Research Article

Computational investigations of folded self-avoiding walks related to protein folding

Jacques M. Bahi[a], Christophe Guyeux[a], Kamel Mazouzi[b], Laurent Philippe[a],*,[1]

[a] FEMTO-ST Institute, UMR 6174 CNRS, University of Franche-Comté, Besançon, France
[b] Mésocentre de calcul (Computing Center), University of Franche-Comté, Besançon, France

ARTICLE INFO

ABSTRACT

Various subsets of self-avoiding walks naturally appear when investigating existing methods designed to predict the 3D conformation of a protein of interest. Two such subsets, namely the folded and the unfoldable self-avoiding walks, are studied computationally in this article. We show that these two sets are equal and correspond to the whole n-step self-avoiding walks for $n \leq 14$, but that they are different for numerous $n \geq 108$, which are common protein lengths. Concrete counterexamples are provided and the computational methods used to discover them are completely detailed. A tool for studying these subsets of walks related to both pivot moves and protein conformations is finally presented.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Self-avoiding walks (SAWs) have been studied over decades for the extent and difficulty of the mathematical problems they provide Madras and Sokal (1988), Bacher and Bousquet-Mélou (2011), Beaton et al. (2012), and for their various contexts of application in physics, chemistry, and biology Slade (2011), de Gennes (1972), Flory (1949). Among other things, they are used to model polymers such as DNA, RNAs, and proteins. Numerous protein structure prediction (PSP) software iterate on self-avoiding walk subsets, often not clearly defined, of various lattices. The last produced SAW S has the length of the targeted protein P. When labeling S with the amino acids of P, S is (one of) the best solution(s) according to a scoring function that associates a value to a 2D or 3D conformation (depending on physical properties of the conformation as hydrophobic neighboring residues, etc.).

In previous studies Bahi et al. (2013a, 2012, 2011), authors of this manuscript have investigated some dynamic protein folding models. They have shown that the possible sets of conformations reachable by these numerous PSP software are not equal. This raises severe questionings on what is indeed really predicted by such software. In particular, they have shown that software that iteratively stretch the conformation from one amino acid until a self-avoiding walk having the length n of the protein, can reach all the n-step SAWs $\mathfrak{G}_n$. Contrarily, the ones that iterate $\pm 90°$ pivot moves on the n-step straight line can only reach what they call the subset of folded self-avoiding walks $fSAW(n)$. It has been clearly established that, for some well-defined small n's, $fSAW(n) \neq \mathfrak{G}_n$. After having obtained this result, the authors' intention was then to investigate more deeply these new kind of self-avoiding walks and other related subsets of walks they called unfolded SAWs, and to determine consequences of these investigations regarding the protein structure prediction problem.

This article is the third of a series of three researches we publish in that field. In Bahi et al. (2013b) we provide a general presentation of folded and unfoldable SAWs, and the collection of results we have obtained on these objects using both theoretical and computational approaches. Article Bahi et al. (2013c) focuses more specifically on the mathematical study of theses subsets of self-avoiding walks, by proving in particular that the number of unfolded SAWs is infinite. This article presents our computational investigations in detail.

After recalling in the next section the basis of self-avoiding walks, of folded SAWs, and of unfoldable SAWs, we explain in Section 3 how the number of folded self-avoiding walks has been computed and how we checked the unfoldable property in practice. The various methods that have been implemented to find the shortest currently known unfoldable SAW are presented too in this section. Then, in Section 4, some heuristics that could be determinant in further studies concerning these subsets of walks are introduced. The next section contains the last contribution of this research work: a free software realized to facilitate the study of folded and unfoldable SAWs. This document ends by a conclusion section, in which all these contributions are summarized and intended future work is proposed.

* Corresponding author. Tel.: +33 3 81 66 66 54; fax: +33 3 81 66 64 50.
E-mail addresses: jacques.bahi@femto-st.fr (J.M. Bahi), christophe.guyeux@femto-st.fr (C. Guyeux), kamel.mazouzi@univ-fcomte.fr (K. Mazouzi), laurent.philippe@femto-st.fr (L. Philippe).
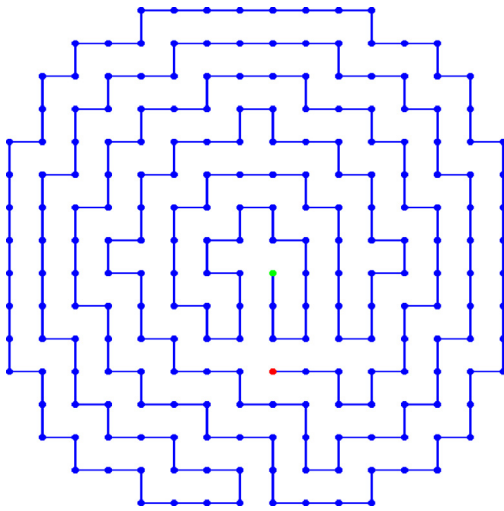[1] Authors in alphabetic order.

Fig. 1. The first SAW shown to be not connected to any other SAW by 90° rotations (Madras and Sokal, 1988).

2. Presentation of folded self-avoiding walks

We recall in this section various notions and properties of self-avoiding walks and of some of theirs folded subsets: folded SAWs obtained by iterating pivot moves on the straight line and unfoldable SAWs on which no pivot move can be applied without breaking the self-avoiding property. Authors that would investigate more deeply these walks are referred to Madras and Slade (1993), Slade (2011), Hughes (1995) for the SAWs in general, and to Bahi et al. (2013b,c) for the folded case.

2.1. Definitions and terminologies

Let $\mathbb{N}$ be the set of all natural numbers, $\mathbb{N}^* = \{1, 2, \ldots\}$ the set of all positive integers, and for $a, b \in \mathbb{N}$, $a \neq b$, the notation $[\![a,b]\!]$ stands for the set $\{a, a+1, \ldots, b-1, b\}$. $|x|$ stands for the Euclidean norm of any vector $x \in \mathbb{Z}^d$, $d \geq 1$, whereas $x_1, \ldots, x_n$ are the n coordinates of x. The nth term of a sequence s is denoted by $s(n)$. Finally, $\sharp X$ is the cardinality of a finite set X. Using this material, self-avoiding walk can be defined as follows Madras and Slade (1993), Slade (2011), Hughes (1995).

Definition 2.1 (*Self-Avoiding Walk*). Let $d \geq 1$. A n−step *self-avoiding walk* from $x \in \mathbb{Z}^d$ to $y \in \mathbb{Z}^d$ is a map $w : [\![0, n]\!] \rightarrow \mathbb{Z}^d$ with:

- $w(0) = x$ and $w(n) = y$,
- $|w(i+1) - w(i)| = 1$,
- $\forall i, j \in [\![0, n]\!]$, $i \neq j \Rightarrow w(i) \neq w(j)$ (self-avoiding property). An illustration of a Self Avoiding Walk is given in Fig. 1.

2.2. Notations

In absolute encoding Hoque et al. (2009), Backofen et al. (1999), a n-step walk $w = w(0), \ldots, w(n) \in (\mathbb{Z}^2)^{n+1}$ with $w(0) = (0, 0)$ is a sequence $s = s(0), \ldots, s(n-1)$ of elements belonging into $\mathbb{Z}/4\mathbb{Z}$, such that:

- $s(i) = 0$ if and only if $w(i+1)_1 = w(i)_1 + 1$ and $w(i+1)_2 = w(i)_2$, that is, $w(i+1)$ is at the East of $w(i)$.
- $s(i) = 1$ if and only if $w(i+1)_1 = w(i)_1$ and $w(i+1)_2 = w(i)_2 - 1$: $w(i+1)$ is at the South of $w(i)$.
- $s(i) = 2$ if and only if $w(i+1)_1 = w(i)_1 - 1$ and $w(i+1)_2 = w(i)_2$, meaning that $w(i+1)$ is at the West of $w(i)$.
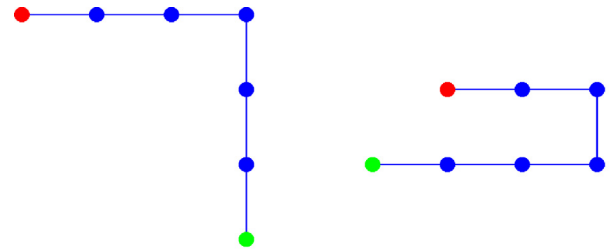
    (a) 000111   (b) $001222 = 00f^{-1}(0)f^{-1}(1)f^{-1}(1)f^{-1}(1)$

Fig. 2. Effects of the clockwise fold function applied on the four last components of an absolute encoding.

- Finally, $s(i) = 3$ if and only if $w(i+1)_1 = w(i)_1$ and $w(i+1)_2 = w(i)_2 + 1$ ($w(i+1)$ is at the North of $w(i)$).

Let us now define the following functions Bahi et al. (2013a).

Definition 2.2. The *anticlockwise fold function* is the function $f : \mathbb{Z}/4\mathbb{Z} \longrightarrow \mathbb{Z}/4\mathbb{Z}$ defined by $f(x) = x - 1 (\mathrm{mod}\, 4)$ and the clockwise fold function is $f^{-1}(x) = x + 1 (\mathrm{mod}\, 4)$.

Using the absolute encoding sequence s of a n-step SAW w that starts from the origin of the square lattice, a pivot move of $+90°$ on $w(k)$, $k < n$, simply consists to transform s into $s(0), \ldots, s(k-1)$, $f(s(k)), \ldots, f(s(n))$. Similarly, a pivot move of $-90°$ consists to apply f^{-1} to the queue of the absolute encoding sequence, like in Fig. 2.

2.3. A graph structure for SAWs folding process

We can now introduce a graph structure that fits the description of iterations for $\pm 90°$ pivot moves on a given self-avoiding walk.

Given $n \in \mathbb{N}^*$, the graph $\mathfrak{G}_n$, formerly introduced in Bahi et al. (2013a), is defined as follows:

- its vertices are the n−step self-avoiding walks, described in absolute encoding;
- there is an edge between two vertices s_i, s_j if and only if s_j can be obtained by one pivot move of $\pm 90°$ on s_i, that is, if there exists $k \in [\![0, n-1]\!]$ s.t.:
  – either $s_j(0), \ldots, s_j(k-1), f(s_j(k)), \ldots, f(s_j(n)) = s_i$
  – or $s_j(0), \ldots, s_j(k-1), f^{-1}(s_j(k)), \ldots, f^{-1}(s_j(n)) = s_i$.

Such a digraph is depicted in Fig. 3. The circled vertex is the straight line whereas strikeout vertices are walks that are not self-avoiding. Depending on the context, and for the sake of simplicity, $\mathfrak{G}_n$ will also refers to the set of SAWs in $\mathfrak{G}_n$ (*i.e.*, its vertices).

Using this graph, the folded SAWs introduced in the previous section can be redefined more rigorously.

Definition 2.3. $fSAW_n$ is the connected component of the straight line $00\ldots0$ (n times) in $\mathfrak{G}_n$, whereas $\mathcal{S}_n$ is constituted by all the vertices of $\mathfrak{G}_n$.

The Fig. 1 shows that the connected component $fSAW(223)$ of the straight line in $\mathfrak{G}_{223}$ is not equal to the whole graph: $\mathfrak{G}_{223}$ is not connected. More precisely, this graph has a connected component of size 1: it is unfoldable whereas the SAW of Fig. 4 can be folded exactly once. Indeed, to be in the same connected component is an equivalence relation $\mathcal{R}_n$ on $\mathfrak{G}_n$, $\forall n \in \mathbb{N}^*$, and two SAWs w, w' are considered equivalent (with respect to this equivalence relation) if and only if there is a way to fold w into w' such that all the intermediate walks are self-avoiding. When existing, such a way is not necessarily unique.

These remarks lead to the following definitions.