Math. Model. Nat. Phenom.
Vol. 6, No. 6, 2011, pp. 136-158
DOI: 10.1051/mmnp/20116608

Self-Assembly of Icosahedral Viral Capsids: the Combinatorial Analysis Approach

R. Kerner [1]

LPTMC, Université Pierre et Marie Curie, CNRS UMR 7600,
Tour 23, 5-ème , Boite 121, 4 Place Jussieu, 75005 Paris, France

Abstract. An analysis of all possible icosahedral viral capsids is proposed. It takes into account the diversity of coat proteins and their positioning in elementary pentagonal and hexagonal configurations, leading to definite capsid size. We show that the self-organization of observed capsids during their production implies a definite composition and configuration of elementary building blocks. The exact number of different protein dimers is related to the size of a given capsid, labeled by its T-number. Simple rules determining these numbers for each value of T are deduced and certain consequences concerning the probabilities of mutations and evolution of capsid viruses are discussed.

Key words: viral capsid growth, self-organized agglomeration, symmetry
AMS subject classification: 92B05

1. Introduction

Viruses are a particular form of biological creations for which it is not totally clear whether their place is among the living organisms or somewhere between these and the organic matter. Viruses do not display some of the characteristic functions of living organisms, like metabolism. They are certainly "offsprings" of living organisms, and interact with them (alas!) very strongly. Both their structure and their operational mode are closely related to those of the *phagocytes* which are produced by advanced multicellular living organisms in order to fight against foreign parasite cells attacking them, which suggests that their possible origin is related to the phagocytes ([1]).

[1]E-mail: richard.kerner@upmc.fr

In principle, a virus is just a package of nucleic acids that can duplicate themselves, like the genetic material of other living organisms. Using the amino acids of the host cell, they duplicate their own genetic code, thus destroying the host cell. Inside the cell they multiply with an extraordinary efficiency, in certain cases about 100 new copies after a couple of hours, then leave the host cell, infecting its neighbors. This provokes a chain reaction that results in a serious, and sometimes fatal, illness of the entire organism.

However, the genetic material stored in the DNA chain which is the viruses' most important part, is very fragile and would be destroyed in a short time were it exposed to ultraviolet radiation and constant chemical attack by active gases like SO_2, O_3 or NO in the air. This is why most of the viruses hide their DNA and other "spare parts" in shells called "capsids" made of special proteins called *coat proteins*.

The capsid also contains certain enzymes that are able to open a hole in cells' membranes and make possible viruses' penetration inside. After that, the capsids open and the DNA chains can start their multiplication. There is also another agent in capsid viruses, the RNA chain that contains information how to construct new capsids. While the viral DNA replicates itself using host cell's genetic material, the RNA orders production of new capsids from proteins available inside the host cell. After some time (usually a couple of hours) new capsids are ready, as are the new DNA chains. The latter start to fold themselves and enter the capsids through special holes left for DNA's penetration. In many cases, the folding of the DNA is helped by a special molecular motor placed at the orifice. After the packing is accomplished, the capsid closes with other coat proteins, and the new virus is ready to leave the cell and penetrate any of its neighbors. Many species cover the capsid with an extra protective coat called the *tegumen*, as shown in the Figure 1 below:

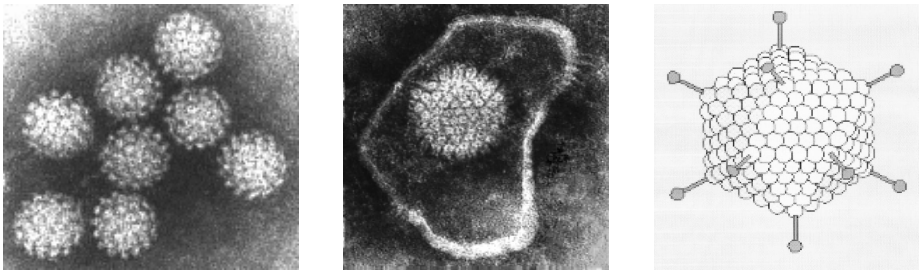

Figure 1: Capsids of the *papilloma* virus as seen with electronic microscopy; A HSV2-virus in a capsid, surrounded by a protective coat called *tegumen* (left) and schematic representation of an *adenovirus* (right).

This numerous group of viruses is called *capsid viruses*. A very important subgroup of these form capsids which display a perfect icosahedral symmetry.

A graphic representation of capsid viruses' replication in an infected cell is shown schematically in Figure 2.

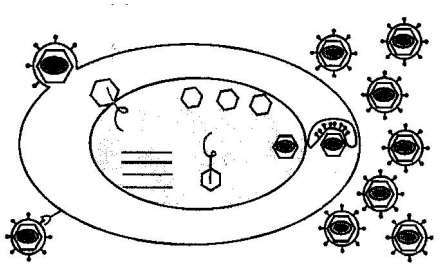

Figure 2: Schematic representation of capsid virus' reproduction cycle. A virus penetrates the cell (upper left), then enters cell's nucleus. Its DNA chain is replicated and capsid shells are produced parallelly (center); then the DNA chains pack into the empty capsids, and new viruses leave the cell (right)

2. Icosahedral capsid viruses

The icosahedral viral capsids are one of the most spectacular examples of self-organization of giant proteins which can build up extremely complicated structures. First theoretical bases of quantitative, physical and mathematical analysis of such processes were set forth by Manfred Eigen in his classical book on biological self-organization [3].

Just like the world of planets and stars, the world of viruses is ruled by numbers. This is particularly true in the case of the numerous group of *spherical viruses*, whose protective protein shells called "capsids" display perfect icosahedral symmetry [1]. It is amazing that these structures, known to mathematicians since Coxeter's classification [2], are also observed in the so-called *fullerenes*, huge molecules composed of 60 carbon atoms, predicted by Smalley and Kroto, and discovered in the eighties.

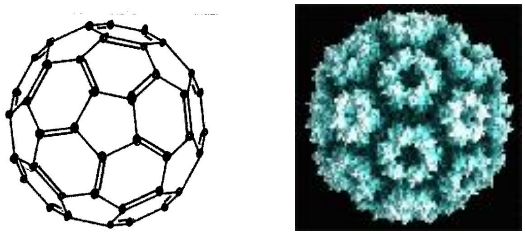

Figure 3: Schematic representation of fullerene C_{60} molecule made of 60 carbon atoms. On the right, the capsid shell of the *Cowpea mottle virus*, displaying internal structure with the same icosahedral symmetry.

.

Since Caspar and Klug [5] introduced simple rules predicting a sequence of observed viral capsids, several models of growth dynamics of these structures have been proposed, e.g. A. Zlotnick's model [6] published in 1994. The common geometrical feature of icosahedral viral capsids is their general structure, with twelve pentagons found on the opposite sides of six five-fold symmetry axes, and an appropriate number of hexagons in between. The number of hexagons is given by the following simple formula: $N_6 = 10\,(T-1)$, with $T = (p^2 + pq + q^2)$, called *the triangular number*, where p and q are two non-negative integers [2].

In capsids, the building blocks made of *coat proteins* are called *monomers*, *dimers*, *trimers*, *pentamers* and *hexamers*, according to their shape, the bigger ones usually being assembled from smaller ones prior to further agglomeration into capsid shells [13] . Sometimes pentameric or hexameric symmetry is displayed despite the direct construction from 60 or 180 smaller subunits, like in the *Cowpea mosaic virus* and the *Cowpea chlorotic mottle virus*, respectively [7]. Although certain virus species grow medium-size capsids corresponding to $N_6 = 20$ (like in the C_{60} fullerene molecule), or $N_6 = 30$ and $N_6 = 60$, some of them form pure dodecahedral capsids (with exclusively pentamers as building blocks), like certain *Comoviridae* or *Cowpea* virus [12], while others, like human *Adenovirus* [10], form very huge capsids with $N_6 = 240$, corresponding to $p = 5,\ q = 0$. Even much greater capsids are also known, with T-numbers up to 219.

In some cases, the similarity with the fullerene structure is striking: for example, the TRSV capsid is composed of 60 copies of a single capsid protein (56 000 Da, 513 amino acid residues) [15], which can be put in a one-to-one correspondence with 60 carbon atoms forming a fullerene C_{60} molecule; the aforementioned *Cowpea* viruses provide another example of the same type.

Capsids are built progressively in liquid medium, from agglomerates of giant protein molecules displaying pentagonal or hexagonal symmetry, or directly from smaller units (*monomers* or *dimers*). It seems that there is no such thing as universal assembly kinetics: the way the capsids are assembled differs from one virus to another. The $T = 7$ phage HK47 appears to build pentamers and hexamers first, then assemble these capsomers to form the final capsid structure, whereas another $T = 7$ phage labeled P22 appears to assemble its capsids directly from individual coat proteins (see [14]) and the references within).

The common point is the presence of pentagons and hexagons in the resulting structure, and the strict topological rules that result from Euler's theorem on convex polyhedra: $V - E + F = 2$, with V number of vertices, E number of edges, and F number of faces. From this one derives the fact that when only pentagonal and hexagonal faces are allowed, the number of pentagons is always $N_5 = 12$, while the hexagon number is $N_6 = 10(T - 1)$. The first three icosahedrons constructed according to these rules are displayed in the Figure 4 below. The triangular numbers are, respectively, $T = 1,\ (p, q) = (1, 0)$ which is a pure dodecahedron composed of twelve pentagons, then $T = 3,\ (p, q) = (1, 1)$ and finally $T = 4,\ (p, q) = (1, 2)$.

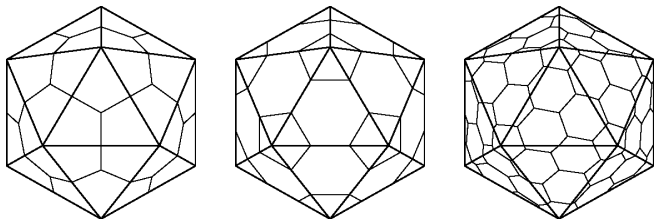

Figure 4: The overall schematic view of T=1, T=3 and T=4 capsids (With kind permission of ViperD)

It is important to stress the extreme efficiency of the capsid building process. Viruses use almost 100% of pentamers and hexamers at their disposal to form perfect icosahedral capsid structures, into which their DNA genetic material is densely packed once the capsid is complete. There is almost no waste in the process.

This means that the initial nucleation ratio of pentamers versus hexamers is very close to its final value in capsids in order to minimize the waste. Secondly, the final size of the capsid must depend on particular assembly rules, which can be fairly well deduced from statistical weights of various agglomeration steps, found by maximizing the final yield. In the following section we investigate the rules that define the type and the size of capsids, simultaneously optimizing the production rate.

3. Combinatorics and statistics of icosahedral capsid growth

We investigate here a model in which statistical factors play the decisive role in ensuring that the "correct" configurations are produced at each consecutive step almost without exception, i.e. practically with a 100% yield. Let us show now how these statistical factors can be evaluated, and what constraints they imply on capsomers' structure.

Let us denote the concentration (or the nucleation rate) of pentamers by c, that of hexamers by $(1-c)$. The agglomeration process can be divided in consecutive steps, the first one being the creation of pairs of capsomers sticking to each other side by side. We shall refer to these couples as *doublets*. The second step consists in creating a *triplet* by adjoining next capsomer (a "*singlet*") to one of the doublets, and so forth. Some of the first steps are represented in Figure 5 below.

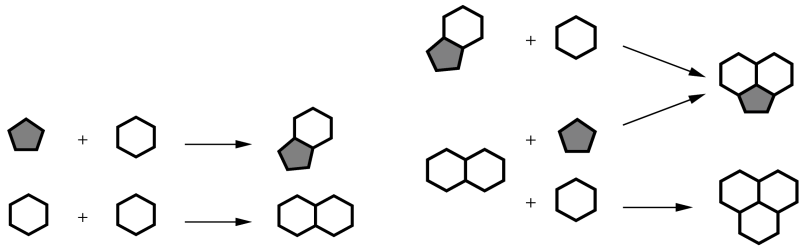

Figure 5: The first two agglomeration steps. Note that the creation of the (55) pairs is not shown here.

It is easy to convince ourselves that the agglomeration of capsids is anything but random. Indeed, if the pairs were formed randomly, then the probabilities of observing the three possible doublets, $(55),\ (56)$ and (66) would be proportional to the products of the probabilities of finding the corresponding singlets. Denoting the probabilities of forming corresponding pairs by $P_{55},\ P_{56}$ and P_{66}, where P_{56} stands for both combinations, (56) and (65), we should have in case of totally random encounters

$$P_{55} = c^2, \quad P_{56} = 2\,c(1-c), \quad P_{66} = (1-c)^2. \tag{3.1}$$

Note that these are genuine probabilities, normed to 1, as

$$P_{55} + P_{56} + P_{66} = (1-c)^2 = c^2 + 2\,c(1-c) + (1-c)^2 = [c + (1-c)]^2 = 1^2 = 1.$$

Similarly, we could go on calculating the probabilities of triplets resulting from encounters of doublets already formed with randomly chosen singlets. Certain triplets can be obtained in two

different ways, e.g. (566) can result from adding a (5) to a (66) pair, or by adding a (6) to a (56) pair. Hence, we get, with obvious notations:

$$P_{555} = c\,P_{55}, \quad P_{556} = c\,P_{56} + (1-c)\,P_{55},$$

$$P_{566} = c\,P_{66} + (1-c)\,P_{56}, \quad P_{666} = (1-c)\,P_{66}. \tag{3.2}$$

The probabilities of particular doublets are multiplied by the probability to encounter a (5)-singlet, equal to c, or a the probability to meet a (6) singlet, $(1-c)$. The resulting expressions are normalized to 1 again, as we have

$$P_{555} = c^3, \quad P_{556} = 3c^2(1-c), \quad P_{566} = 3c(1-c)^2, \quad P_{666} = (1-c)^3.$$

Now, the above expressions are obtained by counting only the probabilities of encounters, without taking into account the number of ways one capsomer can stick to another one by sharing a common edge. As a matter of fact, if the edges were totally undifferentiated, then there would be $5 \times 5 = 25$ possibilities how to stick one pentamer to another, $5 \times 6 = 30$ possibilities of sticking a pentamer to a hexamer, and the same number of possibilities of sticking a hexamer on a pentamer. Finally, there would be 6×6 different choices of two edges when sticking together two hexamers.

With this in mind, the probabilities of producing doublets should contain extra statistical factors:

$$P_{55} \sim 25\,c^2, \quad P_{56} \sim 2 \times 30\,c(1-c), \quad P_{66} \sim 36\,(1-c)^2.$$

These expressions should be now normalized in order to represent genuine probabilities; the normalizing factor Q is the sum of all contributions,

$$Q = 25\,c^2 + 60\,c(1-c) + 36\,(1-c)^2,$$

so that the probabilities of all possible doublets sum up to 1:

$$P_{55}\frac{25\,c^2}{Q}, \quad P_{56} = \frac{60\,c(1-c)}{Q}, \quad P_{66}\frac{36\,(1-c)^2}{Q}. \tag{3.3}$$

In a totally random agglomeration process the average content of pentamers at any stage will reproduce their original concentration in the singlet stage. For example, if we want to evaluate the average number of pentagons versus all capsomers in totally random doublets, we should define the function $c^{(1)}$ as follows:

$$c^{(1)} = \frac{1}{2}\,\left[2P_{55} + P_{56}\right], \tag{3.4}$$

Here the superscript $^{(1)}$ stays for the first agglomeration step (from singlets to doublets). We divide by 2 because there are two capsomers per doublet, of which there are two pentamers in each (55)-doublet, and only one in each (56) doublet.

It is easy to see that in case of totally random pairing one has $c^{(1)} = c$. This is not the case when one takes into account statistical weights due to the geometry of the items to be assembled, as well as the possibility of chemical affinities or barriers excluding certain pairings. Already taking

into account the multiplicities of various pairings due to the number of available edges to share we obtain $c^{(1)}$ different from c:

$$c^{(1)} = \frac{1}{2}\,\frac{2\times 25c^2 + 60c(1-c)}{25c^2 + 60c(1-c) + 36(1-c)^2} \neq c.$$

The difference $c^{(1)} - c$ can be interpreted as an aproximate first derivative of c with respect to a continuous parameter s, ("s" for "step"), measuring the average advancement of agglomeration process. For better understanding of this parameter's nature, let us consider the first stage of agglomeration, when doublets start to form from singlets, then some amount of triplets appears, but one can assume that the bigger clusters are not yet formed. Let the total number of clusters with one, two or three capsomers be respectively N_1, N_2 and N_3. Then the average number of capsomers per cluster will be given by

$$s + 1 = \frac{N_1 + 2N_2 + 3N_3}{N_1 + N_2 + N_3}. \tag{3.5}$$

The parameter s can be treated as a continuous one when the numbers N_i are sufficiently great. With the above definition 3.5 with only doublets present its value will be 1, corresponding to the first agglomeration step; if all clusters are exclusively triplets, $s = 2$, corresponding to the second step, end so on.

Suppose now that the original pentamer rate among the originally produced capsomers is c, whereas in doublets the same rate has a different value, $c^{(1)}$. Then the average pentamer rate in the mixture containing N_1 singlets and N_2 doublets is given by

$$c(s) = (1-s)\,c + s\,c^{(1)}. \tag{3.6}$$

Differentiating this expression with respect to s we get the *master equation*

$$\frac{dc}{ds} = c^{(1)} - c, \tag{3.7}$$

describing the initial evolution of pentamer concentration in newly formed clusters.

To take the example of *Cowpea* capsid virus, it is clear that pentamers cannot touch each other. The remaining two combinations may be more or less preferred due to different chemical potential barriers between different sides of agglomerating capsomers.

Then the probabilities of doublets are readily calculated as follows:

$$P_{56} = 2c(1-c)\cdot W_{56}/Q; \quad P_{66} = (1-c)^2\,W_{66}/Q, \tag{3.8}$$

where W_{jk}, $j,k = 5,6$ are the statistical weights depending on the virus type and chemical barriers between various sides, and $Q = 2c(1-c)\cdot W_{56} + (1-c)^2 W_{66}$ is the normalizing factor. Note that we exclude two pentamers coming together, i.e. $W_{55} = 0$. Similarly, the probabilities of admissible "triplets" in the next agglomeration step displayed in Figure 5 are given by:

$$P_{566} = P_{56} + c\cdot P_{66}W_{66,5}/Q_2, \quad P_{666} = (1-c)\cdot P_{66}W_{66,6}/Q_2, \tag{3.9}$$

where $W_{66,5}$ and $W_{66,6}$ denote the statistical weights of the corresponding agglomeration processes, and $Q_2 = W_{66,5}c + W_{66,6}(1-c)$.

Now, we can evaluate the average pentamer ratio $c^{(k)}$ in clusters of a given size after the k-th agglomeration step. The first three values are given by:

$$c^{(1)} = \frac{1}{2}P_{56}, \quad c^{(2)} = \frac{1}{3}P_{566}, \quad c^{(3)} = \frac{1}{4}(2P_{5665} + P_{5666} + P_{6665}), \text{ etc.}$$

(In the expression for $c^{(3)}$ we use the probabilities for three different allowed clusters, which can be seen on Figure 6).

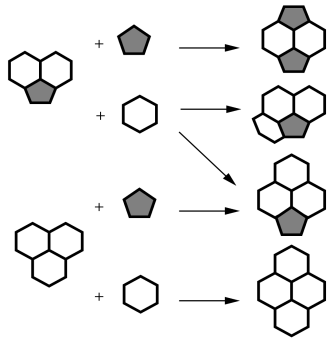

Figure 6: Random agglomeration of 5- and 6-sided polygons

We can use these formulae in two different ways. Either we impose the statistical weights, and determine the consecutive pentamer concentrations in growing clusters, starting from a given initial concentration c. Or we can treat the statistical weights as unknowns and determine them from self-similarity equations for successive pentamer concentrations:

$$\frac{c^{(n+2)} - c^{(n+1)}}{c^{(n+1)} - c^{(n)}} = \frac{c^{(n+1)} - c^{(n)}}{c^{(n)} - c^{(n-1)}}, \quad n = 2, 3, 4..., \tag{3.10}$$

where the general expression for pentamer concentration in all clusters resulting from the n-th agglomeration step is given by the recurrent formula:

$$c^{(n)} = \frac{1}{n}\sum_{k} k\, P_{(k\times 5)((n-k)\times 6)}, \tag{3.11}$$

the summation going up to the maximal admissible number of pentamers in clusters with n capsomers; this maximal number depends on n and we can get it only by direct inspection of all configurations with edge-sharing pentamers excluded. The resulting solutions for the limit values of c and for the auxiliary variables $\xi = (W_{56})/(W_{66}), \eta = (W_{56,6})/(W_{66,5}), \zeta = (W_{66,6})/(W_{66,5})$, etc., although usually not in the form of simple fractions, give very good hints concerning the assembly rules leading to particular capsid structures.

For example, if the sides of all hexamers were equivalent and could stick with equal probability to the sides of the pentamers as well as between themselves, the rate of production of proper (T=3)

capsids would be close to 2^{-13}, which is not the case. Therefore, hexamers and pentamers must display strict sticking rules discriminating against the undesired configurations.

4. The simplest case: the MS2 phage

The $MS2$ phage forms one of the smallest icosahedral capsids. It is often referred to as a "pseudo-$T3$" capsid, because its coat proteins are assembled in dimers of a roughly rhombic shape which assemble to form mutually interpenetrating star-like pentamers and hexamers. The shells produced in this way could be divided into pentagons and heptagons like genuine $T = 3$ capsids, but this will cut the real coat proteins in halves.

A symbolic representation of one of the coat proteins on the $MS2$ phage and the structure of its capsid subdivided into rhombs representing coat proteins of two different types are displayed in the Figure 7 below: Two different rhombic coat proteins serve as the elementary building blocks.

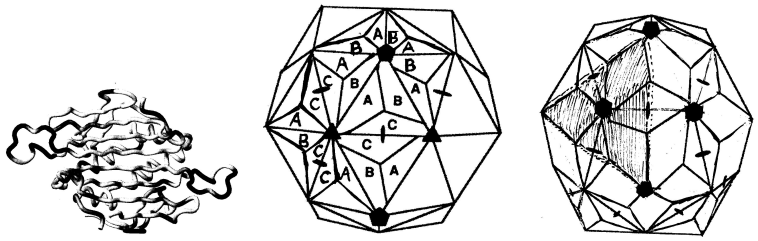

Figure 7: A representation of one of the dimer proteins and the disposal of two different kinds of dimers in the $MS2$ capsid. One of the 20 triangles forming the icosahedral structure is shown in the third drawing

They are denoted as AB-proteins and as CC-proteins. The five-fold and the six-fold symmetry axes are shown, as well as the two-fold symmetry axis in the middle of each CC-protein. It is easy to see that a complete $MS2$-capsid contains sixty AB-rhombs and thirty CC-rhombs, altogether 90 rhombs, so that the relative frequencies of AB and CC species are respectively $\frac{2}{3}$ and $\frac{1}{3}$. It is also easy to check that Euler's formula works here, too. The number of vertices is equal to 92 which can be found by analyzing the smallest triangle containing three entire AB rhombs and three halves of CC rhombs; there are twenty such triangles forming the entire icoahedron. It is easy to check that each such triangle contains one entire vertex (the six-fold symmetry center) in its center, six halves of vertices belonging to the three CC rhombs and three fifths of vertices belonging to the three AB rhombs which form the vertices of the elementary triangle. Then the count is: $20 \times (\frac{3}{5} + \frac{6}{2} + 1) = 20 \times \frac{46}{10} = 92$ The number of edges is equal to 180 because there are 90 rhombs, each of which has four edges, but each edge is shared between the two adjacent rhombs, and the result is $(90 \times 4)/2 = 180$. The final result is $F - E + V = 90 - 180 + 92 = 2$ as it should be.

The next important observation concerns the sticking rules. The AB proteins stick together exclusively through their A and B sides, and stick to the CC dimers with the remaining two sides, denoted by AC and BC, which also stick to corresponding sides of CC dimers, denoted by CA

and CB. The only six allowed pairings are encoded in the *affinity matrix*, containing all possible pairings, with the allowed ones marked with a "1", and the forbidden ones with a "0":

	A	B	AC	BC	CA	CB
A	0	1	0	0	0	0
B	1	0	0	0	0	0
AC	0	0	0	0	1	0
BC	0	0	0	0	0	1
CA	0	0	1	0	0	0
CB	0	0	0	1	0	0

Table I. The affinity matrix for assembly of the MS2 phage capsid with two different rhombs displaying six different side types

The affinity matrix displayed in the Table I represents an ideal situation in which the energy barriers between the two CC dimers are infinite, as well as the barriers prohibiting the formation of all other "wrong" combinations. This leaves as the only possibility the following three different sticking modes: $AB + AB$ forming a doublet with an A side of one AB dimer connecting to the B side of another AB dimer. We shall prove now that this hypothesis corresponds very well to the observed almost 100% efficiency of the capsid building process.

Let us denote the rate of the AB dimers by p and the rate of the CC dimers by $q = 1 - p$.

The probabilities of finding an $AB+AB$ doublet or one of the two possible $AB+CC$ doublets are then, respectively:

$$p_{ABAB} = 2p^2/Q, \quad p_{BAAC} = 2 \times 2\,p(1-p)/Q,$$
$$p_{ABBC} = 2 \times 2\,p(1-p)/Q, \quad Q = 2p^2 + 8p(1-p). \tag{4.1}$$

The three corresponding doublets of dimers are displayed in the Figure 8.

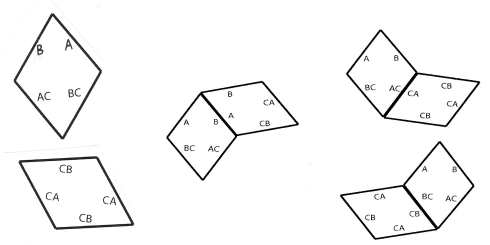

Figure 8: Two types of dimers, the AB and the CC proteines, and the three possible doublets

In the doublets, the new probability distribution can be readily computed to give the following expression:

$$p^{(1)} = \frac{1}{2}\left(2p_{ABAB} + p_{BAAC} + p_{ABBC}\right) = \frac{p + 2\,(1-p)}{p + 4(1-p)}, \tag{4.2}$$

We can use again the master equation describing the evolution of probabilities during the agglomeration process:

$$\frac{dp}{dt} \simeq p^{(1)} - p = \frac{p + 2\,(1-p)}{pe^{-\alpha} + 4(1-p)} - p, \tag{4.3}$$

When put to the common denominator the right-hand side is proportional to the following expression:

$$\frac{dp}{dt} \simeq p + 2(1-p) - p^2 - 4p(1-p) = 3p^2 - 5p + 2, \tag{4.4}$$

which has two solutions: the obvious one when $p_1 = 1$ and when only the AB dimers agglomerate in total absence of the CC dimers, or the second one given by

$$p_2 = \frac{2}{3}. \tag{4.5}$$

By linearizing the differential equation 4.2 in the vicinity of these two singular solutions one can easily see that the point $p = 1$ is repulsive, while the solution $p = \frac{2}{3}$ of 4.5 is an attractive stable point, which coincides with the exact proportion of the AB dimers in the MS2 capsid when it is completed.

Adding another dimer (an AB or a CC) to one of the three doublets leads to the creation of *eight* different triplets. A detailed analysis of probability distributions leads to the same singular point at $p = 2/3$, which confirms the validity of the agglomeration scheme and of the affinity matrix (Table I).

On the other hand, if the CC dimers had undifferentiated sides i.e. if they could stick to anyone of the two accessible sides of an AB dimer, the result would be very different. In such a case, the statistical factor for the creation of a doublet AB+CC would be equal to 8, leading to the following distribution of doublets (neglecting the Boltzmann factors i.e. setting them all equal to 1):

$$p_{ABAB} = \frac{2p^2}{2p^2 + 16p(1-p)}, \qquad p_{ABCC} = \frac{16p(1-p)}{2p^2 + 16p(1-p)}. \tag{4.6}$$

Now the new probability of finding an AB doublet is

$$p^{(1)} = \frac{4 - 3p}{8 - 7p}$$

and the condirion $p^{(1)} = p$ leads to the following characteristic equation:

$$7p^2 - 11p + 4 = 0, \tag{4.7}$$

leading to the following stationary solutions: $p_1 = 1, \qquad p_2 = \frac{4}{7}$, which is quite far away from the expected two-thirds ratio of AB dimers. It should be stressed here that although the relative difference between $2/3$ and $4/7$ is of about 16%, it would be occurring at each consecutive agglomeration stage, almost totally hindering the production of properly constructed capsids. This example shows how simple statistical considerations may give important hints concerning the assembly pathways of capsid construction.

5. The assembly rules for icosahedral capsids

Because 5 is a prime number, it can be divided only by 5 or by 1. This corresponds to two situations: either all sides of a pentamer are identical, or all the five sides are different. The latter case is observed in *Papovaviridae*, whose geometrical structure has been successfully resolved in a model proposed by R. Twarock and co-workers ([16]). In what follows, we shall consider the by far more frequent case, with pentamers composed of five identical dimers, so that their five edges are perfectly equivalent. They also possess a defined orientation, i.e. it is known which one of the two faces will be on the outer side of the capsid. From now on we shall assume that all the five sides of a pentamer are equivalent (identical), because any division into parts will break the symmetry.

Concerning the hexamers, as 6 is divisible by 2 and 3, one can have the following four situations:

- All 6 sides equivalent, $(aaaaaa)$
- Two types of sides, disposed as $(ababab)$
- Three types of sides, disposed as $(abcabc)$
- Six different sides, $(abcdef)$

The hexamers are also *oriented*, with one face becoming external, and the other one turned to the interior of the capsid. The three differentiated hexamers are represented in Figure 9 below.

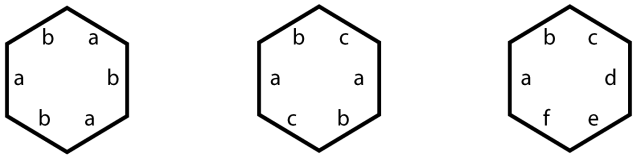

Figure 9: Three differentiations of hexamers

Let us denote pentamers' sides by symbol p, whereas two different kinds of sides on hexamers' edges will be called a and b (Figure 10). Suppose that a hexamer can stick to a pentamer with only $(p + a)$-combination; then two hexamers must stick to each other only through a $(b + b)$ combination, with both $(p + b)$ and $(a + b)$ combinations being forbidden by chemical potential barrier.

Similarly, with a more differentiate hexamer scheme, $(abcabc)$, and with the assembling rules allowing only associations of $p + a$ and $b + c$, we get with a 100% probability the $T = 4$ capsid, with $x = 2/7$, as shown in Figure 10 (right). Note that in both cases we show only one of the "basic triangles" forming the capsid, which is always made with 20 identical triangles sticking together to form a perfect icosahedral shape, as shown in Figure 4.

These examples suggest that strict association rules may exist providing precise agglomeration pathways for each kind of icosahedral capsid. Let us analyze these rules in more detail.

If the viruses were using undifferentiated hexamers with all their sides equivalent, then there would be no reason for not creating any kind of structures as shown in Figure 6 , and the final yield

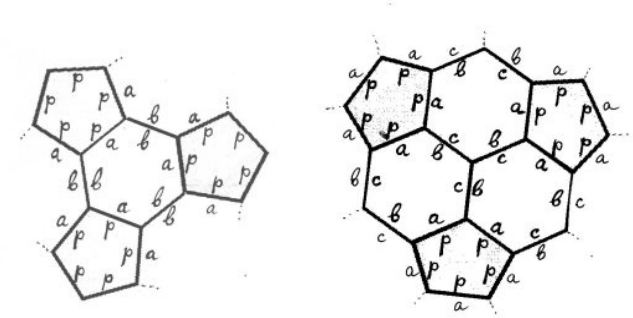

Figure 10: The building formula for the T=3 and T=4 capsids

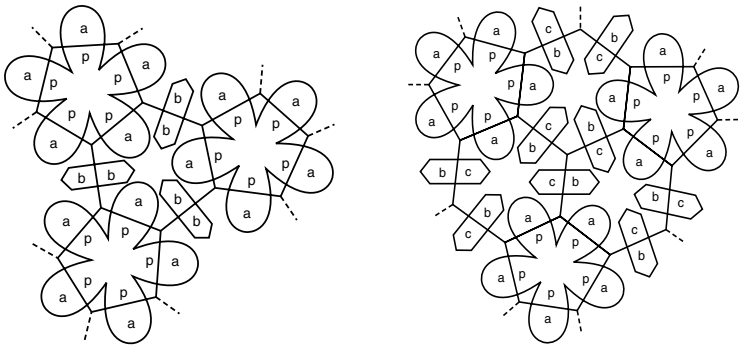

Figure 11: The T=3 and T=4 capsid's basic triangle divided in dimers and trimers

would be very low (at best like in the fullerenes, less than 10%). But with differentiated hexamers of the $(ababab)-$ type simple selection rules excluding the $(p-b)$ and (ab) associations while letting the creation of $(p-a)$ and of $(b-b)$ links, we have seen that the issue becomes determined with practically 100% certainty, as it follows from the Figure 10.

The next case presents itself when one uses the second hexamer type, with a two-fold symmetry: $(abcabc)$. Again, supposing that only a-sides can stick to pentamers' sides p, there is no other choice but the one presented in Figure 10 on the right. The corresponding affinity matrices are displayed in the Table II below:

	p	a	b
p	0	1	0
a	1	0	0
b	0	0	1

	p	a	b	c
p	0	1	0	0
a	1	0	0	0
b	0	0	0	1
c	0	0	1	0

Table II. The affinity matrices for the $T=3$ and $T=4$ capsids

Again, a "0" is put at the crossing of two symbols whose agglomeration is forbidden, and a "1" when it is allowed. By construction, a "1" can occur only once in any line or column.

Finally, let us use the most highly differentiated hexamers of the $(abcdef)$-type. Starting with pentamers surrounded by the hexamers sticking via the $(p-a)$-pairing, we discover that now two choices are possible, leading to left- and right-hand sided versions, as shown in the following Figure 12

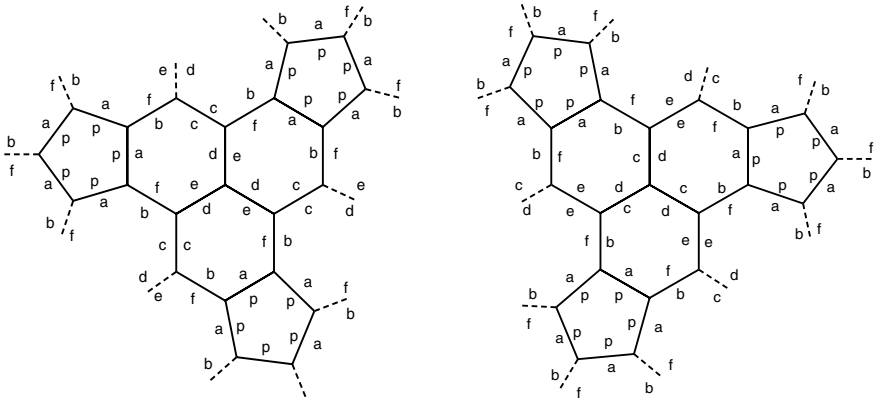

Figure 12: The building formulae for the T=7 capsids; left and right

The corresponding affinity matrices are given below:

	p	a	b	c	d	e	f
p	0	1	0	0	0	0	0
a	1	0	0	0	0	0	0
b	0	0	0	0	0	0	1
c	0	0	0	1	0	0	0
d	0	0	0	0	0	1	0
e	0	0	0	0	1	0	0
f	0	0	1	0	0	0	0

	p	a	b	c	d	e	f
p	0	1	0	0	0	0	0
a	1	0	0	0	0	0	0
b	0	0	0	0	0	0	1
c	0	0	0	0	1	0	0
d	0	0	0	1	0	0	0
e	0	0	0	0	0	1	0
f	0	0	1	0	0	0	0

Table III Affinity matrix for the $T = 7$ (left and right) capsid construction

Now a natural question can be asked: what comes next? In order to grow capsids with T-numbers greater than 7, one has to introduce new types of hexamers that would never stick to pentamers, but being able to associate themselves with certain sides of the former maximally differentiated hexamers. The result is shown in the Figure 13 below:

The corresponding affinity table is given below:

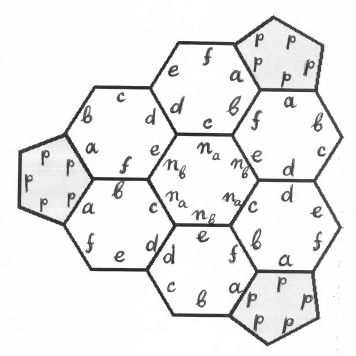

Figure 13: The building formula for the T=9 capsid

	p	a	b	c	d	e	f	n_a	n_b
p	0	1	0	0	0	0	0	0	0
a	1	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	1	0	0
c	0	0	0	0	0	0	0	1	0
d	0	0	0	0	1	0	0	0	0
e	0	0	0	0	0	0	0	0	1
f	0	0	1	0	0	0	0	0	0
n_a	0	0	0	1	0	0	0	0	0
n_b	0	0	0	0	0	1	0	0	0

Table IV: Affinity matrix for the $T = 9$ capsid construction

For bigger capsids, in which the rate of pentamers is lower, one can not obtain proper probabilities unless more than one type of hexamers is present, out of which only one is allowed to agglomerate with pentamers. In the case of two different hexamer types one obtains either the $T = 9$ capsid, or, with more exclusive sticking rules, the $T = 12$ capsid.

Finally, in order to get the $T = 25$ adenovirus capsid, one must introduce no less than *four* hexamer types, out of which only one type can agglomerate with pentamers.

The affinity matrix for $T = 12$ capsid is as follows:

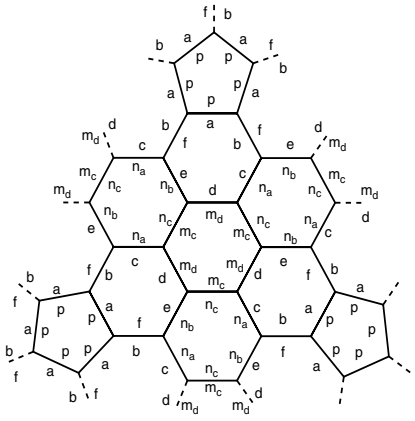

Figure 14: The T=12 capsid's basic triangle

	p	a	b	c	d	e	f	n_a	n_b	n_c	m_c	m_d
p	0	1	0	0	0	0	0	0	0	0	0	0
a	1	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	1	0	0	0	0	0
c	0	0	0	0	0	0	0	1	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0	0	1
e	0	0	0	0	0	0	0	0	1	0	0	0
f	0	0	1	0	0	0	0	0	0	0	0	0
n_a	0	0	0	1	0	0	0	0	0	0	0	0
n_b	0	0	0	0	0	1	0	0	0	0	0	0
n_c	0	0	0	0	0	0	0	0	0	0	1	0
m_c	0	0	0	0	0	0	0	0	0	1	0	0
m_d	0	0	0	0	1	0	0	0	0	0	0	0

Table V: Affinity matrix for the $T = 12$ capsid construction

Now we can organize all these results in a single table that follows. To each value of triangular number T corresponds a unique partition into $1 + (T - 1)$, where the "1" represents the unique pentamer type and $(T - 1)$ is partitioned into a sum of certain number of different hexamer types, according to the formula

$$(T - 1) = 6\,\alpha + 3\,\beta + 2\,\gamma \tag{5.1}$$

with non-negative integers α, β and γ.

Type (p,q)	$T = p^2 + pq + q^2$	$N_6 = 10(T-1)$	T decomposition
(1,1)	3	20	$1+2$
(2,0)	4	30	$1+3$
(2,1)	7	60	$1+6$
(3,0)	9	80	$1+6+2$
(2,2)	12	110	$1+6+2+3$
(3,1)	13	120	$1+6+6$
(4,0)	16	150	$1+6+6+3$
(3,2)	19	180	$1+6+6+6$
(4,1)	21	200	$1+6+6+6+2$
(5,0)	25	240	$1+(4\times 6)$
(3,3)	27	260	$1+(4\times 6)+2$
(3,3)*	27	260	$1+(3\times 6)+2\times 3+2$
(4,2)	28	270	$1+(4\times 6)+3$
(5,1)	31	300	$1+(5\times 6)$
(6,0)	36	350	$1+(5\times 6)+2+3$

Table VI: Classification of icosahedral capsids. The last column gives the number and type of hexamers needed for the construction

6. Classification of Icosahedral Capsids

Let us continue to organize all icosahedral capsids into different classes following the type of symmetry of hexamers involved in their construction. We saw already that to each value of triangular number T corresponds a unique partition into $1+(T-1)$, where the "1" represents the unique pentamer type and $(T-1)$ is partitioned into a sum of certain numbers of different hexamer types according to the formula $(T-1) = 6\,\alpha + 3\,\beta + 2\,\gamma$ (5.1) with non-negative integers α, β and γ. Now, the number γ can take on exclusively the values 0 or 1; this results from the fact that the corresponding hexamers are centered on a *three-fold* axis in the center of basic triangle; sometimes the center of such three-fold symmetry is found between three hexamers, then $\gamma = 0$. The number β corresponds to the *two-fold symmetry axis*, found in the center of an edge between elementary triangles. However, such two-fold symmetric jexamers can be found not only in the middle of an edge, but also placed on both sides of edge's center, when the T-number is sufficiently great, and when the symmetry of the capsid makes such a solution possible. The first possibility of placing *two* two-fold symmetric hexamers on the edges occurs with the capsid $T = 25,\ (p,q) = (5,0)$, as shown in the Figure 15 below. This structure is found in the quite common *adenovirus* capsid, the virus at the origin of one of the forms of "bad cold" in humans.

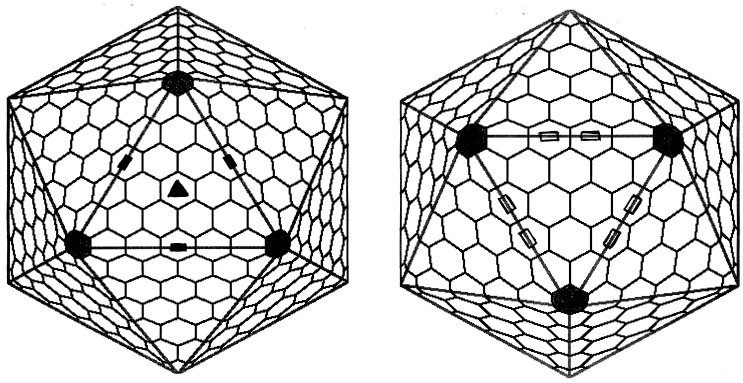

Figure 15: Left: Capsids $T = 36$ with one two-fold hexamer on the edge, and $T = 25$ with two two-fold hexamers on the edge.

The next possibility of having more than one two-fold symmetric hexamer on the edge is found in the capsid with $T = 27$, as shown in the Figure 16 below.

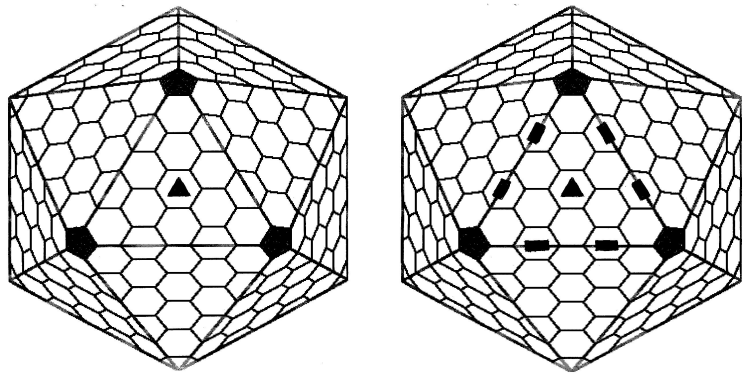

Figure 16: Left: Capsids $T = 36$ with one two-fold hexamer on the edge, and $T = 25$ with two two-fold hexamers on the edge.

The number α of maximally differentiated hexamers follows then from the corresponding partition of a given triangular number, as shown in the following table (Figure17).

It follows that all icosahedral capsids can be divided into four separate groups according to their internal symmetry, dictated by the presence of three- and two-fold centers inside the elementary triangles or on their edges. The result looks like a periodic table of capsids, arranged in four columns, according to the composition of constitutive hexamers: pure $(abcdef)$ hexamers exclusively in the first column; the $(abcdef)$-type hexamers with one $(ababab)$-type hexamer in the second column; with one $(abcabc)$-type hexamer in the third column, and finally with both, $(ababab)$ and $(abcabc)$-type in the fourth column.

Symmetric icosahedral capsids (marked in gold) correspond to specific values of T with $(p, 0)$ or (p, p); the chiral icosahedral capsids, existing in left- and right-handed versions, are marked

T and (p,q)	1 + (k×6)	1 + (k×6) + 2	1 + (k×6) + 3	1 + (k×6) + 2 + 3
1 (1,0)	1 + (0×6)			
3 (1,1)		1 + (0×6) + 2		
4 (2,0)			1 + (0×6) + 3	
7 (2,1)	1 + (1×6)			
9 (3,0)		1 + (1×6) + 2		
12 (2,2)				1 + (1×6) + 2 + 3
13 (3,1)	1 + (2×6)			
16 (4,0)			1 + (2×6) + 3	
19 (3,2)	1 + (3×6)			
21 (4,1)		1 + (3×6) + 2		
25 (5,0)	1 + (4×6)			
27 (3,3)		1 + (4×6) + 2		
28 (4,2)			1 + (4×6) + 3	
31 (5,1)	1 + (5×6)			
36 (6,0)				1 + (5×6) + 2 + 3
37 (4,3)	1 + (6×6)			
39 (5,2)		1 + (6×6) + 2		
43 (6,1)	1 + (7×6)			
48 (4,4)				1 + (7×6) + 2 + 3
49 (7,0)	1 + (8×6)			
49 (5,3)	1 + (8×6)			
52 (6,2)			1 + (8×6) + 3	
57 (7,1)		1 + (9×6) + 2		
61 (5,4)	1 + (10×6)			
63 (6,3)		1 + (10×6) + 2		
64 (8,0)			1 + (10×6) + 3	
67 (7,2)	1 + (11×6)			
73 (8,1)	1 + (12×6)			
75 (5,5)		1 + (12×6) + 2		
76 (6,4)			1 + (12×6) + 3	
79 (7,3)	1 + (13×6)			
81 (9,0)		1 + (13×6) + 2		

Figure 17: Left: Classification of icosahedral capsids up to $T = 81$.

in magenta. The statistics appearing to the naked eye in the Table (17) shows that the most differentiated capsids are the rarest. This tendency continues in the following Tables (Figure18 and 19). There are 46 capsids (including isomers) in the first column, 27 capsid species in the second column, 19 species in the third column and only 11 species in the fourth column.

Trying to define evolutionary trends from these capsid classification tables seems to be a risky endeavour; one can recall however that non-negligible information was drawn from a careful study of animal skeletons, fish scales, and similar secondary features of living organisms.

It seems plausible that the major evolution trend is from smaller towards bigger forms, as it supposes progressive differentiation among the constitutive hexamers. From a purely mathematical point of view the evolution would mean then an addition of a new hexamer type, or a transformation of one of the constitutive hexamers into another one with higher differentiation. An addition of new $(abcdef)$ hexamer to one of the capsids of any column results in one step down the same column; an addition of several new maximally differentiated hexamers results in the same number of steps down the same column.

T and (p,q)	1 + (k×6)	1 + (k×6) + 2	1 + (k×6) + 3	1 + (k×6) + 2 + 3
84 (8,2)				1 + (13×6) + 2 + 3
91 (6,5)	1 + (15 ×6)			
91 (9,1)	1 + (15 ×6)			
93 (7,4)		1 + (15×6) + 2		
96 (8,3)				1 + (15×6) + 2 + 3
100 (10,0)			1 + (16×6) + 3	
103 (9,2)	1 + (17 ×6)			
108 (6,6)				1 + (17 ×6) + 2 + 3
109 (7,5)	1 + (18 ×6)			
111 (10,1)		1 + (18×6) + 2		
112 (8,4)			1 + (18 × 6) + 3	
117 (9,3)		1 + (18×6) + 2		
121 (11,0)	1 + (20×6)			
124 (10,2)			1 + (20×6) + 3	
127 (7,6)	1 + (21×6)			
129 (8,5)		1 + (21×6) + 2		
133 (11,1)	1 + (22×6)			
133 (9,4)	1 + (22×6)			
139 (10,3)	1 + (23×6)			
144 (12,0)				1 + (23 ×6) + 2 + 3
147 (7,7)		1 + (24×6) + 2		
147 (11,2)		1 + (24×6) + 2		
148 (8,6)			1 + (24×6) + 3	
151 (9,5)	1 + (25×6)			
156 (10,4)				1 + (25×6) + 2 + 3
157 (12,1)	1 + (26×6)			
163 (11,3)	1 + (27×6)			
169 (13,0)	1 + (28×6)			
169 ((8,7)	1 + (28×6)			
171 (9,6)		1 + (28×6) + 2		
172 (12,2)			1 + (12×6) + 3	
181 (11,4)	1 + (30×6)			
183 (13,1)		1 + (30×6) + 2		
189 (12,3)		1 + (31×6) + 2		
192 (8,8)				1 + (31×6) + 2 + 3

Figure 18: The classification table, from $T = 84$ up to $T = 192$

In order to better understand the tables in Figs.17, 18 and 19 one should imagine them as cylinders, with the right edge of the right column glued to the left edge of the first column. Then one can imagine all possible transitions from any column to another one, not necessarily its immediate neighbor, but always towards the right side and down. For example, adding a new $(ababab)$-type hexamer to a capsid from the first column creates a species belonging to the second column, etc.

The important question is then, how many mutations are necessary to accomplish one of these transformations? This would give a hint as to how can one conceive the notion of "distance" between different capsids. It seems reasonable to assume that the closest species are those separated by a single addition of a maximally differentiated hexamer, which can be symbolized by the transition $(abcdef) \rightarrow (a'b'c'd'e'f')$, because such a mutation does not alter hexamer's character and can be obtained by a common modification (adding a particular radical, for example). This is why we should expect the viruses with triangulation numbers 7 and 13 to be close parents, both belonging to the first column and both skew-symmetric; we would also expect the Adenovirus ($T = 25$, $(p, q) = (5, 0)$) to be related to one of the isomers of the $T = 49$, $(p, q) = (7, 0)$ capsid,

T and (p,q)	1 + (k×6)	1 + (k×6) + 2	1 + (k×6) + 3	1 + (k×6) + 2 + 3
193 (9,7)	1 + (32×6)			
196 (14,0)			1 + (32 ×6) + 3	
196 (10,6)			1 + (32 ×6) + 3	
199 (13,2)	1 + (33×6)			
201 (11,5)		1 + (33×6) + 2		
208 (12,4)			1 + (34×6) + 3	
211 (14,1)	1 + (35 ×6)			
217 (9,8)	1 + (36×6)			
217 (13,3)	1 + (36×6)			
219 (10,7)		1 + (36 ×6) + 2		
223 (11,6)	1 + (37×6)			
225 (15,0)		1 + (37×6) + 2		
228 (14,2)				1 + (37×6) + 2 + 3
229 (12,5)	1 + (38×6)			
237 (13,4)		1 + (39×6) + 2		
241 (15,1)	1 + (40×6)			
243 (9,9)		1 + (40×6) + 2		
244 (10,8)			1 + (40×6) + 3	
247 (11,7)	1 + (41×6)			
247 (14,3)	1 + (41×6)			
252 (12,6)				1 + (41×6) + 2 + 3
256 (16,0)			1 + (42×6) + 3	
259 (13,5)	1 + (43×6)			
268 (14,4)			1 + (44×6) + 3	
271 (10,9)	1 + (45×6)			
273 (11,8)		1 + (45×6) + 2		
273 (16,1)		1 + (45×6) + 2		
277 (12,7)	1 + (46×6)			
283 (13,6)	1 + (47×6)			
289 (17,0)	1 + (48×6)			
291 (14,5)		1 + (48×6) + 2		
292 (16,2)			1 + (48×6) + 3	
301 (11,9)	1 + (50×6)			
301 (15,4)	1 + (50×6)			
303 (16,3)		1 + (50×6) + 2		
304 (12,8)			1 + (50×6) + 3	

Figure 19: The next classification table, from $T = 193$ up to $T = 304$

whose respective schemes are $25 = 1 + 4 \times 6$ and $49 = 8 \times 6$, which can be obtained by a common modification of all the four differentiated hexamers thus doubling their number. The isomer $T = 49,\ (p, q) = (5, 3)$ should be quite distant frome these two.

It is also plausible that the evolution by mutations keeps the capsid types inside the same column, with the same symmetry type, as the symmetry change $(abcdef) \rightarrow (ababab)$ requires several more specific mutations at once. This is probably why the capsid type $T = 16 = 1+2\times6+3$ in the third column (Figure 17) appearing as isolated corresponds to a single and unique family of *Herpesvirus*, which admits many mutations and variations, but remaining always inside the same capsid species and size ([11]).

Such "islands" exist also among the capsid types corresponding to higher values of T, even when one considers the Tables (17), (18) and (19) as cylinders, with their right and left borders glued together. There are single isolated species corresponding to $T = 124$; $T = 196$ and $T = 268$ (the last one beyond the range of observed types). But there are also "islands" in form of isolated groups of species containing a few neighbors in the table. As an example, one can cite the group

of five species with triangular numbers equal to $T = 52, 57, 61, 63$ and 64; there is another small isolated doublet with $T = 79$ and $T = 81$; an isolated group of four species $T = 67, 73, 75$ and 76; another isolated doublet with $T = 96$ and $T = 100$, and so forth.

It is reasonable to suppose that all such groups represent an increased stability against mutations that would force them to get out of their isolated cluster, because by definition, such evolutionary displacements need more than one mutation at once. To define a notion of a distance between various types of capsids is a challenge for further research in this direction ([18], [19], [20]).

Acknowledgments

The author is greatly indebted to Dr. Nicola Stonehouse for many enlightening discussions and for her extremely useful suggestions and remarks. Thanks are also due to Drs. Reidun Twarock and Adam Zlotnick for interesting discussions, and to Dr. Maja Nowakowski for her help in getting acquainted with current literature and careful reading of the manuscript.

References

[1] D. D. Richman, R. J. Whitley, F. G. Hayden. *Clinical Virology*. (second edition); ASM Press, Washington DC, 2009.

[2] M.C.M. Coxeter. *"Regular polytopes"*, Methuen and C, London, 1948.

[3] M. Eigen, 1971, *Selforganization of matter and the evolution of biological molecules*, Springer-Verlag, *D*ie Natutwissenschaften, 58 heft 10,

[4] H. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, R. E. Smalley. C_{60}*: Buckminsterfullerene*. Nature, 318 (1995), 162–163.

[5] D. L. D. Caspar, A. Klug, *Physical Principles in the Construction of Regular Viruses*. Cold Spring Harbor Symp. Quant. Biology, 27 (1962), No 1, 1–24.

[6] A. Zlotnick. *To Build a Virus Capsid : An Equilibrium Model of the Self Assembly of Polyhedral Protein Complexes*. J. Mol. Biology, 241, (1994), 59–67.

[7] S. B. Larson. *Refined structure of satellite tobacco mosaic virus at 1.8 A resolution*. Journal of Molecular Biology, 277 (1998), 37–59.

[8] D. J. McGeogh, A. J. Davison. *The descent of human herpesvirus*. 8.Semin. Cancer Biology, 9 (1999), 201–209.

[9] D. J. McGeogh, A. J. Davison. *The molecular evolutionary history of the herpesviruses: origins and evolution of viruses*. Academic Press Ltd., London, 1999.

[10] P. L. Stewart, R. M. Burnett, M. Cyrklaff, S. D. Fuller, *Image reconstruction reveals the complex molecular organization of adenovirus*. Cell, 67 (1991), 145–154.

[11] B. L. Trus. *Capsid structure of Kaposi's sarcoma-associated herpesvirus, a gammaherpesvirus, compared to those of an alphaherpesvirus, herpes simplex virus type 1, and a a Betaherpesvirus, Cytomegalovirus*. Journal of Virology, 75 (2001), No 6, 2879–2890.

[12] Q. Wang, T. Lin, L. Tang, J. E. Johnson, M. G. Finn. *Icosahedral Virus Particles as Addressable Nanoscale Building Blocks*. Angewandte Chemie, 114 (2002), No. 3, 477–480.

[13] H. R. Hill, N. J. Stonehouse, S. A. Fonseca, P. Stockley. *Analysis of phage MS2 coat protein mutants expressed from a reconstituted phagemid reveals that proline 78 is essential for viral infectivity*. Journal of Molecular Biology, 266, (1997), 1–7.

[14] P. E. Prevelige, D. Thomas, J. King. *Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells*. Biophys. Journal, 64 (1993), 824–835.

[15] B. Buckley, S. Silva, S. Singh. *Nucleotide sequence and in vitro expression of the capsid protein gene of tobacco ringspot virus*. Virus Research, 30 (1993), 335–349.

[16] R. Twarock. *A tiling approach to virus capsid assembly explaining a structural puzzle in virology*. Journal of Theoretical Biology, 226 (2004), No 4, 477–482.

[17] R. Kerner. *The principle of self-similarity*, in " Current Problems in Condensed Matter", ed. J. Moran-Lopez, (1998), 323–341.

[18] R. Kerner. *Model of viral capsid growth*. Journal Computational and Mathematical Methods in Medicine, 6 (2007), Issue 2, 95–97.

[19] R. Kerner. *Classification and evolutionary trends of icosahedral viral capsids*. Journal Computational and Mathematical Methods in Medicine, 9 (2008), Issue 3 & 4, 175–181.

[20] R. Kerner. *Models of Agglomeration and Glass Transition*. Imperial College Press, 2007.