f-DISTANCE OF KNOTOIDS AND PROTEIN STRUCTURE

AGNESE BARBENSI, DIMOS GOUNDAROULIS

Abstract. Recent studies classify the topology of proteins by analysing the distribution of their projections using knotoids. The approximation of this distribution depends on the number of projection directions that are sampled. Here we investigate the relation between knotoids differing only by small perturbations of the direction of projection. Since such knotoids are connected by at most a single forbidden move, we characterise forbidden moves in terms of equivariant band attachment between strongly invertible knots and of strand passages between θ-curves. This allows for the determination of the optimal sample size needed to produce a well approximated knotoid distribution. Based on that and on topological properties of the distribution, we propose a numerical measure for the determination of deeply knotted proteins that does not require the computationally expensive method of subchain analysis.

1. Introduction

Knotoids provide a generalisation of knots that deals with the problem of classifying knottiness for open curves [30] . They are defined as equivalence classes of diagrams of open arcs, up to isotopies of S^2 and Reidemeister moves performed away from the endpoints. Each equivalence class forms a specific knotoid type. Some examples of knotoid diagrams are shown in Figure 2.1.

In the past few years, knotoids have been used to classify entanglement in proteins [12, 10, 9, 7]. Proteins are long chains of amino acids that fold into specific conformations that can sometimes form open ended knots. The fraction of knotted proteins is fairly small [7], and even if the presence of knots in proteins slows the folding process [21, 6, 26], it is known that the knotted domains of some families of proteins have been conserved through evolution [27]. While studies seem to suggest that knots provide advantages to some proteins [8, 25], the biological purpose of the presence of knots in proteins is still an open and interesting question in biology. For this reason, understanding the topological features of knotted proteins is an important step in investigating the effect of the presence of knots to structure and function of proteins. With the approach provided by knotoids, a protein is represented as an open-ended polygonal chain and it is studied by considering all different projections of it. Subsequently, these projections are analysed as knotoids. The topology of the curve is characterised by a distribution of knotoid types, also called the spectrum of the curve, while the predominate type represents the distribution. Since considering all possible projections of a curve is not computationally feasible, the usual approach is to sample from the knotoid distribution.

In this work, we study the relation between pairs of knotoids that are obtained from projections that differ from one another only by a small perturbation. Indeed, small perturbations in the choice of the direction of projections can either leave the corresponding knotoid type unchanged (*i.e.* by changing the knotoid diagram by isotopies of S^2) or have the effect of performing a sequence of, so-called, *forbidden moves* (see Figure 1.1) on the knotoid diagram. Thus, when the spectrum approximates well the knotoid distribution, a pair of different knotoids whose projections are related by a small perturbation will differ by at most a single forbidden move.

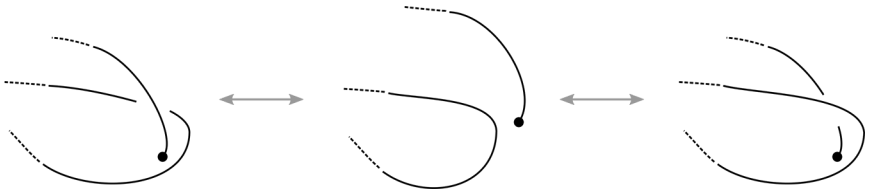

FIGURE 1.1. Forbidden moves between knotoid diagrams. Performing a forbidden move on a knotoid diagram might result in changing the knotoid type. Moreover, any knotoid diagram can be transformed into the trivial one by a finite sequence of forbidden moves.

In analogy to the case of knots and crossing changes, we define a distance on knotoids using forbidden moves.

Definition. *Given two knotoids k_1 and k_2, their* forbidden move-distance *or f-distance $d_f(k_1, k_2)$ is the minimal number of forbidden moves, across all representatives of k_1 and k_2, needed to transform k_1 into k_2.*

We characterise forbidden moves on knotoids in terms of equivariant band attachments on strongly invertible knots, using the correspondence between knotoids and strongly invertible knots proved in [2], and in terms of crossing changes on θ-curves [30]. The main theorem of this paper is the following:

Theorem 1.1. *Consider two equivalence classes of of knotoids k_1 and k_2 up to rotation and reversion. The following are equivalent:*

(i) k_1 and k_2 differ by a single forbidden move;
(ii) their corresponding θ-curves $t_{\approx}(k_1)$ and $t_{\approx}(k_2)$ differ by a strand passage of the edge e_0 over either e_+ or e_-;
(iii) their corresponding strongly invertible knots $\gamma_S(k_1)$ and $\gamma_S(k_2)$ differ by an equivariant band surgery.

This result allows us to produce lower bounds on the f-distance between knotoids. We then compute the total number of strand passages on all the knotoid diagrams with up to 6 crossings. This computation gives us upper bounds for the f-distance between knotoids with up to 6 crossings. By comparing lower and upper bounds we compute the f-distances $d_f(k_1, k_2)$ for each pair of knotoids k_1 and k_2 with minimal crossing number ≤ 4. We then create the *f-distance table* for knotoids with minimal crossing number ≤ 4.

In the second part of this work we apply the main theorem and the table of f-distances to determine an optimal size for the set of sampled projections that not only approximates well the knotoid distribution but also favours computational speed. We further expand on this result by inferring a numerical measure for detecting deeply knotted proteins. A protein is considered deeply knotted if it has at least 20 amino acids at each one of its termini [29]. The knotted core of a protein is believed to have an important biological role in the protein's function [20, 5], and recent studies show that that formation of deep knots with characteristic structural motifs provides a favourable environment for active sites in enzymes [8]. The length of the knotted core of a protein is currently determined by computationally expensive subchain analysis [9, 7]. The advantage of our measure is that it gives an estimate of whether a protein is deeply knotted or not without requiring the full subchain analysis. Finally, we demonstrate our method by analysing all proteins in the Protein Data Bank (PDB, [3]) having the 3_1 as predominate knotoid.

Notation

Throughout this paper knotoids are indicated according to the tabulation created by the second author in [11]. All maps and manifolds are assumed to be smooth, and for maps and sets we will use the notation of [2], namely:

- $\mathbb{K}(S^2)$ and $\mathbb{K}(S^2)/_{\approx}$ are the sets of knotoids and the set of knotoids up to rotation and reversion;
- Θ^s is the set of simple labelled θ-curves in S^3 and $\Theta^s/_{\approx}$ is the set of simple labelled θ-curves in S^3 up to relabelling the vertices, and up to relabelling the vertices and the edges e_- and e_+, respectively;
- $\mathcal{K}SI(S^3)$ is the set of strongly invertible knots (K, τ) in S^3.

2. On forbidden moves, crossing changes and band surgeries

Recall that there are four commuting involutive operations on knotoids in $\mathbb{K}(S^2)$. These operations are called *reversion*, *mirror reflection*, *symmetry* and *rotation*, and are described in Figure 2.1. Reversion has the effect of changing the orientation of a knotoid, and mirror reflection transforms a knotoid into a knotoid represented by the same diagrams with all the crossings changed. Symmetry reflects a knotoid diagram with respect to the line in $\mathbb{R}^2$ passing through the endpoints. The last involution, the rotation, is defined as the composition of symmetry and mirror reflection.

We will sometimes consider knotoids up to these involutions. Indeed, following the notation of [2], we will denote by $\mathbb{K}(S^2)/_{\approx}$ the set of knotoids in S^2 up to rotation and reversion.

2.1. Knotoids, θ-curves and strongly invertible knots.

Definition 2.1. *A* θ-curve *is a graph embedded in* S^3 *with* 2 *vertices,* v_0 *and* v_1*, and* 3 *edges,* e_+, e_- *and* e_0*, each of which joins* v_0 *to* v_1*, taken up to ambient isotopies preserving the labels of the vertices and the edges.*

Note that the curves $e_0 \cup e_-$, $e_- \cup e_+$ and $e_0 \cup e_+$ form knots, called the *constituent knots* of the θ-curve. A θ-curve is called *simple* if its constituent

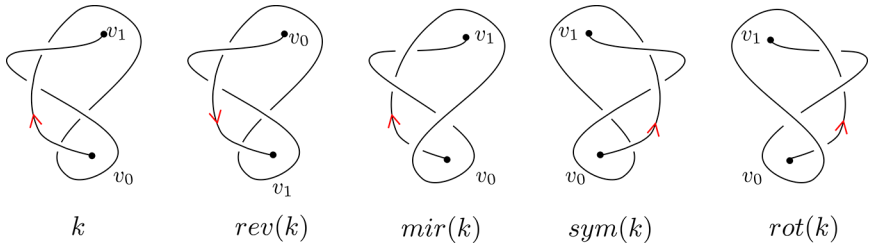

FIGURE 2.1. From left to right, a knotoid k, its reverse $rev(k)$, its mirror reflection $mir(k)$, its symmetric $sym(k)$ and its rotation $rot(k)$.

knot $e_- \cup e_+$ is the trivial knot. An example of a simple θ-curve is shown in Figure 2.2. We will only work with simple θ-curves, since these are the ones corresponding to knotoids (see [30]). Moreover, we will always choose e_+ and e_- to be the ones forming the trivial knot.

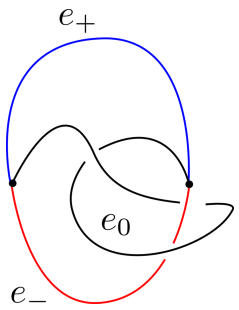

FIGURE 2.2. A simple θ-curve. The constituent knot $e_- \cup e_+$ (the blue&red circle) is the trivial knot.

We will find useful to consider θ-curves up to certain particular symmetries. Indeed, following the notation of [2], we will denote by $\Theta^s/_\approx$ the set of simple θ-curves up to relabelling the vertices v_0 and v_1, and the edges e_- and e_+.

Recall that the symmetry group of a knot K, $Sym^+(S^3, K)$, is the group of diffeomorphisms of the pair (S^3, K) preserving the orientation of S^3, where the diffeomorphisms are taken up to isotopies.

Definition 2.2. *A* strongly invertible knot *is a pair (K, τ), where K is a knot in S^3, and $\tau \in Sym^+(S^3, K)$ is called a* strong inversion*, and it is an orientation preserving element of S^3 that reverses the orientation of K and such that τ^2 is equal to the identity, taken up to conjugacy in $Sym^+(S^3, K)$.*

Note that the positive solution of the Smith Conjecture (see *e.g* [31]) implies that $fix(\tau)$ (*i.e.* the fixed point set of τ) is a trivial knot intersecting K in two points. We will denote by $\mathcal{K}SI(S^3)$ the set of strongly invertible knots. As an example, the trefoil knot 3_1 admits such an inversion in its symmetric group, see Figure 2.3.

The set $\mathbb{K}(S^2)/_\approx$ (*i.e.* the set of unoriented knotoids in S^2, taken up to rotation) is in bijection with the sets $\Theta^s/_\approx$ and $\mathcal{K}SI(S^3)$ (see [30] and [2]

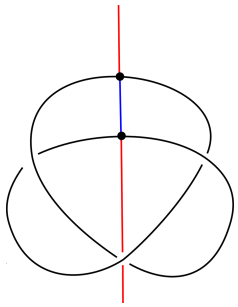

FIGURE 2.3. The trefoil knot admits a unique strong inversion τ. The fixed point set of $fix(\tau)$ is shown in the picture as a vertical line (the blue&red line).

respectively for the proofs):

$$t_{\approx} : \mathbb{K}(S^2)/_{\approx} \longrightarrow \Theta^s/_{\approx}.$$
$$\gamma_S : \mathbb{K}(S^2)/_{\approx} \longrightarrow \mathcal{K}SI(S^3).$$

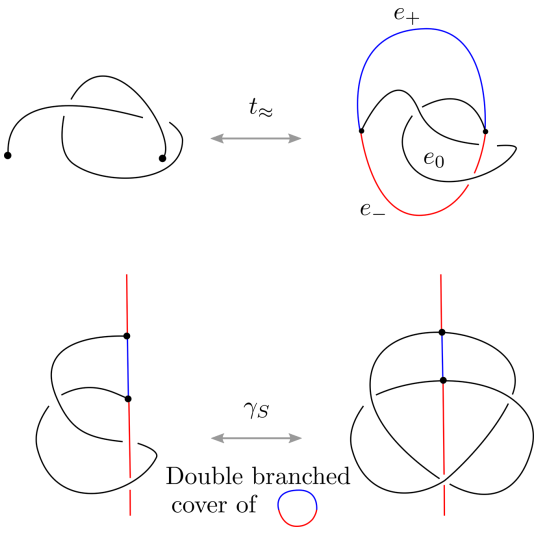

FIGURE 2.4. On the top, a knotoid and its corresponding simple θ-curve. On the bottom, the associated strongly invertible knot obtained by taking the double cover of S^3 branched along $e_- \cup e_+$.

These correspondences work schematically as follows (see also Figure 2.4 for an example).

Step 1. Given a diagram in S^2 representing a knotoid k, we construct an embedded arc in $S^2 \times I$ by pushing the overpasses of the diagram into the upper half-space, and the underpasses into the lower one. The endpoints lie in the lines $v_0 \times I$ and $v_1 \times I$. We then obtain a θ-curve by collapsing $S^2 \times \partial I$ to two points. The vertices of the θ-curve are the endpoints of k, $e_0 = k$, e_+ is the edge containing the image of $S^2 \times \{1\}$ and e_- the one containing the

image of $S^2 \times \{-1\}$.

Step 2. Given a simple, labelled θ-curve, we obtain a strongly invertible knot by taking the double cover of S^3 branched[1] along the constituent trivial knot $e_- \cup e_+$, see Figure 2.4.

The converse follows in a similar fashion:

Step 1. Given a strongly invertible knot (K, τ), we label the two halves of the circle $fix(\tau)$ as $\tilde{e_+}$ and $\tilde{e_-}$. The involution τ induces a projection map $p : S^3 \longrightarrow S^3/\tau \approx S^3$. We then obtain the θ-curve $p(fix(\tau)) \cup p(K)$, where $p(K) = e_0$ and $p(fix(\tau)) = e_- \cup e_+$.

Step 2. Given a simple, labelled θ-curve in S^3, we can think about it as being embedded in $\mathbb{R}^3$. We can isotope it in such a way that e_+ and e_- lie in the upper and lower half-spaces, respectively. We can always do that in such a way that the projection of e_0 into $\mathbb{R}^2 \times \{0\}$ is standard. Such projection gives us a knotoid.

Remark 2.3. Since two knotoids k and $k_{\mathbf{rot}}$ (respectively $rev(k)$) differing by a rotation (respectively reversion) correspond to θ-curves differing by swapping the e_- and e_+ labels (respectively v_0 and v_1), and since the double branched covers of such θ-curves produce equivalent strongly invertible knots, we have the correspondences.

2.2. **Characterisation of forbidden moves.** In what follows we will give a characterisation of forbidden moves in terms of operations on θ-curves and on strongly invertible knots.

2.2.1. *Crossing changes on θ-curves.* A forbidden move on k corresponds to performing a strand passage (*i.e.* a crossing change) on the θ-curve $t_{\approx}(k)$, see Figure 2.7. More precisely, a forbidden move induces a strand passage between the arc e_0 and either e_+ or e_-.

Remark 2.4. Call $K_k^{\pm}$ the constituent knot of $t_{\approx}(k)$ given by $e_0 \cup e_{\pm}$. From the previous construction, it follows that a forbidden move induces a crossing change on exactly one between K_k^+ and K_k^-. In particular, this specific strand passage cannot change simultaneously both these constituent knots of $t_{\approx}(k)$.

Remark 2.5. Note that given a knotoid k, the pair (K_k^+, K_k^-) can be obtained by computing the *overpassing closure* and the mirror image of the *underpassing closure* of k (for a definition see [30]).

2.2.2. *Band surgeries on strongly invertible knots.* A band surgery is an operation which changes a link into another link.

Definition 2.6. *Let L_1 be a link and $b : I \times I \longrightarrow S^3$ an embedding such that $L_1 \cap b(I \times I) = b(I \times \partial I)$. The link $L_2 = (L_1 \setminus b(I \times \partial I)) \cup b(\partial I \times I)$ is said to be obtained from L_1 by a* band surgery *along the band $B = b(I \times I)$, see Figure 2.5.*

[1]For a definition of double branched covers, and for an explanation on how to obtained the double cover of S^3 branched along a trivial knot see *e.g.* [24].

The band surgery is called *coherent* if it respects the orientation of L_1 and L_2, otherwise it is called *non-coherent*, see Figure 2.5. A non-coherent band surgery it is often called a H_2-move (see *e.g.* [1], [14]). Contrary to the case of coherent band surgeries, H_2-moves preserve the number of the components of links. This means that the result of an H_2-move performed on a knot will always be a knot.

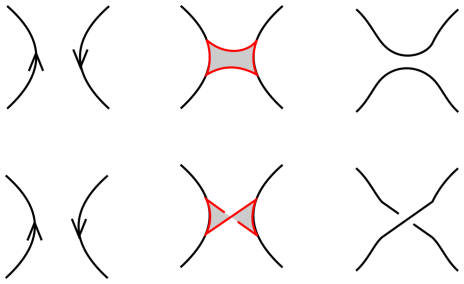

FIGURE 2.5. Local pictures for a coherent (top) and a non-coherent (bottom) band surgery.

As discussed in [2], two knotoids that differ by a forbidden move have lifts that are related by a single H_2-move. Moreover, the band attachment is *equivariant* with respect to the involutions of the two knots (see Figure 2.7).

Definition 2.7. *Consider a strongly invertible knot* (K_1, τ_1)*. We say that the strongly invertible knot* (K_2, τ_2) *is obtained from* (K_1, τ_1) *by an* equivariant band surgery *if the knots* K_1 *and* K_2 *are related by an* H_2*-move, such that:*

- $fix(\tau_1)$ *intersects the band* $b(I \times I)$ *transversally exactly once in its interior and the band is invariant under* τ_1*;*
- (K_2, τ_2) *and* (K_1', τ_1) *are equivalent as strongly invertible knots*[2]*, where* K_1' *is the knot obtained from* K_1 *by doing the band surgery.*

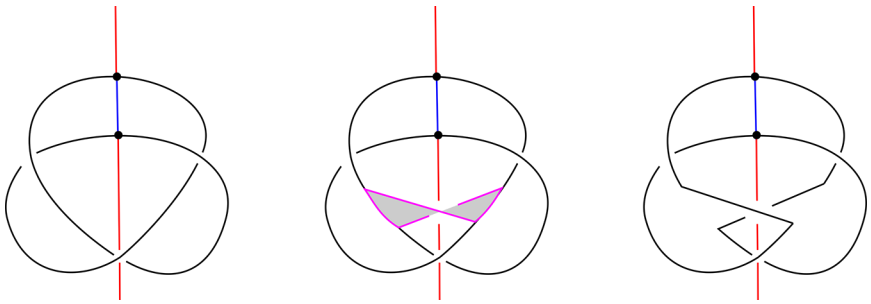

FIGURE 2.6. The strongly invertible knots $(3_1, \tau)$ and $(0_1, \tau')$ are related by an equivariant band surgery.

An example of an equivariant band surgery is shown in Figure 2.6.

[2]Recall from Definition 2.2 that two strongly invertible knots (K, τ) and (K', τ') are *equivalent* if K and K' are equivalent as knots in S^3, and τ and τ' are conjugated in $Sym^+(S^3, K)$.

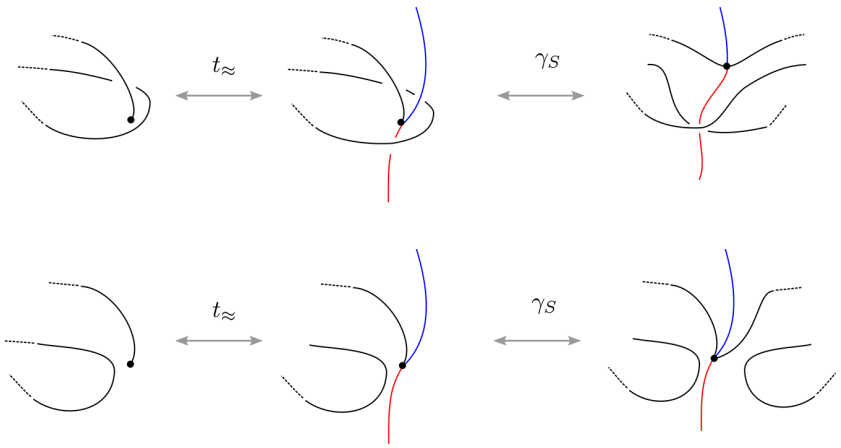

FIGURE 2.7. A forbidden move between two knotoids k_1 and k_2 induces a strand passage between the arcs e_0 and $e_\pm$ between the corresponding θ-curves, and an equivariant band attachment between the corresponding strongly invertible knots.

2.2.3. *Main result.* We are now able to prove our main result, Theorem 1.1 below.

Theorem. *Consider two equivalence classes of of knotoids k_1 and k_2 up to rotation and reversion. The following are equivalent:*

- *k_1 and k_2 differ by a single forbidden move;*
- *their corresponding θ-curves $t_{\approx}(k_1)$ and $t_{\approx}(k_2)$ differ by a strand passage of the edge e_0 over either e_+ or e_-;*
- *their corresponding strongly invertible knots $\gamma_S(k_1)$ and $\gamma_S(k_2)$ differ by an equivariant band surgery.*

Proof. Thanks to the discussion of the previous subsections, it is enough to show the following.

- Given two strongly invertible knots related by an equivariant band attachment, their corresponding θ-curves are related by a strand passage of e_0 through either e_+ or e_-;
- given two θ-curves related by a strand passage of e_0 through either e_+ or e_-, their corresponding knotoids differ by a forbidden move.

Consider then an equivariant band surgery between two strongly invertible knots (K_1, τ_1) and (K_2, τ_2). Up to ambient isotopies fixing the circle $fix(\tau_1)$ the band attachment locally looks like the one in the top part Figure 2.8, with possibly the opposite twists on the band. On the quotient $S^3/\tau_1 \approx S^3$ this results in a strand passage between the arcs e_0 and one between e_+ or e_- in the θ-curve $p(fix(\tau)) \cup p(K)$, as shown in the bottom of Figure 2.8.

Analogously, consider a simple θ-curve. Up to label preserving ambient isotopies fixing the circle $e_- \cup e_+$, any strand passage between the arc e_0 and the arc $e_\pm$ locally looks like the one shown in the top part of Figure 2.9 (up to changing the crossing between e_0 and $e_\pm$). The bottom right part of Figure 2.9 shows how this translates into a forbidden move on the

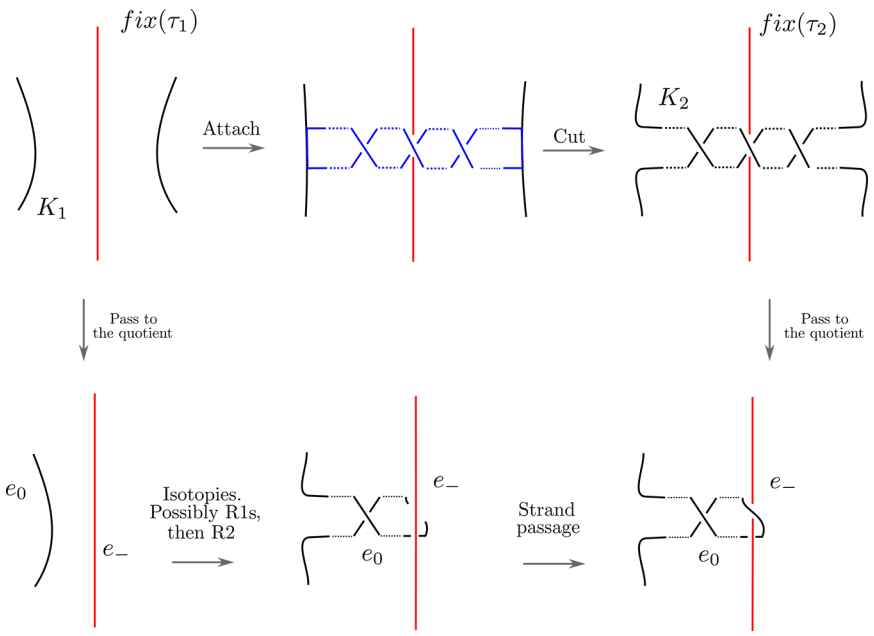

FIGURE 2.8. On the top, two strongly invertible knots (K_1, τ_1) and (K_2, τ_2) related by an equivariant band surgery. Up to ambient isotopies fixing the circle $fix(\tau_1)$ (and up to inverting the crossings), the band looks like the one in the middle of the top row. The band has an odd number of twists. On the bottom, the corresponding effect on the associated θ-curves. If the band has $2n+1$ twists, the θ-curves are related by a sequence of n Reidemeister moves of type I (R1s) followed by a a Reidemeister move of type II (R2) and by a single strand passage.

corresponding knotoid. The case where the crossing between e_0 and $e_\pm$ is the opposite one is similar.

□

2.3. **Lower bounds on the f-distance.** We use Theorem 1.1 to produce lower bounds for the forbidden move-distance between equivalence classes of knotoids up to the four involutions of Figure 2.1. With a little abuse of notation, we will still call "knotoids" these equivalence classes.

The *H_2-Gordian distance* $d_{H_2}(K, K')$ between two knots K and K' is defined as the minimal number of equivariant band attachments connecting K and K' (see [1]). As an immediate consequence of Theorem 1.1, given two knotoids k_1 and k_2, with corresponding strongly invertible knots $\gamma_S(k_1) = (K_1, \tau_1)$ and $\gamma_S(k_2) = (K_2, \tau_2)$, we have that

$$d_f(k_1, k_2) \geq d_{H_2}(K_1, K_2). \tag{2.1}$$

Analogously, given two knotoids k_1 and k_2, consider the pairs $(K^+_{k_1}, K^-_{k_1})$ and $(K^+_{k_2}, K^-_{k_2})$. We can define their *Gordian distance* $d_{\text{pair}}((K^+_{k_1}, K^-_{k_1}), (K^+_{k_2}, K^-_{k_2}))$ as the minimum between $d(K^+_{k_1}, K^+_{k_2}) + d(K^-_{k_1}, K^-_{k_2})$ and $d(K^+_{k_1}, K^-_{k_2}) + d(K^-_{k_1}, K^+_{k_2})$,

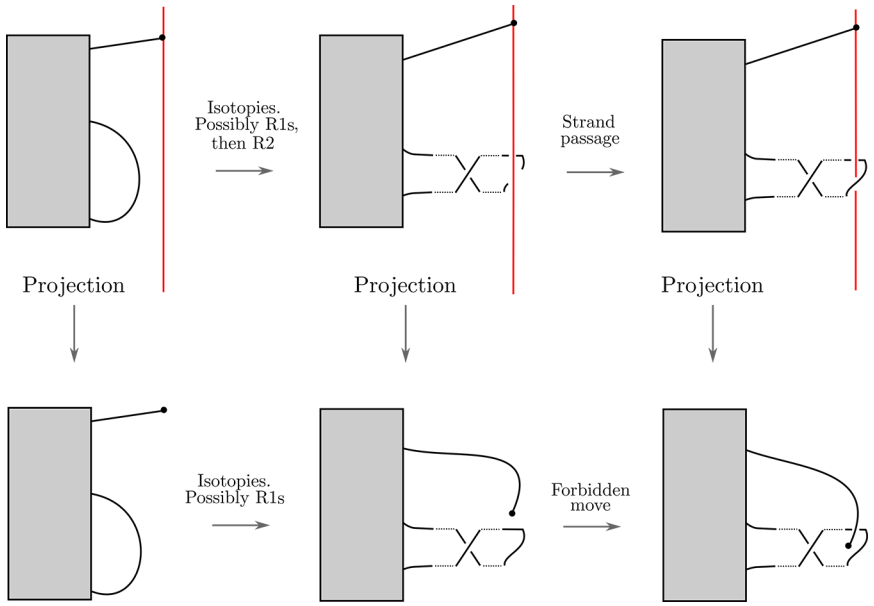

FIGURE 2.9. On the top row, two θ-curves related by a strand passage between the arc e_0 and the arc $e_\pm$. Up to label preserving ambient isotopies fixing the circle $e_- \cup e_+$ we can make the strand passage look like in the picture. The effect on the corresponding projections giving the knotoids is to perform a sequence of Reidemeister moves of type I followed by a single forbidden move.

where d is the usual Gordian distance between knots. From Remark 2.4 it follows that

$$d_f(k_1, k_2) \geq d_{\text{pair}}((K^+_{k_1}, K^-_{k_1}), (K^+_{k_2}, K^-_{k_2})). \tag{2.2}$$

Thus, as a corollary of Theorem 1.1 we have the following.

Corollary 2.8. *Let u denote the trivial knotoid. If k is a non trivial knot type knotoid then $d_f(k, u) \geq 2u(k)$, where $u(k)$ is the unknotting number (see* e.g. [24] *for a definition) of k considered as a knot in S^3.*

Proof. If k is a knot type knotoid, its corresponding constituent knots $K^\pm_k$ are both isotopic to k. Thus, $d_f(k, u) \geq 2u(K^\pm_k)$ where $u(K^\pm_k)$ is the unknotting number of the constituent knot.

□

Note that we do not expect the equality to hold in general, since in $K^\pm_k$ the unknotting crossing change might involve only the arc $e_\pm$, and thus it would not correspond to a forbidden move.

3. Computing f-distances of S^2-knotoids

As mentioned above, the main theorem provides lower bounds for f-distances between isotopy classes of knotoids. Since our aim is to build a

table of f-distances, this is information alone is not sufficient. For this reason, we computed experimentally the f-distance between all non-composite knotoid diagrams, including non-minimal crossing representations, with up to six crossings with the help of a computer program written in `python 3.7`.

In brief our strategy works as follows. First, all 2363766 knotoid diagrams with up to six crossings (both of minimal and non-minimal crossing number representation) [11] are encoded using the oriented Gauss codes for knotoids (see *e.g.* [11] and [2]). We note here that from now on, we shall be using the terms knotoid diagram and oriented Gauss code interchangably. Moreover, we will not take into consideration composite knotoids and we will deal only with prime knotoids. Each knotoid diagram is then identified using the arrow polynomial for knotoids [13] and the classification of S^2-knotoids provided by [11]. Let now $\mathcal{K}$ be the set of all knotoid diagrams with up to six crossings and let $G(V, E)$ be an undirected graph such that:

$$V(G) = \mathcal{K}$$
$$E(G) = \{(v, u) \mid (v, u) \in \mathcal{K}^2,\ v \overset{f}{\sim} u,\ v \neq u\},$$

where $v \overset{f}{\sim} u$ denotes a pair of knotoid diagrams (v, u) that are related by a single forbidden move. In other words, G is the undirected graph whose vertices are knotoid diagrams and two diagrams are related with a single forbidden move if and only if the corresponding vertices of G are connected with an edge. Our program builds G and then searches for all Dijkstra paths between all possible pairs of vertices. Finally, the set of all diagrams is partitioned into isotopy classes and the Dijkstra path of minimal distance between two isotopy classes determines their numerical f-distance, d_f^{num}. From this we can obtain upper bounds for the f-distances between isotopy classes of knotoids by computing their experimental f-distances which are defined as:

$$d_f^{\text{exp}}(v, u) = \min\left\{d_f^{\text{num}}(v, x) \mid x \in (u, mir(u), sym(u), rot(u))\right\}.$$

By comparing the upper bounds with the lower bounds discussed in Section 2.3 we are able to produce Table 1 (shown in Appendix A) containing the f-distances between equivalence classes of knotoids with minimal crossing number ≤ 4. Most of the lower bounds in Table 1 are obtained using the inequality (2.2). Gordian distances between knot types are taken from [22], while H_2-distances from [17] and [18].

Note that this could be improved by considering in the experimental approach a higher threshold for the maximum crossing number. This means that non-minimal representations of higher crossing number for the ambiguous entries in Table 1 will be considered, which may help decreasing their upper bounds. Unfortunately, our available computational power prohibited us from exploring this possibility.

Example 3.1. Computing lower bounds using inequality (2.2) it is quite straightforward. Indeed, given a knotoid k we obtain the constituent knots

$K_k^{\pm}$ as explained in Remark 2.5. Then, using values for the Gordian distance taken from [22] we compute d_{pair} for each pair of knotoids.

To illustrate how our method works in the case of inequality (2.1), we will prove as an example that $d_f(3_1, 4_7) = 2$. Given a knot K, it is well known (see *e.g.* [24]) that the double cover of S^3 branched along K is a closed 3-manifold $\Sigma(K)$ whose homeomorphism class depends solely on the knot K. Let's denote by $\delta(K)$ the dimension of the first homology of $\Sigma(K)$ with coefficients in $\mathbb{Z}_3$, $\delta(K) = \dim(H_1(\Sigma(K), \mathbb{Z}_3))$. The value of the Jones polynomial of K at $t^{1/2} = e^{i\pi/6}$ can be computed as $V(K, \omega) = \pm(i\sqrt{3})^{\delta(K)}$ [17, Proposition 5.1]. If two knots K and K' have H_2-distance 1 then the ratio $V(K, \omega)/V(K', \omega) \in \{\pm 1, \pm i\sqrt{3}^{\pm 1}\}$, which is the content of [17, Lemma 5.2].

We shall apply these to the pair $(3_1, 4_7)$. Following [2] it is straightforward to see that knotoid 3_1 lifts to a connected sum of trefoil knots $3_1\#3_1$. Additionally, it is known (see *e.g.* [17]) that $\delta(3_1\#3_1) = 2$ and, thus, $V(3_1\#3_1, \omega) = \pm 3$. On the other hand, 4_7 lifts to the torus knot 8_{19}, and in this case we have (see *e.g.* [4]) that $H_1(\Sigma(8_{19}), \mathbb{Z}) \cong \mathbb{Z}_3$. Thus, $\delta(8_{19}) = 1$. The ratio $V(3_1\#3_1, \omega)/V(8_{19}, \omega) \notin \{\pm 1, \pm i\sqrt{3}^{\pm 1}\}$ and so 3_1 and 4_7 cannot have H_2-distance equal to 1. Finally, from Table 3 we see that $d_f^{\exp}(3_1, 4_7) = 2$ and therefore we have that $d_f(3_1, 4_7) = 2$. In a similar way we can compute lower bounds for the f-distance of corresponding to the entries in red of table and 2, since in these cases the knotoids lift to knots K with $\delta(K) = 1$.

Remark 3.2. By employing statistical procedures we can re-arrange the table of numerical f-distances so that the isotopy classes of knotoids are ordered with respect to their proximity. More precisely, by considering Tables 3-5 as a 40×40 matrix M and then shifting its empirical mean to zero, we can apply Principal Component Analysis (PCA)[23, 15] to move the data to a new orthogonal coordinate system where the greatest variance of the data appears by projecting along the first coordinate. This corresponds to the eigenvector that is related to the highest eigenvalue of the correlation matrix $\frac{1}{n-1}M^T M$, where n is the number of rows of M. The post-PCA matrix provides a more comprehensive overview of the knotoids space and allows for an easier exploration of potential relations between knotoids. A graphical representation of the results of applying PCA on the set of knotoids is shown below in Figure 3.1. The analysis and the graphic representations below were done using the statistical package `R`.

4. Application to the study of proteins' topology

Proteins are long linear biomolecules that often fold into conformations with non-trivial topology. By tracing the coordinates of their Cα atoms, one can model them as open polygonal curves in 3-space. Until recently, in order to analyse proteins in terms of their knottedness one had first to artificially close the curve since under classical knot theory all open curves are topologically trivial. In [10] the second author and collaborators proposed an alternative approach using the concept of knotoids.

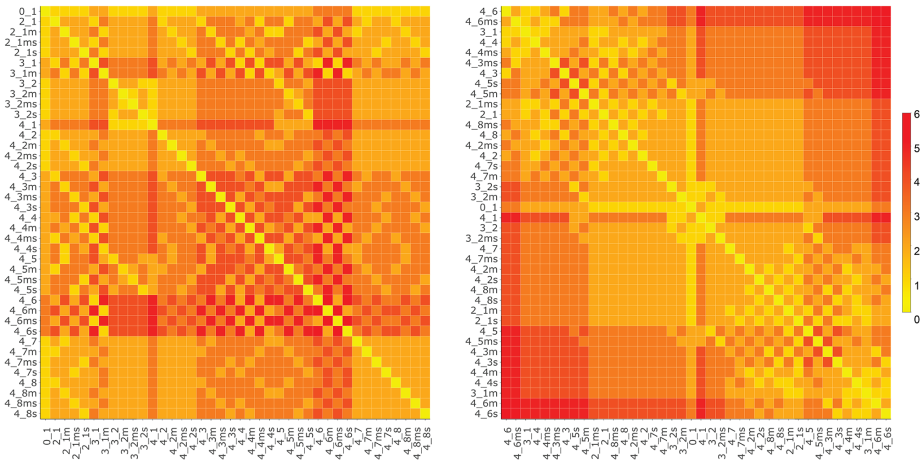

FIGURE 3.1. A graphical representation of the set of all knotoids with up to 4 crossings (on the left) before and after (on the right) using PCA. On the right, the trivial knotoid is placed in the middle of the figure since while each knotoid together with its rotation (here denoted as `ms`) are always on the opposite side of 0_1 and in equal distance from it than its mirror reflection and its symmetric involutions. Note also that the same holds for 4_1 due to its amphichirality and since two knotoids cannot occupy the same spot, it is place immediately to the left of 0_1. The legend on the far right shows the correspondence between distance and colour.

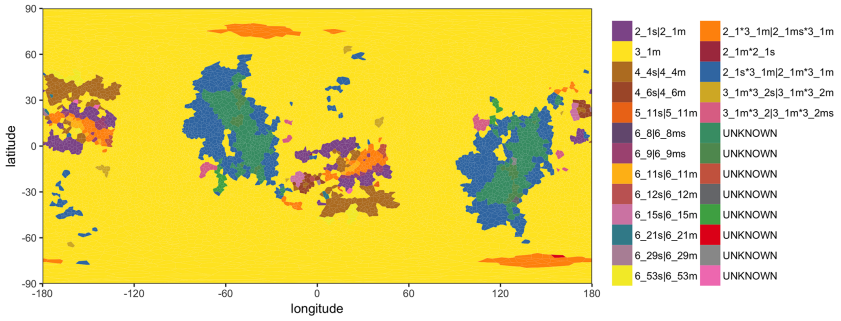

FIGURE 4.1. The projection map for the protein N-acetyl-L-ornithine transcarbamylase complexed with N-acetyl-L-ornirthine (PDB code: 3kzn). For this map we used 10000 projections. We can see that the predominate knotoid type is 3_1^m since it corresponds to the region with the biggest area. The colour scale on the right shows the colour-knotoid type correspondence. Note that the | symbol in the knotoid name stands for "or", meaning that knotoid names separated with | share the same arrow polynomial. The label "UNKNOWN" corresponds to knotoids with minimal crossing number > 6. Finally, knotoid composition is denoted with an asterisk $*$.

The general idea of this approach is to characterise the *global topology* of each protein chain by assigning a knotoid type to it. The term global topology refers to the topology of the whole protein chain. The modelled protein is considered inside a large enough sphere (one can also consider its convex hull), centered at the center of mass of the chain. Choosing a point on the sourranding sphere defines a planar projection, resulting in a knotoid diagram [10]. Note that different planes of projection may yield different knotoid diagrams, hence determining the knotoid type of a protein using a single projection is far from being accurate. In principle, the knotoids approach considers the knottedness of any open-ended curve embedded in 3-space as a distribution (also called *spectrum*) of knotoid types. The knotoid with the highest probability in the distribution of knotoid types over all projections, characterises the protein and is called *predominate*. In order to obtain an unbiased overview of a protein's topology, all possible projections have to be considered. However, since this is not computationally feasible, the distribution is approximated by sampling from the space of all possible projections. To avoid a change of knotoid type under ambient isotopy, two infinite lines are introduced each time a projection plane is chosen. Each line passes through one of the endpoints and they are perpendicular to the projection plane [13, 10]. Additionally, an algorithm similar to KMT [19, 28] that simplifies the curve but preserves its underlying topology is also applied [13, 10] in order to make computations of knotoid types more efficient. The knotoid types are determined using invariants. For this work we have used the arrow polynomial for knotoids in S^2 [13].

The above data is often summarised in a plot called *the projection map* [10]. The projection map is in fact the Voronoi diagram $\mathcal{VD}$ of the corresponding Delaunay triangulation of S^2 with respect to the set of points sampled from S^2. Furthermore, each cell of the projection map is colour-coded according to the knotoid type it produces. By construction, there is a bijection between the number of different colours in a projection map and the number of different knotoids in the spectrum of the analysed curve (see Fig. 4.1).

Both the spectrum and the projection map depend heavily on the sample size of projections; if too few points are sampled, then the overall topology of the analysed curve will not be well approximated. In fact, the optimal size for the set of sampled projections remains an open question.

In several cases the protein chain doubles back right after forming a knot. This results in unknotting the knot previously formed. This type of proteins are called slipknotted. In order to detect these local knots one has to study the *local topology* of any given protein by analysing all possible subchains. During this analysis one can also determine the *knotted core* of a protein, that is the shortest subchain forming a knot. The subchains located before and after the knotted core are called *N-tail* and the *C-tail*, where N and C are the two termini of the protein (see Fig. 4.2). The vast majority of proteins are enzymes where there is an overlap between knotted cores and the respective enzymatic active sites. Indeed, these sites are either located inside or close to the knotted core of the chain. In this context, it was shown that knotted cores of proteins play a vital role in some aspects of a

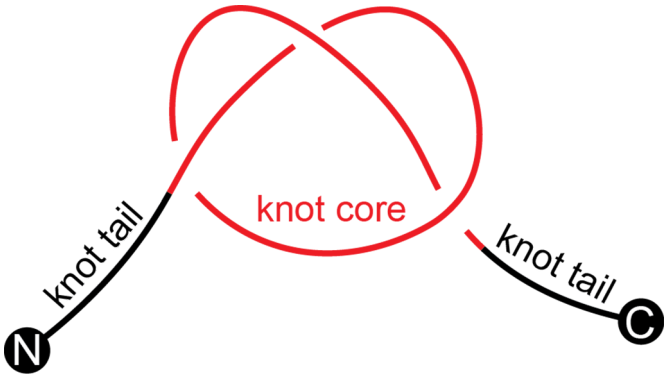

FIGURE 4.2. The core (in red) and the two tails (in black) of an open knot representing a protein with a trefoil knot. The two beads on either side represent the two termini, N and C respectively, of the protein chain.

protein's structure and function [20, 5]. Moreover, in [8], it was observed that formation of deep knots with characteristic structural motifs provides a favourable environment for active sites in enzymes. However, it is important to mention that the knots are not necessary for the formation of regions with increased intra-chain contacts. An open polygonal knot is considered *deep*, if several vertices from either side of the knotted core have to be removed before making the knot trivial.

The subchain analysys can be computationally heavy, depending on the total length of the protein. Therefore, it would be useful if one could determine numerically the tightness of a knot directly from the global topology analysis.

In this section we will use Theorem 1.1 to provide numerical approximations for the two questions that were discussed in this section, namely:

a. How many projections are required in order to have an accurate overview of a protein's topology?
b. Can we define a numerical measure that can indicate whether a protein is deeply knotted or not?

4.1. **Approximating the sample size of projections.** Consider a generic projection of a protein chain on some plane and let k be the corresponding knotoid. If we continuously perturb the projection direction until the knotoid type changes to k', we will obtain a pair of knotoids with $d_f(k, k') > 0$. We will use this idea to make measurements on projection maps that are obtained from sample sets of increasing size, in order to approximate numerically an optimal sample size of projections s for a given protein.

As mentioned earlier, the spectrum of a protein chain depends on the number of projections. Indeed, using a smaller sample will lead to a triangulation of S^2 that produces a Voronoi diagram with wider cells. Therefore, there is a higher chance for cells corresponding to knotoids with $d_f > 1$ to be adjacent. Since the predominate knotoid corresponds to the largest region of the projection map, it suffices to focus on the discrepancies between the region of the predominate and its immediate neighbours. For this, we define

the *interface error*, $er(s)$, associated to the sample of size s. The interface error is the ratio of the total number of pairs of adjacent regions adjacent to the region of the predominate knotoid k_0 that correspond to knotoids k_i, for which Theorem 1.1 gives $d_f(k_0, k_i) > 1$ over the number of all adjacent regions to k_0. We expect that there will be a high correlation between $er(s)$ and $Spec(s)$, the spectrum of knotoids in the projection map associated to the sample of size s. By gradually increasing the number of projections, the triangulation will become progressively finer and so the possibility of having pairs of adjacent cells with $d_f > 1$ will effectively decrease, hence the correlation between $er(s)$ and $Spec(s)$ will become weaker.

For our experiment, we concentrated on proteins with predominate knotoid type 3_1. As of August 2019, there are 457 such proteins in total deposited in the Protein Data Bank [3], according to [16, 7]. All proteins of interest are analysed using 50, 100, 500, 1,000, 5,000 and, finally, 10,000 projections. Each time $er(s)$ for the respective Voronoi diagram is computed. More precisely, for each Voronoi diagram we build a graph whose vertices correspond to the regions of the map and the edges correspond to common boundaries between regions. We compute $er(s)$ by counting the number of graph edges between 3_1 and knotoids that give $d_f > 1$ and taking the ratio over the total number of edges that have 3_1 as one of its endpoints, loops excluded. In Table 2 we present the f-distances of 3_1 from all knotoids with 5 and with 6 crossings. The number of unique knotoid types in the Voronoi diagram corresponds to the size of $Spec(s)$. Recall that in this analysis we don't consider composite knotoids. Finally, we compute the Spearman correlation coefficient, r, between $er(s)$ and $Spec(s)$, for each of the six cases for the sample size of projections.

As shown in Fig. 4.3, we observe positive correlation between $er(s)$ and $Spec(s)$ that is marginally strong since $r = 0.7455$, keeping in mind that for strong correlation one should have $r \in (0.75, 1)$. As the number of projections is increased, r become consistently smaller with its smallest value at 10000 projection. Our analysis suggests that at 10000 projections we will have the most accurate overview of the topology of a protein. However, if we consider computational speed as well, 5000 projections providea an approzimation reliable enough.

In conclusion, the numerical difference between the Spearman correlation coefficients for the 5,000 and 10,000 projections is not that significant and, moreover, both coefficients lie in the interval $(0.25, 0.5)$ indicating a weak positive correlation. For this reason, we suggest that analysing a protein with 5,000 projections provides a good compromise between computational speed and accuracy.

4.2. **A numerical measure for deeply knotted proteins.** In this section we discuss a numerical measure, the *relative area* A_{rel} of a protein chain, that can be used to estimate wether a protein is deeply knotted, without passing through the computational expensive subchain analysis [7].

Definition 4.1. *Let k be a knotoid in S^2. An interface knotoid, denoted by k_{int}, is a knotoid such that:*

$$d_f(k, k_{\text{int}}) = 1 \quad \textit{and} \quad d_f(k_{\text{int}}, 0_1) \leq 1$$

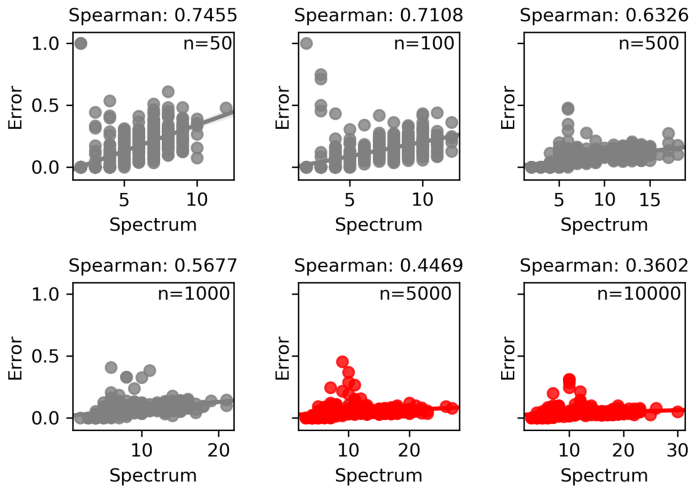

FIGURE 4.3. The diminishing effect of increasing the number of projections on the correlation between the interface error $er(s)$ and the spectrum $Spec(s)$ of a protein chain. The slopes of the lines gradually decrease. In the last two panels (in red) the lines are almost parallel to the x-axis, indicating there is almost no correlation between the two measures.

A deep protein knot is also usually tightly knotted as the length of its knotted core is relatively small, compared to the knot's overall length. The converse is not always true since we can easily find examples of tight knots very close to one the termini of the chain.

The idea behind our strategy is assuming that the deeper a knot is, the smaller the total area of the interface knotoids will be in the $\mathcal{VD}$. This is because the two tails of the knot are less probable to interact with the rest of the chain in a way that will produce a forbidden move. Since the knot is deep, it will also be relatively tight, hence the regions of the corresponding diagram will be smaller. Thus, the probability of having large areas in $\mathcal{VD}$ corresponding to interface knotoids will be smaller. To quantify this assumption we define notion of the *relative area* A_{rel} as the ratio of the sum of areas of interface knotoids in the $\mathcal{VD}$ over the area of the predominate, namely:

$$\mathcal{A}_{\text{rel}} = \frac{1}{\mathcal{A}_p} \sum_{k \in \mathcal{K}_{\text{int}}} \mathcal{A}_k, \tag{4.1}$$

where $\mathcal{K}_{\text{int}}$ is the set of all interface knotoids of the predominate knotoid, $\mathcal{A}_p$ is the area in the $\mathcal{VD}$ of the predominate knotoid and $\mathcal{A}_k$ is the area in the $\mathcal{VD}$ of the knotoid k.

As mentioned earlier, to determine the depth of a protein knot one has to analyse all of its subchains. Computationally this is achieved by progressively trimming the chain from each side and evaluating its knotoid type until the knotted core is obtained. We can define an abstract measure of

depth, denoted by $D(k)$, as follows:

$$D(k) = \frac{\ell_N(k)\,\ell_C(k)}{\ell_T(k)^2}$$

where k is a knotoid, $\ell_T(k)$ is the total length of the chain of k, $\ell_N(k)$ is the length of the N-tail of k and $\ell_C(k)$ is the length of the C-tail of k.

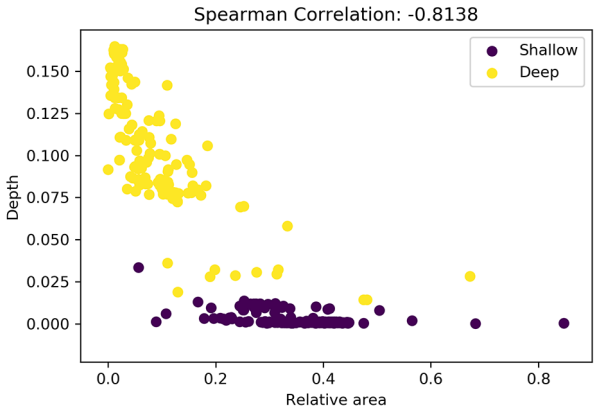

FIGURE 4.4. Scatterplot of values of $\mathcal{A}_{\text{rel}}$ against $D(k)$. The color map on the right indicates the different values of $D(k)$. The higher the value, the deeper a knot is. Two distinct clusters of points, in terms of $D(k)$, are visible in the graph indicating a well defined separation between deeply and shallowly knotted proteins.

In the spirit of the previous section, we will test our measures on all proteins that form a 3_1 knot. From Theorem 1.1 we have that 2_1 is the only knotoid (among knotoids with 6 or fewer crossings) having distance 1 from both 3_1 and 0_1. In fact, it is straightforward to check that all the knotoids k with less than 6 crossings and $d_f(k, 3_1) = 1$ have distance > 1 from 0_1 using inequality (2.2). In this case Equation (4.1) becomes:

$$\mathcal{A}_{\text{rel}} = \frac{\mathcal{A}_{2_1}}{\mathcal{A}_{3_1}}$$

Next, we compute $\mathcal{A}_{\text{rel}}$ and $D(k)$ for all of the 457 studied proteins. In more detail, we first compute the projection map for each protein using the optimal value of 5000 projections that was determined in the previous section. From each projection map we then compute the corresponding $\mathcal{A}_{\text{rel}}$ and $D(k)$. The Spearman correlation between these two measurements is -0.8138, indicating a strong decreasing monotonous relation between them, which can be seen in Figure 4.4. We also observe a partition of the set of all proteins with a 3_1 into two separate clusters, one in upper left corner of the scatterplot and one in the lower-right. The upper cluster includes the deep 3_1-proteins while the lower one is the cluster of shallow 3_1-proteins. From the histograms of the two different groups we can say that, a value $\mathcal{A}_{\text{rel}} \leq 0.2$ is likely to correspond to a protein having a deeply knotted trefoil (see Figure 4.5). It would be interesting to further explore $\mathcal{A}_{\text{rel}}$ for other

proteins knots and determine the appropriate cutoff values that suggest deep knots of specific types.

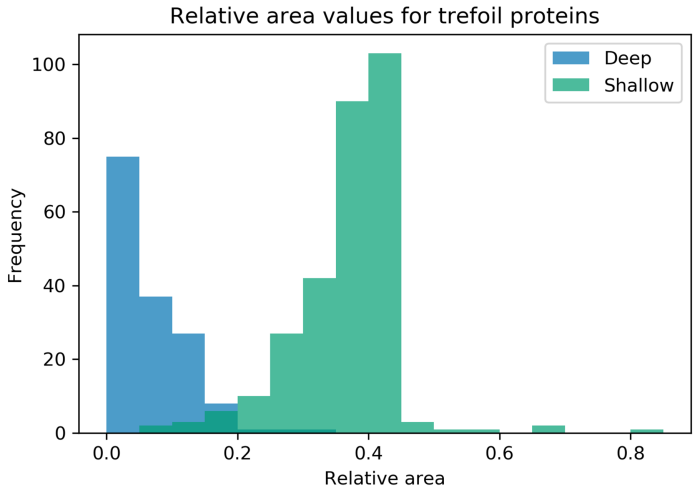

FIGURE 4.5. The two histograms for the deeply knotted protein group and the shallowly knotted group.

There are seven proteins that are shallow and have $\mathcal{A}_{\text{rel}} < 0.2$, namely the proteins with PDB codes 1eku, 1z97, 2obv, 3iml, 4h6v, 4twl and 5yud. From a biological point of view, there is no relation between them (no common function, organism of residence, length or PFAM identifier) that would suggest any particular behaviour and so it is probable that they are just numerical outliers of our method. On the other hand, there are 10 configurations that are deep but have $\mathcal{A}_{\text{rel}} > 0.2$. These are the PDB entries with codes: 2egv, 2egw, 3jyw, 3sig, 3sij, 3v5u, 4k1c, 4kjr, 4kpp and 5jdf. From these, 2egv and 2egw represent homologous protein and the same holds for 3sig and 3sij. Therefore there are also eight non-homologous deep proteins with $\mathcal{A}_{\text{rel}} > 0.2$. Similarly to the previous case, there is no specific biological relation between them and they are probably another outlier group of our method.

Concluding, our computations show how, remarkably, we can infer subtle information about the global geometry of the protein and about its knot depth, directly from a refined topological analysis based on the properties of knotoids and on Theorem 1.1.

ACKNOWLEDGEMENTS

The authors would like to thank Pawel Dabrowski-Tumanski, Dorothy Buck, Heather Harrington, Marc Lackenby and Andrzej Stasiak for several fruitful conversations and guidance throughout the development of this work. A.B. is supported by the RS-EPSRC grant "Algebraic and topological approaches for genomic data in molecular biology" EP/R005125/1. We also thank the COST Action European Topology Interdisciplinary Action (EUTOPIA) CA17139 for supporting collaborative meeting of the authors.

Data availability

The data and the code implemented for this work are available on the GitHub repository: `https://github.com/dgound/f-distance`

References

[1] Tetsuya Abe and Taizo Kanenobu, *Unoriented band surgery on knots and links*, arXiv preprint arXiv:1112.2449 (2011).

[2] Agnese Barbensi, Dorothy Buck, Heather A Harrington, and Marc Lackenby, *Double branched covers of knotoids*, arXiv preprint arXiv:1811.09121 (2018).

[3] Helen M Berman, Philip E Bourne, John Westbrook, and Christine Zardecki, *The protein data bank*, (2003), 394–410.

[4] M Borodzik and S Friedl, *Knotorious world wide web page*, 2011.

[5] Thomas Christian, Reiko Sakaguchi, Agata P Perlinska, Georges Lahoud, Takuhiro Ito, Erika A Taylor, Shigeyuki Yokoyama, Joanna I Sulkowska, and Ya-Ming Hou, *Methyl transfer by substrate signaling from a knotted protein fold*, Nature structural & molecular biology **23** (2016), no. 10, 941.

[6] P Dabrowski-Tumanski, AI Jarmolinska, and JI Sulkowska, *Prediction of the optimal set of contacts to fold the smallest knotted protein*, Journal of Physics: Condensed Matter **27** (2015), no. 35, 354109.

[7] Pawel Dabrowski-Tumanski, Pawel Rubach, Dimos Goundaroulis, Julien Dorier, Piotr Sułkowski, Kenneth C Millett, Eric J Rawdon, Andrzej Stasiak, and Joanna I Sulkowska, *Knotprot 2.0: a database of proteins with knots and other entangled structures*, Nucleic acids research **47** (2018), no. D1, D367–D375.

[8] Pawel Dabrowski-Tumanski, Andrzej Stasiak, and Joanna I Sulkowska, *In search of functional advantages of knots in proteins*, PloS one **11** (2016), no. 11, e0165986.

[9] Julien Dorier, Dimos Goundaroulis, Fabrizio Benedetti, and Andrzej Stasiak, *Knotoid: a tool to study the entanglement of open protein chains using the concept of knotoids*, Bioinformatics **34** (2018), no. 19, 3402–3404.

[10] Dimos Goundaroulis, Julien Dorier, Fabrizio Benedetti, and Andrzej Stasiak, *Studies of global and local entanglements of individual protein chains using the concept of knotoids*, Scientific reports **7** (2017), no. 1, 6309.

[11] Dimos Goundaroulis, Julien Dorier, and Andrzej Stasiak, *A systematic classification of knotoids on the plane and on the sphere*, arXiv preprint arXiv:1902.07277 (2019).

[12] Dimos Goundaroulis, Neslihan Gügümcü, Sofia Lambropoulou, Julien Dorier, Andrzej Stasiak, and Louis Kauffman, *Topological models for open-knotted protein chains using the concepts of knotoids and bonded knotoids*, Polymers **9** (2017), no. 9, 444.

[13] Neslihan Gügümcü and Louis H Kauffman, *New invariants of knotoids*, European Journal of Combinatorics **65** (2017), 186–229.

[14] Jim Hoste, Yasutaka Nakanishi, and Kouki Taniyama, *Unknotting operations involving trivial tangles*, Osaka Journal of Mathematics **27** (1990), no. 3, 555–566.

[15] Harold Hotelling, *Analysis of a complex of statistical variables into principal components.*, Journal of educational psychology **24** (1933), no. 6, 417.

[16] Michal Jamroz, Wanda Niemyska, Eric J Rawdon, Andrzej Stasiak, Kenneth C Millett, Piotr Sułkowski, and Joanna I Sulkowska, *Knotprot: a database of proteins with knots and slipknots*, Nucleic acids research **43** (2014), no. D1, D306–D314.

[17] Taizo Kanenobu, *Band surgery on knots and links*, Journal of Knot Theory and Its Ramifications **19** (2010), no. 12, 1535–1547.

[18] ———, *Band surgery on knots and links, iii*, Journal of Knot Theory and Its Ramifications **25** (2016), no. 10, 1650056.

[19] Kleanthes Koniaris and Murugappan Muthukumar, *Self-entanglement in ring polymers*, J. Chem. Phys. **95** (1991), no. 4, 2873–2881.

[20] Anna L Mallam and Sophie E Jackson, *The dimerization of an α/β-knotted protein is essential for structure and function*, Structure **15** (2007), no. 1, 111–122.

[21] ———, *Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins*, Nature chemical biology **8** (2012), no. 2, 147.

[22] Hyeyoung Moon, *Calculating knot distances and solving tangle equations involving montesinos links*, (2010).
[23] Karl Pearson, *On lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2** (1901), no. 11, 559–572.
[24] Dale Rolfsen, *Knots and links*, vol. 346, American Mathematical Soc., 2003.
[25] Álvaro San Martín, Piere Rodriguez-Aliaga, José Alejandro Molina, Andreas Martin, Carlos Bustamante, and Mauricio Baez, *Knots can impair protein degradation by atp-dependent proteases*, Proceedings of the National Academy of Sciences **114** (2017), no. 37, 9864–9869.
[26] Joanna I Sułkowska, Jeffrey K Noel, and Jose N Onuchic, *Energy landscape of knotted protein folding*, Proceedings of the National Academy of Sciences **109** (2012), no. 44, 17783–17788.
[27] Joanna I Sułkowska, Eric J Rawdon, Kenneth C Millett, Jose N Onuchic, and Andrzej Stasiak, *Conservation of complex knotting and slipknotting patterns in proteins*, Proceedings of the National Academy of Sciences **109** (2012), no. 26, E1715–E1723.
[28] William R. Taylor, *A deeply knotted protein and how it might fold*, Nature **406** (2000), 916–919.
[29] William R Taylor and Kuang Lin, *Protein knots: a tangled problem*, Nature **421** (2003), no. 6918, 25.
[30] Vladimir Turaev et al., *Knotoids*, Osaka Journal of Mathematics **49** (2012), no. 1, 195–223.
[31] Friedhelm Waldhausen, *Über involutionen der 3-sphäre*, Topology **8** (1969), no. 1.

Appendix A. The f-distances tables.

	0_1	2_1	3_1	3_2	4_1	4_2	4_3	4_4	4_5	4_6	4_7	4_8
0_1	0	1	2	1	2	1	2	2	2	3	1	1
2_1	1	0	1	2	3	1	1	1	1	2	1	1
3_1	2	1	0	3	4	2	2	1	2	1	2	2
3_2	1	2	3	0	1	2	2-3	3	1	3-4	2	2
4_1	2	3	4	1	0	3	3-4	4	2	4-5	3	3
4_2	1	1	2	2	3	0	1-2	1	2	2	1-2	1
4_3	2	1	2	2-3	3-4	1-2	0	2	2	1	1-2	1
4_4	2	1	1	3	4	1	2	0	2	1	1-2	1
4_5	2	1	2	1	2	2	2	2	0	2-3	1-2	2
4_6	3	2	1	3-4	4-5	2	1	1	2-3	0	2-3	2
4_7	1	1	2	2	3	1-2	1-2	1-2	1-2	2-3	0	1-2
4_8	1	1	2	2	3	1	1	1	2	2	1-2	0

Table 1. The f-distance table for equivalence classes of knotoids with minimal crossing number ≤ 4. In a few cases (*e.g* for the pair $(4_1, 4_6)$) lower and upper bounds do not coincide. In these cases we write upper and lower bounds separated by a dash, indicating the interval of possible values of the f-distances. Entries in the table are colour coded accordingly to how lower bounds were computed. Lower bounds for the entries in blue are computed using the inequality 2.2, while the ones in red using the inequality 2.1. We are not able to produce lower bounds for entries in orange.

	3_1		3_1		3_1		3_1		3_1
5_1	2	6_6	3	6_{35}	2	6_{64}	3	6_{93}	3
5_2	2	6_7	3	6_{36}	2	6_{65}	2-3	6_{94}	2
5_3	2	6_8	1-2	6_{37}	2-4	6_{66}	2	6_{95}	2
5_4	2	6_9	1-2	6_{38}	2-3	6_{67}	2	6_{96}	2
5_5	3	6_{10}	2	6_{39}	1-2	6_{68}	2	6_{97}	2-3
5_6	2	6_{11}	1-2	6_{40}	2	6_{69}	2	6_{98}	2
5_7	3	6_{12}	2	6_{41}	2	6_{70}	3	6_{99}	2-4
5_8	2	6_{13}	2	6_{42}	3	6_{71}	4	6_{100}	2-3
5_9	2	6_{14}	3	6_{43}	3	6_{72}	2	6_{101}	3
5_{10}	3	6_{15}	1	6_{44}	2	6_{73}	2	6_{102}	3
5_{11}	1	6_{16}	1	6_{45}	1-2	6_{74}	3-4	6_{103}	3
5_{12}	2	6_{17}	2	6_{46}	3-4	6_{75}	3	6_{104}	3
5_{13}	4	6_{18}	2	6_{47}	2	6_{76}	3	6_{105}	2
5_{14}	3	6_{19}	2-3	6_{48}	2-3	6_{77}	2	6_{106}	1
5_{15}	4	6_{20}	3-4	6_{49}	3	6_{78}	3	6_{107}	2
5_{16}	2	6_{21}	2	6_{50}	2-3	6_{79}	3-4	6_{108}	2-3
5_{17}	2	6_{22}	2-3	6_{51}	2	6_{80}	2-4	6_{109}	2-3
5_{18}	2	6_{23}	2-3	6_{52}	3	6_{81}	2	6_{110}	3
5_{19}	2	6_{24}	2-3	6_{53}	1	6_{82}	2-3	6_{111}	2
5_{20}	1	6_{25}	3	6_{54}	4	6_{83}	1-3	6_{112}	1-2
5_{21}	3	6_{26}	2	6_{55}	2	6_{84}	2	6_{113}	3
5_{22}	2	6_{27}	2-3	6_{56}	2	6_{85}	2-3	6_{114}	2
5_{23}	3	6_{28}	2	6_{57}	4	6_{86}	2-4	6_{115}	1-2
5_{24}	1-2	6_{29}	2	6_{58}	2	6_{87}	2	6_{116}	2
6_1	4	6_{30}	1-2	6_{59}	1	6_{88}	1	6_{117}	3
6_2	2	6_{31}	3	6_{60}	3	6_{89}	2	6_{118}	1-3
6_3	2	6_{32}	3	6_{61}	3	6_{90}	2	6_{119}	2-3
6_4	3	6_{33}	3	6_{62}	1-2	6_{91}	2-3	6_{120}	2-3
6_5	2	6_{34}	2-3	6_{63}	2-3	6_{92}	2-3	6_{121}	2-4

TABLE 2. The f-distances between equivalence classes of knotoids with minimal crossing number ≤ 6 and the 3_1 knotoid. In a few cases (*e.g* for the 6_{99} knotoid) lower and upper bounds do not coincide. In these cases we write upper and lower bounds separated by a dash, indicating the interval of possible values of the f-distances. Entries in the table are colour coded accordingly to how lower bounds were computed. Lower bounds for the entries in blue are computed using the inequality 2.2, while the ones in red using the inequality 2.1. We are not able to produce lower bounds for entries in orange.

APPENDIX B. EXPERIMENTAL VALUES

	0_1	2_1	2_1^m	2_1^{ms}	2_1^s	3_1	3_1^m	3_2	3_2^m	3_2^{ms}	3_2^s	4_1	4_2	4_2^m
0_1	0	1	1	1	1	2	2	1	1	1	1	2	1	1
2_1	1	0	2	2	2	1	3	2	2	2	2	3	1	2
2_1^m	1	2	0	2	2	3	1	2	2	2	2	3	2	1
2_1^{ms}	1	2	2	0	2	1	3	2	2	2	2	3	2	2
2_1^s	1	2	2	2	0	3	1	2	2	2	2	3	2	2
3_1	2	1	3	1	3	0	4	3	3	3	3	4	2	3
3_1^m	2	3	1	3	1	4	0	3	3	3	3	4	3	2
3_2	1	2	2	2	2	3	3	0	2	2	1	1	2	2
3_2^m	1	2	2	2	2	3	3	2	0	1	2	1	2	2
3_2^{ms}	1	2	2	2	2	3	3	2	1	0	2	1	2	2
3_2^s	1	2	2	2	2	3	3	1	2	2	0	1	2	2
4_1	2	3	3	3	3	4	4	1	1	1	1	0	3	3
4_2	1	1	2	2	2	2	3	2	2	2	2	3	0	2
4_2^m	1	2	1	2	2	3	2	2	2	2	2	3	2	0
4_2^{ms}	1	2	2	1	2	2	3	2	2	2	2	3	2	2
4_2^s	1	2	2	2	1	3	2	2	2	2	2	3	2	2
4_3	2	1	3	3	3	2	4	3	3	3	3	4	2	3
4_3^m	2	3	1	3	3	4	2	3	3	3	3	4	3	2
4_3^{ms}	2	3	3	1	3	2	4	3	3	3	3	4	3	3
4_3^s	2	3	3	3	1	4	2	3	3	3	3	4	3	3
4_4	2	2	3	1	3	1	4	3	3	3	3	4	3	3
4_4^m	2	3	2	3	1	4	1	3	3	3	3	4	3	3
4_4^{ms}	2	1	3	2	3	1	4	3	3	3	3	4	2	3
4_4^s	2	3	1	3	2	4	1	3	3	3	3	4	3	2
4_5	2	3	1	3	3	4	2	1	3	3	2	2	3	2
4_5^m	2	1	3	3	3	2	4	3	1	2	3	2	2	3
4_5^{ms}	2	3	3	3	1	4	2	3	2	1	3	2	3	3
4_5^s	2	3	3	1	3	2	4	2	3	3	1	2	3	3
4_6	3	2	4	2	4	1	5	4	4	4	4	5	3	4
4_6^m	3	4	2	4	2	5	1	4	4	4	4	5	4	3
4_6^{ms}	3	2	4	2	4	1	5	4	4	4	4	5	3	4
4_6^s	3	4	2	4	2	5	1	4	4	4	4	5	4	3
4_7	1	2	1	2	2	3	2	2	2	2	2	3	2	2
4_7^m	1	1	2	2	2	2	3	2	2	2	2	3	2	2
4_7^{ms}	1	2	2	2	1	3	2	2	2	2	2	3	2	2
4_7^s	1	2	2	1	2	2	3	2	2	2	2	3	2	2
4_8	1	1	2	2	2	2	3	2	2	2	2	3	2	2
4_8^m	1	2	1	2	2	3	2	2	2	2	2	3	2	2
4_8^{ms}	1	2	2	1	2	2	3	2	2	2	2	3	2	2
4_8^s	1	2	2	2	1	3	2	2	2	2	2	3	2	2

TABLE 3. Table of experimental f-distances of all knotoids with up to 4 crossings (part 1).

	4_2^{ms}	4_2^{s}	4_3	4_3^{m}	4_3^{ms}	4_3^{s}	4_4	4_4^{m}	4_4^{ms}	4_4^{s}	4_5	4_5^{m}	4_5^{ms}	4_5^{s}
0_1	1	1	2	2	2	2	2	2	2	2	2	2	2	2
2_1	2	2	1	3	3	3	2	3	1	3	3	1	3	3
2_1^m	2	2	3	1	3	3	3	2	3	1	1	3	3	3
2_1^{ms}	1	2	3	3	1	3	1	3	2	3	3	3	3	1
2_1^s	2	1	3	3	3	1	3	1	3	2	3	3	1	3
3_1	2	3	2	4	2	4	1	4	1	4	4	2	4	2
3_1^m	3	2	4	2	4	2	4	1	4	1	2	4	2	4
3_2	2	2	3	3	3	3	3	3	3	3	1	3	3	2
3_2^m	2	2	3	3	3	3	3	3	3	3	3	1	2	3
3_2^{ms}	2	2	3	3	3	3	3	3	3	3	3	2	1	3
3_2^s	2	2	3	3	3	3	3	3	3	3	2	3	3	1
4_1	3	3	4	4	4	4	4	4	4	4	2	2	2	2
4_2	2	2	2	3	3	3	3	3	2	3	3	2	3	3
4_2^m	2	2	3	2	3	3	3	3	3	2	2	3	3	3
4_2^{ms}	0	2	3	3	2	3	2	3	3	3	3	3	3	2
4_2^s	2	0	3	3	3	2	3	2	3	3	3	3	2	3
4_3	3	3	0	4	4	4	3	4	2	4	4	2	4	4
4_3^m	3	3	4	0	4	4	4	3	4	2	2	4	4	4
4_3^{ms}	2	3	4	4	0	4	2	4	3	4	4	4	4	2
4_3^s	3	2	4	4	4	0	4	2	4	3	4	4	2	4
4_4	2	3	3	4	2	4	0	4	2	4	4	3	4	2
4_4^m	3	2	4	3	4	2	4	0	4	2	3	4	2	4
4_4^{ms}	3	3	2	4	3	4	2	4	0	4	4	2	4	3
4_4^s	3	3	4	2	4	3	4	2	4	0	2	4	3	4
4_5	3	3	4	2	4	4	4	3	4	2	0	4	4	3
4_5^m	3	3	2	4	4	4	3	4	2	4	4	0	3	4
4_5^{ms}	3	2	4	4	4	2	4	2	4	3	4	3	0	4
4_5^s	2	3	4	4	2	4	2	4	3	4	3	4	4	0
4_6	3	4	3	5	3	5	2	5	2	5	5	3	5	3
4_6^m	4	3	5	3	5	3	5	2	5	2	3	5	3	5
4_6^{ms}	3	4	3	5	3	5	2	5	2	5	5	3	5	3
4_6^s	4	3	5	3	5	3	5	2	5	2	3	5	3	5
4_7	2	2	3	2	3	3	3	3	3	2	2	3	3	3
4_7^m	2	2	2	3	3	3	3	3	2	3	3	2	3	3
4_7^{ms}	2	2	3	3	3	2	3	2	3	3	3	3	2	3
4_7^s	2	2	3	3	2	3	2	3	3	3	3	3	3	2
4_8	2	2	2	3	3	3	3	3	2	3	3	2	3	3
4_8^m	2	2	3	2	3	3	3	3	3	2	2	3	3	3
4_8^{ms}	2	2	3	3	2	3	2	3	3	3	3	3	3	2
4_8^s	2	2	3	3	3	2	3	2	3	3	3	3	2	3

TABLE 4. Table of experimental f-distances of all knotoids with up to 4 crossings (part 2).

	4_6	4_6^m	4_6^{ms}	4_6^s	4_7	4_7^m	4_7^{ms}	4_7^s	4_8	4_8^m	4_8^{ms}	4_8^s
0_1	3	3	3	3	1	1	1	1	1	1	1	1
2_1	2	4	2	4	2	1	2	2	1	2	2	2
2_1^m	4	2	4	2	1	2	2	2	2	1	2	2
2_1^{ms}	2	4	2	4	2	2	2	1	2	2	1	2
2_1^s	4	2	4	2	2	2	1	2	2	2	2	1
3_1	1	5	1	5	3	2	3	2	2	3	2	3
3_1^m	5	1	5	1	2	3	2	3	3	2	3	2
3_2	4	4	4	4	2	2	2	2	2	2	2	2
3_2^m	4	4	4	4	2	2	2	2	2	2	2	2
3_2^{ms}	4	4	4	4	2	2	2	2	2	2	2	2
3_2^s	4	4	4	4	2	2	2	2	2	2	2	2
4_1	5	5	5	5	3	3	3	3	3	3	3	3
4_2	3	4	3	4	2	2	2	2	2	2	2	2
4_2^m	4	3	4	3	2	2	2	2	2	2	2	2
4_2^{ms}	3	4	3	4	2	2	2	2	2	2	2	2
4_2^s	4	3	4	3	2	2	2	2	2	2	2	2
4_3	3	5	3	5	3	2	3	3	2	3	3	3
4_3^m	5	3	5	3	2	3	3	3	3	2	3	3
4_3^{ms}	3	5	3	5	3	3	3	2	3	3	2	3
4_3^s	5	3	5	3	3	3	2	3	3	3	3	2
4_4	2	5	2	5	3	3	3	2	3	3	2	3
4_4^m	5	2	5	2	3	3	2	3	3	3	3	2
4_4^{ms}	2	5	2	5	3	2	3	3	2	3	3	3
4_4^s	5	2	5	2	2	3	3	3	3	2	3	3
4_5	5	3	5	3	2	3	3	3	3	2	3	3
4_5^m	3	5	3	5	3	2	3	3	2	3	3	3
4_5^{ms}	5	3	5	3	3	3	2	3	3	3	3	2
4_5^s	3	5	3	5	3	3	3	2	3	3	2	3
4_6	0	6	2	6	4	3	4	3	3	4	3	4
4_6^m	6	0	6	2	3	4	3	4	4	3	4	3
4_6^{ms}	2	6	0	6	4	3	4	3	3	4	3	4
4_6^s	6	2	6	0	3	4	3	4	4	3	4	3
4_7	4	3	4	3	0	2	2	2	2	2	2	2
4_7^m	3	4	3	4	2	0	2	2	2	2	2	2
4_7^{ms}	4	3	4	3	2	2	0	2	2	2	2	2
4_7^s	3	4	3	4	2	2	2	0	2	2	2	2
4_8	3	4	3	4	2	2	2	2	0	2	2	2
4_8^m	4	3	4	3	2	2	2	2	2	0	2	2
4_8^{ms}	3	4	3	4	2	2	2	2	2	2	0	2
4_8^s	4	3	4	3	2	2	2	2	2	2	2	0

TABLE 5. Table of experimental f-distances of all knotoids with up to 4 crossings (part 3).