Integer Sequences in the HP Model of Dill

Slav Angelov(✉) and Latchezar Tomov

New Bulgarian University, Sofia, Bulgaria
{sangelov,lptomov}@nbu.bg

Abstract. We examine a sequence that we derive from Dill's hydrophobic-polar protein folding model in a two-dimensional square lattice. This sequence already exists as A248333 in the OEIS, but the information is scarce. We construct two more sequences not previously listed in the OEIS, A342711 and A342712. Thanks to the perspective given us by Dill's model, we find relationships between sequences 248333, A240025, A000267, A342711, A342712, 000027, and A000217. We use the newly found relationships to obtain an explicit formula for the nth member of A248333 and a formula for the nth member of its partial sums, A342712. Moreover, we find an explicit formula for the nth member of A342711 and present an alternative proof for the nth member of A000267.

Keywords: Dill's HP model · Square lattice · Quarter square · Spiral · Positive integer sequence

1 Introduction

The sequence that we are going to discuss is listed in the OEIS [4] as A248333, first documented by Wesley Hurt in 2014. The first terms of the sequence are as follows:

$$0, 0, 0, 0, 1, 1, 2, 2, 3, 4, 4, 5, 6, 6, 7, 8, 9, 9, 10, 11, 12, 12, 13, 14, 15, 16, 16, 17, 18, 19,$$
$$20, 20, 21, 22, 23, 24, 25, 25, 26, 27, 28, 29, 30, 30, 31, 32, 33, \ldots.$$

Sequence A248333 emerges while trying to count the total number of squares that are formed while adding adjacent points on a square lattice. The complete idea behind the creation of the sequence can be seen in Fig. 1.

Hurt noticed that the pattern fails to add a square for $n > 0$ if n is of the form $k^2 + 1$ (A002522) or $k^2 - k + 1$ (A002061). This is all that was known till now. We will continue this paper by presenting a slightly different way of deriving A248333. This will help us to obtain an explicit formula for the nth member of the sequence and a formula for the sum of the first n members. Moreover, we will derive and analyze some other sequences related to A248333.

T. Zlateva and R. Goleva (Eds.): CSECS 2022, LNICST 450, pp. 154–167, 2022.
https://doi.org/10.1007/978-3-031-17292-2_13

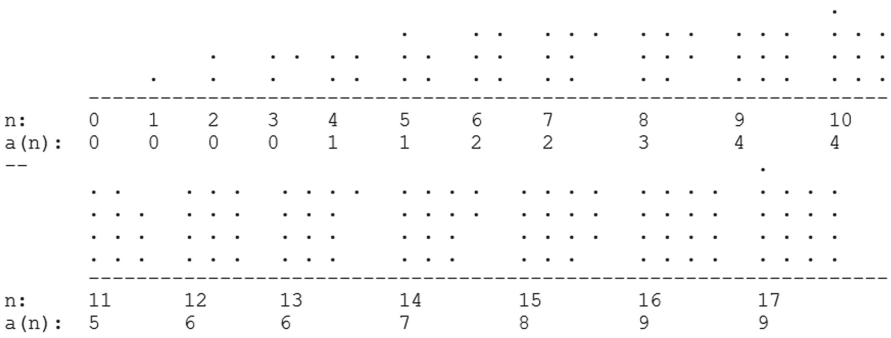

Fig. 1. Counting the squares while observing the added points.

2 Deriving the Sequence

In order to present the new approach for deriving A248333, we will make a brief introduction to a problem in biology. We may observe the proteins as the building blocks of the living organisms. Each protein consists of a strictly determined sequence of amino acids (there might be some additions to the protein's structure, but we will ignore them for simplicity). If we put a sequence of amino acids in a liquid environment, it will always take one and the same shape in space. This shape determines the unique properties of each protein. If the biologists have an away to determine what will be the shape based on the amino acid sequence, this will help them to create methods against incurable diseases. The problem is that finding a protein structure is not an easy task. Dill was one of the first scientists who proposed a possible approach to the protein structure problem, see paper [2]. He noticed that some of the amino acids are hydrophilic (H) while the rest are hydrophobic (P). Then he classified all of the main amino acid types into the two mentioned groups. Thus, a protein can be observed as a finite strictly determined sequence of Hs and Ps, we will refer to them as HP-sequences. Dill observed HP-sequences in a square lattice where he claimed that to obtain the protein structure, we need to locate the HP-sequence in a way that maximizes the number of *contacts* between the Hs. We say that there is a *contact* if two Hs are adjacent on the lattice and are not next to each other on the HP-sequence. The described model is called hydrophobic-polar protein folding model (HP model).

Finding the maximum number of contacts in the HP model is an NP-hard task even on a two-dimensional square lattice. Thus, many heuristic algorithms were created, see papers [1,5–8], and many others. In general, the upper bound for the possible number of contacts on a two-dimensional square lattice is equal to $2 \times min\{ODD, EVEN\}$, where ODD is the total number of Hs in the chosen HP-sequence that are on odd positions (respectively for $EVEN$). This boundary is exploited by the theory, e.g., see this article [3]. The upper bound is obtained thanks to the fact that it is not possible to realize a contact between Hs that are both odd or both even on a square lattice. In order to refine the upper

bound for specific cases, we examined HP-sequences that are entirely from Hs. We maximized the total number of contacts by putting HP-sequences of size n in the form of a spiral, see Fig. 1. We noted that the total number of contacts for a spiral with n elements corresponds to the nth member of A248333. This is not an unexpected fact, if we observe carefully, we see that counting contacts in the spiral is equivalent to counting squares while adding points like Hurt did, see Fig. 1.

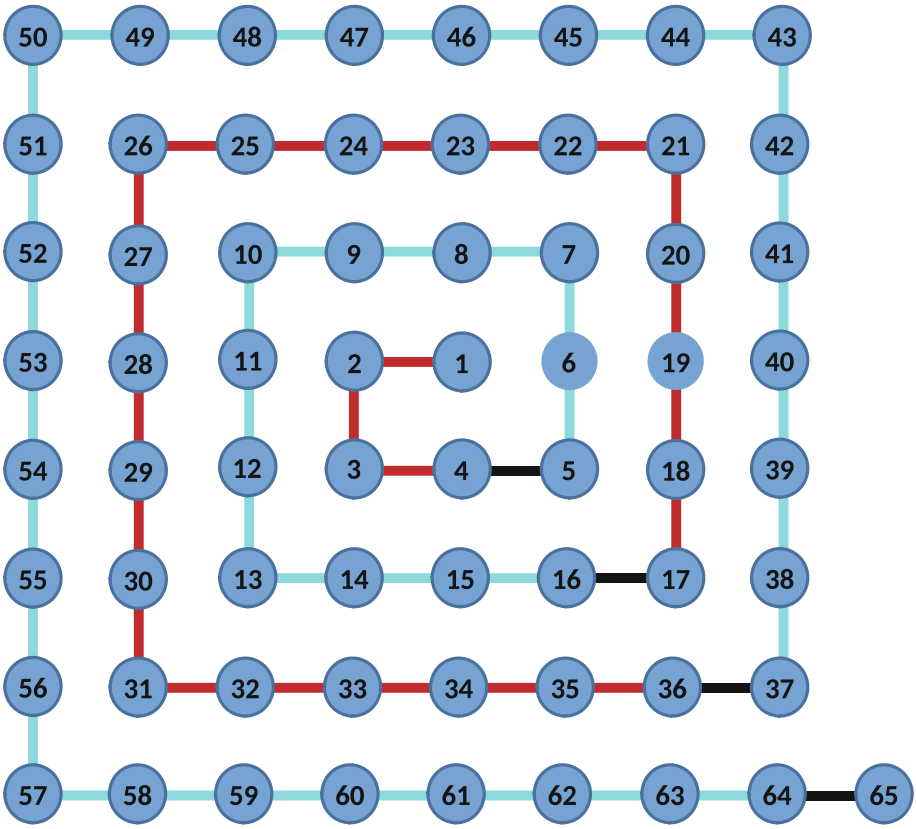

Fig. 2. Aminoacid chain on a square lattice ordered in the form of a spiral.

3 Definitions and Lemmas

We analyzed some structures in the spiral in Fig. 2 to help us for the further analysis in this work. We defined these structures in the next several definitions and described some of their properties.

Definition 1. *We distinguish* layers *in the spiral.* Layer 1 *includes the elements indexed by 1, 2, 3, and 4 in the spiral.* Layer 2 *includes the elements that fully surround* Layer 1*, i.e., 4, 5, 6,...,16.* Layer 3 *includes the elements that fully surround* Layer 2*, and so on. We denote* Layer N *by* L_N*.*

In Fig. 2, layers are easily distinguishable thanks to the colors blue and red, the back color shows the transition between the layers. Note that each layer is in the form of a square. We examined some properties of the newly defined layers.

Lemma 1. *The number of elements in the* N*th layer are* $8N - 4$*.*

Proof. The number of elements in each layer is as follows:

1. $4 \rightarrow 4 + 0 \times 8$;
2. $12 \rightarrow 4 + 1 \times 8$;
3. $20 \rightarrow 4 + 2 \times 8$;
4. $28 \rightarrow 4 + 3 \times 8$;
5. ...

Let the observed relations hold for Layer N, so we have $4+8(N-1)$. If we observe Fig. 2, we see that the elements in each layer form a square. If we transpose the sides of this square to the upper layer, we will need to add exactly 4 more elements to complete the upper layer, and we have duplicated the elements in each corner of the previous layer. Thus, the total number of elements added is 8. For Layer N+1, we have $(4 + 8(N - 1)) + 8 = 4 + 8N$. The expected formula for Layer N+1 is $4 + 8N$, and is identical to the obtained one. We proved the lemma by induction. The only think left is to simplify the formula for Layer N: $4 + 8(N - 1) = 8N - 4$.

Lemma 2. *The total number of contacts in the* N*th layer with the previous ones is* $C(L_N) = \#L_N - 4 = 8(N - 1)$*,* $N > 1$*.*

Proof. Note that the elements in L_N do not form a contact only at the beginning, and on the three consecutive corners of the square, the total number is 4. If we denote the total number of contacts in L_N by $C(L_N)$, we have $C(L_N) = \#L_N - 4 = 8(N - 1)$. The only exception is Layer 1, where we have only one contact, and it is with itself.

Definition 2. *A* level *in the spiral is a structure that includes all of the layers till the index of the level. We denote the* N*th* level *by* Lv_N *or* Level N*, and by* $Lv(n)$ *the maximum possible* level *which has elements less than* n*.*

Level 1 consists of Layer 1; Lv_2 consists of L_1 and L_2; Lv_3 consists of L_1, L_2, and L_3; and so on.

Lemma 3. *The number of elements* $\#Lv_N$ *in the* N*th level is* $(2N)^2$*.*

Proof. The number of elements in the Nth level is equal to the sum of the elements in all of the layers in this level:

$$\#Lv_N = \sum_{i=1}^{N}(8i - 4) = 8\sum_{i=1}^{N} i - 4N = 8\frac{N(N+1)}{2} - 4N = 4N^2.$$

Lemma 4. *The total number of contacts in the Nth level is $C(Lv_N) = (2N - 1)^2$.*

Proof.

$$C(Lv_N) = \sum_{i=1}^{N} C(L_i) = \sum_{i=1}^{N} (8(i-1)) + 1 = 8\sum_{i=1}^{N} i - 8N + 1 =$$
$$= 8\frac{N(N+1)}{2} - 8N + 1 = 4N^2 - 4N + 1 = (2N-1)^2.$$

Corollary 1. *The total number of contacts in the Nth level is equal to the sum of the first m odd numbers, where $m = 2N - 1$.*

Proof. We will use the following well-known formula $1+2+3+5+\ldots+(2m-1) = m^2$. It states that the sum of the first m odd numbers is equal to the square of their number. We can use $m = 2N - 1$ to obtain the desired result.

Lemma 5. *The maximum possible level $Lv(n)$ of a given integer n is equal to $\lfloor \frac{\sqrt{n}}{2} \rfloor$, function $\lfloor \ldots \rfloor$ is the floor function.*

Proof. Note that the expression $\lfloor \frac{\sqrt{n}}{2} \rfloor$ gives an integer for $n = 4N^2$, $N \in N$ even without the floor function, this happens when n coincides with the number of elements in a level, see Lemma 3. Let observe the case where $n = \#Lv_N + e$ is between Lv_N and Lv_{N+1}:

$$4N^2 < n < 4(N+1)^2 \leftrightarrow \frac{\sqrt{4N^2}}{2} < \frac{\sqrt{n}}{2} < \frac{\sqrt{4(N+1)^2}}{2} \leftrightarrow N < \frac{\sqrt{n}}{2} < N + 1.$$

Thus, we obtain N if we use $\left\lfloor \frac{\sqrt{n}}{2} \right\rfloor$ for any n between $4N^2$ and $4(N+1)^2$.

4 Some Other Related Sequences

We, again, observe Fig. 2. Let observe the spiral from the beginning layer by layer, and put 1 where the aminoacids do not form a contact with the previous layer, and 0 elsewhere. We will obtain the following sequence:

$n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, \ldots$
$a(n) = 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, \ldots$

We denote this sequence by A_0, and its nth member by $A_0(n)$. The nth member of sequence A_0 corresponds to the $n-1$th member of sequence A240025 in the OEIS (small difference in the indexing). We keep our indexing because it better reflects the framework of the HP model that we are using.

If we observe the partial sums of A_0, we will obtain the following sequence:

$n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, \ldots$
$a(n) = 1, 2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, \ldots$

Let denote this sequence by S_{A_0} and its nth member by $S_{A_0}(n)$. The nth member of sequence S_{A_0} corresponds to the $n-1$th member of sequence A000267 in the OEIS.

Now, if we observe the partial sums of S_{A_0}:

$$n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, \ldots$$
$$a(n) = 1, 3, 6, 9, 13, 17, 22, 27, 32, 38, 44, 50, 57, 64, 71, 78, 86, 94, 102, 110, 119, \ldots$$

Let denote this sequence by SS_{A_0} and its nth member by $SS_{A_0}(n)$. The nth member of sequence SS_{A_0} corresponds to the $n-1$th member of sequence A342711 in the OEIS.

Finally, we will denote our initial sequence A248333 by A and its nth member by $A(n)$. While observing the partial sums of A, we obtain the following sequence:

$$n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, \ldots$$
$$a(n) = 0, 0, 0, 1, 2, 4, 6, 9, 13, 17, 22, 28, 34, 41, 49, 58, 67, 77, 88, 100, 112, 125, \ldots$$

We denote this sequence by S_A and its nth member by $S_{A(n)}$. Sequence S_A corresponds to A342712 in the OEIS.

5 Some Explicit Formulas for the Presented Sequences

Proposition 1. *The nth member of sequence S_{A_0} is estimated by the following formula:*

$$S_{A_0}(n) = 3N + q - 1 + \left\lceil \frac{\sqrt{n}}{2} \right\rceil,$$

where:

1. $N = \left\lfloor \frac{\sqrt{n}}{2} \right\rfloor$;
2. $q = \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$.

The function $\lfloor \ldots \rfloor$ is the floor function, and the function $\lceil \ldots \rceil$ is the ceil function.

Proof. Sequence A_0 highlights the elements that do not form a contact in the spiral, see Fig. 2, we have 1 there, 0 if there is a contact. Note that A_0 is 1 only on the three of possible four formed corners of a layer (each layer has a square form) of the spiral, and at the first element of the next layer. We represent n as $n = \#Lv_N + R$, where $\#Lv_N = Lv(n)$ is the number of elements of the maximum level which has elements less than n, and $R = n - N$ is what is left from n. Sequence S_{A_0} represents the sum of the first n members of A_0. Let $n = \#Lv_N$, $N = 1, 2, \ldots$. Then $S_{A_0}(n) = 4N - 1$ - for each layer we have 1 four times, in the Nth layer 3 times. Let $n \neq \#Lv_N$. Then $n = \#Lv_N + R$, $0 < R < \#L_{N+1}$. We need to determine the ones in R. The elements in R are part from Layer $N+1$. Each layer can be observed as a square with 4 sides, each side has $\frac{\#L_{N+1}}{4} = 2N + 1$ elements (we used Lemma 1). We have one 1 at the

end of each side. The total number of sides that are covered by the elements of R is equal to R divided by the number of elements in one of the sides of Layer N which is $\left\lfloor \frac{R}{2N+1} \right\rfloor = \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$. Thus, we obtained $S_{A_0}(n) = 4N - 1 + \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$. Because there is one additional 1 at the beginning of Layer $N+1$, we must correct the formula by replacing "1" with $F = \left\lfloor N + 1 - \frac{\sqrt{n}}{2} \right\rfloor$. The fix F is 1 only if n coincides with the number of elements of a level, i.e., $R = 0$, it is 0 in all other cases. We simplified F by using the following link between the floor and ceil functions: $\lfloor -a \rfloor = -\lceil a \rceil$.

$$F = \left\lfloor N + 1 - \frac{\sqrt{n}}{2} \right\rfloor = N + 1 + \left\lfloor -\frac{\sqrt{n}}{2} \right\rfloor = N + 1 - \left\lceil \frac{\sqrt{n}}{2} \right\rceil.$$

$$S_{A_0}(n) = 4N - 1 + \left\lfloor \frac{n - 4N^2}{2N+1} \right\rfloor - F = 3N + q - 1 + \left\lceil \frac{\sqrt{n}}{2} \right\rceil.$$

Theorem 1. *An explicit formula for the nth member of sequence S_{A_0} is as follows:*

$$S_{A_0}(n) = \lceil 2\sqrt{n} \rceil - 1,$$

where the function $\lceil \ldots \rceil$ is the ceil function.

Proof. We count the number of ones in each level in the context of sequences A_0 and S_{A_0}:

1. The total number is 3;
2. The total number is $3 + 4$;
3. The total number is $3 + 2 \times 4$;
4. The total number is $3 + 3 \times 4$;
5. ...

By induction (we skip the induction phase for clarity, see the proof of Lemma 1 for a similar approach), we obtain the following formula for Layer N: $3+4(N-1)$. Now, $\#Lv_N = 4N^2$. Thus, $n = 4N^2$, and $N = \frac{\sqrt{n}}{2}$. Finally, $S_{A_0}(n) = 3+4(\frac{\sqrt{n}}{2} - 1) = 2\sqrt{n} - 1$. The final expression is not always an integer because n cannot always coincide with $\#Lv_N$. However, if we round it to the upper integer, we will obtain the desired result: $S_{A_0}(n) = \lceil 2\sqrt{n} - 1 \rceil = \lceil 2\sqrt{n} \rceil - 1$. Now, we will show that the derived express for $S_{A_0}(n)$ is equivalent to the expression in Proposition 1:

$$\lceil 2\sqrt{n} \rceil - 1 \stackrel{?}{=} 3N + q - 1 + \left\lceil \frac{\sqrt{n}}{2} \right\rceil \Leftrightarrow \left\lceil \frac{4\sqrt{n}}{2} \right\rceil \stackrel{?}{=} 3N + q + \left\lceil \frac{\sqrt{n}}{2} \right\rceil. \quad (1)$$

From Lemma 5, we have that $\lfloor \frac{\sqrt{n}}{2} \rfloor = N \Rightarrow \frac{\sqrt{n}}{2} = N + r, 0 \leq r < 1$. We substitute the derived expression for $\frac{\sqrt{n}}{2}$ in (1):

$$\left\lceil \frac{4\sqrt{n}}{2} \right\rceil \stackrel{?}{=} 3N + q + \left\lceil \frac{\sqrt{n}}{2} \right\rceil \Leftrightarrow \lceil 4(N + r) \rceil \stackrel{?}{=} 3N + q + \lceil N + r \rceil. \quad (2)$$

Additionally, using Lemma 3, we can express n by $n = N^2 + R, R < 4(2N + 1)$. Let observe $R = 4(2N + 1)$. Then $n = (N + 1)^2$, and it coincides with the next level. If we observe $q = \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$ (see Proposition 1), we note that:

- $q = 0$, $0 \le R < 2N + 1$;
- $q = 1$, $2N + 1 \le R < 2(2N + 1)$;
- $q = 2$, $2(2N + 1) \le R < 3(2N + 1)$;
- $q = 3$, $3N + 1 \le R < 4(2N + 1)$;
- $q = 4$, $R = 4(2N + 1)$.

Finally, we need to obtain a relationship between R and r based on $\frac{\sqrt{n}}{2}$:

- For $R = 0$, we have $\frac{\sqrt{n}}{2} = \sqrt{\frac{n}{4}} = \sqrt{\frac{4N^2+0}{4}} = N + 0 = N + r,\ r = 0$;
- For $R = 2N + 1$, $\sqrt{\frac{n}{4}} = \sqrt{\frac{4N^2+2N+1}{4}} = \sqrt{N^2 + \frac{N}{2} + \frac{1}{16} + \frac{1}{16}}$ $= \sqrt{(N + \frac{1}{4})^2 + \frac{1}{16}} > N + \frac{1}{4} = N + r,\ r = \frac{1}{4}$;
- For $R = 2(2N + 1)$, $\sqrt{\frac{n}{4}} = \sqrt{\frac{4N^2+2(2N+1)}{4}} = \sqrt{N^2 + N + \frac{1}{4} + \frac{1}{4}}$ $= \sqrt{(N + \frac{1}{2})^2 + \frac{1}{4}} > N + \frac{1}{2} = N + r,\ r = \frac{1}{2}$;
- For $R = 3(2N + 1)$, $\sqrt{\frac{n}{4}} = \sqrt{\frac{4N^2+3(2N+1)}{4}} = \sqrt{N^2 + \frac{6N}{4} + \frac{9}{16} + \frac{3}{16}}$ $= \sqrt{(N + \frac{3}{4})^2 + \frac{3}{16}} > N + \frac{3}{4} = N + r,\ r = \frac{3}{4}$;
- For $R = 4(2N + 1)$, we have $\frac{\sqrt{n}}{2} = \sqrt{\frac{n}{4}} = \sqrt{\frac{4N^2+4(2N+1)}{4}} = N + 1 = N + r,\ r = 1$.

So, we obtained the following relationships:

- $n = 4N^2 + R$, $0 \le R < 2N + 1 \ \Rightarrow\ \frac{\sqrt{n}}{2} = N + r$, $0 \le r \le \frac{1}{4}$;
- $n = 4N^2 + R$, $2N + 1 \le R < 2(2N + 1) \ \Rightarrow\ \frac{\sqrt{n}}{2} = N + r$, $\frac{1}{4} < r \le \frac{1}{2}$;
- $n = 4N^2 + R$, $2(2N + 1) \le R < 3(2N + 1) \ \Rightarrow\ \frac{\sqrt{n}}{2} = N + r$, $\frac{1}{2} < r \le \frac{3}{4}$;
- $n = 4N^2 + R$, $3(2N + 1) \le R < 4(2N + 1) \ \Rightarrow\ \frac{\sqrt{n}}{2} = N + r$, $\frac{3}{4} < r < 1$.

Now, from Expression 2, we can check $\lceil 4(N + r) \rceil \stackrel{?}{=} 3N + q + \lceil N + r \rceil$:

- For $n = 4N^2 + R$, $0 \le R < 2N + 1$, we have $4N + 1 \stackrel{?}{=} 3N + 0 + N + 1 = 4N + 1$;
- For $n = 4N^2 + R$, $2N + 1 \le R < 2(2N + 1)$, we have $4N + 2 \stackrel{?}{=} 3N + 1 + N + 1 = 4N + 2$;
- For $n = 4N^2 + R$, $2(2N + 1) \le R < 3(2N + 1)$, we have $4N + 3 \stackrel{?}{=} 3N + 2 + N + 1 = 4N + 3$;
- For $n = 4N^2 + R$, $3(2N + 1) \le R < 4(2N + 1)$, we have $4N + 4 \stackrel{?}{=} 3N + 3 + N + 1 = 4N + 4$;
- For $n = 4N^2 + R$, $R = 4(2N + 1)$, we have $n = 4(N + 1)^2$. This means that n coincides with the next level, and the equality is true.

Thus, the expressions in Proposition 1 and Theorem 1 are equivalent.

Corollary 2. *An explicit formula for the nth member of A000267 is as follows:*

$$A000267(n) = \lceil 2\sqrt{n + 1} \rceil - 1,$$

where the function $\lceil \ldots \rceil$ *is the ceil function.*

Proof. It directly follows from the relationship between S_{A_0} and A000267, $S_{A_0}(n) = A000267(n-1)$.

Lemma 6. *The nth member of A248333 can be expressed thanks to sequence* S_{A_0}*:*

$$A(n) = n - S_{A_0}(n) = A000027 - A000267(n-1).$$

Proof. First, note that $A(n) = n$ if each consecutive element adds a contact. Second, correct that expression by subtracting the cases where there are no contacts, they are $S_{A_0}(n)$ by definition (see how we defined A_0 and S_{A_0}).

Proposition 2. *An explicit formula for the nth member of A248333 is as follows:*

$$A(n) = n - 3N - q + 1 - \left\lceil \frac{\sqrt{n}}{2} \right\rceil,$$

where:

1. $N = \left\lfloor \frac{\sqrt{n}}{2} \right\rfloor$;
2. $q = \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$.

The function $\lfloor \ldots \rfloor$ *is the floor function, and the function* $\lceil \ldots \rceil$ *is the ceil function.*

Proof. It directly follows from Proposition 1 and Lemma 6.

Theorem 2. *An explicit formula for the nth member of A248333 is as follows:*

$$A(n) = \lfloor (\sqrt{n} - 1)^2 \rfloor = n + 1 - \lceil 2\sqrt{n} \rceil,$$

where the function $\lfloor \ldots \rfloor$ *is the floor function, and the function* $\lceil \ldots \rceil$ *is the ceil function.*

Proof. We can derive this formula by induction in a similar way like in the pretext of Theorem 1. We can see that the expression in Theorem 2 is equivalent to the expression in Proposition 2 by the following:

$$n + 1 - \lceil 2\sqrt{n} \rceil \stackrel{?}{=} n - 3N - q + 1 - \left\lceil \frac{\sqrt{n}}{2} \right\rceil \Leftrightarrow \lceil 2\sqrt{n} \rceil \stackrel{?}{=} 3N + q + \left\lceil \frac{\sqrt{n}}{2} \right\rceil.$$

Thus, we ended with the already proved equality in the proof of Theorem 1.

We can use Theorem 2 to obtain the nth partial sum $SS_{A_0}(n)$ of sequence A248333:

$$SS_{A_0}(n) = \sum_{i=1}^{n} (i + 1 - \lceil 2\sqrt{i} \rceil) = \frac{n(n+3)}{2} - \sum_{i=1}^{n} (\lceil 2\sqrt{i} \rceil). \tag{3}$$

The problem is that in (3), for a large n, it will be a time-consuming procedure to estimate $\sum_{i=1}^{n} (\lceil 2\sqrt{i} \rceil)$. We treat this problem by introducing an explicit formula for $SS_{A_0}(n)$ without long sum of squared ns that should be rounded to the upper integer.

Theorem 3. *The nth member of sequence SS_{A_0} can be obtained by the following formula:*

$$SS_{A_0}(n) = 4nN - \frac{16}{3}N^3 - 2N^2 + \frac{N}{3} - \frac{q(q+1)}{2}(2N+1) + q(n - 4N^2 + 1),$$

where:

1. $N = Lv(n) = \left\lfloor \frac{\sqrt{n}}{2} \right\rfloor$,
2. $q = \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$.

Proof. We can split any integer n into $n = \#Lv(N) + (n - \#Lv(N))$. Then, we can observe $SS_{A_0}(n)$ in the following way:

$$SS_{A_0}(n) = SS_{A_0}(\#Lv(N)) + S_{A_0}(\#Lv(N)) \times (n - \#Lv_N) + E(n), \quad (4)$$

where $N = Lv(n)$; $S_{A_0}(\#Lv(N))$ is the value of sequence S_{A_0} for the complete Level N; the expression $S_{A_0}(\#Lv(N)) \times (n - \#Lv_N)$ shows what is accumulated after the complete Level N; and the function $E(n)$ is what left to add in order to obtain $SS_{A_0}(n)$.

$$E(n) = \begin{cases} \varepsilon(N,0) & \text{if } 4N^2 < n < 4N^2 + 1(2N+1); \\ 2\varepsilon(N,1)+ & \\ +1(2N+1)+1 & \text{if } 4N^2 + 1(2N+1) \le n < 4N^2 + 2(2N+1); \\ 3\varepsilon(N,2)+ & \\ +1(2N+1)+1+ & \\ +2(2N+1)+1 & \text{if } 4N^2 + 2(2N+1) \le n < 4N^2 + 3(2N+1); \\ 4\varepsilon(N,3)+ & \\ +1(2N+1)+1+ & \\ +2(2N+1)+1+ & \\ +3(2N+1)+1 & \text{if } 4N^2 + 3(2N+1) \le n < 4(N+1)^2. \end{cases}$$

where $\varepsilon(N,q) = n - 4N^2 - q(2N+1)$, $q = \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$, is what is left from n when we omit the elements in the complete Level N, and the elements that cover the complete sides from layer $N+1$ in the spiral. To understand better $E(n)$, see Fig. 3.

We generalize $E(n)$ in the following way:

$$E(n) = (q+1)\varepsilon(N,q) + \sum_{i=0}^{q} q(2N+1) + q = (2N+1)\frac{q(q+1)}{2} + q + (q+1)\varepsilon(N,q).$$

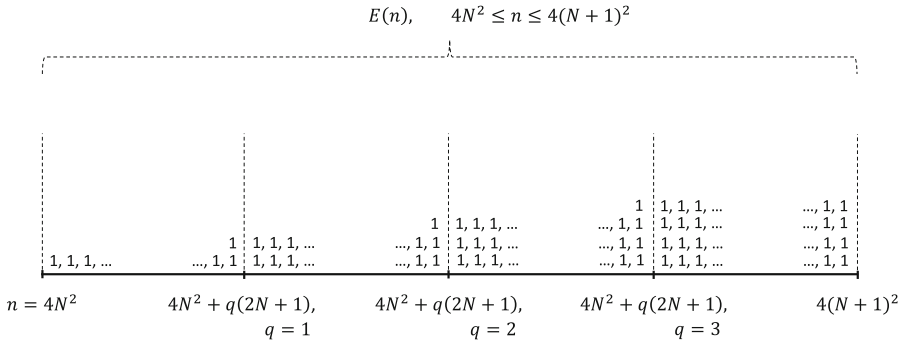

Fig. 3. The relationship between $E(n)$ and n.

Now, we obtain the expression for $E(n)$ when $n = \#Lv(N+1)$, i.e., $n = 4(N+1)^2$, it is equivalent to $E(n)$ where $q = 3$, fixed Level N, and $\varepsilon(N, q) = 2N + 1$. We have:

$$E(\#Lv(N+1)) = 6(2N+1) + 4(2N+1) + 3 = 20N + 13. \tag{5}$$

We want to obtain the sum of the elements of S_{A_0} in Layer $N+1$, we will denote by $SS_{A_0}(L_{N+1})$. We skip $SS_{A_0}(\#Lv(N))$ in Expression 4, use Expression 5, and obtain:

$$SS_{A_0}(L_{N+1}) = S_{A_0}(\#Lv(N)) \times (n - \#Lv_N) + E(\#Lv(N+1)).$$

After we apply Proposition 1 to $S_{A_0}(\#Lv(N))$, we obtain the following expression:

$$SS_{A_0}(L_{N+1}) = (4N-1) \times (n - \#Lv_N) + E(\#Lv(N+1)).$$

Note that $\#L_{N+1} = n - \#Lv_N$, $Lv(n) = N+1$:

$$SS_{A_0}(L_{N+1}) = (4N-1) \times \#L_{N+1} + E(\#Lv(N+1))$$
$$= (4N-1)4(2N+1) + 20N + 13 = (4(N+1)-5)4(2(N+1)-1) + 20(N+1) - 7.$$

Now, we substitute $N+1$ with N. We expand the brackets and obtain $SS_{A_0}(L_N)$:

$$SS_{A_0}(L_N) = 32N^2 - 36N + 13.$$

Finally, note that $SS_{A_0}(\#Lv_N) = \sum_{i=1}^{N} SS_{A_0}(L_i)$. Then:

$$SS_{A_0}(\#Lv_N) = \sum_{i=1}^{N}(32i^2 - 36i + 13) = 32\sum_{i=1}^{N} i^2 + 36\sum_{i=1}^{N} i + \sum_{i=1}^{N} 13$$
$$= 32\frac{N(N+1)(2N+1)}{6} - 36\frac{N(N+1)}{2} + 13N$$
$$= N(N+1)\left(\frac{16}{3}(2N+1) - 18\right) + 13N = N(N+1)\left(\frac{32}{3}N - \frac{38}{3}\right) + 13N$$
$$= \frac{N}{3}(32N^2 - 6N + 1).$$

We use that $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$ which are the *square pyramidal numbers*, A000330. One thing that left is to further simplify the expression for $E(n)$:

$$\begin{aligned}
E(n) &= (2N+1)\frac{q(q+1)}{2} + (q+1)\varepsilon(N,q) + q \\
&= (2N+1)\frac{q(q+1)}{2} + (q+1)(n-4N^2-q(2N+1)) + q \\
&= (2N+1)\frac{q(q+1)}{2} - (2N+1)q(q+1) + (q+1)(n-4N^2) + q \\
&= (n-4N^2)(q+1) - (2N+1)\frac{q(q+1)}{2} + q \\
&= n - 4N^2 - (2N+1)\frac{q(q+1)}{2} + q(n-4N^2+1).
\end{aligned}$$

Finally, we simplify the whole formula:

$$\begin{aligned}
SS_{A_0}(n) &= SS_{A_0}(\#Lv(N)) + S_{A_0}(\#Lv(N)) \times (n - \#Lv_N) + E(n) \\
&= \frac{32}{3}N^3 - 2N^2 + \frac{N}{3} + (4N-1)(n-4N^2) + E(n) \\
&= \frac{32}{3}N^3 - 2N^2 + \frac{N}{3} + 4nN - n - 16N^3 + 4N^2 + E(n) \\
&= 4nN - n - \frac{16}{3}N^3 + 2N^2 + \frac{N}{3} + n - 4N^2 - (2N+1)\frac{q(q+1)}{2} + q(n-4N^2+1) \\
&= 4nN - \frac{16}{3}N^3 - 2N^2 + \frac{N}{3} - (2N+1)\frac{q(q+1)}{2} + q(n-4N^2+1).
\end{aligned}$$

Lemma 7. *The relationship between the sum of the first n elements of sequence A and the sum of the first n elements of sequence S_{A_0} is as follows:*

$$S_A(n) = \frac{n(n+1)}{2} - SS_{A_0}(n) = A000217(n) - A342711(n-1).$$

Proof. From Lemma 6, we have $A(n) = n - S_{A_0}(n)$. Then $S_A(n) = \sum_{i=1}^{n} i - \sum_{i=1}^{n} S_{A_0}(i) = \frac{n(n+1)}{2} - SS_{A_0}(n)$. A well-known fact is that the nth member of the triangular numbers, A000217, is equal to $\frac{n(n+1)}{2}$.

Theorem 4. *The nth member of sequence S_A has the following explicit formula:*

$$S_A(n) = \frac{n(n+1)}{2} - 4nN + \frac{16}{3}N^3 + 2N^2 - \frac{N}{3} + \frac{q(q+1)}{2}(2N+1) - q(n-4N^2+1),$$

where:

1. $N = Lv(n) = \left\lfloor \frac{\sqrt{n}}{2} \right\rfloor$,
2. $q = \left\lfloor \frac{n-4N^2}{2N+1} \right\rfloor$.

Proof. The stated formula directly follows from Lemma 7 and Theorem 3.

6 Application

Let, again, observe the HP model of Dill. A well-known fact is that the maximum number of contacts of one protein sequence S of Hs and Ps cannot exceed $2 \times min(ODD(S), EVEN(S))$, where $ODD(S)$ is the number of Hs with odd index in S and $EVEN(S)$ is the number of even Hs. We will refer to this limit of contacts as *the standard threshold for the number of contacts*. The standard threshold is used as a stop in the heuristic algorithms, which are designed to optimize the number of contacts (an NP-hard task). In his work [2], Dill states that while maximizing the number of contacts, we obtain a compact core in the protein. Let observe a sequence of only Hs. We assume that the maximum number of contacts on a square lattice is obtained when the sequence of Hs is in a square shape. If this holds, then following the analysis in this paper, we have that the maximum number of contacts of a sequence with size n, which has only Hs, is equal to the nth member of A248333. Thus, any protein sequence of size n cannot exceed $n + 1 - \lceil 2\sqrt{n} \rceil$ contacts, see Theorem 2. We will refer to this limit of contacts as *the absolute threshold for the number of contacts*. Note that the absolute threshold depends only on the length of the protein sequence, not on the inner structure of the protein (Ps and Hs). Moreover, for a sequence with only Hs, the standard threshold will never be reached (for even n, this threshold assumes n contacts, but they are less in the spiral). Thus, the absolute threshold can be used as a more precise estimate for the maximum number of contacts in the cases where we have protein sequences that consist dominantly of Hs. Finally, we can further refine the absolute threshold if we observe the position of the Ps in the spiral and in other conformations of the sequence that form a square shape, e.g., to prune the Ps on both sides in order to obtain smaller sequence's length n.

7 Summary

In this paper, we observed A248333 from a slightly different angle - we noticed the sequence while trying to deal with a problem in the Dill's hydrophobic-polar protein folding model on a two-dimensional square lattice. This point of view, a geometrical one, allowed us to distinguish several more integer sequences and to show their geometrical interpretation. We investigated the relationships between them, which led us to explicit formulas for their nth members (including A248333). Moreover, we stated shorter explicit formulas for the nth members of A248333 and A000267 (for this sequence, we presented an alternative prove to an existing formula) and managed to prove them. Moreover, we obtained an explicit formula for the nth partial sum of A248333, sequence A342712 in the OEIS. Finally, we gave a possible application of the explicit formula for the nth member of sequence A248333.

References

1. Chen, M., Huang, W.: A branch and bound algorithm for the protein folding problem in the HP Lattice Model. Genomics Proteomics Bioinf. **3**(4), 225–230 (2005). https://doi.org/10.1016/S1672-0229(05)03031-7
2. Dill, K.: Theory for the folding and stability of globular proteins. Biochemistry **24**, 1501–1509 (1985). https://doi.org/10.1021/bi00327a032
3. Hart, W., Istrail, S.: Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. J. Comput. Biol. **4**, 1–22 (1997). https://doi.org/10.1089/cmb.1997.4.1
4. Sloane, N., et al.: The on-line encyclopedia of integer sequences (2018). https://oeis.org
5. Thachuk, C., Shmygelska, A., Hoos, H.: A replica exchange Monte Carlo algorithm for protein folding in the HP model. BMC Bioinf. **8**, 342–362 (2007). https://doi.org/10.1186/1471-2105-8-342
6. Traykov, M., Angelov, S., Yanev, N.: A new heuristic algorithm for protein folding in the HP model. J. Comput. Biol. **23**, 662–668 (2016). https://doi.org/10.1089/cmb.2016.0015
7. Yanev, N., Traykov, M., Milanov, P., Yurukov, B.: Protein folding prediction in a cubic lattice in hydrophobic-polar model. J. Comput. Biol. **24**(5), 412–421 (2017). https://doi.org/10.1089/cmb.2016.0181
8. Yanev, N., Milanov, P., Mirchev, I.: Integer programming approaches to HP folding. Serdica J. Comput. **5**, 359–366 (2011). http://hdl.handle.net/10525/1633