International Journal of
Molecular Sciences

MDPI

Article

Needles in Haystacks: Understanding the Success of Selective Pairing of Nucleic Acids

Carlos A. Plata [1,2,†], Stefano Marni [3,†], Samir Suweis [1,4], Tommaso Bellini [3] and Elvezia Maria Paraboschi [5,6,*]

1 Dipartimento di Fisica 'G. Galilei', INFN, Università di Padova, Via Marzolo 8, 35131 Padova, Italy; cplata1@us.es (C.A.P.); samir.suweis@unipd.it (S.S.)
2 Física Teórica, Universidad de Sevilla, Apartado de Correos 1065, 41080 Sevilla, Spain
3 Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università di Milano, Via Fratelli Cervi 93, 20054 Segrate, Italy; stefanomarni@gmail.com (S.M.); tommaso.bellini@unimi.it (T.B.)
4 Padova Neuroscience Center, Università di Padova, Via Giuseppe Orus 2, 35131 Padova, Italy
5 Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20072 Pieve Emanuele, Italy
6 IRCCS Humanitas Research Hospital, Via Manzoni 56, 20089 Rozzano, Italy
* Correspondence: elvezia_maria.paraboschi@hunimed.eu
† These authors contributed equally to this work.

Abstract: The effectiveness of several biological and biotechnological processes relies on the remarkably selective pairing of nucleic acids in contexts of molecular complexity. Relevant examples are the on-target binding of primers in genomic PCR and the regulatory efficacy of microRNA via binding on the transcriptome. Here, we propose a statistical framework that enables us to describe and understand such selectivity by means of a model that is extremely cheap from a computational point of view. By re-parametrizing the hybridization thermodynamics on three classes of base pairing errors, we find a convenient way to obtain the free energy of pairwise interactions between nucleic acids. We thus evaluate the hybridization statistics of a given oligonucleotide within a large number of competitive sites that we assume to be random, and we compute the probability of on-target binding. We apply our strategy to PCR amplification and microRNA-based gene regulation, shedding new light on their selectivity. In particular, we show the relevance of the defectless pairing of 3′ terminals imposed by the polymerase in PCR selection. We also evaluate the selectivity afforded by the microRNA seed region, thus quantifying the extra contributions given by mechanisms beyond pairing statistics.

Keywords: nucleic acid interactions; pairing statistics; stat-mech modeling

Citation: Plata, C.A.; Marni, S.; Suweis, S.; Bellini, T.; Paraboschi, E.M. Needles in Haystacks: Understanding the Success of Selective Pairing of Nucleic Acids. *Int. J. Mol. Sci.* **2022**, *23*, 3072. https://doi.org/10.3390/ijms23063072

Academic Editor: Jesus Vicente De Julián Ortiz

Received: 10 February 2022
Accepted: 9 March 2022
Published: 12 March 2022

1. Introduction

The selective pairing of nucleic acids is the key molecular property enabling genetic coding, gene expression and regulation, and heredity transmission. The extent of such selectivity becomes evident in processes in which complementary strands have to selectively pair amid a plethora of other nucleic acid polymers and oligomers. Relevant examples of such a successful "needle in the haystack" search performed by nucleic acids can be found in both biological and technological contexts. For instance, in the biological context, microRNA (miRNA) play a key role in gene expression and regulation. miRNA are short RNA molecules (∼22 nt, where nt stands for nucleotides) typically targeting specific messenger RNAs (mRNA) among the molecular variety present in the cytoplasm, inducing mRNA degradation or translation halting. In the technological context, polymerase chain reaction (PCR) is the most used technique in molecular biology, allowing the exponential amplification of target DNA/cDNA regions thanks to the selective pairing between oligonucleotide primers and entire genomes/transcriptomes. In both cases, one short oligomer (of the

order of 20 nt) has to search and find its complementary counterpart within much longer polymers (e.g., $\sim 10^9$ nt).

Regarding the PCR technique, since Mullis' first publication, the primer length considered effective in PCR was in the range of 20–27 nt [1]. A simple statistical consideration is to evaluate permutations in a strand of length L and compare it to the total length L_0 of the analyzed genome [2]. When $L = 20$, the possible permutation of nucleobases is around 10^{12}, much more than the length of the human genome (around 3×10^9). However, this simple evaluation does not take into account the possibility of forming defected pairings, which is the most relevant form of potential failure in selective targeting. In its current use, primer design is optimized through the use of algorithms that allow us to control for GC content, secondary structure, or internal complementary regions [2].

On the other hand, for miRNA selectivity, the mechanism of action has different layers of complexity. First, miRNAs in cells function within a ribonucleoprotein complex called the RNA-induced silencing complex (RISC). The formation of the mature miRNA–RISC complex is not trivial, and requires the maturation of the miRNA molecule, the association with Argonaute (AGO) proteins, and the selection of the guide strand that takes the RISC to the target mRNAs, usually in its 3′ untranslated region (3′ UTR) [3]. Moreover, although the length of mature miRNAs is ∼22 nt, the "active" region, called the "seed", is only 6–8 nt long [4]. Generally, the seed corresponds to nucleotides 2–7, and it is considered the minimal element to bind and repress mRNA translation potential. This length must have been optimized by nature as a compromise between selectivity on the one hand and fast diffusion and accessibility to the target on the other. Despite the seed being recognized as a critical element in the miRNA mechanism, growing evidence indicates that sequences in the miRNA 3′-end play an important role in mRNA targeting [3]. Interestingly, structural studies have shown that, once the miRNA forms a complex with AGO proteins, only the seed is available to interact with the target site [5]. However, the binding of the miRNA–RISC complex to a target RNA induces a conformational change that unmasks the 3′ end of the miRNA, allowing further pairing outside the seed region [3,5], which can impact the specificity of targeting, the regulatory mechanism, and the stability of the miRNA itself. Finally, the presence of a mRNA–seed (or extended) pairing is not the only determinant of miRNA successful activity. In fact, mRNAs in cells tend to form secondary structures, and to interact with RNA binding proteins, which can limit miRNA accessibility to the target. Site accessibility was demonstrated to be a key feature for miRNA-mediated translational repression: functional miRNA target sites are preferentially located in highly accessible regions, and this feature is conserved across genomes [6]. These notions need to be taken into account in the estimate of the total amount L_0 of sites on which miRNA may bind in competition to its targets.

In spite of the differences and the complexity of the selectivity processes described above, they are both rooted in the selectivity of interactions between nucleic acids. A natural question thus arising from these remarkably successful examples of selectivity is how to model and understand these phenomena on the basis of the well-known thermodynamics of nucleic acid duplex formation [7]. Here, we tackle this problem by elaborating on a re-parametrization on three classes of base pairing error guided by the description of hybridization thermodynamics from the so-called "nearest-neighbor model" [8]. We then develop a mean field method to calculate the probability for the formation of perfect and defected duplexes in these two contexts. In particular, we focus on exploring the effect of the oligomer (primer and miRNA) length L in the efficiency of targeting their cognate sites within long random sequences, gaining new insights into the factors at play in both situations. The dependences on other relevant parameters such as the temperature are also analyzed.

2. Materials and Methods

Our strategy relies on the comparison between the Boltzmann statistical weights for on-target and off-target pairings in order to evaluate the success probability of the process. In the miRNA case, we study the pairing of the miRNA–RISC complex to the mRNA, where mainly the nucleotides within the "seed" region are available for Watson–Crick interactions; on the other hand, we consider the first annealing cycle of the PCR, being the most significant for the success of the technique.

In the following subsections, we present the main ingredients for our physical statistical description of the PCR technique and miRNA gene expression regulation: firstly, we are able to obtain the average free energy of a certain quality of duplex, thanks to a parametrization of the pairing depending on the kinds of mismatches involved. Secondly, the same parametrization allows us to obtain the degeneracy of each kind of duplex, i.e., the total number of sequences with which the primer/miRNA can realize a duplex with the same combination of mismatched bases. For our purposes of general validity of the results, we neglect the sequence specificity of the genomic ssDNA or of the mRNA and we consider them as random sequences, where the 4 nitrogenous bases are equiprobable in each nucleotide of the off-target sites. Finally, we have combined the binding free energy and the degeneracy to compute the Boltzmann weight of the on-target and off-target pairings. Comparing these two terms, we obtain the on-target pairing probability.

2.1. Free Energy for Duplex Formation

Differently from other works on DNA hybridization focusing on the prediction of stable pairings as a function of the temperature, i.e., the study of "melting curves" [9–11], we would like to characterize here the probability of on-target binding of oligonucleotides in the presence of huge numbers of random competitive sites. To do so, we have to describe the binding free energy between any given pair of interacting oligomers, as well as the degeneracy of their potential pairing.

The free energy difference ΔG between a nucleic acid duplex and its free constituent sequences can be split into an enthalpic and an entropic part,

$$\Delta G = \Delta H - T\Delta S. \tag{1}$$

Nevertheless, providing an accurate description of such thermodynamic parameters characterizing the interaction between nucleic acids is not an easy task. In the highly cited review by SantaLucia and Hicks [12], detailed energetic data for several DNA motifs can be found, comprising canonical Watson–Crick pairing and a long catalog of errors, including internal mismatches, terminal mismatches, terminal dangling ends, hairpins, bulges, internal loops, and multibranched loops. The extraction of such thermodynamic parameters, however, necessarily requires the knowledge of the specific bases composing the two strings, and this is information that is not possible to access typically, or it is simply unfeasible to compute when dealing with a multitude of random possible competing pairs. Moreover, since our aim is to unveil some fundamental properties based on thermodynamic arguments with a coarse-grained modeling to explain the effectiveness of selective bindings in nucleic acids, we consider that such properties do not depend on fine details such as the specific bases composing the interacting oligomers. This hypothesis is checked for specific cases (see Appendix A, and Appendix A.5 in particular), proving the robustness and range of applicability of our description. Remarkably, our assumption of two states, i.e., on–off hybridization with no intermediate state between unbound and paired, is justified for short oligomers [13], as it is in the cases considered in this study.

For these reasons, we develop here an effective energetic model that, by considering only three classes of base pairing errors and through a "mean field" approach where all possible combinations of interacting pairs are averaged, yields a simplified but yet quantitatively fair description of DNA (or RNA) hybridization.

For the sake of concreteness, let us focus on the pairing between a generic primer (an oligomer with length L) and a long polymer with length $L_0 \gg L$. Specifically, L_0 is measuring the number of ways in which the first oligomer can couple to the latter (number of sites wherein it can attach). Once L is defined, in our description, the duplex is fully characterized through a three-component parameter vector $\vec{\alpha} = (\alpha_{e1}, \alpha_{e2}, \alpha_i)$. This vector carries the information of the number of external mismatches, α_{e1} and α_{e2}, and internal mismatches, α_i. The definition of $\vec{\alpha}$ thus consists of re-parametrizing the hybridization thermodynamics on three classes of base pairing errors. To this aim, we split the total enthalpy and entropy into different contributions stemming from the different interactions involved in the duplex,

$$\Delta H(L, \vec{\alpha}) = \Delta H_{\text{perf}}(L) + \Delta H_{\text{dang}}(\alpha_{e1}, \alpha_{e2}) + \Delta H_{\text{int}}(\alpha_i), \quad (2)$$

$$\Delta S(L, \vec{\alpha}) = \Delta S_{\text{perf}}(L) + \Delta S_{\text{dang}}(\alpha_{e1}, \alpha_{e2}) + \Delta S_{\text{int}}(\alpha_i) + \Delta S_{\text{salt}}\left(L, \vec{\alpha}, [Na^+]\right). \quad (3)$$

Above, we have separated the contributions from the perfect match, the dangling ends, and internal mismatches. Note that entropy is additionally corrected due to salt concentration $[Na^+]$ [14]. In the following subsections, we account for each contribution in detail.

2.1.1. Perfect Match: Initiation and Nearest-Neighbor Canonical Base Pairs

Our starting point is the contribution of an ideal matched duplex. The nearest-neighbor model has been proven to provide a very good description for the enthalpy and entropy of duplexes [12]. This model starts from initiation values ΔH_0 and ΔS_0, which are complemented by additive contributions coming from each couple of neighboring base pairs. Such contributions depend on the specific bases considered. Nevertheless, in our coarse-grained description, we associate a single averaged contribution ΔH_{n-n} and ΔS_{n-n} to any couple of neighboring matched base pairs (see Appendix A along with Table A1 therein for further details on the averaging). Therefore, in our framework, the enthalpy and entropy of perfectly matched duplexes depend solely on the length L and simply read

$$\Delta H_{\text{perf}}(L) = \Delta H_0 + (L-1)\Delta H_{n-n}, \quad (4)$$

$$\Delta S_{\text{perf}}(L) = \Delta S_0 + (L-1)\Delta S_{n-n}, \quad (5)$$

where we have taken into account that the number of couples of neighboring base pairs is $L-1$.

2.1.2. Dangling Ends: External Mismatches

This contribution takes into account that the duplex may happen with a certain external mismatched base pair. Moreover, if the external base is well paired, there is a stacking contribution to the free energy, due to the base of the long polymer that is next to the pair.

The number of external mismatches in each end is given by α_{e1} and α_{e2}, respectively. Note that, in order to obtain at least one matched base pair, we need to enforce $\alpha_{e1} + \alpha_{e2} \leq L - 1$ (see Figure 1). Our work hypothesis, motivated by the values typically found [12], is that external mismatches can be thought of as two dangling bases at the same end. Therefore, due to external mismatches, (*i*) $\alpha_{e1} + \alpha_{e2}$ neighboring base pairs are canceled out with respect to the perfect match and (*ii*) there is an extra contribution stemming from the first bases within a dangling end. Although, in reality, this contribution would depend on the identity of the bases, we consider an averaged contribution ΔH_d and ΔS_d to any dangling end (see Appendix A along with Table A2 therein for further insight on these values).

Therefore, by summing up the previous discussion, the contribution of external mismatches can be parametrized as follows:

$$\Delta H_{\text{dang}}(\alpha_{e1}, \alpha_{e2}) = c_d(\alpha_{e1}, \alpha_{e2})\Delta H_d - (\alpha_{e1} + \alpha_{e2})\Delta H_{n-n}, \quad (6)$$

$$\Delta S_{\text{dang}}(\alpha_{e1}, \alpha_{e2}) = c_d(\alpha_{e1}, \alpha_{e2})\Delta S_d - (\alpha_{e1} + \alpha_{e2})\Delta S_{n-n}, \quad (7)$$

where $c_d(\alpha_{e1}, \alpha_{e2})$ takes the possible values $\{2, 3, 4\}$ depending on the external mismatches

$$c_d(\alpha_{e1}, \alpha_{e2}) = \begin{cases} 2 & \text{if } \alpha_{e1} = \alpha_{e2} = 0 \\ 3 & \text{if } \alpha_{e1} + \alpha_{e2} > 0 \\ & \text{and } \alpha_{e1}\alpha_{e2} = 0, \\ 4 & \text{if } \alpha_{e1}\alpha_{e2} > 0, \end{cases} \quad (8)$$

corresponding, respectively, to dangling ends without external mismatches, dangling ends plus external mismatches in one end, and dangling and external mismatches in both ends. When writing the cases above, we have kept in mind the binding of a primer inside a specific region of a longer DNA as in Figure 1 . Nevertheless, this has to be modified if one is interested in studying selection by miRNA. As described in the Introduction, the active region of miRNA is finite, as represented in Figure 2. Therefore, when considering miRNA, we always assume $c_d = 4$, regardless of the number of external mismatches.

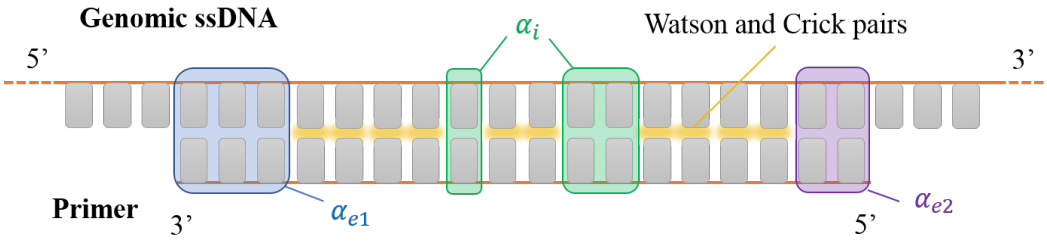

Figure 1. Sketch of a DNA primer interacting with a generic portion of a DNA single strand of a denaturated genome. Gray rectangles represent the nucleobases. Canonical Watson–Crick pairing is marked in yellow. Shaded boxes mark pairing defects: internal mismatches (green shades, counted by α_i), terminal mismatches at the 3′ and 5′ ends (blue shades, α_{e1} and purple shades, α_{e2} respectively). In this sketch, $\alpha_i = 3$, $\alpha_{e1} = 3$ and $\alpha_{e2} = 2$.

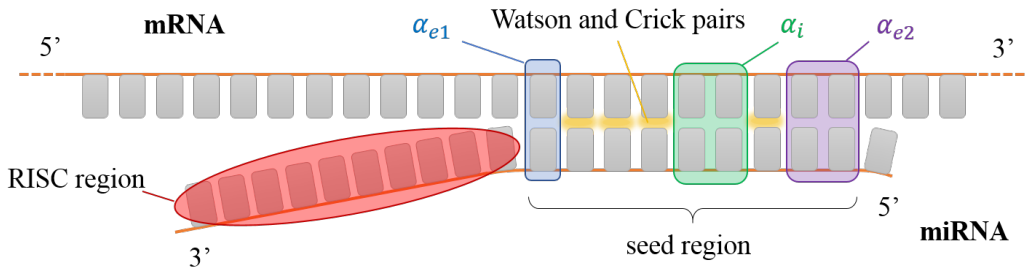

Figure 2. Sketch of a miRNA seed interacting with a generic portion of a mRNA. The nucleobases involved in the interaction with the AGO protein (red shading) are not available for pairing. Colored boxes have the same color code as Figure 1. In this sketch, $\alpha_i = 2$, $\alpha_{e1} = 1$ and $\alpha_{e2} = 2$.

2.1.3. Internal Mismatches

Now, we consider the effect of internal mismatches in the duplex. The integer parameter α_i gives the number of internal mismatches within the duplex. When $\alpha_i > 0$, the set of possible $\vec{\alpha}$ defining a possible duplex, with one matched base pair at least, fulfills the condition $\alpha_{e1} + \alpha_{e2} + \alpha_i \leq L - 2$ (see Figure 1). Besides the corresponding couples of neighboring base pairs that are canceled out, the contribution penalty that stems from

single internal mismatches has been thoroughly studied [12]. This depends on the particular bases. Again, following the philosophy of our coarse-grained approach, we give an averaged contribution ΔH_i and ΔS_i to those eventualities (see Appendix A along with Table A3 therein for further details on the averaging). We assume that such a contribution does not vary when more than one internal mismatch is considered. Moreover, in order to prevent further complexity, we completely neglect the internal structure of the internal mismatches (number, sizes, and separation of adjacent internal mismatches). Specifically, we consider that the effects of additional internal mismatches are equivalent to considering those mismatches to be non-consecutive. Therefore, the thermodynamic parameters associated with internal mismatches are

$$\Delta H_{\text{int}}(\alpha_i) = 2\alpha_i[\Delta H_i - \Delta H_{n-n}], \tag{9}$$

$$\Delta S_{\text{int}}(\alpha_i) = 2\alpha_i[\Delta S_i - \Delta S_{n-n}]. \tag{10}$$

According to our modeling, each internal mismatch replaces two couples of next-neighbor canonical base pairs with two next-neighbor couples of mismatched base pairs. Note that we have carried out a strong approximation in the study of internal mismatches. Nevertheless, since the states with a significant statistical weight are those with low numbers of errors, we can argue that this approximation will not lead to significant errors. The results we present in this work are reasonable and physically sound, corroborating that our assumptions do not seem to misguide our analysis.

2.1.4. Salt Correction

Thermodynamic parameters are computed for a given referential salt concentration, usually 1 M of NaCl. Either excess or a defect of salt, or the presence of other ions, will imply a change in those parameters, affecting mainly the entropic contribution. Salt correction has been studied in detail in the literature [14]. In a nutshell, the most accepted proposals for this contribution assume that ΔS_{salt} is a function of the salt concentration $[Na^+]$, usually through its logarithm. Again, this contribution has a dependence on the specific sequence that we neglect through averaging (see Appendix A for details).

2.1.5. CG Contribution

In order to complement our averaged description, we develop also a more detailed, yet simple, approach that takes into account also the effect of different sequences. Specifically, we assume that the energetic parameters will be a function on the fraction of bases C or G in the DNA sequence, which may change in a significant way the thermal stability of the duplex. This description will be primarily of interest for the PCR pairing statistics, since it will highlight how the choice of specific sequences can influence the success of the PCR.

Herein, we follow the IUPAC-IUB notation, where the bases are classified as either strong bases S = {C, G} or weak bases W = {A, T}. Then, we define f_S as the fraction of S bases in the DNA sequence of interest. Our hypothesis is that the contribution coming from the next-neighbor canonical couples of base pairs is a function of this fraction. Specifically, we consider a linear interpolation (see Appendix A for further details), i.e.,

$$\Delta H_{n-n}(f_S) = f_S\,\Delta H_{(S,S)} + (1 - f_S)\,\Delta H_{(W,W)}, \tag{11}$$

$$\Delta S_{n-n}(f_S) = f_S\,\Delta S_{(S,S)} + (1 - f_S)\,\Delta S_{(W,W)}. \tag{12}$$

When illustrating the effect of differences in the richness of strong bases, we will present the results in terms of the number of S bases $n_{CG} = Lf_S$.

2.2. Degeneracy of Equivalent Duplexes

Given a specific sequence, there is only one well-defined complementary sequence, which corresponds with $\vec{\alpha} = \vec{0}$. On the contrary, with the same specific referential sequence, we can find many duplexes with the same $\vec{\alpha}$, i.e., duplexes with errors are degenerate.

Since there are 4 different possible bases, if we focus on one of them, there is only 1 exact complementary and 3 possible mismatches. Therefore, the degeneracy of a duplex with errors made by a selective molecule (primer or miRNA) of length L within a specific site of a much longer nucleic acid characterized by $\vec{\alpha}$ is

$$d(L,\vec{\alpha}) = 3^{(\alpha_{e1}+\alpha_{e2}+\alpha_i)} \binom{L-2-\alpha_{e1}-\alpha_{e2}}{\alpha_i}, \tag{13}$$

where the binomial coefficient takes into account all possible combinations of the α_i mismatches in the internal region of the duplex. Note that the simplicity of this degeneracy is partially due to our disregarding of the internal structure of internal mismatches.

2.3. Quantifying Selectivity

In the annealing phase of PCR, a short primer of length L can pair to its complementary target or to an off-target site in the two genomic ssDNA. Similarly, this also occurs in the pairing of miRNA, which can pair to its specific target or to other available sites within mRNA different molecules. The specificity of this binding is key to guarantee the success of the selective process. Herein, we compute the probability of having such successful binding using our model.

Let us consider a duplex comprising one selective molecule (primer/miRNA) and a longer nucleic acid. This duplex has, in principle, many ways to be formed. Obviously, we expect that there is a preferred binding, which corresponds with the selective molecule binding to the target region of the longer nucleic acid. For generalization purposes, let us assume that this target region appears N_{tar} times in the longer nucleic acids.

The statistical weight of occurrence for a specific binding j is given by the Boltzmann factor

$$\zeta_j = \exp\left(-\frac{\Delta G_j}{RT}\right), \tag{14}$$

where ΔG_j is the free energy difference corresponding to such binding, R is the gas constant, and T the temperature used in the experiment. Therefore, if we label $j = 0$ as the desirable hybridization of the primer/miRNA with a specific target region, the probability of having a successful selection is

$$\phi_0 = \frac{N_{tar}\zeta_0}{\sum_j \zeta_j}, \tag{15}$$

where the sum is carried out over all possible pairings in the system. Note that ϕ_0 is the conditional probability of having a successful binding, given that a binding occurs. In other words, we implicitly assume that in typical conditions, concentration and temperature grant a good degree of PCR primer (or miRNA) binding to the longer nucleic acids. ϕ_0 should not be confused with a melting curve, e.g., $\phi_0 = 0.1$ means that, out of a total of $n_b = 10$ bound primers/miRNA per long polymer, $n_b\phi_0 = 1$ is on-target and $n_b(1-\phi_0) = 9$ are off-target.

When computing ϕ_0, we have conjectured that the oligomers (primer/miRNA) in the system are mutually independent, i.e., they do not compete for the binding on each specific site. Therefore, we are requiring implicitly that the total number of actual bindings n_b per long polymer measured by the melting curve should not be much larger than N_{tar}/ϕ_0. This constraint means that, on average, the number of primers/miRNA on target computed from ϕ_0, i.e., $\phi_0 \times n_b$, does not exceed the number of target spots on the genome. In Appendix B, we provide a numerical check of n_b in typical genomic PCR conditions, based on the assumption of independence of primers and the computation of the melting curve, validating our hypothesis. Thus, we interpret ϕ_0 as a good estimator of pairing selectivity, expressing the ratio between on-target and off-target bindings.

In order to compute ϕ_0, we need to quantify the different ζ_j. On the one hand, we can compute ζ_0 through ΔG_0 using the formalism introduced in the previous section considering $\vec{\alpha} = \vec{0}$. On the other hand, using a mean field approximation, we assign the averaged Boltzmann factor

$$\zeta_a = \frac{\sum_{\vec{\alpha}} d(L,\vec{\alpha}) \exp\left[-\frac{\Delta G(L,\vec{\alpha},c_{\mathrm{NaCl}})}{RT}\right]}{4^L} \tag{16}$$

to the rest of the possible bindings, where the sum over $\vec{\alpha}$ runs for all possible external and internal mismatches. The denominator in (16) comes from $\sum_{\vec{\alpha}} d(L,\vec{\alpha}) = 4^L$, where the sum includes duplexes without a single complementary base pair. These duplexes can be considered within our energetic framework as impossible bindings to which we associate $\Delta G \to \infty$. These off-target pairs have a weight proportional to the total sites of pairing L_0 available in the system, i.e., the number of bases of the long polymer. Finally, we can rewrite the probability in (15) for pairing to the targets that are found in number N_{tar} in the system as

$$\phi_0 = \frac{N_{tar}\,\zeta_0}{N_{tar}\,\zeta_0 + L_0\zeta_a}. \tag{17}$$

Note that we have used that $L_0 \gg N_{tar}$, which is true for both PCR and miRNA.

This statistical approach allows us to provide a simple theoretical result with no knowledge of the specific sequences involved, which is computationally cheap. Although we are aware of the quantitative limitations of such an approach, we show here that our framework leads to a better understanding of the physics involved in selective processes such as the PCR technique or miRNA.

3. Results

The theoretical framework introduced above enables the evaluation of the probability of successful binding in the two conditions we have identified as especially challenging for the selectivity. In this section, we compute, for both PCR and miRNA, the targeting efficiency as a function of the relevant parameters (i.e., the length of the oligonucleotides L, the temperature T, the number of competing sites L_0, and the number of target sites in the system N_{tar}), by varying one parameter at a time and holding the other values fixed, and chosen to mimic typical real conditions.

The energetic parameters used in the calculations are obtained by averaging over the DNA and RNA thermodynamic dataset of the nearest-neighbor model, as detailed in Appendix A (Tables A1–A3).

3.1. PCR

The application of our general framework to the selectivity of primers in PCR requires some specifications. First, primers are typically designed to pair to a single target position on the genome, i.e., $N_{tar} = 1$. Second, in evaluating PCR efficiency, it is crucial to include the notion that the DNA polymerase needs a correct pairing between the target molecule and the 3′ terminal of the primer in order to start the amplification reaction [15]. This can be included in the model by splitting the average ζ_a of off-target pairings into the weighted combination of the two contributions stemming from $\alpha_{e1} = 0$ and $\alpha_{e1} > 0$,

$$\zeta_{(\alpha_{e1}=0)} = \frac{\sum_{\alpha_{e2}} \sum_{\alpha_i} d(L,(0,\alpha_{e2},\alpha_i))\zeta(0,\alpha_{e2},\alpha_i)}{4^{L-1}}, \tag{18}$$

$$\zeta_{(\alpha_{e1}>0)} = \frac{\sum_{\alpha_{e1}>0} \sum_{\alpha_{e2}} \sum_{\alpha_i} d(L,\vec{\alpha})\zeta(\vec{\alpha})}{3 \times 4^{L-1}}, \tag{19}$$

where the denominator expresses the degeneracy of duplexes in the two cases. Accordingly, the total statistical weight of the pairing of the primer along the genome becomes

$$\sum_j \zeta_j \simeq \zeta_0 + \frac{1}{4} L_0 \zeta_{(\alpha_{e1}=0)} + \frac{3}{4} L_0 \zeta_{(\alpha_{e1}>0)}, \tag{20}$$

where the coefficients $1/4$ and $3/4$ are the frequency with which correct and defected pairing occur in the $3'$ terminal nucleobase, respectively. Thus, the probabilities of the two classes of off-target pairings, with and without correct pairing at the $3'$ terminal, are

$$\phi_{(\alpha_{e1}=0)} = \frac{\frac{1}{4} L_0 \zeta_{(\alpha_{e1}=0)}}{\zeta_0 + \frac{1}{4} L_0 \zeta_{(\alpha_{e1}=0)} + \frac{3}{4} L_0 \zeta_{(\alpha_{e1}>0)}}, \tag{21}$$

$$\phi_{(\alpha_{e1}>0)} = \frac{\frac{3}{4} L_0 \zeta_{(\alpha_{e1}>0)}}{\zeta_0 + \frac{1}{4} L_0 \zeta_{(\alpha_{e1}=0)} + \frac{3}{4} L_0 \zeta_{(\alpha_{e1}>0)}}. \tag{22}$$

Since off-target pairing with errors at the $3'$ terminal inhibits the amplification, the relevant quantity expressing the selectivity of PCR is the ratio $\tilde{\phi}_0$ of on-target pairing over all the defectless $3'$ primer–genome binding,

$$\tilde{\phi}_0 = \frac{\phi_0}{\phi_0 + \phi_{(\alpha_{e1}=0)}} = \frac{\zeta_0}{\zeta_0 + \frac{1}{4} L_0 \zeta_{(\alpha_{e1}=0)}}, \tag{23}$$

i.e., the meaningful ratio is normalized with only defectless $3'$ primer–genome possible pairings, as, for the other cases, the PCR would not even start its amplification process.

Figures 3 and 4 show the primer length dependence for the PCR pairing statistics. Specifically, we display the curves for the on-target binding probability ϕ_0 (blue dots in Figure 3), the probability of off-target binding with and without $3'$ pairing errors $\phi_{(\alpha_{e1}>0)}$ and $\phi_{(\alpha_{e1}=0)}$ (yellow and purple diamonds in Figure 3, respectively), and the renormalized on-target binding probability $\tilde{\phi}_0$ (Figure 4). The computation of the different curves is performed holding fixed $L_0 = 6 \times 10^9$, since the primer can bind to both strands of the genomic DNA double helix, and $T = 55\,^{\circ}\mathrm{C}$, a typical annealing temperature; the salt concentration is $[Na^+] = 55$ mM, representing the standard salt concentration, according to a typical DNA polymerase manufacturer's instructions. Due to the logarithmic dependence on the salt concentration, it is necessary to notably change the salt concentration in order to observe significant changes (see Appendix C for details). Both ϕ_0 and $\tilde{\phi}_0$ exhibit a rather sharp rise, indicating that the selectivity of the primer markedly changes upon lengthening or shortening the primer of a single nucleobase. The significant difference between ϕ_0 and $\tilde{\phi}_0$ is due to the remarkable difference between the probability of off-target binding and its sub-ensemble of off-target with a defectless $3'$ terminal (red and purple diamonds, respectively, in Figure 3). This proves that such a defect is actually quite common in random binding, since terminal defects involve the smallest energy penalties [12] with a limited growth of degeneracy. In Figure 4, we also consider the effect of modifying the fraction of CG bases in the primer. Full dots are computed with a number of CG bases $n_{CG} = L/2$; open dots correspond to $n_{CG} = L/2 \pm 2$.

The temperature dependence for the PCR pairing statistics is analyzed in Figure 5. Therein, $\tilde{\phi}_0$ is plotted for either a balanced or unbalanced proportion of CG bases for $L = 20$ and $L_0 = 6 \times 10^9$. When T increases, the fraction of on-target binding decreases, as expected since the energetic gain for Watson–Crick against defected pairing decreases with T. Again, we find a sharp transition between high and low $\tilde{\phi}_0$, and the typical working temperature $T = 55\,^{\circ}\mathrm{C}$ is indeed in the regime of high selectivity, but close to the transition to low selectivity.

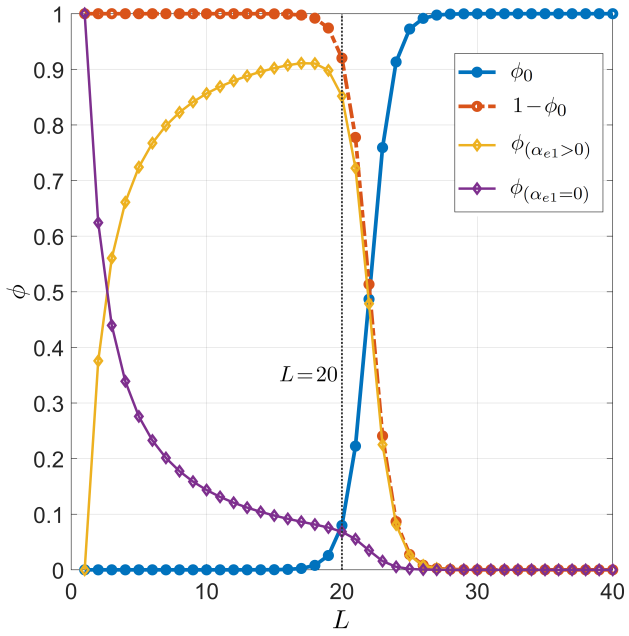

Figure 3. Dependence on the primer length L of the pairing probability for PCR. Fixed values are considered for temperature $T = 55\,^{\circ}$C, total sites $L_0 = 6 \times 10^9$, salt concentration $[Na^+] = 55$ mM, and for CG fraction $n_{CG} = L/2$. Successful target binding (ϕ_0, blue dots). Off-target binding ($1 - \phi_0$, red dots). Off-target binding can be split into 2 contributions: off-target binding with no terminal defects at the 3′ end ($\phi_{(\alpha_{e1}=0)}$, purple diamonds), off-target binding with terminal defects at the 3′ end ($\phi_{(\alpha_{e1}>0)}$, yellow diamonds). The vertical gray line stands for $L = 20$, a typical primer length in PCR.

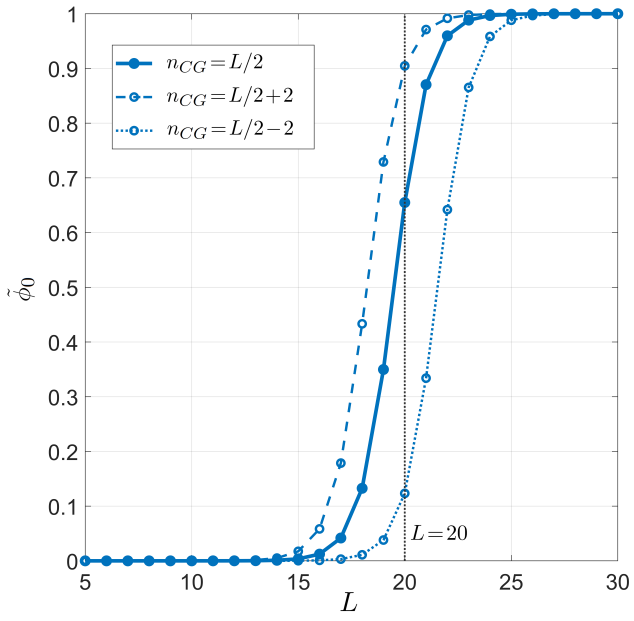

Figure 4. Dependence on the primer length L of the on-target pairing probability conditioned on the well-paired 3′ end for PCR. Fixed values are considered for temperature $T = 55\,^{\circ}$C, salt concentration $[Na^+] = 55$ mM, and total sites $L_0 = 6 \times 10^9$, for different CG fractions in the primer. Full dots and solid line: CG fraction $n_{CG} = L/2$. Open dots and dashed line: CG fraction $n_{CG} = L/2 + 2$. Open dots and dotted line: CG fraction $n_{CG} = L/2 - 2$. Curves are computed using the average energetic description, detailed on the CG fraction (Equations (11) and (12)). The vertical gray line stands for $L = 20$, a typical primer length in PCR.

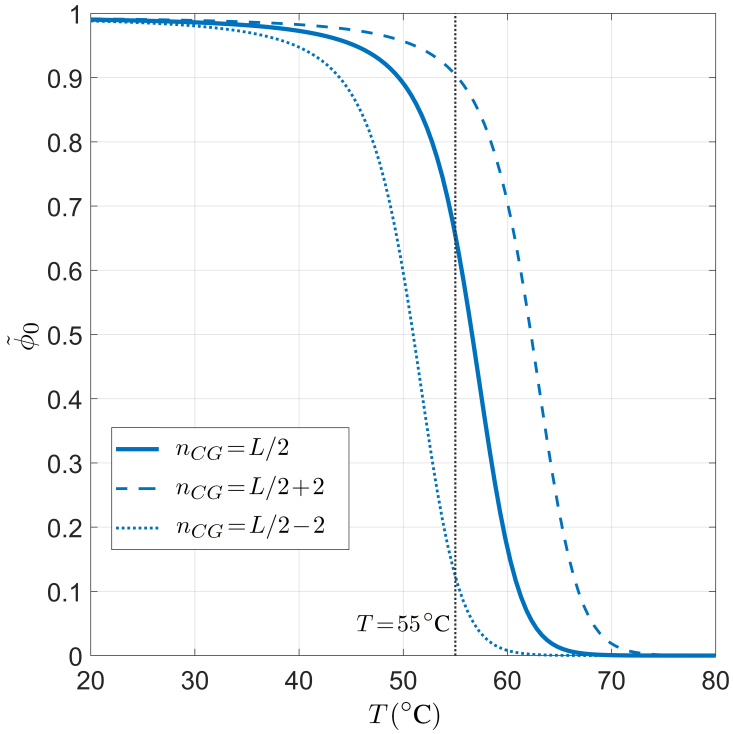

Figure 5. Dependence on the temperature T of the on-target primer pairing probability conditioned on the well-paired 3′ end for PCR. Fixed values are considered for primer length $L = 20$, salt concentration $[Na^+] = 55$ mM, and total sites $L_0 = 6 \times 10^9$, for different CG fractions in the primer. Solid line: CG fraction $n_{CG} = L/2$. Dashed line: CG fraction $n_{CG} = L/2 + 2$. Dotted line: CG fraction $n_{CG} = L/2 - 2$. The gray line marks $T = 55\,^\circ$C, a typical annealing temperature in the PCR experiments.

Finally, we compute the on-target binding probability $\tilde{\phi}_0$ as a function of the number of competing binding sites L_0. This is shown in Figure 6, where we repeat our study for different proportions of CG bases while fixing $L = 20$ and $T = 55\,^\circ$C. In this case, the transition is much smoother and relevant changes in the selectivity appear only when changing L_0 of order of magnitudes. When considering the L_0 of the human genome, the selectivity of the PCR primers is found to be very high, as expected.

3.2. miRNA

To apply our theoretical approach to the selective binding of miRNA, we need first to assess which are the most appropriate values for L_0 and N_{tar}.

miRNAs preferentially target 3′ UTRs, since the coding region is usually bound to other macromolecular complexes, e.g., exon junction complexes and ribosomal machinery, that would displace the RISC complex [16]. For this reason, we choose to include in our analysis only a portion of around 1000 nt, which corresponds to the median length of the 3′ UTR [17]. Moreover, to evaluate the total number of possible binding sites for each miRNA, we have to consider that not all the genes encoded in the human genome are actively transcribed within a cell. Transcriptome data in fact show that approximately 11,000 genes are simultaneously detectable within a specific cell type [18]. Thus, in evaluating the seed selectivity, we consider a reduced transcriptome length given by the product of these two quantities, $L_0 = 1.1 \times 10^7$.

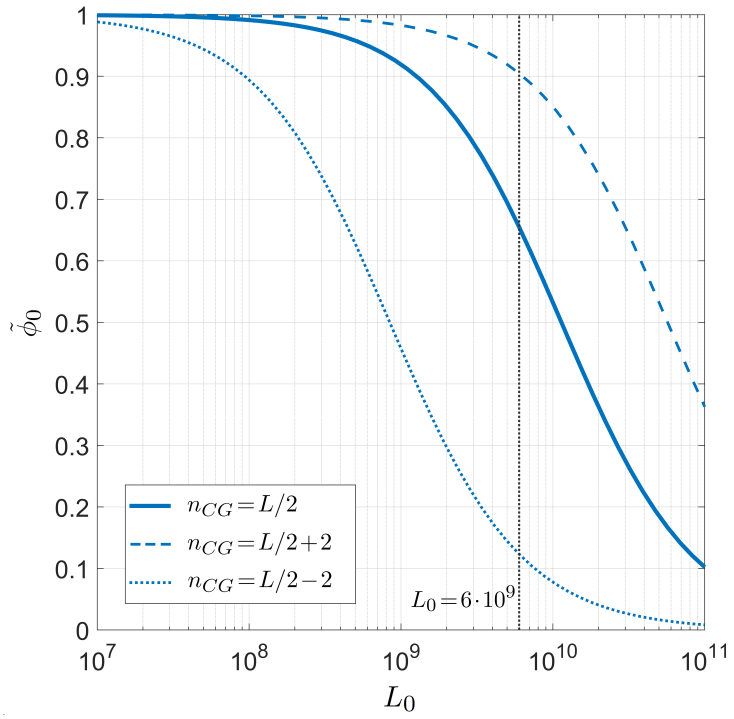

Figure 6. Dependence on the genome length L_0 of the on-target primer pairing probability conditioned on the well-paired 3′ for PCR. Fixed values are considered for temperature $T = 55\,^{\circ}$C, salt concentration $[Na^+] = 55$ mM, and primer length $L = 20$, for different CG fractions in the primer. Solid line: CG fraction $n_{CG} = L/2$. Dashed line: CG fraction $n_{CG} = L/2 + 2$. Dotted line: CG fraction $n_{CG} = L/2 - 2$. The gray line marks twice the length of the human genome, $L_0 = 6 \times 10^9$.

As for the evaluation of N_{tar}, it is relevant to notice that, differently from the PCR situation in which the primer is designed to target a single position in the genome, a single miRNA regulates the expression of several genes simultaneously. In particular, evidence suggests that the "targetome" of a miRNA is not random, but it is generally constituted by transcripts sharing the same biological network. This fact suggests that miRNAs can regulate entire target pathways [19,20]. Thus, in order to provide a reasonable value for the seed length that ensures the required selectivity, we need to consider the mean number of genes targeted by each miRNA family (groups of miRNA sharing the same seed). Analyses of preferential conservation of the seed sequence in mammals against vertebrates have indicated that the average number of targets for each miRNA family is around 300 [16]. More recent studies based on the integration of miRNA target prediction and RNA sequencing data suggest an average of 90 targets for each miRNA, highlighting the high variability among individual miRNAs [21]. Therefore, in the application of our approach to miRNA selectivity, we consider N_{tar} to be in the range 100–300.

The dependence of successful binding ϕ_0 on the length of the miRNA seed region is shown in Figure 7, where the three conditions of $N_{tar} = 1$, $N_{tar} = 100$, and $N_{tar} = 300$ are considered for $T = 37\,^{\circ}$C (temperature in human cell) and $L_0 = 1.1 \times 10^7$; the salt concentration is $[Na^+] = 150$ mM (to mimic the physiological salt concentration). The results obtained for $N_{tar} = 1$ clearly indicate that, if miRNAs were meant to regulate only one specific gene, the seed length should have been 4–5 nucleobases longer in order to have the right selectivity.

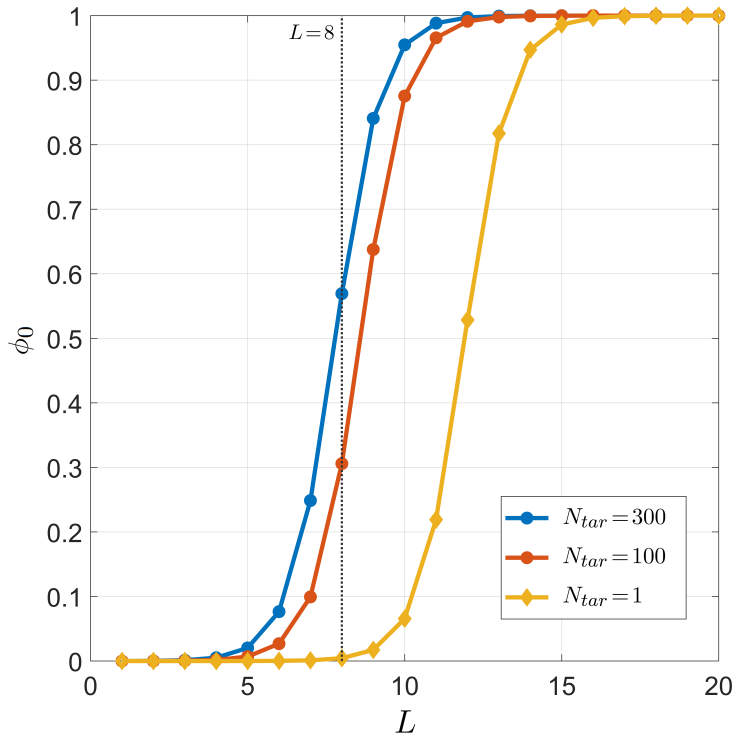

Figure 7. Dependence on the miRNA length L of the target pairing probability ϕ_0 for fixed temperature $T = 37$ °C, total sites $L_0 = 1.1 \times 10^7$, salt concentration $[Na^+] = 150$ mM, and CG fraction $n_{CG} = L/2$. Curves correspond to different numbers of distinct miRNA targets N_{tar}. Yellow dots: $N_{tar} = 1$. Red dots: $N_{tar} = 100$. Blue dots: $N_{tar} = 300$. The gray line marks $L = 8$, the typical length of the seed region of miRNA.

It is possible to recast the effect of L_0 and N_{tar} in Equation (17) in a single parameter $L_{eff} = L_0/N_{tar}$, which is the ratio of the number of off-target over on-target binding sites, quantifying the required selectivity. With such a definition, the equation can be rewritten as

$$\phi_0 = \frac{\zeta_0}{\zeta_0 + L_{eff}\zeta_a}. \tag{24}$$

Therefore, two different systems where N_{tar} and L_0 scale with the same factor, and thus with the same L_{eff}, are completely equivalent in our theoretical framework.

Finally, we present the relation between selectivity and the number of competing binding sites L_{eff} in Figure 8, i.e., the analogous dependence shown in Figure 6 in the case of PCR. As already introduced above, since the role played by the length of the long polymer is always modulated by the number of targets, it suffices to study the dependence on the defined effective length $L_{eff} = L_0/N_{tar}$. The inset shows that the dependence on T of ϕ_0 is not so strong as observed in the PCR case, i.e., the miRNA selectivity is less sensible to the temperature in the range of interest for the human body.

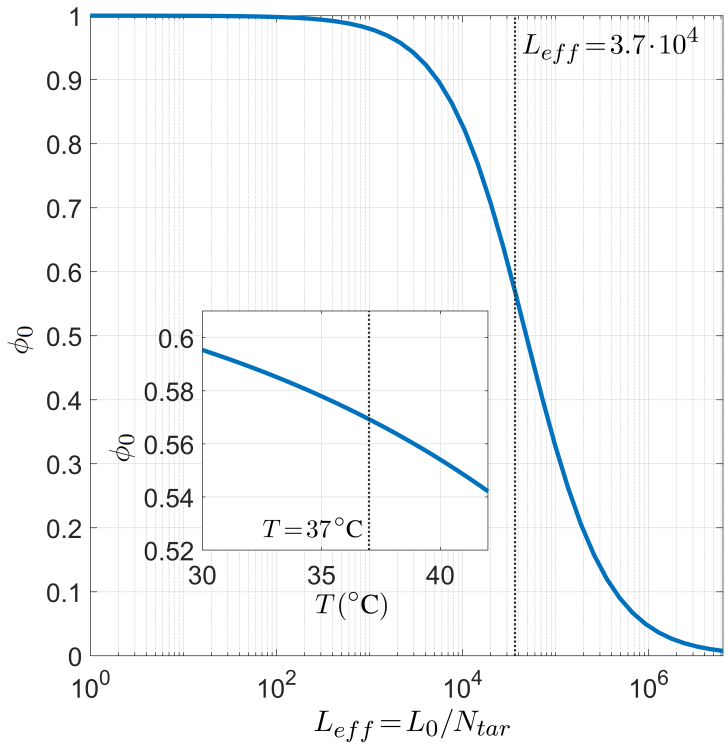

Figure 8. Dependence of the miRNA pairing probability ϕ_0 on the reduced transcriptome length L_{eff} computed with $N_{tar} = 300$ for fixed temperature $T = 37\,^\circ$C, salt concentration $[Na^+] = 150$ mM, CG fraction $n_{CG} = L/2$, and miRNA length $L = 8$. Inset: T dependence of ϕ_0 in the same conditions and $L_{eff} = 3.7 \times 10^4$. The gray lines mark the reference values $T = 37\,^\circ$C (inset) and $L_{eff} = 3.7 \times 10^4$ (main figure).

4. Discussion

The statistical framework developed in this work has allowed the analysis of the effectiveness of selectivity in both PCR and miRNA. In spite of its simplicity, the model has helped to better understand the relevance of the mechanisms behind the selective process, enabling non-trivial predictions that appear to have quantitative agreement with experimental observations. Among these remarkable features, we highlight the steep dependence of selectivity on L in Figures 3, 4, and 7 and the complex L dependence of various families of defect duplexes (Figure 3), which are discussed below separately for the two cases of interest.

4.1. PCR

In the context of PCR, our results convey various insights on the nature of primer selectivity. If the selective mechanism was entirely provided by on-target vs. off-target binding, i.e., expressed by ϕ_0, longer primers would be needed, e.g., selectivity of $\phi_0 > 0.8$ entails $L > 24$ from Figure 3. Nevertheless, we find that the constraint of Watson–Crick pairing at the $3'$ terminal of the primers significantly changes the range of successful binding. This is because, out of the large number of expected off-target pairings (red dots in Figure 3), the fraction of primer that binds off-target with a well-formed $3'$ terminal is small and has a non-trivial dependence on L, with a drop for $L > 20$. When only defectless $3'$ terminal binding is considered, the successful binding of $L = 20$ primers among the plethora of off-target positions offered by the human genome is approximately $\tilde{\phi}_0 \simeq 0.65$. While this figure is still far from 1, we argue that it is sufficient, since the PCR protocol makes use of a combination of two primers, designed to target the complementary strands of the region of interest. The double strands produced at the end of the first replication cycle are much shorter than the initial genome, reducing effectively the value of L_0 in our description.

Thus, in the following replication cycles, the ratio between on-target and off-target position increases, leading to a progressive increment of $\tilde{\phi}_0$.

Another outcome of our approach is the quantification of the effect of unbalancing CG and AT bases in the primer: Figure 6 shows that the reduction or addition of two CG bases markedly affects $\tilde{\phi}_0$. This is in line with the experimental procedures: when the CG content of a primer is low, its length is usually extended to compensate for the loss of selectivity.

The dependence of $\tilde{\phi}_0$ on L_0 found by our model in Figure 6 is weak, i.e., for a moderate change of L_0, the pairing probability does not change. This indicates that the PCR primer's length granting selectivity depends weakly on the complexity of the molecular target, and thus it does not need to be significantly changed depending on the nucleic acid environment.

We have found that the on-target pairing probability decreases as T increases (Figure 5). This behavior is well grounded from a thermodynamic point of view, since increasing the temperature makes the free energy penalty associated with mispairing decrease, and the population of more entropic (defected) states is favored. However, this dependence appears in contradiction with the typical experience of the molecular biologist. In fact, when PCR efficiency is not very high, the annealing temperature is usually raised (typically by 2–3 °C), especially during the first cycles to improve specificity. We argue that this experimental strategy is not rooted in an increment in the selectivity at equilibrium (which is the quantity we compute), but rather it is a strategy to overcome kinetic barriers, i.e., to avoid off-target defected bindings having lifetimes comparable with the annealing time. Indeed, the increase in T by even a few degrees strongly reduces the lifetime of off-target bindings, thus speeding up the dynamics towards equilibrium (see Appendix D).

4.2. miRNA

Now, the application of our description to the miRNA selective process is discussed. The results in Figure 7 demonstrate that, in order to obtain significant selectivity over the targets around 0.8, a seed region of length 9–10 would be required, depending on the number of targets. Differently from what was found in PCR, where we obtained a primer length transition consistent with the typical experimental setting, the estimated length of the miRNA seed is larger than the actual value. This difference is not surprising since miRNAs operate within a much more intricate biological network than in vitro PCR settings.

Many factors contribute to the complexity of this system. miRNAs are part of a ribonuclear particle, where the interaction with the protein component plays an essential role not only in the mechanism of silencing that follows the binding, but also in the target recognition. Experiments exploiting AGO crosslinking and coimmunoprecipitation revealed extensive AGO-bound mRNAs in the absence of miRNA seed complementarity, thus suggesting that AGO proteins might have an RNA-binding property that allow thems to recognize mRNA targets [22]. Moreover, once the RISC complex is bound to the target mRNA, the molecular machinery undergoes a conformational change that exposes a part of the miRNA 3′ region (nts 13–16), thus allowing a supplemental pairing with the target, and providing additional selectivity [3], which could be interpreted as an increment in L in our description. Furthermore, the interaction between a miRNA and its target is not solely dependent on the nucleic acid pairing, but also on the availability of the target in the cell. This implies that the target gene must be transcribed, and that its secondary structure must allow the landing of the RISC complex and the binding of the miRNA to the target region. These elements suggest that the actual seed length is a compromise between the selectivity provided by nucleic acid pairing and the complexity of the cellular environment that calls for a higher degree of flexibility of the system. Overall, the comparison between the estimated and actual miRNA seed length offers a quantification of the extra selectivity brought by the mechanisms at play beyond base pairing.

5. Conclusions

In this study, a theoretical framework has been developed to describe selective processes in complex nucleic acid environments and applied to the PCR technique and miRNA-based gene regulation. In both cases, the selective binding occurs in spite of a huge degeneracy of competing defected pairings. The theory is constructed around two main approximations: (i) the coarse-grained description of the duplex energetics, recast based on three classes of base pairing errors, and (ii) the statistics of competing binding sites, computed by assuming random sequences.

Despite the complexity of the problem, our simple approach has led to a quite cheap model that has enabled quantitative estimates of the selectivity in the two processes, at the same time enlightening features that cannot be recognized in the absence of a quantitative framework: the sharpness of the transition in selectivity as a function of the length L of the oligomers at play; the relevance of the constraint of defectless $3'$ terminals in PCR primer targeting; and the quantitative estimate of the contribution to miRNA target selectivity provided by processes beyond base pairing.

Author Contributions: The research was designed by S.S., T.B., and E.M.P.; C.A.P. and S.M. refined the theoretical model and carried out the numerical calculations. All authors contributed in writing the article. All authors have read and agreed to the published version of the manuscript.

Funding: C.A.P. acknowledges support from project PGC2018-093998-B-I00 funded by FEDER/Ministerio de Ciencia e Innovación–Agencia Estatal de Investigación (Spain), and program PAIDI-DOCTOR funded by Junta de Andalucía and the European Social Fund. S.S. acknowledges support from an INFN Lincoln grant and UNIPD DFA BIRD2020 grant. This work was supported by Regione Lombardia and FESR, Linea Accordi per la Ricerca (NeOn project ID 239047 to S.M. and T.B.), and by a PRIN2017 project from Ministero dell'Istruzione dell'Università e della Ricerca (ID 2017Z55KCW to S.M. and T.B.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code used for this analysis is available upon request to the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Thermodynamic Parameters

In this section, we provide the values, and corresponding references from where they have been taken, which have been used in the main text for our thermodynamic description of the duplexes: either DNA/DNA in the case of the PCR or RNA/RNA in the case of miRNA. As already settled in the captions of all tables, the units we have used for ΔH and ΔS are kcal mol^{-1} and cal mol^{-1} K^{-1}, respectively. If nothing is said regarding the units, we assume that everything is expressed in such units to avoid burdening the writing of the values.

Appendix A.1. Initiation and Canonical Watson–Crick Base Pairs

Table A1 displays the thermodynamic parameters for the canonical Watson–Crick base pairing and the initiation contribution given in terms of the last bases with the end. The data for DNA/DNA interaction are taken from [12] whereas data from RNA/RNA belong to [23]. We obtain ΔH_0, ΔS_0, ΔH_{n-n}, and ΔS_{n-n} averaging within the table. For the initiation term, we take twice the average value of the end contributions, whereas for the next-neighbor term, we have to take into account that the quadruplets that are symmetric under complementary inversion (e.g., AT/TA, TA/AT, CG/GC, and GC/CG for DNA/DNA) have half of the weight in the average since the other combinations appear twice in frequency when we consider random sequences. We obtain $\Delta H_0 = 2.4$, $\Delta S_0 = 1.2$, $\Delta H_{n-n} = -8.2$, and $\Delta S_{n-n} = -22.01$ for the PCR case (DNA/DNA interaction); and $\Delta H_0 = 7.33$, $\Delta S_0 = 9.0$, $\Delta H_{n-n} = -10.78$, and $\Delta S_{n-n} = -27.9$ for the miRNA case (RNA/RNA interaction).

Appendix A.2. External and Internal Mismatches

In Tables A2 and A3, we collect the thermodynamic parameters for dangling ends and internal mismatches, respectively. The assumption made in our work is that using these data for describing the RNA/RNA duplexes does not introduce significant deviations since our expectation is that the main difference between DNA/DNA and RNA/RNA comes from the different energetics of canonical Watson–Crick base pairs. The data for the dangling end have been taken from [12], whereas the information regarding internal mismatches was somewhat more split within the literature [24–28]. For the average description, we have simply averaged over all cells within the tables, obtaining the average contributions $\Delta H_d = -2.5$, $\Delta S_d = -6.7$, $\Delta H_i = 0.1$, and $\Delta S_i = -0.8$.

Appendix A.3. Dependence on the Sequence, Fraction of CG

Herein, we specify how we obtain the average energetic parameters for the next-neighbor couple of base pairs, considering the fraction of C or G base in the DNA sequences. We follow the IUPAC-IUB notation, where the bases are classified into two categories, either strong bases or weak ones. Specifically, we consider S = {C, G} and W = {A, T}. Then, we compute the partial average of the energetic parameters for the next-neighbor couple of base pairs, distinguishing if they are made by (W,W), (S,S), or a mixture (S,W) and (W,S). The results for such averages are displayed in Table A4. Note that $(\Delta H_{(S,S)} + \Delta H_{(W,W)})/2 = -8.25 \simeq \Delta H_{(S,W),(W,S)}$ and $(\Delta S_{(S,S)} + \Delta S_{(W,W)})/2 = -21.9625 \simeq \Delta S_{(S,W),(W,S)}$. These properties justify the development of a simple energetic description of the pairing based on a linear interpolation. Namely, we assign averaged next-neighbor contributions such that $\Delta H_{n-n}(f_S) = f_S \Delta H_{(S,S)} + (1 - f_S)\Delta H_{(W,W)}$ and $\Delta S_{n-n}(f_S) = f_S \Delta S_{(S,S)} + (1 - f_S)\Delta S_{(W,W)}$, where f_S stands for the fraction of strong bases in the duplex of interest.

Appendix A.4. Salt Contribution

Salt corrections, ΔS_{salt}, have been adapted from Equation (22) of the article by Owczarzy et al. [14]. Therein, the effect of the salt is written for the melting temperature T_m. If we take into account that $T_m = \Delta H/(\Delta S + R \ln(c_{DNA}))$, with c_{DNA} being the DNA concentration, and we assume that the salt has no impact on the enthalpy, it is possible to obtain the effect of salt into the entropy. Specifically, we obtain that the salt correction to the pairing entropy depends on the salt concentration $[Na^+]$, on the set of parameter of the duplex presented in the article $(L, \vec{\alpha})$, and the fraction of strong bases f_S

$$\Delta S_{salt}([Na^+], L, \vec{\alpha}) = \Delta H(L, \vec{\alpha})[(4.29 f_S - 3.95) \times 10^{-5} \ln([Na^+]) + 9.4 \times 10^{-6} \times \ln^2([Na^+])], \quad \text{(A1)}$$

being independent of the DNA concentration as expected. When the fully averaged description is used in the main text, we fix $f_S = 0.5$.

Table A1. Thermodynamic parameters for canonical Watson–Crick base pairs for duplexes made by DNA/DNA (left panel) and RNA/RNA (right panel). Energy and entropy units are kcal mol^{-1} and cal mol^{-1} K^{-1}, respectively.

Propagation Sequence	ΔH	ΔS	Propagation Sequence	ΔH	ΔS
AA/TT	−7.6	−21.3	AA/UU	−6.82	−19.0
AT/TA	−7.2	−20.4	AU/UA	−9.38	−26.7
TA/AT	−7.2	−21.3	UA/AU	−7.69	−20.5
CA/GT	−8.5	−22.7	CA/GU	−10.44	−26.9
GT/CA	−8.4	−22.4	GU/CA	−11.4	−29.5

Table A1. *Cont.*

Propagation Sequence	ΔH	ΔS	Propagation Sequence	ΔH	ΔS
CT/GA	−7.8	−21.0	CU/GA	−10.48	−27.1
GA/CT	−8.2	−22.2	GA/CU	−12.44	−32.5
CG/GC	−10.6	−27.2	CG/GC	−10.64	−26.7
GC/CG	−9.8	−24.4	GC/CG	−14.88	−36.9
GG/CC	−8.0	−19.9	GG/CC	−13.39	−32.7
EC(G)/G(C)E	0.1	−2.85	EC(G)/G(C)E	1.805	−0.75
EA(T)/T(A)E	2.3	4.05	EA(U)/U(A)E	5.525	9.75

Table A2. Thermodynamic parameters for dangling ends for DNA/DNA interaction. Energy and entropy units are kcal mol^{-1} and cal mol^{-1} K^{-1}, respectively.

		X							
Dangling	Propagation	A		T		C		G	
End	Sequence	ΔH	ΔS	ΔH	ΔS	ΔH	ΔS	ΔH	ΔS
5′-dangling	XA/T	0.2	2.3	−6.9	−20.0	0.6	3.3	−1.1	−1.5
	XT/A	−2.9	−7.7	−0.2	−0.3	−4.1	−13.2	−4.2	−15.1
	XC/G	−6.3	−17.2	−4.0	−11.0	−4.4	−12.5	−5.1	−14.1
	XG/C	−3.7	−10.1	−4.9	−13.8	−4.0	−11.8	−3.9	−10.8
3′-dangling	AX/T	−0.5	−1.2	−3.8	−12.7	4.7	14.25	−4.1	−13.2
	TX/A	−0.7	−0.7	2.9	10.3	4.4	14.8	−1.6	−3.5
	CX/G	−5.9	−16.4	−5.2	−15.1	−2.6	−7.4	−3.2	−10.3
	GX/C	−2.1	−3.8	−4.4	−13.1	−0.2	0.1	−3.9	−11.2

Table A3. Thermodynamic parameters for internal errors for DNA/DNA interaction. Energy and entropy units are kcal mol^{-1} and cal mol^{-1} K^{-1}, respectively.

		Y							
Propagation		A		T		C		G	
Sequence	X	ΔH	ΔS	ΔH	ΔS	ΔH	ΔS	ΔH	ΔS
AX/TY	A	1.2	1.7	WC	WC	2.3	4.6	−0.6	−2.3
	T	WC	WC	−2.7	10.8	−1.2	6.2	1.0	0.9
	C	5.3	14.6	0.7	0.2	0.0	−4.4	WC	WC
	G	−0.7	−2.3	−2.5	−8.3	WC	WC	−3.1	−9.5
TX/AY	A	4.7	12.9	WC	WC	3.4	8.0	0.7	0.7
	T	WC	WC	0.2	−1.5	1.0	$0.7	−0.1	−1.7
	C	7.6	20.2	1.2	0.7	6.1	16.4	WC	WC
	G	3.0	7.4	−1.3	−5.3	WC	WC	1.6	3.6
CX/GY	A	−0.9	−4.2	WC	WC	1.9	3.7	−0.7	−2.3
	T	WC	WC	−5.0	−15.8	−1.5	−6.1	−4.1	−11.7
	C	0.6	−0.6	−0.8	−4.5	−1.5	−7.2	WC	WC
	G	−4.0	−13.2	−2.8	−8.0	WC	WC	−4.9	−15.3
GX/CY	A	−2.9	−9.8	WC	WC	5.2	14.2	−0.6	−1.0
	T	WC	WC	−2.2	−8.4	5.2	13.5	3.3	10.4
	C	−0.7	−3.8	2.3	5.4	3.6	8.9	WC	WC
	G	0.5	3.2	−4.4	−12.3	WC	WC	−6.0	−15.8

Table A4. Thermodynamic parameters for canonical Watson–Crick base pairs for duplexes made by DNA/DNA averaged over categories of bases. Energy and entropy units are kcal mol^{-1} and cal mol^{-1} K^{-1}, respectively.

Propagation Sequence	ΔH	ΔS
WW	−7.4	−21.1
SS	−9.1	−22.8
SW(WS)	−8.2	−22.1

Table A5. Estimation of the relative decrease in the disassociation time $\tau_{L,\vec{\alpha}}(T)\,/\tau_{L,\vec{\alpha}}(T+\Delta T)$, due to a temperature increment ΔT. The values of the table are computed using Equation (A5), with different values of the increment ΔT, primer length L, and external mismatch α_{e1} in $\vec{\alpha}=(\alpha_{e1},0,0)$, representing different levels of defectiveness of the duplex. $T=55\,^{\circ}$C, as a typical PCR temperature.

		$\tau_{L,\vec{\alpha}}(T)/\tau_{L,\vec{\alpha}}(T+\Delta T)$					
ΔT	L	$\alpha_{e1}=0$	$\alpha_{e1}=1$	$\alpha_{e1}=2$	$\alpha_{e1}=5$	$\alpha_{e1}=10$	$\alpha_{e1}=15$
$\Delta T=+2\,^{\circ}$C	$L=20$	4.4	4.2	3.8	3.1	2.1	1.4
	$L=21$	4.7	4.5	4.2	3.3	2.3	1.5
	$L=22$	5.1	4.8	4.5	3.6	2.4	1.7
$\Delta T=+3\,^{\circ}$C	$L=20$	9.1	8.4	7.5	5.3	3.0	1.7
	$L=21$	10.2	9.5	8.4	6.0	3.4	1.9
	$L=22$	11.5	10.6	9.5	6.7	3.8	2.1

Appendix A.5. Estimation of the Goodness of the Thermodynamic Parameters

Being aware that the pairing energy of two sequences has a significant dependence not only on the CG content but on the specific sequence of the nucleobases, herein, we analyze the error introduced in this simplification. In Figure A1, we show the melting curves of solutions of two 20 mers with perfect complementarity, all computed with the analytical expression of a complementary system with two equipopulated species [11], with different pairing energies. The colored lines are computed with our averaged thermodynamic parameters, with f_S ranging in all the possible values. In addition, we have computed the melting curves of 40 different duplexes with $f_S=0.5$, with the energetic parameters obtained with the standard NN protocol for each duplex; the average of these melting curves and the standard deviation are represented in the purple dashed line and shadow, respectively. We notice that the curve with $f_S=0.5$ of our energetic description approximates the average melting curve with standard Santa Lucia protocol with low discrepancy $\Delta T<1\,^{\circ}$C, being always within the standard deviation region, too. This is clearly a good energetic approach, because of the comparison with the standard protocol and because it provides a simple way to obtain the hybridization thermodynamics ranging across all the possible values of f_S.

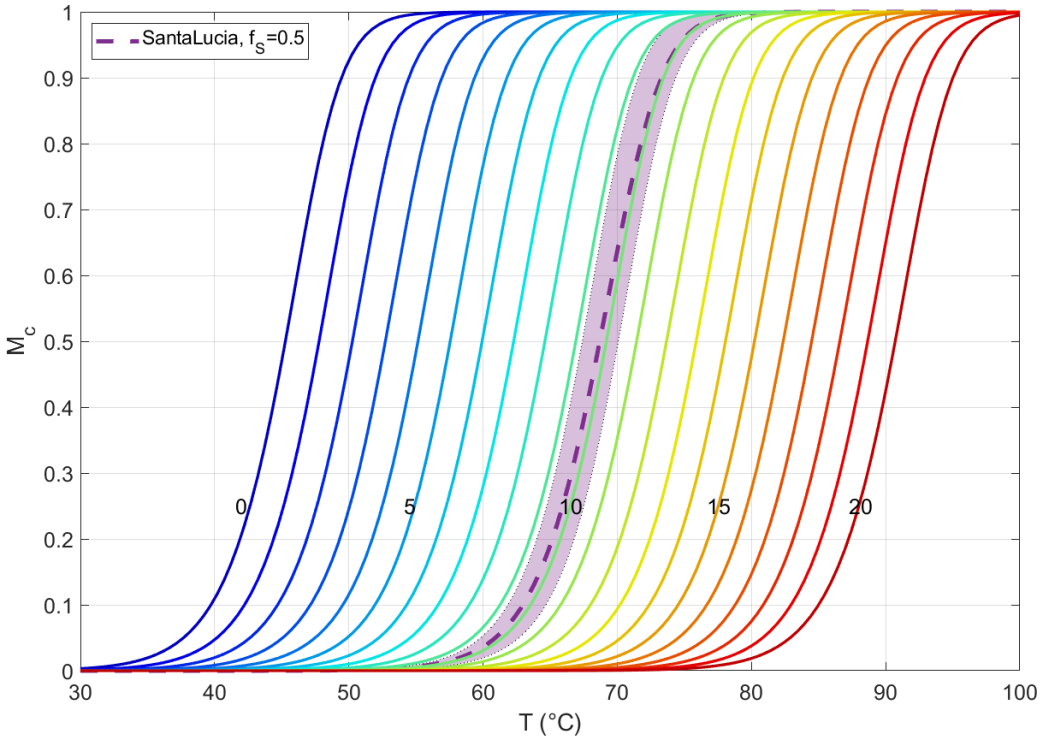

Figure A1. Melting curves of solutions of two 20 mers with perfect complementarity, computed with the analytical expression of a complementary system with two equipopulated species [11], with different pairing energies. The colored lines are computed with our averaged thermodynamic parameters, with number of C or G bases ranging from 0 to 20. Melting curves of 40 different duplexes with $f_S = 0.5$ have been computed with the energetic parameters obtained with the standard NN protocol for each duplex; the average of these melting curves and the standard deviation are represented in the purple dashed line and shadow, respectively. All the solutions are in the same experimental conditions: $c_{DNA} = 100$ nM and $[Na^+] = 1$ M.

Appendix B. Melting Curve

The core of our results relies on the conditional probabilities of a specific binding ϕ_0, given that there is a binding. This probability is obtained as a comparison between the Boltzmann weights $\zeta_j = \exp(-\frac{\Delta G_j}{RT})$ for different bindings. The independence of the primers is an important hypothesis in this respect. Therein, we should guarantee that the primers do not saturate the binding locations since this could make the quantities ϕ lose their meaningfulness. Herein, we will give an estimate of the number of primers bound to any possible location per genome. To do so, we start by deriving the melting curve M_c, i.e., the fraction of free primers in the system. Resting again on the assumption of independence of the primers, which we expect to be a good approximation at least in the limit of high temperature, we take

$$M_c = \frac{1}{1 + c_g(\zeta_0 + L_0\zeta_a)}, \tag{A2}$$

with c_g being the concentration of genomes. Note that, as seen in [11], the Boltzmann factor is accompanied by the concentration. Since, in the main text, we have always considered bound states (without comparing to the free state), the concentration was not necessary. In our approximation, the relevant concentration is the one of the genome, since we are assuming the limit of completely independent primers. Therefore, the number of expected bounded primers per genome is

$$n_b = \frac{c_p}{c_g}(1 - M_c), \tag{A3}$$

where c_p is the concentration of primers. Our rule of thumb is that the product $\phi_0 n_b$ should not be much greater than 1. Since we have obtained that $\phi_0 \simeq 0.1$ in the main text from Figure 3, we want to ensure that n_b does not reach values much greater than 10.

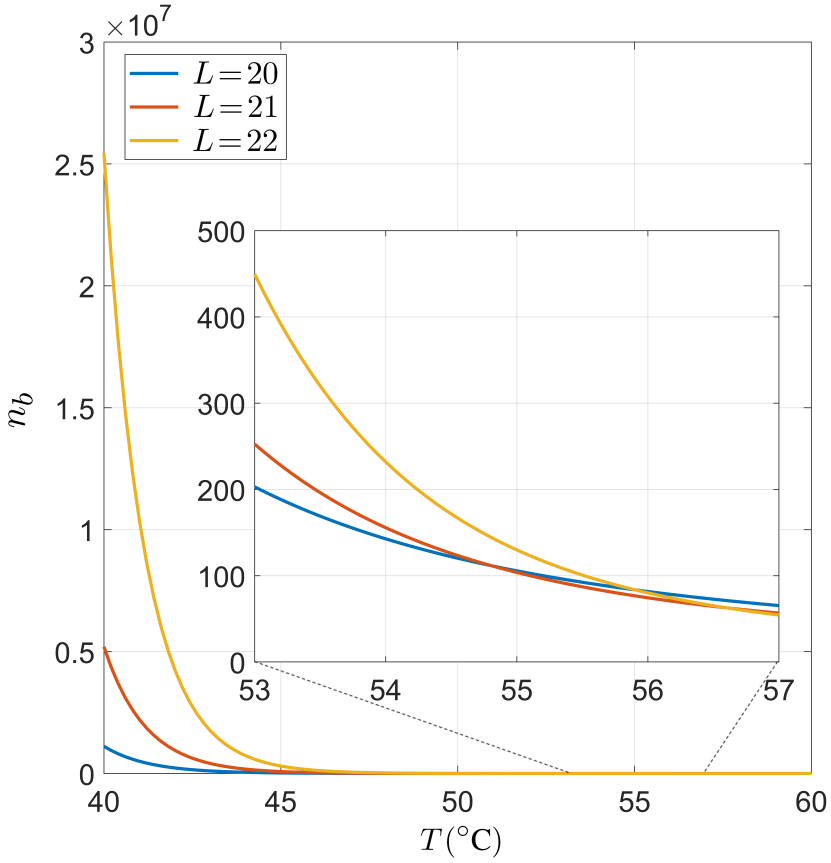

Figure A2. Dependence on the temperature T of the typical number of primers bounded per genome. We have considered fixed values for $c_p = 400$ nM, $c_g = 0.4$ fM, $L_0 = 6 \times 10^9$, whereas we have considered $L = 20$, $L = 21$ and $L = 22$ as typical primer lengths. The drop of n_b around the working temperature justifies the use of the approach used in the main text.

Implementing typical values of the order of magnitude used in the experiments in Equation (A3), we obtain the results shown in Figure A2. Specifically, we have used $c_p = 400$ nM, $c_g = 0.4$ fM, $L_0 = 6 \times 10^9$. Taking into account that ϕ_0 was around 0.1 for $L = 20$ and $T = 55$ °C, and that our estimation of the bounded primers per genome is an overestimation due to the purely independence hypothesis, we can conclude that our approach is consistent with the low saturation of sites in the working temperature. Note that we obtain relatively low values for $T = 55$ °C. This checkpoint validates the results obtained through the conditional probabilities presented in the main text.

Appendix C. Role of the Salt Concentration

Since the dependence of the thermodynamic parameters on salt concentration is logarithmic, it is necessary to consider relatively large changes in the concentration to observe significant changes. In this appendix, we have reobtained the dependence on the primer length of the pairing probability for PCR, multiplying the typical salt concentration, 55 mM, by either a factor 1/4 or 4. The result is shown in Figure A3. As shown, the relevant crossing length between the curves corresponding to ϕ_0 and $\phi_{\alpha_{e1}=0}$ decreases as the salt concentration is increased. This is a reasonable result since, generally speaking, higher salinity involves higher stability and thus significant selectivity is guaranteed even for shorter molecules.

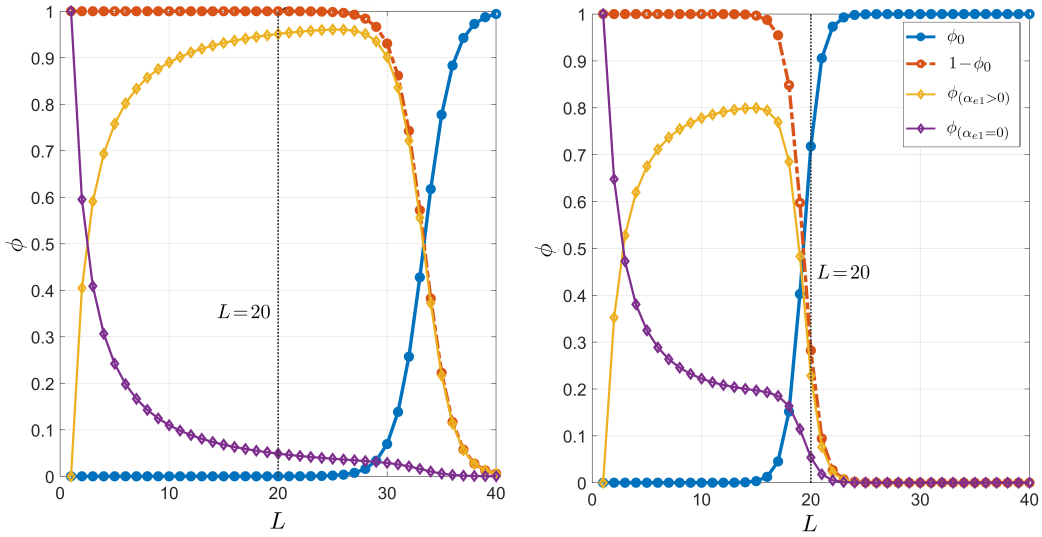

Figure A3. Dependence on the primer length L of the pairing probability for PCR. The plots are completely analogous to that shown in Figure 3 but with $[Na^+] = 55/4$ mM (**left panel**) and $[Na^+] = 4 \times 55$ mM (**right panel**).

Appendix D. Disassociation Times

As highlighted in the main text, it is important to stress that our approach describes a system in thermodynamic equilibrium. Therein, minimum enthalpy prevails for very low temperatures, but entropic states are promoted as soon as the temperature increases. Depending on the experimental situation, it is not trivial to ensure that, for any considered situation, during the annealing stage (with a typical duration lower than 60 s), the thermodynamic equilibrium is fully reached. The disassociation time of duplexes τ, which is the inverse of the rate at which the duplexes detach, depends on the depth of the free energy barrier from the double-strand state *ds* to the transition state *ts*: $\Delta G^\ddagger = \Delta H^\ddagger - T\Delta S^\ddagger = \Delta G_{ts} - \Delta G_{ds}$. Specifically, we assume $\tau \propto \exp\left[\Delta G^\ddagger/(R\,T)\right]$.

Since the enthalpic barrier is comparable with the binding enthalpy [29], i.e., $\Delta H^\ddagger \simeq -\Delta H$, we can estimate the temperature dependence of the typical disassociation time of a duplex as

$$\tau_{L,\vec{\alpha}}(T) = \tau_0 \exp\left[-\Delta H(L,\vec{\alpha})/(R\,T)\right], \tag{A4}$$

where we have expressed the enthalpy using our parametrization of the duplex quality and τ_0 contains the temperature-independent parameters, such as the entropic contribution to the unfolding barrier $\Delta S^\ddagger$. We are interested in studying the variation in the disassociation time of a certain duplex due to a temperature change ΔT,

$$\tau_{L,\vec{\alpha}}(T) \,/\, \tau_{L,\vec{\alpha}}(T+\Delta T) = \exp\left[-\frac{\Delta H(L,\vec{\alpha})}{R}\left(\frac{1}{T} - \frac{1}{T+\Delta T}\right)\right], \tag{A5}$$

where the dependence is affected exclusively by $\Delta H(L,\vec{\alpha})$. This expression enables us to compute the unfolding time variation of the duplexes, which we present in Table A5, ranging within typical conditions of the values ΔT, primer length L, and external mismatch α_{e1}, representing different kinds of defectiveness of the duplex. We set $T = 55\,^\circ$C, the same as in the main text. As we expected, the disassociation time has a significant temperature dependence for the well-paired duplexes, and a lower dependence for the duplexes with few well-paired nitrogenous bases; the main point of interest for PCR experiments is that, increasing the temperature by 2–3 °C, the typical disassociation time may be changed by an order of magnitude, with potential relevant effects on the pairing dynamics in the system.

References

1. Mullis, K.B.; Faloona, F.A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **1987**, *155*, 335–350. [PubMed]
2. Van Pelt-Verkuil, E.; van Belkum, A.; Hays, J.P. PCR primers. In *Principles and Technical Aspects of PCR Amplification*; Springer: Dordrecht, The Netherlands, 2008; pp. 63–90.
3. Chipman, L.B.; Pasquinelli, A.E. miRNA Targeting: Growing beyond the Seed. *Trends Genet.* **2019**, *35*, 215–222. [CrossRef] [PubMed]
4. Grimson, A.; Farh, K.K.; Johnston, W.K.; Garrett-Engele, P.; Lim, L.P.; Bartel, D.P. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol. Cell* **2007**, *27*, 91–105. [CrossRef] [PubMed]
5. Schirle, N.T.; Sheu-Gruttadauria, J.; MacRae, I.J. Structural Basis for microRNA Targeting. *Science* **2014**, *346*, 608–613. [CrossRef] [PubMed]
6. Kertesz, M.; Iovino, N.; Unnerstall, U.; Gaul, U.; Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **2007**, *39*, 1278–1284. [CrossRef]
7. Lane, A.N.; Jenkins, T. Thermodynamics of nucleic acids and their interactions with ligands. *Q. Rev. Biophys.* **2000**, *33*, 255–306 [CrossRef]
8. SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **1998**, *4*, 1460–1465. [CrossRef]
9. Breslauer, K.J.; Frank, R.; Blöcker, H.; Marky, L.A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 3746–3750. [CrossRef]
10. Owczarzy, R.; Vallone, P.M.; Gallo, F.J.; Paner, T.M.; Lane, M.J.; Benight, A.S. Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers* **1997**, *44*, 217–239. [CrossRef]
11. Plata, C.A.; Marni, S.; Maritan, A.; Bellini, T.; Suweis, S. Statistical physics of DNA hybridization. *Phys. Rev. E* **2021**, *103*, 042503. [CrossRef]
12. SantaLucia, J.; Hicks, D. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol.* **2004**, *33*, 415–440. [CrossRef] [PubMed]
13. Vologodskii, A.; Frank-Kamenetskii, M.D. DNA melting and energetics of the double helix. *Phys. Life Rev.* **2018**, *25*, 1–21. [CrossRef] [PubMed]
14. Owczarzy, R.; You, Y.; Moreira, B.G.; Manthey, J.A.; Huang, L.; Behlke, M.A.; Walder, J.A. Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures. *Biochemistry* **2004**, *43*, 3537–3554. [CrossRef]
15. Wu, J.H.; Hong, P.Y.; Liu, W.T. Quantitative effects of position and type of single mismatch on single base primer extension. *J. Microbiol. Methods* **2009**, *77*, 267–275. [CrossRef] [PubMed]
16. Bartel, D.P. MicroRNAs: Target recognition and regulatory functions. *Cell* **2009**, *136*, 215–233. [CrossRef] [PubMed]
17. Available online: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/ (accessed on 21 September 2020).
18. Available online: https://www.proteinatlas.org/humanproteome/cell/cell+line (accessed on 21 September 2020).
19. Kern, F.; Krammes, L.; Danz, K.; Diener, C.; Kehl, T.; Küchler, O.; Fehlmann, T.; Kahraman, M.; Rheinheimer, S.; Aparicio-Puerta, E.; et al. Validation of human microRNA target pathways enables evaluation of target prediction tools. *Nucleic Acids Res.* **2021**, *49*, 127–144. [CrossRef] [PubMed]
20. Satoh, J.; Tabunoki, H. Comprehensive analysis of human microRNA target networks. *BioData Min.* **2011**, *4*, 17. [CrossRef]
21. Liu, W.; Wang, X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol.* **2019**, *20*, 18. [CrossRef]
22. Li, J.; Kim, T.; Nutiu, R.; Ray, D.; Hughes, T.R.; Zhang, Z. Identifying mRNA sequence elements for target recognition by human Argonaute proteins. *Genome Res.* **2014**, *24*, 775–785. [CrossRef]
23. Xia, T.; Santa Lucia, J.; Burkard, M.E.; Kierzek, R.; Schroeder, S.J.; Jiao, X.; Cox, C.; Turner, D.H. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochemistry* **1998**, *37*, 14719–14735. [CrossRef]
24. Allawi, H.T.; Santa Lucia, J. Thermodynamics and NMR of internal G·T mismatches in DNA. *Biochemistry* **1997**, *36*, 10581–10594. [CrossRef] [PubMed]
25. Allawi, H.T.; Santa Lucia, J. Thermodynamics of internal C·T mismatches in DNA. *Nucleic Acid Res.* **1998**, *26*, 2694–2701. [CrossRef] [PubMed]
26. Allawi, H.T.; Santa Lucia, J. Nearest-neighbor thermodynamics of internal A·C mismatches in DNA: Sequence dependence and pH effects. *Biochemistry* **1998**, *37*, 9435–9444. [CrossRef] [PubMed]
27. Allawi, H.T.; Santa Lucia, J. Nearest neighbor thermodynamic parameters for internal G·A mismatches in DNA. *Biochemistry* **1998**, *37*, 2170–2179. [CrossRef]
28. Peyret, N.; Seneviratne, P.A.; Allawi, H.T.; Santa Lucia, J. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A·A, C·C, G·G, and T·T mismatches. *Biochemistry* **1999**, *38*, 3468–3477. [CrossRef] [PubMed]
29. Dupuis, N.F.; Holmstrom, E.D.; Nesbitt, D.J. Single-Molecule Kinetics Reveal Cation-Promoted DNA Duplex Formation Through Ordering of Single-Stranded Helices. *Biophys. J.* **2013**, *105*, 756–766. [CrossRef]