A conjecture about an upper bound of the RMSD between linear chains

Christian L. Müller, Ivo F. Sbalzarini *

Abstract

We combine stochastic global optimization and analytical geometry in order to conjecture an upper bound for the Root Mean Square Deviation (RMSD) between linear chains of N beads with link length b after optimal roto-translational fitting. We report pairs of putative extremal configurations and an analytical expression for the RMSD between them, asymptotically approaching $\frac{1}{4}\sqrt{\frac{5}{3}}bN$ for large N.

1 Introduction

Since the pioneering works of Flory [3] chain models have been instrumental for theoretical studies of polymers. The simplest models are linear chains such as the freely jointed (or ideal) Random Walk (RW) or the Self-Avoiding Walk (SAW). These models provide the theoretical basis for more complex (bio-)polymer and protein models [2]. A linear chain is a configuration of ordered points (beads, atoms) in three-dimensional space, where the Euclidean distance between consecutive beads is constrained to an arbitrary but fixed constant b, the bond or link length.

The advent of efficient algorithms for determining the minimum Root Mean Square Deviation (RMSD) [10, 7] between two linear chains has triggered research in characterizing the configuration space of chain ensembles using RMSD as the standard distance metric in the field. Starting from ideal RW ensembles [11] the analysis has been extended to more complex polymer and protein models [14].

While the minimum RMSD between two configurations reaches a trivial lower bound of 0 for identical chains, a tight upper bound – or the two configurations of linear chains that are most dissimilar from each other – is still unknown.

We address this problem using a combination of global optimization and analytical geometry. We numerically determine the maximum RMSD of RW chains for several N and deduce from these results a general formula for odd N. We conjecture that the asymptotic limit of this formula is valid for all N and that it is an upper bound for the maximum RMSD between general linear chains.

*Institute of Theoretical Computer Science, Department of Computer Science, ETH Zürich, and Swiss Institute of Bioinformatics, christian.mueller@inf.ethz.ch, ivos@ethz.ch

2 Definitions and Methods

2.1 The minimum RMSD

We represent two configurations of N beads each by the matrices X, $Y \in \mathbb{R}^{3\times N}$. Each column in X, Y is denoted $\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$ and contains the three-dimensional Cartesian coordinates of the i^{th} bead of the configuration. In a linear chain model, consecutive beads are connected by links of fixed length b. Calculating the minimum RMSD $D(X,Y)$ between X and Y comprises two steps: (i) translating the centers of mass $\mathbf{x}_{\text{cm}}$ and $\mathbf{y}_{\text{cm}}$ of both configurations to the origin, leading to repositioned chains X_0 and Y_0 with columns $\mathbf{x}_0^{(i)}, \mathbf{y}_0^{(i)}$; (ii) determining the optimal rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$, such that:

$$D^2(X,Y) \doteq \min_{\mathbf{R}} \frac{1}{N} \|\mathbf{R}X_0 - Y_0\|_2^2 \,. \tag{1}$$

The optimal rotation matrix $\mathbf{R}$ can be determined using Singular Value Decomposition (SVD) [6, 7] or quaternions [8]. It is a special case of the *orthogonal Procrustes problem* ([4], pp. 601) where $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}_3$ (the 3×3 identity matrix) and $\det \mathbf{R} = 1$.

$D^2(X,Y)$ can be expressed in terms of the radii of gyration of X and Y, $R_G(X)$ and $R_G(Y)$, as [10, 11]:

$$D^2(X,Y) = R_G^2(X) + R_G^2(Y) - 2\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{x}}_0^{(i)} \cdot \mathbf{y}_0^{(i)} \tag{2}$$

with $\tilde{\mathbf{x}}_0^{(i)} = \mathbf{R}\mathbf{x}_0^{(i)}$ and $R_G^2(X) = \mathrm{tr}(X^{\mathrm{T}}X)$. The term $\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{x}}_0^{(i)} \cdot \mathbf{y}_0^{(i)}$ describes the structural correlation between X and Y after optimal superposition and can be re-written as [1]:

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{x}}_0^{(i)} \cdot \mathbf{y}_0^{(i)} = \frac{\sum\limits_{i=1}^{N} \tilde{\mathbf{x}}_0^{(i)} \cdot \mathbf{y}_0^{(i)}}{\sqrt{\sum\limits_{i=1}^{N} \mathbf{x}_0^{(i)2} \sum\limits_{i=1}^{N} \mathbf{y}_0^{(i)2}}} R_G(X) R_G(Y) \,. \tag{3}$$

Betancourt and Skolnick [1] refer to the fraction in Eq. (3) as the *aligned correlation coefficient* $ACC(X,Y)$. The radius of gyration R_G of a chain X is roto-translation invariant and can be written as:

$$R_G^2(X) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \mathbf{x}_{\text{cm}}\|_2^2 \tag{4}$$

$$= -\mathbf{x}_{\text{cm}} \cdot \mathbf{x}_{\text{cm}} + \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}^{(i)}\|^2 \,.$$

From Eq. (2), McLachlan derived relative lower and upper bounds for $D^2(X,Y)$ of two given chains X and Y [11]:

$$0 \leq D^2(X,Y) \leq R_G^2(X) + R_G^2(Y) \,. \tag{5}$$

2.2 The linear chain RW model

A linear chain X is represented in internal coordinates. We denote by θ_i the angle between three consecutive beads $\mathbf{x}^{(i)}$, $\mathbf{x}^{(i+1)}$, $\mathbf{x}^{(i+2)}$. The dihedral between the two consecutive planes spanned by ($\mathbf{x}^{(i)}$, $\mathbf{x}^{(i+1)}$, $\mathbf{x}^{(i+2)}$) and ($\mathbf{x}^{(i+1)}$, $\mathbf{x}^{(i+2)}$, $\mathbf{x}^{(i+3)}$) is ω_i (see Fig. 1a). A chain of N beads with fixed bond length

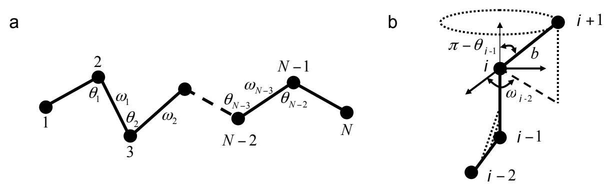

Figure 1: a. Definition of the angles θ_i and dihedrals ω_i characterizing an anchored walk of length N. b. Illustration of the trigonometric map $\mathbf{q}_X \to X$ from internal coordinates $\mathbf{q}_X$ to Cartesian coordinates X.

b has $N-2$ angles and $N-3$ dihedrals, resulting in $M = 2N-5$ degrees of freedom. It is described by the internal coordinate vector $\mathbf{q}_X \doteq \{\theta_i | i = 1 \ldots N-2, \omega_i | i = 1 \ldots N-3\}$. For an ideal RW chain, the direction of each link is chosen uniformly random on the unit sphere by sampling the $\cos(\theta_i)$ uniformly from $[0,1]$ and the ω_i uniformly from $[0,\pi]$ [2]. The link length between consecutive beads is fixed to b, the mass of each bead is $m = \frac{1}{N}$.

In order to avoid redundant chains that can be superimposed by rigid-body translation and rotation, we use "anchored" walks where $\mathbf{x}^{(1)}$ is placed at the origin, $\mathbf{x}^{(2)}$ along the x-axis at $(b,0,0)^{\mathrm{T}}$, and the link between $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(3)}$ is contained in the xy-plane. This uniquely defines the overall position and orientation of the walk [13].

A pair of anchored RW chains (X,Y) of N beads each is represented by $\mathbf{q}_S = (\mathbf{q}_X, \mathbf{q}_Y)$. The transformation from internal coordinates to three-dimensional Cartesian coordinates is denoted by $J(\mathbf{q}_S) = (X,Y)$ (see Fig. 1b).

2.3 The maximum RMSD problem

The *maximum RMSD problem* (MAX-RMSD) is stated as a continuous, non-convex max–min optimization problem. We seek the specific pair of chains ($X_{\max}^N$, $Y_{\max}^N$) of N beads each that maximizes $D^2(X,Y)$ over all possible X and Y, hence:

$$\begin{aligned}(X_{\max}^N, Y_{\max}^N) &= \arg\max_{X,Y} D^2(X,Y) \qquad (6)\\ &= \arg\max_{X,Y} \min_{\mathbf{R}} \frac{1}{N} \|\mathbf{R}X_0 - Y_0\|_2^2 \,.\end{aligned}$$

We refer to the pair $(X_{\max}^N, Y_{\max}^N)$ as the *extremal configurations* of the chain ensemble in the RMSD sense. If both X and Y are anchored linear RW chains, we call the problem RW-MAX-RMSD. Its $D_{\max}(N) = \sqrt{D^2(X_{\max}^N, Y_{\max}^N)}$ is an upper bound for the maximum RMSD of all linear chain ensembles.

The inner minimization problem can be solved analytically by constructing the optimal rotation matrix $\mathbf{R}$ from SVD [6, 7, 4] or using quaternions [8, 9]. The distance constraints on the positions of consecutive beads $\|\mathbf{x}_i - \mathbf{x}_{i+1}\|_2 = b, \quad i = 1, \ldots, N-1$, are satisfied by using internal coordinates $\mathbf{q}$.

The outer maximization problem can be formulated as a constrained, non-convex black-box optimization problem in $n = 2(2N-5) = 4N-10$ dimensions. For convenience we consider the unit hypercube as feasible domain, i.e., candidate solution vectors $\hat{\mathbf{q}}_{\mathrm{S}}$ are in $[0,1]^n$. The unique map $T : \hat{\mathbf{q}}_{\mathrm{S}} \in [0,1]^n \to \mathbf{q}_{\mathrm{S}} \in ([0,1]^{2(N-2)}, [0,\pi]^{2(N-3)})$ transforms any candidate solution vector to internal coordinates of a chain. The black-box objective function f to be maximized then reads:

$$\begin{aligned}f(\hat{\mathbf{q}}_{\mathrm{S}}) &\equiv D^2\left(J(T(\hat{\mathbf{q}}_S))\right) = D^2(X,Y) \qquad (7)\\ &= \min_{\mathbf{R}} \frac{1}{N} \|\mathbf{R}X_0 - Y_0\|^2 \,.\end{aligned}$$

Note that this formulation is known *a priori* to become two-fold degenerate if two consecutive links in a trial configuration are co-linear. First, the corresponding dihedral angles are then undefined, i.e., the configuration remains the same regardless of their values. Second, the optimal rotation matrix has only rank 1, permitting infinitely many rotations that minimize $D^2(X,Y)$ [8].

3 Numerical optimization results

We numerically solve the RW-MAX-RMSD problem for pairs of configurations with $N = 3, \ldots, 16$ beads. The dimensionality of the problem is thus ranging from $n = 2, \ldots, 50$. We use two optimization algorithms: (i) Sequential Quadratic Programming (SQP) and (ii) Best Local Restart Covariance Matrix Adaptation Evolution Strategy (BLR-CMA-ES). For SQP, the MATLAB implementation *fmincon* is used. For box-constrained black-box optimization problems, this implementation uses an active-set SQP algorithm with approximate BFGS and line search. BLR-CMA-ES is a local restart variant of the variable-metric optimizer CMA-ES [5]. Details of

BLR-CMA-ES and the set-up of the numerical experiments have been described elsewhere [12].

The putative optimal solutions $(X^N_{\max}, Y^N_{\max})$ found by SQP and BLR-CMA-ES agree for $N = 3, 5, 7, 11$. For all other instances, BLR-CMA-ES consistently outperforms SQP, finding configurations with larger minimum RMSD than those found by SQP. These extremal configuration are shown in Fig. 2. For odd N, they follow a regular geometric pattern: one configuration always is the fully extended linear rod, the other is a linear rod of half the length with beads $\frac{N+3}{2}$ to N folded back onto beads $\frac{N-1}{2}$ to 1. For even N, the first extremal configuration is again the fully extended linear rod, whereas the other is a *planar hairpin* with crossed ends. For odd N, the *ACC* of the extremal configurations is virtually 0 ($< 10^{-15}$), for even N it is $< 10^{-3}$. These optima found by BLR-CMA-ES suggest a near-linear dependence of $D_{\max}(N)$ on N with a best linear fit of $D_{\max}(N) \approx 0.3251bN - 0.04013$.

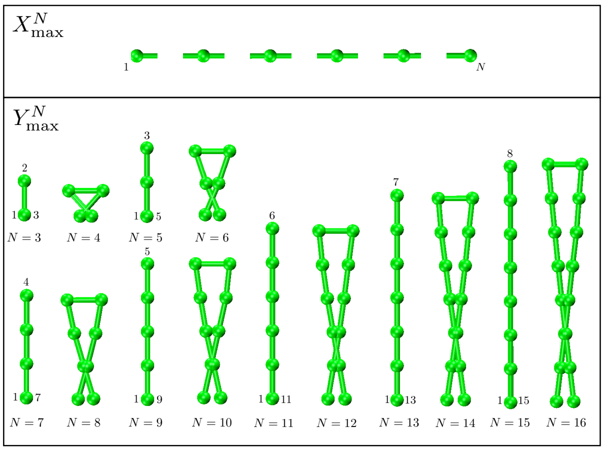

Figure 2: Extremal configurations $(X^N_{\max}, Y^N_{\max})$ of linear RW chains with $N = 3, \ldots, 16$ found by BLR-CMA-ES. The upper box shows the extended configurations $X^N_{\max}$. The lower box shows the corresponding configurations $Y^N_{\max}$. For odd N, $Y^N_{\max}$ is a linear rod of half the length with beads $\frac{N+3}{2}$ to N folded back onto beads $\frac{N-1}{2}$ to 1. For even N, $Y^N_{\max}$ is a planar hairpin where the links from beads $\frac{N+2}{2}$ to N cross the links from beads $\frac{N}{2}$ to 1.

4 The MAX-RMSD conjecture

The above results suggest that the extremal configurations for odd N follow a simple geometric pattern: one configuration is the fully extended linear rod, the other one a linear fold-back of half the length.

Conjecture 1 *The fully extended linear rod and its linear fold-back configuration are the optimal solution of the RW-MAX-RMSD problem for all odd N.*

Under this assumption we derive a general formula for $D_{\max}(N)$ for odd N. Combining Eqs. (2) and (3) we find:

$$\begin{aligned} D^2_{\max}(N) &= D^2(X^N_{\max}, Y^N_{\max}) = R^2_G(X^N_{\max}) + R^2_G(Y^N_{\max}) \\ &- 2ACC(X^N_{\max}, Y^N_{\max}) R_G(X^N_{\max}) R_G(Y^N_{\max}) \\ &= R^2_G(X^N_{\max}) + R^2_G(Y^N_{\max}), \end{aligned} \tag{8}$$

provided that $ACC(X^N_{\max}, Y^N_{\max}) = 0$ for all odd N.

Lemma 1 *The $ACC(X^N_{\max}, Y^N_{\max})$ for odd N is 0.*

Proof. Without loss of generality we assume that $X^N_{\max}$ is the fully extended rod and $Y^N_{\max}$ the back-folded one, and that their centers of mass are at $(0, 0, 0)$. For odd N, the problem of optimal superposition then reduces to a rotation in the xy-plane. We define the x-axis to be aligned with $X^N_{\max}$ after optimal superposition. $Y^N_{\max}$ forms a certain rotation angle α with $X^N_{\max}$ as shown in Fig. 3. We show that for the specific pair of configurations $(X^N_{\max}, Y^N_{\max})$ the $ACC(X^N_{\max}, Y^N_{\max})$ is 0 for any rotation angle α and any odd N by recalling the definition of ACC for two optimally aligned chains X, Y:

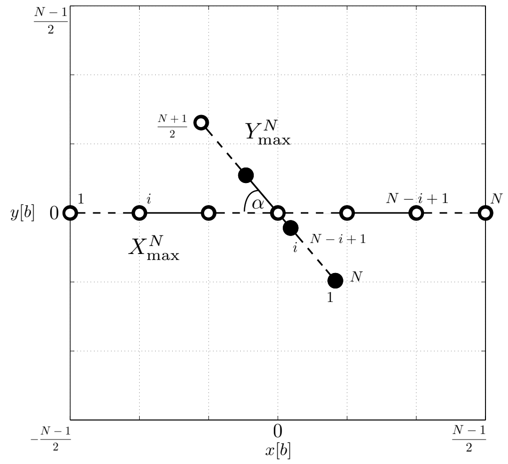

Figure 3: Calculation of the RMSD between $X^N_{\max}$ and $Y^N_{\max}$ for odd N after optimal superposition. $X^N_{\max}$ is the extended configuration and $Y^N_{\max}$ the folded one. Open circles (∘) represent positions that are occupied by single beads, filled circles (•) indicate positions occupied by two beads. The two configurations enclose a planar angle α.

$$ACC(X, Y) = \frac{\sum_{i=1}^N \mathbf{x}^{(i)} \cdot \mathbf{y}^{(i)}}{\sqrt{\sum_{i=1}^N \mathbf{x}^{(i)2} \sum_{i=1}^N \mathbf{y}^{(i)2}}}. \tag{9}$$

The denominator of this expression is always positive since the two factors under the square root are sums over squared bead coordinates.

From Fig. 3 we see that the coordinate vectors $\mathbf{x}^{(i)}_{\max}$ only have non-zero entries in x-direction. Furthermore, the x coordinate of the i^{th} bead in $X^N_{\max}$ is the negative of the x coordinate of the $(N-i+1)^{\text{th}}$ bead. The central bead (i.e., the $\left(\frac{N+1}{2}\right)^{\text{th}}$ bead) in $X^N_{\max}$ is at $(0,0,0)$, so the scalar product with its corresponding bead in $Y^N_{\max}$ is 0. The positions of the i^{th} and $(N-i+1)^{\text{th}}$ beads in $Y^N_{\max}$ are identical (filled circles in Fig. 3) for all α. The numerator in Eq. (9) hence becomes:

$$\sum_{i=1}^{N} \mathbf{x}^{N,(i)}_{\max} \cdot \mathbf{y}^{N,(i)}_{\max} = \sum_{i=1}^{\frac{N+1}{2}-1} \mathbf{x}^{N,(i)}_{\max} \cdot \mathbf{y}^{N,(i)}_{\max} + 0$$
$$+ \sum_{i=\frac{N+1}{2}-1}^{N} \mathbf{x}^{N,(i)}_{\max} \cdot \mathbf{y}^{N,(i)}_{\max} = - \sum_{i=\frac{N+1}{2}-1}^{N} \mathbf{x}^{N,(i)}_{\max} \cdot \mathbf{y}^{N,(i)}_{\max} +$$
$$\sum_{i=\frac{N+1}{2}-1}^{N} \mathbf{x}^{N,(i)}_{\max} \cdot \mathbf{y}^{N,(i)}_{\max} = 0 \tag{10}$$

and, therefore, $ACC(X^N_{\max}, Y^N_{\max}) = 0$ for all odd N and all rotation angles α. □

Observation 1 *The radii of gyration of X^N_{max} and Y^N_{max} for odd N are*

$$R^2_G(X^N_{max}) = \frac{2}{N} b^2 \sum_{i=1}^{M^-} (i)^2 \tag{11}$$

$$R^2_G(Y^N_{max}) = -b^2 \left(\frac{M^- M^-}{N}\right)^2 + \frac{1}{N} b^2 \left((M^-)^2 + 2\sum_{i=1}^{\hat{M}^-}(i)^2\right) \tag{12}$$

with $M^- = \frac{N-1}{2}$ and $\hat{M}^- = \frac{N-3}{2}$.

A derivation of these expressions can be found in Ref. [12]. Combining Eqs. (8), (11), and (12) yields an analytic formula for $D_{\max}(N)$ for odd N, asymptotically approaching [12]:

$$\hat{D}_{\max}(N) = \lim_{N\to\infty} D_{\max}(N) = \frac{1}{4}\sqrt{\frac{5}{3}} bN \,. \tag{13}$$

We conjecture that this asymptotic limit is valid also for even N and, since the maximum RMSD of RW chains is always larger than that of any other chain ensemble, formulate:

Conjecture 2 *$\hat{D}_{max}(N)$ is an asymptotic upper bound on the RMSD between any two linear chains.*

5 Conclusion

We combined stochastic global optimization and analytic geometry in order to conjecture an upper bound for the RMSD between linear chains of N beads with link length b after optimal roto-translational fitting. We reported pairs of putative extremal configurations of RW chains and an analytical expression for the maximum RMSD between these extremal configurations for odd N. This expression asymptotically approaches $\frac{1}{4}\sqrt{\frac{5}{3}}bN$ for large N, which is the conjectured upper bound for any two linear chains for all N. Future research will try proving this conjecture.

Acknowledgments

We thank Dr. Bojan Žagrović and Dr. Philippe Hünenberger for many insightful discussions.

References

[1] M. R. Betancourt and J. Skolnick. Universal similarity measure for comparing protein structures. *Biopolymers*, 59(5):305–309, Oct 2001.

[2] C. R. Cantor and R. Schimmel, Paul. *Biophysical chemistry Part III: The behavior of biological macromolecules.* W. H. Freeman, New York, 1980.

[3] P. J. Flory. *Statistical Mechanics of Chain Molecules.* Wiley-Interscience, New York, 1969.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, 3rd edition, 1996.

[5] N. Hansen. *The CMA Evolution Strategy: A Tutorial*, Nov 2007.

[6] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A.*, 32(SEP1):922–923, 1976.

[7] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A.*, 34(SEP):827–828, 1978.

[8] G. R. Kneller. Superposition of Molecular Structures using Quaternions. *Mol. Simulat.*, 7(1):113–119, 1991.

[9] G. R. Kneller. Comment on "Using quaternions to calculate RMSD" - [J. Comp. Chem. 25, 1849 (2004)]. *J. Comp. Chem.*, 26(15):1660–1662, Nov 2005.

[10] A. D. McLachlan. Mathematical procedure for superimpsoing atomic coordinates of proteins. *Acta Crystallogr. A.*, A 28(NOV1):656–657, 1972.

[11] A. D. McLachlan. How alike are the Shapes of Two Random Chains. *Biopolymers*, 23(7):1325–1331, 1984.

[12] C. L. Müller. *Black-box Landscapes: Characterization, Optimization, Sampling, and Application to Geometric Configuration Problems.* PhD thesis, Institute of Theoretical Computer Science, Department of Computer Science, ETH Zürich, 2010.

[13] C. L. Müller, I. F. Sbalzarini, W. F. van Gunsteren, B. Zagrovic, and P. H. Huenenberger. In the eye of the beholder: Inhomogeneous distribution of high-resolution shapes within the random-walk ensemble. *J. Chem. Phys.*, 130(21), Jun 7 2009.

[14] D. C. Sullivan and I. D. Kuntz. Distributions in protein conformation space: Implications for structure prediction and entropy. *Biophys. J.*, 87(1):113–120, Jul 2004.