ArbAlign: A Tool for Optimal Alignment of Arbitrarily Ordered Isomers Using the Kuhn-Munkres Algorithm

Berhane Temelso,[1,4] Joel M. Mabey,[1] Toshiro Kubota,[2] Nana Appiah-Padi,[1,3] George C. Shields[1,4*]*

[1]Dean's Office, College of Arts and Sciences, and Department of Chemistry,

Bucknell University, Lewisburg, PA 17837, USA

[2]Department of Mathematical Sciences,

Susquehanna University, Selinsgrove, PA, 17870, USA

[3]Lewisburg Area High School,

Lewisburg, PA, 17837, USA

[4]Current Address: Provost's Office and Department of Chemistry, Furman University,

Greenville, SC 29613, USA

[*] Corresponding authors: berhane.temelso@furman.edu, george.shields@furman.edu

1. ABSTRACT

When assessing the similarity between two isomers whose atoms are ordered identically, one typically translates and rotates their Cartesian coordinates for best alignment and computes the pairwise root mean square deviation (RMSD). However, if the atoms are ordered differently or the molecular axes are switched, it is necessary to find the best ordering of the atoms and check for optimal axes before calculating a meaningful pairwise RMSD. The factorial scaling of finding the best ordering by looking at all permutations is too expensive for any system with more than ten atoms. We report use of the Kuhn-Munkres matching algorithm to reduce the cost of finding the best ordering from factorial to polynomial scaling. That allows the application of this scheme to any arbitrary system efficiently. Its performance is demonstrated for a range of molecular clusters as well as rigid systems. The largely standalone tool is freely available for download and distribution under the GNU General Public License v3.0 (GNU_GPL_v3) agreement. An online implementation is also provided via a web server (http://www.arbalign.org) for convenient use.

2. INTRODUCTION

Analyzing the similarity between sets of isomers is one of the most common and fundamental tasks in the chemical and biological sciences.[1] All polyatomic molecules have a large number of nuclear degrees of freedom that allow them to form either constitutional isomers or stereoisomers. In the case of constitutional isomers, the structures in question would have different connectivity or bonding, while stereoisomers such as conformers, rotamers, enantiomers and diastereomers would have the same connectivity. Being able to establish whether two isomers are different, and if so, quantifying the extent of the difference, is of paramount importance in many fields like drug design,[2] cheminformatics,[3] and bioinformatics[4].

These comparisons can be performed on the basis of reduced fingerprints like energies, dipole moments, binding affinities, pharmacophoric properties, mathematical descriptions of shapes, molecular electrostatic potentials (MEP), radii of gyration, and rotational constants that are independent of the representation of the molecules in three-dimensional space. These are called non-superimposing approaches. They are often quick to compute and therefore particularly useful for high throughput screenings in problems like drug design.[5] One of the most popular approaches is the ultrafast shape recognition (USR)[6] method which can screen billions of compounds for shape similarity without needing to align them first. Likewise, Shape-it[7] performs alignments of a reference molecule against other molecules based on shape. It is efficient since it overlays structures based on overall shape and assigns a score rather than superimposing them based on their coordinates and calculating RMSDs. Align-It[8] or PHARAO (PHARmacophore Alignment and Optimization) aligns molecules based on their pharmacophores representing certain drug-receptor interactions such as hydrogen bonding, functional groups, hydrophobic interactions, charge transfer, electrostatic and hydrophobic interactions. LigandScout[9] is a

commercial software capable of aligning structures based on pharmacophores for accurate virtual screening in drug design problems, but it is once again not applicable to a general system like the molecular clusters we often study. These molecular similarity methods are summarized in many reviews.[10,11]

In the case of superimposing methods, it is necessary to superimpose the coordinates of the molecules in question optimally before computing a root-mean-square distance (RMSD) that serves as a quantitative measure of the similarity between the structures. Because the molecular coordinates used during the superimposition differ based on the choice of axes, origin, ordering of atoms, and presence of chiral centers, the resulting RMSD is dependent on the representation of the molecular coordinates. Therefore, it is necessary to perform operations that ensure permutational invariance[12] as well as chirality.[13] Nevertheless, superimposing methods are most often used as measures of structural similarity because they provide a physically meaningful overlay of the structures in question and a useful similarity metric in the form of RMSD values. For the RMSDs to be truly meaningful, however, one needs to ensure that the atoms in the isomers are ordered identically, and the isomers are translated, rotated, and reflected for optimal superimposition.

Many superimposing methods have been developed and reported in the literature.[14-24] In these methods, researchers typically compare structures of the same size by taking their Cartesian coordinates and calculating the RMSD between them. Assuming the two isomers (A and B) have the same size (n = number of atoms), composition (C, H, O, …) and ordering, the RMSD between them should be zero if they are identical. If they are different, their RMSD should be a non-zero number, with larger RMSDs generally reflecting more significant differences. In general, these methods first translate the center of mass (COM) of the two structures to the origin

and rotate the second structure to optimally superimpose over the reference using the Kabsch algorithm[25] or the faster quaternions.[26] After that, the RMSD between the two structures, A and B, is calculated as the root mean square difference of their respective x, y and z coordinates:

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(x_{iA}-x_{iB})^2+(y_{iA}-y_{iB})^2+(z_{iA}-z_{iB})^2\right]} \qquad (1)$$

The above approach would be sufficient if (i) the atoms are ordered identically in both structures, (ii) their x-, y- and z-axis are defined optimally and (iii) the isomers are not related by reflections across multiple planes. In most cases, there are no guarantees that the above conditions are satisfied. For example, **Figure 1** shows two sets of Cartesian coordinates of the water dimer global minimum. Ideally, matching these two structures using conventional RMSD calculations would give a value of zero indicating that they are the same. However, a typical RMSD calculation on that system yields an RMSD of 2.003 Å. To obtain an RMSD that truly reflects the similarity between the two sets of coordinates, one would first need to re-order the atoms, swap the axes and reflect the coordinates across different planes as detailed in **Figure 1**. To take care of the atom ordering problem that makes meaningful RMSD calculations difficult, a naive approach would entail permuting atoms of the same atomic automorphism class in the first structure with every like atom of the same atomic automorphism class in the second structure, calculating the total RMSD, and repeating the process for all present automorphism classes. For a system with *N* atoms in its largest automorphism class, a naïve approach would require *N!* permutations and RMSD calculations, resulting in an exponential cost that is prohibitively expensive for molecules of appreciable size. Although this brute force approach has a steep scaling, many algorithms exist to reduce the scaling to a more manageable polynomial time.[27,28]

There are some tools in the literature that attempt to perform similarity analysis using structural superposition, but none of them are truly general purpose, standalone, free and easy to use. Open3Dalign[29] has the capability to do unsupervised molecular alignment on a large set of structures based on structure or pharmacophores. However, it is mainly intended for large drug design problems and its application to a pair of general structures is non-trivial. It also has a large number of dependencies that make its installation and use challenging. Allen and Rizzo have implemented a symmetry corrected RMSD calculation method including a Hungarian assignment for optimal atom correspondence in Dock6.[30] While their implementation is very efficient and similar to ours, it only performs heavy-atoms alignments and some of its RMSD values are much larger than our benchmark numbers, suggesting that its alignment is not always optimal. Vasquez-Perez[21] and Marquez[22] have devised a method that performs a series of random rotations followed by Hungarian assignments until the RMSD converges sufficiently. Their method consistently yields low RMSDs, however it is limited in two ways. First, the code only provides the optimal RMSD of the pairs of structures, but not the coordinates of the best-aligned structure. Secondly, its cost scales very unpredictably with the system size perhaps because of the large number of rotations needed or Hungarian algorithm. Helmich and Sierka have reported a similar approach combining the Hungarian algorithm with graph theory to achieve a quadratic scaling structural overlapping method.[31] Automatic RMSD (aRMSD) is a tool recently developed by Wagner and Himmel to employ the Hungarian assignment and RMSD calculations to compare experimentally and theoretically determined molecular structures.[32]

As noted above, most of the tools are not truly general purpose, standalone, free, easy to use or they either have a lot of dependencies or are part of a bigger software package. Most of them are intended for ligand-receptor docking in structure-based drug design problems and they cannot be

applied to any arbitrary molecular system. The current tool, ArbAlign, is unique for a few reasons. First, it is virtually a standalone Python script whose only non-standard dependence is a Hungarian assignment module that can be installed following a few simple instructions. Secondly, it can perform alignments on different automorphism classes based on atom label, SYBYL atom type or atom connectivity. Unlike other tools which only perform heavy-atom alignments, it has the option of performing all atom or heavy atom only alignments. Users also have the option of applying the Hungarian algorithm in the original coordinate system or 48 coordinate swaps and reflections thereof. Third, all the method's functionalities are available through a web server, making it very convenient for any user. In short, this work describes an algorithm to find meaningful RMSDs for arbitrarily-ordered molecules of the same composition, i.e. isomers, efficiently using the Kuhn-Munkres[33,34] algorithm for assignment of atoms in their principal coordinate system. It also explores the merits of assessing similarity on the basis of different automorphism classes with and without chemical information like connectivity and hybridization.

3. COMPUTATIONAL DETAILS

3.1. ALGORITHM

The only requirement for this method to work is that the structures in consideration have the same chemical composition. If they share the same connectivity, they would be stereoisomers such as conformers, rotamers, enantiomers and diastereomers. If they have different connectivity, they would be constitutional or structural isomers. The algorithm employed in this work is summarized in **Figure 2** and described in detail below.

1. The Cartesian coordinates of isomers 1 and 2 are read in. The atoms of both isomer 1 and 2 are then ordered alphabetically by atom name, type or connectivity.
2. The coordinates are converted to their respective principal coordinate axis system. First, their coordinates are moved such that their respective center of mass (COM) is at the origin. Then, their moment of inertia (MOI) tensor is calculated and diagonalized. The principal coordinates are found by taking the dot product of the COM coordinates with the eigenvectors of the MOI tensor. This essentially rotates the arbitrary coordinate system to a more intrinsic principal coordinate system. This transformation is necessary because two isomers that have similar shapes would also have comparable principal coordinates. Therefore, it is a good first step towards aligning the isomers.
3. At this point, we can use different criteria to partition each isomer to different atomic automorphism classes. One can take three different approaches depending on whether one wants to account for the atoms' bonding and connectivity. Here, it is helpful to clarify the different forms of isomerism. Two isomers with the same connectivity are stereoisomers whereas those with different connectivity or bonding are constitutional or structural isomers. Aligning isomers by atom label, type or connectivity should give meaningful final RMSDs for stereoisomers. However, if one is dealing with constitutional isomers, aligning isomers by atom type or connectivity will give a more meaningful RMSD because it factors in the connectivity of the atoms. Depending on which approach one takes, the resulting alignments can be different. Some alignments can also be more physically meaningful than others even if they do not yield the lowest RMSD. The three approaches are demonstrated for sulfuric acid in Table 1 and described below.

3.1. The most general case is to match the atoms of the same name. For example, sulfuric acid's atom labels would be S, O, O, O, O, H, and H.

3.2. The atom type can include other information including bonding and local environments as defined in the Tripos SYBYL Mol2 file format implemented in OpenBabel 2.3.1.[35] For example, sulfuric acid's atom types would be S.O2, O.2, O.2, O.3, O.3, H, and H, where S.O2, O.2 and O.3 are defined as sulphone sulfur, *sp2* oxygen and *sp3* oxygen, respectively.

3.3. The atoms can also include connectivity information based on the multilevel neighborhoods of atoms (MNA)[36] file format implemented in OpenBabel 2.3.1.[35] In the current case, we only consider the connectivity to the nearest neighbor (level 1). Accordingly, sulfuric acid's connectivity types are -S(-O-O-O-O), -O(-S), -O(-S), O(-H-S), O(-H-S), -H(-O), and -H(-O).

4. When one switches to a principal coordinate system or works with an isomer with high level of symmetry, it is often necessary to consider coordinate swaps and reflections to obtain optimal reordering and alignments.

4.1. Swaps correspond to the switching of (x,y,z) axis to ([x,y,z], [x,y,z], [x,y,z]). There are six such possible swaps.

4.2. Reflections take (x,y,z) coordinates to (±x, ±y, ±z). There are eight possible reflections.

4.3. We generate coordinates of structure 2 corresponding to all forty-eight combinations of swaps and reflections. Some of these combinations are redundant, but we perform all of them for the sake of simplicity.

5. For each one of the forty-eight coordinates of structure 2, we employ the Kuhn-Munkres algorithm[33,34] on each atom name, type or connectivity to obtain the optimal alignment as prescribed below.

   5.1. For each atom name, type, or connectivity from Step 3,

      5.1.1. Generate a distance matrix, a_{ij}, between atom *i* of structure 1 and atom *j* of structure 2 of the same name, type or connectivity. This would be equivalent to what is referred to as the cost matrix in most Kuhn-Munkres assignment problems.

      5.1.2. Apply the Kuhn-Munkres[33,34] algorithm to obtain the optimal assignment of a given atom name, type, or connectivity.

      5.1.3. Apply the optimal assignment on the given atom name or type of structure 2.

      5.1.4. Calculate the all-atom RMSD of structure 1 with the newly reordered structure 2 and save the RMSD value and optimal assignment. Kromann et al. [37] originally wrote the Python implementation of the Kabsch[25] algorithm used here to calculate RMSDs.

   5.2. Move on to the next atom name, type or connectivity and repeat Step 5.1

6. Once we have compiled the best assignments for each atom name, type or connectivity of the forty-eight swapped and reflected structures of structure 2, we sort the RMSDs and declare the one with the lowest RMSD as having the best alignment and ordering.

A Python implementation of this approach is included in the Supporting Information and a web version is available at http://www.arbalign.org. This implementation requires the following scripts and tools.

1. A Python script to convert isomers from arbitrary to principal coordinate system is included in the Supporting Information.
2. In cases where one wants to use atom types containing hybridization information or atom connectivity information, it is necessary to use OpenBabel[35] to convert the Cartesian coordinates to SYBYL Mol2 (sy2) and MNA (mna) formats, respectively. Shell scripts to convert between the different formats and perform pre- and post-processing of the output are also included in the Supporting Information.
3. A wrapper to a fast C++ implementation (`hungarian`)[38] of the Kuhn-Munkres algorithm in Python is needed to solve the assignment problem. It can be installed in three easy steps. They both require the `Numpy` module[39] on top of other standard packages in Python.

4. USAGE

4.1. Command Line Tool

ArbAlign can be used either as a command line or web tool. The command line tool has a driver script that can take in many options or resort to sensible defaults when necessary.

Usage:
```
ArbAlign-driver.py
          [-b/--by {l, t, c}]
          [-n/--noHydrogens]
          [-s/--simple]
          filename_1.xyz filename_2.xyz
```

`-b {l,t,c},` `--by {l,t,c}`	Match atoms by l-label, SYBYL t-type, or NMA connectivity (-c). The default is by atom label (-l)
`-s,` `--simple`	Perform Kuhn-Munkres assignment reordering without axes swaps and reflections. The default is to perform axes swaps and reflections
`-n,` `--noHydrogens`	Ignore hydrogens. The default is to include all atoms

If the pairs of structures pass a sanity test, the tool will align them optimally and provide the following information.

- The initial Kabsch RMSD,
- The Kuhn-Munkres reorderings for each atom and the corresponding RMSDs,
- The final Kabsch RMSD after the application of the Kuhn-Munkres algorithm, and
- The coordinates corresponding to the best alignment of the second structure with the first.

4.2. Web Tool

As shown in Figure 3, the web interface provides all the functionalities of the command line tool via an installation at http://www.arbalign.org. Once a user uploads the Cartesian coordinates to compare and submits them for alignment, the web server will output the same details as the command line tool. It additionally provides a Chimera[40] figure and session file of the initial and final overlays. It also uses Chimera to generate an interactive 3D rendering of the molecules before and after alignment for users with WebGL-enabled browsers to examine.

The source code and all the examples discussed about are included in the Supporting Information for users who want to explore, use or expand on this work.

5. RESULTS and DISCUSSION

The two most important properties for evaluating a similarity analysis tool are its efficiency and accuracy. Because a lot of similarity analysis tools are used in batch mode in high throughput environments, the tool needs to be efficient. In terms of accuracy, one needs to ensure that the tool can yield a reasonable measure of the similarity between structures and also work universally for any chemical system. Both the efficiency and accuracy of the current tool are discussed below.

A brute force approach to the atom assignment problem is very limited because it scales exponentially with the system size. Therefore, our current approach needs to have substantially lower scaling to be useful. In general, the most expensive step in our method is the atom re-ordering using the Kuhn-Munkres algorithm. This step typically scales as $O(N^3)$ where N is the dimension of the cost matrix to be solved for making an assignment. We use a wrapper to a fast C++ implementation (`hungarian`)[38] of the Kuhn-Munkres algorithm instead of a Python[41] implementation because of its speed, as demonstrated for in Section S2 of the Supporting Information. N is not the total number of atoms in a molecule, but rather the size of a given atomic automorphism class such as the atom name, type or connectivity. The leading term in the cost of this algorithm is the size of most common atomic automorphism class in the isomers being aligned. For example, when aligning two water decamer, $(H_2O)_{10}$, structures, the most expensive step would be optimally assigning the 20 hydrogen atoms by solving a 20x20 distance

matrix using the Kuhn-Munkres algorithm. Next, we would assign the 10 oxygen atoms by solving the 10x10 distance matrix. And we need to perform these operations forty-eight times for every atom name, type or connectivity. In the end, we would have done 96 Kuhn-Munkres assignments and compiled 96 RMSD values that would be used to determine the best assignments

For a set of water clusters,[42,43] Lennard-Jones clusters,[44-47] atmospheric hydrates of sulfuric acid and alklyamines,[48] and peptides,[49,50] pairs of arbitrarily ordered structures were compared using a conventional RMSD calculations as well as one employing our method. Those applications are discussed below.

5.1. Application to Water Clusters

The comparison for water clusters of size n = 2 - 100 is shown in **Figure 4**. In every case, the conventional RMSD value was much higher than the Kuhn-Munkres one, indicating that orderings and axis transformations were necessary to obtain an optimal RMSD. These clusters' coordinates were taken from published literature[42-44] and compared for similarity. The monomers in these water clusters are often interchangeable and it is necessary to permute the monomers, or the oxygen and hydrogen atoms, to optimally align the cluster before computing an RMSD. The comparison of 10-PP1 and 10-PP2 pentagonal prism water decamers in **Figure 5** clearly demonstrates this point. Although these two clusters have the same oxygen skeleton, their hydrogen coordinates and hydrogen bonds are arranged differently. Performing a conventional overlay would lead to a large RMSD (3.266 Å) and an exaggerated misalignment between the two structures. However, applying our approach yields a much lower RMSD (0.338 Å) and an alignment that clearly illustrates the difference between the two structures. Structures 10-PP1 and 10-PP2 only differ in the orientation of the hydrogen atoms and cyclic hydrogen bonds in the bottom pentagonal plane.

5.2. Application to Noble Gas Clusters

Noble gas clusters are held together by weak dispersion forces that allow them to adopt an immense number of configurations. Comparing these clusters rigorously requires reordering of the atoms because interchanging the atoms can lead to better alignments and lower RMSD values. We have employed our algorithm to pairs of neon clusters, $(Ne)_n$, of size n = 10 – 1000 from the Cambridge Cluster Database.[44-47] For every cluster size, we calculated the RMSD between two isomers using conventional and Kuhn-Munkres alignment methods. As shown in **Figure 6**, our approach yields a lower RMSD than a conventional alignment. For $(Ne)_{50}$, **Figure**

7 illustrates the importance of optimally reordering the atoms to obtain a good measure of cluster similarity.

5.3. Application to Atmospheric Hydrates

In the case of water and noble gas clusters, it was sufficient to perform Kuhn-Munkres assignments based only on atom names, because each unique atom name has identical hybridization and connectivity. All oxygen atoms have the same hybridization and connectivity and the same is true for all hydrogen and neon atoms. However, when you have constitutional isomers where atoms of the same name can have different hybridization or connectivity, it becomes necessary to factor in that information to make a physically meaningful alignment. That is precisely the case for atmospheric sulfate hydrates and we examine two such cases here. The first is a complex composed of one bisulfate, one methylammonium and one water molecule. The challenge here is that the bisulfate ion has two types of oxygen atoms just like the sulfuric acid molecule described in Table 1. **Figure 8** shows RMSDs of the global minimum structure with nineteen higher-lying isomers as computed using a conventional alignment as well as three variations of our Kuhn-Munkres (KM) alignment. As expected, all the KM alignments yield a lower RMSD than the conventional one. The more interesting result is the difference in RMSDs between KM alignments based on atom name, type or connectivity. Alignments based on atom name give the lowest RMSD, and those based on atom type and connectivity yield slightly higher RMSDs. These differences are most stark for the alignment of the global minimum with isomer 19 in **Figure 8** and we explore these differences in detail in **Figure 9**. KM alignment based on atom name clearly gives the lowest RMSD (0.767 Å), but this results from assigning an *sp2* oxygen on one structure to an *sp3* oxygen on another. Technically, using atom types as defined in the SYBYL Mol2 format should distinguish between *sp2* (O.2) and *sp3* (O.3) oxygen

atoms, but the format runs into trouble when handling charged monomers like the bisulfate ion. In this case, it identifies the atoms in the bisulfate as having two *sp2* (O.2) and two *sp3* (O.3) oxygens instead of three *sp2* (O.2) and one *sp3* (O.3) oxygens. This misidentification leads to an erroneous assignment and a larger RMSD of 1.166 Å. The most appropriate assignment is the one based on connectivity. Here, the oxygens' connectivity is identified as -O(-S) and -O(-H-S) type, and they are assigned properly to oxygens of the same connectivity. The RMSD of 0.983 Å lies between those based on atom name and type, but it is the most physically meaningful metric of the similarity between these two clusters. The need to use chemical information like atom type and connectivity is even more apparent for studying more complex systems such as peptides.

5.4. Application to Small Peptides

The four alignment approaches were also tested on a set of five di- and tri-peptides containing one aromatic side chain.[49,50] These peptides (FGG, GFA, GGF, WG, and WGG) contain the residues phenylalanine (F), glycine (G), tryptophan (W) and alanine (A), of which F and W have aromatic side-chains. For each peptide, we calculated the RMSD between the global minimum structure and fourteen higher energy isomers using conventional and Kuhn-Munkres methods. The comparison for FGG is shown in **Figure 10** and analogous results for the other four peptides are included in the Supporting Information. If one uses atom names to do the alignment, there are four atoms, namely H, C, O and N. However, if one uses SYBYL Mol2 atom types, there are eight types: H, C.AR, C.3, C.2, O.2, N.AM, O.3, N.3. Likewise, employing connectivity yields fourteen types that can be used as a basis for assignments. Despite these large differences in groups used for alignment, the RMSDs from the different KM approaches are fairly similar. We demonstrate some of these differences for the case of aligning the global minimum with isomer 7 in **Figure 11**. The RMSDs based on conventional, and Kuhn-Munkres approaches, using atom

name, type and connectivity are 2.676, 0.746, 0.746 and 0.964 Å, respectively. Although the alignments based on atom type and connectivity look identical, the one based on atom type matches H4, the hydrogen bonded to a carbon in the first structure, to a hydrogen bonded to a nitrogen in the second structure. Fortunately, exploiting connectivity information distinguishes between different types of hydrogens and yields a more meaningful alignment. All these examples demonstrate that exploring how similar or different structures are is non-trivial and that superimposing methods for determining similarity need to be checked carefully.

6. CONCLUSIONS

We have developed a tool that reorders, swaps, and reflects any two isomers to find their optimal alignment. The reordering is performed using the Kuhn-Munkres algorithm on every atom of the same name, type, or connectivity. The resulting alignment consistently yields a lower RMSD than conventional alignments that only employ the Kabsch method to translate and rotate isomers for the best alignment. We have applied this method to a large number of systems ranging from homonuclear noble gas clusters to peptides. Its implementation is efficient enough to be used on large molecules or screen through a large number of small molecules. The tool can be used from the command line or via a web server at http://www.arbalign.org.

Author Information

Corresponding Authors

*E-mail: berhane.temelso@furman.edu, george.shields@furman.edu,

Notes:

The authors declare no competing financial interests

Acknowledgements

The authors thank Frank Pickard and Andy Simmonett at NIH for productive discussions about this work. BT thanks JLF for the joyful inspiration. Acknowledgment is made to the NSF and Bucknell University and Furman University for their support of this work. This project was supported in part by NSF grants CHE-1213521 and CHE-1508556, and by NSF grant CHE-1229354 as part of the MERCURY consortium (http://www.mercuryconsortium.org).

Associated Content

Supporting Information. Python implementation of the Kuhn-Munkres method, a Python driver script, scripts to calculate principal coordinates, convert molecules to formats containing atom type and connectivity information, calculate Kabsch RMSDs, RMSDs and timing information for water clusters, hydrates containing sulfuric acid, monomethylamine, homonuclear noble gas clusters, and a tripeptide. This information is available free of charge via the Internet at http://pubs.acs.org/.

Table 1. Three Representations of Sulfuric Acid's Coordinates Factoring in the Atom Name (a), Atom Type (b) and Connectivity (c).

(a) By Atom Name				(b) By SYBYL Atom Type				(c) By NMA Atom Connectivity Type			
S	x1	y1	z1	S.O2	x1	y1	z1	-S(-O-O-O-O)	x1	y1	z1
O	x2	y2	z2	O.2	x2	y2	z2	-O(-S)	x2	y2	z2
O	x3	y3	z3	O.2	x3	y3	z3	-O(-S)	x3	y3	z3
O	x4	y4	z4	O.3	x4	y4	z4	-O(-H-S)	x4	y4	z4
O	x5	y5	z5	O.3	x5	y5	z5	-O(-H-S)	x5	y5	z5
H	x6	y6	z6	H	x6	y6	z6	-H(-O)	x6	y6	z6
H	x7	y7	z7	H	x7	y7	z7	-H(-O)	x7	y7	z7

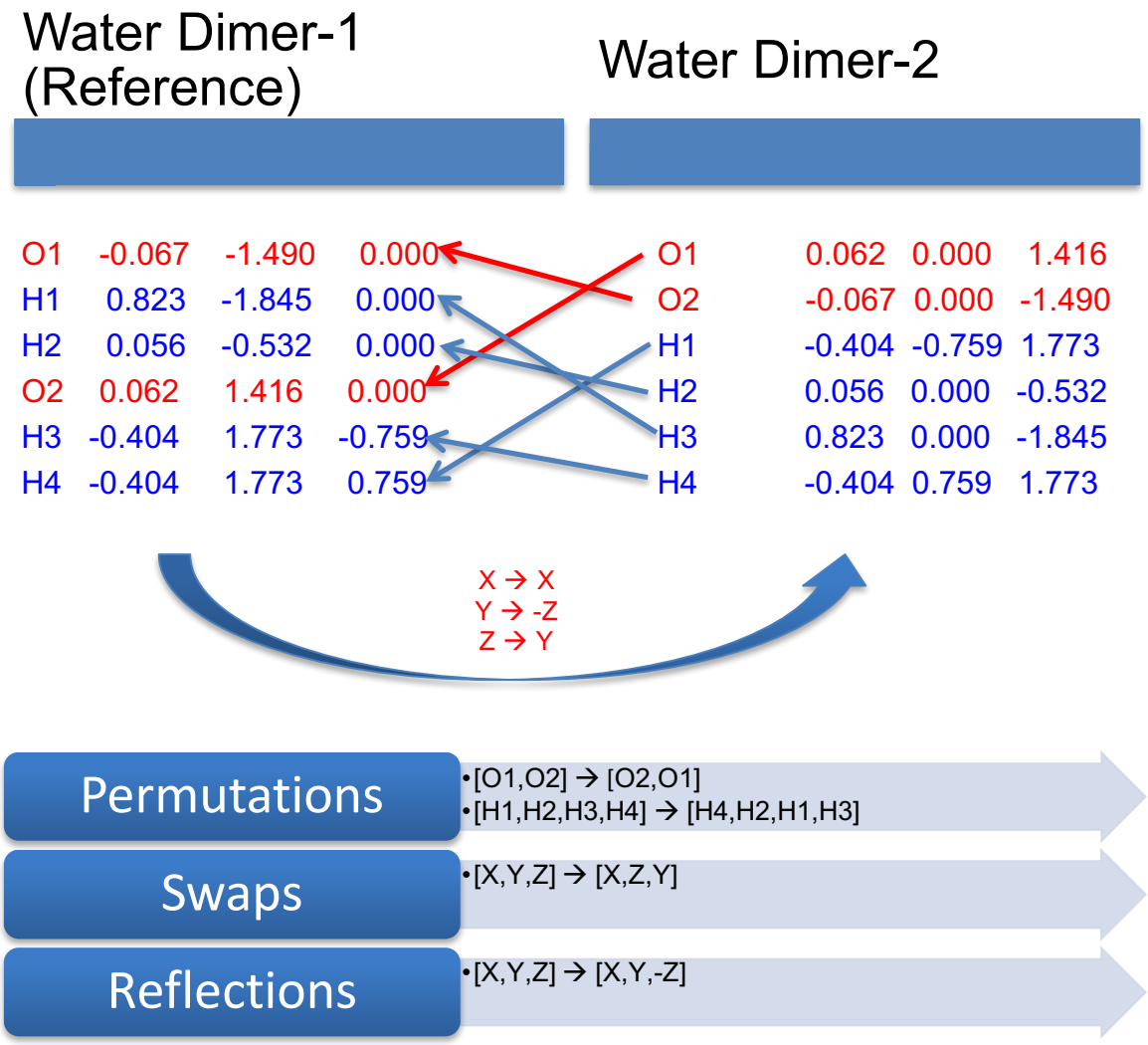


Figure 1. Cartesian coordinates of two identical water dimers and the necessary transformations to yield their optimal alignment. A conventional RMSD calculation yields a large value while performing the above transformations yields the true RMSD of 0.000 Å.

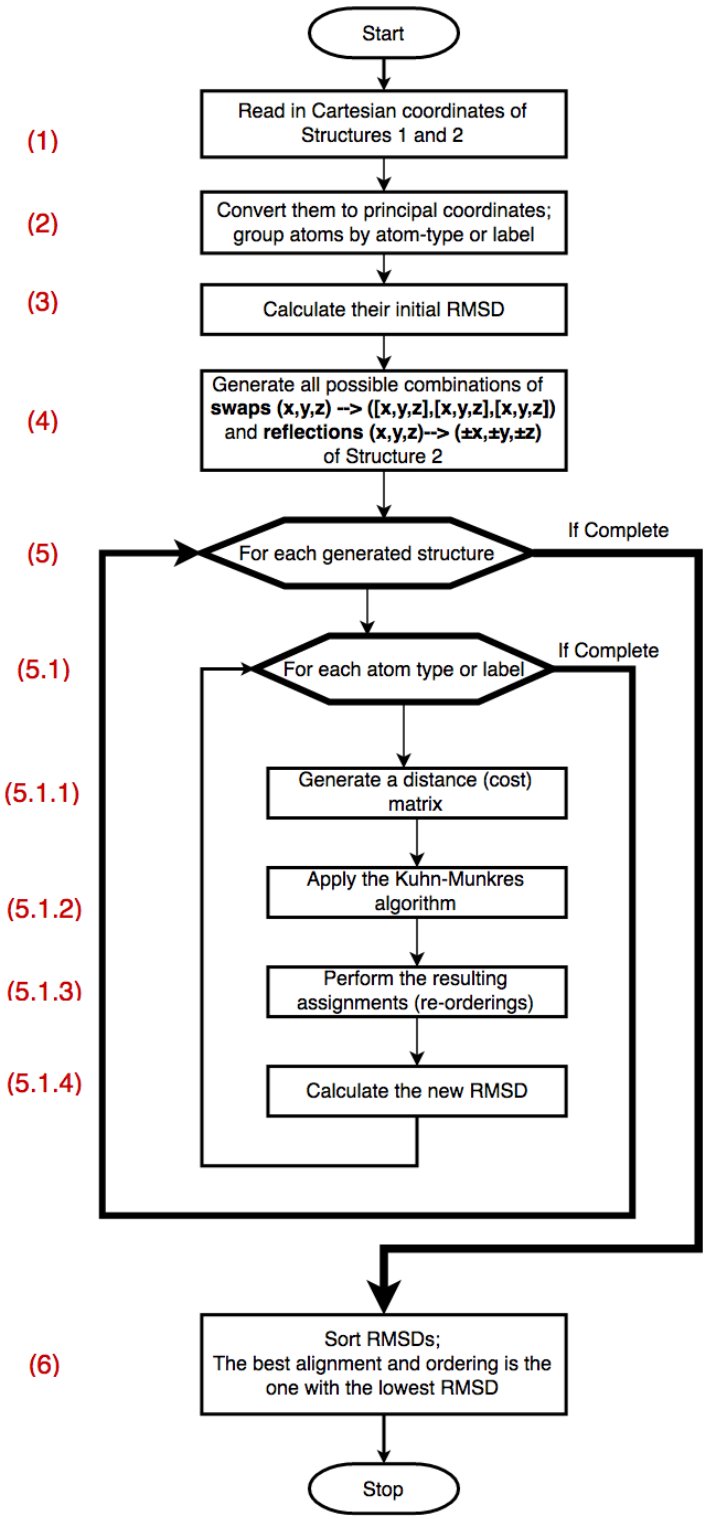


Figure 2. A schema of the algorithm used in this work.

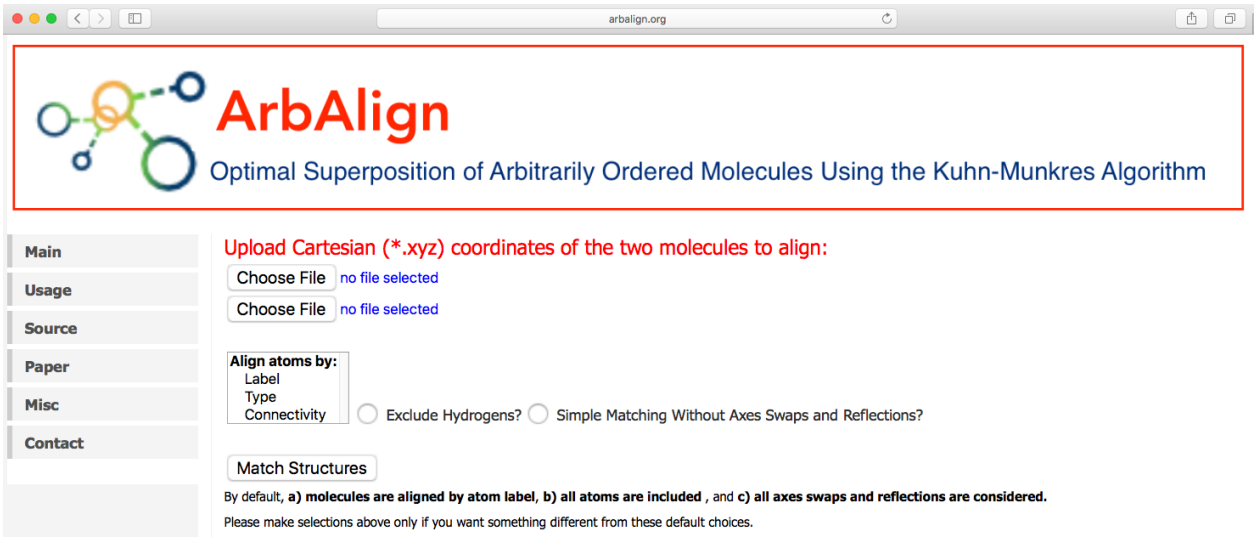


Figure 3. The web interface to ArbAlign at http://www.arbalign.org. It has all the functionalities of the command line tool.

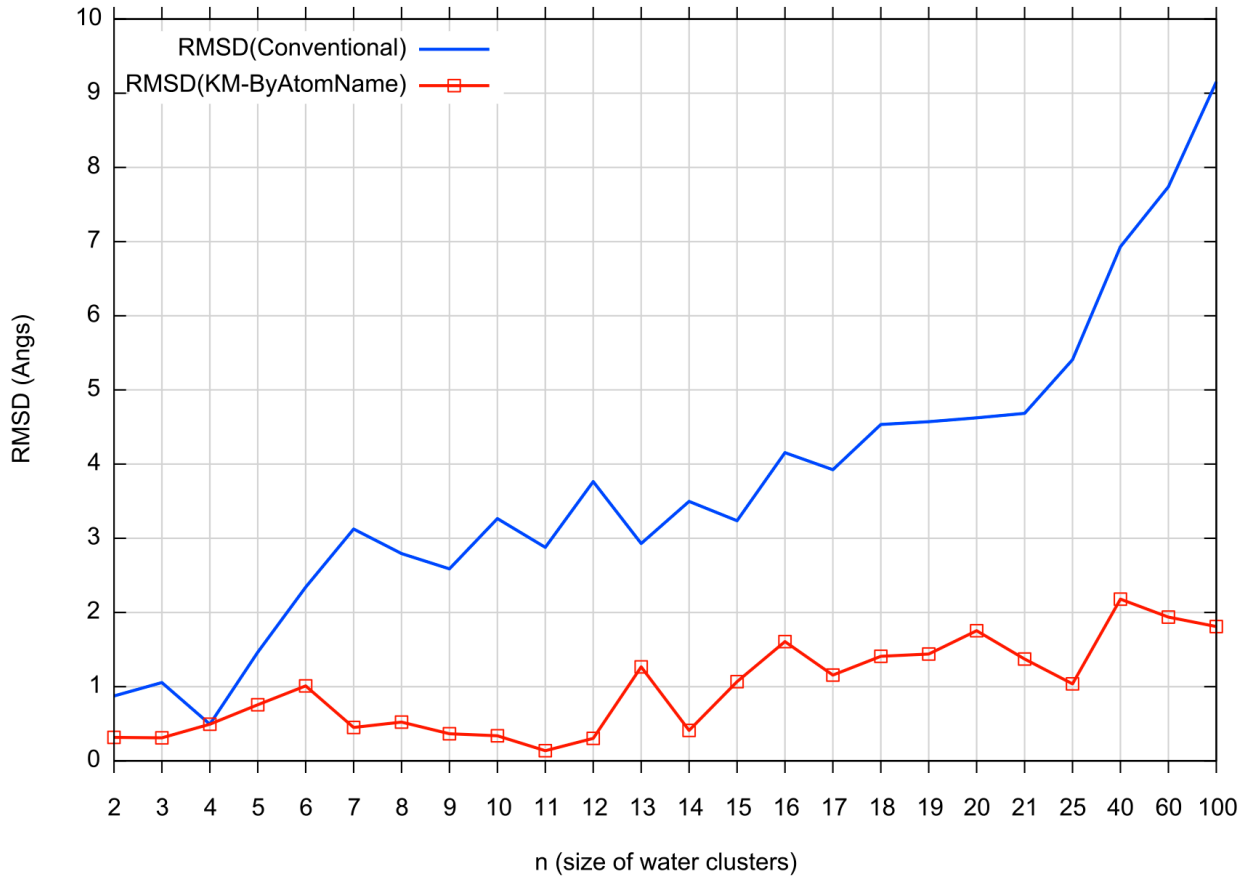


Figure 4. A comparison of conventional and Kuhn-Munkres RMSD values for pairs of water clusters $(H_2O)_n$, n=10-100 from ref. 42-44. In every case, the Kuhn-Munkres method yields a lower RMSD value.

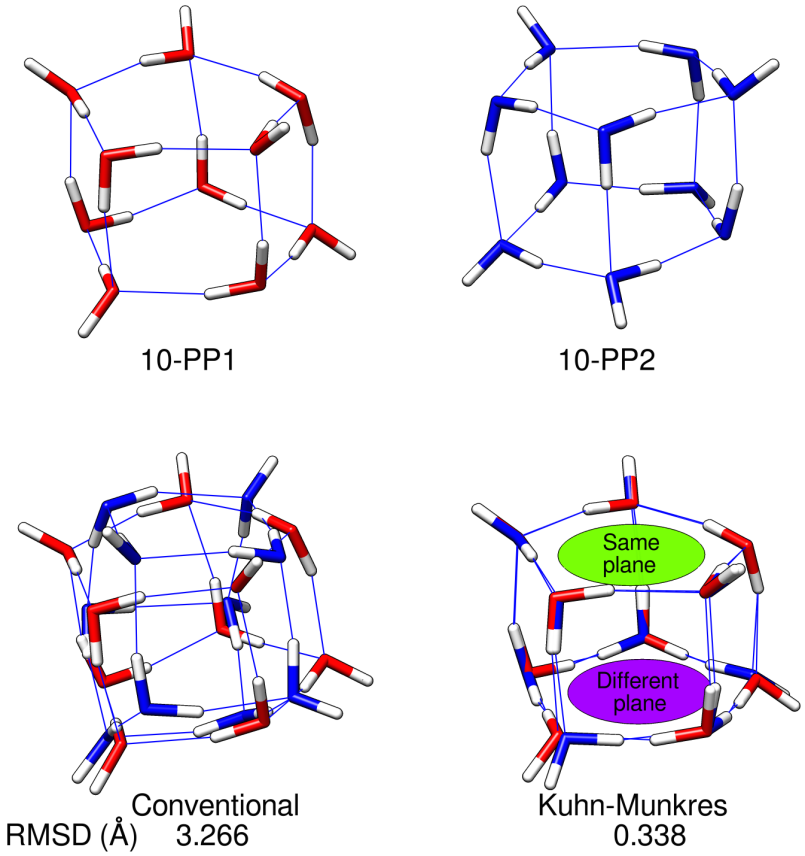


Figure 5. A comparison of conventional and Kuhn-Munkres RMSD matching for 10-PP1 and 10-PP2 water decamers, $(H_2O)_{10}$. While a conventional alignment suggests that the clusters are very different, a Kuhn-Munkres alignment shows that the two decamers actually share an identical pentagonal plane and that the second pentagonal planes only differ in the direction of the hydrogen bonds.

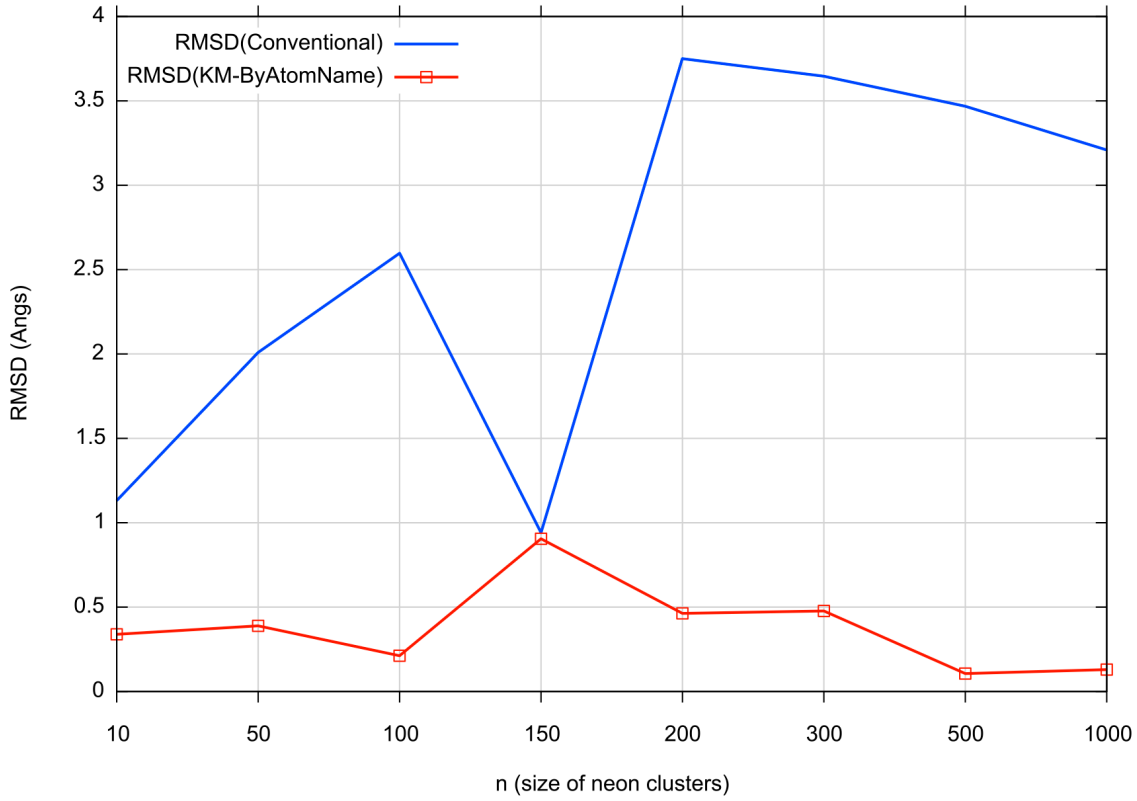


Figure 6. A comparison of conventional and Kuhn-Munkres RMSD values for pairs of neon clusters, $(Ne)_n$, n=10-1000 from the Cambridge Cluster Database (ref. 44-47).

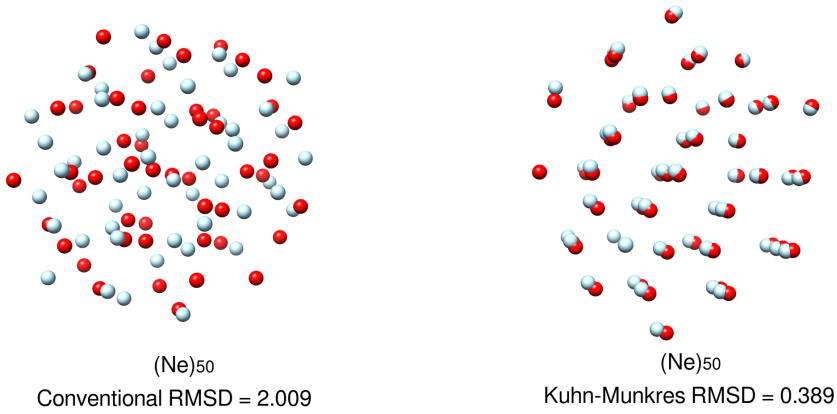


Figure 7. A comparison of conventional and Kuhn-Munkres RMSD matching for a pair of $(Ne)_{50}$ clusters. The Kuhn-Munkres alignment clearly reveals the similarity between the two structures that are not apparent from a conventional alignment.

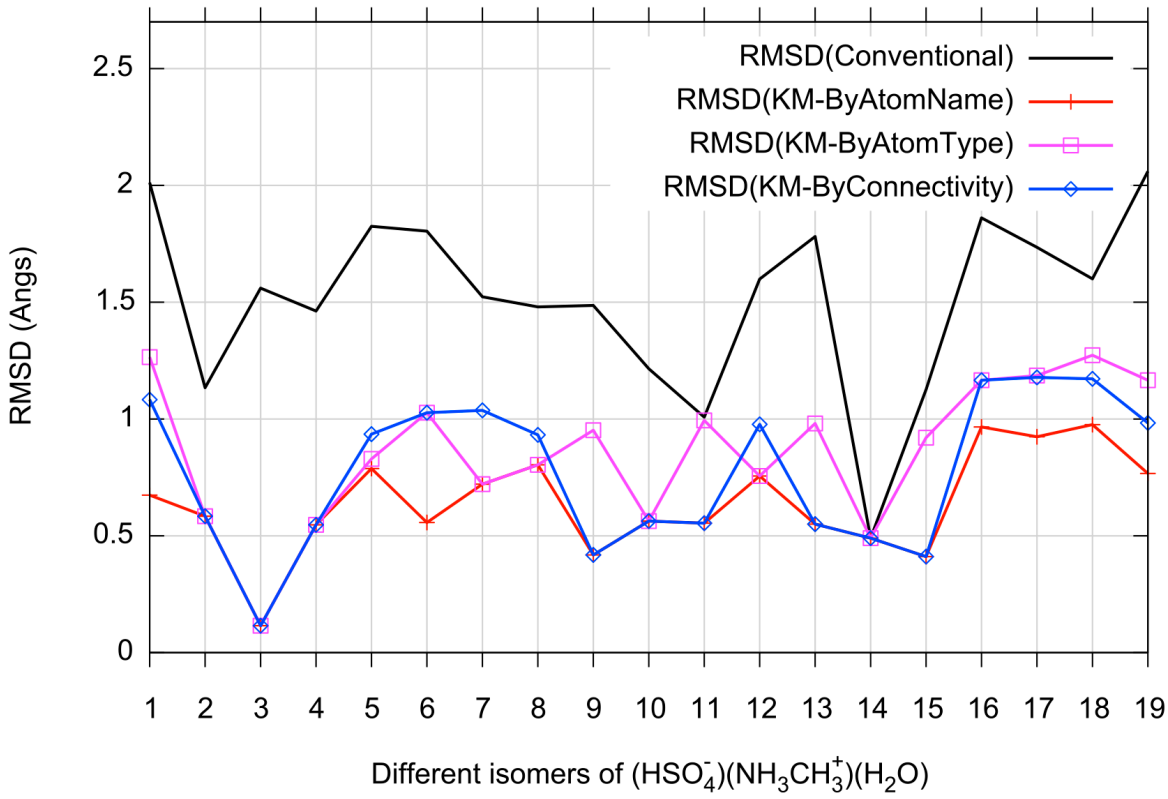


Figure 8. A comparison of conventional and Kuhn-Munkres RMSD matching for isomers of a complex containing one bisulfate, one methylammonium and one water molecule from ref 48. The RMSD is between the most stable isomer and the next nineteen isomers. Assignments based on atom name always give the lowest RMSD, but using atom type or connectivity information often yields a more physically meaningful alignment.

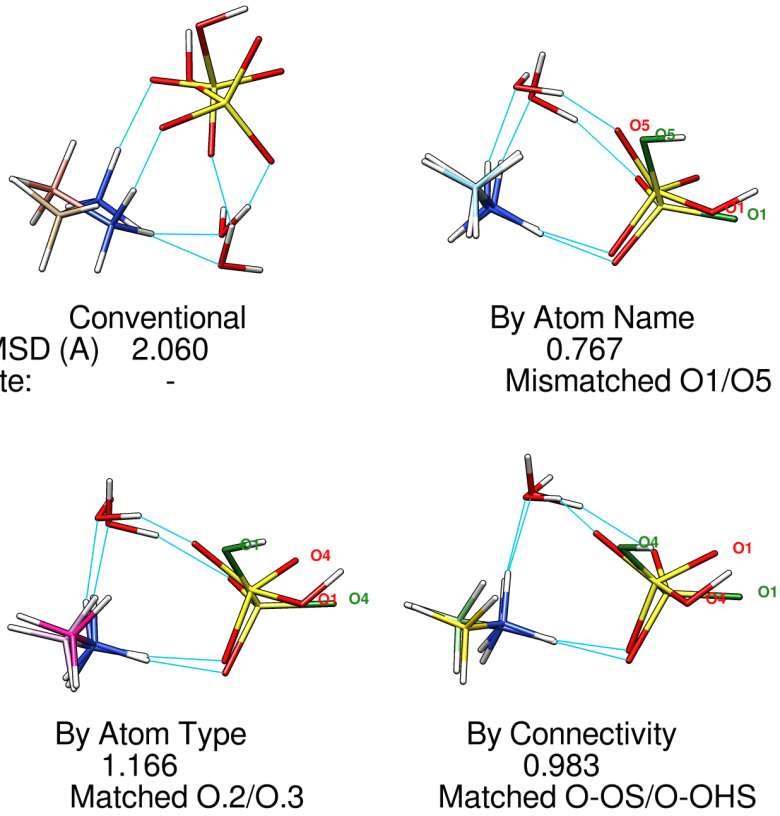


Figure 9. A comparison of conventional and Kuhn-Munkres RMSD matching for a pair of isomers of a complex containing one bisulfate, one methylammonium and one water molecule. Kuhn-Munkres assignments based on atom names alone yield the lowest RMSDs, but wrongly assign *sp3* oxygens to *sp2* oxygens and vice versa. Using atom type and connectivity yields a higher RMSD, but a more physically meaningful assignment.

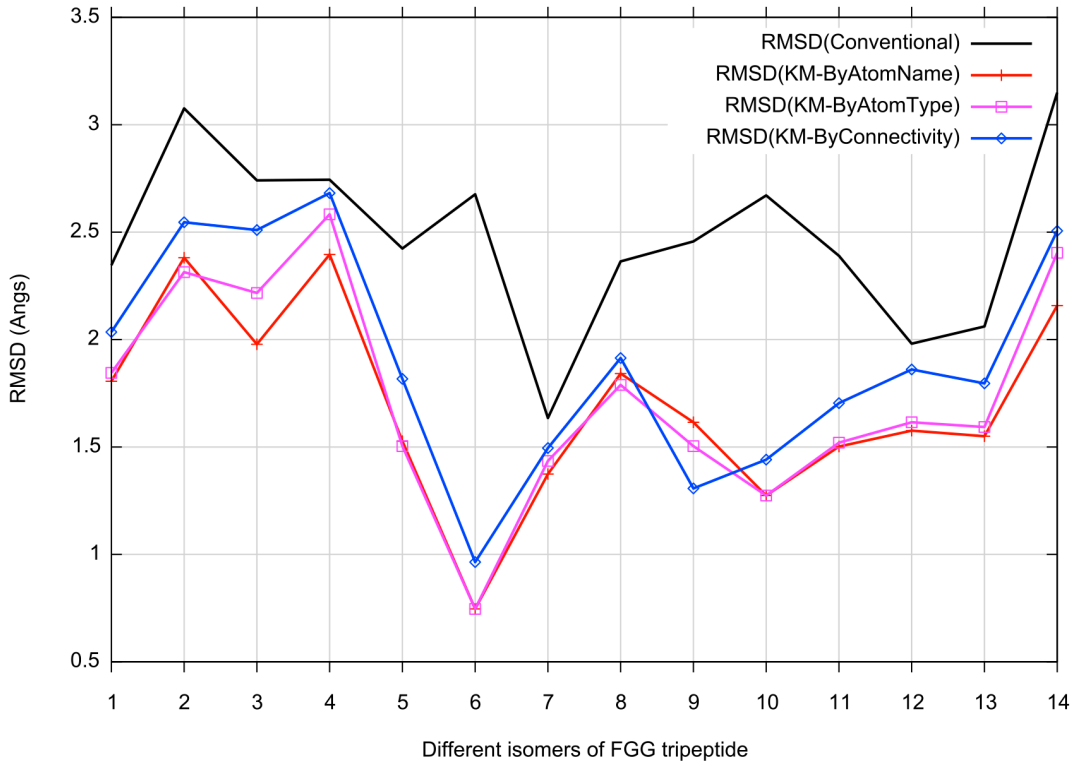


Figure 10. A comparison of conventional and Kuhn-Munkres RMSD matching for isomers of FGG tripeptide. The RMSD is between the most stable isomer and the next fourteen isomers.

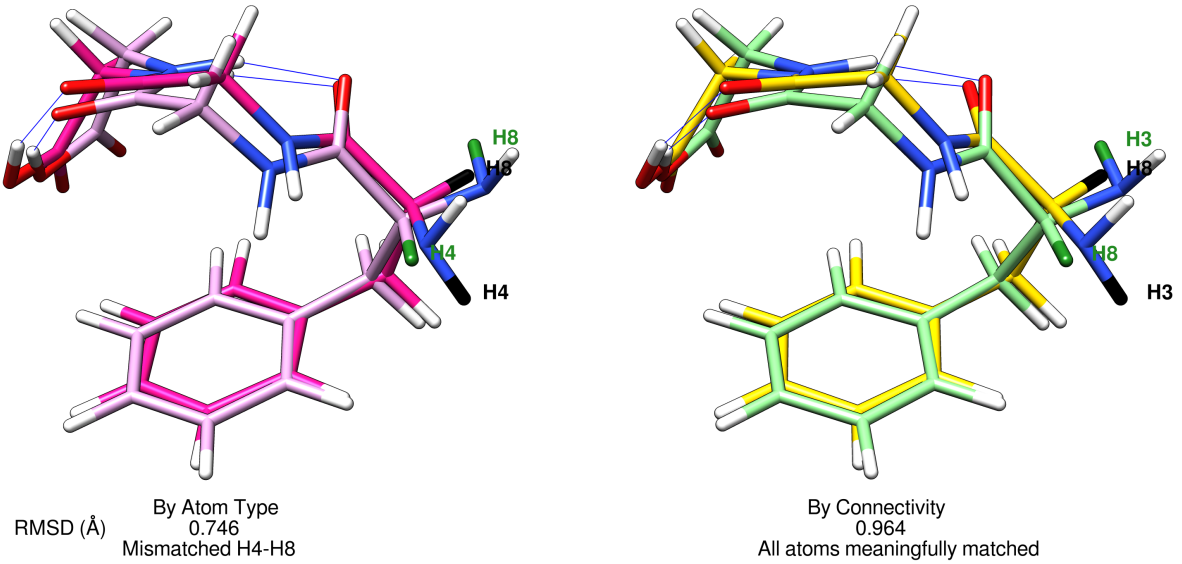
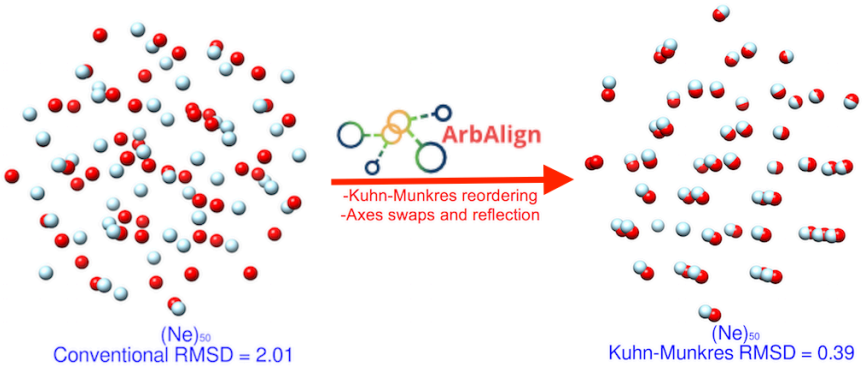


Figure 11. A comparison of conventional and Kuhn-Munkres RMSD matching for a pair of isomers of FGG tripeptide. Assignments based on atom names or types fail to distinguish between different types of hydrogen atoms as shown in the mismatch between H4 and H8 atoms on the left. Using atom connectivity properly assigns H3 and H8. It is a more physically meaningful alignment despite the higher RMSD.

TOC Graphic


ArbAlign
-Kuhn-Munkres reordering
-Axes swaps and reflection
(Ne)50
Conventional RMSD = 2.01
(Ne)50
Kuhn-Munkres RMSD = 0.39

References

(1) Johnson, M. A.; Maggiora, G. M. *Concepts And Applications Of Molecular Similarity*; Wiley: New York, 1990.

(2) Eckert, H.; Bojorath, J. Molecular Similarity Analysis In Virtual Screening: Foundations, Limitations And Novel Approaches. *Drug Discov. Today* **2007**, *12*, 225-233.

(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.

(4) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique In Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204-3218.

(5) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. Shaep: Molecular Overlay Based On Shape And Electrostatic Potential. *J. Chem Inf. Model.* **2009**, *49*, 492-502.

(6) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition To Search Compound Databases For Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28*, 1711-1723.

(7) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method Of Molecular Shape Comparison: A Simple Application Of A Gaussian Description Of Molecular Shape. *J. Comput. Chem.* **1996**, *17*, 1653-1666.

(8) Taminau, J.; Thijs, G.; De Winter, H. Pharao: Pharmacophore Alignment And Optimization. *J. Mol. Graph. Model.* **2008**, *27*, 161-169.

(9) Wolber, G.; Dornhofer, A. A.; Langer, T. Efficient Overlay Of Small Organic Molecules Using 3d Pharmacophores. *J. Comput. Aided Mol. Des.* **2006**, *20*, 773-788.

(10) Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. Field Interaction And Geometrical Overlap: A New Simplex And Experimental Design Based Computational Procedure For Superposing Small Ligand Molecules. *J. Med. Chem.* **2003**, *46*, 1359-1371.

(11) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T. Molecular Similarity And Diversity In Chemoinformatics: From Theory To Applications. *Mol. Divers.* **2006**, *10*, 39-79.

(12) Ferré, G.; Maillet, J.-B.; Stoltz, G. Permutation-Invariant Distance Between Atomic Configurations. *J. Chem. Phys.* **2015**, *143*, 104114.

(13) Petitjean, M. On The Root Mean Square Quantitative Chirality And Quantitative Symmetry Measures. *J. Math. Phys.* **1999**, *40*, 4587-4595.

(14) Barakat, M. T.; Dean, P. M. Molecular-Structure Matching By Simulated Annealing .1. A Comparison Between Different Cooling Schedules. *J. Comput. Aided Mol. Des.* **1990**, *4*, 295-316.

(15) Bayada, D. M.; Simpson, R. W.; Johnson, A. P.; Laurenco, C. An Algorithm For The Multiple Common Subgraph Problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 680-685.

(16) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. Mimic: A Molecular-Field Matching Program. Exploiting Applicability Of Molecular Similarity Approaches. *J. Comput. Chem.* **1997**, *18*, 934-954.

(17) Nissink, J. W. M.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition Of Molecules: Electron Density Fitting By Application Of Fourier Transforms. *J. Comput. Chem.* **1997**, *18*, 638-645.

(18) Flower, D. R. Rotational Superposition: A Review Of Methods. *J. Mol. Graph. Model.* **1999**, *17*, 238-244.

(19) Girones, X.; Robert, D.; Carbo-Dorca, R. Tgsa: A Molecular Superposition Program Based On Topo-Geometrical Considerations. *J. Comput. Chem.* **2001**, *22*, 255-263.

(20) Karney, C. F. F. Quaternions In Molecular Modeling. *J. Mol. Graph. Model.* **2007**, *25*, 595-604.

(21) Vasquez-Perez, J. M.; Martinez, G. U. G.; Koster, A. M.; Calaminici, P. The Discovery Of Unexpected Isomers In Sodium Heptamers By Born-Oppenheimer Molecular Dynamics. *J. Chem. Phys.* **2009**, *131*, 10.

(22) Marques, J. M. C.; Llanio-Trujillo, J. L.; Abreu, P. E.; Pereira, F. B. How Different Are Two Chemical Structures? *J. Chem. Inf. Model.* **2010**, *50*, 2129-2140.

(23) Kawabata, T. Build-Up Algorithm For Atomic Correspondence Between Chemical Structures. *J. Chem. Inf. Model.* **2011**, *51*, 1775-1787.

(24) Teixeira, A. L.; Falcao, A. O. Noncontiguous Atom Matching Structural Similarity Function. *J. Chem. Inf. Model.* **2013**, *53*, 2511-2524.

(25) Kabsch, W. A Solution For The Best Rotation To Relate Two Sets Of Vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922-923.

(26) Coutsias, E. A.; Seok, C.; Dill, K. A. Using Quaternions To Calculate Rmsd. *J. Comput. Chem.* **2004**, *25*, 1849-1857.

(27) Faulon, J.-L. Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 432-444.

(28) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A combinatorial algorithm for calculating ligand binding. *J. Comput. Chem.* **1984**, *5*, 24-34.

(29) Tosco, P.; Balle, T.; Shiri, F. Open3dalign: An Open-Source Software Aimed At Unsupervised Ligand Alignment. *J. Comput. Aided Mol. Des.* **2011**, *25*, 777-783.

(30) Allen, W. J.; Rizzo, R. C. Implementation Of The Hungarian Algorithm To Account For Ligand Symmetry And Similarity In Structure-Based Design. *J. Chem. Inf. Model.* **2014**, *54*, 518-529.

(31) Helmich, B.; Sierka, M. Similarity recognition of molecular structures by optimal atomic matching and rotational superposition. *J. Comput. Chem.* **2012**, *33*, 134-140.

(32) Wagner, A.; Himmel, H. J. aRMSD: A Comprehensive Tool for Structural Analysis. *J. Chem. Inf. Model.* **2017**, *57*, 428-438.

(33) Kuhn, H. W. The Hungarian Method For The Assignment Problem. *Naval Research Logistics Quarterly* **1955**, *2*, 83-97.

(34) Munkres, J. Algorithms For The Assignment And Transportation Problems. *J. Soc. Indust. Appl. Math.* **1957**, *5*, 32-38.

(35) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33.

(36) Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment Through Multilevel Neighborhoods Of Atoms: Definition And Comparison With The Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666-670.

(37) Kromann, J. C.; Steinmann, C.; Bratholm, L. A.; Lauritzen, K. P. *RMSD: Calculate RMSD For Two XYZ Structures*. https://github.com/charnley/rmsd/tree/v1.1; (accessed Dec 1, 2015).

(38) Cooper, H. *Hungarian: Munkres' Algorithm For The Linear Assignment Problem, In Python*. https://github.com/Hrldcpr/Hungarian; (accessed January 15, 2015).

(39) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The Numpy Array: A Structure For Efficient Numerical Computation. *Computing in Science & Engineering* **2011**, *13*, 22-30.

(40) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. Ucsf Chimera - A Visualization System For Exploratory Research And Analysis. *J. Comput. Chem.* **2004**, *25*, 1605-1612.

(41) Clapper, B. M. *Munkres 1.0.8: Munkres Algorithm For The Assignment Problem*. https://pypi.python.org/Pypi/Munkres; (accessed March 15, 2016).

(42) Shields, R. M.; Temelso, B.; Archer, K. A.; Morrell, T. E.; Shields, G. C. Accurate Predictions Of Water Cluster Formation, $(H_2o)_{N=2-10}$. *J. Phys. Chem. A* **2010**, *114*, 11725-11737.

(43) Temelso, B.; Archer, K. A.; Shields, G. C. Benchmark Structures And Binding Energies Of Small Water Clusters With Anharmonicity Corrections. *J. Phys. Chem. A* **2011**, *115*, 12034-12046.

(44) Wales, D. J.; Doye, J. P.; Dullweber, A.; Hodges, M. P.; Naumkin, F. Y.; Calvo, F.; Hernández-Rojas, J.; Middleton, a. T. F. The Cambridge Cluster Database; http://www-wales.ch.cam.ac.uk/CCD.Html; (Accessed March 1, 2015).

(45) Wales, D. J.; Doye, J. P. K. Global Optimization By Basin-Hopping And The Lowest Energy Structures Of Lennard-Jones Clusters Containing Up To 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111-5116.

(46) Xiang; Cheng; Cai; Shao. Structural Distribution Of Lennard-Jones Clusters Containing 562 To 1000 Atoms. *J. Phys. Chem. A* **2004**, *108*, 9516-9520.

(47) Xiang; Jiang; Cai; Shao. An Efficient Method Based On Lattice Construction And The Genetic Algorithm For Optimization Of Large Lennard-Jones Clusters. *J. Phys. Chem. A* **2004**, *108*, 3586-3592.

(48) Bustos, D. J.; Temelso, B.; Shields, G. C. Hydration Of The Sulfuric Acid-Methylamine Complex And Implications For Aerosol Formation. *J. Phys. Chem. A* **2014**, *118*, 7430-7441.

(49) Valdes, H.; Pluháčková, K.; Pitonák, M.; Rezáč, J.; Hobza, P. Benchmark Database On Isolated Small Peptides Containing An Aromatic Side Chain: Comparison Between Wave Function And Density Functional Theory Methods And Empirical Force Field. *Phys Chem Chem Phys* **2008**, *10*, 2747-2757.

(50) Rezac, J.; Jurecka, P.; Riley, K. E.; Cerny, J.; Valdes, H.; Pluhackova, K.; Berka, K.; Rezac, T.; Pitonak, M.; Vondrasek, J.; Hobza, P. Quantum Chemical Benchmark Energy And Geometry Database For Molecular Clusters And Complex Molecular Systems (http://www.begdb.com)/ A Users Manual And Examples. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261-1270.