IDA

INSTITUTE FOR DEFENSE ANALYSES

Towards the Batch Synthesis of Long DNA

Steven Huntsman

October 2002

IDA Document D-2752

Log: H 02-001869

This work was conducted under contract DASW01 98 C 0067, Task DA-2-1955, prepared for DARPA/ATO. The publication of this IDA document does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-2752

Towards the Batch Synthesis of Long DNA

Steven Huntsman

PREFACE

This document was prepared for the Defense Advanced Research Projects Agency under a task entitled "Review and Assessment of Advanced Technologies in Molecular Biology."

Many thanks are due to Bert Barrois, Randy Good, Jim Heagy, and Barry Pallotta. Michael Frank and James Wetmur in particular offered kind guidance and helpful criticism.

We list several references here which were not cited in the main body of the text but which are related on some point or other:

Liu, H., et al., "Effect of electrostatic interactions on the structure and dynamics of a model polyelectrolyte. I. Diffusion," *J. Chem. Phys.* **109**, 7556 (1998). Examines the dynamics of a 20 bp dsDNA in light of electrostatic effects.

Chuang, T. J., and Eisenthal, K. B., "Theory of Fluorescence Depolarization by Anisotropic Rotational Diffusion," *J. Chem. Phys.* **57**, 5094 (1972). Explains the basic theory of the standard experimental technique for the quantitative study of rotational diffusion (notably the origins of the $[I_{\parallel}(t)-I_{\perp}(t)]/[I_{\parallel}(t)+2I_{\perp}(t)]$ expression for the fluorescence polarization anisotropy [FPA] and of the five decay constants in FPA).

Caspi, A., et al., "Diffusion and directed motion in cellular transport," *Phys. Rev. E* **66**, 011916 (2002). Discusses the way microtubules give rise to both sub- and super-diffusion in eukaryotic cells.

Rietman, E. A., *Molecular Engineering of Nanosystems*, Springer/AIP, New York (2001). Introduces many of the relevant concepts of thermodynamics and kinetics in the light of self-assembly.

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

The tradeoffs involved in designing and selecting a batch DNA manipulation (BDM) protocol for DNA synthesis involve

- fault-tolerance threshold (e.g., with respect to propagating mishybridizations)
- engineering complexity (e.g., number of microfluidic interconnects)
- time (e.g., number of [parallel or serial] hybridization/annealing steps).

We develop the tools to evaluate the fault-tolerance threshold while keeping other factors in mind. In particular, time mandates examination of the kinetic processes involved. In the course of examining aspects of synthesis protocols, we implicitly find that fault-tolerance threshold and engineering complexity are the principal considerations: time affects the others principally through (e.g.) batch size and imperfect annealing, and the engineering overhead associated with parallelization, respectively.

There appears to be no serious obstacle to successful implementation of the staggered ligation model for DNA synthesis. A useful test of this claim would be provided by using DNA for which sequencing by hybridization (SBH) works well, as the synthesis of such DNAs can also be expected to go well. That said, a significant amount of fundamental work on the processes involved still needs to be done. For instance, we find that the ultimate rate-limiting kinetic step of DNA hybridization (i.e., the translational component of nucleation) is poorly understood. Experiments using molecular beacons on an ensemble of de Bruijn-type DNAs may offer a way to gain this understanding. In any event, this is a crucial molecular-biological process that deserves a detailed quantitative analysis, which is currently lacking. More generally, it is necessary to obtain experimental data to enable the quantitative use of error models of the sort outlined here.

I. INTRODUCTION

The generic synthesis of long (> 100 base) DNA molecules with specific base sequences is at present infeasible due to error rates of the typical phosphoramidite chemical synthesis method.[1] On the other hand, two ss (single-stranded) DNAs can be joined or ligated into a single ds (double-stranded) DNA with an enzyme called DNA ligase; this technique, therefore, might offer a way to indirectly synthesize very long (>10 kb) DNAs (by staggered ligation of a batch of ssDNAs) with predetermined base sequences.[2]

Indeed, such a strategy was recently patented, and can be expected to represent the preferred method for the de novo laboratory synthesis of long DNA.[3] More generally there are undoubtedly profound clinical (e.g., gene therapeutic) and technological (e.g., DNA nanotechnological)[4] applications of long DNAs. Obtaining these DNAs on a production scale is difficult, however. There is a tradeoff (robustness of design vs. hybridization affinity, for instance) between the number of ligations that must be

[1] This solid-phase synthesis method is based on repeatedly adding monomers with protonated phosphoramidites (essentially, positively charged analogues of the phsophodiester backbones that comprise the DNA backbone) at their 3′ ends and protective groups at their 5′ ends to a growing polynucleotide anchored to a solid resin support. The 3′-phosphoramidite$^+$ + 5′-OH reaction, after being followed by oxidation via I_2 and deprotection of the 5′ end, results in a longer polymer. The yield from each step of this method exceeds 98 percent, but $(0.99)^{100} = 0.37$, and $(0.99)^{1000} = 4.32 \cdot 10^{-5}$. Isolation of the desired product nominally requires gel separation (although we believe that PCR [see footnote 7] might conceivably be used in its stead to effect the same result), and as such there must be a sufficient volume of product for gel extraction. See (e.g.) Stryer, L., *Biochemistry*, 4th ed., W H Freeman, New York (1995).

[2] DNA ligase catalyzes the formation of a covalent phosphodiester bond between two unbonded nucleotides. See (e.g.) Turner, P. C., et al., *Instant Notes in Molecular Biology,* Springer-Verlag, New York (1997). Stryer and Turner are our generic references for biochemistry and molecular biology and any otherwise unreferenced term or claim may safely be assumed to come from one of these. Similarly, our generic references for molecular biophysics are Daune, M., *Molecular Biophysics,* Oxford, New York (1999), and Cantor, C. R., and Schimmel, P. R., *Biophysical Chemistry, Part III: The behavior of biological macromolecules,* W H Freeman, New York (1980). Our generic references for chemical kinetics are Laidler, K. J., *Chemical Kinetics,* 3rd ed., Harper & Row, New York (1987), and van Santen, R. A., and Niemantsverdriet, J. W., *Chemical Kinetics and Catalysis,* Plenum, New York (1995).

[3] http://www.maxygen.com/newsview.php?listid=96 (U.S. Patent No. 6,368,861 announced 9 April 2002), Online, Available: 21 November 2002.

[4] Winfree, E., et al., “Design and self-assembly of two-dimensional DNA crystals,” *Nature* **394**, 539 (1998) and Seeman, N., et al., “New Motifs in DNA Nanotechnology,” *Nanotechnology* **9**, 257–273 (1998).

performed to assemble a long DNA and the length (hence the number) of short DNAs to be ligated (possibly also subject to end cleaving by a few robust sequence-specific exonucleases [which act to truncate certain end sequences]). There are also issues of thermodynamic and electrochemical homogeneity of the set of short oligonucleotides, which is a necessary precondition for a robust synthesis process. Many other, subtler obstacles are also present.

We attempt to address some of these problems here. For convenience, we consider a simple case (which for later reference we term the *baseline protocol*) of a batch of $2n$-mer ssDNAs which are to be used to synthesize an N-mer dsDNA ($N >> 2n$) with sticky ends (i.e., overhanging portions of ssDNA). Throughout this paper unless explicitly stated otherwise, we will assume that perfect complementarity is required for hybridization. For concreteness, let $X \equiv vwxy$ denote a desired $4n$-mer. (Throughout this paper X will denote either a long ssDNA or a product of batch ssDNA assembly [such as a sticky-ended dsDNA]: context will determine which.) Ideally, we would like to have a dsDNA X self-assemble from the batch ssDNAs in the presence of DNA ligase, but even this simple case presents difficulties. In particular, it is vital for some proposed batch assembly of X that the various batch ssDNAs hybridize as desired and not in some other fashion.[5] Below we depict a notionally successful batch assembly, with complementation denoted by an overbar.

$$\begin{array}{cccccccc} & 3'- & & \bar{w}_1 \cdots \bar{w}_n & \bar{x}_1 \cdots \bar{x}_n & \bar{y}_1 \cdots \bar{y}_n & \bar{z}_1 \cdots \bar{z}_n & -5' \\ 5'- & & v_1 \cdots v_n & w_1 \cdots w_n & x_1 \cdots x_n & y_1 \cdots y_n & -3' & \end{array}$$

We use an obvious shorthand for the $2n$-mer batch ssDNAs:

$$_5vw_3;\ _5xy_3;\ _5\tilde{z}\tilde{y}_3;\ _5\tilde{x}\tilde{w}_3$$
$$\left(\tilde{u} \text{ denotes the reverse complement of } u\right)$$

(We will typically drop the 5 and 3 subscripts.) Requiring that batch self-assembly should produce X places constraints upon the various batch ssDNAs. We consider three particular types of hybridization for pairs of equal-length batch ssDNAs: type I, in which there are no sticky ends; type II5 (not to be confused with anything involving restriction enzymes), in which the sticky ends are both 5′; and type II3, in which the sticky ends are both 3′. Types II5 and II3 are collectively referred to as type II. We further distinguish

[5] DNAs amenable to SBH can be expected to be similarly amenable to synthesis by staggered annealing and ligation for combinatorial reasons. See (e.g.) Peuzner, P.A., *J. Biomolecular Structure and Dynamics* **7**, 63 (1989).

type II5a and II3a hybridizations (collectively, type IIa), in which the sticky ends are *n* bases long.

Now type I hybridizations (which of course are undesirable, though not the principal obstacles to synthesis for thermodynamic reasons) can be catalogued: recalling that we assume perfect complementarity for hybridization, a type I hybridization requires one of the following to hold (□ denotes logical AND):

$$v = \tilde{w};\ (v = \tilde{y})\square\underline{(w = \tilde{x})};\ (v = y)\square(w = z);\ v = w = x;\ x = \tilde{y};$$
$$x = y = z;\ w = x = y;\ y = \tilde{z};\ (w = \tilde{z})\square\underline{(x = \tilde{y})};\ w = \tilde{x}$$

where we use an underbar to denote a logical clause that appears alone elsewhere in the listing.

A similar effort yields the following list of identities which would allow for nontrivial (i.e., undesirable) IIa hybridizations by similar cataloguing:

$$v = \tilde{v};\ w = \tilde{w};\ x = \tilde{x};\ y = \tilde{y};\ z = \tilde{z};\ w = \tilde{y};$$
$$v = \tilde{x};\ \underline{w = y};\ v = z;\ \underline{v = x};\ \underline{x = z};\ x = \tilde{z}$$

where we use an underbar here to denote an equality that is also contained in some type I clause.

We can represent all of these potential mishybridization equalities graphically:

mis	v	w	x	y	z	$\tilde{v}$	$\tilde{w}$	$\tilde{x}$	$\tilde{y}$	$\tilde{z}$
v			○	+	○	○	□	○	+	
w				○	+	□	○	□	○	+
x	○				○	○	□	○	□	○
y	+	○				+	○	□	○	□
z	○	+	○				+	○	□	○
$\tilde{v}$	○	□	○	+				○	+	○
$\tilde{w}$	□	○	□	○	+				○	+
$\tilde{x}$	○	□	○	□	○	○				○
$\tilde{y}$	+	○	□	○	□	+	○			
$\tilde{z}$		+	○	□	○	○	+	○		

□□ type I mishybridization;
+ □ type I mishybridization only with 2 equality clauses,
with clause pairing indicated by dotted lines;
○ □ type IIa mishybridization

Of course, as stated previously, it is the type IIa mishybridizations that are really of concern, since it ought to be possible to employ a PCR[6] protocol in which type I mishybridizations are at worst a secondary error source (in each cycle any type I mishybridized dsDNAs will just revert to batch ssDNAs). But we can easily search for these using the identities above. Of subtler and more persistent concern are generic mishybridizations which arise because of frame-shifting (in the special case of pairs of batch ssDNAs [as opposed to semi-ligated "DNA Frankensteins"], these are just type II mishybridizations): for example, for some number $m < n$, it might happen that the terminal m bases of two batch ssDNAs are complementary and a "gapped" mishybridization takes place. If m were taken to be 1, then this obviously would be an unreasonable constraint to consider; however, DNA ligase generally requires "footprints" (which we quantify now by the variable m) of more than a single base for its operation. Usually m can be considered to be at least 4. Indeed the number of base pairs required to initiate hybridization nucleation is about three or four, and so happily these constraints are consistent with reality.

With this in mind, we state a principal constraint of concern to us (a so-called "frame-shift" constraint): for some $m > 1$ and arbitrary batch ssDNAs x, y, we ought to have

$$x_1 \ldots x_l \neq y_1 \ldots y_l;$$

$$x_{n-l+1} \ldots x_n \neq y_{n-l+1} \ldots y_n$$

for $m < l < n$. In particular, the frame-shift constraint here requires that any sticky ends of dsDNAs must be n base pairs long. It is easy to see that it cannot always be satisfied, and so (at least for now) we must make some assumptions about the structure of X. Probably the simplest and best one we can make is that the base sequence of X is chosen at random with respect to the uniform measure.

[6] PCR (the polymerase chain reaction) serves to amplify dsDNA by repeatedly (1) separating or denaturing into two ssDNAs at 95 ºC followed by (2) primer annealing at ~55 ºC and (3) primer-initiated polymerization at 72 ºC in the presence of deoxyribonucleic triphosphates (dNTPs; e.g., dATP) and Mg^{2+}. Typically, ~30 PCR cycles are used for amplification.

II. ANALYSIS OF THE BASELINE PROTOCOL

It appears a reasonable first approximation to the general problem for the baseline protocol to consider only type IIa mishybridizations, along with the frame-shift constraint. It also seems reasonable to assume that if k type IIa mishybridizations *might* occur, then, as long as 2^{-k} is not prohibitively small, we could simply extract X (which is presumably assembled and could be isolated by fluorescent probes) from a pool of $\leq 2^k$ different dsDNAs of the same size (this pool could be obtained via gel electrophoresis, e.g., although in practice we would only need to isolate a single copy of X, which could be facilitated by the aforementioned probes), and so we will concentrate on the frame-shift constraint alone for the remainder of this paper.

Two possible frame-shift scenarios, which we term minus- and plus-shift hybridizations (MSH and PSHs), of length l are respectively depicted below:

$$\begin{array}{l} \qquad\qquad 3'- \quad \bar{w}_{n-l+1}\ldots\bar{w}_n \quad \bar{w}_{l+1}\ldots\bar{w}_n \quad \bar{x}_1\ldots\bar{x}_n \quad -5' \\ 5'- \quad v_1\ldots v_n \quad w_1\ldots w_{n-l} \quad \underbrace{w_{n-l+1}\ldots w_n}_{\text{length } l<n} \qquad -3' \end{array}$$

$$\Rightarrow\ w_1\ldots w_l = w_{n-l+1}\ldots w_n;$$

.

$$\begin{array}{l} \qquad\qquad 3'- \quad \bar{v}_{n-l+1}\ldots\bar{v}_n \quad \bar{w}_1\ldots\bar{w}_n \quad \bar{x}_{n-l}\ldots\bar{x}_n \quad -5' \\ 5'- \quad v_1\ldots v_{n-l} \quad \underbrace{v_{n-l+1}\ldots v_n}_{\text{length } l<n} \quad w_1\ldots w_n \qquad -3' \end{array}$$

$$\Rightarrow\ v_{n-l+1}\ldots v_n w_1\ldots w_n = w_1\ldots w_n x_1\ldots x_l$$

Essentially the same conditions are required if we stagger in the opposite direction, that is, with 3′ sticky ends, and although both plus and minus shifts are potential bogeymen, we concentrate on the minus-shift condition, since the ligation mechanism and PCR cycling will presumably tend to inhibit any deleterious effects of the plus-shift condition on a synthesis protocol. That is, PSHs are hindered from propagating through later stages of a batch self-assembly.

Assuming (as we do throughout unless stated otherwise) that the bases of X are selected independently and uniformly at random,[7] it can be shown[8] that for a word of

[7] This is not always a realistic assumption; for example, the *Alu* and *L1* motifs vary fairly little from instance to instance and together comprise nearly 10 percent of the human genome. Base frequencies will also (for instance) display correlations stemming from broken symmetries in the genetic code and

length l—which of course ends (or begins) with probability 4^{-l} at any given symbol placemarker s—contained in an infinite word, the probability of finding the next *nonoverlapping*[9] occurrence of the word beginning at placemarker $s + t$ is

$$4^{-l}\left(1-4^{-l}\right)^{t-1}.$$

Hence the probability of a possible MSH of length l is

$$p_l \equiv 4^{-l}\left(1-4^{-l}\right)^{n-l-1}.$$

It follows that the probability of M possible MSHs of length l in a batch of B ssDNAs is

$$\binom{B}{M} p_l^M (1-p_l)^{B-M};$$

and we explicitly make the operational assumption that the size of the batch is $N/n + 1$ (i.e., the batch consists of the $2n$-mers intended to self-assemble as X and no other ssDNAs). To get the probability of at least one *possible* MSH of length l, we simply sum over M (which of course gives 1) and subtract the $M = 0$ term, giving

$$1-(1-p_l)^{N/n+1}.$$

Suppose for the moment that if an MSH of length l is possible for a pair of ssDNAs, say as depicted above: then for an MSH of length $l + 1$ to be possible, it is necessary and sufficient that the equalities $w_1 = w_{n-l}$ and $w_{l+1} = w_n$ hold. But it must also be the case that $w_1 = w_{n-l+1}$ and $w_l = w_n$ hold, so that $w_{n-l} = w_{n-l+1}$ and $w_l = w_{l+1}$. Inductively we see that it must in fact be the case that $w_{n-l} = w_{n-l+1} = \ldots = w_n = w_{l+1}$, and it follows that both the initial and terminal portions of the half-oligo w must be poly(·), where · indicates some base or its complement. Similarly, for an MSH of length $l + l'$ to be possible, it is necessary and sufficient that both the initial and terminal portions of the half-oligo w must be poly(·), where · indicates some sequence of length l' or its complement. From

the emergence of structural motifs (both coded by and intrinsic to the DNA itself). More generally, DNA sequences are often modeled as stationary Markov processes, and these display statistical periodicities. See Durbin, R., et al., *Biological Sequence Analysis,* Cambridge University Press, Cambridge (1998).

[8] Percus, J. K., *Mathematics of Genome Analysis,* Cambridge University Press, Cambridge (1998).

[9] The relative impact of self-overlap is not significant, and we neglect it chiefly for simplicity; its inclusion ought not to change hard numbers by much, and it should not affect order-of-magnitude estimates at all.

this we see that there are well-defined least and greatest lengths for a possible MSH between a given pair of ssDNAs: we denote these lengths by l_0 ($\geq m$) and l , respectively.[10] Finally, we note that if an MSH of length greater than $\lfloor n/2 \rfloor$ is possible, then the entire half-oligo w must be of the form poly(·).[11] With this noted, *we neglect such MSHs for the remainder of the text unless explicitly stated otherwise*. Extension to incorporate self-overlapping words is possible, however.[12]

For example, if we try to construct a 10,005-mer using 668 30-mer batch ssDNAs, there is a 92-percent chance of having a *possible* MSH of length 4; a 47-percent chance for length 5; a 15-percent chance for length 6; and a 4-percent chance for length 7 (after this the above formalism breaks down [although it turns out to still be a good approximation] since then we will have overlapping words).

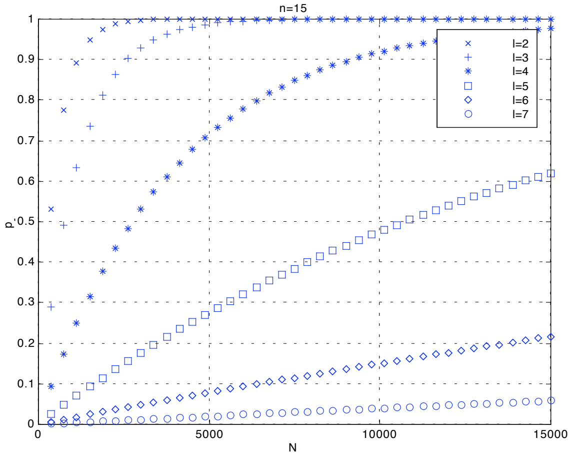


Figure 1. Probabilities of at least one possible MSH for *n* = 15 and varying *l*, *N*

More generally, suppose that the bases of X are selected at random, but with respect to a nonuniform probability distribution which is sufficiently well-behaved. Assume further that W (here a word occurring at an end of some batch ssDNA) is not self-overlapping (or that the effects of overlap are small and hence neglected). Then the probability of a possible MSH of length l is

[10] Heuristically, we might enforce $m = \lceil \log_4 B \rceil$. It is convenient that in the regimes we are interested in (i.e., $n \sim 30$), the RHS is 4 or 5.

[11] Such periodicity will, in general, occur uniformly at random. Structurally significant 3-base periodicities are typically characteristic of protein-coding regions, and 10.5-base periodicities generally indicate α-helix coding or structural motifs for DNA incorporation into chromatin. Trifonov, E. N., "3-, 10.5-, 200- and 400-base periodicities in genome sequences," *Physica A* **249**, 511–516 (1998).

[12] Percus, *Mathematics of Genome Analysis*.

$$p_W = \Pr(W) \cdot \left(1 - \Pr(W)\right)^{n-l-1}.$$

It follows that the probability of at least one possible MSH of length l in a batch of ssDNAs is

$$\max_W \left[1 - \left(1 - p_W\right)^{N/n+1}\right] = 1 - \left(1 - \max p_W\right)^{N/n+1}.$$

From this we immediately see that the assumption of a uniform random distribution of bases gives an optimistic answer (i.e., the probability above induced by any other choice of distribution will inevitably be greater than the earlier estimates). On the other hand, it allows for a pessimistic estimate also: if we assume that the words comprising the first l bases of each half of all the batch ssDNAs are all distinct and not self-overlapping, then we have that

$$p_W \le \frac{1}{\frac{N}{n}+1} = \frac{n}{N+n} \Rightarrow \Pr\left(\text{at least one MSH of length } \ge l\right) \le 1 - \left(\frac{N}{N+n}\right)^{N/n+1}.$$

(However, this is a very crude pessimistic estimate.) For the 10,005-mer with 668 30-mer batch ssDNAs, it gives .63 as an upper bound on the probability.

III. GENERALIZED PROTOCOLS

Let us change focus now, and consider a fixed dsDNA *X*. We want to evaluate the error rates associated with the self-assembly of *X* with batch ssDNAs of variable length. In particular, we want to obtain the optimal (de)composition of *X* into short ssDNAs while taking MSH errors into account. Toward this end (but really so that we can write code that properly treats compositions—see Appendix C), first note the canonical form of such a composition (a double arrow indicates the reverse of a word, and over/under arrows are used simply to indicate orientation):

$$\overrightarrow{w_{[1]}^{1}X} \equiv 5\square\square\, w_{[1]}^{1}w_{[2]}^{1}w_{[1]}^{2}\cdots w_{[2]}^{B/2\square 1}w_{[1]}^{B/2}w_{[2]}^{B/2}\,\square 3\square;$$

$$\underleftarrow{X\overleftrightarrow{w}_{[1]}^{B/2+1}} \equiv 3\square\square\,\overline{w}_{[2]}^{1}\overline{w}_{[1]}^{2}\overline{w}_{[2]}^{2}\cdots\overline{w}_{[1]}^{B/2}\overline{w}_{[2]}^{B/2}\overleftrightarrow{w}_{[1]}^{B/2+1}\,\square 5\square = 5\square\square\,\underbrace{w_{[1]}^{B/2+1}}_{w_{[1]}^{B/2+1}}\underbrace{\tilde{w}_{[2]}^{B/2}}_{w_{[2]}^{B/2+1}}\underbrace{\tilde{w}_{[1]}^{B/2}}_{w_{[1]}^{B/2+2}}\cdots\underbrace{\tilde{w}_{[2]}^{2}}_{w_{[2]}^{B\square 1}}\underbrace{\tilde{w}_{[1]}^{2}}_{w_{[1]}^{B}}\underbrace{\tilde{w}_{[2]}^{1}}_{w_{[2]}^{B}}\,\square 3\square$$

$$\equiv 5\square\square\, w_{[1]}^{B/2+1}w_{[2]}^{B/2+1}w_{[1]}^{B/2+2}\cdots w_{[2]}^{B\square 1}w_{[1]}^{B}w_{[2]}^{B}\,\square 3\square = 3\square\square\,\overleftrightarrow{w}_{[2]}^{B}\overleftrightarrow{w}_{[1]}^{B}\overleftrightarrow{w}_{[2]}^{B\square 1}\cdots\overleftrightarrow{w}_{[1]}^{B/2+2}\overleftrightarrow{w}_{[2]}^{B/2+1}\overleftrightarrow{w}_{[1]}^{B/2+1}\,\square 5\square$$

$$\square\ \left(w_{[2]}^{B}=\tilde{w}_{[2]}^{1}\right)\square\left(w_{[1]}^{B}=\tilde{w}_{[1]}^{2}\right)\square\left(w_{[2]}^{B\square 1}=\tilde{w}_{[2]}^{2}\right)\square\cdots\square\left(w_{[1]}^{B/2+2}=\tilde{w}_{[1]}^{B/2}\right)\square\left(w_{[2]}^{B/2+1}=\tilde{w}_{[2]}^{B/2}\right)$$

$$\square\ w_{[h]}^{B/2+k}=\tilde{w}_{[h]}^{B/2\square k+3\square h}\quad\left(\left(h,k\right)\square\left\{1,2\right\}\square\left\{1,\ldots,B/2\right\}\setminus\left\{\left(1,1\right),\left(1,B/2+1\right)\right\}\right)$$

$$\square\ \vec{X}\equiv 5\square\square\left\{w_{[2]}^{k}w_{[1]}^{k+1}\right\}_{k=1}^{B/2\square 1}w_{[2]}^{B/2}\,\square 3\square;$$

$$\underleftarrow{X}\equiv 5\square\square\left\{w_{[2]}^{B/2+k}w_{[1]}^{B/2+k+1}\right\}_{k=1}^{B/2\square 1}w_{[2]}^{B}\,\square 3\square = 5\square\square\left\{\tilde{w}_{[2]}^{B/2\square k+3\square h}\tilde{w}_{[h]}^{B/2\square k+2\square h}\right\}_{k=1}^{B/2\square 1}\tilde{w}_{[2]}^{1}\,\square 3\square$$

This puts us in position to state the general problem: For *X* fixed but generic (i.e., with a uniformly random base distribution) we might consider two problems: the design of protocols for synthesizing *X* from a batch with a fixed number of ssDNA species, each of variable length—*but as close to equal length as possible*—ssDNAs and from a batch itself of variable size. The former type of protocol is notionally easier to analyze, whereas the latter type is easier to design. Given a particular designed protocol, we can use the earlier analysis in conjunction with a probability distribution on the batch size characteristics to analyze this protocol.

$$X = \begin{array}{llllll} & 3\square\square & \overline{w}^{1}_{[2]} & \overline{w}^{2}_{[1]} & \cdots & \cdots \quad \square 5[\\ 5\square\square & w^{1}_{[1]} & w^{1}_{[2]} & \cdots & \cdots & \square 3\square \end{array}.$$

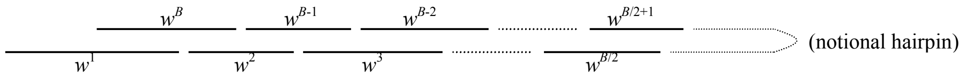


Figure 2. Upper: (de)composition of *X*. Lower: formal treatment of (dsDNA) *X* as hairpinned ssDNA.

We begin with the analysis. Let $L(w)$ denote the length of a word. An MSH of length l between the initial pair of batch ssDNAs can be depicted as

$$\begin{array}{lllll} & 3\square\square & \overline{w}^{1}_{L\left(w^1_{[1]}w^1_{[2]}\right)\square l+1}\square\ \overline{w}^{1}_{L\left(w^1_{[1]}w^1_{[2]}\right)} & \overline{w}^{1}_{L\left(w^1_{[1]}\right)+l}\square\ \overline{w}^{1}_{L\left(w^1_{[1]}w^1_{[2]}\right)} & \overbrace{\overline{w}^{2}_{1}\square\ \overline{w}^{2}_{L\left(w^2_{[1]}\right)}}^{\overline{w}^2_{[1]}} \quad \square 5[\\ 5\square\square\ \underbrace{w^1_1\square\ w^1_{L\left(w^1_{[1]}\right)}}_{w^1_{[1]}} & w^1_{L\left(w^1_{[1]}\right)+1}\square\ w^1_{L\left(w^1_{[1]}w^1_{[2]}\right)\square l} & \underbrace{w^1_{L\left(w^1_{[1]}w^1_{[2]}\right)\square l+1}\square\ w^1_{L\left(w^1_{[1]}w^1_{[2]}\right)}}_{\text{length } l<L\left(w^1_{[2]}\right)} & \square 3\square & \end{array}$$

$$\square\ w^1_{L\left(w^1_{[1]}w^1_{[2]}\right)\square l+1}\square\ w^1_{L\left(w^1_{[1]}w^1_{[2]}\right)} = w^1_{L\left(w^1_{[1]}\right)+1}\square\ w^1_{L\left(w^1_{[1]}\right)+l}$$

,

and the general case is the same after some superscript/subscript replacements.

For a first (probabilistic) stab at it, we will, as before, make the operational assumption that the batch consists only of the ssDNAs intended to self-assemble as X. As before, the probability of a possible MSH of length l ($l \leq L/2$) at the kth batch oligo/word (here $1 \leq k \leq B/2$) is

$$p_{l;k} \equiv 4^{\square l}\left(1\square 4^{\square l}\right)^{L\left(w^k_{[1]}w^k_{[2]}\right)\square l\square 1}.$$

Let the number of batch ssDNAs of length L be denoted by $\#(L)$. Then the probability of M possible MSHs of length l in a sub-batch of $\#(L)$ ssDNAs of length L is

$$\Pr\left(\# MSH_l^{possible} = M\right) = \begin{array}{c}\square\#(L)\square \\ \square\ M\ \square\end{array} p^M_{l;L}\left(1\square p_{l;L}\right)^{\#(L)\square M};$$

where we abusively write $p_{l;L}$ for all of the $p_{l;k}$ accounted for in the expression (since they are all equal). Consequently, the probability of at least one possible MSH of length l in the entire batch is

$$\Pr\left(\# MSH_l^{possible} \mid \left\{\#(L)\right\} > 0\right) = 1 - \left(1 - p_{l;L}\right)^{\#(L)}$$

$$\Pr\left(\# MSH_l^{possible} > 0\right) = \sum_L \underbrace{\Pr\left[\# MSH_l^{possible} \mid \left\{\#\left(L\left(w^k\right)\right)\right\} > 0\right]}_{=1-\left(1-p_{l;L}\right)^{\#(L)}} \cdot \Pr\left[\left\{\#\left(L\left(w^k\right)\right)\right\}\right]$$

$$= \sum_{L=\min L\left(w^k\right)}^{\max L\left(w^k\right)} \frac{1-\left(1-p_{l;L}\right)^{\#(L)}}{\#(L)} = 1 - \sum_{L=\min L\left(w^k\right)}^{\max L\left(w^k\right)} \frac{\left(1-p_{l;L}\right)^{\#(L)}}{\#(L)} \approx \sum_{L=\min L\left(w^k\right)}^{\max L\left(w^k\right)} \frac{1-e^{-p_{l;L}\#(L)}}{\#(L)}.$$

Rewriting, this is

$$\Pr\left(\# MSH_l^{possible} > 0\right) = 1 - \sum_{L=\min L\left(w^k\right)}^{\max L\left(w^k\right)} \frac{\left[1 - 4^{-l}\left(1-4^{-l}\right)^{L-l-1}\right]^{\#(L)}}{\#(L)} \approx \sum_{L=\min L\left(w^k\right)}^{\max L\left(w^k\right)} \frac{1-e^{-\frac{\#(L)}{4^l}\left(1-4^{-l}\right)^{L-l-1}}}{\#(L)}.$$

The probability of *actually* having any MSHs at all (as opposed to *possible* MSHs of some fixed length) is considerably more involved, even in this simple approximation, and requires some gauge of relative hybridization affinities and the actual physics of hybridization. The reason for this is that we have to have a decent way to weigh MSH probabilities for various lengths. This brings us into the realm of chemical physics.

IV. THERMODYNAMIC ASPECTS

The SantaLucia nearest neighbor (NN) model[13] for DNA *N*-mer (proper) duplex thermodynamics is essentially contained in the formula

$$\Delta\square(D) = \sum_k \Delta\square(D(k), D(k+1)) + \Delta\square_{\partial}(D(1)) + \Delta\square_{\partial}(D(N)) + \Delta\square_{SC}\cdot\square(SC(D)),$$

where $\square$ refers variously to the free energy G, the enthalpy H, or the entropy S. D is a proper (i.e., strand-strand complement) duplex with kth base pair $D(k)$; the various component $\Delta\square(\cdot)$ constants are given in Table 1, and $\square(SC(D))$ is unity if and only if D is self-complementary and zero otherwise. We can generally ignore the last term in the NN formula; the second (initiation) and third (termination) terms account for sequence-independent factors such as counterion condensation.

Table 1. Parameters for the SantaLucia Model (@ 37 °C, 1 M Na$^+$)

NN sequence	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	ΔS (cal mol^{-1} K^{-1})
AA/TT	–1.00	–7.9	–22.2
AT/TA	–0.88	–7.2	–20.4
TA/AT	–0.58	–7.2	–21.3
CA/GT	–1.45	–8.5	–22.7
GT/CA	–1.44	–8.4	–22.4
CT/GA	–1.28	–7.8	–21.0
GA/CT	–1.30	–8.2	–22.2
CG/GC	–2.17	–10.6	–27.2
GC/CG	–2.24	–9.8	–24.4
GG/CC	–1.84	–8.0	–19.9
∂(CG)	0.98	0.1	–2.8
∂(AT)	1.03	2.3	4.1
SC	0.43	0	–1.4

The relationship $\Delta G = \Delta H - T\Delta S$ allows for extrapolations from the baseline (37°C or 310°K), provided that the heat capacity differences between folded and denatured states are taken into consideration.[14] The empirical relationships

[13] SantaLucia, J., Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc. Nat'l Acad. Sci.* **95**, 1460 (1998).

$$\Delta G_{oligo}([\mathrm{Na}^+]) \sim \Delta G(1\ \mathrm{M}) - 0.114\cdot\log([\mathrm{Na}^+])\cdot L_{hyb}$$

$$\Delta G_{polymer}([\mathrm{Na}^+]) \sim \Delta G(1\ \mathrm{M}) - 0.75\cdot\log([\mathrm{Na}^+])\cdot L_{hyb} - 0.20$$

$$\Delta S([\mathrm{Na}^+]) \sim \Delta S(1\ \mathrm{M}) + 0.368\cdot\log([\mathrm{Na}^+])\cdot L_{hyb}$$

allow for extrapolation to different ionic environments.[15] The length dependence of the energetics (i.e., the difference between "oligo" and "polymer"), presumably due to counterion condensation effects that arise in the polymeric regime, is negligible for shorter duplexes.

An averaged and simplified form of the SantaLucia model predicts that the Gibbs free energy of hybridization will be a negative-slope linear function of the hybridization length l. Using the Gibbs distribution we get that $\Pr(E_1)/\Pr(E_0) = e^{-\beta\Delta E} = e^{C\Delta l}$, where $\beta = (k_BT)^{-1}$, k_B is the Boltzmann constant ($3.3\cdot10^{-27}$ kcal K^{-1}), and C is a constant depending on temperature.[16] At 37 °C $\beta = 9.8\cdot10^{23}$ kcal^{-1}. The simplified model predicts that the addition of a single base pair yields changes in the free energy, enthalpy, and entropy: $\langle\Delta G_{37}\rangle = -1.4$ kcal mol^{-1}, $\langle\Delta H_{37}\rangle = -8.3$ kcal mol^{-1}, and $\langle\Delta S_{37}\rangle = -22$ cal mol^{-1} K^{-1}, respectively. We also note that $\langle\Delta G_{55}\rangle$ is predicted to be -1 kcal mol^{-1}.

Dangling ends and NN (aka stacking) interactions contribute significantly to the thermodynamic picture for short oligos.[17] The incorporation of dangling ends into the SantaLucia model has been taken into account by Bommarito et al.[18] Dangling ends are typically stabilizing; only adenine dangling ends show much variation in the degree of (de)stabilization. The NN framework appears to give a good approximation for the thermodynamic properties of dangling ends. Below we illustrate the magnitudes of the dangling end NN parameters:

[14] We are loath to do this explicitly: quantitative enthalpic considerations can open a box of worms that remains closed when dealing with free energies. See Naghibi, H., et al., "Significant differences between van't Hoff and calorimetric enthalpies," *Proc. Nat'l Acad. Sci.* **92**, 5597 (1995).

[15] For Mg^{2+} ions we can use an equivalent Na^+ concentration: $[\mathrm{Na}^+]_{\mathrm{equivalent}} = 4\ [\mathrm{Mg}^{2+}]$. Wetmur, J. G., cited as a personal communication (1997) in Hartemink, A. J., and Gifford, D. K., "Thermodynamic Simulation of Deoxynucleotide Hybridization for DNA Computation," in Wood, D., ed., *Proceedings of the 3rd DIMACS Workshop on DNA Based Computers, June 23–25, 1997*, Vol. 48 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, AMS, Providence, R.I. (1999).

[16] β will be used to denote both $(k_BT)^{-1}$ and $(RT)^{-1}$, where R (the ideal gas constant) is roughly $2\cdot10^{-3}$ kcal mol^{-1} K^{-1}. Context (in the guise of molar units) should suffice to determine which is being used.

[17] Wetmur, J. G., "Physical Chemistry of Nucleic Acid Hybridization," in Wood, D., ed., *Proceedings of the 3rd DIMACS Workshop on DNA Based Computers, June 23–25, 1997*, Vol. 48 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, AMS, Providence, R.I. (1999).

[18] Bommarito, S., et al., "Thermodynamic parameters for DNA sequences with dangling ends," *Nuc. Acids Res.* **28**, 1929 (2000).

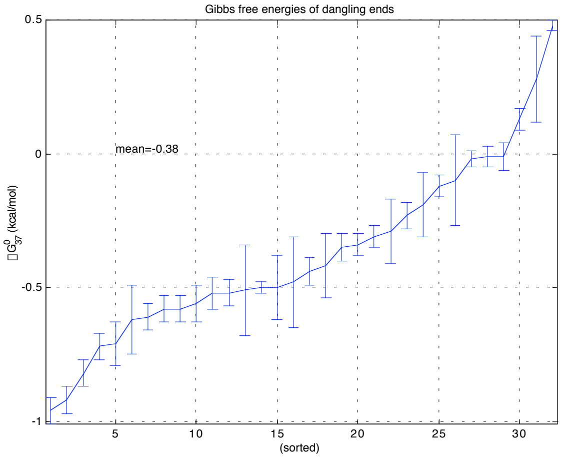


Figure 3. Gibbs parameters from Bommarito et al.

This framework being established, it is reasonable (for *uniformly random* ssDNAs) to make the zeroth-order approximation that

$$\frac{\Pr\left(MSH_l^{actual}\right)}{\Pr\left(MSH_{l\square}^{actual}\right)} = e^{C_\square\cdot(l\square l\square}.$$

That is, the hybridization mechanism provides this as a decent ansatz (statisticians might say Bayesian prior, but we are using physics, and so this has some [albeit tenuous] footing in reality).

Now, neglecting periodicities that give rise to multiple possible MSHs (their impact will be negligible anyway), we have the crude estimate (recall the operational definition of *m* from Chapter 1)

$$\Pr\left(\#\,MSH^{actual}>0\right)\square\Pr\left(MSH_m^{actual}\right)\overset{l_1}{\underset{l=m}{\square}}e^{C_\square\cdot(l\square m)}\Pr\left(\#\,MSH_l^{possible}>0\right)$$

$$=\Pr\left(MSH_m^{actual}\right)\overset{l_1}{\underset{l=m}{\square}}e^{C_\square\cdot(l\square m)}\overset{\max L\left(w^k\right)}{\underset{L=\min L\left(w^k\right)}{\square}}\frac{1\square\square 1\square 4^{\square l}\left(1\square 4^{\square l}\right)^{L\square l\square 1}\square^{\#(L)}}{\#(L)}.$$

For a non-uniform distribution of words this will instead take the form

$$\Pr\left(\#\,MSH^{actual}>0\right)\square\Pr\left(MSH_m^{actual}\right)\overset{l_1}{\underset{l=m}{\square}}e^{C_\square\cdot(l\square m)}\overset{\max L\left(w^k\right)}{\underset{L=\min L\left(w^k\right)}{\square}}\frac{1\square\square 1\square\max_{W_l}\square\Pr\left(W_l\right)\cdot\left(1\square\Pr\left(W_l\right)\right)^{L\square l\square 1}\square^{\#(L)}}{\#(L)}.$$

Operationally, we might want to minimize this quantity subject to fixed X, B, and l_1, while varying the lengths L. But we have not taken the sequences of the batch ssDNAs into account at all here, and so this quantity is useful only as a relative estimate of failure rates. In particular, it gives us no information about the likelihood of error for a particular composition of a particular dsDNA—only a random dsDNA. The reader will probably not be surprised to see that this quantity will be minimized for equal lengths L. If the lengths are all equal, then the inner sum has only one term, and it is easy to see that in fact the likelihood of success is maximized.

This is not to say that the variable-length composition scheme is less likely to succeed. Rather, it is a reminder that *we want to keep the lengths for a composition as close to equal as possible*: that is, it supports one of our operational constraints. It will also be a way to minimize MSHs when the assumption of perfect complementarity for hybridization is weakened. As a penultimate note we remark that for designed sequences it appears that hybridization errors due to base mismatch will decrease exponentially as a function of hybridization length, however, and so this should not present much of a problem in practice.[19]

In any case, as Carbone and Gromov put it: "The Gibbs measure does not tell one how the actual hybridization process develops but only describes the *equilibrium* stage of hybridization. Unlike the statistical ensembles usually studied in physics, the *relaxation time* (i.e., the time needed to reach equilibrium) in biology is relatively long, and the road to equilibrium may be very bumpy"[20] (emphasis in the original). The equilibrium analysis paints a deceptively simple picture, one that obscures many important points. With this in mind, we next consider kinetic processes.

[19] Deaton, R., and Rose, J. A., "Simulations of Statistical Mechanical Estimates of Hybridization Error," Preprint (2000).

[20] Carbone, A., and Gromov, M., "Mathematical slices of molecular biology," Preprint IHES/M/01/03, 2001. This paper also asks the intriguing question, "Consider a random population of strands of various length, i.e., that is a measure on the sequence space. How does this measure evolve in time…?"

V. KINETIC ASPECTS
I—INTRODUCTION, OVERVIEW, AND APPROACH

A. INTRODUCTION

In principle, the batch DNA synthesis goes like

$$\underbrace{\{w\} \equiv \underset{w}{\square}\,\square_w w}_{\text{batch ssDNAs}} \square \underbrace{\{x\} \equiv \square_X X + \underset{MSH}{\square}\,\square_{MSH} MSH + \underset{other}{\square}\,\square_{other} other}_{X\text{, mishybridized partial products, and other waste}};$$

$$\overline{K}^{eq}_{\{w\}\square\,\{x\}} \equiv VK^{eq}_{\{w\}\square\,\{x\}} = \underbrace{\underset{w}{\square}\, Z_w^{\square\square_w}}_{\substack{\text{reactant}\\\text{contributions}}} \cdot \underbrace{Z_X^{\square_X} \cdot \underset{MSH}{\square}\, Z_{MSH}^{\square_{MSH}} \cdot \underset{other}{\square}\, Z_{other}^{\square_{other}}}_{\substack{\text{product}\\\text{contributions}}}.$$

Operationally, of course, this picture is useless. Instead, we will deal with a more concrete picture, focusing chiefly on the pairwise hybridization of ssDNAs.

Following Cantor and Schimmel,[21] we note that often simple bimolecular reaction kinetics are adequate to describe a given duplex formation, with relaxation time given by $\square^{1} = 2k_{\square}\,[w] + k_{\square}$. Measured relaxation kinetic parameters for several oligonucleotides (and their complements) at 21–23 ºC are given in the table below.[22]

The negative forward activation energies imply a decrease of the forward reaction rate with temperature, provided that the kinetics are Arrhenius, that is, that $k_{\square} = \exp(-\square E_a)$. However, the second-order hybridization rate constant is relatively insensitive to temperature for oligos. More generally, temperature dependence arises implicitly

[21] Cantor, C. R., and Schimmel, P. R., *Biophysical Chemistry, Part III: The behavior of biological macromolecules,* WH Freeman, New York (1980).

[22] Adapted from Cantor and Schimmel; taken in turn from Riesner, D., and Römer, R., *Physico-Chemical Properties of Nucleic Acids,* Vol. 2, ed. J. Duchesne, Academic Press, London (1973).

The association rate constant for duplex formation is strongly dependent on ionic conditions and apparently is ~ 10^6–10^7 s^{-1} at 25 ºC in 0.25-1 M Na^+. This is still much less than the rate constant for a diffusion-controlled reaction (~ 10^8–10^9 s^{-1}). See Reynaldo, L. P., et al., "The Kinetics of Oligonucleotide Replacements," *J. Mol. Bio.* **297**, 511 (2000), or Patzel, V., and Sczakiel, G., "Length Dependence of RNA-RNA Annealing," *J. Mol. Bio.* **294**, 1127 (1999).

Table 2. Relaxation kinetic parameters for several oligonucleotides

Sequence	$k_{Æ}$ [M^{-1} s^{-1}]	E_a [kcal mol^{-1}]	Nucleus length
A_9	$5.3 \cdot 10^5$	–8	3
A_{10}	$6.2 \cdot 10^5$	–14	3
A_{11}	$5.0 \cdot 10^5$	–12	3
A_{14}	$7.2 \cdot 10^5$	–17.5	3
A_4U_4	$1.0 \cdot 10^6$	–6	2–3
A_5U_5	$1.8 \cdot 10^6$	–4	2–3
A_6U_6	$1.5 \cdot 10^6$	–3	2–3
A_7U_7	$8.0 \cdot 10^5$	5	2–3
A_2GCU_2	$1.6 \cdot 10^6$	3	1–2
A_3GCU_3	$7.5 \cdot 10^5$	7	1–2
A_4GCU_4	$1.3 \cdot 10^5$	8	1–2
A_5G_2	$4.4 \cdot 10^6$	7	1–2
A_4G_3	$4.2 \cdot 10^6$	9	1–2

through the allowed formation of secondary structure in polymeric ssDNAs.[23] This is one of the first hints that DNA hybridization is more complicated than a garden-variety bimolecular reaction. Of course, elementary kinetic steps must have nonnegative activation energies, and resolving this issue quickly leads to a kinetic scheme in which nucleation is followed by zippering.[24] In any event, this illustrates the link between activation energies and nucleation; in particular, the precise mechanism of nucleation will depend on the base sequence. We will neglect this dependence for the sake of tractability.

B. OVERVIEW

One of the drawbacks of the all-or-none model (which, broadly speaking, uses the bimolecular kinetic explanation above and neglects finer scales) is its experimentally unsupported prediction of a linear increase in forward rate with (hybridization) length for homologous oligomers. Another drawback (which our approach shares in principle, but not in practice, owing to the particular nature of our present context) is that one has for simple second-order kinetics the deceptively complex “C_0t relationship”[25]

$$[\mathrm{ss}]^{-1} \cdot ([\mathrm{ss}] + [\mathrm{ds}]) = 1 + k_{\square}\,\square t \cdot ([\mathrm{ss}] + [\mathrm{ds}]).$$

[23] Wetmur, “Physical Chemistry of Nucleic Acid Hybridization,” and Howorka, S., et al., “Kinetics of duplex formation for individual DNA strands within a single protein nanopore,” *PNAS* **98**, 12996 (2001).

[24] This kinetic scheme is well established. See, e.g., Ross, P. D., and Sturtevant, J. M., “The kinetics of double helix formation from polyriboadenylic acid and polyribouridylic acid,” *PNAS* **46**, 1363 (1960).

[25] A nice discussion of this is in Dieckmann, T., Lecture notes online at http://www-chem.ucdavis.edu/courses/W02/107B/ (2002).

Here □ is a factor reflecting the (Shannon) sequence entropy; it is less (respectively, greater) than unity for high (respectively, low) entropy. This connection between sequence entropies and renaturation rates has historically been exploited to determine genetic complexity in experiments. Now the "complexity" N is given in the present context by

$$C_0 = 2N \cdot ([\mathrm{ss}] + [\mathrm{ds}]),$$

(hence the "C_0t relationship") where C_0 denotes the total initial concentration of *nucleotides* and is consequently proportional to the length of the polynucleotides. More generally (i.e., when considering the renaturation of a ssDNA fragments obtained from shearing/fragmenting a single long duplex),

$$\square C_0 = 2N \cdot ([\mathrm{ss}] + [\mathrm{ds}]).$$

That is, when considering pairwise hybridization, we should set □ = 1 (hence our previous statement about not sharing the problem in practice); when considering batch hybridizations in full generality, we should be more careful.

k_{hyb} has the empirical form[26]

$$k_{hyb} = k_N \square (L_{hyb})^{1/2} N^{-1},$$

where k_N□is the nucleation rate constant (~ $3.5 \cdot 10^5$ $\mathrm{M}^{-1}\mathrm{s}^{-1}$ in typical hybridization buffers ~ 1 M NaCl), L_{hyb} is the hybridization length in nucleotides, and in the present context N equals L_{hyb},[27] giving $k_{hyb} = k_N \square (L_{hyb})^{-1/2}$. The square root dependence is generally presumed to be due to excluded volume effects, a view which we do not share (we will elaborate on this point later). Hybridization is effectively an all-or-none process. Mismatching of up to ~10□percent of the bases (see appendix) has little effect on hybridization rates (although it will surely affect denaturation rates).

In an attempt to make a comprehensive overview (and provide a sanity check) we mention work done on RNA–RNA annealing.[28] At high temperatures the (inverse) square root functional relationship between length and rate constant is preserved; secondary structure appears to cause a sharp decrease in the rate constant at lower (i.e., physio-

[26] Wetmur, "Physical Chemistry of Nucleic Acid Hybridization."

[27] Wetmur, J. G. (personal communication, 2002).

[28] Patzel, V., and Sczakiel, G., "Length Dependence of RNA-RNA Annealing," *J. Mol. Bio.* **294**, 1127 (1999).

logical) temperatures.[29] This relationship evidently holds for polynucleotides as short as 14 nucleotides, although deviations from it occur for very short RNAs. (Evidently, k_{hyb} has a minimum between 10–15 nucleotides) DNA–DNA annealing rates are similar to (although not less than) RNA–DNA or RNA–RNA annealing rates,[30] a fact presumably due to secondary structure effects. For long RNAs and an 800 nucleotide RNA, experiments at physiological ionic strength and temperature (as opposed to, say, PCR regimes) yielded a relationship of the form $k_{hyb} = e^{aL(\mathrm{long})+b}$.

For oligos a relationship of this form can be obtained through considerations of ionic interactions between RNAs.[31] For longer/polymeric RNAs, Patzel and Sczakiel considered the scaling of various non-ionic interactions by modeling RNAs as spheres with some fixed (possibly sequence dependent) fraction of their areas capable of participating in specific non-ionic interactions (e.g., hybridization nucleation). They remark that if the non-specific non-ionic interactions behave similarly (and there seems to be no reason to assume otherwise), then the specific and non-specific non-ionic interactions can be expected to cancel each other out as far as scaling is concerned. In any case, ionic interactions scale as r^{-1} (vs. r^{-6} for nonionic interactions), and so we ought to expect the ionic character to dominate scaling regardless, despite screening effects and the like.

The net effect seems to be that square-root dependence arises from what are generally regarded as excluded-volume effects at high temperatures[32]—*a view we do not agree with, at least in the regime of interest to us here*, whereas the exponential dependence is explained as arising from the dominant role of ionic interactions. One explanation offered for an observed rate limit for annealing is in line with Wetmur's contention that annealing is hydrodynamically controlled through viscosity (this is clearly

[29] For homopolymeric RNAs (which ought not to have secondary structure) the square-root dependence was evidently maintained. Cited in Patzel and Sczakiel as Lee, C. H., and Wetmur, J. G., "On the kinetics of helix formation between complementary ribohomopolymers and deoxyribohomopolymers," *Biopolymers* **11**, 549 (1972).

[30] Galau, G. A., et al., "Studies on nucleic acid reassociation kinetics: Rate of hybridization of excess RNA with DNA, compared to the rate of DNA renaturation," *PNAS* **73**, 1020 (1977), and Galau, G. A., et al., "Studies on nucleic acid reassociation kinetics: Retarded rate of hybridization of RNA with excess DNA," *PNAS* **74**, 2306 (1977).

[31] The so-called "kinetic salt effect" is supposedly manifested through the Brönstedt-Bjerrum relationship log k $z_a z_b I^{1/2}$, where z_a and z_b are the charges of reactands and I is the ionic strength. Since the charges are proportional to the lengths of the RNAs, the claimed exponential dependence on the length (or on the square root of the ionic strength) follows from this.

[32] The role ostensibly played by secondary structure formation in reducing rates seems to us to fall into the same class of phenomenological considerations.

true of the zipping mechanism; see below). That said, several authors also point out that the annealing reaction cannot be diffusion-controlled per se (i.e., the rate constants are not those of a diffusion-controlled reaction). These two lines of thinking are not inconsistent, however.

C. APPROACH

We will introduce a transition-state model in which batch ssDNAs with complementary ends will encounter each other through diffusion. Pairs of batch ssDNAs that are close enough to interact comprise "proximal complexes," which can subsequently form activated-hybrid complexes, and which in turn can form nucleated and finally hybridized complexes.

A naïve (and ultimately, we believe, misleading) cartoon of the kinetics is as follows:

$$w+w' \underset{k_{-d}}{\overset{k_d}{\rightleftharpoons}} \overbrace{w\cdot w'}^{\text{co-located ssDNAs}} \underset{k_{-nuc*}}{\overset{k_{nuc*}}{\rightleftharpoons}} \overbrace{(ww')_{nuc*}}^{\text{activated complex}} \underset{k_{nuc\to nuc*}}{\overset{(\beta h)^{-1}}{\rightleftharpoons}} \underbrace{\left[\overbrace{(ww')_{nuc}}^{\text{nucleated complex}} \underset{k_{nuc\leftarrow hyb}}{\overset{k_{nuc\to hyb}}{\rightleftharpoons}} \overbrace{(ww')_{hyb}}^{\text{hybridized complex}}\right]}_{\equiv \mathrm{R1}^{hyb-nuc}}$$

The component reaction R1$^{hyb\text{-}nuc}$ is essentially (though not exactly) a series of identical reactions (denoted R1) in which individual base pairs are formed. In the course of each such formation there is a transition state (the reaction going from one transition state to the next is denoted R1*):

$$\mathrm{R1}^{hyb-nuc} \equiv (ww')_{nuc}\cdots\underbrace{(n)^* \underset{\substack{(\beta h)^{-1}e^{-\beta\Delta G_{(n)\to(n)^*}}\\ =(\beta h)^{-1}e^{\beta\Delta G_{(n)^*\to(n)}}}}{\overset{(\beta h)^{-1}}{\rightleftharpoons}} (n) \underset{(\beta h)^{-1}}{\overset{(\beta h)^{-1}e^{-\beta\Delta G_{(n)\to(n+1)^*}}}{\rightleftharpoons}} (n+1)^*}_{\mathrm{R1}^*;\quad (n)\to(n+1)\equiv\ \mathrm{R1}}\cdots(ww')_{hyb}:$$

$$\Delta G_{(n)\to(n+1)^*}+\Delta G_{(n+1)^*\to(n+1)} \equiv \Delta G_{(n)\to(n+1)} \equiv \Delta G_{(n)^*\to(n+1)^*}.$$

$$\text{Stable duplexes} \Rightarrow \frac{[(n)]}{[(n)^*]} = e^{-\beta\Delta G_{(n)^*\to(n)}} > 1 > e^{-\beta\Delta G_{(n)\to(n+1)^*}} = \frac{[(n+1)^*]}{[(n)]} \Rightarrow \Delta G_{(n)^*\to(n)} < 0 < \Delta G_{(n)\to(n+1)^*}.$$

The natural reaction coordinate here is just the extent of zipping (it is a spatial coordinate) and the base pairing dipole interaction must have to some (possibly poor) approximation a quadratic minimum,[33] so we invoke the equipartition theorem to get an estimate of the energy barrier:

$$\left.\begin{array}{ll}\text{Equipartition} \ \square & \square G_{(n)\square\,(n+1)^{*}} \ \square\,\frac{1}{2\square} = 0.3\ \text{kcal/mol} \\ \text{NN DNA model} \ \square & \square G_{(n)^{*}\square\,(n+1)^{*}} \underset{\text{average}}{=} \square 1.4\ \text{kcal/mol}\end{array}\right\} \square\ \ \square G_{(n+1)^{*}\square\,(n+1)} \ \square\,\square 1.7\ \text{kcal/mol.}$$

A detailed semimicroscopic analysis, on the other hand, predicts a barrier of slightly over 2 kcal/mol at body temperature, with approximately half coming from each of hydrogen bonding and ss-ds rigidity difference contributions (the latter arise from base-base stacking interactions).[34] However, the level of detail in that analysis is considerably higher, and the use of one or the other value does not substantively affect any of our conclusions. In any case, if we could associate a degree of freedom to each of the hydrogen bonding and base stacking contributions, then the barrier would be roughly 2□kcal/mol. This picture appears to be broadly consistent with the kinetics of hybridization. It is also worth noting that at 55 ºC (the annealing temperature in PCR) we get an estimate of –1.3 kcal/mol for the energy barrier with one effective degree of freedom and –1.6 kcal/mol with two.

It is important to note that such phenomena as enzymatic activity affects the batch uniformly (at least before the emergence of heterogenous DNA structures); therefore (for the purposes of deriving Gibbsian weights), we can (presumably) safely neglect kinetic features which are common to the entire batch. For instance, the kinetics of ~4 base pairs nucleation of hybridization between ssDNAs ought not to depend (at least not strongly) on parts of the ssDNAs distant from the nucleation site. (This view informs our proposed mechanism for nucleation: see below.) In the uniformly random case (i.e., ours) we therefore neglect the aspects of the molecular biology and chemical kinetics which do not vary significantly between instantiations. Our kinetic cartoon is applied with these points in mind.

[33] Cocco, S., et al., “Force and kinetic barriers to initiation of DNA unzipping,” *Phys. Rev. E* **65**, 041907 (2002).

[34] Base pairing interactions are well modeled by a Morse potential $D[(e^{-a(r-R)}-1)^2-1]$ with $D = 5.84\square^{-1}$, a = 6.3 Å^{-1}, and R = 10 Å. The quadratic approximation is actually only accurate within about .1–.2 Å of the minimum (at R). Incorporating the energetic contributions of rigidity/base stacking does not affect this qualitatively, and the net result ought to be that the formation of an activated complex (i.e., achieving partial base stacking without hydrogen bonding) is more energetically unfavorable than equipartition would suggest, as appears to be the case. See note 33. Also, a 0.2 Kcal/mol increase in potential energy (arising because hydrogen bonds break on length scales of ~1 Å) comes into play. (Cocco, S., et al., “Slow nucleic acid unzipping kinetics from sequence-defined barriers,” cond-mat/0207609 (2002).

We are motivated by a parallel batch synthesis protocol—a protocol in which the entire batch of ssDNAs is co-introduced simultaneously. We assume that subprotocols such as ligation, polymerization, and thermal cycling are performed periodically, and we consider the resultant (quasi-) equilibrium resulting from such a process. We expect that the primary mechanism at play is diffusive hybridization.

VI. KINETIC ASPECTS
II—DIFFUSION AND CONFIGURATION KINETICS

We begin our discussion in earnest by noting that there is some apparent inconsistency in the literature about the applicability of models from polymer physics to short ssDNA dynamics. Ansari et al. offer a resolution indicating that short ssDNAs exhibit behavior consistent with models of flexible polymers that can exhibit transient mishybridized loop configurations. Goel et al.[35] point out that ssDNA has Kuhn length[36] ~1.4 nm , and hence persistence length ~.7 nm (whereas the interphosphate distance for ssDNA is typically ~ .75-1 nm). With these bits of information in hand, we can consequently model ssDNAs as purely floppy polymers rather than resorting to more complicated models such as the wormlike chain.[37] That said, at physiological conditions, local helical order should be present in ssDNA, at least in a statistical sense.[38] However, the lack of persistence in this local order allows us to use the floppy model in practice.

On the other hand, double-stranded DNA, is well-described by Hearst's "weakly bending rod" model with 3.4 Å rise/bp and 13 Å radius for the helix; its persistence length[39] is ~500 Å (hence its Kuhn length is ~100 nm) according to both the Hearst

[35] See, e.g., Goddard, N. L., et al., "Sequence dependent rigidity of single stranded DNA," *Phys. Rev. Lett.* **85**, 2400 (2000); Ansari, A., et al., "Misfolded Loops Decrease the Effective Rate of DNA Hairpin Formation," *Phys. Rev. Lett.* **88**, 069801-1 (2002); Goel, A., et al., "Unifying Themes in DNA Replication: Reconciling Single Molecule Kinetic Studies with Structural Data on DNA Polymerases," *J. Biomolecular Structure and Dynamics* **19**, 1 (2002).

[36] The Kuhn length of an N-mer is defined as $l_K \equiv R^2/(a[N\text{-}1])$, where R^2 is its mean square end-to-end distance and a is the bond or monomer length. It can be shown that $l_K = 2l_p$ (see next footnote).

[37] In the wormlike chain model (WLC) a polymer is modeled as a series of N segments of constant length a (hence arclength $L = Na$), each making a fixed angle θ with its predecessor but free to rotate about its predecessor's axis. (Saitô noted that the conformations of the WLC can be identified with diffusion paths on the unit sphere.) The (temperature dependent) *persistence length* $l_p = \beta A$ (A is the stiffness) is defined as the limiting average length of the projection of the end-to-end distance onto the axis of the first segment (i.e., $l_p = a(1\text{-}\cos\theta)$). It can further be shown that the mean square of the end-to-end distance satisfies $\langle R^2\rangle = 2l_p\{L - l_p[1\text{-}\exp(\text{-}L/l_p)]\} \approx 2l_pL$. (See Daune.)

[38] Cantor and Schimmel, *Biophysical Chemistry*.

[39] According to Song, L., and Schnurr, J. M., "Dynamic bending rigidity of DNA," *Biopolymers* **30**, 229 (1990), the persistence length of DNA is determined by $l_p^{-1} = P(sequence)^{-1} + P(slow)^{-1} + P(local)^{-1}$, where the RHS terms refer respectively to sequence-dependent deviations from straight B-DNA, millisecond-timescale structural variations, and quickly relaxing local elastic deformation contributions.

model[40] and the “wormlike chain” Zimm model[41] (a variation on the nonhydrodynamically interacting Rouse model[42]) which is preferable for longer dsDNAs. For dsDNAs with more than ~ 600 base pairs the entire spectrum of relaxation times enters into the intramolecular dynamics, and modeling consequently becomes more difficult.[43]

A. TRANSLATIONAL DIFFUSION

Using the Stokes-Einstein-Smoluchowski theory we can obtain a diffusive rate constant. For example, a floppy polymer should have diffusion and diffusive rate constants governed by its effective radius, and this is indeed the case. For two floppy polymers of lengths L_1 and L_2, it is reasonable to assume that their effective radii scale as the square root of the length,[44] and so we obtain (up to a constant steric factor)

$$k_d = \frac{2}{3□□} \frac{\left(L_1^{1/2} + L_2^{1/2}\right)^2}{L_1^{1/2} L_2^{1/2}}.$$

Here □ is the solvent viscosity. Note that more generally, we have a formula like

$$k_d = 4□(D_1 + D_2)d_{12}$$

where D_1, D_2 denote the two diffusion coefficients, and d_{12} denotes a critical reaction distance. The previous formula is based on this one with (statistically) spherical (floppy) polymers, which react when the spheres defining their statistical extent come in physical contact. Taking L_1 and L_2 to be equal, we get a rate constant of $8(3□□)^{-1} \sim 10^8$–10^9 s^{-1} in

[40] Goel, “Unifying Themes in DNA Replication.”

[41] Zimm, B. H. “Dynamics of Polymer Molecules in Dilute Solution: Viscoelasticity, Flow Birefringence and Dielectric Loss,” *J. Chem. Phys.* **24**, 269 (1956).

[42] Rouse, P. E., Jr., “A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers,” *J. Chem. Phys.* **21**, 1272 (1953).

[43] Diekmann, S., et al., “Orientation Relaxation of DNA Restriction Fragments and the Internal Mobility of the Double Helix,” *Biophys. Chem.* **15**, 263 (1982).

[44] In reality, as Kuhn predicted in 1934, floppy polymers do *not* assume spherical shapes. An experiment was performed (and reported in Haber, C., et al., “Shape anisotropy of a single random-walk polymer,” *PNAS* **97**, 10792 [2000]) in which conformations of fluorescently labeled DNA molecules (T2-DNA, with arclength ~ 56 □m and persistence length ~ 52 nm [and hence ~ 1,075 orientationally uncorrelated statistical segments]) were optically monitored. In the ensemble the average aspect ratio of an ellipsoid approximating the T2–DNA conformation was ~ 4.1:2.3:1. We obtained fits to their data suggesting that the rate (in s^{-1}) of (major/minor) aspect ratio change satisfies an exponential distribution with decay constant ~ 1.9, and similarly that the rate (in s^{-1}) of angular change (more specifically, the “rate at which a polymer rotated by more than 90°”) satisfies an exponential distribution with decay constant ~.8. On the other hand, appealing to ergodicity to go between time and ensemble averages indicates that the spherical assumption is not intrinsically invalid for our purposes. In any case, the relative roles of sphericity and anisotropy raise a subtle point that we proceed to sweep under the rug with a smile.

water at normal temperatures. In general, we expect this to be a tight upper bound on the rate, said to hold in *diffusion-controlled reactions*.[45]

Tirado and García de la Torre have laid most of the theoretical groundwork for diffusion of dsDNA: duplex oligonucleotides can be modeled by rigid rods with translational diffusion coefficient

$$D = (\log p + \square_{trans})/(3\square\square\square L),$$

where p is the axial ratio (length over diameter) and $\square_{trans}$ is a correction factor (which depends on p).[46]

Table 3 gives a "dirty-hands" sense of the translational diffusion characteristics of oligonucleotides:

Table 3. Predicted and measured translational diffusion coefficients for oligonucleotides[47]

Translational diffusion coefficient (10^{-10} m^2/s)	8-mer	12-mer	20-mer
Experimental data	1.53 ± 0.05	1.34 ± 0.03	1.07 ± 0.02
Double bead model	1.614	1.362	1.066

B. ROTATIONAL DIFFUSION

The time-scale for rotational relaxation of a dsDNA is $\square_{long\ axis} = 1/(6\square)$, where $\square$ is the long-axis rotational diffusion coefficient.[48] Tirado and García de la Torre[49] obtain a rotational (tumbling) diffusion coefficient for duplex oligonucleotides, and one also has a spinning diffusion coefficient:[50]

$$\square_{tumble} = 3(\log p + \square_r)/(\square\square\square L^3);$$

$$\square_{spin} = (3.841\square\square\square LR^2(1 + \square_s))^{-1},$$

[45] Exceptions are in promoted reactions (see below with respect to accelerating hybridization kinetics).

[46] Tirado, M. M., and García de la Torre, J., "Rotational dynamics of rigid, symmetric top molecules. Application to circular cylinders," *J. Chem. Phys.* **73**, 1986 (1980).

[47] Reproduced from Banachowitz, E., et al., "Solution Structure of Biopolymers: A New Method of Constructing a Bead Model," *Biophys. J.* **78**, 70 (2000), in turn citing Eimer, W., and Pecora, R., "Rotational and translational diffusion of short rodlike molecules in solution: Oligonucleotides," *J.□Chem. Phys.* **94**, 2324 (1991) and Liu, H., et al., "Effect of electrostatic interactions on the structure and dynamics of a model polyelectrolyte. I. Diffusion," *J. Chem Phys.* **109**, 7556 (1998).

[48] Diekmann, S., et al., "Orientation Relaxation of DNA."

[49] Tirado and de la Torre, "Rotational dynamics."

[50] Chrico, G., et al., "Rotational dynamics of curved DNA fragments studied by fluorescence polarization anisotropy," *Euro. Biophys. J.* (online 2000).

where again p is the axial ratio (length over diameter) and the deltas are correction factors that depend on p. Curvature in dsDNAs (such as is commonplace in, e.g., poly(A)) can be expected to manifest itself in the rotational dynamics as a decrease in the spin diffusion coefficient. Roughly equivalently, spin coefficient variations can be accounted for by larger hydrodynamic radii (which is ~ 9.5 Å for the DNA B-helix).

Eimer and Pecora showed[51] that the Tirado-García de la Torre model is valid for short (8, 12, and 20 base pair) oligonucleotides. Diffusion coefficients at 20 ºC are given in the table below:

Table 4. Data from Eimer and Pecora. Here $D = D_0(1+k[\text{oligo}])$ is an empirical relationship yielding the second virial coefficient of diffusion.

	D_0 (10^{-10} m^2 s^{-1})	k (l kg^{-1})	□ (10^7 s^{-1})	p	L (Å)
8-mer	1.53	8.5	5.18	1.43	28.6
12-mer	1.34	8.0	2.61	2.10	42.1
20-mer	1.09	12.9	1.03	3.59	68.8

Data from a separate study[52] yielded rotational relaxation times:

Table 5. From Banachowitz et al. The authors considered bead models for B–DNA hydrodynamics in which identical overlapping beads of radius 5.0 Å were used for both base and phosphodiesters. Short DNA fragments are slower in reality than the bead model with parameters for longer fragment (the authors attribute this to end effects).

Rotational relaxation time (ns)	8-mer	12-mer	20-mer
Experimental data	3.22 ± 0.16	6.39 ± 0.32	16.2 ± 0.8
Double bead model	2.84	5.58	15.2

C. CONFIGURATIONAL KINETICS

Goel et al. point out that as a rule for applying transition state theory, polymer relaxation times should be much less than the time to cross energy barriers (estimated to be around 10–100 ms). They also remark that for a dsDNA complex "the relaxation times…can increase appreciably with its contour length and fraction of dsDNA composition. Therefore transition state theory must be applied cautiously to describe reactions."[53] Indeed, a 50-percent dsDNA complex would have comparable relaxation and reaction times, signaling that transition state theory may begin to fail in this regime.[54]

[51] Eimer, W., and Pecora, R., "Rotational and translational diffusion of short rodlike molecules in solution: Oligonucleotides," *J. Chem. Phys.* **94**, 2324 (1991).

[52] Reproduced from Banachowitz et al.

[53] Goel, "Unifying Themes in DNA Replication."

[54] Goel, A., et al., "Tuning DNA 'strings,' Modulating the rate of DNA replication with mechanical tension," *PNAS* **98**, 8485 (2001).

This suggests that our point of view for examining batch DNA synthesis may be more difficult to use for long times and more generally hints at a potential obstacle for attempts at understanding the self-assembly of DNA structures.

VII. KINETIC ASPECTS
III—NUCLEATION AND ZIPPING

A. THE GEOMETRY OF NUCLEATION

The formation of an activated complex between two ssDNAs requires something like nucleation of hybridization. In fact, it happens that nucleation is the rate-limiting step in the kinetics of DNA hybridization.[55] However, there are some indications that if hybridization is modeled as a single second-order reaction the forward rate is length-independent (indeed, effectively diffusion-controlled) for very short oligonucleotides[56] (hence the reverse/dissociation/off rate ought to vary exponentially with length)[57] and is roughly given by

$$k_{forward} = A_{forward}\, e^{-\square E(forward)} \sim 6 \cdot 10^5 \text{ M}^{-1}\text{s}^{-1}$$

where $A_{forward} = 5 \cdot 10^8$ M^{-1}s^{-1} and $E_{forward} = 4$ kcal mol^{-1} is the forward activation energy.[58]

In any event nucleation is the crucial event in the hybridization process. With this in mind, we identify a symbol for a batch ssDNA with an arclength parametrization (in units of monomer length, so that a batch ssDNA has a length given by the number of its constituent bases and so that the tangent vectors have unit norm) beginning at the 5□end of an idealized curve representing the ssDNA. Using s and s□to denote a positions along a pair of ssDNAs that together comprise a potential nucleation site, a plausible geometrical constraint for nucleation can be written as

$$\left\langle \dot{w}(s), \square\dot{w}\square(s\square) \right\rangle \geq 1 \square \square_{hyb}\,.$$

[55] Wetmur, "Physical Chemistry of Nucleic Acid Hybridization."

[56] Quartin, R. S., and Wetmur, J. G., "Effect of ionic strength on the hybridization of oligodeoxynucleotides with reduced charge due to methylphosphate linkages to unmodified oligodeoxynucleotides containing the complementary sequence," *Biochem.* **28**, 1040 (1989).

[57] Cocco, S., et al., "Force and kinetic barriers to unzipping of the DNA double helix," *PNAS* **98**, 8608 (2001).

[58] This appears to be at odds with Wetmur's empirical formula $k_{hyb} = k_N\square(L_{hyb})^{1/2}N^{-1}$. In this regime, the "complexity" N and length can be identified, and $k_N\square \sim 3.5 \cdot 10^5$ M^{-1} s^{-1} gives too low a value. This sort of conflicting information about the hybridization kinetics of oligonucleotides is commonplace.

It is not hard to see that

$$\Pr\left(\left\langle \dot{w}(\cdot), \square\dot{w}\square(\cdot)\right\rangle \geq 1 \square \square_{hyb}\right) = \frac{1}{2}\square_{hyb}\,.$$

The values of the critical hybridization distance and critical hybridization orientation parameter $\square_{hyb}$ are governed by electrochemical factors.

There is a –6 cal mol^{-1} K^{-1} empirical initiation entropy which presumably comes about as the thermodynamic cost of strand alignment.[59] We can use this to estimate the probability of strand alignment:

$$A_{=} \equiv \left\{\left\langle \dot{w}(\cdot), \square\dot{w}\square(\cdot)\right\rangle \geq 1 \square \square_{hyb}\right\};\quad \Pr(A_{=}) = \frac{1}{2}\square_{hyb}$$

$$A_{\square} \equiv \left\{\left\langle \dot{w}(\cdot), \square\dot{w}\square(\cdot)\right\rangle < 1 \square \square_{hyb}\right\};\quad \Pr(A_{\square}) \equiv 1 \square \Pr(A_{=}).$$

$$\Pr(A_{=}) + \Pr(A_{\square}) = 1 \square \;\; \Pr(A_{=}) = 1 \square \frac{1}{1 + \dfrac{\Pr(A_{=})}{\Pr(A_{\square})}}\,.$$

$$\frac{\Pr(A_{=})}{\Pr(A_{\square})} = e^{\square\square\square G_{\square\square =}} = e^{\square S_{\square\square =}/R} \square e^{\square 3} \square \;\; \Pr(A_{=}) \square 0.047, \quad \square_{hyb} \square 0.095.$$

(This means “alignment” translates to ~ 18° tolerance in the relative orientation of the strands. Of course, this is meaningless except in a statistical sense.) As we shall see below, the dynamical time-scale for nucleation (the orientational relaxation time of ~3–4□bases) in the regime we are concerned with should be roughly .8–.9 ns. Now it can be shown (see Appendix D) that the so-called average fractional heat (or probability) content of the non-alignment parameter space goes (to first order) as 0.610 + 0.390e$^{-2\square t}$. That is (if we do not initially have alignment already),

$$\Pr(\min_t \square_t \leq \square_{hyb}) \sim 0.390 \cdot (1 - e^{-2\square t}).$$

Using this, we arrive at the total probability of alignment as a function of time as

$$0.419 \cdot (1 - 0.889e^{-.4t})$$

with t the time in nanoseconds. Can things really be so tractable as this? No—but the silver lining is that we can use this to gauge what is really going on in the hybridization process, and we continue with this approach in mind.

[59] Winfree, E., *Algorithmic Self-Assembly of DNA,* Ph.D. thesis, California Institute of Technology (1998). Since this entropy is independent of length, it must be due to either rotational degrees of freedom or translational ones that do not depend on the length (e.g., strand-strand distances). The latter case seems unlikely since it cannot be isolated from diffusion proper.

The first thing to note is that the rate constant in the exponential here is $4 \cdot 10^8$ s^{-1}, which immediately rules out alignment as a rate-limiting step (non-rate-limiting processes such as base-pair addition have constants of 10^6–10^7 s^{-1} [see the discussion below], and any strand-strand interactions [which along with hydrodynamic effects are presumably accounted for by the initiation entropy anyway] could only be expected to speed this alignment process up). Consequently, we can safely assume that translational constraints (but of a type qualitatively different than those associated with the free diffusion of noninteracting ssDNAs) are the rate-limiting step in the overall rate-limiting step of nucleation. Unlike alignment constraints, we can expect these translational constraints to depend explicitly on the lengths of the ssDNAs involved. To recap, the overall lesson to draw here is that the kinetics of the nucleation process will probably exhibit a dependence on DNA length because of translational constraints arising from a qualitatively different process than free diffusion in solution. This conclusion is consistent with the commonly held view cited above and derived from Wetmur that excluded-volume effects are the overall rate-limiting factors in DNA hybridization—but there is another, potentially better explanation.

B. SEARCH PHASE

We suspect that the excluded-volume view (certainly in spirit even if not quite in fact) is inappropriate—at least in the regimes that interest us. (For instance, it is unlikely that excluded-volume effects govern the hybridization kinetics of short ssDNAs.) Instead, we propose a conceptually simpler picture, in which, much like a polymerase,[60] two ssDNAs shift bases relative to each other, evoking a one-dimensional diffusive process that continues until nucleation is initiated or the ssDNAs decouple.[61]

[60] Guthold, M., et al., "Direct Observation of One-Dimensional Diffusion and Transcription by *Escherichia coli* RNA Polymerase," *Biophysical J.* **77**, 2284 (1999).

[61] A similar picture holds for DNA transcription factors: "In their nonspecific binding mode, TFs are still strongly associated with the DNA but are able to diffuse (i.e., slide) randomly along the genome." Gerland, V., et al., "Physical constraints and functional characteristics of transcription factor-DNA interaction," *PNAS* **99**, 12015 (2002). Recent work on RAD52 promotion of DNA annealing also suggests the possibility of such a mechanism: "Once strand annealing is initiated, the two [RAD52-ssDNA] complexes could roll around each other, driven by the energetically favorable formation of the DNA duplex until the DNA becomes base-paired…Specificity is achieved therefore by the favorable energy change that would acompany only the annealing of complementary sequences." Singleton, M.R., et al., "Structure of the single-strand annealing domain of human RAD52 protein," *PNAS* **99**, 13492 (2002).

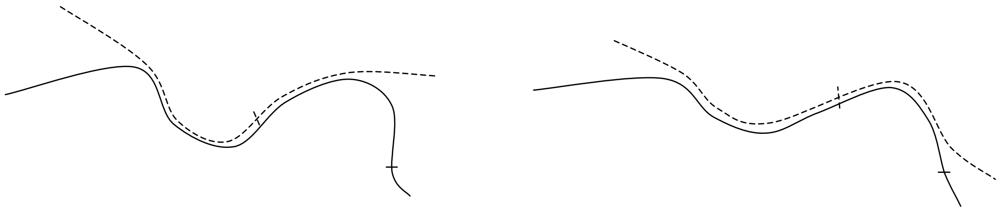


Figure 4. Proposed search phase. (L) Portions of two ssDNAs with a single complementary region indicated by hatch marks form a transient/metastable spatially coherent structure evocative of dsDNA ("2ssDNA"). The energy barrier to mutual one-dimensional diffusion along their respective phosphodiester backbones can easily be overcome by thermal noise. (R) The same two ssDNAs after a brief time period. If the hatch marks line up then a series of Watson-Crick pairs can be formed (this is just the nucleation step) and hybridization can occur.

Suppose that l bases are (correctly or incorrectly) paired between two ssDNAs. As a (very) simple model, suppose that a correct base pairing contributes an energy of ~ 1 unit, and that an incorrect base pairing contributes an energy of << 1 unit. Then (if spatial and sequence characteristics are taken uniformly at random) a typical 2ssDNA configuration donates an energy of ~ $l/4$ + (*correction*) whereas a properly paired configuration gives an energy of ~ l. As such there will tend to be a single deep energy well for a proper configuration and a low-energy barrier to one-dimensional relative translations of the ssDNAs.[62]

The immediate problem with this picture is that it appears to preclude zipping. In fact this need not (and should not) be the case. The fraying of ends and overall shifting of bases noticed in simulations of denaturation dynamics by Drukker at al. lend qualitative support to an explanation of why this picture (despite its novelty and unorthodoxy) can explain the kinetics of DNA hybridization.[63] More generally it is reasonable to assume

[62] Others have had these and complementary thoughts:

> Amazingly, being random is sometimes advantageous for reconstruction of cleaved strands, due to the high degree of specification: take a long strand S and a strand S_0' which is twin to a subsegment S_0 of S. In the random case, the binding energy between S_0' and S_0 is roughly twice as great as the energies of other possible bindings of S_0' to S, since the number of matches of letters for random pairs of sequences is roughly equal to the number of mismatches. But for pure breed strands the binding energy is constant for full hybridization and, in general, proportional to the length of the hybridization overlap. The same consideration applies to the hybridization of two pieces of a cleaved strand on available ligation sites. Since the energy enters the canonical sum under the exponent of the Gibbs-Boltzmann factor, it can beat entropy and, depending on parameters (as temperature, energy, concentration, etc.), the population may evolve towards strands with repeated random motives. One wonders if the tandem repetitions in genomes can be explained by mechanisms of this nature. [Carbone, A., and Gromov, M., "Mathematical slices of molecular biology," Preprint IHES/M/01/03 (2001).]

[63] Drukker, K., et al., "Model simulations of DNA denaturation dynamics," *J. Chem. Phys.* **114**, 579 (2001).

that transient 2ssDNA configurations are qualitatively similar to partially denatured dsDNA configurations—there ought to be "denatured" bubbles, as in the Poland-Scheraga model.[64] This picture allows us to retain the zipper model, albeit in a modified form: it suggests that the rate-limiting factor in zipping could be the breaking of sterically unfavorable transient base pairs due to tension in the shorter strand component of the induced bubble.

Quantitatively, it would be nice to show that this scheme can reproduce observed kinetics. This is not hard. In fact it is too easy. We list a few sketches, any of which might turn out to have a kernel of truth:

- By equipartition, the average kinetic energy of the center of mass of an oligo is constant, and so $\langle E \rangle := mv^2/2$ defines a mean (positive) velocity which is consequently proportional to $m^{-1/2}$—and hence also to $L^{-1/2}$.[65] One expects the net instantaneous velocity in diffusion to behave similarly. If hydrodynamic effects turn out to be negliglible, relative thermally driven diffusion of the strand components of a 2ssDNA complex could be the driving mechanism. This would be tantamount to treating some point on one of the strand components as a random walker on a line.
- For long strands, the amount of "denatured" 2ssDNA ought to be roughly proportional to the number of bases, and the combination of entropic forces and the lack of a significant energy barrier between random 2ssDNA configurations (with the same amount of paired bases) would cause the bubbles to tend to assume random coil configurations. These coils would not be hydrodynamically screened and would therefore (since the diffusion coefficient for a random coil is proportional to inverse square root of its length) bring about observed kinetic forms. Moreover the effective lack of an energy barrier between 2ssDNA configurations would imply that the hydrodynamic diffusion (confined to one dimension) is the driving mechanism.
- In the most general setting, we might have some form of dynamics corresponding to anomalous diffusion,[66] in which we have a relationship of

[64] Poland, D., and Scheraga, H.A., *J. Chem. Phys.* **45**, 1456 (1966).

[65] Frank, M. P., personal communication (2002). This argument does not rely on dimensionality (other than trivially in changing the constant of proportionality in accord with equipartition).

[66] An interesting example of a process giving rise to anomalous diffusion is reptation. The reptation theory (which is central to the theory of gel electrophoresis—see, e.g., Viovy, J., "Electrophoresis of DNA and other polyelectrolytes: Physical mechanisms," *Rev. Mod. Phys.* **72**, 813, 2000) was developed by de Gennes to explain the dynamics of polymer gels by considering the motion of individual polymers in the presence of geometrical constraints imposed by entanglements with other such polymers. In this setting, a given polymer snakes through a reptation tube that represents these constraints. Two important relaxation times play roles: the longest Rouse (in this context,

> the form $\square x^2 \square \mu\ t^{\square}$. (If $\square < 1$ we call the behavior subdiffusive [characterized by long waiting times], whereas if $\square > 1$ we call it superdiffusive [characterized by long steps].) The first passage time (which would govern the initiation of nucleation) for anomalous diffusion on an interval of length L is proportional to $L^{2/\square}$.[67] It could be the case that different regimes of anomalous diffusion come into play in separate stages and combine to give the observed scaling. The excluded-volume interpretation (in which "the longer a DNA strand, the more difficult it is for a second strand to interpenetrate and find complementary sites") may actually imply anomalous dynamics.[68]

Although we find the first of these explanations by far the most appealing, the overall moral here is, as every kineticist knows, that we have to actually determine what the underlying processes and influences *are*—at present we simply do not have a sufficient understanding, either experimental or theoretical, of the processes involved in DNA hybridization. Nothing in the literature clarifies the nature of the rate-limiting translational component of the (itself rate-limiting) nucleation process, and so all we can do is guess and offer the suggestion that simulations and experiments using (e.g.) DNAs with cyclic code or de Bruijn sequences (see Appendix B) and various hybridization detection techniques or some other real-time hybridization detection mechanisms may offer a way to conclusively determine what really goes on.[69]

hydrodynamic interactions are typically assumed to be screened, and so the polymers obey Rouse dynamics) relaxation time $\square_R$ (proportional to the length squared; polymeric viscosity comes about as a result of dynamics on timescales like $\square_R$), and the de Gennes reptation time $\square_{dG}$ (proportional to the length cubed [Peters, F., "Polymers in flow: modeling and simulation," Ph.D. thesis, Delft (2000)]. In fact, we have a relationship of the form $\square_{dG} = 3L\square_R/a$, where a is the characteristic monomer length. Broadly speaking, we can interpret $\square_R$ as an intra-polymer relaxation time, and here $\square_{dG}$ as a polymer-polymer relaxation time.). More concretely, reptation theory predicts—and experiments confirm—that a reptating monomer ought to have a mean square displacement that goes as $t^{1/4}$ for intermediate timescales and as $t^{1/2}$ at other times (versus a displacement that goes as t for diffusion proper). Diffusion along the tube goes as $\square x^2 \square \mu\ t/L_{tube}$, which is consistent with a search phase: we might expect two ssDNAs to diffuse in an effective reptation tube determined by their common backbones. That is, insofar as we might expect them to reptate along each other until nucleation is initiated. Indeed the difference between simple one-dimensional diffusion and "reptation" appears to be largely one of interpretation in this particular context. See de Gennes, P. G., "Reptation of a Polymer Chain in the Presence of Fixed Obstacles," *J. Chem. Phys.* **55**, 572 (1971), Ebert, U., et al., "Short Time Behavior in de Gennes' Reptation Model," *Phys. Rev. Lett.* **78**, 1592 (1997), and Smith, D. E., et al., "Self-Diffusion of an Entangled DNA Molecule by Reptation," *Phys. Rev. Lett.* **75**, 4146 (1995).

[67] Gitterman, M., "Mean first passage time for anomalous diffusion," *Phys. Rev. E,* **62**, 6065 (2000).

[68] Chuang, J., et al., "Anomalous dynamics of translocation," *Phys. Rev. E,* **65**, 011802 (2001).

[69] Tyagi, S., and Kramer, F. R., "Molecular beacons: probes that fluoresce upon hybridization," *Nature Biotechnology* 14, 303 (1996); and McKendry, R., et al., "Multiple Label-Free Detection and Quantitative DNA-Binding Assays on a Nanomechanical Cantilever Array," *PNAS* **99**, 9783 (2002).

The idea of a search phase is not trivially reconcilable with second-order kinetics: if the two ssDNAs are chemically bound, the reaction is not second-order, whereas if they are unbound, it is difficult to see how a searching mechanism can come into play.[70] That said, transient base shifting in a 2ssDNA complex could happen through a number of ways. Ionic contributions of the phosphates tend to destabilize dsDNA, but entropic effects arising from counterion condensation (*not* generic entropic effects) tend to stabilize it.[71] These factors compete without a definite outcome and the net result for non-complementary regions of ssDNAs can be to keep them co-located in a 2ssDNA complex for a time-scale that would, for energy budgeting reasons, probably have to depend on the hybridization length.[72]

There are other, subtler issues that our proposed search phase raises. For one thing, we ought to expect a prefactor arising from the possible disengagement of complementary ssDNAs to manifest itself in the hybridization rate. For another, the nucleation rate is evidently unaffected by circular permutation of linear ssDNAs but is decreased roughly threefold if one of the ssDNAs is circular.[73] We attempt to give a sketch here of why this is not inconsistent with a search phase. The strands of a 2ssDNA complex might be hydrodynamically screened relative to each other, or they might not be. But the circular DNA cannot be hydrodynamically screened from the linear one it tries to hybridize to, and the moment of inertia can be expected to play a nontrivial role (especially when compared to the linear-linear case).

For instance (although such an argument is not really appropriate other than as a cartoon), the moment of inertia of a thin wheel[74] with mass $M = L$ and radius $r = L/(2\pi)$ is $L^3/(4\pi^2)$. If we assume that the linearized angular velocity of the ring equals the linear

[70] Wetmur, J. G., personal communication (2002).

[71] Gelbart, W. M., et al., "DNA-Inspired Electrostatics," *Physics Today* **53** (9), 38 (2000).

[72] Alternatively, local regions of partial bonding and base stacking could maintain a 2ssDNA complex as the unbound regions reptated. Although we can expect to see partial bonding and base stacking, we think it unlikely that such phenomena can maintain 2ssDNA complexes. We consider it much more likely that electrostatic effects would mediate the putative metastability of (the also basically putative) 2ssDNA complexes.

[73] Kinberg-Calhoun, J., and Wetmur, J. G., "Circular, but not circularly permuted, deoxyribonucleic acid reacts slower than linear deoxyribonucleic acid with complementary linear deoxyribonucleic acid," *Biochemistry* **20**, 2645 (1981).

[74] Even relaxed (let alone supercoiled) circular DNA will not act like this. Here we might hand-wave and say that a careful application of the ergodic hypothesis could conceivably justify this, although the truth or even relevance of this claim is far from clear. Our aim throughout this portion of the discussion is chiefly to illustrate a potentially feasible alternative to the excluded-volume hypothesis for the rate-limiting step of DNA hybridization.

velocity of a rod (or the curvilinear velocity of a suitably well-behaved polymeric object) with the same length (and hence mass), then the ratio of the rotational kinetic energy of the ring to the translational kinetic energy of the rod turns out to be approximately 4. (On the other hand, if the two kinetic energies were equal, then the linearized angular velocity would be $1/\sqrt{2}$ times the rod's velocity.) Rescaling one of the time-scales so that the kinetic energies equal one another (i.e., appealing to equipartition) implies a time dilation factor of about 2; throwing in a factor of $\sqrt{2}$ to compensate for the velocity gives a net factor of roughly 3. Of course, this does not actually make any sense. That said, a more careful non-hand-waving exercise of this sort might be able to explain the decrease in rate. In any case these arguments alone seem sufficient to justify reasonable doubt of the excluded volume view.

C. INTERNAL SSDNA RELAXATION DYNAMICS

Goel et al.[75] point out that the orientational relaxation time (which is effectively given as the shortest relaxation time in the Zimm model) for a single segment of ssDNA is ~ .7 ns. For even moderately long ssDNAs we expect to be in the "non-free-draining" regime, in which the relaxation times are of the form

$$\tau_k \propto \left(\frac{N}{k}\right)^{3/2} \left(1 - \frac{1}{2\sqrt{k}}\right)^{-1}.$$

The nucleation sites will be ~ 3–4 bases regardless of the lengths of DNA involved (which should lead to a zipping rate that is essentially independent of the lengths)[76] and so we expect the nucleation step proper to have a time constant (governed by if not exactly equal to)

$$\tau_{nuc} \approx \tau_{short}\left(1 + \frac{2.5}{N - 2.5}\right)^{3/2}.$$

For N ~ 20–40 this implies a time-scale of ~ .8–.9 ns. Since it seems plausible that the reverse-nucleation process should have a shorter time constant, we might simply put it at ~ .7 ns. In any event the small variations in these time-scales indicate that we can safely dispense with any presumed need for a more detailed model of hybridization (especially since the zipping stage generally occurs over much longer [microsecond] time-scales due to hydrodynamic effects [see below]).

[75] Goel, et al., "Tuning DNA 'strings.'"

[76] Cocco, et al., "Force and kinetic barriers to unzipping of the DNA double helix."

D. SINGLE-MOLECULE DNA HYBRIDIZATION DATA

The only single-molecule study of hybridization kinetics that we are aware of indicates that two different event types (i.e., two clearly distinct binding timescales) took place.[77] these events were only able to be characterized by two qualitatively different association/dissociation reactions.[78] The longer-lifetime events dominated the equilibrium constant, and these alone are reflected in the adapted tables (reproduced from the authors at 20 ºC with 2 M KCl and 12 mM $MgCl_2$ and using association-rate constants of $10^7\ M^{-1}\ s^{-1}$):

Nanopore data:

Untethered sequence	k_{on} [M^{-1} s^{-1}]	K_d [M]	$\Delta G°$ [kcal mol^{-1}]
5′-GGTGAATG-3′	$1.3 \cdot 10^7$	$9.2 \cdot 10^{-8}$	–9.2
5′-TACGTGGA-3′	$2.2 \cdot 10^7$	$1.5 \cdot 10^{-7}$	–8.9
5′-GGTGAAT-3′	$1.1 \cdot 10^7$	$1.5 \cdot 10^{-6}$	–7.7

Bulk (melting profile) values (k_{on} values assumed by authors):

Untethered sequence	k_{on} [M^{-1} s^{-1}]	K_d [M]	$\Delta G°$ [kcal mol^{-1}]
5′-GGTGAATG-3′	10^7	$3.6 \cdot 10^{-8}$	–9.8
5′-TACGTGGA-3′	10^7	$1.7 \cdot 10^{-7}$	–8.9
5′-GGTGAAT-3′	10^7	$8.3 \cdot 10^{-7}$	–8.0

The hybridization rate constant was found to vary only weakly with temperature. Data obtained from nanopore recordings can evidently serve as a useful approximator for kinetic/thermodynamic behavior of DNA in solution. The short-lifetime association events had rate constants of roughly 10^6 M^{-1} s^{-1} (an order of magnitude less than that of the long-lived events). These events may or may not correspond to nucleation followed by dissociation; we think it most likely that they are unsuccessful search phases, but we cannot support this. That said, it may be possible to use the data on the short-lived events to reach some quantitative conclusions about the microkinetics of nucleation. This idea is probably worth examining in some detail.

[77] Howorka et al., "Kinetics of duplex formation." It is demonstrated in this paper that data obtained from nanopore recordings can evidently serve as a useful approximator for kinetic/thermodynamic behavior of DNA in solution, despite the possible influences of steric constraints, applied electric potential, etc.

[78] Although Howorka et al. do not speculate on the explicit nature of these events, it seems clear that the short and long events could be characterized by the kinetic avenues taken in lieu of and directly after nucleated hybridization of a few (~3–4) base pairs: to wit, that the short events might correspond to a search phase which failed to initiate nucleation and hence also hybridization (in view of the oligonucleotides and experimental protocols they used, we think it conceivable but unlikely that the short events are bona fide mishybridizations or even transient nucleations), whereas the long events might correspond to fully zipped duplexes (zipping occurs on a much faster time-scale than nucleation and indeed could not be resolved by the experiment). This explanation (which is broadly consistent with the "all-or-none" model) seems especially suitable in the context of transition-state theory.

E. KINETICS OF ZIPPING

Recall the series of identical reactions R1 in which individual base pairs are formed:

$$\text{R1}:\quad (n) \xrightleftharpoons[(\beta h)^{-1}]{(\beta h)^{-1} e^{-\beta\Delta G_{(n)\to(n+1)^*}}} (n+1)^* \xrightleftharpoons[(\beta h)^{-1} e^{\beta\Delta G_{(n)^*\to(n)}}]{(\beta h)^{-1}} (n+1)$$

At steady - state :

$$k_1(\text{R1}) \equiv k_1 = \frac{k_{(n)\to(n+1)^*}k_{(n+1)^*\to(n+1)}}{k_{(n)\to(n+1)^*}+k_{(n+1)^*\to(n+1)}} = \frac{(\beta h)^{-1} e^{-\beta\Delta G_{(n)\to(n+1)^*}}(\beta h)^{-1}}{(\beta h)^{-1}+(\beta h)^{-1}} = \frac{1}{2\beta h}e^{-\beta\Delta G_{(n)\to(n+1)^*}};$$

$$k_{-1} = \frac{k_{(n+1)^*\leftarrow(n+1)}k_{(n)\leftarrow(n+1)^*}}{k_{(n+1)^*\leftarrow(n+1)}+k_{(n)\leftarrow(n+1)^*}} = \frac{(\beta h)^{-1} e^{\beta\Delta G_{(n)^*\to(n)}}(\beta h)^{-1}}{(\beta h)^{-1}+(\beta h)^{-1}} = \frac{1}{2\beta h}e^{\beta\Delta G_{(n)^*\to(n)}} \equiv \frac{1}{2\beta h}e^{-\beta\Delta G_{(n)^*\leftarrow(n)}}.$$

$$\text{R1} \Rightarrow \quad (n)\xrightleftharpoons[k_{-1}]{k_1}(n+1);\quad (n)\xrightleftharpoons[k_{-1}]{k_1}(n+1)\cdots(n+p-1)\xrightleftharpoons[k_{-1}]{k_1}(n+p)\equiv(n)\xrightleftharpoons[k_{-p}]{k_p}(n+p)\equiv \text{R1}^p.$$

$$\frac{[(n+1)]}{[(n)]} = \frac{k_1}{k_{-1}} \equiv K_1 = e^{-\beta\Delta G_{(n)\to(n+1)}} \Rightarrow k_2 = \frac{k_1k_1}{k_1+k_{-1}} = \frac{k_1}{1+K_1^{-1}} = K_1\frac{k_1}{1+K_1};\quad k_{-2} = \frac{k_{-1}k_{-1}}{k_1+k_{-1}} = \frac{1}{K_1}\cdot\frac{k_1}{1+K_1}.$$

Now

$$\frac{[(n+p)]}{[(n)]} = \frac{k_p}{k_{-p}} = \left(\frac{k_1}{k_{-1}}\right)^p = K_1^p \Rightarrow k_p = \frac{k_{p-q}k_q}{k_{-(p-q)}+k_q};\quad k_{-p} = \frac{k_{-(p-q)}k_{-q}}{k_{p-q}+k_{-q}}.$$

A little work shows that

$$k_{2^m} = \frac{k_{2^{m-1}}k_{2^{m-1}}}{k_{2^{m-1}}+k_{-2^{m-1}}} = \frac{k_{2^{m-1}}}{1+\dfrac{k_{-2^{m-1}}}{k_{2^{m-1}}}} = \frac{k_{2^{m-1}}K_1^{2^{m-1}}}{K_1^{2^{m-1}}+K_1^{2^{m-1}}\dfrac{k_{-2^{m-1}}}{k_{2^{m-1}}}}$$

$$= k_{2^{m-1}}\frac{K_1^{2^{m-1}}}{K_1^{2^{m-1}}+1} = k_1\prod_{j=0}^{m-1}\frac{K_1^{2^j}}{1+K_1^{2^j}} = k_1K_1^{2^m-1}\frac{1-K_1}{1-K_1^{2^m}} = \frac{K_1^{2^m}}{K_1^{2^m}-1}(k_1-k_{-1}).$$

A similar calculation gives an expression for the backwards rate constant, in turn giving

$$k_{-2^m} = \frac{k_{-2^{m-1}}k_{-2^{m-1}}}{k_{2^{m-1}}+k_{-2^{m-1}}} = \frac{k_{-2^{m-1}}}{1+\dfrac{k_{2^{m-1}}}{k_{-2^{m-1}}}} = \frac{k_{-2^{m-1}}}{1+K_1^{2^{m-1}}} = k_{-1}\prod_{j=0}^{m-1}\frac{1}{1+K_1^{2^j}} = k_{-1}\frac{1-K_1}{1-K_1^{2^m}}.$$

$$\Rightarrow k_{2^m}-k_{-2^m} = k_{-2^m}\left(K_1^{2^m}-1\right) = k_{-1}\frac{1-K_1}{1-K_1^{2^m}}\left(K_1^{2^m}-1\right) = k_{-1}(K_1-1) = k_1-k_{-1}.$$

We might compare the overall rate with an expression of the form

$$\frac{1}{\beta h} e^{-\beta \Delta G_{n \to n+1}} \approx 6.4 \cdot 10^{13}\ \mathrm{s}^{-1}.$$

Indeed,

$$k_{2^m} \approx k_{-2^m} \equiv k_p \approx k_{-p} = \frac{1}{2\beta h} e^{-\beta \Delta G_{(n)^* \to (n)}} \left[e^{-\beta \Delta G_{(n) \to (n+1)}} + 1 \right] \approx 1.7 \cdot 10^{12}\ \mathrm{s}^{-1};$$

$$\tau_p \equiv \left(k_p + k_{-p} \right)^{-1} \approx 5 \cdot 10^{-13}\ \mathrm{s}.$$

When viewed in light of experimental data (e.g., the base-pair addition rate constant is a very much smaller 10^6–10^7 s^{-1} at 25 ºC in 0.05–0.10 M Na^+ for short oligos),[79] it is clear that hydrogen-bond formation in zipping is not rate-limiting (indeed, intermediate complexes between Watson-Crick complementary nucleotides with one hydrogen bond should be very short-lived)[80] and therefore does not drive the kinetics of hybridization.[81] This does not render the transition state model useless,[82] but rather makes it clear that the zipping mechanism viewed in light of hydrodynamic effects is more complex. Viscosity evidently becomes a rate-limiting factor in base pair formation,[83] but does not affect the dynamics of hybridization strongly: addition of viscosity-increasing agents such as dextran sulfate [or phenol emulsions; see Pontius, B. W., and Berg, P.,

[79] Pörschke, D., "A direct measurement of the unzippering rate of a nucleic acid double helix," *Biophys. Chem.* **2**, 97 (1974).

[80] Cantor, and Schimmel, *Biophysical Chemistry*.

[81] If we include the inverse self-diffusion time for a base pair as a prefactor, however, we obgtain 10^6–$10^7 \mathrm{s}^{-1}$, as required. Cocco, S., et al., "Force and kinetic barriers to initiation of DNA unzipping."

[82] As an amusing aside, the (grossly inaccurate) numbers for base pair-formation without self-diffusion derived above are essentially the same as the hydrogen-bond vibration time-scales in liquid water. Hydrogen bonds in both DNA and water have energies of roughly –5 kcal/mol, and although (e.g.) viscosity clearly retards the kinetics of zipping (which is our main point in this subsection), there is a decent analogy between DNA hybridization and water freezing. (To further stretch the analogy, we cite Matsumoto, M., et al., "Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing," *Nature* **416**, 409 (2002).) This point of view brings up the issue of thermodynamic phases. We expect a 2ssDNA complex in the search phase to exhibit glassy behavior (i.e., a behavior between those liquid and a crystal. Glassy dynamics are typically characterized by energetic frustration, long time-scales, and significant fluctuations of an order parameter—something like the density of paired bases—near the glass temperature, some temperature (not too much) higher than T_m. See Mézard, M., "First Steps in Glass Theory," cond-mat/0005173, 2000. The notion of characterizing 2ssDNA as a glass gains some credence from the existence of an RNA glassy phase; see Pagnani, A., et al., "Glassy transition in a disordered model for the RNA secondary structure," cond-mat/9907125, 2000). It is at least conceivable that we could actually describe hybridization as a one-dimensional glass-to-crystal transition. It is potentially interesting (even if also potentially specious) to consider hybridizing DNA as a model system for condensed-matter physics—and vice versa—in this light. See, e.g., Kiang, C. , and Ramos, R., "The Percolation Transition in the DNA-Gold Nanoparticle system," physics/0111002 (2001).

[83] Wetmur, "Physical Chemistry of Nucleic Acid Hybridization."

PNAS **87**, 8403 (1990)] effectively concentrates the reacting strands and thereby accelerates the hybridization kinetics.[84]

Finally, it is worth noting that the role of viscosity highlights hydrodynamic interactions in hybridization. Bearing this in mind, in the excluded-volume interpretation it is hard to see how length effects could not govern the rate of base-pair formation (through time-varying moments of inertia as a duplex was zipped, etc.), whereas in the search-phase interpretation hydrodynamic effects should not lead to length-dependent kinetics of base-pair formation.

[84] Similarly, agents which shield repulsive ionic interactions (and also effectively deny the formation of secondary structure) between the phosphate groups on nucleic acid backbones can speed up hybridization, as with *E. coli.* single-stranded DNA binding (SSB) protein (Christiansen, C., and Baldwin, R. L., *J. Mol. Bio.* **115**, 441 (1977)) and cetyltrimethylammonium bromide (CTAB) (Pontius, B. W., and Berg, P., *PNAS* **88**, 8237 (1991)). A1 promotes increased associations between unpaired strands (Pontius, B. W., and Berg, P., *PNAS* **87**, 8403 (1990)), as does p53 for sticky-ended duplexes (Bakalkin, G. B., et al, *PNAS* **91**, 413 (1990)). There are many more such examples (such as Mg^{2+}, whose phosphate affinity features prominently in PCR and RAD52 protein [Mortensen, V.H., et al., *PNAS* **93**, 10729 (1996) and Sugiyama, T., et al., *PNAS* **95**, 6049 (1998)].

VIII. KINETIC ASPECTS
IV—UNZIPPING AND DISSOCIATION

Howorka et al.[85] give experimental values of denaturation rate (a.k.a. dissociation rate, thermal off-rate, reverse rate, etc.) constants:

Nanopore data:

Untethered sequence	k_{off} [s^{-1}]	K_{d} [M]	$\Delta G°$ [kcal mol^{-1}]
5′-GGTGAATG-3′	1.2	$9.2 \cdot 10^{-8}$	–9.2
5′-TACGTGGA-3′	3.4	$1.5 \cdot 10^{-7}$	–8.9
5′-GGTGAAT-3′	16	$1.5 \cdot 10^{-6}$	–7.7

Bulk (melting profile) values (k_{on} values of 10^7 M^{-1} s^{-1} assumed by authors):

Untethered sequence	k_{off} [s^{-1}]	K_{d} [M]	$\Delta G°$ [kcal mol^{-1}]
5′-GGTGAATG-3′	0.4	$3.6 \cdot 10^{-8}$	–9.8
5′-TACGTGGA-3′	1.7	$1.7 \cdot 10^{-7}$	–8.9
5′-GGTGAAT-3′	8	$8.3 \cdot 10^{-7}$	–8.0

The denaturation rate constant varied exponentially with temperature, as also remarked by Reynaldo et al.[86] Ionic strength does not appear to influence the unzipping rate strongly.[87]

Cantor and Schimmel cite measured relaxation kinetic parameters for complementary oligonucleotides at 21–23 ºC given in Table 6.[88] The large reverse-activation energies are evidently due to the energy cost of broken base pairs.

The kinetic equations for unzipping are difficult to deal with because of coupling between phenomena occurring at different relaxation time-scales[89] although unzipping of DNAs with heterogeneous sequences can be expected on theoretical grounds to go

Table 6. Relaxation kinetic parameters (see text)

[85] Howorka, et al., "Kinetics of duplex formation."

[86] Reynaldo, L. P., et al., "The Kinetics of Oligonucleotide Replacements," *J. Mol. Bio.* **297**, 511 (2000).

[87] Pörschke, "A direct measurement."

[88] Cantor and Schimmel, *Biophysical Chemistry*. The table is adapted in turn from Riesner, D., and Römer, R., *Physico-Chemical Properties of Nucleic Acids*, Vol.□, ed., J. Duchesne. Academic Press, London (1973).

[89] Pörschke, D., "A direct measurement."

Sequence	k_{-} [s^{-1}]	E_a [kcal mol^{-1}]
A_9	640	30
A_{10}	175	45
A_{11}	28	53
A_{14}	1	75
A_4U_4	3000	37
A_5U_5	150	50
A_6U_6	8	60
A_7U_7	0.8	65
A_2GCU_2	450	33
A_3GCU_3	3	50
A_4GCU_4	1.5	26
A_5G_2	340	43
A_4G_3	5	44

through L bases in a time $t_L \sim \exp(□\ L)$.[90] Indeed, the dissociative kinetics of DNA as a whole is still poorly understood, since (as we shall see) the relevant dynamics vary with length scale. Reynaldo et al. argue that the dissociation of a short duplex should have a rate constant of the form $\sim 2k_f\,(L_{hyb} - 2.5)K_1^{-L(hyb)+2.5}$, where k_f is the base pair formation rate constant ($\sim 10^6$–10^7 s^{-1}), K_1 is the base-pair equilibrium constant (~ 10), and the 2.5 numbers arise from ~3–4 bp for nucleation.[91] This model at first seems inconsistent with data from other experiments as well as theoretical work focusing on dissociation; we (and they) believe that their model does not apply to oligos of more than ~ 20 base pairs.[92] Since their model for the kinetics of oligonucleotide replacement is largely based on this quantitative underpinning we are loath to use it. That said, Reynaldo et al. obtained experimental data that we can draw lessons from, and their work can clearly support the general statement that at physiological temperatures the displacement pathway dominates the kinetics, although within PCR regimes dissociation should dominate.[93]

[90] Lubensky, D. K., and Nelson, D. R., "Single Molecule Statistics and the Polynucleotide Unzipping Transition," cond-mat/0107423 (2001). This paper also remarks on the similarities between unzipping and wetting transitions.

[91] Reynaldo, et al., "The Kinetics of Oligonucleotide Replacements."

[92] Strunz, T., et al., "Dynamic force spectroscopy of single DNA molecules," *PNAS* **96**, 11277 (1999), and Cocco et al., "Force and kinetic barriers to unzipping of the DNA double helix."

[93] Quartin, R. S., et al., "Branch migration mediated DNA labeling and cloning," *Biochem.* **28**, 8676 (1989), cited in Reynaldo et al.

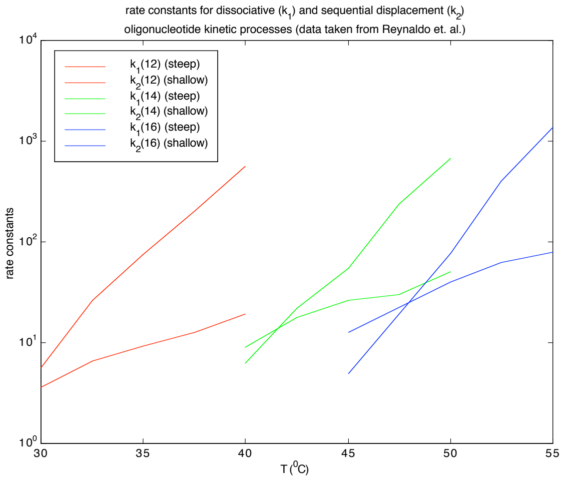


Figure 5.

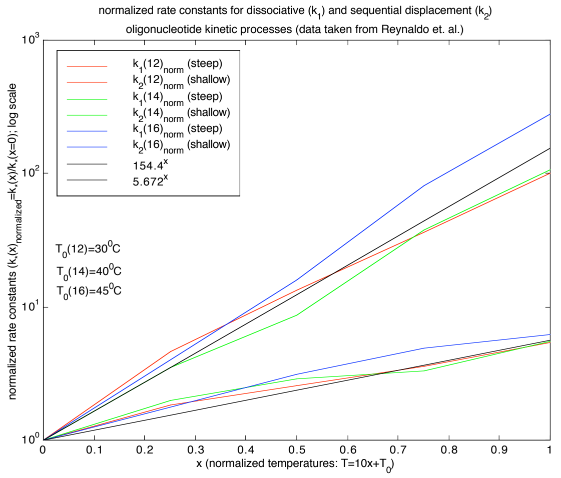


Figure 6.

Figures 5 and 6. Data taken from Reynaldo et al. Rate constants in Figure 5 are in units of 10^{-6} s^{-1} ($\boldsymbol{k_1}$) and M^{-1} s^{-1} ($\boldsymbol{k_2}$).

Table 7. Data from Reynaldo et al.

Activation energies (kcal mol^{-1})	12 base pairs duplex	14 base pairs duplex	16 base pairs duplex
Dissociation	85	94	118
Replacement	30	32	39

A treatment of dissociative kinetics more appropriate for scales of interest to us was given by Strunz et al.[94] By performing driven unbinding experiments on DNA and extrapolating to 0 unbinding force, they sought to determine the thermal off-rate. Cocco et al. remark that Strunz et al.'s interpreted "thermal off-rate" is really not that at all, but

[94] Strunz, "Dynamic force spectroscopy."

rather is related to the activation energy for dissociation. However, we are interested chiefly with the scaling of the dissociative kinetic rate, and so their extrapolation is useful□even in light of this objection. The “thermal off-rate” was found to go as $10^{[3\pm1]-[0.5\pm0.1]L(hyb)}$ s^{-1}; accordingly, the energy gap between the barrier and minimum increased linearly. A frequency prefactor included in this strongly increases with L_{hyb} due to the inclusion of extra degrees of freedom introduced by extra bases. Since the base pair equilibrium constant is ~ 10 this prefactor is evidently proportional to $10^{[0.5\pm0.1]L(hyb)}$ s^{-1}.[95]

Boundaries between two ss and a single ds region of a dsDNA complex (2ss-ds boundaries, such as in nucleation of hybridization or a denatured “bubble”) all consist of approximately four (for an end of a strand) or eight (for an interior segment of a dsDNA complex, i.e., for denaturated bubbles) base pairs, irrespective of the lengths of DNA involved. At a 2ss-ds boundary the hydrogen bonds may be broken but partial base-stacking contributions to the free energy remain.[96] Since base stacking contributes roughly half of the energy barrier, and the base-pair equilibrium constant is ~ 10, we offer this (combined with the cooperativity of the melting transition) as a “mickey-mouse” explanation of the observed and predicted ~ $10^{const\text{-}[0.5-0.6]L(hyb)}$ s^{-1} off rates (which will conflict with the model used by Reynaldo et al.): Pörschke[97] found an off-rate of $10^{8-0.5L(hyb)}$ s^{-1} for L_{hyb} ~ 8-18, and Cocco et al. predict an off-rate of $10^{6.3-0.6L(hyb)}$ s^{-1}. The combination of non-hydrogen bonded bases (and the concomitant high enthalpy contribution) with high rigidity (and hence low entropy contribution) in a boundary region results in a large free-energy barrier to unzipping. [98]

On shorter scales, off-rates for short oligomers can be computed from the equilibrium relation $k_{reverse} = k_{forward}\, e^{-\square\square G}$.[99] We remark that this must be done consistently; that is, for oligonucelotides that are short enough that (1) what are generally thought to be excluded-volume effects do not come into play, (2) the forward reaction is diffusion-controlled, and (3) the frequency prefactor above does not manifest itself (i.e., there ought not to be more than a few relaxation modes of the oligos). All in all this tack appears to work decently, but only up to (at most) ~ 20 base pairs. However, it does offer an explanation of Reynaldo et al.’s results and theoretical interpretations. It also serves to

[95] Cocco et al., “Force and kinetic barriers to initiation of DNA unzipping.”

[96] Ibid.

[97] Cited in several of our references as Pörschke, D. J., *Mol. Bio.* 62, 361 (1971).

[98] Cocco et al., “Force and kinetic barriers to initiation of DNA unzipping.”

[99] Winfree, E., *Algorithmic Self-Assembly of DNA,* and Cocco et al., “Force and kinetic barriers to unzipping of the DNA double helix.”

illustrate the differences in dynamics over a fairly narrow length scale of 10–40 base pairs. Since this is the whole regime in which we are interested, it is important to take all of these dynamical phenomena into account.

IX. KINETIC ASPECTS
V—SUMMARY AND CONCLUSIONS

The net result of this is that it is probably necessary to use a more complicated model of hybridization (with some intermediate 2ssDNA complex[100] between proximal and activated complexes). Now we have the kinetic cartoon

$$w + w' \underset{k_{-d}}{\overset{k_d}{\rightleftharpoons}} \overbrace{w \cdot w'}^{\text{co-located ssDNAs}} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \overbrace{(ww')_{2ss}}^{\text{2ssDNA complex}} \underset{k_{-2}}{\overset{k_{+2}e^{-\beta\Delta G_{n*}}}{\rightleftharpoons}} \overbrace{(ww')_{nuc*}}^{\text{activated complex}} \underset{k_{-3}}{\overset{(\tau_h)^{-1}}{\rightleftharpoons}} \overbrace{(ww')_{nuc}}^{\text{nucleated complex}} \underset{k_{-4}}{\overset{k_{+4}e^{-\beta\Delta G_h}}{\rightleftharpoons}} \overbrace{(ww')_{hyb}}^{\text{hybridized complex}}.$$

Neglecting sequence variations, this scheme takes into account three length scales: the lengths of the two ssDNAs and their overlap. In light of the theory work and the experimental data available, it appears to be the simplest model that can address the framework of diffusive hybridization in a self-consistent way.

It can in principle be extended to incorporate mixed ss-dsDNA linear structures by regarding them as collections of rigid rods joined by floppy springs (although we can expect such a tactic to meet with some difficulty; see above with respect to the difficulty of applying transition-state theory to heterogenous ss/dsDNA complexes). Force-extension measurements on ssDNAs confirm that a WLC model can be used to describe their elastic response, and by building the force relationship into such a model we could predict diffusion behavior for mixed structures.[101]

If the search phase holds, then it is not unreasonable to expect that (since for long overlaps the search can proceed in either of two directions) two ssDNAs that support a possible MSH will, in the ensemble, assume the MSH configuration with a probability not necessarily strictly related to the nominal Boltzmann-Gibbs distribution. For instance, an ssDNA could be designed that supports multiple possible hybridizations (not, strictly speaking, MSHs because of the context) but with different free energies. This may be a way to settle the issue conclusively and can presumably help in the design and evaluation of hybridization protocols.

[100] The kinetic avenues to and from which ought to embody our proposed search phase, for example.

[101] Bustamante, C., et al., “Single-molecule studies of DNA mechanics,” *Curr. Opin. Struct. Biol.* **10**: 279 (2000).

It is vital to keep in mind the heretofore implicit disconnect between the equilibrium/thermodynamic and nonequilibrium/kinetic pictures. We expect that the MSH probabilities ought to look something like

$$\Pr\left(\# MSH_{neq}^{actual}(t) > 0\right) \approx A \sum_{l} e^{-Cl} \underbrace{\left(1 - \frac{1}{1+[w]_0 \tilde{C}\, t/\sqrt{l}}\right)}_{\substack{\text{fraction of ssDNAs}\\ \text{that have hybridized}}} \Pr\left(\# MSH_{l}^{possible} > 0\right).$$

In the limit of long time this gives the equilibrium result, but only experiment can establish anything—the validity of the model, values of the constants, etc. The crossover between nonequilibrium and equilibrium regimes (which this formula only begins to hint at, since the dynamics of the system will change drastically as more batch DNAs anneal into progressively longer dsDNA complexes) will of course depend on operating characteristics of any protocol, and so (especially when contemplating microfluidic systems, as we are implicitly) it is impossible to say in general when one can use the equilibrium thermodynamic description.

But this is about engineering, so we can live with that sort of uncertainty. In practice we will want to try to use this to help find a desirable tradeoff between time and average batch length scales to mitigate error rates in experimental attempts at synthesis. As an example, we show a function proportional to the probability of actual MSHs as a function of time and half-word length n (i.e., with a batch of ssDNAs of equal length $2n$) for the synthesis of a 10,000-mer with an initial concentration of 10^{-5} M, nucleation constant of $5 \cdot 10^{5}$ M^{-1} s^{-1}, and temperature of 55 ºC.

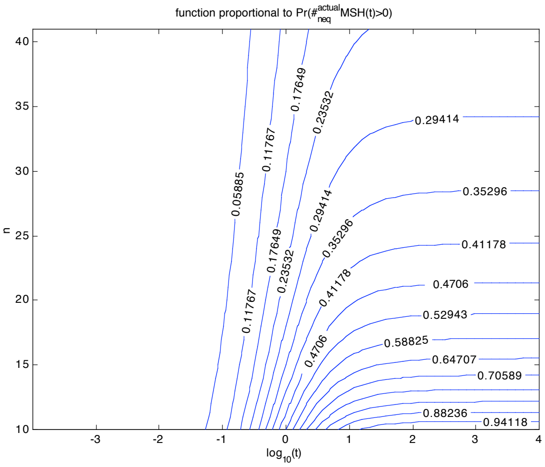


Figure 7.

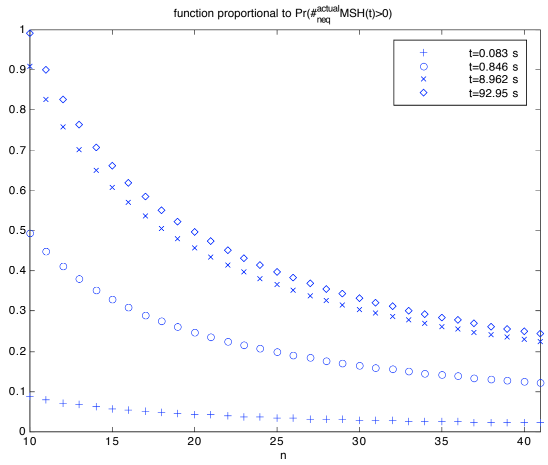


Figure 8.

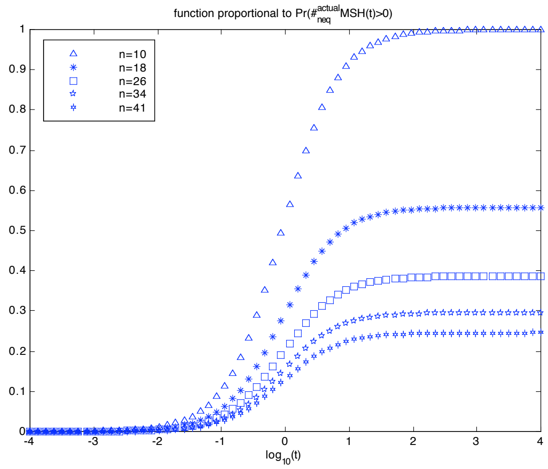


Figure 9.

Figures 7, 8, 9. Function proportional to the probability of actual MSHs as a function of time and half-word length *n* (i.e., with a batch of ssDNAs of equal length 2*n*) for the synthesis of a 10,000-mer with an initial concentration of 10^{-5} M, nucleation constant of $5 \cdot 10^5$ M^{-1} s^{-1}, and temperature of 55 °C. Time is given in seconds.

X. OPTIMAL VARIABLE-LENGTH DECOMPOSITIONS

In general, even if we really knew how, finding an optimal variable-length decomposition appears intractable: there is very good reason to believe, for instance, that the problem is NP-complete,[102] and tricky at that. Without attempting to prove this but bearing it in mind, we outline a simulated annealing[103] algorithm for the variable-length problem. Let the sticky-ended dsDNA X be represented as a (notionally hairpinned) ssDNA. Given a partition or composition □of X, we can compute $\#MSH_l(X;\ \square)$: let $init_l(X;\ \square)$ denote the set of l-prefixes; similarly, let $term_l(X;\ \square)$ denote the set of l-suffixes. For each l-prefix (respectively, l-suffix) arising from the partition □, count the number of times its reverse complement occurs in $init_l(X;\ \square)$ (respectively, $term_l(X;\ \square)$). Then $\#MSH_l(X;\ \square)$ is the sum of these counts over both $init_l(X;\ \square)$ and $term_l(X;\ \square)$. (Depending on symmetry considerations, we might divide this number by two, but it will not matter for our purposes here.) Put (for instance)

$$E(X;\square)\equiv\mathscr{F}\,\square\left\langle\Big(L(X;\square)\,\square\,\big\langle L(X;\square)\big\rangle\Big)^{2}\right\rangle,\square_{l}\,l\cdot\#\,MSH_{l}^{possible}(X;\square)\Big[\ ,$$

where $\mathscr{F}$ denotes a nonnegative (possibly nonlinear) functional. With an appropriate formal energy such as this, a standard simulated annealing procedure will give near-optimal decompositions (but it will take time). More generally, a functional incorporating free energies of duplex formation for partial compositions and undesirable configurations could be implemented in principle.

The problem with such a construction is that it will be inefficient. Indeed, we might expect it to encounter some of the same difficulties as simulated annealing attacks

[102] Roughly speaking, this means that we expect (i.e., it is conjectured but unproven that) the difficulty of finding an optimal decomposition in general ought to increase exponentially with the overall length. The notion of NP complexity and the conjectures surrouding it are central to theoretical computer science. See (e.g.) Cormen, T. H., et al., *Introduction to Algorithms*, MIT, Cambridge, Massachusetts (1990).

[103] Kirkpatrick, S., et al., "Optimization by simulated annealing," *Science* **220**, 671 (1983). Simulated annealing works by analogy with the cooling of metals (i.e., annealing) into stable/low-energy ground states (the optimization part).

on the (NP-complete) number partition problem (NPP).[104] These notorious difficulties stem from a superabundance of local metastable minima;[105] the biological analogy with energy landscapes in protein folding (for which simulated annealing is similarly problematic) is an appropriate one in this context. Moreover, a common rule of thumb in optimization is that genetic algorithms will have problems if simulated annealing does; the consequent “method of last resort” tag attached to evolutionary optimization discourages its application here (i.e., simulated annealing should work at least as well as genetic algorithms, and it will run faster). The potential ineffectiveness of simulated annealing (as of this writing we have not drawn a conclusion on this point, although we have a simulated annealing implementation in code; see Appendix C) suggests that we consider heuristic approaches such as the Karmarkar-Karp differencing method.[106] In any event, a refined software implementation taking into account the various dynamical regimes of hybridization and computational obstructions would probably be very complex and run very slowly (which is already the case with our code).

[104] Johnson, D. S., et al., “Optimization by simulated annealing: an experimental evaluation; part II, graph coloring and number partitioning,” *Operations Research* **39**, 378 (1991).

[105] Ferreira, F. F., and Fontanari, J. F., “Probabilistic Analysis of the Number Partitioning Problem,” adap-org/9801002 (1998).

[106] See (e.g.) Cormen, T. H., et al., *Introduction to Algorithms,* MIT, Cambridge, Massachusetts (1990).

XI. CONCLUSION

Despite all the uncertainties, caveats, and differences between relevant regimes, the overall picture is generally agreed upon. While on the one hand there "is no satisfactory quantitative theory predicting the statistical distribution of hybridization,"[107] at the same time, "the self-assembly of aperiodic [DNA] structures should also be considered…progress in this field will require detailed knowledge of the physical, kinetic, structural, dynamic, and thermodynamic parameters that characterize DNA self-assembly. Additionally, improved methods for error reduction and purification must be developed."[108] What we have attempted to do here is to show the way forward in a general light.

With respect to comparing generic synthesis protocols in practice, we can make some easy qualitative engineering comparisons:

Techniques	Advantages	Disadvantages
Baseline batch protocol	Simple theory and implementation, somewhat scalable	Slow, fault prone
Parallel baseline ("binary tree" flow) batch protocol	Simple theory, somewhat scalable, fast	Somewhat complex implementation, fault prone
Parallel baseline batch subprotocols followed by ligation	Simple theory, somewhat scalable, fast, somewhat fault tolerant	Complex implementation
Fully variable-length batch protocol	Scalable, fault tolerant	Complex theory and somewhat complex implementation, slow
Fully variable-length batch subprotocols followed by ligation	Scalable, fast, fault tolerant	Complex theory and implementation

Ultimately, however, the rubber has to hit the road, and of course everything in our discussion defers to that central truth. Although MSHs are clearly the preeminent specter haunting the road to the production-scale synthesis of long DNAs, there is every reason to expect them to be surmountable as an engineering obstacle, given sufficient care. Likewise, although (as we have both seen and demonstrated) much of the science

[107] Carbone, A., and Gromov, M., "Mathematical slices of molecular biology," Preprint IHES/M/01/03 (2001).

[108] Winfree, E., et al., "Design and self-assembly of two-dimensional DNA crystals."

behind the engineering is still uncertain, we have tried to outline the key issues—mathematical, thermodynamic, and kinetic—in such a way that such approaches to synthesis and generic batch DNA manipulation can be more fully understood.

We conclude with a quotation and a caveat:

> The link between computation and SA [self-assembly] may find most use in defining the structures accessible by SA, both theoretically and practically. Theoretically, because the mathematics of computation may be used to classify self-assembled structures…or to analyze the resources… needed to create a particular structure…. Practically, because tilings that encode computations provide new synthetic targets—structures more complex, in general, than any considered by chemists so far. [109]

Although taken out of context, this sentiment—as we have seen—is only partially true: there is more to the analysis of self-assembly than the mathematics of computation. Despite all the old hype about DNA computation, there is still a need for detailed experimentation to elucidate the nature and parameters of the processes involved in DNA self-assembly. This is why it has taken the better part of a decade to see any real progress in DNA computation.[110] Although our discussion is generic and may be plagued with similar missteps, we trust that it sheds some light on the problems that need to be answered to facilitate the synthesis of long DNAs and of more complex DNA nanostructures.

[109] Rothemund, P. W. K., "Using lateral capillary forces to compute by self-assembly," *PNAS* 97, 984 (2000).

[110] Adleman, L. M., "Molecular Computation of Solutions to Combinatorial Problems," *Science* **266**, 1021 (1994) and Braich, R. S., et al., "Solution of a 20-Variable 3-SAT Problem on a DNA Computer," published online in *Science Express* (10.1126/science.1069197), 14 March 2002.

APPENDIX A

NOTES ON PERSISTENCE LENGTH AS A STATISTICAL FUNCTION OF ANGLE VARIANCE

APPENDIX A
NOTES ON PERSISTENCE LENGTH AS A STATISTICAL FUNCTION OF ANGLE VARIANCE

It can be shown that the persistence length depends only on the variance of the angle distribution of a fixed step-length random walk. We illustrate this for such a general walk in the plane:

$$\mathbf{x}_0 := \mathbf{0}; \quad \mathbf{x}_1 := e_1; \quad \theta_1 := 0$$

$$\mathbf{x}_{k+1} := \mathbf{x}_k + R_{\theta_{k+1}}\left(\mathbf{x}_k - \mathbf{x}_{k-1}\right); \quad R_{\theta_{k+1}} := \begin{pmatrix} \cos\theta_{k+1} & -\sin\theta_{k+1} \\ \sin\theta_{k+1} & \cos\theta_{k+1} \end{pmatrix}$$

$$\theta_{k+1} \sim p : [-\pi, \pi) \to \mathbf{R} \quad \text{s.t.} \quad \int_{[-\pi,\pi)} p = 1; \quad \left\langle \theta_{k+1} \right\rangle \equiv 0$$

$$\dot{\mathbf{x}}_k := \mathbf{x}_{k+1} - \mathbf{x}_k \quad \Rightarrow \quad \dot{\mathbf{x}}_k = R_{\theta_{k+1}}\left(\mathbf{x}_k - \mathbf{x}_{k-1}\right) = \prod_{j=1}^{k+1} R_{\theta_j} e_1 = R_{\left(\sum_{j=1}^{k+1}\theta_j\right)} e_1$$

$$\Rightarrow \quad \left\langle \dot{\mathbf{x}}_k \cdot \dot{\mathbf{x}}_0 \right\rangle = \left\langle R_{\left(\sum_{j=1}^{k+1}\theta_j\right)} e_1 \cdot e_1 \right\rangle = \left\langle \cos\left(\sum_{j=1}^{k+1}\theta_j\right) \right\rangle.$$

Now, by the central limit theorem, we have that

$$\Pr\left(a\sqrt{n\sigma_p^2} \le \sum_{j=1}^{n}\theta_j \le b\sqrt{n\sigma_p^2} \right) \xrightarrow{n\to\infty} \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\tfrac{x^2}{2}\right) dx$$

$$\Rightarrow \quad \Pr\left(r \le \sum_{j=1}^{n}\theta_j \le r + dr \right) \approx \frac{1}{\sqrt{2\pi}} \int_{\frac{r}{\sqrt{n\sigma_p^2}}}^{\frac{r+dr}{\sqrt{n\sigma_p^2}}} \exp\left(-\tfrac{x^2}{2}\right) dx = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{r^2}{2n\sigma_p^2}\right) dr :$$

$$\left\langle \cos\left(\sum_{j=1}^{n}\theta_j\right) \right\rangle = \int \cos\left(\sum_{j=1}^{n}\theta_j\right) dF_{\sum_{j=1}^{n}\theta_j} = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} \exp\left(-\frac{r^2}{2n\sigma_p^2}\right) \cos r\, dr$$

$$= \sqrt{n\sigma_p^2}\, \exp\left(-\frac{n\sigma_p^2}{2}\right)$$

(Landau and Lifshitz[111] give an elegant calculation which reproduces the exponential term here and which is more tractable, but which is also more involved.) For this model we have a correlation function which yields the persistence length and a concomitant energy functional:

[111] §151 in Landau, L. D., and Lifshitz, E. M., *Statistical Physics,* 2nd ed., Addison-Wesley, Reading, Massachusetts (1969).

$$\left\langle \dot{\mathbf{x}}_k \cdot \dot{\mathbf{x}}_0 \right\rangle = \sqrt{(k+1)\square_p^2}\,\exp\square\square\frac{(k+1)\square_p^2}{2}\square \sim \exp\square\square\frac{(k+1)}{l_p}\square\square \quad l_p = \frac{2}{\square_p^2};$$

$$E \equiv \frac{1}{\square\square_p^2}\overset{L}{\underset{0}{\square}}\left|\ddot{\mathbf{x}}(s)\right|^2 ds = \frac{1}{\square\square_p^2}\overset{L}{\underset{0}{\square}}\square\square^2(s)ds.$$

(Here, the stiffness would be defined as twice the constant prefactor.)

As for the mean square end-to-end length, we have

$$\mathbf{x}_k = \overset{k\square 1}{\underset{i=0}{\square}} R_{\square_{j=1}^{i+1}\square_j} e_1 \;\square\; \mathbf{x}_k \cdot \mathbf{x}_k = \left\langle \overset{k\square 1}{\underset{i=0}{\square}} R_{\square_{j=1}^{i+1}\square_j} e_1, \overset{k\square 1}{\underset{m=0}{\square}} R_{\square_{l=1}^{m+1}\square_l} e_1 \right\rangle$$

$$= \overset{k\square 1}{\underset{i=0}{\square}}\,\overset{k\square 1}{\underset{m=0}{\square}} \left\langle R_{\left(\square_{j=1}^{i+1}\square_j\right)\square\left(\square_{l=1}^{m+1}\square_l\right)} e_1, e_1 \right\rangle = k + 2\underset{i<m}{\square}\left\langle R_{\square_{l=i+2}^{m+1}\square_l} e_1, e_1 \right\rangle = k + 2\underset{i<m}{\square}\cos\square R_{\square_{l=i+2}^{m+1}\square_l}\square$$

$$\square\; \left\langle \left\|\mathbf{x}_k\right\|^2 \right\rangle = k + 2\underset{i<m}{\square}\left\langle \cos\square R_{\square_{l=i+2}^{m+1}\square_l}\square \right\rangle = k + 2\overset{k\square 1}{\underset{n=1}{\square}}(k\square n)\left\langle \cos\square R_{\square_{l=1}^{n}\square_l}\square \right\rangle$$

$$= k + 2\overset{k\square 1}{\underset{n=1}{\square}}(k\square n)\sqrt{n\square_p^2}\,\exp\square\square\frac{n\square_p^2}{2}\square$$

This broadly agrees with the expression in Landau and Lifshitz.

APPENDIX B

NOTES ON DE BRUIJN SEQUENCES, CYCLIC CODES, AND FRAME SHIFTS

APPENDIX B
NOTES ON DE BRUIJN SEQUENCES, CYCLIC CODES, AND FRAME SHIFTS

A de Bruijn sequence[112] of order n is a sequence of 2^n binary words of length n such that for any word w in the sequence, the next word is of the form $w_2 \ldots w_n z$, where z is a generic Boolean variable. Such sequences can be shown to exist, and in fact their number can be computed, as we shall see below. For instance, the sequence 000 001 010 101 011 111 110 100 (alternatively denoted 00010111, in what we term a block representation) is a de Bruijn sequence of order 3. Indeed any such sequence can be shown to arise as a Eulerian circuit (i.e., a circuit that traverses every edge exactly once) on a directed graph[113] such as the one below (the de Bruijn graph of order 3):

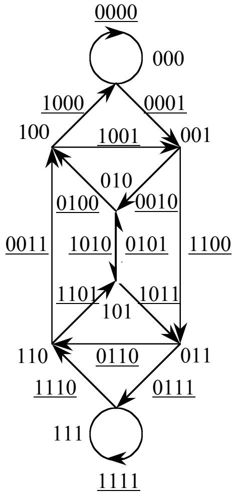


Figure B-1. The de Bruijn Graph of Order 3.
Underlined binary words indicate edges.

It is not hard to see that the de Bruijn sequence above defines a Hamiltonian path (i.e., a path that visits each vertex exactly once) on the graph shown. Similarly, an Eulerian circuit on this graph will define a de Bruijn sequence of order 4 (e.g., 000

[112] See, e.g., van Lint, J. H., and Wilson., *A Course in Combinatorics,* Cambridge University Press, Cambridge (1992), or Lempel, A., "On a Homomorphism of the de Bruijn Graph and Its Applications to the Design of Feedback Shift Registers," *IEEE Trans. Computers* **19**, (1970).

[113] See, e.g., Bollobás, B., *Modern Graph Theory,* Springer, New York (1998).

000 □ 001 □ 011 □ 111 □ 111 □ 110 □ 100 □ 001 □ 010 □ 101 □ 011 □ 110 □ 101 □ 010 □ 100 □ 000, which can be denoted 0000111100101101). It is interesting that the technique of pushing Eulerian circuits (which are easy to deal with) into Hamiltonian paths (which are notoriously difficult to deal with) on special directed graphs appears in another context for fault-tolerant batch DNA manipulation—namely, in the construction of thermodynamically homogeneous oligos that are statistically regular in a particular sense,[114] as well as in the analysis of sequencing by hybridization.[115] The correspondence between Eulerian circuits and Hamiltonian paths is one of the "pointy sticks" of bioinformatics.

The particular quality of de Bruijn sequences that merits attention here is that a block representation is resistant to frame-shifting; that is, any *n* consecutive digits in a de Bruijn sequence of order *n* are uniquely determined. As such, quaternary de Bruijn sequences are natural models for ssDNAs that will be resistant to shift hybridizations and MSHs in particular.[116] The number of binary de Bruijn sequences of order *n* + 1 (equivalently, the number of Eulerian circuits on the graph of order *n*) is

$$2^{2^{n\square 1}\square n}.$$

This can be proved by the BEST and matrix-tree theorems.[117] For instance, there are 16 de Bruijn sequences of order 4 (equivalently, Eulerian circuits on the graph depicted above). The number of quaternary de Bruijn sequences of order *n* + 1 (hence of length 4^{n+1}) can be calculated. In any event, it follows that there are 324 quaternary de□Bruijn sequences of order 1/length 16, which will be the only case of concern to us here. Even so, the probability that a uniformly random cyclic quaternary sequence of length 16 is a de Bruijn sequence is 324/268439590 □ $1.207 \cdot 10^{-6}$. [118] This is too small to

[114] Huntsman, S., in preparation.

[115] See Pevzner, P. A., "L-tuple DNA sequencing: computer analysis," *J. Biomolecular. Structure and Dynamics* **7**, 63 (1989); Kandel, D., et al., "Shuffling Biological Sequences," Preprint (1995); and Arratia, R., et al., "Euler circuits and DNA sequencing by hybridization," *Disc. Appl. Math.* **104**, 63 (2000).

[116] Ben-Dor, A., et al., "Universal DNA Tag Systems: A Combinatorial Design Scheme," Preprint (2000).

[117] Bollobás, B., *Modern Graph Theory*, Springer, New York (1998).

[118] By the Burnside-Frobenius lemma, there are (8·4 + 4·16 + 2·256 + 48 + 416)/16 = 268439590 cyclic quaternary sequences of length 16. See van Lint, J. H. and Wilson, R. M. A Course in Combinatorics. Cambridge, Cambridge (1992), whose treatment we follow here. The Burnside-Frobenius lemma can be stated as follows: let G □ Sn, let □ : G □ X □ X be an action on X, write Fixg(□) for the set of fixed points {x : □g(x) = x} and write Orb(□) for its space of orbits. Then #Orb(□)·#(G) = □g□G #Fixg(□). Consider the set of linear k-ary sequences of length n and the natural action □□of Z/nZ on it, so that Orb(□□ is the set of cyclic k-ary sequences of length n. Let g denote here a generic element of Z/nZ. If dln then #{g : (n, g) = d} = #{h : (n/d, h) = 1} = □(n/d), where the last identity is tantamount to a

expect to see very often; in particular, we cannot reasonably hope to preempt generic MSHs by assembling with sticky ends incarnating de Bruijn sequences.

Since there are 324 quaternary de Bruijn sequences of length 16 (and they are cyclic), there are 16·2·324 = 10,368 linear sequences of length 8 (vs. 65,536 generic sequences of length 8) obtained by taking half of a de Bruijn cycle (i.e., 8 consecutive positions). (It is not clear that these sequences are unique, and we have not determined this.) Lempel illustrated that the preimage of a binary de Bruijn sequence of order *n* (with respect to a homomorphism between de Bruijn graphs) is a pair of disjoint "half-de□Bruijn sequences" of "order" *n*+1; operationally this might suggest incorporating these half-de□Bruijn sequences within sticky ends.[119] It is not clear how to proceed along these lines, however.

Coding theory also bears on the MSH issue. We consider quaternary codes, that is, codes over an alphabet of four elements and presumably with some internal structure. For instance, the symbols might be represented by the elements of the finite field with four elements, *GF*(4).[120] This field has the addition and multiplication tables:

+	**0**	**1**	**□**	***s***
0	0	1	□	*s*
1	1	0	*s*	□
□	□	*s*	0	1
s	*s*	□	1	0

□	**0**	**1**	**□**	***s***
0	0	0	0	0
1	0	1	□	*s*
□	0	□	*s*	1
s	0	*s*	1	□

Using these it is not hard to verify that *GF*(4) is isomorphic to the set {(0, 0), (0, 1), (1,□0), (1, 1)} (parentheses indicate ordered *n*-tuples) ≡ {00, 01, 10, 11}, with field operations given by componentwise addition and multiplication mod 2. That is, *GF*(4) □ *GF*(2)⊕*GF*(2). This early abstraction allows us to deal with binary coding schemes in the

definition of the Euler function. By inspection, #Fixg(□) = kd, so the Burnside-Frobenius lemma gives #Orb(□□) = n-1·□dln □(n/d)·kd.

[119] Lempel, A. "On a Homomorphism of the de Bruijn Graph and Its Applications to the Design of Feedback Shift Registers," *IEEE Trans. Computers* **19**, (1970). See also Annexstein, F. S., "Generating de Bruijn sequences: An Efficient Implementation," *IEEE Trans. Computers* **46**, (1997).

[120] The finite field with four elements (*GF*(4) or, alternatively, $\mathbf{F}_4$) is (essentially defined to be) isomorphic to the splitting field of $x^2 + x + 1$ over *GF*(2) (the finite field with two elements, which is itself isomorphic to {0, 1} under addition mod 2 and multiplication). That is, *GF*(4) is isomorphic to the set {*a* + *b*□| *a*, *b* □ *GF*(2)}, with □a root of $x^2 + x + 1$. Since by definition □² = □+1, multiplication goes as (*a* + *b*□)(*c* + *d*□) = (*ac*+*bd*) + (*ad*+*bc*+*bd*)□, where *a*, *b*, *c*, *d* □ *GF*(2). Dummit, D. S., and Foote, R. M., *Abstract Algebra*, Prentice-Hall, Englewood Cliffs, N.J. (1991).

same context of abstract coding schemes over $GF(4)$.[121] If instead of $GF(4)$ we want to consider $\mathbf{Z}_4$ (the group [or ring] of integers mod 4) then we will use the same correspondence without distinction.

The simplest example of an error-correcting code[122] is the binary *triplet parity code*: 0 is encoded as the codeword 000 and 1 as 111. A received triplet other than these is weighted: either it has two zeroes or two ones, according to which it is changed to 000 or 111 accordingly. This is a specific (but trivial) instance (3, 1) of the more general notion of a *linear binary* $(n, 2^k)$ *or* $(n, 2^k, d)$ *code*. Here, n (the *length*) denotes the number of symbols used to encode a sequence of k symbols (of which there are 2^k total—this is the code's *size*, typically denoted by M), and d refers to the minimum *distance* (from zero), or number of ones, in a codeword. It can be shown that a linear (n, M, d) code can correct $(d\text{-}1)/2$ or fewer errors; the (integral) number t of errors a code can correct is referred to as its *weight.* Finally, if there is no risk of confusion—or if it is more convenient—such a code may also be described as an (n, k) code.

Such a code $\mathcal{C}$ is specified (for example) by a *generator matrix* G which can be assumed to be in the form $(Id|A)$ where Id is the k-by-k identity matrix and A is a k-by-$(n–k)$ matrix (equivalently, the *dual code* $C^{\perp}$ may be characterized by the *parity check matrix* $(–A^T|Id)$). The rows of the matrix G are then the basis codewords, and a generic bit string x of length k is encoded by producing the linear combination of basis codewords whose first k bits equal x.

[121] An alternative framework using 3-bit base encodings (000, 010, 101, and 111) is outlined in Li, Z., "Algebraic properties of DNA operations," *Biosystems* 52, 55 (1999). While this is a natural way to accommodate string reversal in binary form, it does not allow for as tractable a means of leveraging the well developed theory of binary error-correcting codes.

[122] We employ all of the following references without distinction (or further citation) for general results on coding theory: MacWilliams, F. J., and Sloane, N. J. A., *The Theory of Error-Correcting Codes,* North-Holland, Amsterdam (1977); McEliece, R. J., *The Theory of Information and Coding,* Addison-Wesley, London (1977); Peterson, W. W., and Weldon, E. J., Jr., *Error-Correcting Codes,* 2nd ed., MIT Press, Cambridge, Massachusetts (1972); and Pless, V., *Introduction to the Theory of Error-Correcting Codes,* John Wiley, New York (1982). MacWilliams and Sloane is generally regarded as the definitive reference; McEliece contains valuable technical results on both coding and Shannonean information theory proper; Peterson and Weldon is an especially good reference for burst error-correcting codes; and Pless is an easy introduction to the subject.

$$\begin{matrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{matrix}$$

The standard generator matrix of the (7, 16, 3) (or (7, 4, 3) Hamming code.

Hence, a linear code can also be described by the span of its basis codewords.[123] It turns out that the (*n*-*k*, *k*, 3) Hamming and (23, 12, 7) Golay codes are the only nontrivial binary perfect (i.e., capable of correcting *t* errors) error-correcting codes. This surprising fact illustrates that the existence theory of classical error-correcting codes is deep and complex.

The minus-shift condition will always be satisfied for batch ssDNA halves that instantiate quaternary cyclic error-correcting codes (i.e., a cyclic permutation of a codeword is again a codeword); moreover, there are reverse-complement quaternary cyclic codes with (optimal) parameters as impressive as, for example, (15, 7, 7), (15, 9, 5), (15,□1, 4), and (15, 13, 2).[124] We argued elsewhere[125] that ssDNAs instantiating quaternary cyclic codes are well suited for batch DNA manipulation for physical as well as mathematical reasons. We briefly sketch some of the ideas behind this claim here.

Any de Bruijn sequence has the same thermodynamic profile (in the NN framework) as any other, and their cyclic permutations are also de Bruijn sequences. It is therefore appropriate to consider them (as a class) as a protocode generically formed from concatenating cyclic protocodes. In this context, it seems that we ought to treat the quaternary order 2 de Bruijn squences as a quaternary (17, 10) cyclic-invariant set of words contained in some (17, *k*, *d*) code with *k* close to 10 and *d* as large as possible.

[123] The decoding process is generally difficult: each codeword has a large coset of errorwords that (unless the code were engineered with viable algorithmic decoding schemes, the construction of which is largely the point of coding theory) has to be exhaustively searched. However, special decoding techniques exist (e.g., syndrome and Hamming decoding) that can dramatically reduce the computational effort involved. Still, when *n* is large enough an (*n*, *k*) code is typically infeasible to implement classically.

[124] Bogdanova, G. T., et al., "Error-Correcting Codes over an Alphabet of Four Elements," preprint (1999); Marathe, A., Condon, A. E., and Corn, R. M., "On Combinatorial DNA Word Design," in Winfree, E., and Gifford, D. K., eds. *DIMACS Workshop: DNA Based Computers V, June 14–15, 1999*, Vol. 54, American Mathematical Society, Providence, R.I. (2000); and Rykov, V. V., et al., "DNA Sequences Constructed on the Basis of Quaternary Cyclic Codes," *Proc. 4th World Multiconference on Systematics, Cybernetics, and Informatics*, SCI 2000/ISAS 2000, (2000).

[125] Huntsman, S., in preparation.

Happily, quaternary cyclic codes of length $16K + 1$ are automatically reversible,[126] and since the de Bruijn sequences are invariant as a set under complementation, it follows that a reverse-complementarity constraint holds, at least up to the minimum distance of any enveloping code. A frame-shift constraint can be subsequently incorporated as desired by selecting an appropriate subset of the de Bruijn sequences. The net import of all this is that the de Bruijn sequences appear to provide a natural starting point for building sequence sets for batch DNA manipulation (if not synthesis proper). Hence both the batch synthesis and manipulation of DNAs appears to go easier with cyclic coding schemes involved.

[126] MacWilliams and Sloane, cited as Theorem 4 in Rykov, V. V., et al., "DNA Sequences Constructed on the Basis of Quaternary Cyclic Codes," *Proc. 4th World Multiconference on Systematics, Cybernetics, and Informatics*, SCI 2000/ISAS 2000, (2000).

APPENDIX C

NOTES ON SIMULATED ANNEALING ATTACKS ON PARTITIONING PROTOCOLS

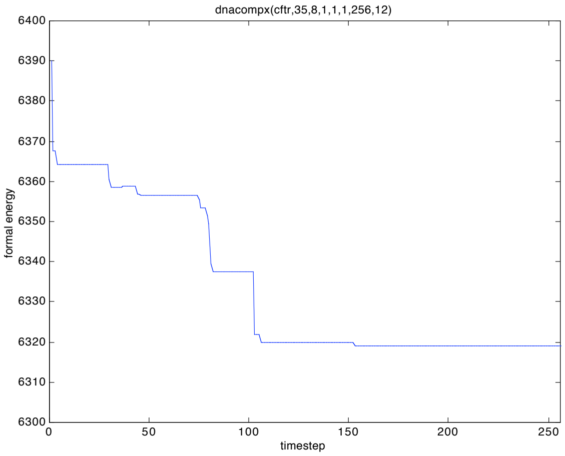
APPENDIX C
NOTES ON SIMULATED ANNEALING ATTACKS ON PARTITIONING PROTOCOLS

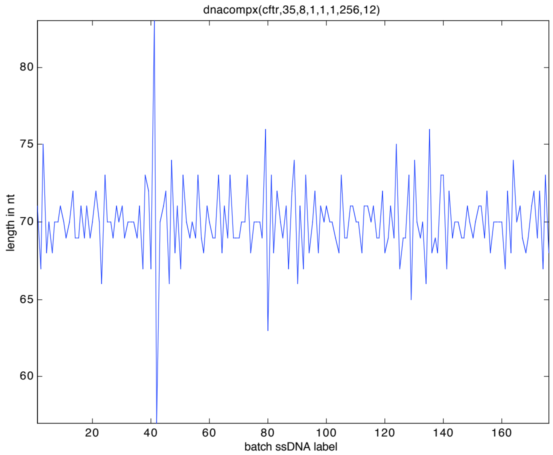
We first show an example output of the code given below on the clone DNA for human cystic fibrosis mRNA encoding a presumed transmembrane conductance regulator (CFTR).[127] The < and > marks are spacer nucleotides. Although a value of $n = 35$ such as we used appears to be a reasonable value (even if it implies longer batch ssDNAs than what we would typically envision), it is inappropriate to regard the output as realistic for many reasons, of which we list a few.

- It is unclear what the proper values of the coefficients in the energy functional (assuming that its form is itself generically appropriate) ought to be; we use equal coefficients.
- Our overlap value of 8 was picked on a whim.
- Our annealing temperature and schedule (logarithmic) are similarly unjustified, as is our protocol in which the extent of a configuration change depends explicitly on these factors.
- 256 timesteps are not very many for a batch of 176 ssDNAs.
- We present only a single run here, rather than an ensemble average.

Moreover, our code is computationally expensive (although porting to C would greatly reduce runtimes): this run took over 2 hours on a desktop computer. We include it mainly as a proof of principle. Obvious extensions of its functionality would include incorporating energy terms accounting for type IIa mishybridizations and sequence particulars.

[127] Online at http://opal.msu.montana.edu/cftr/cftrsequence.htm





Figures C-1 and C-2. Results from a Run of dnacompx([human cystic fibrosis CFTR sequence],35,8,1,1,1,256,12). (See below also.)

Part of the batch decomposition was as indicated:

```
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>AATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAG
GGTTGAGCGGCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAG
CAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTTTT
CAGCTGGACCAGACCAATTTTGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGACATATACCAAA
TCCCTTCTGTTGATTCTGCTGACAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAGCTGGCTTC
...
TAATTTTTATATTTGAAATATTGACTTTTTATGGCACTAGTATTTTTATGAAATATTATGTTAAAACTGG
GACAGGGGAGAACCTAGGGTGATATTAACCAGGGGCCATGAATCACCTTTTGGTCTGGAGGGAAGCCTT
GGGGCTGATCGAGTTGTTGCCCACAGCTGTATGATTCCCAGCCAGACACAGCCTCTTAGATGCAGTTCTGA
AGAAGATGGTACCACCAGTCTGACTGTTTCCATCAAGGGTACACTGCCTTCTCAACTCCAAACTGAC
TCTTAAGAAGACTGCATTATATTTATTACTGTAAGAAAATATCACTTGTCAATAAAATCCATACATTTGTGT

<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<ACACAAATGTATGGATTTTATTGACAAGTGATATTTTCTTACA
GTAATAAATATAATGCAGTCTTCTTAAGAGTCAGTTTGGAGTTGAGAAGGCAGTGTACCCTTGATG
GAAACAGTCAGACTGGTGGTACCATCTTCTTCAGAACTGCATCTAAGAGGCTGTGTCTGGCTGGGAATCAT
ACAGCTGTGGGCAACAACTCGATCAGCCCCAAGGCTTCCCTCCAGACCAAAAGGTGATTCATGGCCC
CTGGTTAATATCACCCTAGGTTCTCCCCTGTCCCAGTTTTAACATAATATTTCATAAAAATACTAGTGCCATA
...
GGGCATTAATGAGTTTAGGATTTTTCTTTGAAGCCAGCTCTCTATCCCATTCTCTTTCCAATTTTTCAG
ATAGATTGTCAGCAGAATCAACAGAAGGGATTTGGTATATGTCTGACAATTCCAGGCGCTGTCTGTATCCTT
TCCTCAAAATTGGTCTGGTCCAGCTGAAAAAAAGTTTGGAGACAACGCTGGCCTTTTCCAGAGGCGA
CCTCTGCATGGTCTCTCGGGCGCTGGGGTCCCTGCTAGGGCCGTCTGGGCTCAAGCTCCTAATGCCAAAGACC
TACTACTCTGGGTGCCTGCCGCTCAACCCTTTTTCTCTGACCTGCTGTGATGTCATTTGCTTCCAATT
```

Finally, the code:

```
function y=dnacompx(D,n,m,a,b,T,time,nb)

% Uses simulated annealing to stab at good partitions
% D is a linear ssDNA strand (e.g., ACGTACTA...)
% ...T is the initial 1/beta (i.e., temperature):
% we cool by beta=ln(t+1)/T; time is um, time and
% a and b are constants for the formal energy function

% Calls rc.m for reverse-complementation. For example,
% rc(‘ACGGATCTA’) gives the string TAGATCCGT.

% DIFFERS from dnacomp (an older, buggy version) in that it
% can move multiple bases per SA step. The number of bases per
% step is given as a uniformly random variable on [1,nb/t].

N=length(D);
% We’ll keep the number of batch ssDNAs fixed...
B=2*ceil(N/(2*n));
% Length of notional hairpinned DNA (see below) will be n*B:
% sticky ends therefore 2*n*B/2-N long...

% We can ignore sticky ends of the dsDNA complex;
% the concomitant batch ssDNAs will therefore not really
% be considered here. (Hence the negatives)
% Form a NOTIONAL hairpin:
E=[repmat(‘>‘,[1 n*B-N]),D,repmat(‘<‘,[1 n*B-N]),rc(D)];
% convert E back to a char array (doubles are bad for us)

% Proto-placeholder...
F=[(1:n*B-N),(n*B-N+1:n*B),fliplr(n*B+1:n*B+n*B-N),fliplr(n*B-N+1:n*B)];

% Now divvy up the complex into baseline batch ssDNAs: establish placeholders
for k=1:B
        batch{k}=E(2*n*(k-1)+1:2*n*k);
        place{k}=F(2*n*(k-1)+1:2*n*k);
end

% here’s the sum of l*MSH(l)...(Evaluate n as a viable upper limit IN GENERAL. Here it’s OK.)

mis=zeros(B,B,n-m+1);
for l=m:n
        for j=1:B
                ss1=batch{j};
                for k=1:B
                        ss2=batch{k};
                        init1=ss1(1:l);
                        init2=ss2(1:l);

                        rcinit2=rc(init2);

                        term1=ss1(end-l+1:end);
                        term2=ss2(end-l+1:end);

                        rcterm2=rc(term2);

                        % Is there a 3’-sticky ended mishybridization of length l?
                        % Only if init1 = rcinit2.
                        % We could augment this part to check init1 v term2 etc, but not now...
                        mis3=1;
                        if strcmp(init1,rcinit2)~=1
                                mis3=0;
                        end
                        mis5=1;
                        % What about 5’?
                        if strcmp(term1,rcterm2)~=1
                                mis5=0;
                        end
                        mis(j,k,l-m+1)=mis3+mis5;
                end
        end
end

% portion of the formal energy coming from mishybridizations
% (which is all there is for the baseline configuration)
% is of the form Hzeta = sum((m:n).*squeeze(sum(sum(mis)))’).
% Don’t worry about the proper hybridizations nominally
% referred to as mishybridizations; their contribution
% is invariant (constant zero point is all we care about).

        Hzeta=sum((m:n).*squeeze(sum(sum(mis)))’);
        energy(1)=a*Hzeta;
```

```
% The thermal noise will be chunks of bases going from one batch
% ssDNA to another: at each timestep, a single batch ssDNA is
% selected uniformly at random, and then one of its ends is
% selected with probability 1/2 (unless we're supposed to be
% at the end of the dsDNA construct), and then it yoinks a chunk
% of bases (the size of the chunk is proportional to the effective
% temperature: i.e., prop. to 1/t...see annealing schedule below)
% from its neighbor (we make sure the neighbor really does
% lose the base that our fella gains). We don't allow moves that
% leave sticky ends of fewer than m bases...

% We also try to be (a little bit) clever: only the two altered batch
% ssDNAs are reevaluated for mishybridizations. This means, however,
% that we've gotta keep mishybridization counts in memory (see above):
% for each batch ssDNA, we have an array like [mis_m(batch{k}) ... ].
% This is tough too, since we've gotta remember what an MSH means
% (i.e., we've gotta keep our placeholders straight). So OK here goes:

% initial batch is the baseline decomposition:
initialbatch=batch;
initialplace=place;

for t=1:time
        numbases=ceil(nb*rand/t);

        yoinker=ceil(B*rand);
        if yoinker==1
                yoinkee=2;
                ind=2;
        elseif yoinker==B/2
                yoinkee=B/2-1;
                ind=1;
        elseif yoinker==B/2+1
                yoinkee=B/2+2;
                ind=2;
        elseif yoinker==B
                yoinkee=B-1;
                ind=1;
        else
                ind=ceil(2*rand);
                nei=[yoinker-1,yoinker+1];
                yoinkee=nei(ind);
        end

        batchyr=batch{yoinker};
        placeyr=place{yoinker};
        batchye=batch{yoinkee};
        placeye=place{yoinkee};

        if ind==1 % yoinker comes after yoinkee
                bases=batchye(end-numbases+1:end);
                placebases=placeye(end-numbases+1:end);
                batchyr=[bases, batchyr]; % It is better to receive
                placeyr=[placebases, placeyr];
                batchye(end-numbases+1:end)=[];
                placeye(end-numbases+1:end)=[];
        else % yoinker comes before yoinkee
                bases=batchye(1:numbases);
                placebases=placeye(1:numbases);
                batchyr=[batchyr, bases]; % It is better to receive
                placeyr=[placeyr, placebases];
                batchye(1:numbases)=[];
                placeye(1:numbases)=[];
        end

        batch{yoinker}=batchyr;
        place{yoinker}=placeyr;
        batch{yoinkee}=batchye;
        place{yoinkee}=placeye;

                % Now if 1≤k≤B/2, batch{k} is supposed to join up with batch{B-k+1} and batch{B-k+2}.
                % Similarly, if B/2+1≤k≤B, batch{k} joins up with...batch{B-k+1} and batch{B-k+2}.
                % Just watch out for the ends!
                % We gotta make sure, therefore, that batch{k} overlaps both of these
                % other batch ssDNAs by at least m bases...
                % ...but this happens iff the placeholders have m or more common values...
                if yoinker==1
                        if length(intersect(place{2},place{B})) < m
                                % just gotta check place{B} and place{2} when yoinker=1
                                batch=initialbatch;
                                place=initialplace;
                        else
                        end
                elseif yoinker==B/2+1
                        if length(intersect(place{B/2},place{B/2+2})) < m
                                % just gotta check place{B/2} and place{B/2+2} when yoinker=B/2+1
                                batch=initialbatch;
                                place=initialplace;
                        else
                        end
                elseif yoinkee==1
                        if length(intersect(place{1},place{B})) < m % etc
                                batch=initialbatch;
                                place=initialplace;
                        else
                        end
                elseif yoinkee==B/2+1
                        if length(intersect(place{B/2},place{B/2+1})) < m % etc
                                batch=initialbatch;
                                place=initialplace;
                        else
                        end
                elseif length(intersect(place{yoinkee},place{B-yoinkee+1})) < m
                        % Now by presumption we had OK sticky ends last time, so just check yoinkee
                        batch=initialbatch;
                        place=initialplace;
                elseif length(intersect(place{yoinkee},place{B-yoinkee+2})) < m % etc
                        batch=initialbatch;
                        place=initialplace;
                else %%%% This else added recently.
                         %%%% We could put the rest of the for loop in here but it’d be obscure %%%%
                end

        % Now tweak entries in mis:
        batchyr=batch{yoinker};
        batchye=batch{yoinkee};
        for l=m:n
                inityr=batchyr(1:l);
                termyr=batchyr(end-l+1:end);
                initye=batchye(1:l);
                termye=batchye(end-l+1:end);

                % check yoinker, (esp. [here meaning only]) for 3’ sticky ends (hence init);
                % and check yoinkee, (esp. [here meaning only]) for 5’ (hence term)
                for j=1:B
                        ss=batch{j};

                        initss=ss(1:l);
                        rcinitss=rc(initss);
                        termss=ss(end-l+1:end);
                        rctermss=rc(termss);

                        mis3er=1;
                        if strcmp(inityr,rcinitss)~=1
                                mis3er=0;
                        end
                        mis3ee=1;
                        if strcmp(initye,rcinitss)~=1
                                mis3ee=0;
                        end
                        mis5er=1;
                        if strcmp(termyr,rctermss)~=1
                                mis5er=0;
                        end
                        mis5ee=1;
                        if strcmp(termye,rctermss)~=1
                                mis5ee=0;
                        end
                        mis(j,yoinker,l-m+1)=mis3er+mis5er;
                        mis(yoinker,j,l-m+1)=mis3er+mis5er;
                        mis(j,yoinkee,l-m+1)=mis3ee+mis5ee;
                        mis(yoinkee,j,l-m+1)=mis3ee+mis5ee;
                end
        end

        Hzeta=sum((m:n).*squeeze(sum(sum(mis)))’);
        % Now compute the lengths of the batch ssDNAs (part of the energy biz);
        for j=1:B
                L(j)=length(batch{j});
        end

        newbatch=batch;
        newplace=place;
        % the energy is
        energy(t+1)=a*Hzeta+b*var(L);
```

```
        % if the change in energy is positive, make the move according to a B-G likelihood;
        % or, if the change in energy is negative, make the move automatically
        if rand(1) > exp(-(energy(t+1)-energy(t))*log(t+1)/(log(2.71828)*T)) % reject the move
        energy(t+1)=energy(t);
        else
                initialbatch=newbatch;
                initialplace=newplace;
        end

end

y=batch;

figure;plot(energy)
```

APPENDIX D

NOTES ON ROTATIONAL DIFFUSION

APPENDIX D
NOTES ON ROTATIONAL DIFFUSION

Here we look at the nucleation process in the context of rotational diffusion of ~□–4 nt segments. As a shorthand in this discussion we take "nuclear satisfaction" to mean satisfaction of the angular constraint on nucleation for which rotational diffusion comes into play. We fix one of these ssDNA segments with orientation given by the standard basis vector $\mathbf{e}_3$ and denote the (antisense or reverse) orientation of the other as a function of time by $\mathbf{x}_t$. Let □ denote the standard azimuthal/longitudinal coordinate, and let □ denote the standard polar/colatitudinal coordinate. Put $\square_{hyb} := \cos^{-1}(1 - \square_{hyb})$ and $\square_t := \cos^{-1}(\mathbf{e}_3 \cdot \mathbf{x}_t)$. We might assume that $\square_0 \sim U[S^2]$: that is, that $\square_0$ is a random variable with uniform density on the unit sphere. However, without loss of generality, we assume that $\mathbf{x}_0$ lies in the plane $\square = 0$. Finally, we put

$$\square_0 := \{(\square, \square) \square S^2 : \cos \square \geq 1 - \square_{hyb}\}.$$

Now the quantity we want to get a handle on is

$$\Pr(\max_t \cos \square_t \geq 1 \text{-} \square_{hyb}) = \Pr(\cos \min_t \square_t \geq 1 \text{-} \square_{hyb}) = \Pr(\min_t \square_t \leq \square_{hyb}).$$

The way to get there from here is via the rotational diffusion or heat equation, which has the same form as the translational diffusion equation (aka the heat equation), but using the Laplacian on the sphere S^2:

$$\square_{\mathbf{R}^n} = \partial_{rr} + \frac{n \square 1}{r} \partial_r + \frac{1}{r^2} \square_{S^{n \square 1}} :$$

$$\square_{\mathbf{R}^3} = \underbrace{\frac{1}{r^2} \partial_r \left(r^2 \partial_r \right)}_{= \partial_{rr} + \frac{2}{r} \partial_r} + \underbrace{\frac{1}{r^2 \sin \square} \partial_\square \left(\sin \square \cdot \partial_\square \right) + \frac{1}{r^2 \sin^2 \square} \partial_{\square\square}}_{= \frac{1}{r^2} \square_{S^2}} .$$

The well-known connection of the heat/diffusion equation with the theory of random walks is useful to us: we will use the solution to a heat equation to tackle the problem. The formal attack on the initial value problem goes through just as for the diffusion equation on $\mathbf{R}^n$.[128] In that case one uses Fourier analysis to transform differentiation into

[128] Taylor, M., *Partial Differential Equations. Basic Theory*, Springer, New York (1996).

multiplication, yielding an ordinary differential equation for the Fourier transform of the solution, which is given by a momentum-space Gaussian times the Fourier transform of the initial value function.[129] The inverse Fourier transform turns this multiplication into convolution, and if the initial value problem is

$$\partial_t p(\mathbf{x},t) = \Delta_{\mathbf{R}^n} p(\mathbf{x},t); \quad p(\mathbf{x},0) := f(\mathbf{x}) \ ,$$

then the (unique/nice) solution is given by the convolution of the initial condition by the heat kernel:

$$\frac{e^{-|\cdot|^2/4t}}{(4\pi t)^{n/2}} * f \ .$$

It turns out that the appropriate way to generalize this approach is by using the spectral theory (i.e., eigenstuff) of the Laplacian and the formal identity for the heat kernel (which we will just write and not try to explain)[130]

$$e^{t\Delta}\delta(\cdot) = \frac{e^{-|\cdot|^2/4t}}{(4\pi t)^{n/2}} \Rightarrow \ e^{t\Delta} f = p.$$

The convolution bit follows from this. Following these lines for the spherical diffusion equation, we recall the s*pherical harmonics* (i.e., the eigenfunctions of the spherical Laplacian):[131]

$$\Delta_{S^2} Y_{l,m} = -\lambda_{l,m} Y_{l,m} = -l(l+1) Y_{l,m} : \quad l \in \mathbf{Z}; \quad m \in \{-l, -l+1, \ldots, l-1, l\}$$

$$Y_{l,m}(\theta,\phi) := (-1)^m \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^{(m)}(\cos\theta) e^{im\phi},$$

$$\text{where} \quad P_l^{(m)}(x) := \left(1-x^2\right)^{m/2} D_x^m P_l(x) \quad \text{and} \quad P_l(x) := \frac{1}{2^l \, l!} D_x^l \left(x^2 - 1\right)^l .$$

Using this we obtain a solution to the initial value problem by the superposition principle (for the diffusion equation is linear):

[129] Dym, H., and McKean, H. P., *Fourier Series and Integrals*, Academic Press, San Diego (1972).

[130] See, e.g., Rosenberg, S., *The Laplacian on a Riemannian Manifold*, Cambridge, Cambridge (1997).

[131] Arfken, G., *Mathematical Methods for Physicists*. Academic Press, Orlando (1985); see also Egorov, Yu. V., and Shubin, M. A., *Foundations of the Classical Theory of Partial Differential Equations*, Springer, Berlin (1998).

$$\partial_t p(\theta,\phi,t) = \Delta_{S^2} p(\theta,\phi,t); \quad p(\theta,\phi,0) \overset{\text{(wlog)}}{=} \delta_{x_0} := \delta(0,\phi_0);$$

$$p \equiv p_{x_0}(\theta,\phi,t) = \sum_{l,m} e^{-t\cdot\lambda_{l,m}} \left\langle \delta_{x_0}, Y_{l,m} \right\rangle Y_{l,m} \overset{\text{formally}}{=} e^{t\Delta_{S^2}} \delta_{x_0}.$$

This gives us an idea of how to proceed in earnest. Enforcing a Dirichlet (absorbing) boundary condition at $\theta = \theta_{hyb}$ will ensure that the resultant diffusion equation will give us what we want.[132] Indeed, the probability of nuclear satisfaction will be one minus the probability of its negation, or one minus the integral of the density over $M := S^2 \setminus \Omega_0$. Depending on x_0, the probability of nuclear satisfaction should either begin (and stay) at unity, or begin at zero and increase over time. Developing the framework to characterize this is our ultimate goal here.

Separation of variables gives us

$$\left[\frac{1}{\sin\theta}\partial_\theta\left(\sin\theta\cdot\partial_\theta\right) + \frac{1}{\sin^2\theta}\partial_{\phi\phi}\right] h(\theta)e^{im\phi} = -\lambda h(\theta)e^{im\phi}$$

$$\Rightarrow \; h''(\theta) + \frac{\cos\theta}{\sin\theta}h'(\theta) + \left[-\frac{m^2}{\sin^2\theta} + \lambda\right]h(\theta) = 0.$$

$$b.c. \quad h \in C^2\left(\left[\theta_{hyb},\pi\right]\right); \quad h\left(\theta = \theta_{hyb}\right) := 0; \quad h'\left(\theta = \theta_{hyb}\right) \neq 0.$$

(The last boundary condition is not really as vague as it seems, since the equation is linear and we will normalize its solutions anyway.) The functions h can be obtained numerically and used to get an orthonormal basis of eigenfunctions $\{\mathcal{Y}_{l,m}\}$ of the Laplacian Δ_M on M. By construction, the eigenvalues of Δ_M are identical to those of the spherical Laplacian; that is, we have that the *truncated spherical harmonics* satisfy

$$\Delta_M \mathcal{Y}_{l,m} = -\lambda_{l,m}\mathcal{Y}_{l,m} = -l(l+1)\mathcal{Y}_{l,m} : \quad l \in \mathbf{Z}; \quad m \in \{-l,-l+1,\ldots,l-1,l\};$$

$$\mathcal{Y}_{l,m} \equiv C_{l,m}h_{l,m}(\theta)e^{im\phi}; \quad \left\|h_{l,m}\right\|_2 \equiv \left(\int_{\theta_{hyb}}^{\pi} \left|h_{l,m}(\theta)\right|^2 \sin\theta \, d\theta\right)^{1/2} := 1.$$

$$\int_M \mathcal{Y}_{l,m}^* \mathcal{Y}_{l,m} \, dS^2 = \left|C_{l,m}\right|^2 \int_M h_{l,m}^* h_{l,m} \, dS^2 = 2\pi\left|C_{l,m}\right|^2 \Rightarrow \left|C_{l,m}\right| = \frac{1}{\sqrt{2\pi}}.$$

As before, we have that

$$\partial_t p(\theta,\phi,t) = \Delta_M p(\theta,\phi,t); \quad p(\theta,\phi,0) \overset{\text{(wlog)}}{=} \delta_{x_0};$$

$$p \equiv p_{x_0}(\theta,\phi,t) = \sum_{l,m} e^{-t\cdot l(l+1)} \left\langle \delta_{x_0}, \mathcal{Y}_{l,m} \right\rangle \mathcal{Y}_{l,m} \overset{\text{formally}}{=} e^{t\Delta_M} \delta_{x_0}.$$

[132] Heagy, J., unpublished notes (2001).

We recall the unit measures

$$\Box_{M}\frac{dM}{2\Box\left(2\Box\Box_{hyb}\right)}=1; \quad \Box_{\Box_{hyb}}^{\Box}\frac{\sin\Box d\Box}{2\Box\Box_{hyb}}=1.$$

With these firmly in hand, we define the average *heat content*: [133]

$$\bar{q}(t):=\frac{1}{2\Box\Box_{hyb}}\Box_{\Box_{hyb}}^{\Box}\Box\frac{1}{2\Box\left(2\Box\Box_{hyb}\right)}\Box_{M}pdM\Box\sin\Box_0 d\Box_0.$$

This quantity will give us a handle on the probability of nuclear satisfaction. But most (i.e., those $\mathscr{Y}_{l,m}$ with $m\neq0$) of the truncated spherical harmonics trivially integrate to zero, and it is easy to show that the average heat content is given by

$$\bar{q}(t)=\frac{1}{2\Box\left(2\Box\Box_{hyb}\right)^2}\Box_{l}e^{\Box t\Box l(l+1)}\Box\Box_{\Box_{hyb}}^{\Box}\Box h_{l,0}(\Box)\sin\Box d\Box\Box^{2}.$$

Thus we see that the decay of the heat content is determined by the principal eigenvalues of the Laplacian.[134] (However, this observation alone would not have told us anything.) For completeness we list computed values of the squared integrals from the expression above:

l	–1, 0	–2, 1	–3, 2	–4, 3	–5, 4	–6, 5
$[\Box h_{l,0}(\Box)\ \sin\Box \mathrm{d}\Box]^2$	1.5241	0.8759	0.0048	0.0657	0.0096	0.0107

Numerical weights for the average heat content.

We note in passing that self-consistent numerical analysis of the heat equation on the sphere via eigenmethods is nontrivial.[135] Below we depict the fractional average heat content.

[133] McDonald, P., and Meyers, R., “Dirichlet spectrum and heat content,” math.SP/0205098 (2002), and Desjardins, S., “Asymptotic expansions for the heat content,” *Pacific J. Math.* 183, 279 (1998).

[134] Burchard, A., and Schmuckenschläger, M., “Comparison theorems for exit times,” *Geometric and Functional Analysis* 11, 651 (2001).

[135] For such a technique, see Le Gia, Q. T., et al., “Solving parabolic PDEs on unit spheres by collocation,” Preprint (2001).

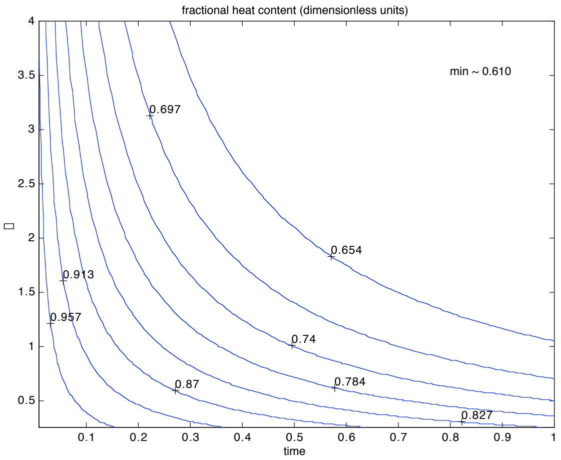


Figure D-1. Fractional average heat content as a function of time and diffusion coefficient

The MATLAB code we used is included below:

```
function hc=heatcontent(n, D, tmax, steps);
% Calculates heat content for a truncated sphere
% w/ delta ic w/ phi = phi and theta = 0 (wlog).
% D is the diffusion coefficient. Time runs in
% steps number of steps from time=0 to 1

a=acos(.905); % a=0.4394
phispan=[a 3.14]; % don’t use pi for the numerics...
time=linspace(0,tmax,steps);
phi=linspace(0,pi,4*2^n);
marker=ceil( (a/pi) *4*2^n);
        % no error handling for phi less than a
dphi=max(diff(phi));

ylm=zeros(steps,4*2^n,4*2^n);
p=zeros(steps,4*2^n,4*2^n);

for l=-n:n
        yo=[0;1];               % By linearity/L^2 normalization it doesn’t
                                % matter what we set the derivative ic to
    options = odeset(‘RelTol’,1e-5,’AbsTol’,1e-8);
        [t,y]=ode45(‘myodef’,phispan,yo,[],l,0,a);
                  % get something associated-Legendreish
        yy=interp1(t,y(:,1),phi);
                % interpolate to linearly spaced points
                nanyy=find(isnan(yy));
                nyy=find(~isnan(yy));
                yy(nanyy)=max(yy(nyy));
                        % kill any ODE solver NaNs

                yy(1:marker)=0;

        l2y=sum((yy.^2).*sin(phi)*dphi);
        yy=yy/sqrt(l2y); % normalize to unit norm in L^2:
        iy2=(sum(yy.*sin(phi)*dphi))^2 % read these off

        for t=1:steps
                f(l+n+1,t)=exp(-time(t)*D*l*(l+1))*iy2;
        end

end

hc=squeeze(sum(f,1));

%- - - - - - - - - - - - -

function ydot = myodef(phi,y,flag,l,m,a)
        % called to generate truncated spherical harmonics
```

```
if isempty(flag)
        ydot = [y(2);( (m.*csc(phi)).^2 - l*(l+1) ).*y(1) - cot(phi).*y(2)]; % use for phi
else
        ‘error’
end
```

APPENDIX E

NOTES ON SINGLE BASE MISMATCHES

APPENDIX E
NOTES ON SINGLE BASE MISMATCHES

In a series of papers,[136] the SantaLucia NN model[137] of Watson-Crick paired DNA thermodynamics was successfully extended to incorporate mismatched base pairs. Several features of interest besides the thermodynamic parameters were noted: firstly, the idea of testing the NN model by using (what turn out to be members of) de Bruijn equivalence classes of DNA sequences;[138] secondly, that mismatches at positions near the ends of duplexes can induce fraying at the ends and that this is a stabilizing effect (in particular, terminal, penultimate, and "penpenultimate" mismatches can typically all be treated as terminal mismatches, whereas mismatches any further within the duplex structure contribute independently of their exact position, and in line with the generalized NN formalism).

The overall trend (in order of decreasing stability) is

$$\mathrm{G{\cdot}C > A{\cdot}T > G{\cdot}G > G{\cdot}T \sim G{\cdot}A > A{\cdot}C^{+} > T{\cdot}T \sim A{\cdot}A \sim C{\cdot}C^{+} > T{\cdot}C \geq A{\cdot}C \geq C{\cdot}C.}$$

The first six of these are (broadly speaking) stabilizing, and the last six destabilizing. In general, it is probably fair to say that a generic single base mismatch will have a negligible net contribution to duplex stability, whereas properly matched bases will enhance duplex stability. The picture rapidly deteriorates past the NN framework: one recent paper explicitly remarks that "we do not understand the basic physics of [single mismatch] hybridization."[139]

[136] Allawi, H., and SantaLucia, J., Jr., "Thermodynamics and NMR of Internal G·T Mismatches in DNA," *Biochemistry* **36**, 10581 (1997); "Nearest Neighbor Thermodynamic Parameters for Internal G·A Mismatches in DNA," *Biochemistry* **37**, 2170 (1998); "Nearest Neighbor Thermodynamics of Internal A·C Mismatches in DNA: Sequence Dependence and pH Effects," *Biochemistry* **37**, 9435 (1998); "Thermodynamics of internal C·T mismatches in DNA," *Nuc. Acids. Res.* **26**, 2694 (1998); and "Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A·A, C·C, G·G, and T·T Mismatches," *Biochemistry* **38**, 3468 (1998).

[137] SantaLucia, "A unified view."

[138] Cited in particular as Kierzek, R., et al., *Biochemistry* **25**, 7840 (1986), and Sugimoto, N., et al., *Biochemistry* **34**, 11211 (1995).

[139] Naef, F., et al., "DNA hybridization to mismatched templates: a chip study," *Phys. Rev. E* **65**, 040902 (2002).

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE October 2002	2. REPORT TYPE Final	3. DATES COVERED *(From–To)* January 2002—August 2002

4. TITLE AND SUBTITLE Towards the Batch Synthesis of Long DNA	5a. CONTRACT NUMBER DAS W01 98 C 0067
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Steven Huntsman	5d. PROJECT NUMBER
	5e. TASK NUMBER DA-2-1955
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 4850 Mark Center Drive Alexandria, VA 22311-1882	8. PERFORMING ORGANIZATION REPORT NUMBER IDA Document D-2752
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) DARPA/ATO 3701 N. Fairfax Drive Arlington, VA 22203-1714	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release: distribution unlimited (18 December 2002).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Production-scale synthesis of long DNAs with specific base sequences by conventional methods is generally infeasible due to error rates. On the other hand, sticky-ended double-stranded (ds) DNA complexes can be joined into a single dsDNA complex with an enzyme called DNA ligase. This idea offers a way to indirectly synthesize very long (>10kb) DNAs—by hybridization of a batch of single-stranded (ss) DNAs followed by ligation—with predetermined base sequences. We analyze such synthesis protocols. In the natural course of this development, we take a reasonably self-contained tour through much of what is known about the physics of DNA hybridization as it relates to the issues here. This paper is therefore a primer on DNA self-assembly, using the staggered ligation model for DNA synthesis as the working example.

15. SUBJECT TERMS

DNA synthesis, DNA hybridization, minus shift hybridization

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Emil Martinsek
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.	SAR	85	19b. TELEPHONE NUMBER *(include area code)* 703-526-4776