Available online at www.sciencedirect.com

ScienceDirect

Journal of Theoretical Biology 242 (2006) 713–721

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

Master equation approach to the assembly of viral capsids

T. Keef[a], C. Micheletti[b], R. Twarock[a,c,*]

[a]Department of Mathematics, University of York, York, UK
[b]International School for Advanced Studies (S.I.S.S.A.) and INFM, Via Beirut 2–4, I-34014 Trieste, Italy
[c]Department of Biology, University of York, York YO10 5DD, UK

Received 23 December 2005; received in revised form 6 April 2006; accepted 26 April 2006
Available online 16 May 2006

Abstract

The distribution of inequivalent geometries occurring during self-assembly of the major capsid protein in thermodynamic equilibrium is determined based on a master equation approach. These results are implemented to characterize the assembly of SV40 virus and to obtain information on the putative pathways controlling the progressive build-up of the SV40 capsid. The experimental testability of the predictions is assessed and an analysis of the geometries of the assembly intermediates on the dominant pathways is used to identify targets for anti-viral drug design.

Keywords: Viral capsid assembly; Master equation; SV40 virus

1. Introduction

Manipulating the assembly of viral capsids is one way of interfering with the viral replication cycle and hence a possible avenue for anti-viral drug design. Despite its importance (Casjens, 1985; Caspar and Klug, 1962; Zandi et al., 2004) the theory of viral capsid assembly is still in its infancy. A first model for the self-assembly of a small plant virus was pioneered by Zlotnick (1994), exploring the assembly of a dodecagonal shape by a cascade of single order reactions. It has since been extended to more involved scenarios (Endres and Zlotnick, 2002; Zlotnick et al., 1999, 2000), including a study of the energy landscape underlying assembly (Endres et al., 2005), that is similar to approaches in protein folding (Brooks et al., 1998) or the energy landscape description of association reactions (Wales, 1987, 1996; Wolynes, 1996). These results have been used to investigate the possibility of inhibiting assembly via an anti-viral drug in the case of Herpes Virus (Zlotnick et al., 2002). Related approaches include molecular dynamics studies of viral capsid assembly (Rapaport et al., 1999, 2004), and a molecular dynamics-like formalism that is implemented in connection with a "local rules" mechanism that regulates capsid assembly (Berger et al., 1994; Schwartz et al., 1998, 2000). Moreover, a stochastic model of icosahedral capsid growth, that is based on results on the agglomeration processes in fullerenes (Kerner, 1994; Kerner et al., 1992; Kerner and dos Santos, 1988), has been developed in Kerner (2005).

A characteristic feature of our approach is the fact that it introduces different association energies for different bonds, and at the same time models the local bonding structure in a way that takes differences in the bonds around the pentamers into account, and in this way is in particular adapted to the situation of the Papovaviridae (Modis et al., 2002; Rayment et al., 1982), which are linked to cancer and are hence of particular interest for the public health sector. For example, the (pseudo-) $T = 7$ capsids in this family are known to be composed of pentamers, i.e. clusters of five protein subunits, that attain two different conformations in the capsid that are distinguished by their local bonding structure. This requires a mathematical representation of these building blocks that takes the differences in the local bonding environments into account. The tiling approach for the description of viral capsids

*Corresponding author. Department of Mathematics, University of York, York, UK. Tel.: +44 1904 434160; fax: +44 1904 433071.
E-mail addresses: tk506@york.ac.uk (T. Keef), michelet@sissa.it (C. Micheletti), rt507@york.ac.uk (R. Twarock).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.04.023

(Twarock, 2004, 2005a, b) provides an appropriate mathematical framework for this. It encodes the locations of proteins and inter-subunit bonds in terms of tilings, i.e. tessellations that represent the surface structure of the capsids. Since the vertex atlas of these tilings, i.e. the collection of all distinct local configurations around vertices in the tiling, encodes the different types of pentamers and their bonding structures, it provides appropriate building blocks for the construction of assembly models. An assembly model for (pseudo-) $T=7$ capsids in the family of Papovaviridae has been introduced along these lines in Keef et al. (2005). In this reference, graphs, called *assembly graphs*, have been determined that encode all energetically preferred pathways of assembly, and it has been analysed how the structures of these graphs change in dependence on the association constants. Moreover, the concentrations of the statistically dominant assembly intermediates (i.e. inequivalent shapes at various stages of capsid construction that are located on all pathways) have been computed.

For applications to anti-viral drug design, it is important to control not only the statistically dominant assembly intermediates, but also the concentrations of all other assembly intermediates, and, based on this, to determine the most probable assembly pathway(s). This issue is addressed in this paper, where we adopt a master equation approach for the computation of the concentrations of all assembly intermediates in the assembly graph and use this information to determine the putative pathways controlling the progressive build-up of the capsid. In particular, in Section 2 we introduce the master equation approach as a tool for the computation of the concentrations of the assembly intermediates for arbitrary viruses from a thermodynamical point of view. In Section 3 we apply this formalism to SV40 virus and determine the equilibrium concentrations of the various assembly intermediates. We discuss how this information can be used to determine the dominant pathway of assembly, and show that the dominant pathways have intermediates with a characteristic structure that may potentially be exploited in the framework of anti-viral drug design. In the final section we summarize our results and assess the implications for other families of viruses.

2. The master equation approach

The formalism presented in this section allows one to compute the probability distribution of the inequivalent configurations (also called species or assembly intermediates) that appear during self-assembly of the major capsid protein of a virus in thermodynamic equilibrium. Assume that the different species are indexed from 1 to N, where species 1 corresponds to the fundamental building block of the capsid, N to the final capsid, and every other assembly intermediate, i, is formed by n_i copies of building block 1. As customary, we consider capsid assembly as a sequence of low order reactions. Consistently with the experimental evidence for the assembly process of several viral capsids, we assume that the formation of the capsid occurs from the attachment or detachment of single fundamental units to or from the partially formed capsids. A notable exception to this simple building scheme is provided by CCMV (Johnson et al., 2005), where pentamers of dimers appear as an intermediate step towards capsid assembly.

From a phenomenological point of view, the equilibrium thermodynamics of the process is described through second-order association constants. Indicating with $[a]$ and $[b]$ the concentrations of two species whose number of constitutive building blocks is, respectively, n_a and $n_b = n_a + 1$, we obtain their association constants as

$$K_{b,a} = \frac{[b]_{eq}}{[a]_{eq}[1]_{eq}}, \tag{1}$$

where $[1]_{eq}$ denotes the concentration of the fundamental building block and the subscripts are used to stress that the concentrations pertain to the equilibrium (stationary) state. This phenomenological relationship can be related to the fundamental entropic and energetic aspects of the association process through the following factorization (as in Zlotnick, 1994; Keef et al., 2005):

$$K_{b,a} = \frac{1}{c_0} S_1 S_{ba} \mathrm{e}^{-\Delta G(b,a)/RT}, \tag{2}$$

where S_1 denotes the geometric degeneracy of the fundamental "incoming" subunit, $S_{ba} := O(b)/O(a)$ corresponds to the ratio of the orders of the discrete rotational symmetry groups of the two species $O(b)$ and $O(a)$, $\Delta G(b,a)$ is the free-energy difference associated to the bonds formed by the incoming building block, and R and T denote the gas constant ($R = 1.987\,\mathrm{cal\,K^{-1}\,mol^{-1}}$) and, respectively, temperature (chosen as room temperature $T = 298\,\mathrm{K}$). The quantity c_0, having the dimension of a concentration, can be put in unique correspondence with the total concentration of elementary blocks present in the system, as will be shown below.

The hierarchy of association constants of the various pairs of intermediate species differing by one fundamental building block can be used in recursive schemes for obtaining the equilibrium probabilities of any species. In fact, by combining Eqs. (1) and (2) and introducing the adimensional quantity $\widetilde{[1]}_{eq} := [1]_{eq}/c_0$ we obtain the fundamental relationship:

$$\frac{[b]_{eq}}{[a]_{eq}} = S_1 S_{ba} \widetilde{[1]}_{eq} \mathrm{e}^{-\Delta G(b,a)/RT} \tag{3}$$

which, used recursively, yields a formal expression of the equilibrium concentration of a generic species, $[i]$, $(i \neq 1)$,

$$[i]_{eq} = S_1^{n_i-1} \frac{O(1)}{O(i)} \mathrm{e}^{-\Delta G(i,1)/RT} [1]_{eq} \widetilde{[1]}_{eq}^{\,n_i-1}, \tag{4}$$

where n_i is the number of fundamental building blocks in species i and $O(1)$ and $O(i)$ refer to the orders of discrete rotational symmetry of subunit 1 and species i, respectively. Note that Eq. (4) depends implicitly on c_0 through $\widetilde{[1]}_{eq}$. This gives the possibility of relating c_0 to the total concentration of fundamental building blocks, $[c^*]$, through the relationship

$$[c^*] = \sum_{i=1}^{N} n_i[i]. \tag{5}$$

Notice that this relationship expresses the law of conservation of the total number of fundamental building blocks present in the system (be they "free" or assembled in intermediate species) and therefore is valid not only in equilibrium. The association constants of Eq. (1) can be used beyond the equilibrium framework since they constitute the starting point for formulating phenomenological kinetic equations apt to capture the time evolution of the system given the initial concentration of the various species. Within a vanishingly small time interval the concentration of a given species $[i]$ (we assume $i \neq 1$) can change only due to the gain or loss of one fundamental building block:

$$\begin{aligned}\frac{\mathrm{d}[i]}{\mathrm{d}t} = {} & \sum_m [m] W_{m,i} + \sum_l [l][1] W_{l,i} \\ & - \sum_m [i][1] W_{i,m} - \sum_l [i] W_{i,l},\end{aligned} \tag{6}$$

where we have omitted the explicit time dependence of the species concentrations. The indices l and m in the sums refer to the species formed by one less, respectively, one more, fundamental building block than species i. Finally, $W_{i,j}$ denotes the *time-independent* rate at which transitions are made from configuration i to configuration j. The dynamics of the system is thus fully described by the set of coupled equations (6) for each $i \neq 1$, supplemented with the conservation law in Eq. (5). The W's must be appropriately related to the association constants, K, to ensure that the correct equilibrium conditions (1) are recovered at large times when the stationary regime is reached (i.e. when $\mathrm{d}[i]/\mathrm{d}t = 0$ for all species i).

Among all possible initial conditions for the above-mentioned kinetic evolution a particularly appealing and interesting one is represented by the case where the only species being present is the one associated with the fundamental building blocks. At this initial time ($t = 0$) the state of the system would then be described as $[1]_{t=0} = [c^*]$ and $[j]_{t=0} = 0$ for all species $j > 1$. The lack of precise experimental characterization of the equilibrium concentrations of the various species in biologically relevant conditions has lead previous theoretical studies to focus on a particularly simple, and biologically relevant, equilibrium situation, namely the one in which the only dominant species are those of the fully formed capsid and of the fundamental building blocks; both species are considered as equiprobable so that $[N]_{eq} = [1]_{eq}$ while, for all other species j, $[j]_{eq} \approx 0$.

Under these assumptions, the concentration of the fundamental species [1] is therefore expected to take on a rather limited range of values. We build on this observation to simplify the description of the assembly process of Eq. (6) through a set of effective first-order reactions. The key ingredient in our analysis is to modify the right-hand side of Eq. (6) so to neglect the time-dependence of the concentration [1] and absorb it in new effective *time-independent* transition rates, W. It is convenient to recast the kinetics obtained by this simplification of Eq. (6) not in terms of the concentration of the ith species but of the equivalent probability of occurrence, $P_i(t) = [i]/\sum_j [j]$. The discrete time evolution of P_i is therefore governed by the following master equation:

$$P_i(t + \Delta t) = P_i(t) + \Delta t \left(\sum_j P_j(t) W_{j,i} - \sum_j P_i(t) W_{i,j} \right), \tag{7}$$

where Δt is the time step of the discretized evolution (assumed to be sufficiently small to justify the linearization of the continuous time evolution). As before, the only non-zero entries in the transition matrix W are those connecting species which differ by the addition/removal of one fundamental building block. The matrix $W = (W_{i,j})$ has to satisfy a number of properties (see Itzykson and Drouffe,):

- $\Delta t \sum_j W_{i,j} = 1$ (normalization condition),
- there exists a finite integer l such that $[W^l]_{i,j} > 0 \; \forall i,j$ (ergodic condition).

It is easy to check that the first condition ensures that $\sum_j P_j(t)$ is constant at all times while the second one requires that any two configurations must be connected by a finite number of transitions. The above conditions are sufficient to ensure the onset of equilibrium at $t \to \infty$, irrespective of the initial condition of the system. From the stationarity of the equilibrium distribution we obtain, from Eq. (7), the generalized balance condition

$$\sum_j (P_j^{eq} W_{j,i} - P_i^{eq} W_{i,j}) = 0. \tag{8}$$

The constraints entailed by the balance equation are not sufficient to identify the matrix W uniquely. We solve this ambiguity by adopting the commonly employed Metropolis criterion within a detailed balance scheme (Itzykson and Drouffe,). The detailed balance condition requires that each term in the sum of Eq. (8) is zero. The Metropolis criterion further specifies the precise form of the W matrix elements. Accordingly, for two different species i and j, which differ by the addition/removal of one fundamental building block (otherwise $W_{ij} = 0$), one has

$$W_{i,j} = \begin{cases} 1 & \text{if } P_i^{eq} < P_j^{eq}, \\ P_j^{eq}/P_i^{eq} & \text{otherwise.} \end{cases} \tag{9}$$

The diagonal elements are instead obtained from the normalization condition

$$W_{ii} = \frac{1}{\Delta t} - \sum_{j \neq i} W_{i,j}. \tag{10}$$

It is easy to check that with this choice of W the equilibrium distribution is stationary under the evolution of Eq. (7). In the particular context of capsid assembly, the ratio of probabilities in Eq. (9) is straightforwardly obtained from Eq. (3) given the proportionality of $[i]$ and P_i.

Some caveats must be borne in mind when interpreting the outcome of the master equation as a kinetic process. While the equilibrium distribution is insensitive to the choice of the W's, as long as they satisfy Eq. (8), the kinetics is strongly affected by the form of the W matrix. Our choice follows the common practice of adopting the Metropolis criterion, but remains only one of the equivalent possibilities in terms of correct asymptotic behaviour. Future advancements in atomistic simulations may provide the possibility to parametrize the effective transition rates by more fundamental approaches, such as Kramers' theory. Also, we stress again that recasting Eq. (6) into the master equation in (7) was possible upon neglecting the time dependence of [1] in (6).

3. Application of the formalism to SV40 virus

In this section we apply the master equation formalism to the assembly of Simian Virus 40 (SV40) (Liddington et al., 1991). The capsid of SV40 is composed of 72 pentamers that adopt two different types of local configurations (with respect to their local bonding structure) in the capsid. In the framework of the tiling approach, viral capsids are represented as tessellations in terms of a finite number of shapes that encode the locations of the protein subunits and of the inter-subunit bonds between them. This is illustrated for the case of SV40 in Fig. 1, which is adapted from Twarock (2004). It is a tessellation in terms of the two types of shapes, or tiles, that are shown in Fig. 2. They encode the locations of the protein subunits as follows: angles subtending vertices of the tiles that meet at 5-coordinated vertices of the tiling are marked by dots and encode the locations of the protein subunits. The locations of the inter-subunit bonds (C-terminal arm extensions in the case of SV40) correspond to the lines (respectively, geodesics if viewed as a spherical tiling) connecting these dots. In particular, the dimer- and trimer interactions in the capsid, that is the interactions between two, respectively, three, protein subunits, are represented by the rhomb, respectively, kite, tiles.

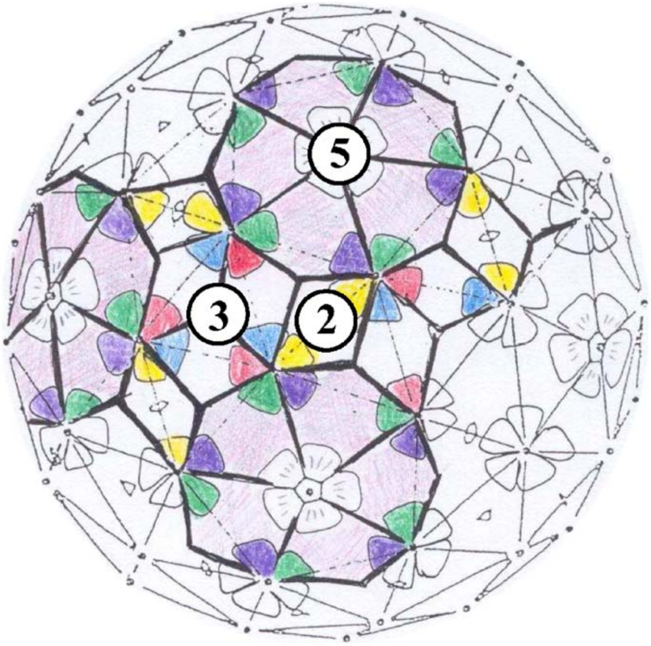

Fig. 1. The tiling representing the viral capsid of SV40.

SV40 has an icosahedrally symmetric capsid, and correspondingly also the tiling has this symmetry. We have indicated the locations of a 5-, a 3- and a 2-fold rotational symmetry axes in Fig. 1. One can see, based on the interpretation of the tiles outlined above, that the 12 pentamers located on the 5-fold axes of the capsid are surrounded by trimer-interactions, and the 60 other pentamers by a combination of dimer- and trimer-interactions. There are hence two different types of local environments, which are shown in Fig. 3.

Instead of using the tiles themselves, it is more convenient—and mathematically equivalent—to work with the pentagons and hexagons shown superimposed on the tiles in Fig. 3 (left). The edges of these shapes are perpendicular to the intersubunit bonds and will be labelled by letters that encode the types of the corresponding virtual bonds (same figure, right). For example, for SV40 there are three different types of bonds, which correspond to the association energies a for a single C-terminal arm in a kite tile, association energy b for a quasi-dimer bond ("yellow–yellow" rhombs in Fig. 1, named after their location on a local 2-fold symmetry axis), and c for a strict dimer bond ("blue–red" rhombs in the figure, named after their location on a 2-fold symmetry axis). They have to be taken into account when computing the free energies ΔG in Eq. (3).

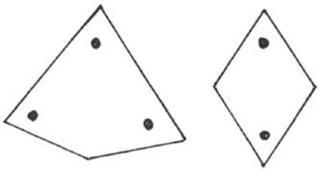

Fig. 2. Tiles corresponding to the tiling in Fig. 1.

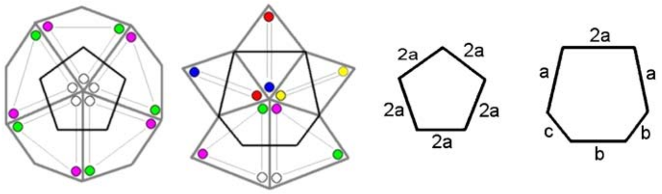

Fig. 3. The building blocks for SV40 capsid assembly.

In Keef et al. (2005) SV40 assembly is considered based on these pentagonal and hexagonal building blocks. It is modelled as a cascade of low order reactions by association of a single building block at a time. The complete characterization of the assembly thermodynamics would require the full classification of all possible species of correctly connected pentamers and hexamers, that is of all combinatorially possible combinations of these building blocks, even those comprising more than 72 blocks necessary to form the full capsid (and may correspond to malformed capsids or other types of closed or open structures). Obviously, this exhaustive enumeration cannot be accomplished. Consequently, it is necessary to reduce the number of building blocks to a manageable size by discarding configurations that are supposed to play an unimportant role in the assembly process. To illustrate the reduction procedure employed in this study it is convenient to regard the various species as the nodes of a graph. The links in the graph connect species which differ by the attachment/removal of a pentamer or hexamer. The graph containing the species to which we restrict our attention is constructed through the following procedure already employed in Zlotnick (1994) and Endres et al. (2005). Without loss of generality the first node is constituted by the fundamental pentamer. We then consider all the geometrically inequivalent species obtained via the addition of an extra building unit. Of these species we retain only the one (or ones in case of degeneracy) having the lowest free energy. Each of the retained species become new nodes of the graph (and are linked to the parent node). In correspondence of each of these offspring nodes we carry out the search of minimum free-energy descendants, as before. The process ultimately ends when the node corresponding to the full capsid is reached. We remark that this commonly employed procedure weeds out the possible intermediates through a directed pruning. In fact, the selection of most favourable intermediates introduces an asymmetry between offsprings nodes and parents nodes. We will explore at the end of this study how the initial stages of the capsid assembly are affected by relaxing the pruning criterion.

Within this limited set of species, the possible assembly pathways are represented as walks on the graph along linked nodes. We stress again that, due to the selection criterion based on the free-energy minimization, the graph we obtain in this way represents only a subset of the combinatorially possible nodes. The resulting "minimal graph" is uniquely encoded by the energy parameters, a, b and c. Since one of these three quantities can be taken as the unit of energy, we have that the structure of the graph depends only on two-adimensional parameters: a/c and b/c. In Keef et al. (2005), the phase diagram of the system corresponding to these parameters is depicted. It is shown that one can identify convex regions in this two-dimensional parameter space where parameters can be varied without affecting the assembly graph (but, obviously, the probability of occurrence of the various species will change from point to point in the same region).

While no direct measurement of the association energies a, b and c is available, an estimate for the assembly free energies is provided by the VIPER database. We discuss the assembly tree associated with the point $x \equiv b/c \approx 0.92$ and $y \equiv a/c \approx 0.47$, in the phase space, because it corresponds to the *ratios* of the association energies listed on the VIPER webpages for SV40.

In Fig. 4 we portray the portion of the assembly graph for SV40 limited to species with up to 16 building blocks. It contains 19 assembly intermediates, out of a *total number* of 505 assembly intermediates that are encoded by the minimal free-energy assembly graph (Keef et al., 2005). In this figure, we label assembly intermediates as **a**, b, where **a** denotes the iteration step and hence the number of pentamers that constitute that intermediate, and b, running from 01 upwards, distinguishes different intermediates with the same number of pentameric building blocks. The complicated structure of the assembly graph makes the use of relation (4) impractical for computing the relative concentrations of the assembly intermediates. Therefore, only the concentrations of the dominant assembly intermediates, i.e. those located on all paths in the assembly

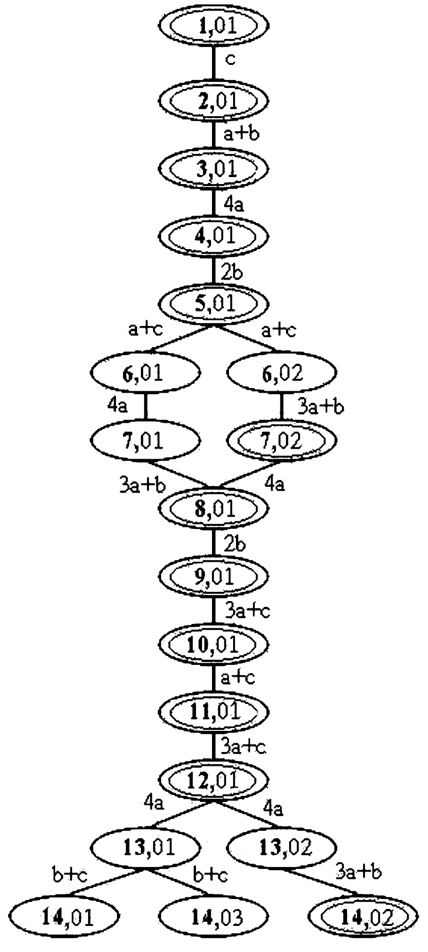

Fig. 4. The start of the assembly tree for SV40.

graph such as, for example, the intermediates denoted as **1**, 01 to **5**, 01 and **8**, 01 to **12**, 01 in Fig. 4, have been computed prior to this work (see Keef et al., 2005). However, the concentrations of all assembly intermediates are needed in order to obtain clues about the putative pathways of assembly, and hence about possible mechanisms for anti-viral drug design.

To compute the equilibrium probabilities of occupation of the various species, and to identify the dominant kinetic pathway within the assembly graph of Fig. 4, we must go beyond the mere specification of the adimensional parameters a/c and b/c and consider the absolute values of the energies a, b, c. The absolute value of the nominal free energies provided by the VIPER website are more than an order of magnitude larger than the typical interaction energies of biomolecules (usually of the order of a few $\mathrm{kcal\,mol^{-1}}$) (Horton and Lewis, 1992; Reddy et al., 1998, 2001). The particularly high VIPER values may indeed reflect the shortcoming of the rigid-unit approximations involved in the potential extraction scheme. We shall therefore obtain an indication of the absolute scale for the SV40 free energies and of the concentration c_0, by following some guidelines inspired by previous theoretical and experimental work.

The first quantitative experimental input pertains to the overall concentration of fundamental units present in solution, $[c^*]$, which is, typically, of the order of $10\,\mu\mathrm{M}$. Secondly, as anticipated, we wish to describe the situation where the dominant species in equilibrium are [1] and $[N]$.[1] Assuming that $[N]_{eq}$ and $[1]_{eq}$ are equiprobable one has

$$1 = \frac{[N]_{eq}}{[1]_{eq}} = 12\left(\frac{\widetilde{[1]}_{eq}}{5}\right)^{71} \mathrm{e}^{\kappa a/RT}, \qquad (11)$$

where $\kappa := 180 + 60b/a + 30c/a \approx 360.8571$.[2]

The above requirement provides a relationship through a, the free-energy scale, and c_0, entering implicitly through $\widetilde{[1]}_{eq}$. The second condition on c_0 and a is obtained by requiring that the concentration of species [2] is significantly smaller than [1]. This requirement implies, through the chain relations of Eq. (4), that the dominant species are indeed [1] and $[N]$, so that

$$[c^*] = \sum_{k=1}^{505} n_k \frac{5^{n_k}}{O(k)} \mathrm{e}^{-\Delta G(1,k)/RT} \widetilde{[1]}_{eq}^{\,n_k-1} [1]_{eq} \approx [1]_{eq} + 72[N]_{eq} = 73[1]_{eq}. \qquad (12)$$

We discuss here the case where $[2] = [1]/10$, which is satisfied when a takes on the realistic value $a \approx -0.7\,\mathrm{kcal\,mol^{-1}}$. All our conclusions about the dominant pathway in the assembly graph are unchanged if much larger values of a (in modulus) are used, although these might result in unrealistically low concentrations of intermediates.

Therefore, the assembly graph for SV40 has been computed for the values $a = -0.7\,\mathrm{kcal\,mol^{-1}}$, $b = -1.37\,\mathrm{kcal\,mol^{-1}}$ and $c = -1.49\,\mathrm{kcal\,mol^{-1}}$. The shortest pathways in the graph that connect [1] and $[N]$ (i.e. those without loops) contains precisely 72 species, one for each possible value of building blocks, see Fig. 3.

We have computed the concentrations of the assembly intermediates in thermodynamic equilibrium as shown in Fig. 5.

In particular, one observes that concentrations are highest at the beginning and at the end of the assembly pathways, and are strictly and rapidly decreasing (at the beginning) or increasing (at the end). For the intermediates at the start of the assembly graph shown in Fig. 4 one observes furthermore the following scenario: in the cases where more than one intermediate of the same number of building blocks exists (such as for example **6**, 01 and **6**, 02, or, **7**, 01 and **7**, 02) their concentrations are either identical as a consequence of degeneracy (such as for **6**, 01 and **6**, 02 and all other nodes springing out from a parent node), or vary strongly (**7**, 02 having a larger concentration than **7**, 01). In the latter case, we indicate the intermediate with the larger concentration by a double circle in Fig. 4. Since dead-ends and traps do not occur in the assembly graph, the pathways containing the two largely dominant configurations **7**, 02 and **14**, 02 in Fig. 4 must be the dominant pathways during the initial stages of the assembly.

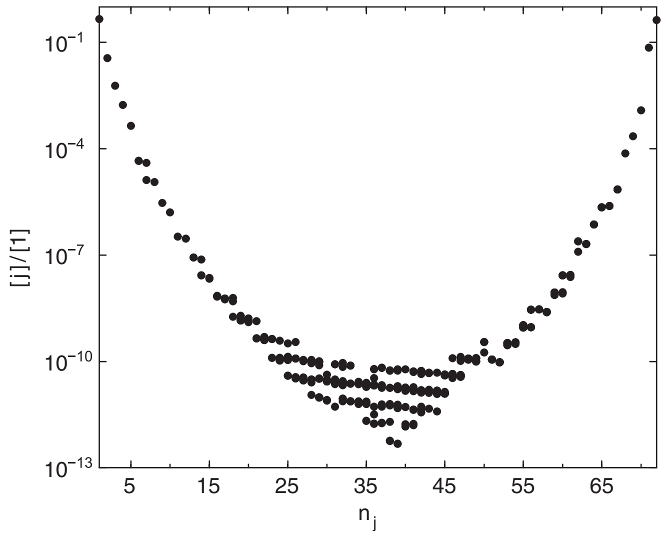

Fig. 5. SV40 capsid assembly: scatter plot for the concentrations of assembly intermediates, $[j]$, as a function of the number of building blocks, n_j. The concentrations were normalized with respect to that of the fundamental unit, [1]. Notice that more than a species may exist for given n_j. The plot refers to the situation where $[c^*] = 10\,\mu\mathrm{M}$ and $a = -0.7\,\mathrm{kcal\,mol^{-1}}$.

[1] For pentamers in solution we do not distinguish between the two different types of building blocks in Fig. 3 as their C-terminal arms are dangling freely and the building blocks are a way of encoding local bonding structures when bound in the capsid.

[2] Note that this equation relates the association energy a with the equilibrium concentrations of pentamers, $[1]_{eq}$, and hence changes in $[1]_{eq}$ may be engineered by changing a. The latter can be achieved for example via alterations in the polypeptide chain of the proteins (see e.g. Johnson et al., 2005).

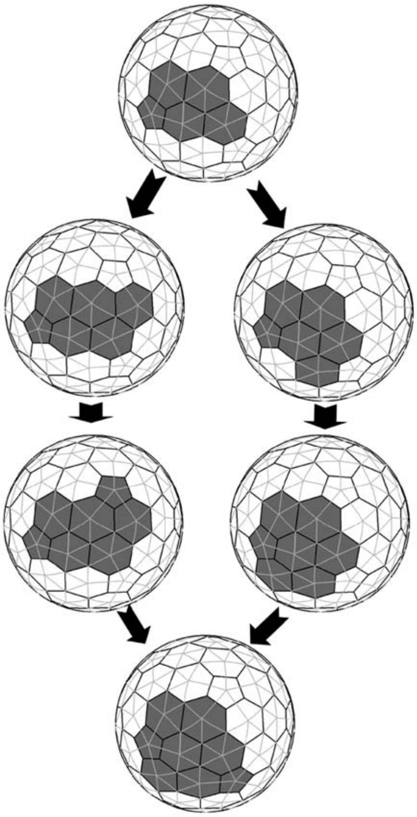

Fig. 6. The geometries of the intermediates **7**, 02 and **14**, 02.

The geometries of the intermediates **7**, 02 and **14**, 02 are shown in Fig. 6. One observes that in each case, the dominant configuration is obtained from the previous intermediate via the simultaneous formation of bonds with association energies 2a, a and b. The fact that the formation of this constellation of bonds is important is corroborated further by the following: we have increased the association energies of the bonds a and b individually, and have compared the ratio of final capsids to pentamers in equilibrium. In both cases the yield of final capsids has increased, with a stronger increase in response to an increase of the association energy related to the bond with association energy b.

These considerations suggest that SV40 capsid assembly is driven more by the details of the association free energies rather than differences in the geometrical entropy associated to rotational symmetries of the various species. This fact can be conveniently checked by setting to zero all association free energies, a, b and c and computing the contribution of the factors S_1 and S_n to the concentrations of the various species. One observes that all assembly intermediates (on different assembly pathways) of an equal number of building blocks have the same probability, with the exception of assembly intermediates that have a discrete rotational symmetry. Obviously, symmetry effects do not influence primary assembly intermediates i.e. those which are on all assembly pathways in the assembly graphs. Rather symmetry effects manifest particularly within large populations of intermediates, a situation that is encountered halfway the assembly process (species with 30–40 building blocks). Notable, intermediates with the highest symmetry weight typically do not coincide with those of lowest free energy, a fact that exemplifies the strong dependence of the dominant pathways on the association free energies.

In order to analyse whether this phenomenon occurs also for other assembly scenarios, we use the phase space formalism developed in Keef et al. (2005). It is a graphical method of analysing and visualizing the dependence of the assembly scenario on the values of the association constants. In particular, in the case of SV40 with three different association energies a, b and c, the phase space is two-dimensional, with axes corresponding to the ratios a/c and b/c, respectively. This phase space, which corresponds only to the first quadrant since a, b and c are of the same sign, can be partitioned into areas in which the qualitative behaviour of assembly, as encoded in the assembly graph, is indistinguishable. For any choice of association energies with ratios falling into any given area in this partition, the same assembly graphs occur. This implies that all choices of a, b and c that have ratios falling into the same area in the partition as SV40 must have the same assembly behaviour.

We explore a different assembly scenario by choosing a point in a different area in the partition of phase space. In order to depart as little as possible from the SV40 scenario, we pick a point in an area adjacent to the one that represents SV40. From the many possibilities suggested by Keef et al. (2005) in Fig. 5, we pick area 1, but any other area could also have been chosen. A representative point of this area is $x = 0.75$ and $y = 0.45$. The complete assembly graph consists of 281 species, and the start of the graph is shown in Fig. 7. The assembly intermediates with the larger concentrations are again marked by double circles. As in the case of SV40 assembly, the occurrence of assembly intermediates with concentrations larger than that of the other intermediates with the same number of building blocks is related to the simultaneous formation of bonds with association energies 2a, a and b. This phenomenon hence seems pertinent to the selection of the dominant pathways, and may therefore provide insight into those aspect of viral capsid assembly that may become targets of anti-viral drugs.

We finally investigate in how far the pruning of the assembly graphs to the energetically most favourable assembly intermediates affects the result. For this, we have constructed an assembly graph that contains all assembly intermediates that have energies within 20% of those of the energetically most favourable ones. Ideally, one would like to consider all combinatorially possible assembly intermediates, but their number is too large (about the size of the entire assembly graph of SV40 after only 10 construction steps) so that this scenario is computationally too costly. However, the restricted setting is sufficient to investigate the effects of pruning.

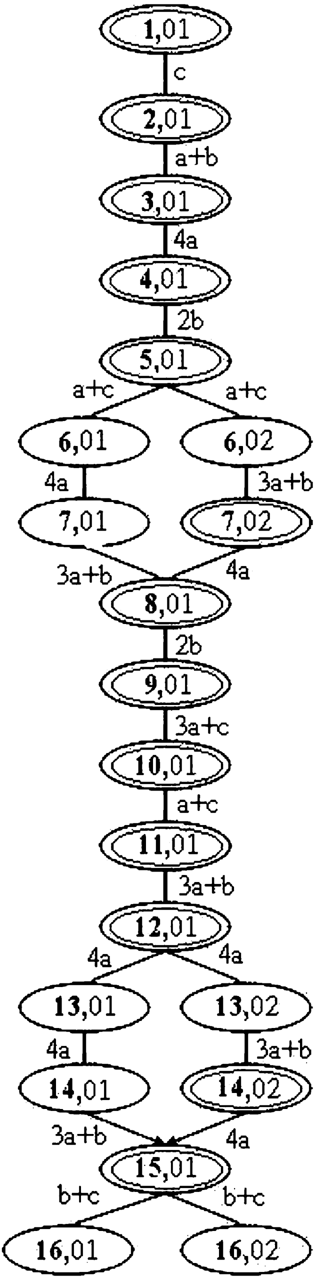

Fig. 7. The start of the assembly graph for a representative of a different area in phase space.

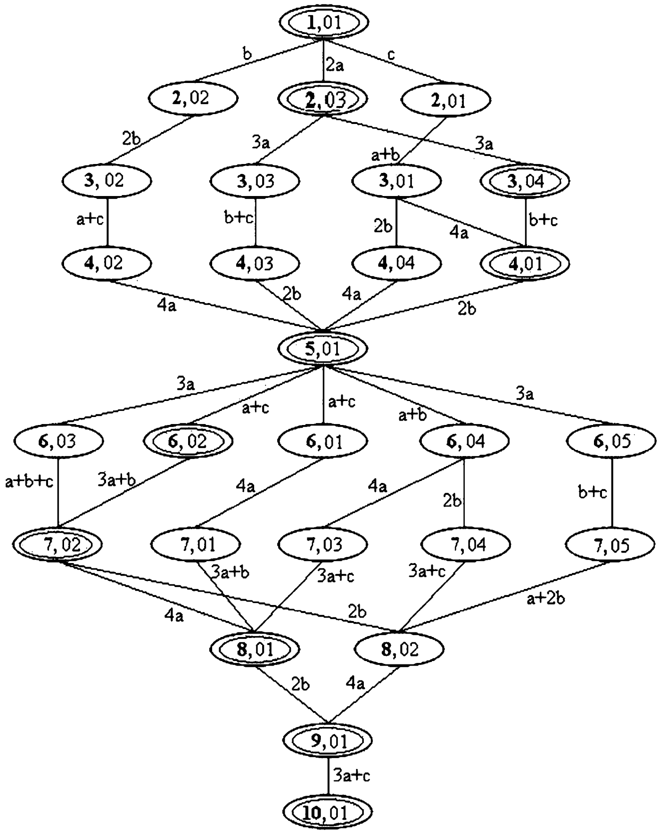

Fig. 8. The start of the assembly graph for a scenario that takes all assembly intermediates within 20% of the energetically most favourable ones into account.

In Fig. 8 we show the start of the corresponding assembly graph. The labelling has been chosen such that the energetically most favourable intermediates have the same labels as in Fig. 4 in order to facilitate comparison, while the additional intermediates are numbered in increasing order in each row. Instead of the 12 assembly intermediates at the start of the assembly graph in Fig. 4 (up to and including **10**, 01), one obtains 37 intermediates in this extended scenario. We have computed the probability distribution of the intermediates in equilibrium, and have marked the most probable pathway by double circles. A comparison with Fig. 4 shows that there is a deviation only at iteration steps 2 and 3, with **2**, 02 and **3**, 04 on the pathway instead of **2**, 01 and **3**, 01 as before. Otherwise, the pathway coincides with our earlier results. The geometric implications of this difference at the beginning of the pathway are illustrated in Fig. 9. The first and last geometry correspond to the assembly intermediates **1**, 01, respectively, **4**, 01, which are located on the dominant pathway both in the restricted and in the extended setting. The dominant pathway in the new scenario is shown on the left (intermediates **2**, 02 and **3**, 04), and in the previous one on the right (intermediates **2**, 01 and **3**, 01). Since the relaxation of the pruning rule appears to affect only a limited portion of the initial assembly tree, it appears a posteriori that adopting the commonly employed pruning rule does not alter the salient features of the assembly scenario.

4. Discussion

We have demonstrated that a combination of the master equation and the tiling approach allows us to determine the putative pathways for SV40 capsid assembly and sheds light on the mechanisms that drive this process. In particular, we have demonstrated that the more important assembly pathways are those where a particular constellation of bonds is formed at an early

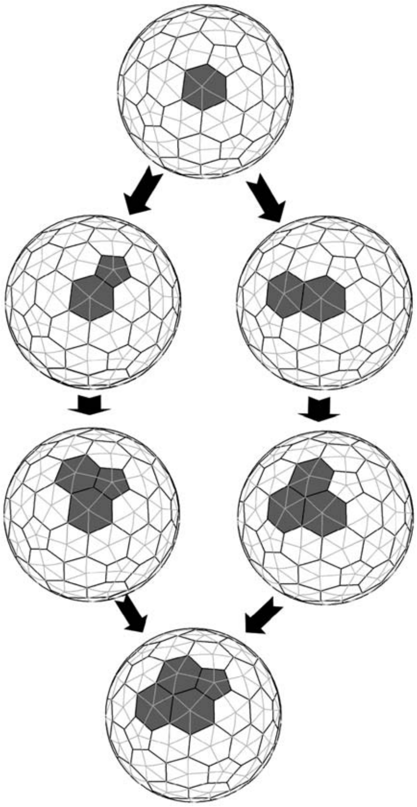

Fig. 9. A comparison of the geometries of the assembly intermediates on the most probable pathway in the extended scenario (left) and the scenario restricted to the energetically most favourable intermediates (right).

stage (see Fig. 6). Hence, this constellation of bonds could be a possible target for anti-viral drug design. For example, it suggests to search for a drug that binds to the sites related to these bonds, hence preventing their formation.

Moreover, our analysis has shown that SV40 capsid assembly is strongly driven by the details of the association free energies of the tiles and is only slightly affected by the entropic aspects associated with the rotational symmetries of the various species. This result provides a justification for simplifying capsid assembly models by neglecting certain types of combinatorially possible intermediates. We remark, however, that this conclusion may not hold for general families of viruses, as demonstrated for example in Endres et al. (2005), and in particular may not extend to RNA viruses, because the interactions between the RNA and the protein building blocks of the capsids play a leading role in guiding capsid assembly (Keef and Twarock, 2006).

Acknowledgements

We are indebted to Adam Zlotnick for valuable comments in response to a careful reading of the manuscript. R.T. has been supported by an EPSRC Advanced Research Fellowship. T.K. has been supported by the EPSRC Grant GR/T26979/01.

References

Berger, B., et al., 1994. Proc. Natl Acad. Sci. 91, 7732.
Brooks III, C.L., et al., 1998. Proc. Natl Acad. Sci. 95, 11037.
Casjens, S., 1985. Virus Structure and Assembly. Jones and Bartlett, Boston, Massachusets.
Caspar, D.L.D., Klug, A., 1962. Cold Spring Harbor Symposium on Quantization Biology, vol. 27, p. 1.
Endres, D., Zlotnick, A., 2002. Biophys. J. 83, 1217.
Endres, D., Miyahara, M., Moisant, P., Zlotnick, A., 2005. Protein Sci. 14, 1518.
Horton, N., Lewis, M., 1992. Protein Sci. 1, 169.
Itzykson, C., Drouffe, J.-M. Statistical Field Theory, vol. 2.
Johnson, J.M., Tang, J., Nyame, Y., Young, M.J., Zlotnick, A., 2005. Nano Lett. 5, 765.
Keef, T., Twarock, R., 2006. MS2 assembly guided by interactions between capsid proteins and RNA, in preparation.
Keef, T., Taormina, A., Twarock, R., 2005. J. Phys. Biol. 2, 175.
Kerner, R., 1994. Comput. Mater. Sci. 2, 500.
Kerner, R., 2005. J. Theor. Med. 6, 95.
Kerner, R., dos Santos, D.M.L.F., 1988. Phys. Rev. B 37, 3881.
Kerner, R., Bennemann, K.H., Penson, K.A., 1992. Europhys. Lett. 19, 363.
Liddington, R.C., et al., 1991. Nature 354, 278.
Modis, Y., et al., 2002. EMBO J. 21, 4754.
Rapaport, D.C., et al., 1999. Comput. Phys. Commun. 121, 231.
Rapaport, D.C., et al., 2004. Phys. Rev. E 70, 051905.
Rayment, I., et al., 1982. Nature 295, 110.
Reddy, V.S., et al., 1998. Biophys. J. 74, 546.
Reddy, V.S., et al., 2001. J. Virol. 75, 11943.
Schwartz, R., et al., 1998. Biophys. J. 75, 2626.
Schwartz, R., et al., 2000. Virology 268, 461.
Twarock, R., 2004. J. Theor. Biol. 226, 477.
Twarock, R., 2005a. Bull. Math. Biol. 67, 973.
Twarock, R., 2005b. J. Theor. Med. 6, 87.
Wales, D.J., 1987. Chem. Phys. Lett. 141, 478.
Wales, D.J., 1996. Science 271, 925.
Wolynes, P.G., 1996. Proc. Natl Acad. Sci. 93, 14249.
Zandi, R., et al., 2004. Proc. Natl Acad. Sci. 101, 15556.
Zlotnick, A., 1994. J. Mol. Biol. 241, 59.
Zlotnick, A., et al., 1999. Biochemistry 38, 14644.
Zlotnick, A., et al., 2000. Virology 277, 450.
Zlotnick, A., et al., 2002. J. Virol. 76, 4848.