Soft Matter

PAPER



Cite this: *Soft Matter*, 2021, **17**, 10736

# Building up DNA, bit by bit: a simple description of chain assembly

R. Foffi, †[a] F. Sciortino, [a] J. M. Tavares [bc] and P. I. C. Teixeira *[bc]

Received 3rd August 2021,
Accepted 4th November 2021

DOI: 10.1039/d1sm01130h

rsc.li/soft-matter-journal

We simulate the assembly of DNA copolymers from two types of short duplexes (short double strands with a single-stranded overhang at each end), as described by the oxDNA model. We find that the statistics of chain lengths can be well reproduced by a simple theory that treats the association of particles into ideal (*i.e.*, non-interacting) clusters as a reversible chemical reaction. The reaction constants can be predicted either from SantaLucia's theory or from Wertheim's thermodynamic perturbation theory of association for spherical patchy particles. Our results suggest that theories incorporating very limited molecular detail may be useful for predicting the broad equilibrium features of copolymerisation.

## I. Introduction

The assembly of multifunctional units into linear or branched architectures is a key ingredient of copolymerisation. In turn, the properties of copolymers depend crucially on how these units are arranged, as in alternating, random or block copolymers.[1] Examples are manifold, and we mention just a few: the stacking transition of single-strand DNA;[2] the nature of the de-mixing instabilities in both coil–coil[3] and coil–rod[4] polymer blends undergoing polycondensation reactions; the ability of urethane–urea elastomers to exhibit strain-induced periodic textures;[5] the self-healing nature of poly(methyl methacrylate)/*n*-butyl acrylate over a narrow range of compositions;[6] and the association of DNA duplexes by stacking interactions.[7–9]

The actual sequence of building blocks on individual copolymer molecules is experimentally inaccessible and must be inferred indirectly, *e.g.*, from the comonomer ratios, or from details of the synthesis method employed. It would be most desirable to have a predictive theory for this information that might be used as input to theories for macroscopic properties, *e.g.*, elastic or rheological. Such a theory could be readily validated by computer simulations of copolymerisation. From a more practical point of view, this approach would have the added bonus of enabling the 'reverse engineering' of desired polymers: by elucidating which properties of building blocks (*e.g.*, size, interaction energies) produce which architectures, polymer synthesis could be more effectively directed towards specific outcomes.

In this paper we propose a theory of association that is able to predict the frequency of any given sequence of monomers in the aggregates formed, *i.e.*, the statistics of block lengths. This extends our earlier work on self-assembly in patchy colloids,[10] in which we concerned ourselves with phase separation and percolation as determined by functionalities and patch interaction strengths, but not with the internal structures of aggregates. The theory tries to be as economical as possible and so forgoes most microscopic detail and treats the bonding of polymerising units as reversible chemical reactions, governed by reaction constants. This is then applied to the assembly of DNA chains from two types of monomers, where each monomer is a DNA duplex, consisting of a double-stranded core with a short single strand at each end. The reaction constants are taken either from SantaLucia's treatment of the nearest-neighbour model for DNA,[11,12] or from Wertheim's thermodynamic perturbation theory (TPT) of association for spherical patchy particles.[13,14] Besides its biological relevance, this model has two advantages with respect to testing our theory: first, it is the simplest case of linear aggregation, hence no percolation can occur; second, by tuning the model parameters we can adjust the interaction energies between different bifunctional building blocks and thus generate a large variety of chain architectures, allowing for a more thorough and detailed comparison between simulation and theoretical predictions. Several theoretical approaches to equilibrium polymerisation[15,16] and to the statistics of blocks in copolymer chains[17] exist.

[a] *Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Moro 5, I-00185 Rome, Italy. E-mail: francesco.sciortino@uniroma1.it*
[b] *ISEL – Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Rua Conselheiro Emídio Navarro 1, 1959-007 Lisboa, Portugal. E-mail: jose.tavares@isel.pt*
[c] *Centro de Física Teórica e Computacional, Faculdade de Ciências da Universidade de Lisboa, P-1749-016 Lisboa, Portugal. E-mail: piteixeira@fc.ul.pt*
† Present address: Institute of Environmental Engineering, Department of Civil, Environmental and Geomatic Engineering, ETH Zürich, Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland. E-mail: rfoffi@ethz.ch

Isodesmic equilibrium polymerisation ideas have been applied to living polymers formed by DNA blunt-end duplexes, aggregating under the effect of stacking interactions.[8,9] To the best of our knowledge none has made a connection with a simple yet microscopic model, nor allowed for two types of monomer interacting *via* more than two bonding energy scales (or reaction constants), as we do. Moreover, experimental work on DNA constructs has demonstrated their value as a model system for investigating unconventional aggregation phenomena of patchy colloidal particles.[18,19]

This paper is organised as follows: in Section II we expound our theory and show how it is related to Wertheim's TPT. We also discuss briefly how it can be generalised to more than two types of monomer. Then in Section III we describe the microscopic DNA model used in our simulations. Results are presented in Section IV, and conclusions drawn in Section V.

# II. Theory

### A. A minimal description of linear aggregation

Our system consists of a binary mixture of $N_A$ particles of species $A$, each decorated with two bonding sites ('patches') of type $A$, and $N_B$ particles of species $B$, each decorated with two bonding sites of type $B$, in a volume $V$. The total number of particles is thus $N = N_A + N_B$, and their mole fractions are $x_A \equiv N_A/N$ and $x_B \equiv N_B/N = 1 - x_A$.

Each of the two sites can participate in at most one bond to another site, so a given particle (of species $A$ or $B$) can bond to at most two other particles (linear aggregation). We shall regard the formation of an $\alpha\beta$ bond as a reversible chemical reaction between an unreacted site of type $\alpha$ and one of type $\beta$ ($\alpha, \beta = A, B$). If we now assume that both sites and bonds behave as ideal gases, then the equilibrium constant for this reaction is given by[20]

$$K_{\alpha\beta} = \frac{P^*_{\alpha\beta}}{P^*_\alpha P^*_\beta} \equiv \exp\left(-\beta\Delta G_{\alpha\beta}\right), \tag{1}$$

where $P^*_k$ is the ratio of the partial pressure of sites or bonds of type $k$ ($k = \alpha$, $\beta$ or $\alpha\beta$) at equilibrium to some reference pressure $P_{\text{ref}}$, $\beta = 1/k_{\text{B}}T$ with $k_{\text{B}}$ the Boltzmann constant and $T$ the temperature, and $\Delta G_{\alpha\beta}$ is the change in Gibbs free energy on forming an $\alpha\beta$ bond. Because we are assuming that all sites and bonds $k$ behave as ideal gases, we have

$$P^*_{\alpha\beta} = \frac{N_{\alpha\beta} k_{\text{B}} T}{P_{\text{ref}} V}, \tag{2}$$

$$P^*_\alpha = \frac{M_{\bar\alpha} k_{\text{B}} T}{P_{\text{ref}} V}, \tag{3}$$

where $N_{\alpha\beta}$ is the number of $\alpha\beta$ bonds and $M_{\bar\alpha}$ is the number of unreacted sites of type $\alpha$. Using eqn (2) and (3), eqn (1) can be rewritten as

$$K_{\alpha\beta} = \frac{N_{\alpha\beta}}{M_{\bar\alpha} M_{\bar\beta}} \frac{P_{\text{ref}} V}{k_{\text{B}} T}. \tag{4}$$

If we further assume that $P_{\text{ref}}$ is the pressure of the system when no 'chemical reaction' has occurred (*i.e.*, when there are no aggregates but the same total number of sites is present) then $P_{\text{ref}} = 2Nk_{\text{B}}T/V$ (recall there are two sites per particle), whence

$$K_{\alpha\beta} = 2N \frac{N_{\alpha\beta}}{M_{\bar\alpha} M_{\bar\beta}}. \tag{5}$$

We note that this result can be given a microscopic interpretation: in terms of the partition functions of species $\alpha$, $Q_\alpha$, and of $\alpha\beta$ dimers $Q_{\alpha\beta}$, the condition for chemical equilibrium is[21,22]

$$\frac{N_{\alpha\beta}}{M_{\bar\alpha} M_{\bar\beta}} = \frac{Q_{\alpha\beta}}{Q_\alpha Q_\beta} = \frac{K_{\alpha\beta}}{2N} = \frac{\exp\left(-\beta\Delta G_{\alpha\beta}\right)}{2N}, \tag{6}$$

*i.e.*, the partition functions of particles and bonds are subsumed in the equilibrium constants, for which we need some prescription. We shall come back to this point later.

From the above we can now derive the laws of mass action for the three reactions: $\bar A + \bar A \rightleftharpoons AA$, $\bar B + \bar B \rightleftharpoons BB$ and $\bar A + \bar B \rightleftharpoons AB$, where the overline denotes an unreacted site. Using the constraints

$$2N_A = 2N_{AA} + N_{AB} + M_{\bar A}, \tag{7}$$

$$2N_B = 2N_{BB} + N_{AB} + M_{\bar B}, \tag{8}$$

and the usual definitions for the fractions of unbonded $A$ and $B$ sites,

$$X_\alpha = \frac{M_{\bar\alpha}}{2N_\alpha}, \tag{9}$$

we arrive at the following laws of mass action:

$$1 - X_A = 2x_A K_{AA} {X_A}^2 + 2x_B K_{AB} X_A X_B, \tag{10}$$

$$1 - X_B = 2x_B K_{BB} {X_B}^2 + 2x_A K_{AB} X_A X_B, \tag{11}$$

Note that in the above expressions we are assuming that no rings are formed. In the thermodynamic limit ($N \to \infty$, $V \to \infty$), $p_\alpha = 1 - X_\alpha$ is the probability that a site of type $\alpha$ has reacted (*i.e.*, is bonded to another site). Noting that the total number of sites of type $\alpha$ that participate in $\alpha\beta$ bonds is $(1 + \delta_{\alpha\beta})N_{\alpha\beta}$ (with $\delta_{\alpha\beta}$ the Kronecker delta) and the total number of sites of type $\alpha$ is $2N_\alpha$, then the probability of bonding a site of type $\alpha$ to one of type $\beta$ is.[23]

$$p_{\alpha\beta} = \left(1 + \delta_{\alpha\beta}\right) \frac{N_{\alpha\beta}}{2N_\alpha}. \tag{12}$$

(Notice that, although $N_{\alpha\beta} = N_{\beta\alpha}$ always holds, if $N_\alpha \neq N_\beta$ then $p_{\alpha\beta} \neq p_{\beta\alpha}$.) From these probabilities, which can be obtained by solving the laws of mass action, eqn (10) and (11), we can compute a number of interesting structural quantities. In particular, we shall derive the statistics of 'blocks', *i.e.*, the probabilities of assembling sequences of contiguous identical bonds ('blocks') of length $\ell_{\alpha\beta}$, defined as the number of $\alpha\beta$ bonds in the sequence (block).

Let us consider first blocks of identical particles. To make an $A$ block of length $\ell_{AA}$, one starts with a particle of species $A$ that has one $A$ site not bonded to another $A$ site: there are $N_A(1 - p_{AA})$ such particles (notice that an $A$ site that is not bonded to another $A$ site could be either unbonded to any site, or bonded to a $B$ site). Then one needs to make $\ell_{AA}$ bonds, each

with probability $p_{AA}$, which gives a factor of $p_{AA}^{\ell_{AA}}$. Finally, the block ends with an $A$ site that is not bonded to any other $A$ site, hence another factor of $(1 - p_{AA})$. It follows that the number of $A$ blocks of length $\ell_{AA}$ is

$$n(\ell_{AA}) = N_A p_{AA}^{\ell_{AA}} (1 - p_{AA})^2. \tag{13}$$

Likewise, the number of $B$ blocks of length $\ell_{BB}$ is

$$n(\ell_{BB}) = N_B p_{BB}^{\ell_{BB}} (1 - p_{BB})^2. \tag{14}$$

Now consider $AB$ blocks of block size $\ell_{AB}$, *i.e.*, alternating sequences of $A$ and $B$ particles. Two cases must be distinguished: blocks with either $A$ or $B$ sites at both ends have even $\ell_{AB}$, whereas blocks with an $A$ site at one end and a $B$ site at the other end have odd $\ell_{AB}$. The number of $AB$ blocks with odd $\ell_{AB}$ is

$$\begin{aligned} n(\ell_{AB}, \text{odd}) &= N_A (1 - p_{AB}) p_{AB}^{(\ell_{AB}+1)/2} p_{BA}^{(\ell_{AB}-1)/2} (1 - p_{BA}) \\ &\quad + N_B (1 - p_{BA}) p_{BA}^{(\ell_{AB}+1)/2} p_{AB}^{(\ell_{AB}-1)/2} (1 - p_{AB}). \end{aligned} \tag{15}$$

The first term on the right-hand side (rhs) of this equation is derived as follows: if an $AB$ block starts with an $A$ particle, then one of its $A$ sites is not bonded to a $B$ site, which gives the factor $N_A(1 - p_{AB})$; then, there follow $(\ell_{AB} + 1)/2$ $AB$ bonds alternating with $(\ell_{AB} - 1)/2$ $BA$ bonds, which gives the factor $p_{AB}^{(\ell_{AB}+1)/2} p_{BA}^{(\ell_{AB}+1)/2}$; finally, the block ends with a $B$ site not connected to an $A$ site, which gives the factor $(1 - p_{BA})$. The second term is obtained by just exchanging $A$ and $B$ in the preceding argument: it corresponds to counting the $AB$ and $BA$ bonds for an $AB$ block that starts with a $B$ particle and ends with an $A$ particle. Eqn (15) can be simplified using the definitions of $p_{AB}$ and $p_{BA}$, with the result

$$n(\ell_{AB}, \text{odd}) = 2 N_A^{1/2} N_B^{1/2} (1 - p_{AB})(1 - p_{BA}) p_{AB}^{\ell_{AB}/2} p_{BA}^{\ell_{AB}/2}. \tag{16}$$

By the same reasoning, the number of $AB$ blocks with even $\ell_{AB}$ is

$$\begin{aligned} n(\ell_{AB}, \text{even}) &= N_A (1 - p_{AB}) (p_{AB} p_{BA})^{\ell_{AB}/2} (1 - p_{AB}) \\ &\quad + N_B (1 - p_{BA}) (p_{AB} p_{BA})^{\ell_{AB}/2} (1 - p_{BA}). \end{aligned} \tag{17}$$

We reiterate that, in order to fulfil symmetry under sequence inversion, *i.e.*, the requirement that the identity of a block should be independent of the order in which its sequence is read, eqn (15) and (17) include a contribution from both: $AB\ldots AB$ and $BA\ldots BA$ sequences, for $\ell_{AB}$ odd; and from $AB\ldots BA$ and $BA\ldots AB$ sequences, for $\ell_{AB}$ even.

The mean block lengths can now be calculated. For $AA$ and $BB$ blocks, we have

$$\langle \ell_{AA} \rangle = \frac{1}{1 - p_{AA}}, \tag{18}$$

$$\langle \ell_{BB} \rangle = \frac{1}{1 - p_{BB}}. \tag{19}$$

Notice that these expressions are general, in the sense that they apply even when $N_A \neq N_B$. The mean length of $AB$ blocks is

$$\begin{aligned} \langle \ell_{AB} \rangle &= \frac{\sum_{i=1}^{\infty} [(2i-1) n(2i-1,\ \text{odd}) + 2i n(2i,\ \text{even})]}{\sum_{i=1}^{\infty} [n(2i-1, \text{odd}) + n(2i, \text{even})]} \\ &= \frac{1}{1 - p_{AB} p_{BA}} \frac{1 + p_{AB} p_{BA} + F(p_{AB}, p_{BA})}{1 + \frac{1}{2} F(p_{AB}, p_{BA})}, \end{aligned} \tag{20}$$

where

$$F(x, y) = \frac{x(1-y)^2 + y(1-x)^2}{(1-x)(1-y)}. \tag{21}$$

It is readily seen that $\langle \ell_{AA} \rangle$ and $\langle \ell_{BB} \rangle$ are functions of, respectively, $p_{AA}$ and $p_{BB}$ only, whereas $\langle \ell_{AB} \rangle$ is a function of $p_{AB}$ and $p_{BA}$ only. It follows that results do not depend on whether these blocks are isolated or part of longer chains. Further note that, by construction, the minimum length of an $\alpha\beta$ block is 1, when $p_{\alpha\beta} \to 0$: this is because, in this limit, $N_{\alpha\beta} \to 0$ and both the number of $\alpha\beta$ blocks and their length $\to 0$, but the ratio of these two quantities $\to 1$.

If $N_A = N_B$, in which case $p_{AB} = p_{BA}$, eqn (20) simplifies to

$$\langle \ell_{AB} \rangle = \frac{1}{1 - p_{AB}}. \tag{22}$$

Furthermore, eqn (15) and (17) become identical, leading to

$$n(\ell_{AB}) = 2 N_A (1 - p_{AB})^2 p_{AB}^{\ell_{AB}}. \tag{23}$$

### B. Extension to multiple species

The theory of linear aggregation of the preceding section can be straightforwardly extended to the case where we have $n$ distinct chemical species $\alpha = \mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_n$, each decorated with two identical bonding sites. If, as before, $N_\alpha$ is the number of particles of species $\alpha$, the total number of particles in the system will be $N = \sum_\alpha N_\alpha$, and their mole fractions $x_\alpha = N_\alpha/N$. Therefore we have a set of $n(n + 1)/2$ coupled chemical reactions:

$$\bar{\alpha} + \bar{\beta} \rightleftharpoons \alpha\beta, \quad \alpha, \beta = \mathrm{S}_1, \ldots, \mathrm{S}_n, \tag{24}$$

subject to the $n$ constraints

$$2N_\alpha = M_{\bar{\alpha}} + 2N_{\alpha\alpha} + \sum_{\beta \neq \alpha} N_{\alpha\beta}, \quad \alpha = \mathrm{S}_1, \ldots, \mathrm{S}_n. \tag{25}$$

In the ideal gas of clusters approximation, the condition of chemical equilibrium between any pair of species $\alpha$ and $\beta$ is given by eqn (6) and similarly the equilibrium constant of each reaction by eqn (1). The $n$ laws of mass action are obtained as

$$1 - X_\alpha = \sum_\beta (1 + \delta_{\alpha\beta}) x_\beta X_\alpha X_\beta K_{\alpha\beta}, \tag{26}$$

where $X_\alpha = M_{\bar{\alpha}}/2N_\alpha$. Accordingly, the results for the block statistics could be extended to account for the assembly of more complex architectures.

As mentioned above, we require some prescription for finding the equilibrium constants $K_{\alpha\beta}$, and thence the probabilities $p_\alpha$ and $p_{\alpha\beta}$. For DNA, perhaps the simplest way is to compute them using the second equality in eqn (1) with $\Delta G_{\alpha\beta}$ given by SantaLucia for the nearest-neighbour model.[11,12] Alternatively, one can map a microscopic, off-lattice theory of self-assembly onto the above minimal description. This we do in the next section; a similar approach has been proposed by Reinhardt and Frenkel.[22]

### C. Wertheim's thermodynamic perturbation theory

Wertheim's thermodynamic perturbation theory (TPT) is a microscopic theory for the self-assembly of particles interacting *via* strong, short-ranged attractions.[13,14] It has found novel applications in the description of the phase behaviour of patchy colloidal particles.[24–26] As in ref. 27, we rather bluntly approximate the solution of DNA sequences as a binary mixture of $N_A$ and $N_B$ equisized hard spheres (HSs) of diameter $\sigma$, contained in a volume $V$; the total number density is thus $\rho = (N_A + N_B)/V \equiv N/V$. The solvent is not explicitly considered. Both species are divalent: particles of species $A$ are decorated with two attractive sites, or 'patches', of type $A$, and particles of type $B$ are decorated with two patches of type $B$: these represent the single strands at the end of the DNA sequences. We make the usual assumption that the patches are distributed over the spheres' surfaces in such a way that each patch can only take part in at most one bond, which is a short-ranged attractive interaction between two patches, as is appropriate for DNA bases. We take these inter-patch attractions to be square wells of depth $\varepsilon_{\alpha\beta}$ and range chosen such that the volume available to an $\alpha\beta$ bond is $v_b^{\alpha\beta}$ ($\alpha$, $\beta$ = $A$, $B$).

Following[28,29] the bonding probabilities $p_{\alpha\beta}$ are given by

$$p_{AA} = 2\xi x_A \Delta_{AA} X_A{}^2, \tag{27}$$

$$p_{BB} = 2\xi x_B \Delta_{BB} X_B{}^2, \tag{28}$$

$$p_{AB} = 2\xi x_B \Delta_{AB} X_A X_B, \tag{29}$$

$$p_{BA} = 2\xi x_A \Delta_{AB} X_A X_B, \tag{30}$$

where $\xi = \rho v_{HS}$, with $v_{HS} = \pi\sigma^3/6$ the volume of a HS, is the (total) packing fraction, $x_\alpha = N_\alpha/N$ is the mole fraction of component $\alpha$ ($\alpha$ = $A$, $B$), and $\Delta_{\alpha\beta}$ are the bond partition functions. In the low-density, strong-interaction limit, which as we shall see is appropriate to our simulations, we have

$$\Delta_{\alpha\beta} \approx \frac{v_b^{\alpha\beta}}{v_{HS}} \exp(\beta\varepsilon_{\alpha\beta}), \tag{31}$$

In eqn (27)–(30), $X_\alpha$, the fractions of unbonded sites of type $\alpha = A, B$, are given by the following laws of mass action:

$$1 - X_A = 2\xi x_B \Delta_{AB} X_A X_B + 2\xi x_A \Delta_{AA} X_A{}^2, \tag{32}$$

$$1 - X_B = 2\xi x_B \Delta_{BB} X_B{}^2 + 2\xi x_A \Delta_{AB} X_A X_B. \tag{33}$$

Comparing eqn (10) and (32), (11) and (33), and further noting that the fraction of sites of type $\alpha$ is the same as the mole fraction of particles of species $\alpha$, we conclude that the reaction constants in our minimal description are very simply related to the bond partition functions:

$$K_{\alpha\beta} = \xi \Delta_{\alpha\beta}. \tag{34}$$

How can we now relate the parameters of Wertheim's TPT to those of SantaLucia? Start by noting that, from eqn (1), (31) and (34), we have

$$\beta\Delta G_{\alpha\beta} = -\log(\xi\Delta_{\alpha\beta}) = -\log(\rho v_b^{\alpha\beta}) - \beta\varepsilon_{\alpha\beta}. \tag{35}$$

Recalling that $\Delta G_{\alpha\beta} = \Delta H_{\alpha\beta} - T\Delta S_{\alpha\beta}$, where $H_{\alpha\beta}$ and $S_{\alpha\beta}$ are, respectively, the enthalpy and entropy of an $\alpha\beta$ bond, we can identify

$$\Delta H_{\alpha\beta} = -\varepsilon_{\alpha\beta}, \quad \Delta S_{\alpha\beta} = \log(\rho v_b^{\alpha\beta}), \tag{36}$$

*i.e.*, the change in enthalpy is related to the bond strength, and the change in entropy to the volume available to the bond. In actual systems it is often the case that the pressure and volume vary very little, and the change in enthalpy can thus be equated to a change in internal energy.[30]

Wertheim's theory thus provides an inexpensive alternative description of the self-assembly statistics in block copolymer systems, on the basis of very simple model – patchy particles – whose interaction parameters can be readily related to hybridisation enthalpies and entropies. It has exactly the same structure as the minimal theory of linear aggregation of the preceding section, so the same accuracy can be expected.

## III. Model for DNA

We describe DNA using the oxDNA model.[31,32] This is a coarse-grained model with implicit solvent, which has been shown to capture the basic thermodynamics, as well as the essential structural properties, of DNA. It consists of rigid nucleotides, interacting *via* pairwise interactions that comprise non-linear elastic, stacking, cross-stacking, excluded-volume and hydrogen bonding contributions; see ref. 32 for details.

The system we investigate is a binary mixture of DNA nanoparticles.[33,34] Each 'particle' is made up of a complementary double helix core **X**, decorated with identical single strands, of types $a$ or $b$, at either end, *i.e.*, the two particle species are $A = a\mathbf{X}a$ and $B = b\mathbf{X}b$. The binding enthalpy of the $a$–$a$, $b$–$b$ and $a$–$b$ pairings can be tuned *via* a judicious choice of the sticky-end binding sequences. In what follows, we shall take $A(B)$ to refer interchangeably to either a single strand of type $a(b)$ or a particle of species $A(B)$. To break the symmetry of the model, we want to favour $BB$ bonds over $AA$ or $AB$ bonds. We thus need to find pairs of short self-complementary DNA sequences $A$ and $B$ that can bind to each other. Preliminary calculations using Wertheim's theory suggest rich behaviour is realised if their hybridisation enthalpies satisfy the conditions

$$\frac{\Delta H_{AB}}{\Delta H_{AA}} \simeq 1, \quad \frac{\Delta H_{BB}}{\Delta H_{AA}} \simeq 1.25. \tag{37}$$

No condition is set on their hybridisation entropies. The nearest-neighbour model of SantaLucia[11] allows us to compute

Table 1 Relative hybridisation enthalpies and entropies of DNA single strands used in our simulations, calculated according to ref. 11 and 12

| $A$ | $B$ | $\frac{\Delta H_{BB}}{\Delta H_{AA}}$ | $\frac{\Delta H_{AB}}{\Delta H_{AA}}$ | $\frac{\Delta S_{BB}}{\Delta S_{AA}}$ | $\frac{\Delta S_{AB}}{\Delta S_{AA}}$ |
|---|---|---|---|---|---|
| CGATCG | TCGATCGA | 1.27 | 1.09 | 1.24 | 1.09 |

these quantities for any given sequence: the values of $\Delta H$ and $\Delta S$ can be assumed to be temperature-independent at least in a 'narrow' range around $T = 310$ K. In order to fulfil conditions (37), $B$ is chosen to be the same as $A$ with an extra pair of complementary nucleotides, one at each end, so that $B$ will bind to $B$ with greater energy than $A$ to $A$, but will still be able to bind to $A$ with roughly the same energy as $A$ to $A$. The dangling ends in an $AB$ bond will actually provide an extra contribution, usually increasing stability of $AB$ with respect to $AA$, as evaluated in ref. 12; without this stabilizing mechanism there would be no reason for $A$ and $B$ to form this mixed complex instead of only $AA$ and $BB$. This also implies that the condition $\frac{\Delta H_{AB}}{\Delta H_{AA}} \simeq 1$ is not really attainable, so we should just look for this ratio to be as close as possible to unity. Two sequences that come close to these target values are $A$ = CGATCG and $B$ = TCGATCGA, whose hybridisation enthalpies and entropies are reported in Table 1. Note that $A$ can bind to $A$ (being 'palindromic') with six bases, $B$ can bind to $B$ with eight bases, and $A$ can bind to $B$ with six bases.

Cartoons of the particles and of a short chain are shown in Fig. 1. The bottom panel displays the full base sequences of particles $A$ and $B$, using the same colour code as in parts (a)–(c). The double-stranded core **X** is 21 base long, significantly longer

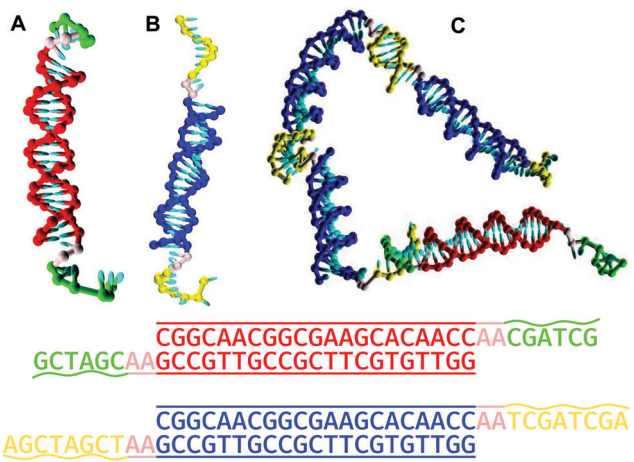


Fig. 1 Cartoons of particles of (A) species $A$-and (B) species $B$, made up of a core of double-stranded DNA and two identical single strands at the ends acting as bonding sites. On both particle species, two extra bases (adenine, coloured light pink) on each side of the core have been included to decouple the free ends from the core. (C) A chain of four particles, one of species $A$ and three of species $B$. This $BBBA$ chain is composed of one $BB$ block of length 2 and one $AB$ block of length 1. The bottom panel displays the full base sequences of particles $A$ and $B$, using the same colour code as in parts (A)–(C).

than the six and eight bases composing the sticky sequences. The hybridisation temperature of short DNA sequences strongly depends on the the number of bases, which guarantees that the particles' cores are stable well above the temperature at which the sticky sequences start to bind. The $AA$ pairs are spacers used to decouple core and free ends, thus allowing the latter freely to explore space without mechanical hindrance from the former. These extra bases are commonly added in DNA-nanotechnology designs.[18,19] If they were not present, binding would result in a continuous double helix, In order for such a helix to form, the relative orientations of two interacting particles would need to be very precisely aligned, which would significantly affect the binding probability. The insertion of one or two 'inactive' bases allows greater freedom in the relative orientation of pairs of bonded particles. This orientational freedom is also expected to yield better agreement with SantaLucia's estimate of the binding free energy, which takes into account only the sticky sequences.

The DNA double helix has a large persistence length, $\sim$100 base pairs in relevant conditions,[35] so the double-stranded core of the particles, which is only 21 bases long, is quite stiff. We do not expect its length (within any reasonable range), or particular base sequence, to have any effect on results. Note also that the large persistence length typical of the DNA double helix disfavours the formation of ring aggregates (as opposed to linear ones).

## IV. Results

We ran constant-volume molecular dynamics (MD) simulations using the oxDNA open source code for coarse-grained simulations of DNA.[31] Using the tools provided by the oxDNA package, we generated the coordinates of $N_A = N_B = 200$ bifunctional particles of types a**X**a and b**X**b, for a total of 800 sticky ends on 400 particles. These 400 particles were then randomly placed in the simulation box to generate a concentration of single strands of each species in solution of $c \approx 392$ μM, a typical experimental value. We mimicked (recall there is no explicit solvent) an aqueous solution of salt concentration $[Na^+] = 0.4$ M at temperatures in the range $[40, 58]$ °C, where binding is expected to take place and equilibration can be achieved with the available computational resources. Individual simulations required from $\sim$200 to $\sim$350 microseconds, depending on temperature. At the lowest investigated temperature (40 °C), the equations of motion were integrated for $6 \times 10^{10}$ steps (with a time step of 0.006 ps), which required eight months on one Tesla P100 GPU. Equilibration was assessed by monitoring the time evolution of the potential energy and of the bonding probabilities The latter are very sensitive indicators, since equilibration requires multiple bond formation and bond breaking events at each binding site. We have noticed that during equilibration $AB$ bonds form first (since $A$ patches are more attracted to $B$ patches than to other $A$ patches), followed by a relaxation stage in which some $AB$ bonds will break to reach their equilibrium value. The equilibration timescale of

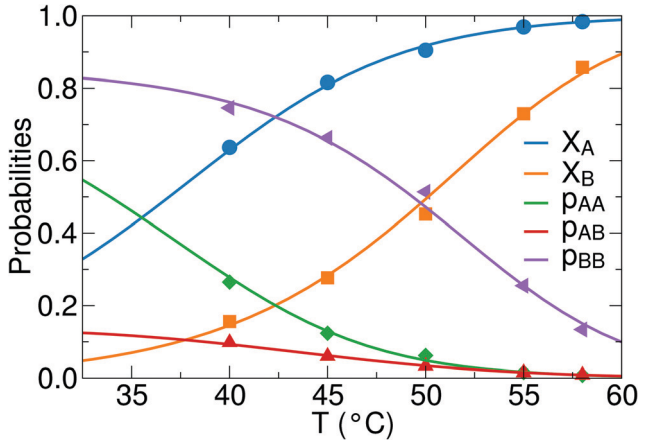


Fig. 2 Probabilities $p_{AA}$, $p_{AB} = p_{BA}$ and $p_{BB}$ of the different types of bonds between patches, and probability $X_A(X_B)$ that a $A(B)$ site is unbonded, *vs.* temperature, from simulations (symbols) and theory (solid lines). The solid lines were obtained by solving eqn (40) numerically using $\Delta H_{ij}$ and $\Delta S_{ij}$ in Table 2.

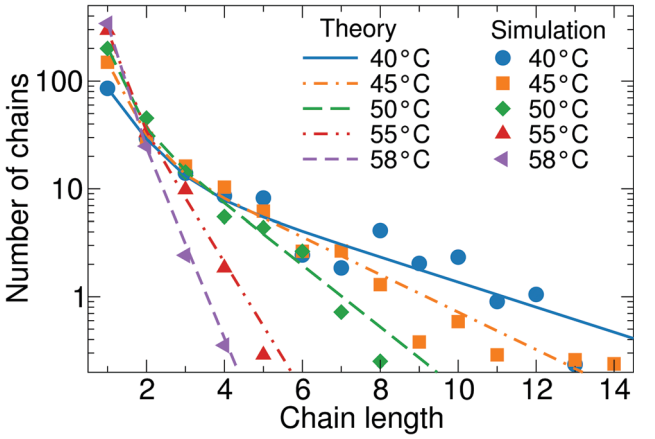


Fig. 3 Comparison between expected and simulated chain length distributions at all temperatures studied.

the system is thus affected by bond lifetimes, which increase exponentially with decreasing temperature. We consider a simulation to have reached equilibrium when the total number of bonds and all the bonding probabilities oscillate around a constant value without showing any evident drift.

From the simulation, we estimate the number of $A$ patches unbonded, bonded with $A$ ($N_{AA}$) and bonded with $B$ ($N_{AB}$). Similar quantities are calculated for $B$ patches. Using the relations previously introduced we can then estimate $p_{AA} \equiv N_{AA}/N_A$, $p_{BB} \equiv \boldsymbol{N_{BB}}/N_B$ and $p_{AB} \equiv N_{AB}/N_A$, as well as $X_A$ and $X_B$. The resulting probabilities are shown in Fig. 2 as symbols; the curves are the theory predictions computed as explained below.

From $p_{AA}$, $p_{BB}$ and $p_{AB}$ we can then predict the chain length distribution $P(n)$, where $n$ is the number of particles in the chain. Let $s$ be a sequence of bonded monomers ($A$ or $B$ particles); the probability of observing that sequence is

$$P(s) = (1-p_A)^{m_{\bar{A}}}(1-p_B)^{m_{\bar{B}}} p_{AB}^{n_{AB}} p_{BA}^{n_{BA}} p_{AA}^{n_{AA}} p_{BB}^{n_{BB}}, \qquad (38)$$

where $m_{\bar{\alpha}}$ is the number of free ends of type $\alpha$ and $n_{\alpha\beta}$ is the number of sites of type $\alpha$ bonded to sites of type $\beta$ (with the constraints $m_{\bar{A}} + m_{\bar{B}} = 2$, $n_{AA} + n_{AB} + n_{BB} = n - 1$). Then the probability $P(n)$ of observing a cluster (literally a linear chain) of length $n$) is found by summing $P(s)$ over all possible sequences of $n$ monomers:

$$P(n) = \sum_{s=1}^{2^n} P(s). \qquad (39)$$

Fig. 3 compares theoretical predictions and simulation data for the chain length distribution at all temperatures studied.

To evaluate $\Delta G_{\alpha\beta}$ we first combine eqn (1), (27)–(29) and (34) to obtain, for $x_A = x_B = 0.5$,

$$p_{\alpha\beta} = X_\alpha X_\beta \exp(-\beta \Delta G_{\alpha\beta}). \qquad (40)$$

Eqn (40) can now be solved using the data in Fig. 2, at each $T$. Results are plotted in Fig. 4. In all cases, a linear dependence of $\Delta G_{\alpha\beta}$ on $T$ is observed. The slope and intercept provide the best-fit values for $\Delta H_{\alpha\beta}$ and $\Delta S_{\alpha\beta}$, which can then be used backwards to predict the bond probabilities. These predictions are shown as solid lines in Fig. 2.

Table 2 compares the best-fit values of $\Delta H_{\alpha\beta}$ and $\Delta S_{\alpha\beta}$ with the predictions of the SantaLucia model. Though not excellent, agreement is reasonable, considering that the oxDNA model is a parametrisation based on SantaLucia estimates for the melting temperature. Specifically, oxDNA predictions for the melting temperatures have been found to deviate on average 1.4 °C from those of SantaLucia.[36]

Finally, Fig. 5 plots the mean block lengths $\langle \ell_{\alpha\beta} \rangle$, *vs.* either temperature (Fig. 5a) or bond probability $p_{\alpha\beta}$ (Fig. 5b). For our choice of parameters, both *AA* and *AB* blocks are very short and the simulation data are very noisy; agreement between theory and simulation is encouraging for the longer *BB* blocks.

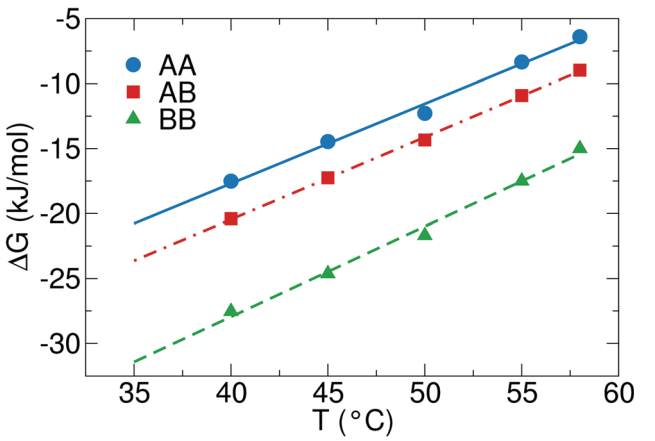


Fig. 4 Binding free energy of the three complexes *AA*, *BB* and *AB*. Symbols are the $\Delta G_{\alpha\beta}$ values obtained by solving eqn (40) using the $p_{\alpha,\beta}$ from simulation. Solid lines are the linear fits $\Delta G_{\alpha\beta} = \Delta H_{\alpha\beta} - T\Delta S_{\alpha\beta}$, where $\Delta H_{\alpha\beta}$ and $\Delta S_{\alpha\beta}$ are taken as temperature-independent fit parameters (values reported in Table 2).

Table 2 Comparison of best-fit values of $\Delta H_{\alpha\beta}$ and $\Delta S_{\alpha\beta}$ with the predictions of the SantaLucia model. The uncertainties are the standard errors of the fits

| | *AA* | *BB* | *AB* |
|---|---|---|---|
| SantaLucia | | | |
| $\Delta S$ (J mol$^{-1}$ K$^{-1}$) | −126.2 | −156.8 | −137.1 |
| $\Delta H$ (kJ mol$^{-1}$) | −44.6 | −56.6 | −48.6 |
| Simulation | | | |
| $\Delta S$ (J mol$^{-1}$ K$^{-1}$) | −147 ± 8 | −166 ± 9 | −151 ± 2 |
| $\Delta H$ (kJ mol$^{-1}$) | −50 ± 3 | −59 ± 3 | −52 ± 1 |

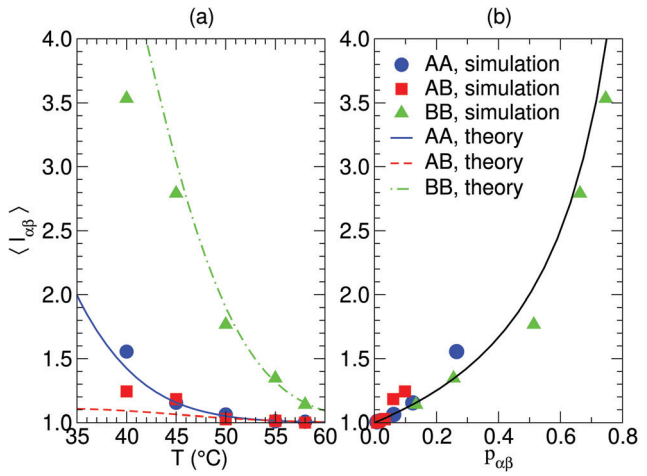


Fig. 5 Mean block lengths $\langle \ell_{\alpha\beta} \rangle$ *vs.* (a) temperature; (b) bond probability $p_{\alpha\beta}$. In (b) the black solid line is the theory prediction $1/(1 - p_{\alpha\beta})$ for all block types, *e.g.*, eqn (18)–(20) when $p_{AB} = p_{BA}$.

# V. Conclusions

We have proposed a minimal theoretical framework for the assembly of linear block copolymers. This makes very few assumptions on the nature of the monomers, namely: (i) assembly is assimilated to reversible chemical reactions between short-ranged bonding sites; (ii) each site can participate in at most one bond; and (iii) the overall concentration is low enough that sites and bonds behave as ideal gases. The theory requires as inputs the reaction constants for the polymerisation reactions. Importantly, these can be derived from theories that incorporate only very limited detail of the actual molecular processes.

The theory was tested against simulation results for the assembly of DNA chains from two types of short duplexes, as described by the oxDNA model, using reaction constants calculated from SantaLucia's theory of a lattice model of DNA. This was found to reproduce the equilibrium block size distributions, mean block sizes, and fractions of unreacted monomers fairly well.

The theory is easily generalised to any number of associating particle species in any proportion. In our view it has the potential to become a useful tool to predict or reverse-engineer the architectures of multi-block copolymers or polycolloids, and even provide some insight into the kinetics of association.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

R. F. and F. S. acknowledge support from MIUR PRIN 2017 (Project 2017Z55KCW). J. M. T. and P. I. C. T. acknowledge financial support from the Portuguese Foundation for Science and Technology (FCT) under Contracts no. UIDB/00618/2020 and UIDP/00618/2020.

## References

1 I. W. Hamley, *The Physics of Block Copolymers*, Oxford University Press, Oxford, 1998.
2 W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, 1984.
3 P. I. C. Teixeira, D. J. Read and T. C. B. McLeish, Demixing instability in polymer blends undergoing polycondensation reactions, *Macromolecules*, 2000, **33**, 3871–3878.
4 P. I. C. Teixeira, D. J. Read and T. C. B. McLeish, Demixing instability in coil-rod blends undergoing polycondensation reactions, *J. Chem. Phys.*, 2007, **126**, 074901.
5 M. H. Godinho, J. L. Figueirinhas, P. Brogueira and P. I. C. Teixeira, in *Tuneable micro- and nano-periodic structures in urethane/urea networks, in Biomimetic and Supramolecular Systems*, ed. A. H. Lima, Nova Science Publishers, New York, 2008.
6 M. W. Urban, D. Davydovich, Y. Yang, T. Demir, Y. Zhang and L. Casabianca, Key-and-lock commodity self-healing copolymers, *Science*, 2018, **362**, 220–225.
7 M. Nakata, G. Zanchetta, B. D. Chapman, C. D. Jones, J. O. Cross, R. Pindak, T. Bellini and N. A. Clark, End-to-end stacking and liquid crystal condensation of 6- to 20-base pair DNA duplexes, *Science*, 2007, **318**, 1276–1279.
8 C. De Michele, L. Rovigatti, T. Bellini and F. Sciortino, Self-assembly of short DNA duplexes: from a coarse-grained model to experiments through a theoretical link, *Soft Matter*, 2012, **8**, 8388–8398.
9 K. T. Nguyen, A. Battisti, D. Ancora, F. Sciortino and C. De Michele, Self-assembly of mesogenic bent-core DNA nanoduplexes, *Soft Matter*, 2015, **11**, 2934–2944.
10 P. I. C. Teixeira and J. M. Tavares, Phase behaviour of pure and mixed patchy colloids – theory and simulation, *Curr. Opin. Colloid Interface Sci.*, 2017, **30**, 16–24.
11 J. SantaLucia Jr, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 1460–1465.
12 S. Bommarito, N. Peyret and J. SantaLucia Jr, Thermodynamic parameters for DNA sequences with dangling ends, *Nucleic Acids Res.*, 2000, **28**, 1929–1934.

13 M. S. Wertheim, Fluids with highly directional attractive forces. I. Statistical thermodynamics, *J. Stat. Phys.*, 1984, **35**, 19–34.
14 M. S. Wertheim, Fluids with highly directional attractive forces. II. Thermodynamic perturbation theory and integral equations, *J. Stat. Phys.*, 1984, **35**, 35–47.
15 P. J. Flory, *Principles of Polymer Chemistry*, Cornell University Press, Ithaca, New York, 1953.
16 M. S. Wertheim, Fluids with highly directional attractive forces. IV. Equilibrium polymerization, *J. Stat. Phys.*, 1986, **42**, 477–492.
17 S. Kuchanov, H. Slot and A. Stroeks, Development of a quantitative theory of polycondensation, *Prog. Polym. Sci.*, 2004, **29**, 563–633.
18 F. Bomboi, F. Romano, M. Leo, J. Fernandez-Castanon, R. Cerbino, T. Bellini, F. Bordi, P. Filetici and F. Sciortino, Re-entrant DNA gels, *Nat. Commun.*, 2016, **7**, 13191.
19 F. Bomboi, D. Caprara, J. Fernandez-Castanon and F. Sciortino, Cold-swappable DNA gels, *Nanoscale*, 2019, **11**, 9691–9697.
20 P. W. Atkins and J. de Paula, *Physical Chemistry*, Oxford University Press, Oxford, 10th edn, 2014.
21 T. L. Hill, *Introduction to Statistical Thermodynamics*, Dover, New York, 1986.
22 A. Reinhardt and D. Frenkel, DNA brick self-assembly with an off-lattice potential, *Soft Matter*, 2016, **12**, 6253–6260.
23 J. M. Tavares, P. I. C. Teixeira and M. M. Telo da Gama, Percolation of colloids with distinct interaction sites, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2010, **81**, 010501(R).
24 E. Bianchi, J. Largo, P. Tartaglia, E. Zaccarelli and F. Sciortino, Phase diagram of patchy colloids: towards empty liquids, *Phys. Rev. Lett.*, 2006, **97**, 168301.
25 J. Russo, J. M. Tavares, P. I. C. Teixeira, M. M. Telo da Gama and F. Sciortino, Reentrant Phase Diagram of Network Fluids, *Phys. Rev. Lett.*, 2011, **106**, 085703.
26 D. J. Audus, F. W. Starr and J. F. Douglas, Valence, loop formation and universality in self-assembling patchy particles, *Soft Matter*, 2018, **14**, 1622–1630.
27 E. Locatelli and L. Rovigatti, An accurate estimate of the free energy and phase diagram of all-DNA bulk fluids, *Polymers*, 2018, **10**, 447.
28 D. de las Heras, J. M. Tavares and M. M. Telo da Gama, Phase diagrams of binary mixtures of patchy colloids with distinct numbers and types of patches: The empty fluid regime, *J. Chem. Phys.*, 2011, **134**, 104904.
29 D. de las Heras, J. M. Tavares and M. M. Telo da Gama, Phase diagrams of binary mixtures of patchy colloids with distinct numbers of patches: The network fluid regime, *Soft Matter*, 2011, **7**, 5615–5626.
30 F. Sciortino, E. Bianchi, J. F. Douglas and P. Tartaglia, Self-assembly of patchy particles into polymer chains: A parameter-free comparison between Wertheim theory and Monte Carlo simulation, *J. Chem. Phys.*, 2007, **126**, 194903.
31 T. E. Ouldridge, A. A. Louis and J. P. K. Doye, DNA Nanotweezers Studied with a Coarse-Grained Model of DNA, *Phys. Rev. Lett.*, 2010, **104**, 178101.
32 T. E. Ouldridge, A. A. Louis and J. P. K. Doye, Structural, mechanical and thermodynamic properties of a coarse-grained DNA model, *J. Chem. Phys.*, 2011, **134**, 085101.
33 S. Biffi, R. Cerbino, G. Nava, F. Bomboi, F. Sciortino and T. Bellini, Equilibrium gels of low-valence DNA nanostars: a colloidal model for strong glass formers, *Soft Matter*, 2015, **11**, 3132–3138.
34 J. Fernandez-Castanon, S. Bianchi, F. Saglimbeni, R. Di Leonardo and F. Sciortino, Microrheology of DNA Hydrogels Gelling and Melting On Cooling, *Soft Matter*, 2018, **14**, 6431–6438.
35 S. Guilbaud, L. Salomé, N. Destainville, M. Manghi and C. Tardin, Dependence of DNA persistence length on ionic strength and ion type, *Phys. Rev. Lett.*, 2019, **122**, 028102.
36 P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye and A. A. Louis, Sequence-dependent thermodynamics of a coarse-grained DNA model, *J. Chem. Phys.*, 2012, **137**, 135101.