Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 6 May 2002
Accepted 2 July 2002

A revised proof of the metric properties of optimally superimposed vector sets

Boris Steipe

Department of Biochemistry and Program in Proteomics and Bioinformatics, University of Toronto, Toronto, Canada. Correspondence e-mail: boris.steipe@utoronto.ca

A revised proof is given that the root-mean-square deviation between more than two vector sets after optimal superposition induces a metric. This corrects an error in a previous manuscript [Kaindl & Steipe (1997). *Acta Cryst.* A**53**, 809].

The $\mathrm{RMSD}_{\mathrm{opt}}$ (root mean-square deviation after optimal superposition) is commonly used as a measure of the similarity of two n-dimensional vector sets (*e.g.* molecular structures). An analytic solution to the calculation of the rotation and translation that minimizes the RMSD of vector sets has been given by Kabsch (1976). If more than two vector sets are to be compared, the question arises whether the $\mathrm{RMSD}_{\mathrm{opt}}$ is a metric in the mathematical sense. Metric properties are often implicit in data-analysis procedures like clustering, sampling or averaging. A previous communication to this effect (Kaindl & Steipe, 1997) contains a technical error that invalidates the proof that the $\mathrm{RMSD}_{\mathrm{opt}}$ is a metric (see Steipe, 2002, for an Erratum). Here I provide a revised proof.

A measure d on a set X, $d\colon X \times X \Rightarrow R_+$ is called a metric, and the pair (X,d) is called a metric space, if

(i) $\forall x,y \in X\colon\ d(x,y) \geq 0,\ d(x,y) = 0 \Leftrightarrow x = y$ (positivity),
(ii) $d(y,x) = d(x,y)$ (symmetry), and
(iii) $\forall z \in X\colon\ d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality).

We consider $X \coloneqq R^{3,n}$ and sets $x \in X\colon \{x_1, \ldots, x_n\},\ x_i \in R^3$. The $\mathrm{RMSD}(x,y) \coloneqq (1/n)[\sum_{i=1}^{n} ||x_i - y_i||^2]^{1/2}$ is equivalent to the Euclidean norm $\|x - y\|$ in $R^{3,n}$, save for the common dividing factor n, thus it is obviously a metric. But it is not obvious whether $\mathrm{RMSD}_{\mathrm{opt}}(x,y)$, the RMSD after optimal superposition, is a metric.

After optimal superposition, vector set centroids coincide (Kabsch, 1976). Let $c_x \in R^3$ be the centroid of x, $t_x \in R^{3,n}$ be $\{c_x, \ldots, c_x\}$, and $M, M_{x\Rightarrow y} \in R^{3,3}$ be proper rotation matrices.

$$\mathrm{RMSD}_{\mathrm{opt}}(x,y) \coloneqq \min ||(y - t_y) - \{M(x_1 - c_x), \ldots, M(x_n - c_x)\}|| \Leftrightarrow ||(y - t_y) - \{M_{x\to y}(x_1 - c_x), \ldots, M_{x\to y}(x_n - c_x)\}||$$

and we write $\tilde{x}_y$ for an x that has been transformed by optimal superposition on y:

$$\tilde{x}_y \coloneqq \{M_{x\to y}(x_1 - c_x), \ldots, M_{x\to y}(x_n - c_x)\} + t_y$$

(*etc.* for y and z).

If two vector sets differ only by an arbitrary rotation and translation, their components coincide after optimal superposition and their $\mathrm{RMSD}_{\mathrm{opt}}$ is 0. We may call such vector sets equivalent and define an equivalence class $\mathbf{x}$: $\{x, x' \in R^{3,n},\ \mathrm{RMSD}_{\mathrm{opt}}(x,x') = 0\}$. Since $x, \tilde{x}_y \in \mathbf{x}$, the $\mathrm{RMSD}_{\mathrm{opt}}(x,y)$ does not depend on the particular choice of $x \in \mathbf{x}$, $y \in \mathbf{y}$; any x', y' can be superimposed identically on an arbitrary reference pair x, y. It follows that metric properties of vector sets after optimal superposition are stated with respect to the entire equivalence class $\mathbf{x}$, not just individual vector sets. We may state (i)–(iii) for the $\mathrm{RMSD}_{\mathrm{opt}}(x,y)$, with reference to their respective equivalence classes, by considering $X \coloneqq R^{3,n}$ and the mapping $d\colon R^{3,n} \times R^{3,n} \to R_+$, $d(\mathbf{x},\mathbf{y}) \coloneqq \|x - \tilde{y}_x\|$ with arbitrary choice of $x \in \mathbf{x}$:

(i′) $\forall x,y \in R^{3,n}\colon\ d(\mathbf{x},\mathbf{y}) \geq 0,\ d(\mathbf{x},\mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$,
(ii′) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$, and
(iii′) $\forall z \in R^{3,n}\colon\ d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$.

Condition (i′) can be restated as $\forall x,y \in R^{3,n}\colon \|x - \tilde{y}_x\| \geq 0,\ \|x - \tilde{y}_x\| = 0 \Leftrightarrow x,\ y \in \mathbf{x}$, and follows from the metric properties of $\|x_i - y_i\|$, $x_i, y_i \in R^3$ and the definition of $\mathbf{x}$.

Condition (ii′) can be restated as $\|x - \tilde{y}_x\| = \|y - \tilde{x}_y\|$, which follows from the metric properties of $\|x_i - y_i\|$, $x_i,\ y_i \in R^3$.

Condition (iii′) can be restated as $\forall z \in R^{3,n}\colon \|x - \tilde{z}_x\| \leq \|x - \tilde{y}_x\| + \|y - \tilde{z}_y\|$.

We note that $\|y - \tilde{z}_y\| = \|\tilde{y}_x - \tilde{z}_{\tilde{y}_x}\|$, since $y, \tilde{y}_x \in \mathbf{y}$ and $\tilde{z}_y, \tilde{z}_{\tilde{y}_x} \in \mathbf{z}$. Thus superimposing $\tilde{z}_y$ on $\tilde{y}_x$ lets us rewrite (iii′) as $\|x - \tilde{z}_x\| \leq \|x - \tilde{y}_x\| + \|\tilde{y}_x - \tilde{z}_{\tilde{y}_x}\|$.

The triangle inequality in $R^{3,n}$ on the triple $(x, \tilde{y}_x, \tilde{z}_{\tilde{y}_x})$ implies that $\|x - \tilde{z}_{\tilde{y}_x}\| \leq \|x - \tilde{y}_x\| + \|\tilde{y}_x - \tilde{z}_{\tilde{y}_x}\|$.

Finally, the definition of $\tilde{z}_x$ implies $\|x - \tilde{z}_x\| \leq \|x - \tilde{z}_{\tilde{y}_x}\|$. **q.e.d.**

The author is indebted to Dr Xu Huafeng for pointing out a technical error in the previously published manuscript (Kaindl & Steipe, 1997). Thanks to Gerald Ted Quon for discussions and an anonymous referee for pointing out the necessity to operate with equivalence classes.

References

Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kaindl, K. & Steipe, B. (1997). *Acta Cryst.* A**53**, 809.
Steipe, B. (2002). *Acta Cryst.* A**58**, 507.