicted and actual efficiencies of a short primer at different sites.

For each experiment we used a short primer (an 8mer with two degenerate positions) that was complementary to several sites in a single-stranded phage template of known sequence (M13mp18). During the extension reaction, such a primer primes simultaneously at those sites. To compare the relative priming efficiencies of the primer at different sites, we designed a technique of differential extension with nucleotide subsets, in which some of the four dNTPs are absent during the extension of the primer (more details below).

The length of the primer extension depends on the template sequence downstream from the priming site, because it determines how soon the nascent strand extension requires a dNTP that is not present in the reaction. This sequence dependence makes the extension length vary from site to site, hence the extension is termed *differential*. The relative priming efficiencies at different sites can therefore be assessed on the same gel by comparing the band brightness of the differential extension products, which are of known sizes because the template sequence is known. The uncertain parameters of DNA folding are then optimized by comparing the ranked priming efficiencies of the 8mer primer at different sites with rankings predicted from secondary structure calculations.

*To whom correspondence should be addressed at present address. Tel: +1 630 252 3940; Fax: +1 630 252 3387; Email: levy@anl.gov

Present addresses: [+]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA and [§]Center for Mechanistic Biology and Biotechnology, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA

Table 1. Separation of the reaction sites into two groups: 43 sites with favorable calculated secondary structure ($\Delta\Delta G$ *below* the threshold) and 48 sites with unfavorable calculated secondary structure ($\Delta\Delta G$ *above* the threshold)

	No. of reactions	Successful reactions	Failed reactions	Success rate (%)
Entire set of DENS reactions performed	91	61	30	67
Below the –13 kcal/mol threshold	43	39	4	91
Above the –13 kcal/mol threshold	48	22	26	46

The probability of obtaining the same or better improvement in the success rate by chance is 7×10^{-5} (by the probability estimate method in Materials and Methods).

Our computation model was somewhat crude, involving relatively few parameters, and is used here only to demonstrate the conceptually novel approach rather than to finalize the DNA folding parameters to any significant degree of accuracy. However, the parameter values optimized within this simple model have been found to significantly improve the prediction of the priming efficiency variation of 8mer primers from site to site, as compared with unoptimized values. We then successfully used this prediction model to select 8mer primers for DNA sequencing by DENS (differential extension with nucleotide subsets; 11) and raised the DENS sequencing success rate appreciably.

MATERIALS AND METHODS

Octamer oligonucleotides (containing two degenerate positions at the 5′ end) were from a DENS (11) sequencing library supplied by DNAgency (Malvern, PA). SequiTherm polymerase and related reagents were from the Epicentre Technologies sequencing kit (catalogue no. S20100). The single-stranded M13mp18 DNA template was from Amersham (UK) (catalogue no. US 70704). Deoxribonucleotide triphosphates (dNTPs) were from Pharmacia LKB (Sweden). Fourteen different dNTP mixes (A, C, G, T, A+C, A+G, A+T, C+G, C+T, G+T, A+C+G, A+C+T, A+G+T and C+G+T) were premixed and stored at –20°C.

The primer extension reactions were performed as follows. The reaction volume was 12 µl, containing 0.25 pmol of single-stranded M13mp18 template, 150–200 pmol of the degenerate 8mer primer; 5 pmol of each of the selected dNTPs, one of which was spiked (1:10) with α-^{32}P label (3000 Ci/mmol; Amersham), and 1.5 µl of the reaction buffer concentrate from the Epicentre Technologies sequencing kit. The reaction buffer contained 2 mM Mg^{2+}, which is equivalent to 0.18 M Na^+ (assuming the formula for conversion of $[Mg^{2+}]$ into $[Na^+]$ to be $[Na^+] = 4 \times [Mg^{2+}]^{1/2}$; see *Oligo Primer Analysis Software Version 5.0 Manual*, 1994). The reaction mixture was incubated at 90°C for 3 min and placed immediately in a 20°C water bath. SequiTherm (5 U) was then added, initiating the primer extension reaction, which was allowed to proceed at 20°C for 10 min. The reaction was stopped by adding 12 µl of stop solution (10 mM EDTA in formamide). The differential extension products were then electrophoresed on a denaturing 12% polyacrylamide gel. The efficiencies of priming were assessed by the intensity of the corresponding bands, taking into account the number of radioactive labels contributing to each band (radiolabeled bases in the extension sequence).

DENS sequencing reactions were performed as described earlier (11).

Free energies (ΔG) associated with folding were computed using a home-made program (in C++ in UNIX), which is a dynamic programming algorithm very similar to that in Zuker (12). We analyzed foldings of the 200 nt-long segment of DNA centered on each priming site. This length of segment was chosen as the longest that we could deal with within a feasible computation time. We used simplified energy rules (as compared with those used in Jaeger *et al.*; 13), where all the energy contributions of perfectly matched double helices were calculated using the nearest neighbor model for base stacking (1). The end effects were taken into account by adding a sequence-independent term to the free energy of the folding per double helix end. Internal mismatches were treated as internal loops containing one base in each strand. The destabilizing energies for loops were assigned according to the type of the loop (hairpin, internal or bulge) and its length. We applied no penalty for internal loops being asymmetric. Our loop energies were assumed to be independent of their sequences. An exception was made only for internal loops of length two, single-base mismatches, which were assigned energy values according to the eight possible mismatches, disregarding the neighbor effects. The energies for multibranched loops were calculated using a linear model as in Jaeger *et al.* (13). In constrained foldings the segments that were supposed to remain unfolded were assigned prohibitively high positive energy values for their base pairing, as in Zuker and Stieger (14). In all the calculations the temperature was assumed to be 20°C, in accordance with our experimental conditions. The free energy of primer–template annealing, $\Delta G_{annealing}$, was calculated using the nearest neighbor model (1), except that the initialization term was omitted.

The statistical significance of our prediction of the success of the DENS sequencing reaction was estimated as follows. Our set of 91 reactions had a $\Delta\Delta G$ value associated with each priming site (equation **1**) and was divided by the threshold of $\Delta\Delta G$ (–13.0 kcal/mol) into two groups: 43 reactions below the threshold (predicted to work) and 48 reactions above the threshold (predicted to fail). Among the 43 reactions in the first group, 39 worked (91%), while among the 48 reactions in the second group, 22 worked (46%) (Table 1). To calculate the statistical probability that our division was as good as it was by chance, we assumed that our 91 reactions (priming sites) were assigned $\Delta\Delta G$ values randomly and ordered by these wrong (random) $\Delta\Delta G$ values. Then, if we take N reactions from the top of the list of our total set of 91 reactions (of which 61 worked and 30 failed), the probability that m or more of them work is:

$$P_N(m) = \frac{\sum_{i=m}^{N} \binom{61}{i}\binom{30}{N-i}}{\binom{91}{N}} = \sum_{i=m}^{N} \frac{61!\cdot 30!}{i!(61-i)!(N-i)!(30-N+i)!} \Big/ \frac{91!}{N!(91-N)!}$$

For our data ($N = 43$ and $m = 39$), this probability equals $P_{43}(39) = 4 \times 10^{-6}$, which is a measure of how good our chosen division is. There are 90 ways to divide our 91 reactions (ranked by the $\Delta\Delta G$ values) into two groups of sizes N and $91 - N$ by setting a $\Delta\Delta G$ threshold. Calculation of the statistical significance of the results of

our DENS sequencing experiments should take into account this multiplicity. The probability that at least one of these 90 divisions (for any of the values of N from 1 to 90) will have $P_N(m) \leq 4 \times 10^{-6}$ was calculated to be 7×10^{-5} by a program that exhaustively scanned all the possible combinations (pairs) of values of N and m.

RESULTS

Ranking priming efficiency using differential extension experiments

The experimental data we used for assessing the unknown parameters of DNA folding were the relative priming efficiencies of an 8mer primer (with two degenerate positions at the 5′-end) at different sites on a single-stranded M13mp18 template. We used the following technique, which allowed us to obtain a complete set of such data in a single priming experiment using differential extension with nucleotide subsets. Each 8mer was extended on a single-stranded M13mp18 template in several separate reactions. Each reaction contained one of the 14 possible subsets of dNTPs (A, C, G, T, A+C, A+G, A+T, C+G, C+T, G+T, A+C+G, A+C+T, A+G+T and C+G+T), but none of them contained all four dNTPs (Fig. 1 and Table 2). Since there was only a partial set of dNTPs in each tube, the extension of the primer stopped at the first occurrence in the template of a base that was not complementary to any of the dNTPs present. Therefore, the length of the extension product at each site was sequence- and dNTP set-dependent (Table 2 and Fig. 1). After separation of the products by 12% PAGE, we could identify bands corresponding to particular priming sites by the length of the product (because the sequence of the M13mp18 template is known) (Fig. 2). The strength of the band indicated the relative efficiency of priming at the corresponding site; in this way we ranked the priming sites by their efficiencies. We took into account the difference in the numbers of the radiolabeled nucleotides in the extension sequences at different sites.

Primers and priming sites. We compared every priming site with every other site (in terms of band brightness) wherever they could be compared in the same lane on the gel. We wrote a program that calculates the number of site pairs that can be compared using the same dNTP set (lane) for a given primer on a given template. We used this number to select the most informative primers for the experiment (an example is given in Table 2). A set of primers and their priming sites selected in this way was used to optimize DNA folding parameters and for this reason is referred to as the *training set*. Our training set consisted of seven sets of ranked sites (for seven different two-base-degenerate 8mer primers, each having three to eight complementary sites on M13mp18) containing, overall, 40 priming sites and 95 ranked pairs of sites that could be compared on a gel. The sequences of the seven primers were as follows (5′→3′): NNGGGAAG; NNGGAAGG; NNGCCAGC; NNGAAACA; NNGATAAA; NNGGAATT; NNGAGTAA.

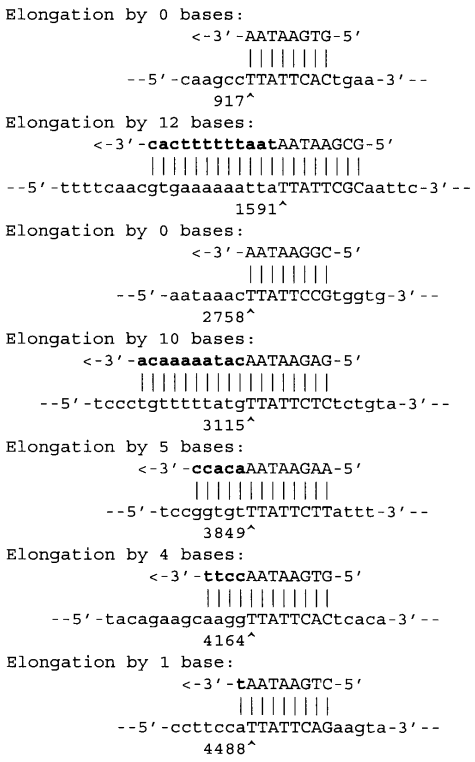

Figure 1. Differential extension of the primer 5′-NNGAATAA-3′ at all complementary sites on single-stranded M13mp18 template using the A+C+T subset of dNTPs. N stands for a degenerate position. The primer is one of the primers we used for the test set. The figure shows the extension of the primer at each of the sites when only dATP, dCTP and dTTP (but not dGTP) are present in the reaction mixture. The upper line for each elongation indicates the primer strand; the lower line indicates the priming site in the template. The primers and their complementary sites are shown in upper case and the bases added during the extension are shown in bold lower case. The number followed by ^ indicates the position in the M13 template.

Table 2. Differential extension lengths (in nt) of a particular primer (5′-NNGAATAA-3′) when extended on a single-stranded M13mp18 template with all possible dNTP subsets

Position	Various dNTP subsets													
in M13	A	C	G	T	A+C	A+G	A+T	C+G	C+T	G+T	A+C+G	**A+C+T**	**A+G+T**	**C+G+T**
917	0	0	2	0	0	2	0	3	0	2	2	**0**	**2**	**11**
1591	0	0	0	1	0	0	9	0	1	1	0	**12**	**9**	**1**
2758	0	0	1	0	0	1	0	1	0	4	1	**0**	**11**	**4**
3115	0	1	0	0	2	0	0	1	1	0	2	**10**	**0**	**1**
3849	1	0	0	0	5	1	1	0	0	0	9	**5**	**1**	**0**
4164	0	2	0	0	2	0	0	2	4	0	2	**4**	**0**	**12**
4488	0	0	0	1	0	0	1	0	1	3	0	**1**	**14**	**3**

The primer has seven complementary sites on the template. The last three columns are in bold to emphasize the fact that they contain the necessary information for ranking all the sites. This is in contrast to the first two columns, where only two sites can be ranked (sites at positions 3115 and 4164 can be ranked in the lane of the extension using dCTP alone).

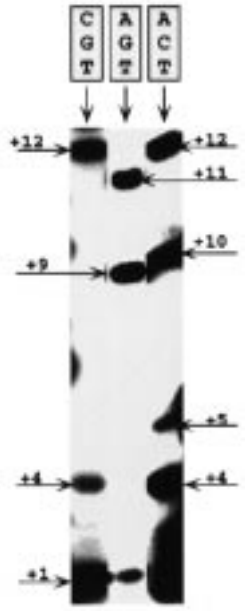

Figure 2. PAGE autoradiogram of three extension reactions of primer 5′-NNGAATAA-3′ on M13mp18 template with the A+C+T, A+G+T and C+G+T subsets of dNTPs. The identified bands are marked by arrows. The numbers above the arrows indicate the differential extension lengths (in nt).

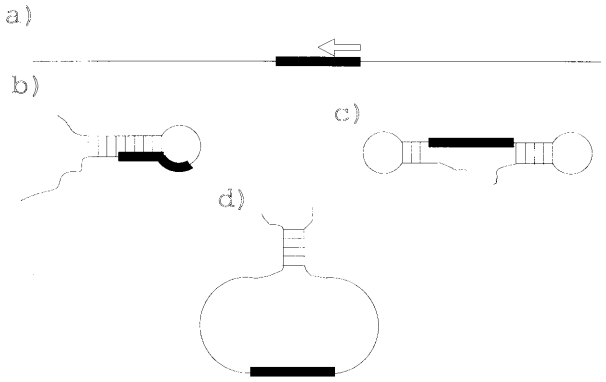

Figure 3. Schematic representation of the clean site. (**a**) Stretched segment of the template. The arrow indicates the primer. The black bar in the middle indicates the 'clean site', which is essential for priming and hence should not be involved in the secondary structure. The clean site includes the 8 nt-long primer-complementary site and a downstream stretch whose length was optimized and found to be 8 nt (16 nt in total). (**b**) The best folding minimizes the free energy but may be incompatible with priming by the primer under consideration. Here the case is illustrated by the priming site being involved in the secondary structure. (**c**) The best folding compatible with priming. It has the minimal free energy among the foldings that leave the 'clean site' (black bar) free for priming. (**d**) Minimal loop length. The double-stranded region is rigid and therefore cannot be part of a loop that is smaller than twice the length of the primer. Also, the dimensions of DNA polymerase probably add some restriction on the minimal loop length, which according to our results is ~40 nt.

The influence of local secondary structure on priming efficiency. We assumed the following link between priming efficiency and secondary structure. A segment of the template can theoretically be folded in many different ways. Each folding is associated with a free energy value. The lower that energy value the higher the proportion of template molecules folded that way.

We refer to the lowest free energy of all possible foldings as the *best folding* energy, because for thermodynamic reasons the template molecules tend to assume this best folding. Not all of the possible foldings are compatible with priming by a short oligonucleotide at a particular priming site. For successful priming, the region that includes the priming site and several bases immediately downstream of the primer should not be involved in secondary structure (Fig. 3a–c). (We refer to this region as the *clean site*.) Nor should the priming site be part of a single-stranded loop shorter than a certain length (referred to as the *minimal loop*) (Fig. 3d). A folding that obeys these rules is referred to as a *constrained* folding (constrained by the priming requirements). A particular constrained folding that provides the lowest free energy among all the constrained foldings is referred to below as the *best constrained folding*. Thus the priming requires the following change in free energy:

$$\Delta\Delta G = \Delta G_{\text{constrained}} - \Delta G_{\text{best}} + \Delta G_{\text{annealing}} \qquad \mathbf{1}$$

where $\Delta G_{\text{compatible}}$ is the free energy of the best constrained folding (compatible with priming), ΔG_{best} is that of the best folding (unconstrained) and $\Delta G_{\text{annealing}}$ is that of primer–template annealing. The larger the $\Delta\Delta G$, the lower the fraction of template molecules that allows annealing of the primer. Consequently, this free energy difference should reflect the strength of the priming reaction at a particular site. The actual character of the dependence of the priming efficiency on $\Delta\Delta G$ (e.g. linear, exponential) does not affect the predicted ranking of priming efficiencies at different sites.

Initial values. We started the optimization process from the following initial thermodynamic parameter values. The stability of double helical regions was calculated using the nearest neighbor model, with thermodynamic parameter values for the 10 dinucleotides from Breslauer *et al*. (1). For single-stranded loops (internal, bulge, hairpin and multibranched), we used the loop values for RNA from Jaeger *et al*. (13) (which gave results similar to 15), disregarding the sequence dependence of the loop energies. Also, no corrections were made for asymmetric loops. Loop closure was assumed to have enthalpy $\Delta H = 0$, as in Jaeger *et al*. (13). Thus, ΔG values for the loops were recalculated from 37°C, as in Jaeger *et al*. (13), to our 20°C by multiplying by the ratio of the absolute temperatures, 293/310. All of the possible mismatches were assigned the same initial value of energy: the energy value of an internal RNA loop of two bases in length, one base in each strand.

The objective function for optimization of parameters. For each priming site we calculated $\Delta\Delta G = \Delta G_{\text{constrained}} - \Delta G_{\text{best}}$, where ΔG_{best} is the lowest free energy for the folded unconstrained template and $\Delta G_{\text{constrained}}$ is that for the constrained one. Here we dropped the last term of equation **1** because we assumed that primer–template binding energy of a particular primer was the same at all its annealing sites. Thus, we neglected the slight variance due to the two degenerate bases in the primer. Indeed, we found no correlation between the template bases matching the degenerate positions and the priming efficiency at different sites of the same primer. In other words, the two degenerate positions make the primer a mixture of 16 sequences, but the differences between the sequences appear to have negligible effect on the priming efficiency when compared with that of the secondary structure of the template.

Having obtained both predicted and actual rankings of a given set of priming sites (for the same degenerate 8mer primer), we then counted the number of misranked site pairs (whose predicted intra-pair rank order was different from the experimentally measured one). Our *objective function* equaled this number (of wrongly predicted site pairs) and hence measured how well the values of the parameters were optimized. The aim of the optimization was to find a set of parameter values that minimized the objective function. We optimized the non-thermodynamic parameters first and then the thermodynamic ones.

Non-thermodynamic parameters. The non-thermodynamic parameters included the clean site, the shift of the 3′-end of the clean site relative to the 3′-end of the primer and the minimal loop length. To optimize them we fixed all the thermodynamic parameters at the initial values (described above). Since the non-thermodynamic parameters are integers, we exhaustively scanned all reasonable combinations and found the exact optimal values. The values for the clean site, the shift of the 3′-end of the clean site relative to the 3′-end of the primer and the minimal loop length have been found to be 16, 8 and 40 nt, respectively. In all subsequent optimizations of the thermodynamic parameters, the non-thermodynamic parameters were fixed at these values.

Thermodynamic parameters. We used the values of base stacking and base pairing energies known from the nearest neighbor model studies (1) without optimizing them. The thermodynamic parameters that we optimized were free energies for single-stranded DNA loops and for mismatches in the double helical portions. The optimization was performed using the 95 ranked pairs of priming sites in the training set. To find the optimal values of the thermodynamic parameters, we varied one parameter at a time, fixed it at the best value and then varied the next one, eventually returning to the first parameter. We continued that process iteratively until no further improvement in the objective function could be reached. After optimization, the number of misrankings (wrongly ranked pairs of sites) dropped from 31 (with initial parameters) to 15 (with optimized parameters). The following thermodynamic parameters were optimized: single-stranded loops, penalty for the ends of double helices, and mismatches.

Single-stranded loops. To reduce the overall number of loop parameters, we assumed that, in general, the behavior of DNA loops is similar to that of RNA loops. Therefore, because our initial values for loop parameters were based on RNA studies, we used them as a skeleton and optimized only a relatively small set of corrections. The correction parameters we defined were the weights of small (<10 nt long) and large (≥10 nt long) loops, denoted $CONST_{small}$ and $CONST_{large}$, respectively. While doing this, we assumed for simplification that the DNA parameters for loops within each of the two classes were, at the first approximation, proportional to those known for RNA (13) (averaged over various sequences). Therefore, the energy of a small loop was calculated as:

$$G_{small} = CONST_{small} \times G_{RNA} \quad \mathbf{2}$$

where G_{small} is the energy of a small DNA loop and G_{RNA} is the energy of the corresponding RNA loop. The energy of a large loop was calculated as:

$$G_{large} = G_9 + CONST_{large} RT \ln(L_{large}/9) \quad \mathbf{3}$$

where G_{large} is the energy of DNA loop of length L_{large}, G_9 is the energy of a 9 nt-long RNA loop, T is temperature and R is the gas constant. These factors ($CONST_{small}$ and $CONST_{large}$) were applied to all the loop types, internal, hairpins, bulges and multibranched, without distinction. The optimized parameters were found to be $CONST_{small} = 1.1 \pm 0.05$ and $CONST_{large} = 1.9 \pm 0.1$.

Penalty for the ends of double helices. Another parameter that can be regarded as one of the loop parameters is a constant penalty for every double helix end. Initially it was assumed to be zero and upon optimization was found to be –0.05 ± 0.01 kcal/mol.

Mismatches. We assigned a separate parameter to the energy of each of the eight possible mismatches (A*A, A*C, A*G, C*C, C*T, G*G, G*T and T*T). The optimized destabilization energies in kcal/mol for the mismatches were found to be: A*A = 2.5 (undefined upper limit, lower limit = 2.3); A*C = –0.6 ± 0.05; A*G = 2.5 (undefined upper limit, lower limit = 2.4); C*C = –0.2 ± 0.05; C*T = 1.1 ± 0.3; G*G = 2.1 (undefined upper limit, lower limit = 2.0); G*T = –0.7 ± 0.1; T*T = 0.4 ± 0.05.

Error margins. The error margins for the obtained parameters were assessed by varying each parameter (one at a time) in both directions and defining the end of the error bar as the point at which the objective function grew by one. For some mismatches one of the two error margins remained undefined (open from one of the two sides). In other words, no change in the objective function was found for A*A, A*G and G*G varying in one direction, probably because of under-representation of the corresponding mismatches in the training set.

Test set. To test the validity of the optimized parameter values, we performed an additional set of reactions (using different primers and priming sites) that we called the test set. The test set included eight primers, which yielded 147 experimentally ordered pairs of priming sites. The primers had the following sequences (5′→3′): NNGAATAA, NNGTGAAT, NNCAGAGC, NNCAAAAG, NNCCACCA, NNCCAGTA, NNGAAAGG and NNCACCAG. Comparison of the computer predicted ranking of the test set sites with the ranking obtained experimentally revealed 56 pairs of sites with wrongly predicted intra-pair ranks (misrankings) when using the initial, unoptimized thermodynamic parameters, and 41 misrankings when using the parameters optimized for DNA (an example is given in Table 3). (In both cases we used the optimized values for the non-thermodynamic parameters listed above, because there were no initial values for them.)

To assess the statistical significance of this result, we wrote a program that assigned $\Delta\Delta G$ randomly (by a random number generator) to all the sites. Running this program 1 000 000 times revealed that:

- the average number of misrankings was 73.5, which is, as expected, 50% of the overall number of pairs;
- out of the 1 000 000 runs, 3×10^4 times the number of misrankings was ≤56 (the number of misrankings when our initial values were used for unknown parameters);
- out of the 1 000 000 runs, 175 times the number of misrankings was ≤41 (the number of misrankings after our optimization for DNA), thus the probability that pure chance caused the test set results to be as good as they were is 175 out of 1 000 000, or $\sim 2 \times 10^{-4}$.

Table 3. Measured and predicted (calculated) rankings of the efficiencies of the seven priming sites of the primer 5′-NNGAATAA-3′, with initial (left) and optimized (right) folding parameters

Predicted ranking with the *initial* thermodynamic parameters yields four misrankings: 3115*1591, 4164*1591, 3849*2758 and 4488*2758			Predicted ranking with the *optimized* thermodynamic parameters yields two misrankings: 3115*1591 and 3115*4164		
Calculated $\Delta\Delta G$ (kcal/mol)	Rank based on gel band brightness	Position in M13 template	Calculated $\Delta\Delta G$ (kcal/mol)	Rank based on gel band brightness	Position in M13 template
0	1	3115	0	1	3115
0.2	1	4164	0.9	0	1591
1.8	0	1591	2.1	0	4164
1.9	3	3849	2.5	2	2758
2.2	4	4488	2.7	3	3849
4.0	2	2758	5.9	4	4488
5.3	4	917	8.3	4	917

For this primer, the optimization of the parameters decreased the number of misrankings from four to two. Misranking is a pair of sites with wrongly predicted intra-pair order of their priming efficiencies. The priming sites are listed in the order of predicted efficiency (the smaller the calculated $\Delta\Delta G$ value, the higher the predicted efficiency).

Success predictions for DENS sequencing reactions

The DENS sequencing technique (11) is based on initial extension (at 20–30°C) of a short primer by a DNA polymerase with only two out of the four possible dNTPs present in the reaction mix. The primer (e.g. a partially degenerate 8mer, complementary to several sites in the template) is too short to prime uniquely. In the presence of only two dNTPs, the polymerase extends the 8mer by several bases until it encounters a template base that cannot form a base pair with either of the two available dNTPs. Therefore, at different priming sites, the same primer is extended to different lengths as determined by the template sequence, thus making the extension ‘differential’. DENS requires the freedom to choose the intended priming site within a span of dozens of bases (as in primer walking, where the last 100–200 bases of the previous sequence run are available for placing the primer). This freedom is used to choose both the intended priming site and the two-dNTP subset so as to maximize the extension length at that site. In contrast, alternative priming sites are located in the template randomly and therefore at these sites the differential extensions are likely to be substantially shorter than at the intended site (with the selected two-dNTP subset). This procedure is repeated using thermocycling with a thermostable polymerase.

A subsequent higher temperature termination reaction is thermocycled with all four dNTPs present, similarly to regular cycle sequencing. The annealing/extension temperature of the termination stage (usually 60–65°C) is selected so as to allow the product of the differential extension at the intended site to be further extended. In contrast, the differential extension products of the same primer at alternative sites are shorter than at the intended one and thus are unlikely to anneal and be extended, because most of them are shorter than the threshold (5 base-long extension) imposed by the temperature (Fig. 4).

Our computer program for predicting the efficiencies of priming sites was tested on 91 sequencing reactions performed on M13mp18 and on chicken virus inserts cloned in Bluescript vectors using the DENS technique with fluorescent dye terminators (11). The sensitivity of our sequencing machines (ABI-373) allowed detection of strong reactions only, thus separating all the reactions into two groups: detectable (strong enough) and undetectable (too weak).

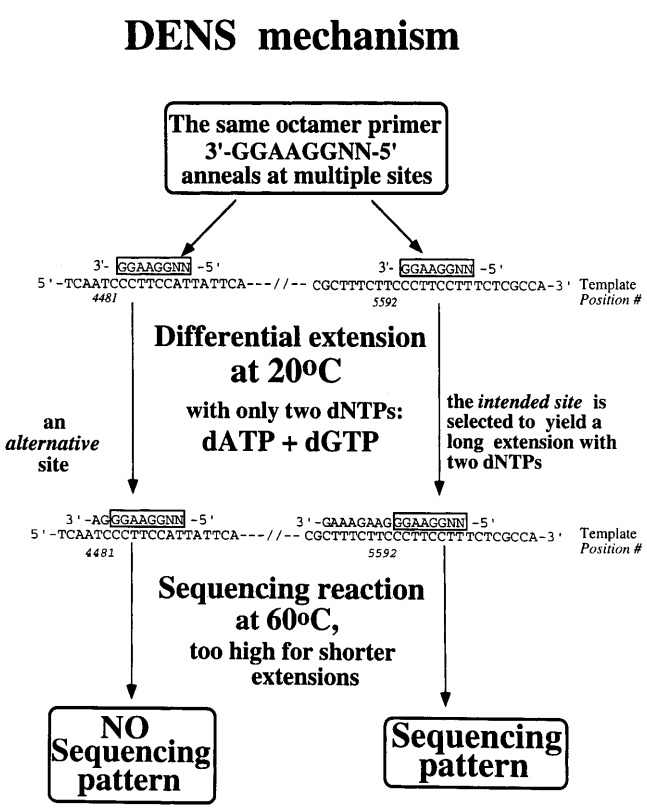

Figure 4. Flow-chart of the DENS sequencing technique (11) illustrating the mechanism of specific priming by an otherwise non-specific 8mer primer. The 8mer shown here has five complementary sites in M13mp18 single-stranded template, and without DENS gives an unreadable sequence pattern. In DENS, only two (out of the four possible) dNTPs are used for the initial ‘differential’ extension step. If the dNTP subset is A+G, then this 8mer can be extended at two positions only, 4481 and 5592, by 2 and 8 nt, respectively. Of these two products, 10 and 16 bases long, only the latter (position 5592) is long enough to prime in the subsequent termination reaction at 60°C. This discrimination by temperature makes the intended priming site unique (position 5592). N stands for a fully degenerate position (A+C+G+T).

The first step in the DENS reaction is the differential extension of an 8mer primer; low priming efficiency at this step is the main reason for weak reactions. Assuming that some first step reactions were weak because of unfavorable secondary structure (i.e. relatively high $\Delta\Delta G$), we found a threshold for $\Delta\Delta G$ above which the reactions were predicted to fail. Thus, the reactions were divided into two groups: predicted to work ($\Delta\Delta G$ below the threshold) and predicted to fail ($\Delta\Delta G$ above the threshold).

Table 1 represents the results of dividing by the $\Delta\Delta G$ threshold compared with the experimental data. It can be seen that by selecting the priming sites predicted to be good (e.g. 43 sites below the threshold of $\Delta\Delta G = -13$ kcal/mol), the rate of success can be raised from 67 to 91%. We performed statistical analysis and estimated that the probability that this result could be obtained by chance is $P < 7 \times 10^{-5}$ (probability calculation method in Materials and Methods). The sequencing failures were due to either weak priming or alternative priming sites. Our predictions deal only with the former cause, thus even a hypothetical 100% accurate $\Delta\Delta G$ ranking of the sites would not translate into a 100% agreement with sequencing results.

DISCUSSION

The method reported here predicts the site-specific efficiency of 8mer priming by computing the change in the free energy of the local secondary structures of the template associated with the priming. We have successfully applied the method to select 8mer primers for DENS sequencing (raising the success rate from 67 to 91%, with a statistical significance of $P < 7 \times 10^{-5}$) and anticipate that the method can probably be applied to other sequencing techniques that are based on short primers. Conversely, to optimize the energy parameters of DNA folding, we used the priming efficiency ranking of an 8mer at different sites in the same template, minimizing the discrepancy between the computed and the actual efficiency rankings. This approach dramatically reduces the benchwork required to estimate the parameters involved in calculating DNA folding. The priming efficiencies of a degenerate 8mer at different sites are compared in the same reaction, on the same template, on the same PAGE gel and in the same lane. Such uniformity minimizes experimental variations and provides informative data from a single experiment. A typical experiment can be accomplished within ~30 min plus a 2 h PAGE. This technique can yield values for folding parameters for a variety of conditions and oligonucleotide modifications.

The priming efficiency ranking of several sites per primer turned out to be a highly informative set of data. For example, eight sites can be ranked in 8! = 40 320 ways and only one of these rankings is found in the actual experiment. These rankings are much more informative than conventional melting data obtained from an experiment of a similar size for deducing thermodynamic parameters of DNA folding. It remains to be seen whether another approach, such as informative pause data (16), can be as efficient for deducing DNA secondary structure parameters.

Our method is based on the free energy of folding rather than on the folding structure and that makes it relatively robust. Indeed, there is great uncertainty about the secondary structure; many very different suboptimal structures have free energies that are virtually the same as that of the optimal structure. Fortunately, we use only the free energy value here, while the structure difference between optimal and suboptimal foldings need not be considered.

We calculated the secondary structure using a static equilibrium model, even though the process is generally considered kinetic. Support for the static model comes from the observation that the relative priming efficiencies were not affected by the concentration of the primer. In experiments not described in this paper, we varied the concentration of the 8mer primer by more than two orders of magnitude without observing any change in relative priming efficiencies.

As a starting point for the optimization we took Breslauer's nearest neighbor parameters for base stacking (1) and RNA parameters for loops (13). We found them to provide better agreement between the predicted and actual rankings of priming sites than did other sets of parameters (2,4), when checked on the 'training set'.

Minimizing a non-linear objective function of a multidimensional parameter set is known to be problematic and researchers have taken different approaches to this ubiquitous problem. Due to the long computation time required to calculate our objective function (~15 min for each point in the parameter space using a DEC 564 alpha computer), we limited the size of the training set and the number of parameters to be optimized. For this reason, our model in this work was very simplistic and unsophisticated. Thus, for mismatches we had to neglect the effect of the flanking sequences (nearest neighbor influence), even though this effect can be significant (5–10). Therefore, the energy values of mismatches that we have optimized can be regarded only as averaged over a spectrum of the nearest neighbors that may be biased in our sample towards more represented neighbors. The rudimentary nature of the model we used and the small size of the training set data make the accuracy of the obtained results difficult to estimate.

Our training set was not large enough to provide tight error margins for all the mismatches. The occurrence of specific mismatches is less common than that of loops. Therefore, unlike mismatches, the loop stability parameters $CONST_{small}$, $CONST_{large}$ and the duplex end penalty have well-defined margins. In contrast to the loop parameters, we could not find the upper margins for some of the mismatches (A*A, A*G and G*G). These results indicate that, while our DNA loop parameters can be considered reliable, many of our mismatch energies cannot. An interesting question is whether these results also indicate that the loop parameters are more critical than mismatches for predicting priming efficiency and, possibly, for computing secondary structure.

Our free energy values for G*T and C*T mismatches are within the range reported in the literature. For G*T our free energy is –0.7 kcal/mol, compared with +1.05 to –1.05 kcal/mol (depending on the neighbors) reported in Allawi and SantaLucia (8). For C*T our free energy is 1.1 kcal/mol, compared with +1.02 to –1.95 kcal/mol (depending on the neighbors) reported in Allawi and SantaLucia (10).

The free energy we obtained for the A*G mismatch was 2.5 kcal/mol and seems too large when compared with the +1.16 to –0.78 kcal/mol reported in Allawi and SantaLucia (9). Partial data about the other mismatches can be found in Aboul-ela *et al.* (5), Werntges *et al.* (6) and Gaffney *et al.* (7). The data in these papers were obtained for very few (out of the 16 possible) neighbors and thus may be biased. This bias could be one of the reasons for discrepancies between these papers and our results.

Other possible reasons for discrepancies between our mismatch free energies and those in the literature include the following.

- Our training set is not big enough to represent equally all the neighbors of each mismatch, thus the obtained parameters may be biased toward the most represented combinations.
- In mismatch studies reported in the literature, internal mismatches were usually measured inside relatively long (>10 nt) double helices. In contrast, we deal with DNA with sequences close to random; where most double helical segments in the DNA folding are much shorter. The structure of base arrangement can be different in short duplexes (as compared with long ones) and so can the free energies of mismatches.
- Most studies used buffers containing Na^+, a monovalent cation, whereas we used Mg^{2+}, which is divalent. The behavior of base-pairing thermodynamics in the presence of divalent cations is likely to be more complicated (3).
- A different ionic strength. Our 2.0 mM Mg^{2+} is equivalent to 0.18 M Na^+ (Materials and Methods), which is much lower than the 1.0 M Na^+ used in most studies.
- The minimum that we found for our objective function may be local rather than global. A more sophisticated minimization technique might find a better solution.

Still, the statistical significance of the validation test results (test set and DENS predictions) shows the robustness of the approach. The reactions using the test set (independent from the training one) have shown that the optimized parameter values predict the priming efficiency much better than the initial ones, which were partially borrowed from RNA studies ($P < 2 \times 10^{-4}$). Moreover, the optimized values turned out to be useful in practice for predicting the primer success ($P < 7 \times 10^{-5}$ in Table 1) in the DENS sequencing technique and thus can be used for DENS primer selection. These two independent validations of the optimized parameter values indicate that our experimental approach may be quite useful in spite of its simplicity. It cannot be ruled out that the folding of single-stranded DNA of natural sequence under physiological conditions (e.g. in DNA sequencing) is better predicted by parameter values found using our approach (perhaps refined in the future) than by parameters found in artificially designed (long) duplexes under artificial conditions, as in most previous studies.

Our algorithm can be adjusted for parallel computing (which we did not use), in which case the CPU time bottleneck should be eased by the use of a supercomputer. With fully parallel computation, the program can be speeded up by a factor of 8000 or so, reducing the required time from 15 min to ~0.1 s. This speed should allow use of a training set large enough to optimize a more complete set of parameters within a more sophisticated model and more advanced minimization algorithms to avoid local minima and find more precisely optimized values for the parameters.

ACKNOWLEDGEMENTS

We thank Edward Trifonov for valuable comments and Maura Devine and Cathy Kaicher for help in editing. This work was supported by DOE grant no. DE-FG02-94ER61831 and DOE contract no. W-31-109-ENG-38.

REFERENCES

1 Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
2 Delcourt,S.G. and Blake,R.D. (1991) *J. Biol. Chem.*, **266**, 15160–15169.
3 SantaLucia,J.,Jr, Allawi,H.T. and Seneviratne,A.P. (1996) *Biochemistry*, **35**, 3555–3562.
4 SantaLucia,J.,Jr (1998) *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
5 Aboul-ela,F., Koh,D., Tinico,I.,Jr and Martin,F.H. (1985) *Nucleic Acids Res.*, **13**, 4811–4823.
6 Werntges,H., Steger,G., Riesner,D. and Fritz,H.J. (1986) *Nucleic Acids Res.*, **14**, 3773–3791.
7 Gaffney,B.,L. and Jones,R.A. (1989) *Biochemistry*, **28**, 5881–5889.
8 Allawi,H.,T. and SantaLucia,J.,Jr (1996) *Biochemistry*, **36**, 10581–10594.
9 Allawi,H.,T. and SantaLucia,J.,Jr (1998) *Biochemistry*, **37**, 2170–2179.
10 Allawi,H.,T. and SantaLucia,J.,Jr (1998) *Nucleic Acids Res.*, **26**, 2694–2701.
11 Raja,M.C., Zevin-Sonkin,D., Shwartzburd,J., Rozovskaya,T.A., Sobolev,I.A., Chertkov,O., Ramanathan,V., Lvovsky,L. and Ulanovsky,L.E. (1997) *Nucleic Acids Res.*, **25**, 800–805.
12 Zuker,M. (1989) In Waterman,M.S. (ed.), *Mathematical Methods in DNA Sequences*. CRC Press, Boca Raton, FL, pp. 159–184.
13 Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 7706–7710.
14 Zuker,M. and Stieger,P. (1981) *Nucleic Acids Res.*, **9**, 133–201.
15 Walter,A.E., Turner,D.H., Kim,J., Lyttle,H.M., Muller,P., Mathews,D.H. and Zuker,M. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
16 Reckmann,B., Grosse,F., Urbanke,C., Frank,R., Blocker,H. and Krauss,G. (1985) *Eur. J. Biochem.*, **152**, 633–643.