On the Salient Limits of Strings of Assembly Theory

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

Posted Date: 3 October 2024

doi: 10.20944/preprints202409.1581.v2

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

Article

On the Salient Limits 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: We consider the formation of strings of any natural radix b within the framework of assembly theory. In particular, we show that the upper bound of the assembly index depends on the radix b both quantitatively and qualitatively, 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 $k = \{1, 2, 3\}$. We also provide two particular forms of such strings. For $k = 1$ such odd length strings are nearly balanced and there are four such different strings if $b = 2$ and seventy-two if $b = 3$. Since these results are valid also 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* $P_0(b) \coloneqq \{0, 1, \ldots, b-1\}$ *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 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* $\{1, 2, \ldots, s-1\}$ *the working assembly pool.*

An assembly step s may consist of

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

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 $C_k \in P_s(b)$. 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 Appendix G).

 doi:10.20944/preprints202409.1581.v2

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 $C_{l,\min}^{(4,b)} = [*\star*\star]$ 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 pol 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 $\star$. Furthermore, we consider the degenerate case of $b = 1$.

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 $a^{(N)}(C_{\min}) = \min_k \left(\{a^{(N,b)}(C_k)\}\right), \forall k \in \{1, 2, \ldots, b^N\}$ 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 $l \leq k < j$. 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 $[*\star]$ 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 $\{0, 1, 2\}$. □

At least some of the following six 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 asembly 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

 doi:10.20944/preprints202409.1581.v2

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 $C_{\min}^{(2^k,b)}$

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

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. □

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 $*\star \neq 01$ and the same distributions of other repetitions (if there are any other repetitions)

$$C_k = [\ldots 01 \ldots 01 \ldots 01 \cdots * \star \ldots], \quad C_l = [\ldots 01 \ldots 01 \ldots 22 \ldots 22 \ldots], \tag{4}$$

where $*\star \neq 01$. 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

$$C_k = [\ldots 01 \ldots 01 \ldots 01 \ldots], \quad C_l = [\ldots 010 \ldots 010 \ldots]. \tag{5}$$

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. □

Theorem 6. *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.*

 doi:10.20944/preprints202409.1581.v2

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

$$C_k = [\ldots 0101 \ldots 0101 \cdots \star \ldots], \quad C_l = [\ldots 010 \ldots 010 \ldots 010 \ldots]. \tag{6}$$

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. □

Theorem 7. *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

$$C_k = [\ldots 0001 \ldots 0001 \ldots], \quad C_l = [\ldots 110 \ldots 10 \ldots 110 \ldots]. \tag{7}$$

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. □

Theorem 8. *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 $N+8$ with the same distributions of other repetitions

$$C_k = [\ldots 110 \ldots 00 \ldots 110 \ldots 00 \ldots], \quad C_l = [\ldots 110 \ldots 10 \ldots 110 \cdots ** \star \ldots], \tag{8}$$

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

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

$$[* * * \star \star \star *] \quad \text{and} \quad [\star * * * \star \star \star], \tag{9}$$

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, Method A and Method B). In both methods, we start with an initial balanced string of length $3b$ containing b clear triplets ordered as

$$[0001112 \ldots (b-2)(b-1)(b-1)(b-1)]. \tag{10}$$

 doi:10.20944/preprints202409.1581.v2

The doublets that can be inserted into the initial string (10) can be arranged in a $b \times b$ matrix

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

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 (10), we can also cross out the entries in the first superdiagonal of the matrix (11). 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 9 ($C_{(N-1)}$ string)**.** *The longest length of a string that has the ASI of $N-1$ is given by*

$$N_{(N-1)} = 3b + (b-1)^2 = b^2 + b + 1 \tag{12}$$

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

$$N_{(N-1)} = bN_c + 1, \tag{13}$$

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

$$\begin{aligned} H(C_{(N-1)}) = -\sum_{c=0}^{b-1} p_c \log_2(p_c) = -(b-1)\frac{N_{(N-1)}-1}{bN_{(N-1)}} \log_2\left(\frac{N_{(N-1)}-1}{bN_{(N-1)}}\right) - \frac{N_{(N-1)}-1+b}{bN_{(N-1)}} \log_2\left(\frac{N_{(N-1)}-1+b}{bN_{(N-1)}}\right) = \\ = \frac{1-b^2}{b^2+b+1} \log_2\left(\frac{b+1}{b^2+b+1}\right) - \frac{b+2}{b^2+b+1} \log_2\left(\frac{b+2}{b^2+b+1}\right) \lesssim \log_2(b). \end{aligned} \tag{14}$$

The proof of the Theorem 9 is given in Appendix C. 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), the formula (12) 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). Thus, AT subsumes information theory.

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

Theorem 10 ($C_{(N-2)}$ string)**.** *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*

$$N_{(N-2)} = b^2 + b + 4, \tag{15}$$

and

$$N_{(N-2)} = (b-2)N_c + (N_c+1) + (N_c+3) = bN_c + 4, \tag{16}$$

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

$$H(C_{(N-2)}) = -\frac{b^2-b-2}{b^2+b+4} \log_2\left(\frac{b+1}{b^2+b+4}\right) - \frac{b+2}{b^2+b+4} \log_2\left(\frac{b+2}{b^2+b+4}\right) - \frac{b+4}{b^2+b+4} \log_2\left(\frac{b+4}{b^2+b+4}\right). \tag{17}$$

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

 doi:10.20944/preprints202409.1581.v2

The proof of the Theorem 10 is given in Appendix E. In general, $C_{(N-2)}$ string must contain a clear quadruplet ($bbbb$) and a pattern binding the symbols adjoining this quadruplet, such as $[\ldots abbbbc\ldots abc\ldots]$, $[\ldots abbbbaba\ldots]$, etc., so that any $C_{(N-2)}$ string contains only one pair of repeated doublets ab, bb, or $\{bc, ba\}$. For example, for $N = 10$, sixteen bitstrings

$$\begin{array}{cccc} [0100011110], & [0111100010], & [0111101000], & [\underline{0100001110}], \\ [0001011110], & [0001111010], & [0101111000], & [0111000010] \end{array} \tag{18}$$

(an additional eight are given by swapping 0 with 1) have the ASI $a = N - 2 = 8$, where the underlined string (18) is the one that is created for $b = 2$ in Appendix E.

Theorem 11 ($C_{(N-3)}$ string)**.** *For all $b > 2$ the longest length of a string that has the ASI of $N - 3$ is given by $N_{(N-3)} = N_{(N-1)} + 6$. or equivalently by*

$$N_{(N-3)} = b^2 + b + 7, \tag{19}$$

and

$$N_{(N-3)} = (b-3)N_c + (N_c+1) + (N_c+2) + (N_c+4) = bN_c + 7, \tag{20}$$

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

$$\begin{aligned} H(C_{(N-3)}) = &-\frac{b^2-2b-3}{b^2+b+7}\log_2\left(\frac{b+1}{b^2+b+7}\right) - \frac{b+2}{b^2+b+7}\log_2\left(\frac{b+2}{b^2+b+7}\right) \\ &- \frac{b+3}{b^2+b+7}\log_2\left(\frac{b+3}{b^2+b+7}\right) - \frac{b+5}{b^2+b+7}\log_2\left(\frac{b+5}{b^2+b+7}\right). \end{aligned} \tag{21}$$

The entropy $H(C_{(N-3)}) \lessapprox \log_2(b)$ for $b \gtrapprox 2.4033$.

The proof of the Theorem 11 is given in Appendix F.

3. Conclusions

We have shown that the shape of the upper bound of the assembly index depends on the size b of the initial assembly pool P_0. The behavior of the upper ASI bound for larger b requires further research. There is one string of length $N_{(N-1)}(1) = 3$, four strings of length $N_{(N-1)}(2) = 7$, seventy-two strings of length $N_{(N-1)}(3) = 13$ (cf. Appendix D). Determining $N_{(N-1)}(b)$ for $b \geq 4$ requires further research.

Author Contributions: WB: First concept of a general method for constructing the string of length $N_{(N-1)}$ leading to Theorem 9; outline of the general Method A; proposition of Theorem 8 numerous clarity corrections and improvements; PM: outline of the general Method B; numerous clarity corrections and improvements; AT: formal proof of Theorem 3; proof that the Shannon entropy (14) can be approximated by $\log_2(b)$ for large b; numerous clarity corrections and improvements; SŁ: The remaining part of the study.

Funding: This research received no external funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

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).

 doi:10.20944/preprints202409.1581.v2

Acknowledgments: The authors thank Mariola Bala for 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 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

We start with a string of clear triplets (10). In the 1st step, we create a string containing doublets on the first subdiagonal of the matrix (11) starting with 10

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

and we append it to the string (10). 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, 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

$$\begin{aligned}
&[000111222|10|20],\\
&[000111222333|102132|03|0],\\
&[000111222333444|10213243|0314|20|40],\\
&[0001112223334445555|1021324354|031425|0415|2053|0],\\
&[00011122233344455566|102132435465|03142536|041526|2064|0516|30],\\
&[000111222333444555666777|10213243546576|0314253647|04152637|2075|051627|306174|0],\\
&[\ldots|1021324354657687|031425364758|0415263748|2086|05162738|30617285|0718|40],\\
&[\ldots|102132435465768798|03142536475869|041526374859|2097|0516273849|\\
&3061728396|071829|408195|0],\\
&[\ldots|10213243546576879\mathit{a}9|031425364758697\mathit{a}|0415263748596\mathit{a}|20\mathit{a}8|\\
&05162738495\mathit{a}|3061728394\mathit{a}7|0718293\mathit{a}|408192\mathit{a}6|091\mathit{a}|50],\\
&[\ldots|10213243546576879\mathit{a}9\mathit{ba}|031425364758697\mathit{a}8\mathit{b}|0415263748596\mathit{a}7\mathit{b}|20\mathit{b}9|\\
&05162738495\mathit{a}6\mathit{b}|3061728394\mathit{a}5\mathit{b}8|0718293\mathit{a}4\mathit{b}|408192\mathit{a}3\mathit{b}7|091\mathit{a}2\mathit{b}|50\mathit{a}1\mathit{b}6|0],\\
&[\ldots|10213243546576879\mathit{a}9\mathit{bacb}|031425364758697\mathit{a}8\mathit{b}9\mathit{c}|0415263748596\mathit{a}7\mathit{b}8\mathit{c}|20\mathit{ca}|\\
&05162738495\mathit{a}6\mathit{b}7\mathit{c}|3061728394\mathit{a}5\mathit{b}6\mathit{c}9|0718293\mathit{a}4\mathit{b}5\mathit{c}|408192\mathit{a}3\mathit{b}4\mathit{c}8|091\mathit{a}2\mathit{b}3\mathit{c}|50\mathit{a}1\mathit{b}2\mathit{c}7|0\mathit{b}1\mathit{c}|60].
\end{aligned} \tag{A3}$$

 doi:10.20944/preprints202409.1581.v2

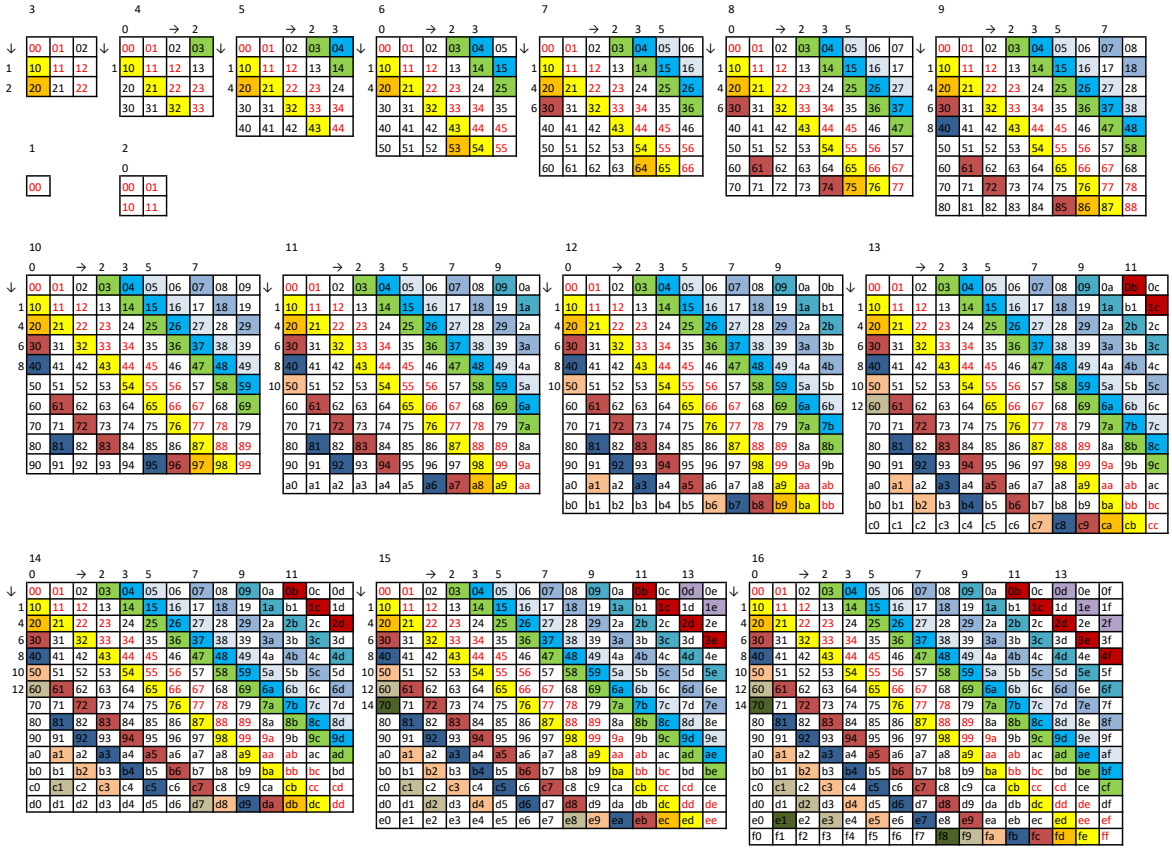

Figure A1. Doublet matrices for $1 \leq b \leq 16$ showing the creation of $N_{(N-1)}$ strings according to Method A. Colored doublets are appended to the initial string of clear triplets in the order indicated by arrows starting from the 1st column or row. Finally, 0 is appended at the end, if b is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

Appendix B. Method B

This method is similar to the Method A. We also start with a string of clear triplets (10) and the matrix of doublets (11) 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 (11)) at the end of the string (10). 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.

 doi:10.20944/preprints202409.1581.v2

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

[000111222|0210],
[000111222333|03|102132|0],
[000111222333444|04|10213243|0314|20],
[0001112223334445555|05|1021324354|031425|304152|0],
[00011122233344455566|06|102132435465|03142536|405162|041526|30],
[000111222333444555666777|07|10213243546576|0314253647|3041526374|051627|506172|0],
[...|08|102132435465768|031425364758|304152637485|05162738|607182|061728|40],
[...|09|1021324354657687|03142536475869|30415263748596|0516273849|5061728394|071829|708192|0],
[...|0a|1021324354657687a9|031425364758697a|30415263748596a7|05162738495a|
60718293a4|061728394a|8091a2|08192a|50],
[...|0b|1021324354657687a9ba|031425364758697a8b|30415263748596a7b8|05162738495a6b|
5061728394a5b6|0718293a4b|708192a3b4|091a2b|90a1b2|0],
[...|0c|1021324354657687a9bacb|031425364758697a8b9c|30415263748596a7b8c9|05162738495a6b7c|
5061728394a5b6c7|0718293a4b5c|8091a2b3c4|08192a3b4c|a0b1c2|0a1b2c|60].

(A4)

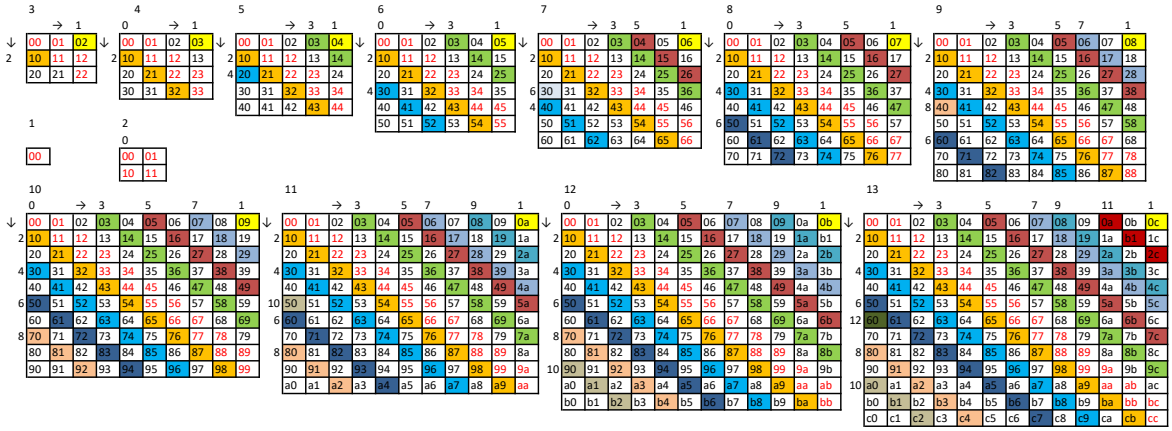

Figure A2. Doublet matrices for $1 \leq b \leq 13$ showing the creation of $N_{(N-1)}$ strings according to Method B. Colored doublets are appended to the initial string of clear triplets in the order indicated by arrows starting from the 1st column or row. Finally, 0 is appended at the end, if b is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

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

The $N_{(N-1)}$ given by the formula (12) is an odd number for all b. The first element $3b$ is the length of the initial string (10) containing b clear triplets and $b^2 - b - (b-1)$ is the number of doublets available in the matrix (11) after crossing out b doublets on its diagonal and $b-1$ doublets on its superdiagonal that are present in the starting string (10). By definition, a $C_{(N-1)}$ string cannot have any repetitions. To be the longest, it must contain all doublets in the matrix (11) 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 (13) the fractions in the Shannon entropy are

$$p_0 = \frac{N_c + 1}{N_{(N-1)}}, \quad p_{1,2,\ldots,b-1} = \frac{N_c}{N_{(N-1)}}, \tag{A5}$$

 doi:10.20944/preprints202409.1581.v2

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 (14) 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

$$\log_2\left(\frac{b+1}{b^2+b+1}\right)<0. \tag{A6}$$

Similarly the second term,

$$\log_2\left(\frac{b+2}{b^2+b+1}\right)<0. \tag{A7}$$

Hence, the entropy (14) can be approximated by the dominant contribution from the first term, which is $\log_2(b)$.

The strings given by (12) are not the shortest possible ones. Strings satisfying the equation (13) 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 (11). Furthermore, their entropies are smaller than the entropies of the strings given by the equation (12). Namely $\forall b>1$

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

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 (12) maximizes the Shannon entropy if it must additionally satisfy the relation (13). Thus, Theorem 9 is proven.

Appendix D. Number of $C^{(13,3)}_{(N-1)}$ Strings

For $b=3$, only two doublets can be introduced without repetitions into the initial string (10), leading to twelve unique strings of length $N_{(N-1)}=13$

$$\begin{aligned}&[000111222|0210],[000111222|1020],[20|21|000111222],[21|02|000111222],[0001112|02|22|10],[0001112|10|22|20],\\&[21|000|20|111222],[000|20|111222|10],[02|000111222|10],[20|00|21|0111222],[21|0001112|02|22],[21|000111222|02].\end{aligned} \tag{A9}$$

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 E. 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

$$C_k=[*000111222\ldots], \quad a_{\max}^{(N_{(N-1)}+1,b)}(C_k)=N-2. \tag{A10}$$

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

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

 doi:10.20944/preprints202409.1581.v2

leading to the repetition of the doublet $\star*$ or $*0$ but not both of them (here we allow $\star = *$). 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 $\star = 1$ and $* = 0$ and the string becomes

$$C_{(N-2)} = [010000111222\ldots], \quad a_{\max}^{(N_{(N-1)}+3,b)}(C_{(N-2)}) = N-2, \tag{A12}$$

leading to the mutually exclusive repetition of the doublet 01, 10 or 00, so that also both length and the ASI of this string increase by one. An insertion of another symbol into the string (A12) 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 (A12)

$$[0010000111222\ldots]. \tag{A13}$$

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 (A12). Thus, Theorem 10 is proven.

For the string of the form (16) the fractions in the Shannon entropy are

$$p_0 = \frac{N_c+3}{N_{(N-2)}}, \quad p_1 = \frac{N_c+1}{N_{(N-2)}}, \quad p_{2,\ldots,b-1} = \frac{N_c}{N_{(N-2)}}, \tag{A14}$$

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 (17).

Appendix F. 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 (A12) decreases of the ASI of this longer string to $a = N-4$ (cf. Appendix E. Thus, w.l.o.g. we append $* \neq 0$ at the start of the $C_{(N-2)}$ string (A12)

$$C_k = [*010000111222\ldots], \quad a_{\max}^{(N_{(N-1)}+4,b)}(C_k) = N-3. \tag{A15}$$

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

$$C_l = [\star * 010000111222\ldots], \quad a_{\max}^{(N_{(N-1)}+5,b)}(C_l) = N-3, \tag{A16}$$

has an increased length and ASI. If $b = 2$ we have

$$C_l^{(12,2)} = [\star * 0100001110], \tag{A17}$$

and

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

Hence, we must take $\star* = 01$ in the string (A17) to obtain the string

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

 doi:10.20944/preprints202409.1581.v2

that cannot be further extended along with the ASI. Therefore, $N_{(N-3)}(2) = N_{(N-1)}(2) + 5 = 12$ for $b = 2$.

For $b > 2$, w.l.o.g. we assume $\star = 1$, $* = 2$ and append 0 at the start of the string (A17) to obtain

$$C_{(N-3)} = [012010000111222\ldots], \quad a_{\max}^{(N_{(N-1)}+6,b)}(C_{(N-3)}) = N - 3, \tag{A20}$$

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. Thus, Theorem 11 is proven.

Appendix G. Misunderstanding Assembly Pools

Consider the following mapping [13] between a working assembly pool $P_3(5)$ containing five basic symbols and three strings made of these symbols and the initial assembly pool of radix $b = 8$

$$\begin{aligned} P_3(5) &\leftrightarrow P_0(8) \\ 0 &\leftrightarrow 0 \\ 1 &\leftrightarrow 1 \\ 2 &\leftrightarrow 2 \\ 3 &\leftrightarrow 3 \\ 4 &\leftrightarrow 4 \\ 20 &\leftrightarrow 5 \\ 201 &\leftrightarrow 6 \\ 2012 &\leftrightarrow 7 \end{aligned} \tag{A21}$$

Now consider the string

$$C_k^{(11,5)} = [20123242012] \tag{A22}$$

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

$$C_l^{(8,8)} = [20123247] \tag{A23}$$

of length $N = 8$ in 7 steps with the initial assembly pool $P_0(8)$ and then, using the mapping (A21), it will correspond to the string (A22). However, as we have shown in Section 2, $N_{(N-1)}(8) = 73 \neq 7$. In fact the latter string (A23) should be assembled as

$$C_m^{(5,8)} = [73247] \tag{A24}$$

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 (A21). 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 $s > 1$ ($P_1(1) = \{0, 00\}$).

References

1. Marshall, S.M.; Murray, A.R.G.; Cronin, L. A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2017**, *375*, 20160342. doi:10.1098/rsta.2016.0342.

2. Imari Walker, S.; Cronin, L.; Drew, A.; Domagal-Goldman, S.; Fisher, T.; Line, M. Probabilistic Biosignature Frameworks. In *Planetary Astrobiology*; Meadows, V.; Arney, G.; Schmidt, B.; Des Marais, D.J., Eds.; University of Arizona Press, 2019; pp. 1–1. doi:10.2458/azu_uapress_9780816540068-ch018.
3. Meadows, V.S.; Arney, G.N.; Schmidt, B.E.; Des Marais, D.J., Eds. *Planetary astrobiology*; University of Arizona space science series, The University of Arizona Press ; Houston : Lunar and Planetary Institute: Tucson, 2020. OCLC: 1151198948.
4. Liu, Y.; Mathis, C.; Bajczyk, M.D.; Marshall, S.M.; Wilbraham, L.; Cronin, L. Exploring and mapping chemical space with molecular assembly trees. *Science Advances* **2021**, *7*, eabj2465. doi:10.1126/sciadv.abj2465.
5. Marshall, S.M.; Mathis, C.; Carrick, E.; Keenan, G.; Cooper, G.J.T.; Graham, H.; Craven, M.; Gromski, P.S.; Moore, D.G.; Walker, S.I.; Cronin, L. Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nature Communications* **2021**, *12*, 3033. doi:10.1038/s41467-021-23258-x.
6. Marshall, S.M.; Moore, D.G.; Murray, A.R.G.; Walker, S.I.; Cronin, L. Formalising the Pathways to Life Using Assembly Spaces. *Entropy* **2022**, *24*, 884. doi:10.3390/e24070884.
7. Sharma, A.; Czégel, D.; Lachmann, M.; Kempes, C.P.; Walker, S.I.; Cronin, L. Assembly theory explains and quantifies selection and evolution. *Nature* **2023**, *622*, 321–328. doi:10.1038/s41586-023-06600-9.
8. Jirasek, M.; Sharma, A.; Bame, J.R.; Mehr, S.H.M.; Bell, N.; Marshall, S.M.; Mathis, C.; MacLeod, A.; Cooper, G.J.T.; Swart, M.; Mollfulleda, R.; Cronin, L. Investigating and Quantifying Molecular Complexity Using Assembly Theory and Spectroscopy. *ACS Central Science* **2024**, *10*, 1054–1064. doi:10.1021/acscentsci.4c00120.
9. Łukaszyk, S.; Bieniawski, W. Assembly Theory of Binary Messages. *Mathematics* **2024**, *12*, 1600. doi:10.3390/math12101600.
10. Raubitzek, S.; Schatten, A.; König, P.; Marica, E.; Eresheim, S.; Mallinger, K. Autocatalytic Sets and Assembly Theory: A Toy Model Perspective. *Entropy* **2024**, *26*, 808. doi:10.3390/e26090808.
11. Łukaszyk, S.; Tomski, A. Omnidimensional Convex Polytopes. *Symmetry* **2023**, *15*. doi:10.3390/sym15030755.
12. Book of John [1.3], c90.
13. Ozelim, L.; Uthamacumaran, A.; Abrahão, F.S.; Hernández-Orozco, S.; Kiani, N.A.; Tegnér, J.; Zenil, H. Assembly Theory Reduced to Shannon Entropy and Rendered Redundant by Naive Statistical Algorithms, 2024. doi:10.48550/ARXIV.2408.15108.