RMSD and Symmetry

Evangelos A. Coutsias[*], Michael J. Wester[†]

February 4, 2019

SUPPORTING INFORMATION

A C Version of frmsd

frmsd was originally written in Fortran 95. We decided to translate the routines into C in order to make a fair comparison with qcp in our timing studies as well as to make our code available to a wider audience. In the process, we tried to optimize frmsd as much as possible.

Some of the optimizations we performed were: loop reordering as C stores arrays by rows while Fortran stores by columns; loop unrolling for loops over the three Cartesian coordinates; replacing Fortran `x = x ± ...` constructions with C increment and decrement operators; minimizing the number of divisions which tend to be computationally expensive (on all the Linux and MacOS systems and compilers that we tried, C double precision floating point divisions were more than twice as slow as multiplications[1]); replacing small arrays by individual variables (computing array element offsets, especially in deeply nested loops, can be expensive); saving pointer references as variables if needed more than a couple of times (in C, 2D arrays are referenced `a[i][j]` where `a[i]` is a pointer to the i^{th} row); in conjunction with the previous, defining and explicitly incrementing array pointer references in heavily used loops rather than performing array indexing (for example, `a_i = a; for (...) { s += f[a_i]; ++a_i; }`); writing explicit code on a case by case basis for dot products involving possibly transposed matrices; inlining small subroutines in order to avoid call overhead; and general code cleanup.

The C frmsd implementation[2] has a number of features. The program operates on a collection of .pdb files in a directory specified on the command line. It is assumed that a file called `files` with the names of all the .pdb files, one entry listed per line, is present in that directory.[3] frmsd has options to compare all the structures with each other [a (array)

[*] Department of Applied Mathematics and Statistics and Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794

[†] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131

[1] frmsd can be compiled for this common situation by defining -DSLOW_DIVISIONS, which we did for all the timings performed here. This option replaces code like `x /= z; y /= z;` by `r = 1.0 / z; x *= r; y *= r;`, which is faster when multiplications are more than twice as fast as divisions and the extra assignment takes negligible relative time, which is common behavior on current hardware.

[2] Available at `http://www.ams.stonybrook.edu/~coutsias/codes/frmsd.tgz`.

[3] In Unix based systems, this file can be created via `ls *.pdb > files`.

and `s` (single)][4], or compare all but the first structure with the first [`l` (list)]. frmsd can then produce just the RMSD, the quaternion form of the rotation matrix, the usual rotation matrix for aligning the structures to minimize the RMSD, and the gradient of the RMSD with respect to the input coordinates. There is an option (for list) to apply the rotations and produce the aligned molecules (either the rest aligned with the first molecule in the list or the first molecule aligned with all the others). Molecules can be filtered in various ways to select a subset of the available atoms for comparison purposes, such as only the Cα, backbone or heavy atoms.[5] Typically, only ATOM records in the .pdb files are used to collect coordinates, but optionally HETATM records can be looked through as well. Various checks and verbose outputs are also available.

A.1 Timings and Consistency

Here, we look at the results of some timings that we performed with the C version of frmsd, along with comparable timings performed with qcp[1–3], which is also written in C. We analyzed three different datasets: 7ring_0.1deg, vhp_mcmd:1vii and lattice_ssfit:1beo. 7ring_0.1deg is a seven membered ring dataset generated using an inverse kinematics algorithm.[4] Specifically, all bonds were set to equal length, all angles to the canonical value 114^o, and one of the torsions (t_1) was sampled over the range $[0^o, 360^o]$ in 0.1^o increments. The remaining six torsions $\{t_i\}_{i=2}^{7}$ were determined by solving the ring closure polynomial, producing none to as many as six alternative sets of values for each torsion t_1. The density of states is high for certain values of t_1, corresponding to very closely spaced (in the sense of RMSD) solutions. Symmetry was not considered for comparisons in the seven member ring dataset, as qcp does not account for this possibility.

vhp_mcmd:1vii[5] is a decoy set[6], that is, a collection of computer generated protein structures, from the chicken villin headpiece thermostable domain (PDB identifier: 1vii). lattice_ssfit:1beo[7] is another decoy set, this one assembled from fragments of unrelated protein structures with similar local sequences. Table SI1 details the basic numerical characteristics of the datasets: the number of molecules (m) and atoms (n), and the number of comparisons performed between molecules (each molecule was compared with every other molecule, so $m(m-1)/2$ comparisons).

To make the comparisons fairer, we modified qcp to compute the atom coordinate norms and barycenters once per molecule rather than during each comparison, which is the way the routines are set up in the distribution. There was also a problem with qcp's handling of identical files (present in the vhp_mcmd:1vii dataset) for which the RMSD is obviously zero. qcp always performs at least one Newton iteration in its quest to determine the maximum eigenvalue of the $\mathcal{F}$ matrix, which is then used to compute the RMSD. In some cases,

[4] The differences between these options are slight, just the way the frmsd routines are invoked to illustrate different calling sequences, that is, either providing an array of coordinates for all structures at once or the coordinates for two structures at a time.

[5] `-fx` filters the atoms in the .pdb files being compared, where `x` may be one of

c	Cα atoms,
b	backbone atoms (N, Cα, C),
h	heavy atoms (non-H) with no special symmetries,
s	the above along with atoms with special symmetries,
a	all atoms.

roundoff would result in the difference of the initial estimate with the first iterate being negative, leading to the square root of a negative number. We modified qcp to use the absolute value of this difference in the RMSD computation.

Table SI2 presents timings for the various datasets under GNU (gcc 4.4.3) and Intel (icc 12.0.5) compilers on a lightly loaded machine with 8 processors (64-bit 2.4 GHz Xeon) and 24 Gb memory running Ubuntu 10.04.3 LTS. For each dataset, one series of runs computed RMSDs only, while the other also produced rotation matrices.

In Table SI3, RMSD consistency is examined. Let $\mathcal{U}$ be the rotation matrix that produces the RMSD e between the coordinate matrices $\mathcal{X}$ and $\mathcal{Y}$, so that e and $\mathcal{U}$ are related by $e = ||\mathcal{Y} - \mathcal{U}\mathcal{X}||$. Call e_{code} and $\mathcal{U}_{\text{code}}$ the outputs of frmsd or qcp. We compared e_{code} with $e' \equiv ||\mathcal{Y} - \mathcal{U}_{\text{code}}\mathcal{X}||$, that is, we checked how well $\mathcal{U}_{\text{code}}$ reproduced e_{code} for each comparison of atomic coordinates in the indicated datasets. The maximum values of the absolute error, $|e_{\text{code}} - e'|$, and relative error, $|e_{\text{code}} - e'|/e_{\text{code}}$, and the number of instances when one or the other error was $> 10^{-10}$ are shown in the table.

A.2 Operation Counts

Table SI4 compares the basic operation counts for frmsd and qcp. The operation counts for qcp reflect the first modification noted in the timings section (the second was only used for timings on the vhp_mcmd:1vii dataset and would simply add one absolute value to the RMSD calculation). Therefore, any initializations along with the reduction to barycentric coordinates and the computation of the norms of the reduced coordinate matrices is a preprocessing step for both codes, indicated by the terms not multiplied by M in the table. Both algorithms try to recompute the eigenvector of $\mathcal{F}$ associated with the maximal eigenvalue if the norm is too small, but the operation counts presented here are simply the basic (best case) calculations in the situation of no degeneracy in the eigenspace of $\mathcal{F}$.

In Table SI5, additional operation counts for computing the rotation matrix with frmsd and qcp are presented for scenarios where there is degeneracy in the eigenspace of $\mathcal{F}$. qcp tests for the case of a simple leading eigenvalue ($\operatorname{rank} \mathcal{A} = \operatorname{rank}(\mathcal{F} - \lambda_1 I) = 3$) only, while frmsd can deal with all possible deficiencies of $\mathcal{F}$'s eigenspace. The operation counts in each section are for the worst possible case of computing all possible sets of appropriate minors in the quest to find a nonzero set in order to construct a nontrivial eigenvector.

The operation counts for frmsd's linear combinatorial algorithm are given in Table SI6. g is the number of residue symmetry groups and B is half the total number of atoms involved. In general, the term involving A counts the number of operations needed to compute the RMSD given the matrix $\mathcal{R}$, while the term involving B tabulates the work involved in modifying $\mathcal{R}$ for each new permutation of atoms. Note that for the simple case when all the ν_i, the number of atoms in the i^{th} symmetry group, are 2, $B = g$. Also, if $g = 0$, then the operation counts here reduce to the basic numbers given in Table SI4.

Table SI6 is also the operation counts for the recursive algorithm given in Figure 5 that handles general atom symmetries. In this situation, A and B take on different values as noted in the table, but all other counts are otherwise exactly the same. A and B are 1 and 0, respectively, in the limit when $g \to 0$. If all the atoms are indistinguishable, then $g = 1$, $\nu_1 = n$ and so $A = n!$ and $B = (n-1)!$.

A.3 A Small Molecule Test Set

The distribution of frmsd includes a set of small molecules with various types of symmetries together with setup files. The compounds are listed in Table SI7 and their symmetry types are given in Table SI8, reproduced from Coutsias, Lexa et al.[8]. Table SI8 also shows RMSD values found when comparing a dataset of $10,000$ alternative conformations of each compound produced by an inverse kinematics algorithm to that compound's crystal structure. The three values given for each compound on the right are the maximum and minimum RMSD between the target structure and the one structure among the dataset with the best minimal RMSD fit to the target, while the middle value is what a typical comparison would give, without accounting for symmetry. In most cases, accounting for symmetry allowed finding a much better fitting structure in each set than would have otherwise been possible. Figure SI1 shows the compounds in each set with a color scheme identifying identical building blocks in each compound.

References

1. D. L. Theobald, Acta Cryst. **A61**, 478 (2005).

2. P. Liu, D. K. Agrafiotis, and D. L. Theobald, Journal of Computational Chemistry **31**, 1561 (2010).

3. P. Liu, D. K. Agrafiotis, and D. L. Theobald, Journal of Computational Chemistry **32**, 185 (2011), ISSN 1096-987X, URL http://dx.doi.org/10.1002/jcc.21606.

4. E. A. Coutsias, C. Seok, M. P. Jacobson, and K. A. Dill, **25**, 510 (2004).

5. F. Fogolari, S. C. E. Tosatto, and G. Colombo, BMC Bioinformatics **6**, 301 (2005).

6. R. Samudrala and M. Levitt, Protein Science **9**, 1399 (2000), http://dd.compbio.washington.edu/.

7. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, Journal of Molecular Biology **268**, 209 (1997).

8. E. A. Coutsias, K. W. Lexa, M. J. Wester, S. N. Pollock, and M. P. Jacobson, Journal of Chemical Theory and Computation **12**, 4674 (2016).

dataset	molecules	atoms	comparisons
7ring_0.1deg	8736	7	38154480
vhp_mcmd:1vii	6255	295	19559385
lattice_ssfit:1beo	1998[a]	716	1995003

Table SI1 The number of molecules, atoms, and comparisons performed for each dataset.

[a] In the distribution from Decoys 'R' Us, the number of decoys listed was 2000, but one of the files was empty and one was an empty directory.

	frmsd		qcp	
	gcc	icc	gcc	icc
7ring_0.1deg	6.63740	6.29700	8.49800	7.61500
$+U$	9.64740	9.18570	10.93130	10.05750
1vii	39.23080	32.19600	51.30010	46.87990
$+U$	40.64030	34.00150	52.86920	48.43790
1beo[a]	7.18110	5.04520	11.02270	9.10200
$+U$	7.39080	5.25350	11.18120	9.28810

Table SI2 Timings (in seconds) of the two codes frmsd and qcp under gcc and icc (Intel C) with -O2 optimization on the various datasets (similar trends were observed for -O3 optimization). All timings were performed on the same computer and are averages over 100 runs. Standard deviations were no more that 0.05% of the means and are omitted. The calculations are listed in pairs, where in the first entry only RMSDs were computed, while in the second the rotation matrices were also produced ($+U$).

[a] Some of the files in this dataset were identical (the RMSD was identically zero), which caused qcp to crash in some cases, so runs were performed using a version of qcp fixed to avoid this error.

	max (1) relerr (2) abserr		err $> 10^{-10}$	
	frmsd	qcp	frmsd	qcp
7ring_0.1deg	$5.982 \cdot 10^{-9}$	$5.237 \cdot 10^{-9}$	16623	12854
	$3.440 \cdot 10^{-12}$	$2.553 \cdot 10^{-12}$		
1vii	$2.156 \cdot 10^{-13}$	$1.371 \cdot 10^{-13}$	0	0
	$4.583 \cdot 10^{-13}$	$4.228 \cdot 10^{-13}$		
1beo	$1.004 \cdot 10^{-13}$	$6.276 \cdot 10^{-14}$	28	27
	$8.788 \cdot 10^{-7}$	$7.808 \cdot 10^{-7}$		

Table SI3 Maximum relative error (first line of pair) and absolute error (second line of pair) of RMSD consistency and number of instances when the relative or absolute error was $> 10^{-10}$ for the two codes on the datasets. Absolute error is expressed in Å.

	RMSDs (unweighted)	
$+$	$6mn + M(9n + 18 + 5\bar{I})$	$6mn + M(9n + 35 + 5\bar{I})$
$-$	$3mn + M(8 + \bar{I})$	$3mn + M(19 + 2\bar{I})$
$\times$	$3m(n + 1) + M(9n + 38 + 6\bar{I}) + 1$	$3mn + M(9n + 50 + 6\bar{I}) + 1$
$\div$	$M(\bar{I}) + 1$	$3m + M(1 + \bar{I})$
$\sqrt{}$	$M(1)$	$M(1)$
$>$	$M(3\bar{I})$	$M(2 + \bar{I})$
$\lvert\,\rvert$	$M(3\bar{I} + 1)$	$M(2\bar{I} + 1)$
	RMSDs (additional with weights)	
$\times$	$3mn$	$mn + M(3n)$
	rotation matrices (best case)	
$+$	$M(17)$	$M(14)$
$-$	$M(21)$	$M(27)$
$\times$	$M(40)$	$M(44)$
$\div$	$M(2)$	$M(4)$
$\sqrt{}$	$M(1)$	$M(1)$
$>$	$M(4)$	$M(1)$
$\lvert\,\rvert$	$M(4)$	
	frmsd	qcp

$$M \equiv \tfrac{1}{2}(m^2 - m)$$

Table SI4 Operation counts for frmsd and qcp (modified as noted in the text) computing RMSDs and then rotation matrices (the operation counts for the latter are in addition to those needed to compute the RMSD). Here, m molecules, each with n atoms, are compared with each other (so $M = \frac{1}{2}(m - 1)m$ comparisons). $\bar{I}$ is the average number of iterations (≤ 50 for frmsd and qcp) that are needed to compute the maximum eigenvalue of $\mathcal{F}$ in each comparison. For the datasets 7ring_0.1deg, 1vii and 1beo, $\bar{I}$ was approximately 5.428, 8.346 and 9.509 (frmsd), 5.428, 8.346 and 9.510 (qcp), respectively, with a standard deviation of at most 0.01%. The operation counts given are for the best case of no recomputed eigenvectors.

rotation matrices (worst case additional)		
simple leading eigenvalue of $\mathcal{F}$		
$+$	$M(1)$	$M(21)$
$-$	$M(3)$	$M(18)$
$\times$	$M(9)$	$M(60)$
$>$	$M(2)$	$M(3)$
multiple leading eigenvalue of $\mathcal{F}$		
$\times$	$M(1)$	not available
$>$	$M(3)$	
$\lvert\,\rvert$	$M(3)$	
	frmsd	qcp

Table SI5 Additional operation counts for frmsd and qcp computing rotation matrices in the worst case of degeneracy in the eigenspace of $\mathcal{F}$ (see text). Here, m molecules are compared with each other (so $M = \frac{1}{2}(m-1)m$ comparisons). Note that qcp does not handle cases where the leading eigenvalue of $\mathcal{F}$ is double or triple.

RMSDs (combinatorial)		
$+$	$6mn+$	$M(9n + A(18 + 5\bar{I}) + B(9))$
$-$	$3mn+$	$M(\quad A(8 + \bar{I}) \quad + B(18))$
$\times$	$1 + 3m(n+1) +$	$M(9n + A(38 + 6\bar{I}) + B(9))$
$\div$	$1+$	$M \quad A(\bar{I})$
$\sqrt{}$		$M \quad A(1)$
$>$		$M \quad A(3\bar{I})$
$\lvert\,\rvert$		$M \quad A(3\bar{I} + 1)$

	A	B
LRS	$g + 1$	$\frac{1}{2}\sum_{i=1}^{g} \nu_i$
GAS	$\prod_{i=1}^{g} \nu_i!$	$\prod_{i=1}^{g}(\nu_i - 1)!$

Table SI6 Operation counts for computing RMSDs using either the linear combinatoric algorithm to handle residue symmetries (LRS) or the recursive algorithm to handle general atom symmetries (GAS). Here, m molecules, each with n atoms, are compared with each other (so $M = \frac{1}{2}(m-1)m$ comparisons). $\bar{I}$ is the average number of iterations needed to compute the maximum eigenvalue of $\mathcal{F}$ in each comparison. Given g symmetry groups containing $\nu_i, i = 1, \ldots, g$ atoms per group, the values of A and B for the two algorithms are as provided above.

compound	name	database ID (CSDS/PDB) or article source
1P2	cyclic pyrrole-imidazole polyamide	3OMJ
AAS	a-Thr ascidiacyclamide	NEPMIE
ACX	alpha-Cyclodextrin	2XFY
BIX	cryptophane (±)-anti-1	BIMXUR
BPH	biphenyl derivative	PEWTAO
CD4	cryptophane E (±)-2[SbF6]6	RAYFED
CRP	cryptophane E analogue	PICKUH
HAX	sandramycin	HAXMOI10
KCR	hemicryptophane	KASGAO
MST	compound 5a	xtal from Mastalerz et al. Angew Chem Int Ed 2013 v52 p3611
RH9	rhizopodin	2VYP
SWI	swinholide A	1YXQ
TET	compressed tetrahedron	xtal from SI in Wang, Day, and Bowman-James JACS 2013 v135 p392

Table SI7 All compounds used in this study, listed alphabetically by short name, with source and full compound name.

	number of	symmetry	RMSD	RMSD	RMSD
compound	alignments	symmetry	max	given	min
MST	24	tetrahedral	3.04	3.02	0.42
TET	24	tetrahedral	2.19	2.08	1.87
BIX	6	dihedral(3)	2.80	0.75	0.46
CD4	6	dihedral(3)	4.41	0.73	0.68
CRP	6	dihedral(3)	2.59	0.30	0.30
ACX	6	cyclic(6)	0.71	0.65	0.55
KCR	3	cyclic(3)	1.55	0.61	0.57
1P2	2	cyclic(2)	1.39	1.39	1.33
AAS	2	cyclic(2)	0.22	0.22	0.22
BPH	2	cyclic(2)	0.52	0.52	0.48
HAX	2	cyclic(2)	0.57	0.57	0.57
RH9	2	cyclic(2)	1.27	1.24	1.24
SWI	2	cyclic(2)	1.39	1.37	1.37

Table SI8 For each compound in the study, the number of possible alignments and symmetry type, as well as the maximum and minimum RMSDs computed over all the alignments. The given RMSDs are computed ignoring atom symmetries. All RMSDs are given in Å.

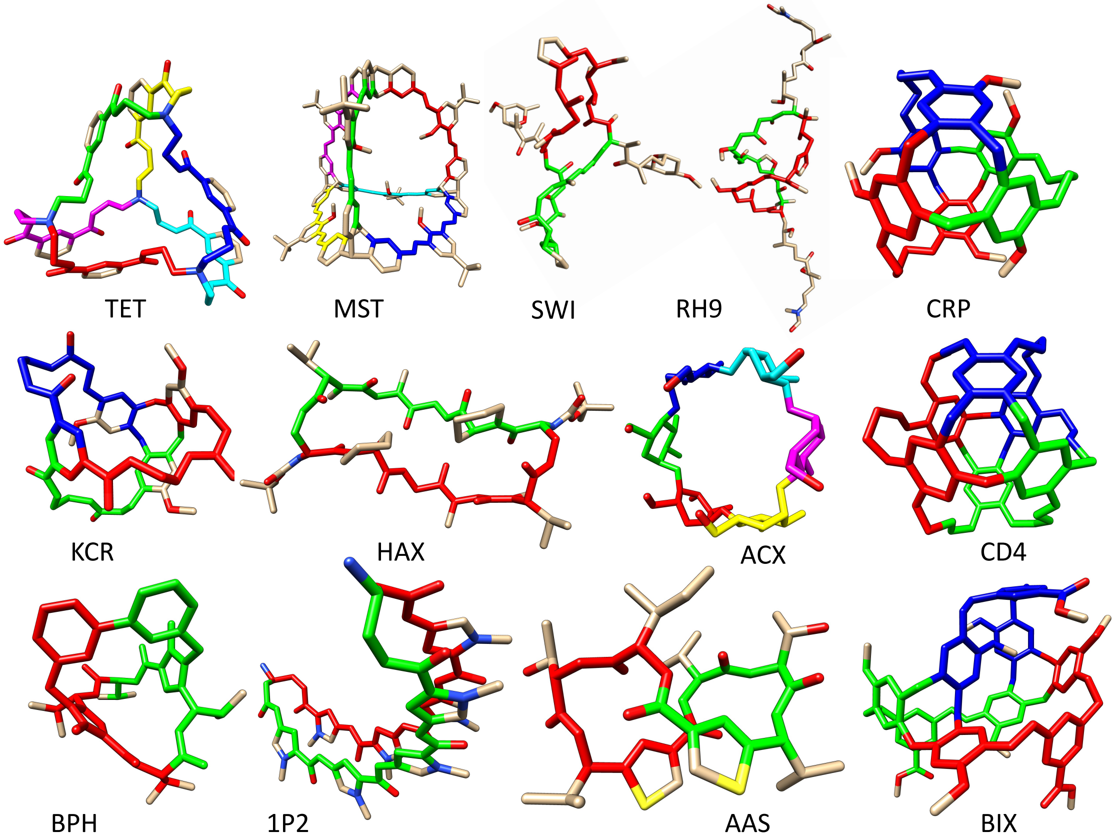

Figure SI1 The symmetric molecules in the dataset. The colored components depict the main backbone atoms included in the RMSD computation from each identical building block. Hydrogens are not shown, while non-backbone atoms not included in RMSD computations are shown in gray.