Article

Not peer-reviewed version

On the Salient Regularities of Strings of Assembly Theory

Wawrzyniec Bieniawski , Piotr Masierak , Andrzej Tomski , Szymon Łukaszyk *

Posted Date: 8 November 2024

doi: 10.20944/preprints202409.1581.v5

Keywords: assembly theory; information theory; complexity measures; information entropy; mathematical physics

 doi:10.20944/preprints202409.1581.v5

Article

On the Salient Regularities of Strings of Assembly Theory

Wawrzyniec Bieniawski [1], Piotr Masierak [1], Andrzej Tomski [2] and Szymon Łukaszyk [3]

[1] Łukaszyk Patent Attorneys, ul. Głowackiego 8, 40-052 Katowice, Poland
[2] University of Silesia, Institute of Mathematics, Bankowa 14, 40-007 Katowice, Poland
[3] Łukaszyk Patent Attorneys, ul. Głowackiego 8, 40-052 Katowice, Poland
* Correspondence: szymon@patent.pl

Abstract: Using assembly theory of strings of any natural radix b we find some of their salient regularities. In particular, we show that the upper bound of the assembly index depends quantitatively on the radix b and the longest length N of a string that has the assembly index of $N-k$ is given by $N_{(N-1)} = b^2+b+1$ and by $N_{(N-k)} = b^2+b+2k$ for $2 \leq k \leq 9$. We also provide particular forms of such strings. Knowing the latter bound, we conjecture that the maximum assembly index of a string of length $N_{(N-2)} \leq N \leq N_{\max}$ is given by $a_{\max}^{(N,b)} = \lfloor N/2 \rfloor + b(b+1)/2$, where $N_{\max} = 4b^4$ if b is even and $N_{\max} = 4(b^4+1)$ otherwise. For $k=1$ such odd length strings are nearly balanced and there are four such different strings if $b=2$ and seventy-two if $b=3$. We also show that each k copies of an n-plet contained in a string decrease its assembly index at least by $k(n-1)-a$, where a is the assembly index of this n-plet. Finally, we show that the assembly depth of a minimum assembly index string is equal to the assembly index of this string, the assembly depth of a maximum assembly index string satisfies $d_{a_{\max}}^{(N,b)} \geq \lceil \log_2(N) \rceil$. Since these results are, in general, also valid for $b=1$, assembly theory subsumes information theory.

Keywords: assembly theory; information theory; complexity measures; information entropy; mathematical physics

1. Introduction

Assembly theory (AT), formulated in 2017, introduced the concept of an *initial pool* [1].

Definition 1. *We call a set* $P_0^{(b)} := \{0, 1, \ldots, b-1\}$ *that contains* $b \in \mathbb{N}$ *different basic symbols* c*, the initial assembly pool.*

The reader will find numerous results on AT in refs. [1–10], for example. Here, we extend the results of our previous study [9] concerning bitstrings to strings of any natural radix b. We consider the formation of strings $C_k^{(N,b)}$ of length N containing symbols from the initial assembly pool $P_0^{(b)}$ within the AT framework in consecutive assembly steps from basic symbols c and strings (doublets, triplets, n-plets) assembled in previous steps. The ancient Greek verb *symbállein* means putting only two *things* ("symbols") together [11].

In fact, any embodiment of AT, with basic symbols representing LEGO® blocks, chemical bonds, graphs, monomers, etc. assembled in any n-dimensional space ($n \in \mathbb{C}$) [12] corresponds to the string AT version. This is because in AT an assembly step always consists in joining two parts only, which can be thought of as the left and right fragments of the newly formed string. Put simply, AT explains and quantifies selection and evolution [7] but it is through the word (aka string or *message*), in particular a nucleotide sequence in the case of $b=4$, all AT *things* come into existence [13].

Definition 2. *We call a set* $P_s^{(b)}$ *that contains basic symbols and strings assembled in previous steps* $\{1, 2, \ldots, s-1\}$ *the working assembly pool.*

 doi:10.20944/preprints202409.1581.v5

An assembly step s may consist of

$$c_1 \circ c_2 = C_k^{(2,b)}, \quad C_l^{(N_l,b)} \circ c_2 = C_k^{(N_l+1,b)}, \quad c_1 \circ C_m^{(N_m,b)} = C_k^{(1+N_m,b)}, \quad C_l^{(N_l,b)} \circ C_m^{(N_m,b)} = C_k^{(N_l+N_m,b)}, \tag{1}$$

where $c_1, c_2 \in P_0^{(b)}$, $C_l^{(N_l,b)}, C_m^{(N_m,b)} \in P_{s-1}^{(b)}$, and $C_k \in P_s^{(b)}$. Using Definitions 1 and 2, the assembly index (ASI) of a string is the minimal achievable value of a difference between the cardinalities of the working and initial assembly pools (ASPs) leading to this string, since at each assembly step the cardinality of the working ASP increases by one. Therefore, the working ASP 2 cannot be identified with the initial ASP 1; the initial ASP 1 must not contain strings of basic symbols (see Appendix G).

2. Results

Theorems 1 and 2 were already stated in our previous study [9] for $b = 2$. We restate them here $\forall b$ for clarity.

Theorem 1. *A quadruplet is the shortest string that allows for more than one ASI for all b.*

Proof. $N = 2$ provides b^2 available doublets with unit ASI. $N = 3$ provides b^3 available triplets with ASI equal to two. Only $N = 4$ provides b^4 quadruplets that include b^2 quadruplets with ASI equal to two, that is b quadruplets $C_{k,\min}^{(4,b)} = [****]$ and $b(b-1)$ quadruplets $C_{l,\min}^{(4,b)} = [*\star**]$, while the ASI of the remaining $b^4 - b^2$ quadruplets is three. □

For example, to assemble the quadruplet $C_{k,\min}^{(4,4)} = [0202]$, we need to assemble the doublet $[02]$ and reuse it from the first step ASP P_1, while there is nothing available to reuse, in the case of the quadruplet $C_l^{(4,4)} = [0123]$.

Where the symbol value can be arbitrary, we write $*$ assuming that it is the same within the string. If we allow for the 2nd possibility different from $*$, we write $\star$. Thus, $C_k^{(2,b)} = [**]$, for example, is a placeholder for all b strings, while $C_l^{(2,b)} = [*\star]$ a placeholder for all $b(b-1)$ strings. Furthermore, we consider the degenerate case of just one basic symbol ($b = 1$).

Theorem 2. *The minimum ASI $a^{(N)}(C_{min})$ as a function of N corresponds to the shortest addition chain for N (OEIS A003313) for all b.*

Proof. Strings $C_{\min}$ for which $a^{(N)}(C_{\min}) = \min_k\left(\{a^{(N,b)}(C_k)\}\right)$, $\forall k \in \{1, 2, \ldots, b^N\}$ can be formed in subsequent steps s by joining the longest string assembled so far with itself until $N = 2^s$ is reached. Therefore, if $N = 2^s$, then $\min_k\left(\{a^{(2^s)}(C_k)\}\right) = s = \log_2(N)$. Only b^2 strings have such ASI if $N = 2^s$, including respectively b and $b(b-1)$ strings

$$C_k^{(2^s,b)} = [**\ldots], \quad C_l^{(2^s,b)} = [*\star*\star\ldots], \tag{2}$$

and the assembly pathway of each of the strings (2) is unique. At each assembly step, its length doubles.

An addition chain for $N \in \mathbb{N}$ having the shortest length $s \in \mathbb{N}$ (commonly denoted as $l(N)$) is defined as a sequence $1 = a_0 < a_1 < \cdots < a_s = N$ of integers such that $\forall\, j \geq 1$, $a_j = a_k + a_l$ for $k \leq l < j$. Thus, an addition chain starts with one, not zero, as zero is the neutral element of addition. For the same reason, two is considered the smallest prime, as one is the neutral element of multiplication. Hence, $j = 1 \implies k = l = 0$ and the first step in creating an addition chain for N is always $a_1 = 1 + 1 = 2$; the ASI of any doublet is one. The second step in creating an addition chain can be $a_2 = 1 + 1 = 2$, $a_2 = 1 + 2 = 3$, or $a_2 = 2 + 2 = 4$. The 1st case does not represent the shortest addition chain, the 2nd one corresponds to assembling a triplet based on the previously assembled doublet, and the 3rd one corresponds to assembling a quadruplet from this doublet. Therefore, four is

 doi:10.20944/preprints202409.1581.v5

the smallest number achievable in two ways since $a_2 = 2+2 = 4$ and $a_3 = 3+1 = 4$, where the latter case corresponds to assembling a quadruplet by joining a basic symbol to a triplet, which is not the shortest way for assembling a quadruplet having a minimum ASI.

Thus, finding the shortest addition chain for N corresponds to finding the ASI of a string containing basic symbols and/or doublets and/or triplets containing these doublets for $N \neq 2^s$ since due to Theorem 1 only they provide the same assembly indices $\{0,1,2\}$. □

The assembly pathways of strings $a_{\text{min}}^{(N)}$ of length $N \neq 2^s$ are not unique. For example, a string $C_{\text{min}}^{(5,b)} = [01010]$ can be assembled in three steps from three working ASPs $P_3^{(2)} = \{0,1,01,0101\}$, $P_3^{(2)} = \{0,1,10,1010\}$, and $P_3^{(2)} = \{0,1,01,010\}$.

Theorem 3. *The strings $C_{min}^{(2^s,b)}$ can contain at most two symbols if $b > 1$. Other minimum ASI strings of length $N \neq 2^s$ can contain at most three symbols if $b > 2$.*

Proof. Minimum ASI strings of length $N = 2^s$ are formed by joining the newly assembled string to itself, where a clear or mixed doublet is created in the first step. Minimum ASI strings of other lengths admit a doublet and a triplet containing this doublet and an additional basic symbol.

To formally prove the first part, we can also use mathematical induction on the assembly step s. If $s = 1$, then the minimum ASI strings $C_{\text{min}}^{(2,b)}$ are doublets of the form $[c_1c_2]$, where $c_1, c_2 \in P_0^{(b)}$. If $c_1 = c_2$, the string contains one distinct symbol, and if $c_1 \neq c_2$, the string contains two distinct symbols. In both cases, the number of distinct symbols does not exceed two. Now assume that for some $k \in \mathbb{N}$, all minimum ASI strings $C_{\text{min}}^{(2^k,b)}$ contain at most two distinct symbols. We must show that $C_{\text{min}}^{(2^{k+1},b)}$ also contains at most two distinct symbols. Consider constructing $C_{\text{min}}^{(2^{k+1},b)}$ by joining two identical minimum ASI strings $C_{\text{min}}^{(2^k,b)}$

$$C_{\text{min}}^{(2^k,b)} \circ C_{\text{min}}^{(2^k,b)} = C_{\text{min}}^{(2^{k+1},b)}, \tag{3}$$

with each other. By the inductive hypothesis, each $C_{\text{min}}^{(2^k,b)}$ contains at most two distinct symbols. Therefore, their concatenation also contains at most two distinct symbols. By induction, for all $s \in \mathbb{N}$, the minimum ASI string $C_{\text{min}}^{(2^s,b)}$ contains at most two distinct symbols.

We will now show that other minimum ASI strings of length $N \neq 2^s$ can contain at most three distinct symbols if $b > 2$. We provide the construction of minimum ASI strings with three symbols. In the first step $s = 1$, we create a doublet $[c_1c_2]$ where $c_1, c_2 \in P_0^{(b)}$ and $c_1 \neq c_2$. Next, we combine the existing doublet $[c_1c_2]$ with a new symbol $c_3 \in P_0^{(b)}$ where $c_3 \notin \{c_1, c_2\}$. This forms a triplet $[c_1c_2c_3]$, introducing a third distinct symbol and further increasing the ASI by 1. We continue assembling by joining the longest string formed so far with itself or with previously formed strings, maintaining the minimal increase in ASI.

Assume *a contrario* that there exists a minimum ASI string $C_{\text{min}}^{(N,b)}$ of length $N \neq 2^s$ that contains four or more distinct symbols. To incorporate a fourth symbol, at least one additional assembly step is required beyond what is needed for the three symbols. This additional step implies an increase in ASI, which contradicts the minimality of $C_{\text{min}}^{(N,b)}$. Thus, Theorem 3 is proven. □

The strings having non-minimum ASI can contain all symbols. For example, the string [14]

$$C_k = [01234012340123401234], \tag{4}$$

has ASI $a^{(20,5)}(C_k) = 6 = a_{\text{min}}^{(20)} + 1$ and contains all five basic symbols $P_0^{(5)} := \{0,1,2,3,4\}$.

Theorem 4. *A string containing the same three doublets has the same ASI as a string containing two pairs of the same doublets, provided that both strings have the same distributions of other repetitions and have the same lengths.*

 doi:10.20944/preprints202409.1581.v5

Proof. Without loss of generality (w.l.o.g.), consider the following two strings of the same length $N+8$ with $**\neq 01$ and the same distributions of other repetitions (if there are any other repetitions)

$$C_k = [\ldots 01 \ldots 01 \ldots 01 \cdots * \star \ldots], \quad C_l = [\ldots 01 \ldots 01 \ldots 22 \ldots 22 \ldots], \tag{5}$$

where $**\neq 01$. Creating a doublet takes one assembly step. Each appending of a doublet to an assembled string counts as another assembly step. Hence, in a general case (i.e., for strings C_k, C_l containing also other symbols), the string C_k requires six additional assembly steps, the same as the string C_l, which completes the proof. □

Theorem 5. *A string containing the same three doublets has the same ASI as a string containing the same two triplets, provided that both strings have the same distributions of other repetitions.*

Proof. W.l.o.g. consider the following two strings of the same length $N+6$ with the same distributions of other repetitions

$$C_k = [\ldots 01 \ldots 01 \ldots 01 \ldots], \quad C_l = [\ldots 010 \ldots 010 \ldots]. \tag{6}$$

Creating a triplet takes two assembly steps. Hence, in the general case, the string C_k requires four additional assembly steps, the same as the string C_l, which completes the proof. □

Theorem 6. *A string containing the same two triplets has the same ASI as a string containing two pairs of the same doublets, provided that both strings have the same distributions of other repetitions and have the same lengths.*

Proof. The proof stems from Theorems 4 and 5. □

Theorem 7. *A string containing the same two quadruplets of the minimum ASI has the same ASI as a string containing the same three triplets, provided that both strings have the same distributions of other repetitions and have the same lengths.*

Proof. W.l.o.g. consider the following two strings of the same length $N+9$ with the same distributions of other repetitions

$$C_k = [\ldots 0101 \ldots 0101 \cdots \star \ldots], \quad C_l = [\ldots 010 \ldots 010 \ldots 010 \ldots]. \tag{7}$$

Creating such a quadruplet takes two assembly steps. Hence, in a general case, the string C_k requires five additional assembly steps, the same as the string C_l, which completes the proof. □

Theorem 8. *A string containing the same two quadruplets of the maximum ASI has the same ASI as a string containing a doublet and the same two triplets based on this doublet, provided that both strings have the same distributions of other repetitions.*

Proof. W.l.o.g. consider the following two strings of the same length $N+8$ with the same distributions of other repetitions

$$C_k = [\ldots 0001 \ldots 0001 \ldots], \quad C_l = [\ldots 110 \ldots 10 \ldots 110 \ldots]. \tag{8}$$

Creating such a quadruplet takes three assembly steps. Hence, in a general case, the string C_k requires five additional assembly steps, the same as the string C_l, which completes the proof. □

 doi:10.20944/preprints202409.1581.v5

Theorem 9. *A string containing the same two doublets and the same two triplets not based on this doublet has the same ASI as a string containing a doublet and the same two triplets based on this doublet, provided that both strings have the same distributions of other repetitions and have the same lengths.*

Proof. W.l.o.g. consider the following two strings of the same length $N+10$ with the same distributions of other repetitions

$$C_k = [\dots 110 \dots 00 \dots 110 \dots 00 \dots], \quad C_l = [\dots 110 \dots 10 \dots 110 \cdots * \star \dots], \tag{9}$$

where $*\star \notin \{11, 10\}$. In a general case, the string C_k requires seven additional assembly steps, the same as the string C_l, which completes the proof. □

In general, Theorems 1-9 show that

- k copies of a doublet in a string decrease the ASI of this string at least by $k-1$;
- k copies of a triplet in a string decrease the ASI of this string at least by $2k-2$;
- k copies of a minimum ASI quadruplet in a string decrease the ASI of this string at least by $3k-2$;
- k copies of a maximum ASI quadruplet in a string decrease the ASI of this string at least by $3k-3$;

where, the phrase "at least" is meant to indicate that other repetitions, such as e.g. doublets forming multiple quadruplets, etc. can further decrease the ASI of the string. This observation allows us to state the following theorem.

Theorem 10. *Each k_r copies of an n_r-plet $C_r^{(n_r,b)}$ contained in a string $C_m^{(N,b)}$ decrease its ASI at least by $\left[k_r(n_r-1) - a^{(n_r,b)}(C_r)\right]$. That is*

$$a^{(N,b)}(C_m) \leq N - 1 - \sum_{r=1}^{R} \left[k_r(n_r-1) - a^{(n_r,b)}(C_r)\right], \tag{10}$$

where R is the total number of repeated n_r-plets.

Proof. W.l.o.g. consider the following string

$$C_m^{(N,b)} = [\dots [c_1 c_2 \dots c_n] \dots [c_1 c_2 \dots c_n] \dots], \tag{11}$$

containing two copies of an n-plet $C_l^{(n,b)} = [c_1 c_2 \dots c_n]$. The n-plet $C_l^{(n,b)}$ can be assembled in $a^{(n,b)}(C_l)$ steps and appended to the assembled string C_m in one step. Consider that the ASI of the n-plet $C_l^{(n,b)}$ is $a^{(n,b)}(C_l) = n-1$, i.e. the n-plet does not have any repetitions that can be reused. Then one copy of this n-plet - as expected - does not decrease the ASI of the string $C_m^{(N,b)}$, as $1(n-1) - (n-1) = 0$, while more copies k decrease it by $(n-1)(k-1)$. On the other hand, if $a^{(n,b)}(C_l) < n-1$ then even a single copy of this n-plet will decrease the ASI of C_m. □

For example, due to the presence of three copies of a 5-plet $[01001]$, each with $a^{(5,6)}([01001]) = 3$, in a string

$$C_k^{(24,6)} = [12|01001|21|01001|235|01001|52], \tag{12}$$

its ASI amounts to $a^{(24,6)}(C_k) = 24 - 1 - (3 \cdot (5-1) - 3) = 14$. The relation (10) provides the upper bound on ASI as it does not describe a situation in which n-plet for $n > 2$ is assembled on a doublet also present in one copy in the string. For example, the string $a^{(14,9)}([56101781014301]) = 10$, while $14 - 1 - (2(3-1) - 2) = 11$. We note that the maximum ASI decrease is provided by 2^s-plets of the minimum ASI and amounts to $k(n-1) - \log_2(n) = k(2^s - 1) - s$.

 doi:10.20944/preprints202409.1581.v5

Another quantity quantifying the complexity of a string is the assembly depth (ASD) defined [15] as

$$d_s^{(N_k,b)}(C_k) := \max\left(d^{(N_l,b)}(C_l), d^{(N_m,b)}(C_m)\right) + 1, \tag{13}$$

where $d_0^{(1,b)}(c) := 0$, and $d^{(N_l,b)}(C_l)$ and $d^{(N_m,b)}(C_m)$ are the ASDs of two substrings C_l, C_m of the string C_k that were joined in step s, where for $N \geq 4$, and if there are more assembly pathways with different depths w_j leading to a string, which happens if at least two independent assembly steps are possible, the minimum pathway depth is the ASD of this string. Hence, the ASD captures the notion of an *independent assembly step*.

Theorem 11. *If a working ASP contains strings having the same ASD they were assembled in independent assembly steps.*

Proof. W.l.o.g. assume *a contrario* that two strings C_l, C_m in the working ASP have the same ASD, i.e., $d^{(N_l,b)}(C_l) = d^{(N_m,b)}(C_m)$, but C_m was used in the assembly of C_l along with a basic symbol c. Then

$$d_s^{(N_l,b)}(C_l) = \max\left(d^{(N_m,b)}(C_m), d^{(1,b)}(c)\right) + 1 = d^{(N_m,b)}(C_m) + 1 \neq d^{(N_m,b)}(C_m), \tag{14}$$

which contradicts our assumption and completes the proof. □

In other words, if two strings C_l, C_m in the working ASP have the same ASD, their assembly pathways are unrelated to each other; by the defining equation (13) neither of them could have been used in the assembly pathway of the other.

Theorem 12. *The ASD of any minimum ASI string $C_{min}^{(N,b)}$ is equal to the ASI of this string, $d_{a_{min}}^{(N,b)} = a_{min}^{(N)}$.*

Proof. We need to show that $d_{a_{\min}}^{(N,b)} = a_{\min}^{(N)}$. While constructing the minimum ASI string, we start with a doublet and follow the shortest addition chain for N, joining this doublet with itself or with a basic symbol to form a triplet. At each assembly step, the ASD increases by one, as we join the assembled string with a string or a basic symbol from the working ASP and we cannot perform independent assembly steps. Since, by Theorem 2, the minimum ASI corresponds to the length of the shortest addition chain $l(N)$, we have

$$d_s{}^{(N,b)}\left(C_{\min}^{(N,b)}\right) = l(N) = a_{\min}^{(N)}. \tag{15}$$

This completes the proof (see Appendix F for additional comments). □

Theorems 11 and 12 show that

- the working ASP of a minimum ASI string cannot contain strings assembled in independent assembly steps,
- the working ASP of a non-minimum ASI string must contain at least two such strings, and
- the assembly pathway of a maximum ASI string will tend to maximize their number in the working ASP, and hence to minimize the possible ASD, taking into account the saturation of the working ASP, as the number of distinct n-plets in the working ASP cannot exceed b^n.

Theorem 13. *The ASD of any maximum ASI string $C_{max}^{(N,b)}$ satisfies*

$$d_{a_{max}}^{(N,b)} = \lceil \log_2(N) \rceil. \tag{16}$$

Proof. Let $d^{(N)} := d_{a_{\max}}^{(N,b)}$. For $N = 2$ we have $d^{(2)} = 1$, as we are joining basic symbols from the initial ASP. This is the base case. In an assembly tree of ASD $d^{(N)}$, the maximum number of leaves that can be

 doi:10.20944/preprints202409.1581.v5

combined is $2^{d^{(N)}}$, because at each assembly step, we join two substrings. Therefore, the maximum length $N_{\max}$ of a $C_{\max}$ string that can be assembled with ASD $d^{(N)}$ satisfies:

$$N_{\max} \leq 2^{d^{(N)}}. \tag{17}$$

This implies that

$$d^{(N)} \geq \log_2(N_{\max}), \tag{18}$$

and leads to the relation (16), since both $d^{(N)}$ and $N_{\max}$ are natural numbers and the latter does not have to be a power of two. We can also use mathematical induction. For $N \geq 2$ and for $N+1$ we have respectively

$$\begin{aligned} d^{(N)} = \lceil \log_2(N) \rceil \quad &\Longrightarrow \quad 2^{d^{(N)}-1} < N \leq 2^{d^{(N)}}, \\ d^{(N+1)} = \lceil \log_2(N+1) \rceil \quad &\Longrightarrow \quad 2^{d^{(N+1)}-1} - 1 < N \leq 2^{d^{(N+1)}} - 1, \\ \max\left(2^{d^{(N)}-1}, 2^{d^{(N+1)}-1} - 1\right) &< N \leq \min\left(2^{d^{(N)}}, 2^{d^{(N+1)}} - 1\right), \end{aligned} \tag{19}$$

where $d^{(N)} \in \mathbb{N}$ implies that either $d^{(N+1)} = d^{(N)}$ or $d^{(N+1)} = d^{(N)} + 1$. Hence,

$$\begin{aligned} d^{(N+1)} = d^{(N)} \quad &\Longrightarrow \quad 2^{d^{(N)}-1} < N \leq 2^{d^{(N)}} - 1, \\ d^{(N+1)} = d^{(N)} + 1 \quad &\Longrightarrow \quad 2^{d^{(N)}} - 1 < N \leq 2^{d^{(N)}}, \end{aligned} \tag{20}$$

which completes the proof. □

Theorems 12 and 13 are somehow counterintuitive. For example, the string $C_{\max}^{(11,2)} = [10100001110]$ has the ASI $a_{\max}^{(11,2)} = 8$ and the ASD $d_{a_{\max}}^{(11,2)} = 4$, while the string $C_{\min}^{(11,2)} = [10101010101]$ has a smaller ASI $a_{\min}^{(11)} = 5$ but a larger ASD $d_{a_{\min}}^{(11,2)} = 5$.

For example, the ASD of a string $C_{\max}^{(7,2)} = [0001110]$ is $d_{a_{\max}}^{(7,2)} = \lceil \log_2(7) \rceil = 3$ as

$$\begin{array}{llll} 00 \quad d_1 = 1, & 00 \quad w_1 = 1, & 00 \quad w_1 = 1, & 00 \quad w_1 = 1, \\ 01 \quad d_2 = 1, & 01 \quad w_2 = 1, & 01 \quad w_2 = 1, & 000 \quad w_2 = 2, \\ 11 \quad d_3 = 1, & 11 \quad w_3 = 1, & 0001 \quad w_3 = 2, & 0001 \quad w_3 = 3, \\ 110 \quad d_4 = 2, & 0001 \quad w_4 = 2, & 00011 \quad w_4 = 3, & 00011 \quad w_4 = 4, \\ 0001 \quad d_5 = 2, & 000111 \quad w_5 = 3, & 000111 \quad w_5 = 4, & 000111 \quad w_5 = 5, \\ 0001110 \quad d_6 = 3, & 0001110 \quad w_6 = 4, & 0001110 \quad w_6 = 5, & 0001110 \quad w_6 = 6, \end{array} \tag{21}$$

even though this string can be assembled with three larger pathway depths $w_6 = \{4, 5, 6\}$ and the ASD of a minimum ASI string $C_{\min}^{(7,2)} = [0101010]$ is

$$01 \quad d_1 = 1, \quad 0101 \quad d_2 = 2, \quad 010101 \quad d_3 = 3, \quad 0101010 \quad d_4 = 4. \tag{22}$$

Similarly, the ASD of a string $C_{\max}^{(8,2)} = [00011101]$ is $d_{a_{\max}}^{(8,2)} = \lceil \log_2(8) \rceil = 3$ as

$$\begin{array}{llll} 00 \quad d_1 = 1, & 00 \quad w_1 = 1, & 00 \quad w_1 = 1, & 01 \quad w_1 = 1, \\ 01 \quad d_2 = 1, & 01 \quad w_2 = 1, & 01 \quad w_2 = 1, & 001 \quad w_2 = 2, \\ 11 \quad d_3 = 1, & 11 \quad w_3 = 1, & 0001 \quad w_3 = 2, & 0001 \quad w_3 = 3, \\ 0001 \quad d_4 = 2, & 0001 \quad w_4 = 2, & 00011 \quad w_4 = 3, & 00011 \quad w_4 = 4, \\ 1101 \quad d_5 = 2, & 000111 \quad w_5 = 3, & 000111 \quad w_5 = 4, & 000111 \quad w_5 = 5, \\ 00011101 \quad d_6 = 3, & 00011101 \quad w_6 = 4, & 00011101 \quad w_6 = 5, & 00011101 \quad w_6 = 6. \end{array} \tag{23}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2024 doi:10.20944/preprints202409.1581.v5

However, the non-maximum ASI string $C_k^{(8,2)} = [01001011]$ has only two doublets that can be assembled in independent steps. Hence, its ASD cannot be decreased to $\lceil \log_2(8) \rceil$

$$\begin{array}{llll} 01 & d_1 = 1, & 01 & w_1 = 1, \\ 11 & d_2 = 1, & 010 & w_2 = 2, \\ 010 & d_3 = 2, & 010010 & w_3 = 3, \\ 010010 & d_4 = 3, & 0100101 & w_4 = 4, \\ 01001011 & d_5 = 4, & 01001011 & w_5 = 5. \end{array} \tag{24}$$

The seven-bit string is the longest string that can have the maximum ASI $a_{\max}^{(7,2)} = 7 - 1 = 6$. There are four such bitstrings containing two clear triplets and the starting bit at the end or the ending bit at the start, that is

$$[* * * * * * *] \quad \text{and} \quad [\star * * * * * \star], \tag{25}$$

and their lengths cannot be increased without a repetition of a doublet, which keeps the ASI at the same level $a_{\max}^{(8,2)} = 8 - 2 = 6$.

This observation and Theorem 2 motivated us to develop a general method to construct the longest possible string having the ASI $a_{\max}^{(N,b)}(C_{(N-1)}) = N - 1$, as a function of the radix b. We denote the length of this string by $N_{(N-1)}$ or $N_{(N-1)}(b)$, and we call this string a $C_{(N-1)}$ string.

After a few groping try-outs, we eventually reached two stable methods (cf. Appendices, Methods A and B). In both methods, we start with an initial balanced string of length $3b$ containing b clear triplets ordered as

$$[0001112 \ldots (b-2)(b-1)(b-1)(b-1)]. \tag{26}$$

The doublets that can be inserted into the initial string (26) can be arranged in a $b \times b$ matrix

$$\begin{bmatrix} \cancel{00} & \cancel{01} & 02 & \ldots & 0(b-1) \\ 10 & \cancel{11} & \cancel{12} & \ldots & 1(b-1) \\ 20 & 21 & \cancel{22} & \ldots & 2(b-1) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (b-2)0 & (b-2)1 & (b-2)2 & \ldots & \cancel{(b-2)(b-1)} \\ (b-1)0 & (b-1)1 & (b-1)2 & \ldots & \cancel{(b-1)(b-1)} \end{bmatrix}, \tag{27}$$

where the crossed out entries on a diagonal cannot be reused, as they would create repetitions in this string. If we assume that we shall not insert doublets between the clear triplets of the string (26), we can also cross out the entries in the first superdiagonal of the matrix (27). The strings of odd lengths generated by these general methods are not only the longest but also the most balanced. This can be stated in the following theorem.

Theorem 14 ($N_{(N-1)}$). *The longest length of a string that has the ASI of $N - 1$ is given by*

$$N_{(N-1)} = 3b + (b-1)^2 = b^2 + b + 1 \tag{28}$$

(OEIS A353887) and this string is nearly balanced, that is

$$N_{(N-1)} = bN_c + 1, \tag{29}$$

where $N_c = b + 1$ is the number of occurrences of all but one symbol within the string, and its Shannon entropy is

$$\begin{aligned} H(C_{(N-1)}) = -\sum_{c=0}^{b-1} p_c \log_2(p_c) &= -(b-1)\frac{N_{(N-1)}-1}{bN_{(N-1)}} \log_2\left(\frac{N_{(N-1)}-1}{bN_{(N-1)}}\right) - \frac{N_{(N-1)}-1+b}{bN_{(N-1)}} \log_2\left(\frac{N_{(N-1)}-1+b}{bN_{(N-1)}}\right) = \\ &= \frac{1-b^2}{b^2+b+1} \log_2\left(\frac{b+1}{b^2+b+1}\right) - \frac{b+2}{b^2+b+1} \log_2\left(\frac{b+2}{b^2+b+1}\right) \lesssim \log_2(b). \end{aligned} \tag{30}$$

 doi:10.20944/preprints202409.1581.v5

The proof of Theorem 14 is given in Appendix D. A $C_{(N-1)}$ string must contain all clear triplets and all doublets and if it is generated by Method A or B it is terminated with 0 and has a form

$$C_{(N-1)} = [000111222\ldots0]. \tag{31}$$

Although the case for $b = 1$ is degenerate, as no information can be conveyed using only one symbol ($H(C_{(N-1)}) = 0$ in this case), nothing precludes the assembly of such defunct strings and the formula (28) yields the correct result; the string [000] is the longest string with $a_{\max}^{(N,1)} = N - 1$ by Theorem 1, as for $b = 1$ the upper and the lower bound on the ASI are the same, $a_{\max}^{(N,1)} = a_{\min}^{(N)}$ (OEIS A003313). This is the only case where the maximum ASI is not a monotonically nondecreasing function of N.

For $b = 3$, only two doublets can be introduced without repetitions into the initial string (26), leading to twelve unique strings of length $N_{(N-1)} = 13$

$$\begin{aligned}&[000111222|0210], [000111222|1020], [20|21|000111222], [21|02|000111222], [0001112|02|22|10], [0001112|10|22|20],\\ &[21|000|20|111222], [000|20|111222|10], [02|000111222|10], [20|00|21|0111222], [21|0001112|02|22], [21|000111222|02].\end{aligned} \tag{32}$$

Finally, we have to multiply the cardinality of this set by $3! = 6$ to account for permutations. For example, the first string [0001112220210], is equivalent to five strings [0002221110120], [1110002221201], [1112220001021], [2220001112102], and [2221110002012]. Hence, there are seventy-two different strings of length $N_{(N-1)}(3) = 13$.

Subsequently, we considered other $C_{(N-k)}$ strings of length $N_{(N-k)}$ with the maximum ASI $a_{\max}(C_{(N-k)}) = N - k$ for $k > 1$.

Theorem 15 ($N_{(N-k)}$). *For all $b > 1$ and $2 \leq k \leq 9$ the longest length of a string that has the ASI of $N - k$ is given by*

$$N_{(N-k)} = b^2 + b + 2k. \tag{33}$$

The proof of Theorem 15 is given in Appendix E. This result disproves our upper bound Conjecture 1 for $b = 2$ stated in our previous study [9]. If the strings of Theorem 15 are based on strings generated by Method A or B, for $b > 2$ they owe their properties to the following distributions of symbols

$$\begin{aligned}
C_{(N-2)} &= [010000111222\ldots10\ldots0],\\
C_{(N-3)} &= [01010000111222\ldots10\ldots0],\\
C_{(N-4)} &= [0101010000111222\ldots10\ldots0],\\
C_{(N-5)} &= [010101000000111222\ldots10\ldots0],\\
C_{(N-6)} &= [0101010000011111222\ldots10\ldots0],\\
C_{(N-7)} &= [01010100000000111111222\ldots10\ldots0],\\
C_{(N-8)} &= [010101000000011011111222\ldots10\ldots0],\\
C_{(N-9)} &= [0101010010000000110111111222\ldots10\ldots0].
\end{aligned} \tag{34}$$

For the strings of the form (34) the fractions in the Shannon entropy are

$$p_0 = \frac{b+k+f_0}{b^2+b+2k}, \quad p_1 = \frac{b+k+f_1}{b^2+b+2k}, \quad p_{2,\ldots,b-1} = \frac{b+1}{b^2+b+2k}, \tag{35}$$

where $f_0 = 3$, $f_1 = -1$ if $k = 5$ and $f_0 = 2$, $f_1 = 0$ otherwise, as [00] is inserted into $C_{(N-5)}$, [11] into $C_{(N-6)}$ and [01] or [10] otherwise. This leads to Shannon entropy

$$H(C_{(N-k)}) = -\frac{b^2-b-2}{b^2+b+2k}\log_2\left(\frac{b+1}{b^2+b+2k}\right) - \frac{b+k+f_1}{b^2+b+2k}\log_2\left(\frac{b+k+f_1}{b^2+b+2k}\right) - \frac{b+k+f_0}{b^2+b+2k}\log_2\left(\frac{b+k+f_0}{b^2+b+2k}\right). \tag{36}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2024 doi:10.20944/preprints202409.1581.v5

The entropies (30) and (36) are shown in Figure 1. Radix $b = 4$ is the smallest one at which the entropy (36) is a monotonically decreasing function. For $b \in \{2,3\}$ there is a local entropy minimum for $k = 5$ and for $b = 2$ an additional local entropy minimum for $k = 2$.

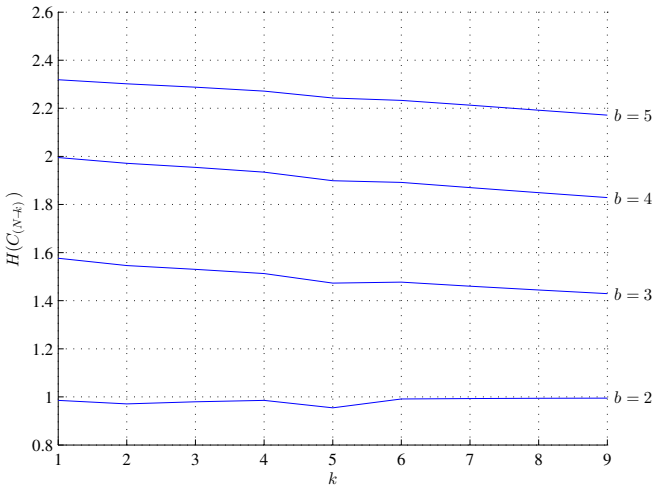

Figure 1. Shannon entropies $H(C_{(N-k)})$ for $1 \leq k \leq 9$ and $2 \leq b \leq 5$.

Conjecture 16 ($N_{\max} > N_{(N-k)}$)**.** *If $b > 1$ and $N_{(N-2)} \leq N \leq N_{max}$ then*

$$a_{max}^{(N,b)} = \begin{cases} a_{max}^{(N-1,b)} + 1 & \text{iff} \quad N = 2l, \\ a_{max}^{(N-1,b)} & \text{iff} \quad N = 2l+1, \end{cases} . \tag{37}$$

or equivalently

$$a_{max}^{(N,b)} = \left\lfloor \frac{N}{2} \right\rfloor + \frac{b(b+1)}{2}, \tag{38}$$

where

$$N_{max} = \begin{cases} 4b^4 & \text{iff} \quad b = 2l, \\ 4(b^4+1) & \text{iff} \quad b = 2l+1, \end{cases} . \tag{39}$$

In other words, if $N \geq N_{(N-2)}$, then ASI increases by one, where N increases by two ($b(b+1)/2$ are triangular numbers, OEIS A000217).

First, we note that maximum ASI must rise. If it were constant for $N > \hat{N}_{max}$, then at some even larger N it would inevitably become lower than the minimum ASI bound 2 which also rises, and this would be a contradiction. W.l.o.g. we aim to prove this conjecture for $b = 2$. We note that inserting any doublet into a $C_{(N-3)}^{(12,2)}$ string (A19) at any position creates a triplet. Using the equation (10) of Theorem 10 we have

 doi:10.20944/preprints202409.1581.v5

$$
\begin{aligned}
a_s &= a_{s-2} + 1, \quad N_s = N_{s-2} + 2, \\
a_s &= N_s - 1 - \sum_{r=1}^{R_r} \left[k_r(n_r - 1) - a\left(C_r^{(n_r,b)}\right) \right], \\
a_{s-2} &= N_{s-2} - 1 - \sum_{p=1}^{R_{s-2}} \left[k_p(n_p - 1) - a\left(C_p^{(n_p,b)}\right) \right], \\
a_s - a_{s-2} &= (N_{s-2} + 2) - 1 - \sum_{r=1}^{R_r} \left[k_r(n_r - 1) - a\left(C_r^{(n_r,b)}\right) \right] - \left(N_{s-2} - 1 - \sum_{p=1}^{R_p} \left[k_p(n_p - 1) - a\left(C_p^{(n_p,b)}\right) \right] \right) = \\
&= 2 - \sum_{r=1}^{R_r} \left[k_r(n_r - 1) - a\left(C_r^{(n_r,b)}\right) \right] + \sum_{p=1}^{R_p} \left[k_p(n_p - 1) - a\left(C_p^{(n_p,b)}\right) \right] = 1, \\
&\sum_{r=1}^{R_r} \left[k_r(n_r - 1) - a\left(C_r^{(n_r,b)}\right) \right] = \sum_{p=1}^{R_p} \left[k_p(n_p - 1) - a\left(C_p^{(n_p,b)}\right) \right] + 1,
\end{aligned}
\tag{40}
$$

for any step s if only $N_{(N-2)} \leq N_s \leq N_{\max}$. Now, assume that $\forall r$, $a\left(C_r^{(n_r,b)}\right) = n_r - 1$ and $\forall p$, $a\left(C_p^{(n_p,b)}\right) = n_p - 1$. Then

$$
\begin{aligned}
\sum_{r=1}^{R_r} [(k_r - 1)(n_r - 1)] &= \sum_{p=1}^{R_p} [(k_p - 1)(n_p - 1)] + 1, \\
\sum_{r=1}^{R_r} n_r k_r - \sum_{r=1}^{R_r} n_r - \sum_{r=1}^{R_r} k_r + R_r &= \sum_{p=1}^{R_p} n_p k_p - \sum_{p=1}^{R_p} n_p - \sum_{p=1}^{R_p} k_p + R_p + 1.
\end{aligned}
\tag{41}
$$

The proof of the Conjecture 16 must show the conditions for the equations (40) and (41) to hold. We note that the assumption used in the equation (41) is valid only for $n_r \leq N_{(N-1)}$ and $n_p \leq N_{(N-1)}$. The bounds of Theorems 14 and 15 and Conjecture 16 are illustrated in Figure 2.

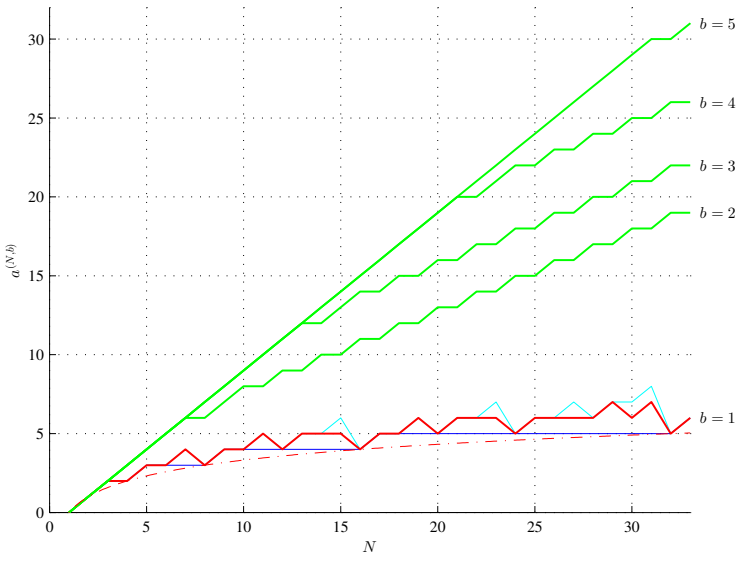

Figure 2. Lower assembly index bound (red) and upper bounds (green) for $1 \leq b \leq 4$, lower assembly depth bound (blue) of $C_{\max}^{(N,b)}$ strings for $b > 1$, $\log_2(N)$ (red, dash-dot), and OEIS A014701 sequence (cyan) for $0 < N \leq 33$.

The results thus far led us to a simple method of determining the ASI of a maximum ASI and a minimum ASD string and strengthened our Conjectures 3 and 4 stated in the previous study [9]. The method is based on unique 2^s-plets and powers of two, as shown in Table 1. First, a maximum ASI

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2024 doi:10.20944/preprints202409.1581.v5

string is sequenced, every two symbols to find the number n_{UAD} of unique adjoining doublets $\times 2_{(b)}$. In particular, a $C_{(N-1)}$ string (A3) or (A4) contain the maximum of $\left\lfloor N_{(N-1)}/2 \right\rfloor$ unique adjoining doublets, a $C_{(N-2)}$ string (A13) contains the maximum of $N_{(N-2)}/2-1$ unique adjoining doublets, and so on. In general, a $C_{(N-k)}$ string contains the maximum of

$$n_{UAD} = \left\lfloor \frac{N_{(N-k)}}{2} \right\rfloor - k + 1 = \begin{cases} b(b+1)/2 = \sum_{l=1}^{b} l & \text{iff } k = 1, \\ b(b+1)/2 + 1 = \sum_{l=1}^{b} l + 1 & \text{iff } k \neq 1, \end{cases} . \tag{42}$$

unique adjoining doublets, where $N_{(N-k)}$ is given by the relations (28) or (33), which is independent of k.

Table 1. Distributions of n-plets in strings of maximum ASI.

N	$\times 2_{(b=1)}$	$\times 2_{(b=2)}$	$\times 2_{(b=3)}$	$\times 2_{(b=4)}$	$\times 4_{(b)}$	$\times 8_{(b)}$	$\times 16_{(b)}$	$\times 32_{(b)}$	last $\times 8$	last $\times 4$	last $\times 2$	last $\times 1$	$a_{\max}^{(N,1)}$	$a_{\max}^{(N,2)}$	$a_{\max}^{(N,3)}$	$a_{\max}^{(N,4)}$
1	0	0	0	0	0	0	0	0	N	N	N		0	0	0	0
2	1	1	1	1	0	0	0	0	N	N		N	1	1	1	1
3	1	1	1	1	0	0	0	0	N	N		Y	2	2	2	2
4	1	2	2	2	1	0	0	0	N		N	N	2	3	3	3
5	1	2	2	2	1	0	0	0	N		N	Y	3	4	4	4
6	1	3	3	3	1	0	0	0	N		Y	N	3	5	5	5
7	1	3	3	3	1	0	0	0	N		Y	Y	4	6	6	6
8	1	3	4	4	2	1	0	0		N	N	N	3	6	7	7
9	1	3	4	4	2	1	0	0		N	N	Y	4	7	8	8
10	1	4	5	5	2	1	0	0		N	Y	N	4	8	9	9
11	1	3	5	5	2	1	0	0		N	Y	Y	5	8	10	10
12	1	4	6	6	3	1	0	0		Y	N	N	4	9	11	11
13	1	3	6	6	3	1	0	0		Y	N	Y	5	9	12	12
14	1	4	6	7	3	1	0	0		Y	Y	N	5	10	12	13
15	1	3	6	7	3	1	0	0		Y	Y	Y	6	10	13	14
16	1	4	7	8	4	2	1	0	N	N	N	N	4	11	14	15
17	1	3	6	8	4	2	1	0	N	N	N	Y	5	11	14	16
18	1	4	7	9	4	2	1	0	N	N	Y	N	5	12	15	17
19	1	3	6	9	4	2	1	0	N	N	Y	Y	6	12	15	18
20	1	4	7	10	5	2	1	0	N	Y	N	N	5	13	16	19
21	1	3	6	10	5	2	1	0	N	Y	N	Y	6	13	16	20
22	1	4	7	10	5	2	1	0	N	Y	Y	N	6	14	17	20
23	1	3	6	10	5	2	1	0	N	Y	Y	Y	7	14	17	21
24	1	4	7	11	6	3	1	0	Y	N	N	N	5	15	18	22
25	1	3	6	10	6	3	1	0	Y	N	N	Y	6	15	18	22
26	1	4	7	11	6	3	1	0	Y	N	Y	N	6	16	19	23
27	1	3	6	10	6	3	1	0	Y	N	Y	Y	7	16	19	23
28	1	4	7	11	7	3	1	0	Y	Y	N	N	6	17	20	24
29	1	3	6	10	7	3	1	0	Y	Y	N	Y	7	17	20	24
30	1	4	7	11	7	3	1	0	Y	Y	Y	N	7	18	21	25
31	1	3	6	11	7	3	1	0	Y	Y	Y	Y	8	18	21	25
32	1	4	7	11	8	4	2	1	N	N	N	N	5	19	22	26
33	1	3	6	11	8	4	2	1	N	N	N	Y	6	19	22	26

Subsequently, these doublets form $\times 4_{(b)}$ unique adjoining quadruplets, quadruplets form $\times 8_{(b)}$ unique adjoining octuples, and so on depending on the length of the string N and the radix b, as there can be at most b^{2^s} unique 2^s-plets. The columns "last 2^s" indicate if the assembled string should be terminated with a single substring of length 2^s in descending order. The empty fields in the respective columns for $N > 1$ indicate that a given $\times 2^s$ substring can be interpreted as either a "regular" single $\times 2^s$ substring or a last $\times 2^s$ substring if $\times 2^s = 1$.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2024 doi:10.20944/preprints202409.1581.v5

For example, the $N_{(N-3)}$ string (A20) of length $N_{(N-3)} = 18$ for $b = 3$ can be assembled as

$$\begin{aligned}
0 \circ 1 = [01], 0 \circ 0 = [00], 1 \circ 1 = [11], 1 \circ 2 = [12], & \\
2 \circ 2 = [22], 1 \circ 0 = [10], 2 \circ 0 = [20] \quad & (\times 2_{(b=3)} = 7), \\
[01] \circ [01] = [0101], [00] \circ [00] = [0000], [11] \circ [12] = [1112], [22] \circ [10] = [2210] \quad & (\times 4 = 4), \\
[0101] \circ [0000] = [01010000], [1112] \circ [2210] = [11122210] \quad & (\times 8 = 2), \\
[01010000] \circ [11122210] = [0101000011122210] \quad & (\times 16 = 1), \\
[0101000011122210] \circ [20] = [010100001112221020] \quad & (\text{last} \times 2), \\
\hline
7 + 4 + 2 + 1 + 1 = 15 \text{ steps.} &
\end{aligned} \tag{43}$$

Similarly, the $N_{(N-1)}$ string (A3) of length $N_{(N-1)} = 21$ for $b = 4$ can be assembled, as shown in Table 1 as

$$\begin{aligned}
0 \circ 0 = [00], 0 \circ 1 = [01], 1 \circ 1 = [11], 2 \circ 2 = [22], 2 \circ 3 = [23], & \\
3 \circ 3 = [33], 1 \circ 0 = [10], 2 \circ 1 = [21], 3 \circ 2 = [32], 0 \circ 3 = [03] \quad & (\times 2_{(b=4)} = 10), \\
[00] \circ [01] = [0001], [11] \circ [22] = [1122], [23] \circ [33] = [2333], & \\
[10] \circ [21] = [1021], [32] \circ [03] = [3203] \quad & (\times 4 = 5), \\
[0001] \circ [1122] = [00011122], [2333] \circ [1021], [23331021] \quad & (\times 8 = 2), \\
[00011122] \circ [23331021] = [0001112223331021] \quad & (\times 16 = 1), \\
[0001112223331021] \circ [3203] = [00011122233310213203] \quad & (\text{last} \times 4), \\
[00011122233310213203] \circ 0 = [000111222333102132030] \quad & (\text{last} \times 1), \\
\hline
10 + 5 + 2 + 1 + 1 + 1 = 20 \text{ steps.} &
\end{aligned} \tag{44}$$

For $N < 15$ and for other small N this combinatorics is valid also for $b = 1$, where obviously $\max(\times 2^s) = 1$. For example, the string of length $N = 15$ can be assembled in six steps as

$$\begin{aligned}
0 \circ 0 = [00], \quad & (\times 2_{(b=1)} = 1), \\
[00] \circ [00] = [0000] \quad & (\times 4_{(b=1)} = 1), \\
[0000] \circ [0000] = [00000000] \quad & (\times 8_{(b=1)} = 1), \\
[00000000] \circ [0000] = [000000000000] \quad & (\text{last} \times 4), \\
[000000000000] \circ [00] = [00000000000000] \quad & (\text{last} \times 2), \\
[00000000000000] \circ [0] = [000000000000000] \quad & (\text{last} \times 1), \\
\hline
1 + 1 + 1 + 1 + 1 + 1 = 6 \text{ steps.} &
\end{aligned} \tag{45}$$

However, this is the 1st exception for $b = 1$ as the ASI of this string is five if it is assembled using doublet [00] and triplet [000]. For $b = 1$ the method produces OEIS A014701 sequence corresponding to the number of steps to reach 1 starting from N_0 and assigning $N_{s+1} = N_s - 1$ if N_s is odd and $N_{s+1} = N_s/2$ otherwise.

We further note that the method illustrated in Table 1 cannot be used to construct the maximum ASI string. For example, both the following two distributions of doublets for $N = 6$ satisfy the

 doi:10.20944/preprints202409.1581.v5

distributions of Table 1. However, only the left one correctly reflects the maximum ASI of the assembled string.

$$\begin{array}{rr} 0\circ 0=[00], 0\circ 1=[01], 1\circ 1=[11] & (\times 2_{(b=2)}=3),\\ [00]\circ[01]=[0001] & (\times 4=1),\\ [0001]\circ[11]=[000111] & (\text{last}\times 2),\\ 3+1+1= & \underline{\quad 5 \text{ steps},} \end{array} \qquad \begin{array}{rr} 0\circ 0=[00], 1\circ 0=[10], 1\circ 1=[11] & (\times 2_{(b=2)}=3),\\ [00]\circ[10]=[0010] & (\times 4=1),\\ [0010]\circ[11]=[001011] & (\text{last}\times 2),\\ 3+1+1= & \underline{\quad 5\neq 4 \text{ steps},} \end{array} \tag{46}$$

as the right one can be assembled in four steps with $P_4^{(2)}=\{0,1,01,\dots\}$. Similarly, only the top distribution of doublets below correctly reflects the maximum ASI of the assembled string for $N=10$

$$\begin{array}{rr} 0\circ 1=[01], 0\circ 0=[00], 1\circ 1=[11], 1\circ 0=[10] & (\times 2_{(b=2)}=4),\\ [01]\circ[00]=[0100], [00]\circ[11]=[0011] & (\times 4=2),\\ [0100]\circ[0011]=[01000011] & (\times 8=1),\\ [01000011]\circ[10]=[0100001110] & (\text{last}\times 2),\\ 4+2+1+1= & \underline{\quad 8 \text{ steps},} \end{array} \tag{47}$$

$$\begin{array}{rr} 0\circ 0=[00], 0\circ 1=[01], 1\circ 0=[10], 1\circ 1=[11] & (\times 2_{(b=2)}=4),\\ [00]\circ[01]=[0001], [10]\circ[11]=[1011] & (\times 4=2),\\ [0001]\circ[1011]=[00011011] & (\times 8=1),\\ [0001011]\circ[11]=[0001101111] & (\text{last}\times 2),\\ 4+2+1+1 & \underline{\quad 8\neq 6 \text{ steps},} \end{array}$$

as the bottom one can be assembled in six steps with $P_6^{(2)}=\{0,1,11,011,\dots\}$. Furthermore, this method tends to exaggerate the estimated maximum ASI value, that is,

$$a_{\max}^{(N,b)} \leq a_{\text{method}}^{(N,b)}(C_k), \tag{48}$$

where $a_{\text{method}}^{(N,b)}$ is the ASI of a string C_k determined by the method illustrated in Table 1. For example, the first six strings below contain four unique doublets instead of the required three. Therefore

$$\begin{aligned} C_1&=[00|10|01|11], \quad a^{(8,2)}(C_1)=5, \quad a_{\text{method}}^{(8,2)}(C_1)=7,\\ C_2&=[00|10|11|01], \quad a^{(8,2)}(C_2)=5, \quad a_{\text{method}}^{(8,2)}(C_2)=7,\\ C_3&=[00|01|10|11], \quad a^{(8,2)}(C_3)=5, \quad a_{\text{method}}^{(8,2)}(C_3)=7,\\ C_4&=[00|01|11|10], \quad a_{\max}^{(8,2)}(C_4)=6, \quad a_{\text{method}}^{(8,2)}(C_4)=7,\\ C_5&=[00|11|10|01], \quad a^{(8,2)}(C_5)=5, \quad a_{\text{method}}^{(8,2)}(C_5)=7,\\ C_6&=[00|11|01|10], \quad a^{(8,2)}(C_6)=5, \quad a_{\text{method}}^{(8,2)}(C_6)=7,\\ \\ C_7&=[00|01|11|00], \quad a_{\max}^{(8,2)}(C_7)=6= \quad a_{\text{method}}^{(8,2)}(C_7)=6. \end{aligned} \tag{49}$$

Further research should consider researching the formula equivalent to (28) that captures a quadruplet repetition, similarly as $b^2+b^1+b^0$ captures a doublet repetition.

3. Discussion

Applications of AT seem to be promising. It offers a new lens for studying the construction of biological molecules like DNA and proteins. By analyzing the steps needed to assemble these molecules from basic building blocks, researchers can gain deeper insights into the evolutionary constraints and optimizations that shape biological pathways. This perspective also sheds light on the efficient construction of cellular structures and helps to identify the minimum number of assembly steps that define biological complexity, reinforcing the idea that life is characterized by highly

 doi:10.20944/preprints202409.1581.v5

organized pathways. Furthermore, AT provides an essential tool for understanding the growth of complexity in biological systems over evolutionary time. By quantifying the assembly steps required to form increasingly complex organisms, scientists can map the trajectory of evolutionary development and identify key transitions that lead to higher levels of structural and functional complexity. It can guide the design and optimization of synthetic biological systems by minimizing the number of steps required to build new biological pathways, making bioengineering more efficient and scalable. The ability to model and simplify complex biological processes using AT could lead to the development of more robust and adaptable synthetic organisms.

Strings having lengths $N_{(N-1)}$ (e.g. (A3) or (A4)) are necessarily the most balanced: all but one symbol occur $b+1$ times and one symbol occurs $b+2$ times within a string $C_{(N-1)}$. However, if the length of a string is constant, it will tend to evolve to decrease the Shannon entropy [16,17] and, hence, to become less balanced. As the energy of a black hole that can be thought of as a balanced bitstring [18] can be two times the energy of the entropy variation sphere that it generates [19], this tendency to imbalance seems to be associated with the minimum energy condition. For example, the Shannon entropy of the SARS-CoV genome containing $N = 29903$ nucleobases decreased from $H = 1.3565$ to 1.3562 within two years after the Wuhan outbreak [9,16]. The minimum ASI for this length of the string, given by the OEIS A003313, is $a_{\min}^{(29903)} = 19$. Perhaps, entropy (36) has other local entropy minima for $b < 4$ and for $k > 9$ and is a monotonically decreasing function only for $b \geq 4$. This could be the reason nature has chosen the non-binary radix $b = 4$ and four nucleobases to encode genetic information.

Author Contributions: WB: first concept of a general method for constructing the string of length $N_{(N-1)}$ leading to Theorem 14; the concept of the doublet matrix (27); outline of the general Method A; proposition of Theorem 9; a string with exactly two copies of all doublets idea and the formula for its length; numerous clarity corrections and improvements; PM: outline of the general Method B; the hint for ASI combinatorics; creation of a software supporting Conjecture 16; creation of a string $C_{\max}^{(24,2)}$; numerous clarity corrections and improvements; AT: formal proof of Theorem 3; proof that the Shannon entropy (30) can be approximated by $\log_2(b)$ for large b; proof of the Theorem 12; conceptualization of the proof of the Theorem 13 and equation (17); the 1st paragraph of the discussion section 3; numerous clarity corrections and improvements; SŁ: The remaining part of the study.

Funding: This research received no external funding.

Data Availability Statement: The public repository for the code written in the MATLAB computational environment and C++ is given under the link https://github.com/szluk/Evolution_of_Information (accessed on 19 September 2024).

Acknowledgments: The authors thank Mariola Bala for her motivation and Rafał Winiarski for noting that the relation (10) is inequality. SŁ thanks his wife, Magdalena Bartocha, for her everlasting support, and his partner and friend, Renata Sobajda, for her prayers.

Conflicts of Interest: Authors Wawrzyniec Bieniawski and Piotr Masierak were employed by the company Łukaszyk Patent Attorneys. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Method A for Generating $C_{(N-1)}$ String

We start with a string of clear triplets (26). In the 1st step, we create a string containing doublets on the first subdiagonal of the matrix (27) starting with 10

$$[102132\ldots(b-2)(b-3)(b-1)(b-2)], \tag{A1}$$

and we append it to the string (26). With this step, we also eliminate the doublets on the second superdiagonal starting with the doublet 02, as well as the doublet $(b-1)1$. In the 2nd step, we create a string containing doublets on the third superdiagonal beginning with the doublet 03

$$[0314\ldots(b-5)(b-2)(b-4)(b-1)], \tag{A2}$$

and append it to the string created so far. With this step, we also remove the doublet $(b-2)0$ and the middle part of the second subdiagonal containing $\{31, 42, \ldots, (b-2)(b-4)\}$. And so on. Finally,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2024 doi:10.20944/preprints202409.1581.v5

we append 0 if b is even. This process is illustrated in Figure A1 and for $3 \leq b \leq 13$ generates the following $C_{(N-1)}$ strings

$$
\begin{aligned}
&[000111222|10|20],\\
&[000111222333|102132|03|0],\\
&[000111222333444|10213243|0314|20|40],\\
&[0001112223334445555|1021324354|031425|0415|2053|0],\\
&[00011122233344455566 6|102132435465|03142536|041526|2064|0516|30],\\
&[000111222333444555666777|10213243546576|0314253647|04152637|2075|051627|306174|0],\\
&[\ldots|1021324354657687|031425364758|0415263748|2086|05162738|30617285|0718|40],\\
&[\ldots|1021324354657687 98|031425364758 69 7a|0415263748 59 6a|2086|05162738|30617285|0718|40], \qquad \text{(A3)}\\
&[\ldots|10213243546576 98|0314253647 58|0415263748 59 6a|2086|05162738|30617285|0718|40],\\
&[\ldots|1021324354657687 98 ba|031425364758 69 7a 8b|0415263748 59 6a 7b|2086|05162738|30617285|0718|40],\\
&[\ldots|10213243546576 98 ba|0314253647 58 8b|0415263748 59 6a 7b|2086|05162738|30617285|0718|40],\\
&[\ldots|1021324354657687 98 ba cb|031425364758 69 7a 8b 9c|0415263748 59 6a 7b 8c|2086|05162738|30617285|0718|40],\\
&[\ldots|10213243546576 98 ba cb|0314253647 58 8b 9c|0415263748 59 6a 7b 8c|2086|05162738|30617285|0718|40].
\end{aligned}
$$

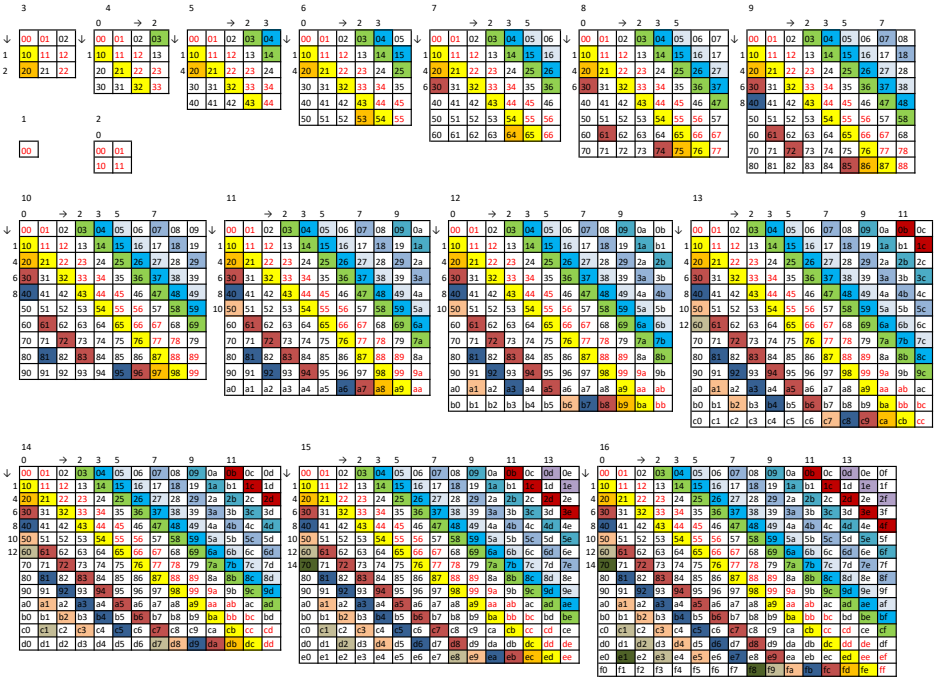

Figure A1. Doublet matrices for $1 \leq b \leq 16$ that illustrate the generation of $N_{(N-1)}$ strings according to Method A. Colored doublets are appended to the initial string of clear triplets in the order indicated by arrows starting from the 1st column or row. Finally, 0 is appended at the end, if b is even.

Appendix B. Method B for Generating $C_{(N-1)}$ String

This method is similar to the Method A. We also start with a string of clear triplets (26) and the matrix of doublets (27) with a crossed diagonal and the first superdiagonal. In the first step, we append the doublet $0(b-1)$ (top right doublet of the matrix of doublets (27)) at the end of the string (26). Next, we generally perform the following pairs of iterations:

 doi:10.20944/preprints202409.1581.v5

1. we check subsequent subdiagonals until we find one that does not contain a doublet present in the string created so far, we append it at the end of this string and proceed to step 2;
2. we check subsequent superdiagonals until we find one that does not contain a doublet present in the string created so far, we append it at the end of this string and proceed to step 1.

Finally, we append 0 if b is even. The method is illustrated in Figure A2 and for $3 \leq b \leq 13$ generates the $C_{(N-1)}$ strings in the form

$$
\begin{aligned}
&[000111222|0210],\\
&[000111222333|03|102132|0],\\
&[000111222333444|04|10213243|0314|20],\\
&[0001112223334445 55|05|1021324354|031425|304152|0],\\
&[00011122233344455566|06|102132435465|03142536|405162|041526|30],\\
&[000111222333444555666777|07|10213243546576|0314253647|3041526374|051627|506172|0],\\
&[\ldots|08|102132435465768 7|031425364758|304152637485|05162738|607182|061728|40],\\
&[\ldots|09|1021324354657687 98|03142536475869|30415263748596|0516273849|5061728394|071829|708192|0],\\
&[\ldots|0a|10213243546576879 8a9|031425364758697a|30415263748596a7|05162738495a|\\
&60718293a4|061728394a|8091a2|08192a|50],\\
&[\ldots|0b|10213243546576879 8a9ba|031425364758697a8b|30415263748596a7b8|05162738495a6b|\\
&5061728394a5b6|0718293a4b|708192a3b4|091a2b|90a1b2|0],\\
&[\ldots|0c|10213243546576879 8a9bacb|031425364758697a8b9c|30415263748596a7b8c9|05162738495a6b7c|\\
&5061728394a5b6c7|0718293a4b5c|8091a2b3c4|08192a3b4c|a0b1c2|0a1b2c|60].
\end{aligned}
\tag{A4}
$$

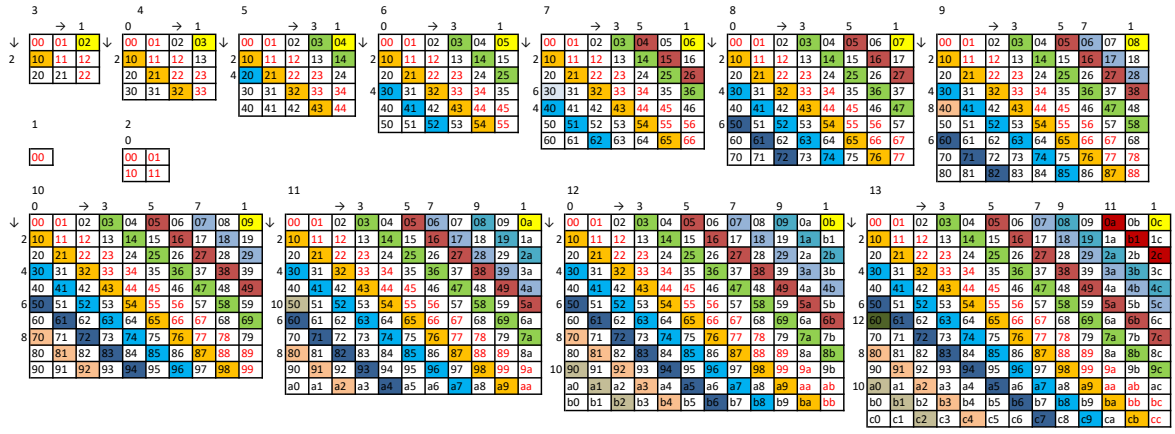

Figure A2. Doublet matrices for $1 \leq b \leq 13$ that illustrate the generation of $N_{(N-1)}$ strings according to Method B. Colored doublets are appended to the initial string of clear triplets in the order indicated by arrows starting from the 1st column or row. Finally, 0 is appended at the end, if b is even.

Appendix C. A String with Exactly Two Copies of All Doublets and No Repeated Triplets

A string that has exactly two copies of all doublets and no repeated triplets can have a form (for $b = \{1, 2, 3, 4, 5\}$)

$$
\begin{aligned}
&[0000]\\
&[00001111|010]\\
&[000011112222|1021|202010]\\
&[0000111122223333|102132|101202303203130]\\
&[00001111222233334444|10213243|10120230340413042421432031 40]
\end{aligned}
\tag{A5}
$$

and has a length of

$$N_{2D} = 2b^2 + b + 1. \tag{A6}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2024 doi:10.20944/preprints202409.1581.v5

A suboptimal method for its generating (with repeated triplets) is illustrated in Figure A3.

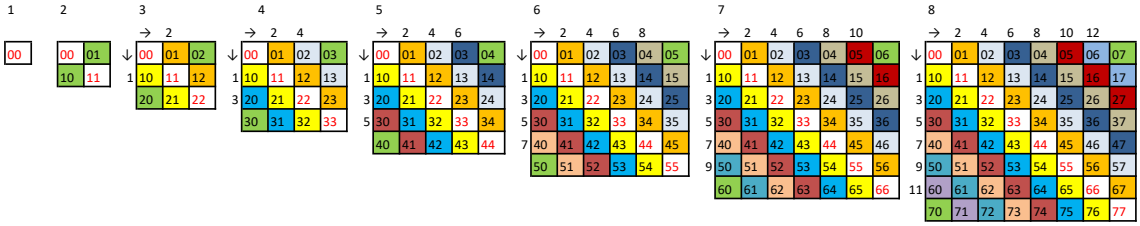

Figure A3. Doublet matrices for $1 \leq b \leq 8$ that illustrate the generation of N_{2D} strings containing exactly two copies of all doublets. Colored doublets are appended to the initial string of clear quadruplets in the order indicated by arrows starting from the 1st column or row. Finally, $0(b-1)0$ is appended at the end. The 1st superdiagonal is appended as 01234 . . . .

Appendix D. Proof of $C_{(N-1)}$ String Theorem

The $N_{(N-1)}$ given by the formula (28) is an odd number for all b. The first element $3b$ is the length of the initial string (26) containing b clear triplets and $b^2 - b - (b-1)$ is the number of doublets available in the matrix (27) after crossing out b doublets on its diagonal and $b-1$ doublets on its superdiagonal that are present in the starting string (26). By definition, a $C_{(N-1)}$ string cannot have any repetitions. To be the longest, it must contain all doublets in the matrix (27) and all clear triplets. Furthermore, to be the most patternless, this string must maximize Shannon entropy; must be the most balanced. For the string of the form (29) the fractions in the Shannon entropy are

$$p_0 = \frac{N_c + 1}{N_{(N-1)}}, \quad p_{1,2,\ldots,b-1} = \frac{N_c}{N_{(N-1)}}, \tag{A7}$$

where w.l.o.g. we assume that the symbol occurring $N_c(b) + 1$ times within the string is $c = 0$. To see that the Shannon entropy (30) of a $C_{(N-1)}$ string can be approximated by $\log_2(b)$ for large b, first notice that $1 - b^2 < 0$ and $b^2 + b + 1 > 0, \forall b > 1$. Furthermore, $\forall b > 0$, $b + 1 \ll b^2 + b + 1$, which implies that the first term

$$\log_2\left(\frac{b+1}{b^2+b+1}\right) < 0. \tag{A8}$$

Similarly the second term,

$$\log_2\left(\frac{b+2}{b^2+b+1}\right) < 0. \tag{A9}$$

Hence, the entropy (30) can be approximated by the dominant contribution from the first term, which is $\log_2(b)$.

The strings given by the relation (28) are not the shortest possible ones. Strings satisfying the equation (29) and satisfying $\min(bN_c(b) + 1) > N_{(N-1)}(b-1)$ are given by $b^2 + 1$ (OEIS A002522). They can be constructed to contain all possible doublets but without any triplets, starting with an initial balanced string of length $2b$ containing b clear doublets ordered from the main diagonal of the doublet matrix (27). Furthermore, their entropies are smaller than the entropies of the strings given by the equation (28). Namely $\forall b > 1$

$$\frac{1-b^2}{b^2+b+1}\log_2\left(\frac{b+1}{b^2+b+1}\right) - \frac{b+2}{b^2+b+1}\log_2\left(\frac{b+2}{b^2+b+1}\right) > \frac{b(1-b)}{b^2+1}\log_2\left(\frac{b}{b^2+1}\right) - \frac{b+1}{b^2+1}\log_2\left(\frac{b+1}{b^2+1}\right). \tag{A10}$$

Now, assume *a contrario* that a string $C'_{(N-1)}$ longer than $N_{(N-1)}$ can be constructed, say of length $N'_{(N-1)} = N_{(N-1)} + 1$. But in this case, the corresponding $H(C'_{(N-1)}) < H(C_{(N-1)})$. The string of the length given by the formula (28) maximizes the Shannon entropy if it must additionally satisfy the relation (29). Thus, Theorem 14 is proven.

 doi:10.20944/preprints202409.1581.v5

Appendix E. Proof of $C_{(N-k)}$ String Theorem

We start by noting that for $b = 1$, $N_{(N-2)}(1) = 5$, as the ASI of $[00000]$ is the same as the ASI of $[000000]$, $N_{(N-3)}(1) = 7$, as the ASI of strings of seven and eight same symbols is three, there is no $N_{(N-4)}(1)$, and so on. Hence, Theorem 15 does not hold for $b = 1$.

A $C_{(N-1)}$ string contains all doublets. Hence, inserting any basic symbol into any position inevitably leads to a repetition of a doublet. W.l.o.g. we append it at the start of the $C_{(N-1)}$ string, obtaining a string

$$C_k = [*000111222\dots], \quad a_{\max}^{(N_{(N-1)}+1,b)}(C_k) = N-2. \tag{A11}$$

Another symbol can be introduced to this string without an additional doublet repetition provided that it adjoins the previously introduced symbol, which gives a string

$$C_l = [\star * 000111222\dots], \quad a_{\max}^{(N_{(N-1)}+2,b)}(C_l) = N-2, \tag{A12}$$

leading to the repetition of the doublet $\star*$ or $*0$ but not both of them (here we allow $\star = *$). Hence, both the length and the ASI of this string increase by one. Finally, 0 can be appended at the start of this string without an additional doublet repetition provided that $\star \neq 0$ and $* = 0$ and the string becomes

$$C_{(N-2)} = [0 \star 0000111222\dots], \quad a_{\max}^{(N_{(N-1)}+3,b)}(C_{(N-2)}) = N-2, \tag{A13}$$

leading to the mutually exclusive repetition of the doublet $0\star$, $\star0$ or 00, so that also both length and the ASI of this string increase by one. An insertion of another symbol into the string (A13) at any position will maintain or even decrease the ASI of this newly formed string. For example, appending 0 at the start of the $C_{(N-2)}$ string (A13), where $\star = 1$

$$[0010000111222\dots]. \tag{A14}$$

creates a 001 triplet based on 00 doublet leading to a decrease of the ASI of this longer string to $a = N-4$ as compared to $a = N-2$ of the string (A13).

$C_{(N-2)}$ string (A13) must contain only two copies of a doublet. Hence, a clear quadruplet ($bbbb$) and a pattern binding different symbols adjoining this quadruplet, such as $[\dots abbbbc\dots abc\dots]$, $[\dots abbbbaba\dots]$, etc. must be present, so that any $C_{(N-2)}$ string contains only one pair of repeated doublets ab, bb, or $\{bc, ba\}$ (See also Appendix C). For example, for $N = 10$, sixteen bitstrings

$$\begin{array}{cccc} [0100011110], & [0111100010], & [0111101000], & [\underline{0100001110}], \\ [0001011110], & [0001111010], & [0101111000], & [0111000010] \end{array} \tag{A15}$$

(an additional eight are given by swapping 0 with 1) have the ASI $a = N-2 = 8$, where the underlined string (A15) is the one that we created for $b = 2$. Each string $C_{(N-2)}$ (A15) contains three pairs of doublets $[01]$, $[10]$, and $[**]$ overlapped in such a way that only one pair can be reused from the ASP to decrease the maximum $N-1$ ASI by one.

Searching for a $C_{(N-3)}$ string, w.l.o.g. we append $* \neq 0$ at the start of the $C_{(N-2)}$ string (A13)

$$C_k = [*010000111222\dots], \quad a_{\max}^{(N_{(N-1)}+4,b)}(C_k) = N-3. \tag{A16}$$

If $* = 1$, we have the same three doublets 10. Otherwise, we have two pairs of the same doublets $*0$ and 10. Both cases are equivalent by Theorem 4. An insertion of another symbol to this string may maintain or even decrease the ASI of this newly formed string. To maximize its ASI, another symbol must adjoin $*$. Hence, we append $\star$ at the start, where $\forall\star$ and $\forall* \neq 0$, a string

$$C_l = [\star * 010000111222\dots], \quad a_{\max}^{(N_{(N-1)}+5,b)}(C_l) = N-3, \tag{A17}$$

 doi:10.20944/preprints202409.1581.v5

has an increased length and ASI. W.l.o.g. for $b = 2$ we have four bitstrings (A17), wherein three of them

$$\begin{aligned} C_1^{(12,2)} &= [000100001110], \quad a\left(C_1^{(12,2)}\right) = 12 - 4 = 8, \\ C_2^{(12,2)} &= [110100001110], \quad a\left(C_2^{(12,2)}\right) = 8, \\ C_3^{(12,2)} &= [100100001110], \quad a\left(C_3^{(12,2)}\right) = 8, \end{aligned} \tag{A18}$$

have the same non-maximum ASI and only one have the maximum ASI

$$C_{(N-3)}^{(12,2)} = [010100001110], \quad a_{\max}^{(N_{(N-1)}+5,2)}(C_{(N-3)}^{(12,2)}) = 12 - 3 = 9, \tag{A19}$$

and cannot be further extended along with the increment of the ASI. Therefore

$$C_{(N-3)}^{(N,b)} = [01010000111222\ldots10\ldots], \quad a_{\max}^{(N_{(N-1)}+5,b)}(C_{(N-3)}^{(N,b)}) = N - 3, \tag{A20}$$

and the ASI of this newly formed string increases again. However, the insertion of another symbol into this string will maintain or even decrease the ASI of this newly formed string. Any $C_{(N-3)}$ string must contain only three copies of a doublet, two copies of a triplet, or two pairs of different doublets. W.l.o.g. we have found the following $C_{(N-k)}$ strings for $b = 2$ and $4 \leq k \leq 8$

$$\begin{aligned} C_{(N-2)}^{(10,2)} &= [0100001110], \quad a_{\max}^{(10,2)} = 8, \\ C_{(N-3)}^{(12,2)} &= [010100001110], \quad a_{\max}^{(12,2)} = 9 \left([01] \text{ to } C_{\max}^{(10,2)}\right), \\ C_{(N-4)}^{(14,2)} &= [01010100001110], \quad a_{\max}^{(14,2)} = 10 \left([01] \text{ to } C_{\max}^{(12,2)}\right), \\ C_{(N-5)}^{(16,2)} &= [0101010000001110], \quad a_{\max}^{(16,2)} = 11 \left([00] \text{ to } C_{\max}^{(14,2)}\right), \\ C_{(N-6)}^{(18,2)} &= [010101000000111110], \quad a_{\max}^{(18,2)} = 12 \left([11] \text{ to } C_{\max}^{(16,2)}\right), \\ C_{(N-7)}^{(20,2)} &= [0101010000000111110], \quad a_{\max}^{(20,2)} = 13 \left([01] \text{ to } C_{\max}^{(18,2)}\right), \\ C_{(N-8)}^{(22,2)} &= [01010100000000110111110], \quad a_{\max}^{(22,2)} = 14 \left([10] \text{ to } C_{\max}^{(20,2)}\right), \\ C_{(N-9)}^{(24,2)} &= [010101001000000110111110], \quad a_{\max}^{(24,2)} = 15 \left([01] \text{ to } C_{\max}^{(22,2)}\right), \end{aligned} \tag{A21}$$

which led us to the strings (34) for all $b > 1$. Thus, Theorem 15 is proven.

Appendix F. Additional Comments for the Proof of Theorem 12

We can also use mathematical induction on the length N of the string, if is is a power of two. For the base case ($N = 2^0 = 1$) the string consists of a single basic symbol $c \in P_0^{(b)}$. Hence, its ASI is $a_{\min}^{(1)} := 0$ and its ASD ${d_s}^{(1,b)} := 0$. Therefore, ${d_s}^{(1,b)} = a_{\min}^{(1)} = 0$. Assume now that for all strings of length 2^k less than N, the ASD equals the minimum ASI, that is

$$d_{a_{\min}}^{(2^k,b)} = a_{\min}^{(2^k)} \quad \forall 2^k < N. \tag{A22}$$

For some integer k, we construct the minimum ASI string as follows. First, we assemble a doublet from two basic symbols:

$$c_1 \circ c_2 = C^{(2,b)}, \quad c_1, c_2 \in P_0^{(b)}. \tag{A23}$$

 doi:10.20944/preprints202409.1581.v5

Its ASI is $a_{\min}^{(2)} = 1$ and its ASD is $d_s^{(2,b)} = 1$. Then for each $k \geq 2$ we have $C^{(2^{k-1},b)}$ with $a_{\min}^{(2^{k-1})} = k-1$ and ${d_s}^{(2^{k-1},b)} = k-1$ and we construct $C^{(2^k,b)}$ by joining two copies of $C^{(2^{k-1},b)}$

$$C^{(2^{k-1},b)} \circ C^{2^{k-1},b)} = C^{(2^k,b)}. \tag{A24}$$

The ASI of $C^{(2^k,b)}$ is equal to

$$a_{\min}^{(2^k)} = a_{\min}^{(2^{k-1})} + 1 = k, \tag{A25}$$

and the ASD is equal to

$$d_s^{(2^k,b)} := \max\left(d_{(s-1)L}^{(2^{k-1},b)},\, d_{(s-1)R}^{(2^{k-1},b)}\right) + 1 = (k-1) + 1 = k. \tag{A26}$$

Therefore, ${d_s}^{(2^k,b)} = a_{\min}^{(2^k)} = k$ in this case.

Appendix G. Misunderstanding Assembly Pools

Consider the following mapping [20] between a working ASP $P_3^{(5)}$ containing five basic symbols and three strings made of these symbols in three steps and the initial ASP of radix $b = 8$

$$\begin{aligned} P_3^{(5)} &\leftrightarrow P_0^{(8)} \\ 0 &\leftrightarrow 0 \\ 1 &\leftrightarrow 1 \\ 2 &\leftrightarrow 2 \\ 3 &\leftrightarrow 3 \\ 4 &\leftrightarrow 4 \\ 20 &\leftrightarrow 5 \\ 201 &\leftrightarrow 6 \\ 2012 &\leftrightarrow 7 \end{aligned} \tag{A27}$$

Now consider the string

$$C_k^{(11,5)} = [20123242012] \tag{A28}$$

assembled beginning with the initial ASP $P_0^{(5)}$ and having the ASI $a^{(11,5)}(C_k) = 7$ only two steps above $a_{\min}^{(11)} = 5$, as we can assemble this string as the string

$$C_l^{(8,8)} = [20123247] \tag{A29}$$

of length $N = 8$ in 7 steps with the initial ASP $P_0^{(8)}$ and then, using the mapping (A27), it will correspond to the string (A28). However, as we have shown in Section 2, $N_{(N-1)}(8) = 73 \neq 7$. In fact the latter string (A29) should be assembled as

$$C_m^{(5,8)} = [73247] \tag{A30}$$

with the ASI $a^{(5,8)}(C_m) = 5 - 1 = 4$ and with the initial ASP $P_0^{(8)}$, as $2012 \leftrightarrow 7$ according to the mapping (A27). Hence, considering a set $P_3^{(5)}$ as the *initial ASP* is a gross misunderstanding; there is only one initial ASP for a given b and many different working ASPs for $b > 1$ and $s > 1$ ($P_1^{(1)} = \{0, 00\}$). Furthermore, basic objects must have the same vanishing ASD (13).

 doi:10.20944/preprints202409.1581.v5

References

1. S. M. Marshall, A. R. G. Murray, and L. Cronin, "A probabilistic framework for identifying biosignatures using Pathway Complexity," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 375, p. 20160342, Dec. 2017.
2. S. Imari Walker, L. Cronin, A. Drew, S. Domagal-Goldman, T. Fisher, and M. Line, "Probabilistic biosignature frameworks," in *Planetary Astrobiology* (V. Meadows, G. Arney, B. Schmidt, and D. J. Des Marais, eds.), pp. 1–1, University of Arizona Press, 2019.
3. V. S. Meadows, G. N. Arney, B. E. Schmidt, and D. J. Des Marais, eds., *Planetary astrobiology*. University of Arizona space science series, Tucson: The University of Arizona Press ; Houston : Lunar and Planetary Institute, 2020. OCLC: 1151198948.
4. Y. Liu, C. Mathis, M. D. Bajczyk, S. M. Marshall, L. Wilbraham, and L. Cronin, "Exploring and mapping chemical space with molecular assembly trees," *Science Advances*, vol. 7, p. eabj2465, Sept. 2021.
5. S. M. Marshall, C. Mathis, E. Carrick, G. Keenan, G. J. T. Cooper, H. Graham, M. Craven, P. S. Gromski, D. G. Moore, S. I. Walker, and L. Cronin, "Identifying molecules as biosignatures with assembly theory and mass spectrometry," *Nature Communications*, vol. 12, p. 3033, May 2021.
6. S. M. Marshall, D. G. Moore, A. R. G. Murray, S. I. Walker, and L. Cronin, "Formalising the Pathways to Life Using Assembly Spaces," *Entropy*, vol. 24, p. 884, June 2022.
7. A. Sharma, D. Czégel, M. Lachmann, C. P. Kempes, S. I. Walker, and L. Cronin, "Assembly theory explains and quantifies selection and evolution," *Nature*, vol. 622, pp. 321–328, Oct 2023.
8. M. Jirasek, A. Sharma, J. R. Bame, S. H. M. Mehr, N. Bell, S. M. Marshall, C. Mathis, A. MacLeod, G. J. T. Cooper, M. Swart, R. Mollfulleda, and L. Cronin, "Investigating and Quantifying Molecular Complexity Using Assembly Theory and Spectroscopy," *ACS Central Science*, vol. 10, pp. 1054–1064, May 2024.
9. S. Łukaszyk and W. Bieniawski, "Assembly Theory of Binary Messages," *Mathematics*, vol. 12, p. 1600, May 2024.
10. S. Raubitzek, A. Schatten, P. König, E. Marica, S. Eresheim, and K. Mallinger, "Autocatalytic Sets and Assembly Theory: A Toy Model Perspective," *Entropy*, vol. 26, p. 808, Sept. 2024.
11. P. Francis, "Dilexit nos: Encyclical letter on the human and divine love of the heart of jesus christ," 2024. Accessed: 2024-11-01.
12. S. Łukaszyk and A. Tomski, "Omnidimensional Convex Polytopes," *Symmetry*, vol. 15, mar 2023.
13. "Book of John [1.3]," c90.
14. L. Cronin, "Exploring assembly index of strings is a good way to show why assembly & entropy are intrinsically different.." https://x.com/leecronin/status/1850289225935257665, 2024. Accessed: 2024-11-01.
15. S. Pagel, A. Sharma, and L. Cronin, "Mapping Evolution of Molecules Across Biochemistry with Assembly Theory," 2024.
16. M. M. Vopson, "The second law of infodynamics and its implications for the simulated universe hypothesis," *AIP Advances*, vol. 13, p. 105308, Oct. 2023.
17. S. Łukaszyk, "Shannon entropy of chemical elements," *European Journal of Applied Sciences*, vol. 11, p. 443–458, Jan. 2024.
18. S. Łukaszyk, *Black Hole Horizons as Patternless Binary Messages and Markers of Dimensionality*, ch. 15, pp. 317–374. Nova Science Publishers, 2023.
19. S. Łukaszyk, "Life as the explanation of the measurement problem," *Journal of Physics: Conference Series*, vol. 2701, p. 012124, Feb 2024.
20. L. Ozelim, A. Uthamacumaran, F. S. Abrahão, S. Hernández-Orozco, N. A. Kiani, J. Tegnér, and H. Zenil, "Assembly Theory Reduced to Shannon Entropy and Rendered Redundant by Naive Statistical Algorithms," 2024.