

Article

Not peer-reviewed version

On the Salient Regularities of Strings of Assembly Theory

[Wawrzyniec Bieniawski](#), Piotr Masierak, [Andrzej Tomski](#), [Szymon Łukaszyk](#)*

Posted Date: 23 October 2024

doi: 10.20944/preprints202409.1581.v4

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



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

On the Salient Regularities of Strings of Assembly Theory

Wawrzyniec Bieniawski ¹, Piotr Masierak ¹, Andrzej Tomski ² and Szymon Łukaszyk ^{1,*}

¹ Łukaszyk Patent Attorneys, ul. Głowackiego 8, 40-052 Katowice, Poland

² University of Silesia, Institute of Mathematics, Bankowa 14, 40-007 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 $N_{(N-k)} = b^2 + b + 3k - 2$ for $k = \{1, 2\}$ and by $N_{(N-3)} = b^2 + b + 6$. We also provide particular forms of such strings. Knowing the latter bound we conjecture that the maximum assembly index of a string of length $N > N_{(N-2)}$ is given by $a_{\max}^{(N,b)} = \lfloor N/2 \rfloor + b(b+1)/2$. 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 a string containing k copies of an n -plet has the assembly index of at most $N - 1 - (n-1)k - a$, 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, while the assembly depth of a maximum assembly index string has a value between $\lceil \log_2(N) \rceil$ and the assembly index of this string. Since these results are 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 $P_0(b) := \{0, 1, \dots, b-1\}$ 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.

In fact, any embodiment of AT, with basic symbols representing LEGO® blocks, chemical bonds, graphs, monomers, etc. assembled in any n -dimensional space ($n \in \mathbb{C}$) [11] 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 [12].

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

An assembly step s may consist of

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

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 $C_k \in P_s(b)$. We note that the joining operator " $\circ$ ", in general, does not commute. 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 leading to this string, since at each assembly step the cardinality of the working assembly pool increases by one. Therefore, the working assembly pool 2 cannot be identified with the initial assembly pool 1; the initial assembly pool 1 must not contain strings of basic symbols (see Section H).

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 quadruplets $C_{k, \min}^{(4, b)} = [***]$ and $b(b-1)$ quadruplets $C_{l, \min}^{(4, b)} = [**]**$ with ASI equal to two, while the ASI of the remaining 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 pool P_1 , while there is nothing available to reuse, in the case of the quadruplet $C_{l, \min}^{(4, 4)} = [0123]$. $\square$

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 $\star$. Furthermore, we consider the degenerate case of just one basic symbol ($b = 1$).

Theorem 2. *The smallest 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 $a^{(N)}(C_{\min}) = \min_k \left(\{a^{(N, b)}(C_k)\} \right), \forall k \in \{1, 2, \dots, b^N\}$ 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

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

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 $1 = a_0 < a_1 < \dots < a_s = N$ of integers such that $\forall j \geq 1, a_j = a_k + a_l$ for $l \leq k < j$. The first step in creating an addition chain for N is always $a_1 = 1 + 1 = 2$ and this corresponds to assembling a doublet $[**]$ or $[**]$ from the initial assembly pool $P_0(b)$. Thus, the lower bound for s of the addition chain for $N, s \geq \log_2(N)$ is achieved for $N = 2^s$ by b^2 strings (2).

The second step in creating an addition chain can be $a_2 = 1 + 1 = 2$ or $a_2 = 1 + 2 = 3$. 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 $\{0, 1, 2\}$. $\square$

At least some of the following seven simple theorems are useful for further consideration.

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

Proof. Minimal 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. Minimal 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 minimal strings $C_{\min}^{(2,b)}$ 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 minimal strings $C_{\min}^{(2^k,b)}$ contain at most two distinct symbols. We must show that $C_{\min}^{(2^{k+1},b)}$ also contains at most two distinct symbols. Consider constructing $C_{\min}^{(2^{k+1},b)}$ by joining two identical minimal strings $C_{\min}^{(2^k,b)}$

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

with each other. By the inductive hypothesis, each $C_{\min}^{(2^k,b)}$ 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 minimal string $C_{\min}^{(2^s,b)}$ contains at most two distinct symbols.

We will now show that other minimal ASI strings of length $N \neq 2^s$ can contain at most three distinct symbols if $b > 2$. We provide the construction of minimal 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 minimal ASI string $C_{\min}^{(N,b)}$ 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 $C_{\min}^{(N,b)}$. Thus, Theorem 3 is proven. $\square$

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.*

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)

$$C_k = [\dots 01 \dots 01 \dots 01 \dots ** \dots], \quad C_l = [\dots 01 \dots 01 \dots 22 \dots 22 \dots], \quad (4)$$

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. $\square$

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

$$C_k = [\dots 01 \dots 01 \dots 01 \dots], \quad C_l = [\dots 010 \dots 010 \dots]. \quad (5)$$

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. $\square$

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. $\square$

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

$$C_k = [\dots 0101 \dots 0101 \dots \star \dots], \quad C_l = [\dots 010 \dots 010 \dots 010 \dots]. \quad (6)$$

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. $\square$

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

$$C_k = [\dots 0001 \dots 0001 \dots], \quad C_l = [\dots 110 \dots 10 \dots 110 \dots]. \quad (7)$$

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. $\square$

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

$$C_k = [\dots 110 \dots 00 \dots 110 \dots 00 \dots], \quad C_l = [\dots 110 \dots 10 \dots 110 \dots \star \star \dots], \quad (8)$$

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

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. A string $C_m^{(N,b)}$ containing k_1 copies of an n_1 -plet $C_1^{(n_1,b)}$, k_2 copies of an n_2 -plet $C_2^{(n_2,b)}$, etc. has the ASI of at most.

$$a^{(N,b)}(C_m) = N - 1 - \sum_{r=1}^R \left[(n_r - 1)k_r - a(C_r^{(n_r,b)}) \right], \quad (9)$$

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

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

$$C_m^{(N,b)} = [\dots [c_1 c_2 \dots c_n] \dots [c_1 c_2 \dots c_n] \dots], \quad (10)$$

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 $a(C_l^{(n,b)})$ 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 $a(C_l^{(n,b)}) = n - 1$, 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 $(n - 1)1 - (n - 1) = 0$, while more copies k decrease it by $(n - 1)(k - 1)$. On the other hand, if $a(C_l^{(n,b)}) < n - 1$ then even a single copy of this n -plet will decrease the ASI of C_m . $\square$

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

$$C_k^{(24,6)} = [120100121010012350100152], \quad (11)$$

its ASI amounts to $a^{(24,6)}(C_k) = 24 - 1 - (4 \cdot 3 - 3) = 14$. We note that the maximum ASI decrease is provided by 2^s -plets of the minimum ASI and amounts to $(n - 1)k - \log_2(n) = (2^s - 1)k - s$.

Another quantity related to the string assembly is the assembly depth (AD) defined [13] as

$$d_s^{(N,b)} := \max(d_{(s-1)L}^{(N,b)}, d_{(s-1)R}^{(N,b)}) + 1, \quad (12)$$

where $d_{(s-1)L}^{(N,b)}$ and $d_{(s-1)R}^{(N,b)}$ are the ADs of two parts of this string that were joined in step s , where $d_0^{(N,b)} := 0$. If there are more assembly paths leading to a string with different ADs, which happens if at least two assembly steps can occur independently, in ref. [13] the minimum AD is assigned to the string. Here, we relax this assumption. Any string has a unique assembly index but can have different ADs if its ASI is not minimal.

Theorem 11. The AD of a minimal ASI string $C_{min}^{(N)}$ is equal to the ASI of this string, $d_{a_{min}}^{(N,b)} = a_{min}^{(N)}$.

Proof. We will prove this by mathematical induction on the length N of the string. For the base case ($N = 1$) the string consists of a single basic symbol $c \in P_0(b)$. Hence, its ASI is $a_{min}^{(1)} := 0$ and its AD $d_s^{(1,b)} := 0$. Therefore, $d_s^{(1,b)} = a_{min}^{(1)} = 0$.

Assume now that for all strings of length k less than N , the AD equals the minimal ASI, that is

$$d_{a_{min}}^{(k,b)} = a_{min}^{(k)} \quad \forall k < N. \quad (13)$$

We need to show that $d_{a_{min}}^{(N,b)} = a_{min}^{(N)}$. Consider two cases based on whether N is a power of two or not.

If N is a power of two, i.e., $N = 2^k$ for some integer k , we construct the minimal ASI string as follows. First, we assemble a doublet from two basic symbols:

$$c_1 \circ c_2 = C^{(2,b)}, \quad c_1, c_2 \in P_0(b). \quad (14)$$

Its ASI is $a_{\min}^{(2)} = 1$ and its AD 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 $C^{(2^{k-1},b)}$

$$C^{(2^{k-1},b)} \circ C^{(2^{k-1},b)} = C^{(2^k,b)}. \quad (15)$$

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

$$a_{\min}^{(2^k)} = a_{\min}^{(2^{k-1})} + 1 = k, \quad (16)$$

and the AD is equal to

$$d_s^{(2^k,b)} = \max\left(d_{(s-1)L}^{(2^{k-1},b)}, d_{(s-1)R}^{(2^{k-1},b)}\right) + 1 = (k-1) + 1 = k. \quad (17)$$

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

If N is not a power of two, we construct the minimal ASI string as follows. We take the shortest addition chain for N . Next, we build the string by combining previously assembled strings according to the addition chain. At each assembly step, the AD increases by one, as we are combining previously assembled strings (i.e. we do not perform independent assembly steps) Since, by Theorem 2, the minimal ASI corresponds to the length of the shortest addition chain $l(N)$, we have

$$d_s^{(N,b)} = l(N) = a_{\min}^{(N)}, \quad (18)$$

also in this case.

In both cases, whether N is a power of two or not, we have:

$$d_s^{(N,b)} = a_{\min}^{(N)}. \quad (19)$$

This completes the proof. $\square$

Hence, for all b , a string with $a_{\min}^{(N)}$ can be constructed only with the AD $d_{a_{\min}}^{(N)} = a_{\min}^{(N)}$.

Theorem 12. The AD of a maximal ASI string $C_{\max}^{(N)}$ satisfies

$$\lceil \log_2(N) \rceil \leq d_{a_{\max}}^{(N,b)} \leq a_{\max}^{(N,b)}, \quad (20)$$

where $\lceil x \rceil$ denotes the ceiling function.

Proof. Since in each assembly step, we combine two previously assembled substrings or basic symbols, the AD increases by at most one. Therefore, the AD cannot exceed the total number of assembly steps and we obtain

$$d_{a_{\max}}^{(N,b)} \leq a_{\max}^{(N,b)}. \quad (21)$$

Now we will prove that $d_{a_{\max}}^{(N,b)} \geq \lceil \log_2(N) \rceil$. Assume, for contradiction, that there exists a maximal ASI string $C_{\max}^{(N,b)}$ such that:

$$d_{a_{\max}}^{(N,b)} < \lceil \log_2(N) \rceil. \quad (22)$$

Let $d := d_{a_{\max}}^{(N,b)}$, so $d < \lceil \log_2(N) \rceil$. In an assembly tree of AD d , the maximum number of leaves that can be combined is 2^d , because at each assembly step, we can combine at most two subassemblies. Therefore, the maximum length $N_{\max}$ of a string that can be assembled with AD d satisfies:

$$N_{\max} \leq 2^d. \quad (23)$$

But since $d < \lceil \log_2(N) \rceil$, we have:

$$2^d < 2^{\lceil \log_2(N) \rceil} \leq 2^{\log_2(N)+1-\epsilon} = N \cdot 2^{1-\epsilon}, \quad (24)$$

for some $\epsilon > 0$. This implies that

$$N_{\max} < N \cdot 2^{1-\epsilon}. \quad (25)$$

However, since $\epsilon > 0$, $2^{1-\epsilon} < 2$, so:

$$N_{\max} < N \cdot 2^{1-\epsilon} < 2N. \quad (26)$$

But this contradicts the fact that $N_{\max} \geq N$ (since we are assembling a string of length N). More precisely, since $d < \lceil \log_2(N) \rceil$, we have $d \leq \lceil \log_2(N) \rceil - 1$, which implies:

$$2^d \leq 2^{\lceil \log_2(N) \rceil - 1} = 2^{\log_2(N)} \cdot 2^{-1} = \frac{N}{2}. \quad (27)$$

Thus, the maximum length of a string that can be assembled with AD d is $N_{\max} \leq N/2$, which is less than N , contradicting the assumption that we are assembling a string of length N . This completes the proof. $\square$

For example, the string $C_{\max}^{(8,2)} = [00011101]$ can be assembled with AD $\lceil \log_2(8) \rceil \leq d_{a_{\max}}^{(8,2)} \leq a_{\max}^{(8,2)}$ as

$$\begin{array}{llll} 00 & d_1 = 1, & 00 & d_1 = 1, & 00 & d_1 = 1, & 01 & d_1 = 1, \\ 01 & d_2 = 1, & 01 & d_2 = 1, & 01 & d_2 = 1, & 001 & d_2 = 2, \\ 11 & d_3 = 1, & 11 & d_3 = 1, & 0001 & d_3 = 2, & 0001 & d_3 = 3, \\ 0001 & d_4 = 2, & 0001 & d_4 = 2, & 00011 & d_4 = 3, & 00011 & d_4 = 4, \\ 1101 & d_5 = 2, & 000111 & d_5 = 3, & 000111 & d_5 = 4, & 000111 & d_5 = 5, \\ 00011101 & d_6 = 3, & 00011101 & d_6 = 4, & 00011101 & d_6 = 5, & 00011101 & d_6 = 6. \end{array} \quad (28)$$

Conjecture 1. For N such that $a_{\min}^{(N)} > a_{\min}^{(N-1)}$ the AD $d_s^{(N,b)}$ of a string $C_k^{(N,b)}$ has a range of values between the minimum ASI for N and the ASI of this string, that is

$$\lceil \log_2(N) \rceil \leq d_s^{(N,b)}(C_k) \leq a^{(N,b)}(C_k). \quad (29)$$

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

$$[*****] \quad \text{and} \quad [*****], \quad (30)$$

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

$$[0001112 \dots (b-2)(b-1)(b-1)(b-1)]. \quad (31)$$

The doublets that can be inserted into the initial string (31) can be arranged in a $b \times b$ matrix

$$\begin{bmatrix} \cancel{00} & \cancel{01} & 02 & \dots & 0(b-1) \\ 10 & \cancel{11} & \cancel{12} & \dots & 1(b-1) \\ 20 & 21 & \cancel{22} & \dots & 2(b-1) \\ \dots & \dots & \dots & \dots & \dots \\ (b-2)0 & (b-2)1 & (b-2)2 & \dots & \cancel{(b-2)(b-1)} \\ (b-1)0 & (b-1)1 & (b-1)2 & \dots & \cancel{(b-1)(b-1)} \end{bmatrix}, \quad (32)$$

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 (31), we can also cross out the entries in the first superdiagonal of the matrix (32). 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 13 ($N_{(N-1)}$). *The longest length of a string that has the ASI of $N-1$ is given by*

$$N_{(N-1)} = 3b + (b-1)^2 = b^2 + b + 1 \quad (33)$$

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

$$N_{(N-1)} = bN_c + 1, \quad (34)$$

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

$$\begin{aligned} H(C_{(N-1)}) &= -\sum_{c=0}^{b-1} p_c \log_2(p_c) = -(b-1) \frac{N_{(N-1)}-1}{bN_{(N-1)}} \log_2\left(\frac{N_{(N-1)}-1}{bN_{(N-1)}}\right) - \frac{N_{(N-1)}-1+b}{bN_{(N-1)}} \log_2\left(\frac{N_{(N-1)}-1+b}{bN_{(N-1)}}\right) \\ &= \frac{1-b^2}{b^2+b+1} \log_2\left(\frac{b+1}{b^2+b+1}\right) - \frac{b+2}{b^2+b+1} \log_2\left(\frac{b+2}{b^2+b+1}\right) \lesssim \log_2(b). \end{aligned} \quad (35)$$

The proof of Theorem 13 is given in Appendix [D](#). A $C_{(N-1)}$ string must contain all clear triplets and all doublets. 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 (33) 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 .

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

Theorem 14 ($N_{(N-2)}$). *For all $b > 1$ the longest length of a string that has the ASI of $N-2$ is given by $N_{(N-2)} = N_{(N-1)} + 3$ or equivalently by*

$$N_{(N-2)} = b^2 + b + 4, \quad (36)$$

and

$$N_{(N-2)} = (b-2)N_c + (N_c+1) + (N_c+3) = bN_c + 4, \quad (37)$$

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

$$H(C_{(N-2)}) = -\frac{b^2 - b - 2}{b^2 + b + 4} \log_2 \left(\frac{b+1}{b^2 + b + 4} \right) - \frac{b+2}{b^2 + b + 4} \log_2 \left(\frac{b+2}{b^2 + b + 4} \right) - \frac{b+4}{b^2 + b + 4} \log_2 \left(\frac{b+4}{b^2 + b + 4} \right). \quad (38)$$

The entropy $H(C_{(N-2)}) \lesssim \log_2(b)$ for $b \gtrsim 1.6398$.

The proof of Theorem 14 is given in Appendix F. $C_{(N-2)}$ string must contain only two copies of a doublet. Hence, a clear quadruplet ($bbbb$) and a pattern binding different symbols adjoining this quadruplet, such as $[\dots abbbbc \dots abc \dots]$, $[\dots abbbbab a \dots]$, 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

$$\begin{aligned} &[0100011110], \quad [0111100010], \quad [0111101000], \quad [0100001110], \\ &[0001011110], \quad [0001111010], \quad [0101111000], \quad [0111000010] \end{aligned} \quad (39)$$

(an additional eight are given by swapping 0 with 1) have the ASI $a = N - 2 = 8$, where the underlined string (39) is the one that is created for $b = 2$ in Appendix F.

Theorem 15 ($N_{(N-3)}$). The longest length of a string that has the ASI of $N - 3$ is given by $N_{(N-3)} = N_{(N-1)} + 5$. or equivalently by

$$N_{(N-3)} = b^2 + b + 6, \quad (40)$$

and

$$N_{(N-3)} = (b - 2)N_c + (N_c + 2) + (N_c + 4) = bN_c + 6, \quad (41)$$

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

$$H(C_{(N-3)}) = -\frac{b^2 - b - 2}{b^2 + b + 6} \log_2 \left(\frac{b+1}{b^2 + b + 6} \right) - \frac{b+3}{b^2 + b + 6} \log_2 \left(\frac{b+3}{b^2 + b + 6} \right) - \frac{b+5}{b^2 + b + 6} \log_2 \left(\frac{b+5}{b^2 + b + 6} \right). \quad (42)$$

The entropy $H(C_{(N-3)}) \lesssim \log_2(b)$ for $b \gtrsim 1.8317$.

The proof of Theorem 15 is given in Appendix G. $C_{(N-3)}$ string must contain only three copies of a doublet, or two pairs of different doublets.

In general, the strings of Theorem 14, generated by Method A or B, are terminated with 0 and owe their properties to the following distributions of symbols

$$\begin{aligned} C_{(N-2)} &= [0100001 \dots 10 \dots 0], \\ C_{(N-3)} &= [010100001 \dots 10 \dots 0], \end{aligned} \quad (43)$$

$C_{(N-2)}$ contains three pairs of doublets $[00]$, $[01]$, and $[10]$ which are connected in such a way that only one pair can be reused from the assembly pool to decrease the maximum $N - 1$ ASI by one.

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

$$a_{\max}^{(N,b)} = \begin{cases} a_{\max}^{(N-1,b)} + 1 & \text{iff } N = 2l \\ a_{\max}^{(N-1,b)} & \text{iff } N = 2l + 1 \end{cases}. \quad (44)$$

Equivalently, if $b > 1$ and $N > N_{(N-2)}$ then

$$a_{max}^{(N,b)} = \left\lfloor \frac{N}{2} \right\rfloor + \frac{b(b+1)}{2}, \quad (45)$$

$(b(b+1)/2)$ are triangular numbers, [A000217](#)). In other words, if $N > N_{(N-2)}$, then ASI increments by one every time the string length N increments by two.

First, we note that maximum ASI must rise. Otherwise, if it was constant for N larger than some N_{max} then at some even larger N it would become inevitably lower than minimum ASI bound which also rises, and this would be a contradiction.

W.l.o.g. we aim to prove this theorem for $b = 2$. We note that inserting any doublet into a $C_{(N-3)}^{(12,2)}$ string ([A20](#)) creates a triplet. Using the equation (9) of Theorem 10 we have

$$\begin{aligned} a_s &= a_{s-2} + 1, \quad N_s = N_{s-2} + 2, \\ a_s &= N_s - 1 - \sum_{s=1}^{R_s} [(n_s - 1)k_s - a(C_s^{(n_s,b)})], \\ a_{s-2} &= N_{s-2} - 1 - \sum_{p=1}^{R_{s-2}} [(n_p - 1)k_p - a(C_p^{(n_p,b)})], \\ a_s - a_{s-2} &= (N_{s-2} + 2) - 1 - \sum_{s=1}^{R_s} [(n_s - 1)k_s - a(C_s^{(n_s,b)})] - \left(N_{s-2} - 1 - \sum_{p=1}^{R_p} [(n_p - 1)k_p - a(C_p^{(n_p,b)})] \right) = \\ &= 2 - \sum_{s=1}^{R_s} [(n_s - 1)k_s - a(C_s^{(n_s,b)})] + \sum_{p=1}^{R_p} [(n_p - 1)k_p - a(C_p^{(n_p,b)})] = 1, \\ \sum_{s=1}^{R_s} [(n_s - 1)k_s - a(C_s^{(n_s,b)})] &= \sum_{p=1}^{R_p} [(n_p - 1)k_p - a(C_p^{(n_p,b)})] + 1, \end{aligned} \quad (46)$$

for any step s if only $N_s \geq N_{(N-2)}$.

Assume that $\forall s, a(C_s^{(n_s,b)}) = n_s - 1$ and $\forall p, a(C_p^{(n_p,b)}) = n_p - 1$. Then

$$\begin{aligned} \sum_{s=1}^{R_s} [(n_s - 1)(k_s - 1)] &= \sum_{p=1}^{R_p} [(n_p - 1)(k_p - 1)] + 1, \\ \sum_{s=1}^{R_s} n_s k_s - \sum_{s=1}^{R_s} n_s - \sum_{s=1}^{R_s} k_s + R_s &= \sum_{p=1}^{R_p} n_p k_p - \sum_{p=1}^{R_p} n_p - \sum_{p=1}^{R_p} k_p + R_p + 1, \end{aligned} \quad (47)$$

The proof of the Conjecture 2 must show the conditions for the equations (46) and (47) to hold.

The bounds of Theorems 14-15 and Conjecture 2 are illustrated in Figures 1 and 2.

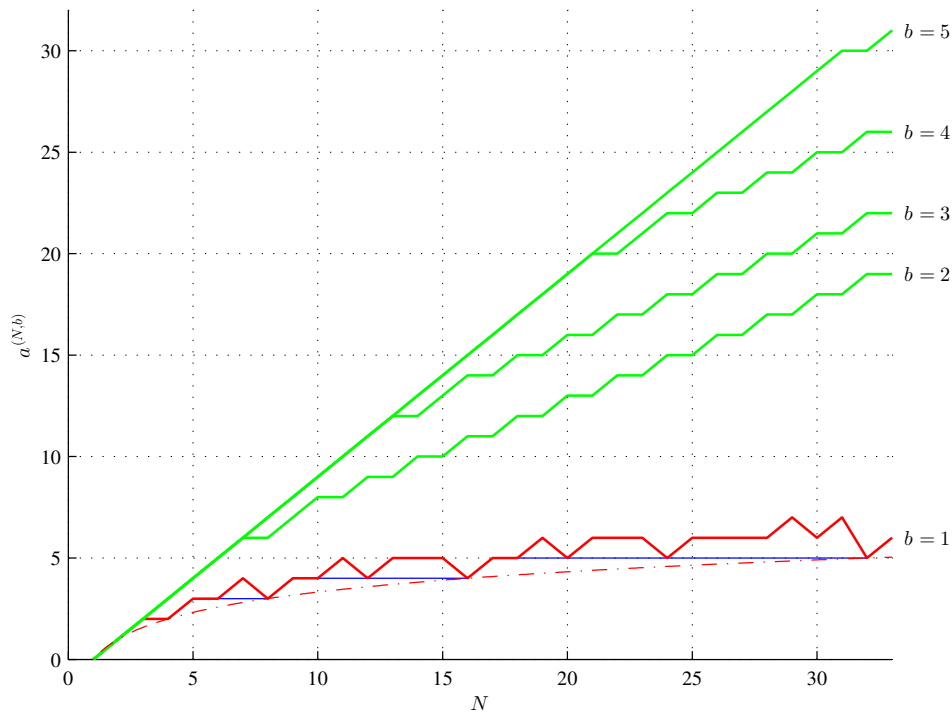


Figure 1. Lower assembly index bound (red) and upper bounds (green) for $1 \leq b \leq 4$, lower assembly depth bound (blue) of $C_{\max}^{(N,b)}$ strings for $b > 1$, and $\log_2(N)$ (red, dash-dot) for $0 < N \leq 33$.

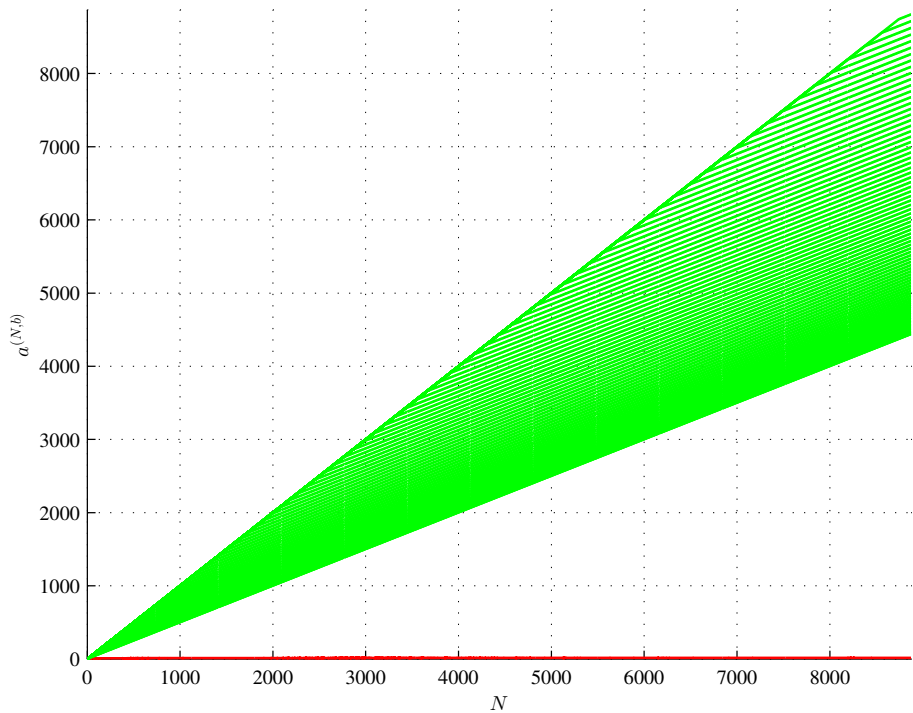


Figure 2. Lower assembly index bound (red) and the upper bounds (green) for $1 \leq b \leq 94$ and $0 < N \leq 8873$.

These results disprove our upper bound Conjecture 1 for $b = 2$ stated in ref. [9]. However, along with Theorem 12 they led us to the method of determining the ASI of a maximum ASI string, based on independent doublets and powers of two, as shown in Table 1. First, a string is sequenced, every two symbols to find the number of unique doublets $\times 2_{(b)}$. Subsequently, these doublets form $\times 4$ quadruplets, quadruplets form $\times 8$ octuples, and so on depending on the length of the string N . Only the first quantity, $\times 2_{(b)}$, depends on the radix b . 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 $\times 2^s = 1$. It turns out that for $b > 1$, the ASI depends only on the distribution of doublets ($\times 2_{(b)}$). In particular, the $N_{(N-1)}$ strings (A3) or (A4) contain the maximum of $\lfloor N_{(N-1)}/2 \rfloor$ independent doublets, and the $N_{(N-2)}$ string (A14) contains the maximum of $N_{(N-2)}/2 - 1$ independent doublets.

For example, the $N_{(N-3)}$ string (A21) of length $N_{(N-3)} = 18$ for $b = 3$ can be assembled as

$$\begin{aligned}
0 \circ 1 &= [01], 0 \circ 0 = [00], 1 \circ 1 = [11], 1 \circ 2 = [12], \\
2 \circ 2 &= [22], 1 \circ 0 = [10], 2 \circ 0 = [20] \quad (\times 2_{(b=3)} = 7), \\
[01] \circ [01] &= [0101], [00] \circ [00] = [0000], [11] \circ [12] = [1112], [22] \circ [10] = [2210] \quad (\times 4 = 4), \\
[0101] \circ [0000] &= [01010000], [1112] \circ [2210] = [11122210] \quad (\times 8 = 2), \\
[01010000] \circ [11122210] &= [0101000011122210] \quad (\times 16 = 1), \\
[0101000011122210] \circ [20] &= [010100001112221020] \quad (\text{last} \times 2),
\end{aligned} \tag{48}$$

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

$$\begin{aligned}
0 \circ 0 &= [00], 0 \circ 1 = [01], 1 \circ 1 = [11], 2 \circ 2 = [22], 2 \circ 3 = [23], \\
3 \circ 3 &= [33], 1 \circ 0 = [10], 2 \circ 1 = [21], 3 \circ 2 = [32], 0 \circ 3 = [03] \quad (\times 2_{(b=4)} = 10), \\
[00] \circ [01] &= [0001], [11] \circ [22] = [1122], [23] \circ [33] = [2333], \\
[10] \circ [21] &= [1021], [32] \circ [03] = [3203] \quad (\times 4 = 5), \\
[0001] \circ [1122] &= [00011122], [2333] \circ [1021], [23331021] \quad (\times 8 = 2), \\
[00011122] \circ [23331021] &= [0001112223331021] \quad (\times 16 = 1), \\
[0001112223331021] \circ [3203] &= [00011122233310213203] \quad (\text{last} \times 4), \\
[00011122233310213203] \circ 0 &= [000111222333102132030] \quad (\text{last} \times 1),
\end{aligned} \tag{49}$$

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

N	$\times 2_{(b=1)}$	$\times 2_{(b=2)}$	$\times 2_{(b=3)}$	$\times 2_{(b=4)}$	$\times 4$	$\times 8$	$\times 16$	$\times 32$	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

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

$$\begin{aligned}
 0 \circ 0 &= [00], & (\times 2_{(b=1)} &= 1), \\
 [00] \circ [00] &= [0000] & (\times 4_{(b=1)} &= 1), \\
 [0000] \circ [0000] &= [00000000] & (\times 8_{(b=1)} &= 1), \\
 [00000000] \circ [0000] &= [000000000000] & (\text{last} \times 4), \\
 [000000000000] \circ [00] &= [00000000000000] & (\text{last} \times 2), \\
 [00000000000000] \circ [0] &= [0000000000000000] & (\text{last} \times 1),
 \end{aligned} \tag{50}$$

$$1 + 1 + 1 + 1 + 1 + 1 = 6 \text{ steps.}$$

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].

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

$$\begin{array}{llll}
 0 \circ 0 = [00], 1 \circ 0 = [01], 1 \circ 0 = [11] & (\times 2_{(b=2)} = 3), & 0 \circ 0 = [00], 1 \circ 0 = [10], 1 \circ 0 = [11] & (\times 2_{(b=2)} = 3), \\
 [00] \circ [01] = [0001] & (\times 4 = 1), & [00] \circ [10] = [0010] & (\times 4 = 1), \\
 [0001] \circ [11] = [000111] & (\text{last} \times 2), & [0010] \circ [11] = [001011] & (\text{last} \times 2), \\
 3 + 1 + 1 = & 5 \text{ steps,} & \underline{3 + 1 + 1} = & \underline{5 \text{ steps.}}
 \end{array} \tag{51}$$

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 minimal assembly steps that define biological complexity, reinforcing the idea that life is characterized by highly 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 [14,15] and, hence, will become less balanced. This tendency to imbalance seems to be associated with some minimum energy condition. The energy of a black hole that can be thought of as a balanced bitstring [16] can be two times the energy of the entropy variation sphere that it generates [17]. 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 [9,14]. The maximum ASI of a $C_{\max}^{(29903,4)}$ string given by the equation (45) is $a_{\max}^{(29903,4)} = 14961$.

Authors' contributions—WB: first concept of a general method for constructing the string of length $N_{(N-1)}$ leading to Theorem 13; the concept of the doublet matrix (32); 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 2; numerous clarity corrections and improvements; AT: formal proof of Theorem 3; proof that the Shannon entropy (35) can be approximated by $\log_2(b)$ for large b ; proof of the Theorem 11; proof of the Theorem 12; 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. 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 Łukaszuk 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

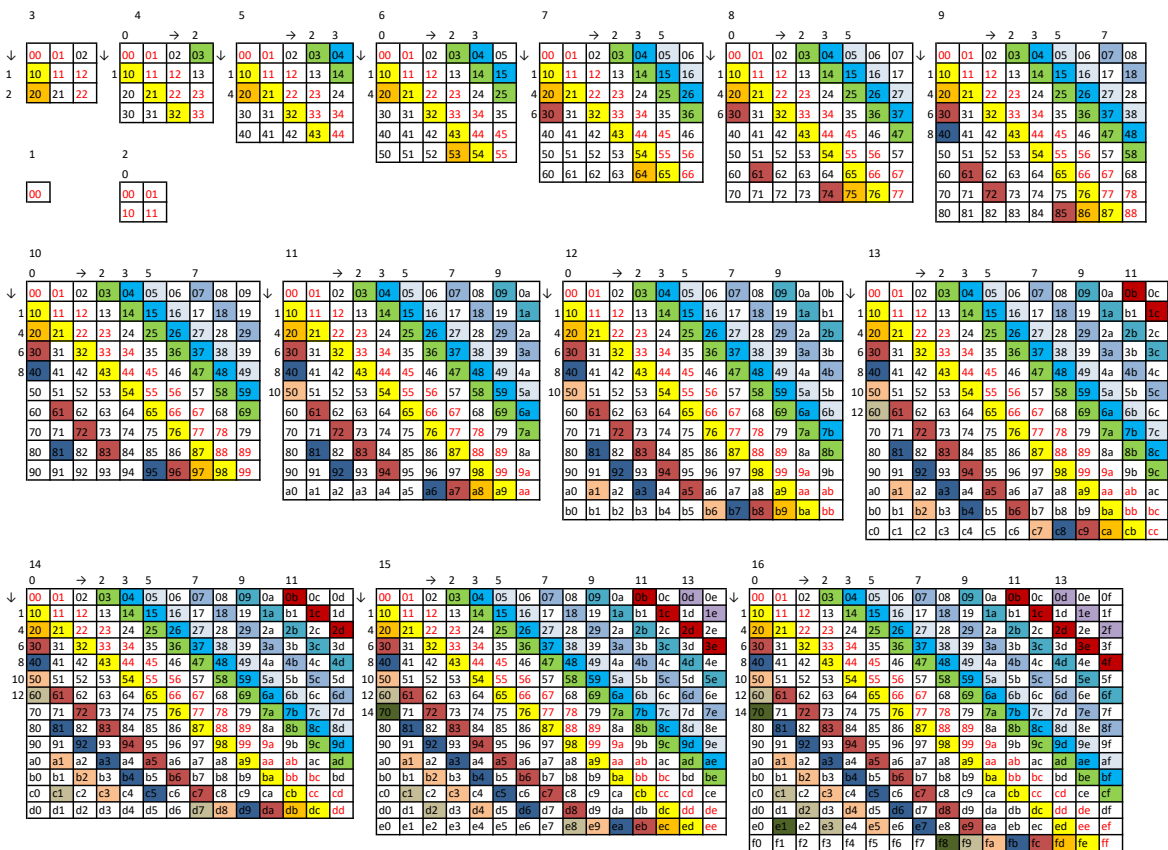


Figure A1. Doublet matrices for $1 \leq b \leq 16$ that illustrate the generation of $N_{(N-1)}$ strings according to Method A. Colored doublets are appended to the initial string of clear triplets in the order indicated by arrows starting from the 1st column or row. Finally, 0 is appended at the end,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

We start with a string of clear triplets (31). In the 1st step, we create a string containing doublets on the first subdiagonal of the matrix (32) starting with 10

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

and we append it to the string (31). 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

$$[0314 \dots (b-5)(b-2)(b-4)(b-1)], \quad (\text{A2})$$

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 $\{31, 42, \dots, (b-2)(b-4)\}$. And so on. Finally, 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

$$\begin{aligned} & [0001112221020], \\ & [000111222333102132030], \\ & [0001112223334441021324303142040], \\ & [0001112223334445551021324354031425041520530], \\ & [000111222333444555666102132435465031425360415262064051630], \\ & [0001112223334445556667771021324354657603142536470415263720750516273061740], \\ & [\dots 1021324354657687031425364758041526374820860516273830617285071840], \\ & [\dots 1021324354657687980314253647586904152637485920970516273849 \\ & 30617283960718294081950], \\ & [\dots 102132435465768798a9031425364758697a0415263748596a20a8 \\ & 05162738495a3061728394a70718293a408192a6091a50], \\ & [\dots 102132435465768798a9ba031425364758697a8b0415263748596a7b20b9 \\ & 05162738495a6b3061728394a5b80718293a4b408192a3b7091a2b50a1b60], \\ & [\dots 102132435465768798a9bacb031425364758697a8b9c0415263748596a7b8c20ca \\ & 05162738495a6b7c3061728394a5b6c90718293a4b5c408192a3b4c8091a2b3c50a1b2c70b1c60]. \end{aligned} \quad (\text{A3})$$

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

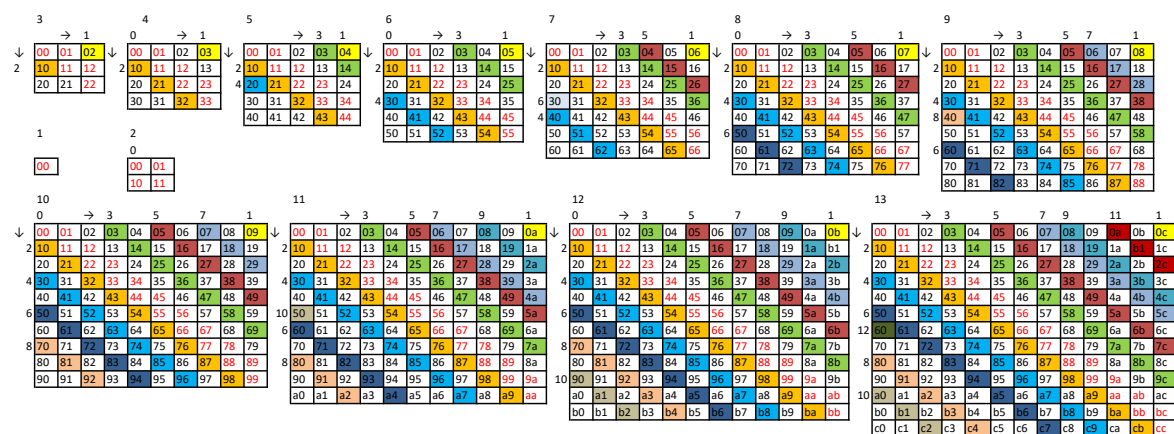


Figure A2. Doublet matrices for $1 \leq b \leq 13$ that illustrate the generation of $N_{(N-1)}$ strings according to Method B. Colored doublets are appended to the initial string of clear triplets in the order indicated by arrows starting from the 1st column or row. Finally, 0 is appended at the end,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

This method is similar to the Method A. We also start with a string of clear triplets (31) and the matrix of doublets (32) 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 (32)) at the end of the string (31). Next, we generally perform the following pairs of iterations:

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

$$\begin{aligned}
 &[0001112220210], \\
 &[000111222333031021320], \\
 &[0001112223334440410213243031420], \\
 &[0001112223334445550510213243540314253041520], \\
 &[000111222333444555666061021324354650314253640516204152630], \\
 &[0001112223334445556667770710213243546576031425364730415263740516275061720], \\
 &[\dots 0810213243546576870314253647583041526374850516273860718206172840], \\
 &[\dots 091021324354657687980314253647586930415263748596051627384950617283940718297081920], \\
 &[\dots 0a102132435465768798a9031425364758697a30415263748596a705162738495a \\
 &60718293a4061728394a8091a208192a50], \\
 &[\dots 0b102132435465768798a9ba031425364758697a8b30415263748596a7b805162738495a6b \\
 &5061728394a5b60718293a4b708192a3b4091a2b90a1b20], \\
 &[\dots 0c102132435465768798a9bacb031425364758697a8b9c30415263748596a7b8c905162738495a6b7c \\
 &5061728394a5b6c70718293a4b5c8091a2b3c408192a3b4ca0b1c20a1b2c60].
 \end{aligned}
 \tag{A4}$$

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 $b = \{1, 2, 3, 4, 5\}$)

$$\begin{aligned}
 &[0000] \\
 &[00001111010] \\
 &[0000111122221021202010] \\
 &[0000111122223333102132101202303203130] \\
 &[00001111222233334444102132431012023034041304242143203140]
 \end{aligned}
 \tag{A5}$$

and has a length of

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

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

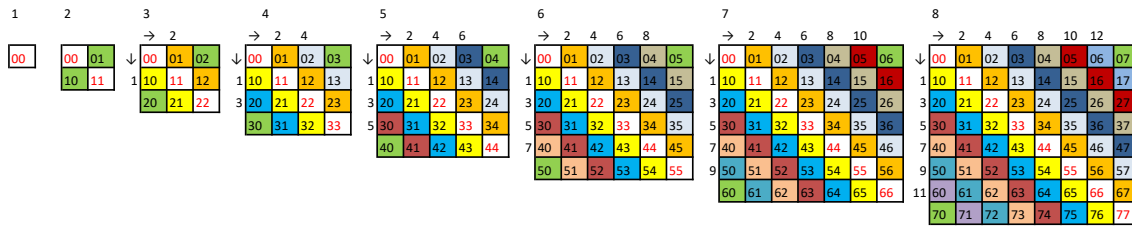


Figure A3. Doublet matrices for $1 \leq b \leq 8$ that illustrate the generation of N_{2D} strings containing exactly two copies of all doublets. Colored doublets are appended to the initial string of clear quadruplets in the order indicated by arrows starting from the 1st column or row. Finally, $0(b-1)0$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

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

The $N_{(N-1)}$ given by the formula (33) is an odd number for all b . The first element $3b$ is the length of the initial string (31) containing b clear triplets and $b^2 - b - (b-1)$ is the number of doublets available in the matrix (32) after crossing out b doublets on its diagonal and $b-1$ doublets on its superdiagonal that are present in the starting string (31). By definition, a $C_{(N-1)}$ string cannot have any repetitions. To be the longest, it must contain all doublets in the matrix (32) 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 (34) the fractions in the Shannon entropy are

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

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 (35) 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

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

Similarly the second term,

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

Hence, the entropy (35) can be approximated by the dominant contribution from the first term, which is $\log_2(b)$.

The strings given by (33) are not the shortest possible ones. Strings satisfying the equation (34) 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 (32). Furthermore, their entropies are smaller than the entropies of the strings given by the equation (33). Namely $\forall b > 1$

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

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 (33) maximizes the Shannon entropy if it must additionally satisfy the relation (34). Thus, Theorem 13 is proven.

Appendix E. Number of $C_{(N-1)}^{(13,3)}$ Strings

For $b = 3$, only two doublets can be introduced without repetitions into the initial string (31), leading to twelve unique strings of length $N_{(N-1)} = 13$

$$[0001112220210], [0001112221020], [2021000111222], [2102000111222], [0001112022210], [0001112102220], [2100020111222], [0002011122210], [0200011122210], [2000210111222], [2100011120222], [2100011122202]. \quad (A11)$$

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 $N_{(N-1)}(3) = 13$.

Appendix F. Proof of $C_{(N-2)}$ String Theorem

For $b = 1$, $N_{(N-2)}(1) = N_{(N-1)}(1) + 2 = 5$, as the ASI of $[00000]$ is the same as the ASI of $[000000]$.

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

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

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

$$C_l = [* * 000111222 \dots], \quad a_{\max}^{(N_{(N-1)}+2,b)}(C_l) = N - 2, \quad (A13)$$

leading to the repetition of the doublet $**$ or $*0$ but not both of them (here we allow $* = *$). Hence, both 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 $* \neq 0$ and $* = 0$ and the string becomes

$$C_{(N-2)} = [0 * 000111222 \dots], \quad a_{\max}^{(N_{(N-1)}+3,b)}(C_{(N-2)}) = N - 2, \quad (A14)$$

leading to the mutually exclusive repetition of the doublet $0*$, $*0$ or 00 , so that also both length and the ASI of this string increase by one. An insertion of another symbol into the string (A14) 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 (A14), where $* = 1$

$$[0010000111222 \dots]. \quad (A15)$$

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 (A14). Thus, Theorem 14 is proven.

For the string of the form (37) the fractions in the Shannon entropy are

$$p_0 = \frac{N_c + 3}{N_{(N-2)}}, \quad p_1 = \frac{N_c + 1}{N_{(N-2)}}, \quad p_{2,\dots,b-1} = \frac{N_c}{N_{(N-2)}}, \quad (A16)$$

where w.l.o.g. we assume that the symbol occurring $N_c(b) + 1$ times within the string is $c = 0$, which leads to Shannon entropy (38).

Appendix G. Proof of $C_{(N-3)}$ String Theorem

$N_{(N-3)}(1) = N_{(N-1)}(1) + 4 = 7$, as the ASIs of strings of seven and eight same symbols is three. The appending 0 at the start of the $C_{(N-2)}$ string (A14) decreases of the ASI of this longer string to $a = N - 4$ (cf. F). Thus, w.l.o.g. we append $\star \neq 0$ at the start of the $C_{(N-2)}$ string (A14)

$$C_k = [\star 010000111222 \dots], \quad a_{\max}^{(N_{(N-1)}+4,b)}(C_k) = N - 3. \quad (\text{A17})$$

If $\star = 1$, we have the same three doublets 10. Otherwise, we have two pairs of the same doublets $\star 0$ 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 $\star$. Hence, we append $\star$ at the start, where $\forall \star$ and $\forall \star \neq 0$, a string

$$C_l = [\star \star 010000111222 \dots], \quad a_{\max}^{(N_{(N-1)}+5,b)}(C_l) = N - 3, \quad (\text{A18})$$

has an increased length and ASI. W.l.o.g. for $b = 2$ we have four bitstrings (A18), wherein three of them

$$\begin{aligned} C_1^{(12,2)} &= [0001000011110], \quad a(C_1^{(12,2)}) = 12 - 4 = 8, \\ C_2^{(12,2)} &= [1101000011110], \quad a(C_2^{(12,2)}) = 8, \\ C_3^{(12,2)} &= [1001000011110], \quad a(C_3^{(12,2)}) = 8, \end{aligned} \quad (\text{A19})$$

have the same non-maximum ASI and only one

$$C_{(N-3)}^{(12,2)} = [0101000011110], \quad a_{\max}^{(N_{(N-1)}+5,2)}(C_{(N-3)}^{(12,2)}) = 12 - 3 = 9, \quad (\text{A20})$$

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

$$C_{(N-3)}^{(N,b)} = [01010000111222 \dots 10 \dots], \quad a_{\max}^{(N_{(N-1)}+5,b)}(C_{(N-3)}^{(N,b)}) = N - 3, \quad (\text{A21})$$

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. Thus, Theorem 15 is proven.

Appendix H. Misunderstanding Assembly Pools

Consider the following mapping [18] between a working assembly pool $P_3(5)$ containing five basic symbols and three strings made of these symbols in three steps and the initial assembly pool of radix $b = 8$

$$\begin{aligned} P_3(5) &\leftrightarrow P_0(8) \\ 0 &\leftrightarrow 0 \\ 1 &\leftrightarrow 1 \\ 2 &\leftrightarrow 2 \\ 3 &\leftrightarrow 3 \\ 4 &\leftrightarrow 4 \\ 20 &\leftrightarrow 5 \\ 201 &\leftrightarrow 6 \\ 2012 &\leftrightarrow 7 \end{aligned} \quad (\text{A22})$$

Now consider the string

$$C_k^{(11,5)} = [20123242012] \quad (\text{A23})$$

assembled beginning with the initial assembly pool $P_0(5)$ 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

$$C_l^{(8,8)} = [20123247] \quad (\text{A24})$$

of length $N = 8$ in 7 steps with the initial assembly pool $P_0(8)$ and then, using the mapping (A22), it will correspond to the string (A23). However, as we have shown in Section 2, $N_{(N-1)}(8) = 73 \neq 7$. In fact the latter string (A24) should be assembled as

$$C_m^{(5,8)} = [73247] \quad (\text{A25})$$

with the ASI $a^{(5,8)}(C_m) = 5 - 1 = 4$ and with the initial assembly pool $P_0(8)$, as $2012 \leftrightarrow 7$ according to the mapping (A22). Hence, considering a set $P_3(5)$ as the *initial assembly pool* is a gross misunderstanding; there is only one initial assembly pool for a given b and many different working assembly pools for $b > 1$ and $s > 1$ ($P_1(1) = \{0, 00\}$). Furthermore, basic objects must have the same vanishing AD (12).

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. Raubitsek, 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. S. Łukaszyk and A. Tamski, "Omnidimensional Convex Polytopes," *Symmetry*, vol. 15, mar 2023.
12. "Book of John [1.3]," c90.
13. S. Pagel, A. Sharma, and L. Cronin, "Mapping Evolution of Molecules Across Biochemistry with Assembly Theory," 2024.
14. M. M. Vopson, "The second law of infodynamics and its implications for the simulated universe hypothesis," *AIP Advances*, vol. 13, p. 105308, Oct. 2023.

15. S. Łukaszyk, "Shannon entropy of chemical elements," *European Journal of Applied Sciences*, vol. 11, p. 443–458, Jan. 2024.
16. S. Łukaszyk, *Black Hole Horizons as Patternless Binary Messages and Markers of Dimensionality*, ch. 15, pp. 317–374. Nova Science Publishers, 2023.
17. S. Łukaszyk, "Life as the explanation of the measurement problem," *Journal of Physics: Conference Series*, vol. 2701, p. 012124, Feb 2024.
18. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.