Article

Not peer-reviewed version

On the Salient Limits of Strings in Assembly Theory

Wawrzyniec Bieniawski , Piotr Masierak , Szymon Łukaszyk *

Posted Date: 20 September 2024

doi: 10.20944/preprints202409.1581.v1

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

Article

On the Salient Limits of Strings in Assembly Theory

Wawrzyniec Bieniawski, Piotr Masierak and Szymon Łukaszyk *

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

* szymon@patent.pl

Abstract: We show that the longest length N of a string, composed according to the principles of assembly theory of b different basic symbols, which has the assembly index of $N-1$ is given by b^2+b+1 and this odd-length string is nearly balanced.

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

---

1. Introduction

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

Definition 1. *We call a set $P_0(b) := \{0, 1, \ldots, b-1\}$ that contains different basic symbols c, where b is a finite natural radix, the initial assembly pool.*

The reader will find numerous results on AT published since 2017 in refs. [1–9].

In this short note, we extend the results of our previous study [9] 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 within the AT framework in consecutive assembly steps from basic symbols c and strings assembled in previous steps.

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

Using the Definitions 1 and 2, the assembly index 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, in contrast to the working assembly pool 2, the initial assembly pool 1 must not contain strings of basic symbols. To illustrate this, consider the following mapping between such a faulty initial assembly pool containing five basic symbols and three strings of these symbols and the initial assembly pool of radix $b=8$

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

Now consider the string

$$C_k^{(11,5)} = [20123242012] \tag{2}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2024 doi:10.20944/preprints202409.1581.v1

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

$$C_l^{(8,8)} = [20123247] \tag{3}$$

of length $N = 8$ in 7 steps with the initial assembly pool $P_0(8)$ and then, using the mapping (1), it will correspond to the string (2). However, as we shall show in the following section, $N_{\text{max}}(8) = 73 \neq 7$. In fact the latter string (3) should be assembled as

$$C_m^{(5,8)} = [73247] \tag{4}$$

with the assembly index $a^{(5,8)}(C_k) = 5 - 1 = 4$ and with the initial assembly pool $P_0(8)$, as $2012 \leftrightarrow 7$ according to (1).

The following two theorems were already stated in our previous study [9] for $b = 2$. We restate them here $\forall b$ for clarity.

Theorem 1. *A string of length $N = 4$ is the shortest string that allows for more than one assembly index for all b.*

Proof. $N = 2$ provides b^2 available strings with unit assembly indices. $N = 3$ provides b^3 available strings with assembly indices equal to two. Only $N = 4$ provides b^4 strings that include b strings $C_k^{(4,b)} = [* * **]$ and $b(b-1)$ strings $C_k^{(4,b)} = [* \star *\star]$ with assembly indices equal to two, while the assembly index of the remaining strings is three. For example, to assemble the string $C_k^{(4,4)} = [0202]$, we need to assemble the string $[02]$ and reuse it from P_1, while there is nothing available to reuse, in the case of the string $C_l^{(4,4)} = [0123]$. □

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

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

Proof. Strings C_{min} for which $a^{(N)}(C_{\text{min}}) = \min_k\left(\{a^{(N,b)}(C_k)\}\right)$, $\forall k \in \{1, 2, \ldots, b^N\}$ 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 an assembly index in this case, including b strings

$$C_k^{(2^s,b)} = [* * \ldots], \tag{5}$$

and $b(b-1)$ strings

$$C_k^{(2^s,b)} = [* \star * \star \ldots], \tag{6}$$

and the assembly pathway of each of the strings (5) and (6) 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. Thus, the lower bound for s of the addition chain for N, $s \geq \log_2(N)$ is achieved for $N = 2^s$ by strings (5) and (6) .

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 an assembly index 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 $\{0, 1, 2\}$. $\square$

2. Results

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

$$[*******] \quad \text{and} \quad [\star\star\star\star\star\star\star], \tag{7}$$

and their lengths cannot be increased without a repetition of a doublet, which inevitably reduces the assembly index to $a_{\max}^{(8,2)} = 8 - 2 = 6$.

This observation and Theorem 2 motivated us to develop a general procedure to construct the longest possible string that has the assembly index $a_{\max}^{(N,b)} = N - 1$, as a function of the radix $b \geq 3$. We denote the length of this string by $N_{\max}(b)$.

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

$$[0001112\ldots(b-2)(b-1)(b-1)(b-1)]. \tag{8}$$

The doublets that can be inserted into the initial string (8) can be arranged in a $b \times b$ matrix

$$\begin{bmatrix} \xcancel{00} & 01 & 02 & \ldots & 0(b-1) \\ 10 & \xcancel{11} & 12 & \ldots & 1(b-1) \\ 20 & 21 & \xcancel{22} & \ldots & 2(b-1) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (b-2)0 & (b-2)1 & (b-2)2 & \ldots & (b-2)(b-1) \\ (b-1)0 & (b-1)1 & (b-1)2 & \ldots & \xcancel{(b-1)(b-1)} \end{bmatrix}, \tag{9}$$

where the crossed out entries on 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 (8) and hence we can also cross out the entries on the first superdiagonal in the matrix (9).

In the 1$^{\text{st}}$ step, we create a string containing doublets on the first subdiagonal of the matrix (9) starting with 10

$$[102132\ldots(b-2)(b-3)(b-1)(b-2)], \tag{10}$$

and we append it to the string (8). 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 2$^{\text{nd}}$ step, we create a string containing doublets on the third superdiagonal beginning with the doublet 03

$$[0314\ldots(b-5)(b-2)(b-4)(b-1)], \tag{11}$$

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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2024 doi:10.20944/preprints202409.1581.v

We shall illustrate this process for $b = 8$. The matrix

$$\begin{bmatrix} \cancel{00} & \cancel{01} & \cancel{02} & 03 & 04 & 05 & \cancel{06} & \cancel{07} \\ 10 & \cancel{11} & \cancel{12} & \cancel{13} & 14 & 15 & 16 & \cancel{17} \\ 20 & 21 & \cancel{22} & \cancel{23} & \cancel{24} & 25 & 26 & 27 \\ 30 & \cancel{31} & 32 & \cancel{33} & \cancel{34} & \cancel{35} & 36 & 37 \\ \cancel{40} & \cancel{41} & \cancel{42} & 43 & \cancel{44} & \cancel{45} & \cancel{46} & 47 \\ \cancel{50} & \cancel{51} & \cancel{52} & \cancel{53} & 54 & \cancel{55} & \cancel{56} & \cancel{57} \\ \cancel{60} & 61 & \cancel{62} & \cancel{63} & \cancel{64} & 65 & \cancel{66} & \cancel{67} \\ \cancel{70} & \cancel{71} & \cancel{72} & \cancel{73} & 74 & 75 & 76 & \cancel{77} \end{bmatrix}, \tag{12}$$

contains all the doublets that were used to create the string of length $N_{\max}(8) = 73$

$$\begin{gathered} [000111222333444555666777|10213243546576| \\ 0314253647|04152637|2075|051627|306174|0], \end{gathered} \tag{13}$$

For $b = 7$ we would obtain the string of length $N_{\max}(7) = 57$

$$\begin{gathered} [000111222333444555666|102132435465| \\ 03142536|041526|2064|0516|30]. \end{gathered} \tag{14}$$

for $b = 6$ we would obtain the string of length $N_{\max}(6) = 43$

$$[000111222333444555|1021324354|031425|0415|2053|0], \tag{15}$$

for $b = 5$ we would obtain the string of length $N_{\max}(5) = 31$

$$[000111222333444|10213243|0314|04|20], \tag{16}$$

for $b = 4$ we would obtain the string of length $N_{\max}(4) = 21$

$$[000111222333|102132|03|0], \tag{17}$$

and $b = 3$ leads to the following string of length $N_{\max}(3) = 13$

$$[000111222|10|20]. \tag{18}$$

The final string is always terminated by 0.

The strings of odd lengths generated by the general procedure outlined above are not only the longest, but also the most balanced. This leads to the following theorem.

Theorem 3. *The longest length $N_{max}(b)$ of a string composed of b different basic symbols that has the assembly index of $N - 1$ is given by*

$$N_{max}(b) = 3b + (b-1)^2 = b^2 + b + 1 \tag{19}$$

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

$$N_{max}(b) = bN_c(b) + 1, \tag{20}$$

where $N_c = b + 1$ is the number of occurrences of all but one symbol within the string.

Proof. The $N_{\max}(b)$ given by formula (19) is an odd number for all b. As shown in Table 1, the first element $3b$ is the length of the initial string (8) containing b clear triplets and $(b-1)^2$ is the number

 doi:10.20944/preprints202409.1581.v

of entries in the doublet matrix (9) of the previous b. By definition, a string of length $N_{\max}(b)$ cannot have any repetitions; it can only contain doublets and clear triplets that do not contain these doublets. Therefore, to be the most patternless, this string must maximize Shannon entropy; must be the most balanced. For the string of the form (20) the fractions in the Shannon entropy are

$$p_0 = \frac{N_c(b)+1}{N_{\max}(b)}, \quad p_{1,2,\dots,b-1} = \frac{N_c(b)}{N_{\max}(b)}, \tag{21}$$

(where without loss of generality we assume that the symbol occuring $N_c(b)+1$ times within the string is $c=0$) and the Shannon entropy is

$$\begin{aligned} H_{\max}(b) = -\sum_{c=0}^{b-1} p_c \log_2(p_c) = -(b-1)\frac{N_{\max}(b)-1}{bN_{\max}(b)} \log_2\left(\frac{N_{\max}(b)-1}{bN_{\max}(b)}\right) - \frac{N_{\max}(b)-1+b}{bN_{\max}(b)} \log_2\left(\frac{N_{\max}(b)-1+b}{bN_{\max}(b)}\right) = \\ = \frac{1-b^2}{b^2+b+1} \log_2\left(\frac{b+1}{b^2+b+1}\right) - \frac{b+2}{b^2+b+1} \log_2\left(\frac{b+2}{b^2+b+1}\right) \approx \log_2(b). \end{aligned} \tag{22}$$

The strings given by the equation (19) are not the shortest possible ones. Strings satisfying the equation (20) and satisfying $\min(bN_c(b)+1) > N_{\max}(b-1)$ are given by b^2+1 (OEIS A002522). However, they do not contain all the possible doublets and furthermore their entropies are smaller than the entropies of the strings given by the equation (19).

Now, assume *a contrario* that a string longer than $N_{\max}(b)$ can be constructed, say of length $N_{\max}(b)' = N_{\max}(b)+1$. But in this case, the corresponding $H_{\max}(b)' < H_{\max}(b)$. The string of the length given by the formula (19) maximizes the Shannon entropy if it must additionally satisfy the relation (20). □

Although the case for $b=1$ (only one symbol) is degenerate, the formula (19) yields correct result; the string [000] is the longest string with $a_{\max}^{(N,1)} = N-1$, as for $b=1$ we simply have $a_{\max}^{(N,1)} = a_{\min}^{(N)}$ (OEIS A003313).

Table 1. The maximum length of a string having the assembly index $a_{\max}^{(N_{\max}(b),b)} = N_{\max}(b)-1$ and their Shannon entropies, as a function of the radix b.

b	$N_{\max}(b)$	$H_{\max}(b)$
1	$3+0=3$	0
2	$6+1=7$	0.9852
3	$9+4=13$	1.5766
4	$12+9=21$	1.9952
5	$15+16=31$	2.3190
6	$18+25=43$	2.5831
7	$21+36=57$	2.8061
8	$24+49=73$	2.9991
9	$27+64=91$	3.1692
10	$30+81=111$	3.3214
11	$33+100=133$	3.4590
12	$36+121=157$	3.5846
13	$39+144=183$	3.7002

3. Conclusions

There is one string of length $N_{\max}(1)=3$, four strings of length $N_{\max}(2)=7$, seventy-two strings of length $N_{\max}(3)=13$ (cf. Appendix A). Their number for $b \geq 4$ requires further research.

Author Contributions: WB: First concept of a general procedure for constructing the string of length $N_{\max}(b)^{(A)}$; determining $N_{\max}(b)$ for $3 \leq b \leq 9$; third concept of a general procedure for constructing the string of length $N_{\max}(b)$ leading to theorem 3; noting that $N_{\max}(b)$ must be more balanced than $N_{\max}(b)^{(B)}$; numerous clarity corrections and improvements; PM: Second concept of a general procedure for constructing the string of length $N_{\max}(b)^{(B)}$; numerous clarity corrections and improvements; SŁ: The remaining part of the study.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2024 doi:10.20944/preprints202409.1581.v1

Acknowledgments: The authors thank Andrzej Tomski for numerous clarity corrections and improvements and Mariola Bala for motivation. SŁ thanks his wife, Magdalena Bartocha, for her everlasting support.

Appendix A. Method A

In the first method of creating the longest, patternless string we developed, we started with the string of clear triplets (8) which we augmented with $b-1$ doublets $\{02, 03, \ldots, 0(b-1), 10\}$ to form the string

$$[0001\ldots(b-2)(b-1)(b-1)(b-1)0203\ldots0(b-1)10], \tag{A1}$$

of length

$$N_{\max}(b)^{(\mathrm{A})} = 3b + 2(b-1) = 5b-2. \tag{A2}$$

The introduction of $b-1$ doublets from the first row of the (9) and the doublet 10 into the string (8) also introduces other doublets. For $b=3$ the augmented string (A1) is has the length $N_{\max}(3) = 13$ as the insertion of 0210 at the end of the string (8) introduces the doublet 21. Thus, by construction, doublet $(b-1)1$ (last row, 2^{nd} column) cannot be reused. For $b=3$ only two doublets can be introduced without repetitions, leading to twelve unique strings of length $N = 13$

$$\begin{aligned}
&[000111222|0210], [000111222|1020], [20|21|000111222],\\
&[21|02|000111222], [0001112|02|22|10], [0001112|10|22|20],\\
&[21|000|20|111222], [000|20|111222|10], [02|000111222|10],\\
&[20|00|21|0111222], [21|0001112|02|22], [21|000111222|02],
\end{aligned} \tag{A3}$$

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 $N_{\max}(3) = 13$. This method turned out to be valid for $b = \{1, 3\}$ only, as $N_{\max}(4)^{(\mathrm{A})} = 18 \neq 21$ and not all available doublets were used.

Appendix B. Method B

The second method we developed is an extension of the first one A. We start with the augmented string (A1) and the doublet matrix

$$\begin{bmatrix}
\cancel{00} & \cancel{01} & \cancel{02} & \cancel{03} & \ldots & \cancel{0(b-1)}\\
\cancel{10} & \cancel{11} & \cancel{12} & 13 & \ldots & 1(b-1)\\
\cancel{20} & 21 & \cancel{22} & \cancel{23} & \ldots & 2(b-1)\\
\cancel{30} & 31 & 32 & \cancel{33} & \ldots & 3(b-1)\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
\cancel{(b-2)0} & (b-2)1 & (b-2)2 & (b-2)3 & \ldots & \cancel{(b-2)(b-1)}\\
\cancel{(b-1)0} & \cancel{(b-1)1} & (b-1)2 & (b-1)3 & \ldots & \cancel{(b-1)(b-1)}
\end{bmatrix} \tag{A4}$$

For $b > 3$, there are $(b-3)(b-2)/2$ and $(b-2)(b-1)/2 - 1$ (in total $(b-3)(b-1)$) doublets available, respectively, in the upper triangle (beginning with 13) and lower triangle (beginning with 21) of the doublet matrix (A4). We can insert the doublets from the upper triangle to the augmented string (A1) as follows

$$[0001112\underline{13}223\underline{1424}334\ldots(b-3)(b-2)\underline{1(b-1)2(b-1)\ldots(b-3)(b-1)}(b-2)(b-2)(b-1)(b-1)(b-1)0203\ldots0(b-1)10], \tag{A5}$$

creating a string of length

$$N_{\max}(b)^{(\mathrm{B})} = 3b + 2(b-1) + (b-3)(b-2) = b^2 + 4. \tag{A6}$$

This method also turned out to be valid for $b = 3$ only, as $N_{\max}(4)^{(B)} = 20 \neq 21$ even though the strings of length $N_{\max}(b)^{(B)}$ contain all doublets of the matrix (9) without repeating. However, the strings (A5) created by this method are non-balanced and do not contain all available b clear triplets. For example, for $b = 4$, the non-balanced string (A5) of length $N_{\max}(4)^{(B)} = 20$ that contains all possible doublets (but no clear triplet 222) is

$$[00011121322333020310]. \tag{A7}$$

This led us to the third method described in Section 2.

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.