J Comput Aided Mol Des (2011) 25:777–783
DOI 10.1007/s10822-011-9462-9

Open3DALIGN: an open-source software aimed at unsupervised ligand alignment

Paolo Tosco · Thomas Balle · Fereshteh Shiri

Received: 13 June 2011 / Accepted: 13 July 2011 / Published online: 27 July 2011

Abstract An open-source, cross-platform software aimed at conformer generation and unsupervised rigid-body molecular alignment is presented. Different algorithms have been implemented to perform single and multi-conformation superimpositions on one or more templates. Alignments can be accomplished by matching pharmacophores, heavy atoms or a combination of the two. All methods have been successfully validated on eight comprehensive datasets previously gathered by Sutherland and co-workers. High computational performance has been attained through efficient parallelization of the code. The unsupervised nature of the alignment algorithms, together with its scriptable interface, make Open3DALIGN an ideal component of high-throughput, automated cheminformatics workflows.

Keywords Cheminformatics · Alignment · Superposition · Pharmacophore

Introduction

The unsupervised alignment of a structurally diverse series of biologically active ligands is a classical cheminformatics task, which preludes many ligand-based drug design methodologies, most importantly pharmacophore elucidation and three-dimensional quantitative structure–activity relationship (3D-QSAR) studies [1, 2]. Alignment consists of two operations, feature matching and conformational search [3]. Feature matching can be accomplished through several approaches, which have been object of comprehensive reviews [1, 2]; briefly, field-based, pharmacophore-based and atom-based methods can be used, according to whether a match between molecular interaction fields (MIFs), a collection of pharmacophoric points or heavy atom pairs is sought. About the conformational search strategy, candidate ligands may either be flexibly aligned on the template(s), or the best-fitting conformer for each ligand may be extracted from pre-built conformer libraries following rigid alignment on the template(s) [3]. While a number of closed-source tools exist pursuing the different strategies (BRUTUS [4], ShaEP [5], CATALYST [6], PHASE [7], GALAHAD [8] to mention but a few of them), the open-source arena is not very populated [9], with Pharao [10] and the Chemistry Development Kit (CDK) [11] being the most outstanding representatives. As part of our continuing effort to build a comprehensive, open-source, cross-platform software suite for 3D-QSAR and pharmacophore modelling [12, 13], herein we describe and validate Open3DALIGN, a package dedicated to conformational search and rigid-body, multi-conformer pharmacophore/atom-based ligand alignment.

Methods

Conformational searches were implemented through quenched molecular dynamics (QMD) [14], using the Merck force field and TINKER [15] as the molecular

P. Tosco (✉) · F. Shiri
Department of Drug Science and Technology,
University of Turin, Via Pietro Giuria 9, 10125 Turin, Italy
e-mail: paolo.tosco@unito.it

T. Balle
Department of Medicinal Chemistry,
The Faculty of Pharmaceutical Sciences,
University of Copenhagen,
2 Universitetsparken, 2100 Copenhagen, Denmark

F. Shiri
Faculty of Chemistry, Razi University,
Baghabrisham, 6714967346 Kermanshah, Iran

mechanics engine. Assignment of MMFF94 atom types, bond types and charges in the format required by TINKER was accomplished using SDF2XYZ2SDF [16], our open-source tool based on OpenBabel [17]. In addition to being generated by the built-in QMD engine, conformational databases can be easily imported as SD files from external sources (e.g., MOE [18], OMEGA [19]). Pharmacophore-based alignments were implemented calling Pharao routines [10], while for atom-based alignments a new method was developed, inspired by the LAMDA algorithm [20]. In detail, a cost matrix is computed for each possible heavy atom pair match between template and candidate molecules, using a cost function similar to the one proposed by Richmond and co-workers [20], based on MMFF94 partial charges and MMFF94 atom types (Eq. 1):

$$c_{ij} = w_1\left|q(T)_i - q(C)_j\right| + w_2 s_{ij} + \sum_{k=1}^{K}\frac{\left(h(T)_{ik} - h(C)_{jk}\right)^2}{h(T)_{ik} + h(C)_{jk}}, \quad (1)$$

where c_{ij} is the cost of matching atoms i,j between template (T) and candidate (C) molecules; $q(T)_i$ and $q(C)_j$ are MMFF94 charges of atoms i,j in T and C, respectively; s_{ij} is the ij element of an empirically determined symmetric matrix **S** which scores the degree of chemical/electronic similarity between MMFF94 atom types i and j (1: identity; 20: maximum dissimilarity); $h(T)$ and $h(C)$ are K-binned histograms whose k-th bin represents the number of atoms in T and C lying in a $(k - 1, k)$ distance range from the i-th and j-th atom, respectively; w_1 and w_2 are empirical weights; K was set to 20 as in the cited work [20].

The total heavy atom matching cost is minimised by the Jonker-Volgenant algorithm [21], thus obtaining the **E** matrix, with n rows describing intermolecular atom equivalences. As in the LAMDA algorithm, atom equivalences are ranked according to the degree to which intramolecular atomic distances are respected between template and candidate molecules. Starting from the three top-ranked ones, p atom pairs (with $3 < p < n$) are progressively extracted from the initial **E** matrix into the **D** matrix, and used to perform a first tentative alignment with the RMS algorithm described by Kearsley [22]; each pair of matching atoms is weighted according to their degree of MMFF94 similarity. The quality of the superposition is scored by a simple Gaussian function (Eq. 2):

$$F = \sum_{k=1}^{p}\left(\alpha - \left|q(T)_{\mathbf{D}[k,0]} - q(C)_{\mathbf{D}[k,1]}\right| + \beta \cdot f(T)_{\mathbf{D}[k,0]} \cdot f(C)_{\mathbf{D}[k,1]}\right) \cdot \exp\left(-r^2_{\mathbf{D}[k,0]\mathbf{D}[k,1]}\right) \quad (2)$$

where $q(*)$ are the MMFF94 charges of the k-th pair; $f(*)$ are the formal charges of the k-th pair; r is the Euclidean distance between atoms of the k-th pair; α and β are empirical coefficients (currently, 2.0 and 5.0).

The scoring function was designed to encourage the match between atoms with similar electrostatic charge, giving an additional bonus to atom pairs bearing a non-zero formal charge of the same sign and a penalty to oppositely charged centres. The atom equivalence matrix **D** and the candidate molecule coordinates are then refined by sequentially calling the SDM algorithm [23] (with a distance threshold t) and the RMS algorithm until the scoring function reaches a maximum. The whole procedure is iteratively repeated while parameters w_1, w_2, t and p are tweaked to maximize the quality of the alignment as evaluated through Eq. 2. The rationale behind the iterative optimization of w_1 and w_2 parameters is that different molecules benefit from different weighting of the charge-based and atom-type-based terms in Eq. 1; in our experience with the MMFF94 force-field, the linear combination between the two performs better than the sole charge term as in the original LAMDA algorithm formulation. The same consideration also applies to the threshold t value (which was originally fixed to 0.7 Å) and to the number of rows p in the **D** matrix. The top-ranked atom equivalence matrix $\mathbf{D}_{\text{best}}$ is finally used to derive the roto-translation matrix which transforms the candidate molecule coordinates to best fit the template.

A mixed alignment scheme was also implemented, in which the pose yielded by Pharao is refined by iterating the SDM/RMS algorithms until convergence is reached, then scored with Eq. 2; after comparison with the pose given by the atom-based algorithm, the best scoring of the two is retained. This procedure allows correcting initial misalignments due to imperfect atom matching by the atom-based cost function, which could not be fixed by parameter tweaking and subsequent SDM/RMS refinement. These misalignments occur on a minority of molecules characterised by having many similar atom types close to each other, such as in condensed heteroaromatic systems, which are often better handled by the pharmacophore-based algorithm.

Open3DALIGN was written in C and linked to high-performance BLAS and LAPACK libraries; parallel algorithms were implemented whenever possible, in order to attain high computational performance when using multi-processor architectures.

Validation

As a validation suite we chose eight benchmark datasets gathered from the literature by Sutherland and co-workers [24]. These datasets include inhibitors of the angiotensin converting enzyme (ACE), acetylcholinesterase (AChE),

benzodiazepine receptor (BZR), cyclooxygenase-2 (COX2), dihydrofolate reductase (DHFR), glycogen phosphorylase b (GPB), thermolysin (THERM) and thrombin (THR). They are quite large (from 66 to 397 compounds) and characterized by a wide range of size, flexibility and stereoelectronic properties; since their biological activities cover at least four orders of magnitude, they also constitute excellent datasets for QSAR. All datasets were supplied by Sutherland in the Supporting Information as 3D geometries aligned according to the original literature, namely by flexible alignment on one or more templates obtained by crystallographic enzyme-inhibitor complexes. When multiple templates were available for a certain dataset, the authors of the original work aligned each compound onto the most similar template; to accomplish the same task in an unsupervised fashion, Open3DALIGN attempts to superimpose each compound on all available templates, then chooses the best scoring one. To carry out the first round of validation, the following automated, script-driven workflow was applied to each dataset:

- aligned 3D coordinate SD files were imported;
- alignments were scrambled by random rigid-body roto-translation in Cartesian space;
- datasets were re-aligned by means of the three schemes implemented in Open3DALIGN, i.e., purely pharmacophore-based, purely atom-based, mixed;
- average RMSD values were computed between the original alignments and the ones generated by Open-3DALIGN to assess the performance of the different superposition schemes (Fig. 1).

Figure 1 shows that Open3DALIGN has a remarkable success in reproducing literature alignments; except in the case of the GPB dataset, the best results are obtained with the mixed alignment scheme, with average RMSD values consistently below 0.8 Å, and for half of the datasets even below 0.2 Å. While the overall performance of the pure atom-based scheme is superior to the pure pharmacophore-based one in seven cases out of eight, certain datasets (BZR, DHFR) benefit from the contribution of Pharao on selected compounds. As previously mentioned, structures characterised by condensed aromatic systems with few or no substituents are not always correctly aligned by the first step of the atom-based algorithm, namely the cost function minimisation which yields the initial atom equivalence matrix **E**. A misalignment may arise from the existence in the cost matrix of multiple paths with similar cost, the lowest cost path not necessarily being the best choice. Complementing an atom-based and a pharmacophore-based initial alignment method appears as a more efficient solution to this issue compared to the implementation of a linear assignment algorithm capable of finding multiple low-cost paths, which would cause a substantial increase in CPU time, as already experienced by previous investigators [20].

For the second round of validation a much more challenging task was chosen: for each compound of the dataset a conformational database was supplied along with the original geometry, then a multi-conformational alignment was performed, in order to assess Open3DALIGN's ability to pick the conformation which best fits the template among many different ones. For this purpose, a QMD conformational search was carried out on each structure (MMFF94 force-field, GB/SA implicit solvent model, 1,000 5-ps molecular dynamics runs at 1000 K followed by energy minimization), keeping the most stable conformations in a 8-kcal mol^{-1} range from the global minimum; pairs of conformers whose heavy atom RMSD was below 0.2 Å were considered duplicate and the higher energy one was discarded. The multi-conformational alignment procedure sequentially attempts superimposing all available conformations for each compound to the template; finally, the conformation which yields the best alignment score is chosen. Atom-based and mixed superimpositions are scored with Eq. 2, while the Tanimoto score implemented in Pharao [10] is used for the pure pharmacophore-based approach. As in the first validation run, the performance of the three alignment schemes was evaluated.

Figure 2 shows that again Open3DALIGN succeeded in reproducing the original alignment to a good extent, with the same trend of accuracy among the different algorithms observed for the corresponding single-conformation validation run (mixed $\geq$ atom-based $>>$ pharmacophore-based, except for GPB). Compared to the single-conformation test, average RMSD values are roughly double; this is not surprising, since the availability of different conformations in addition to the original one may give rise to equally good or better alternative superimpositions, especially for the compounds which have only an extremely feeble resemblance with the template(s), as is the case for many molecules in the ACE and THERM datasets. For such compounds there is a high chance that in the conformational database a conformer may exist which has a good local match with part of the template, therefore being preferred to the one used in the original literature, which has a uniformly low global match. However, while for ACE and THERM the atom-based algorithm yields only a few misaligned compounds in the context of a generally sound alignment, the absence of a common pharmacophore across the dataset makes it almost impossible for Pharao to obtain a good quality alignment (Fig. 2). Rather than pointing to specific flaws or limitations in the algorithms implemented in Open3DALIGN, these results suggest that such highly flexible and diverse datasets are not the most suited to be aligned by rigid superposition of pre-computed conformations.

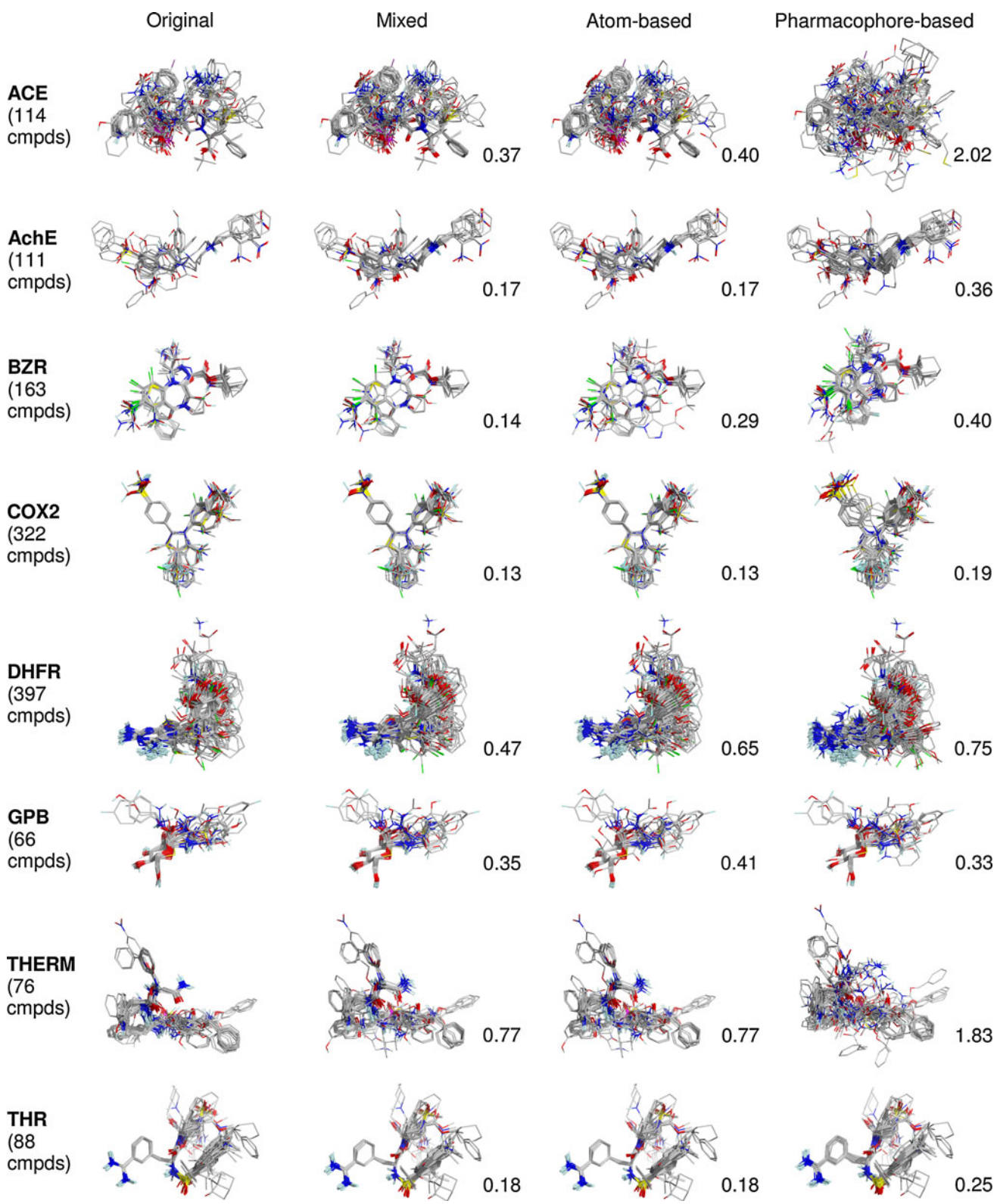

Fig. 1 Results for the *single-conformation* validation suite; heavy-atom RMSD from the original alignment is reported for each dataset

Timings for the aforementioned validation suites are reported in Table 1. While the single-conformation alignment of a dataset constituted by a few hundred compounds is a matter of seconds, the whole multi-conformation validation set (8 datasets, over $1.5 \cdot 10^5$ conformations) can be aligned with the most effective, mixed algorithm in less

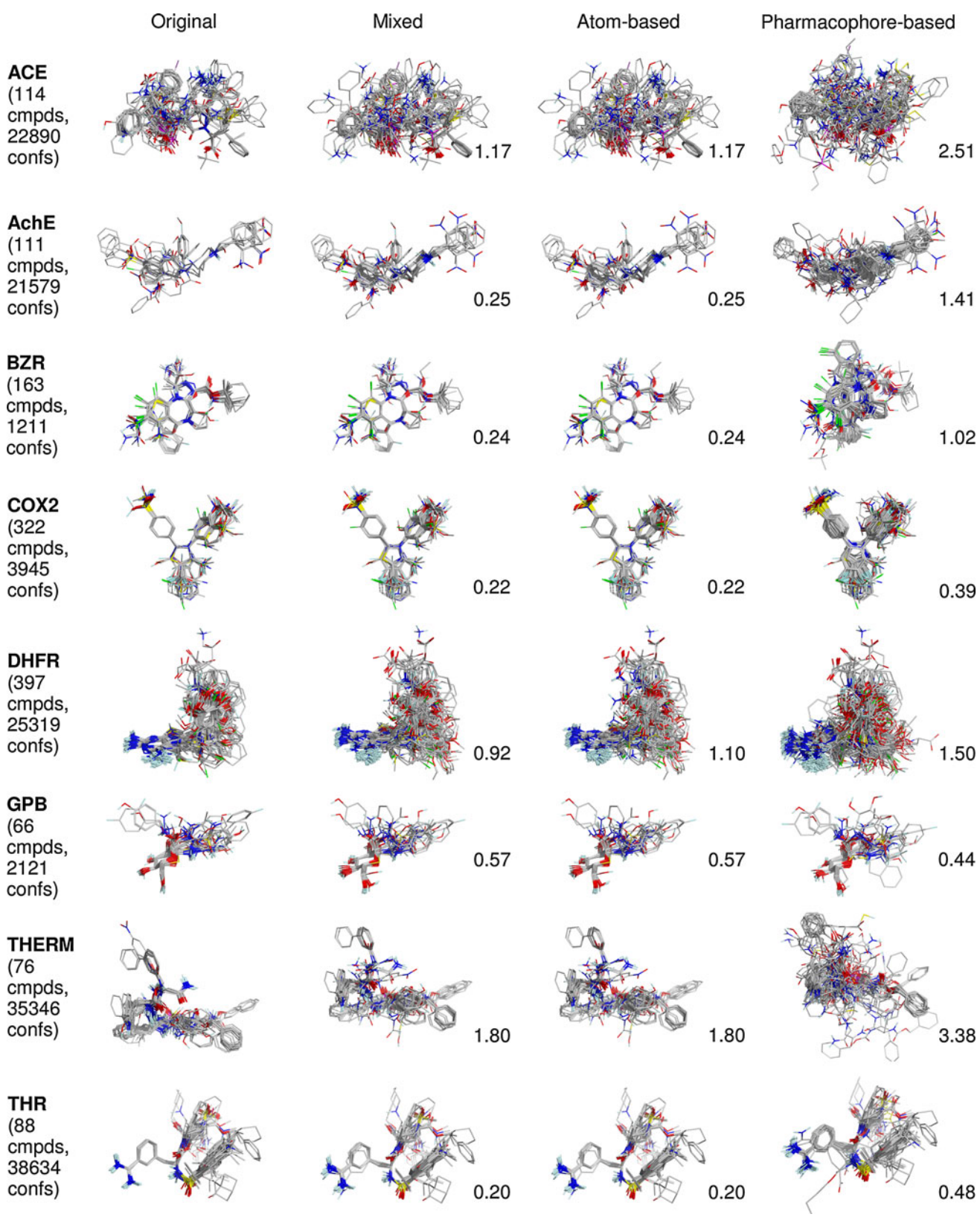

Fig. 2 Results for the *multi-conformation* validation suite; heavy-atom RMSD from the original alignment is reported for each dataset

Table 1 Timings for Open3DALIGN validation suites

	Mixed	Atom-based	Pharmacophore-based
Single-conformation			
validation suite[a]	0:01:25	0:00:35	0:00:49
Multi-conformation			
validation suite[b]	2:50:26	2:21:44	0:29:39
Single alignment[c]	0.0203	0.0169	0.0035

Timings were computed on a 2.4 GHz 8-core Intel Xeon E5620 workstation

[a] Time (h:mm:ss) required to carry out the whole single-conformation validation suite (8 datasets, 1,337 compounds)

[b] Time (h:mm:ss) required to carry out the whole multi-conformation validation suite (8 datasets, 1,337 compounds, 151,045 conformations, using an average of 3.34 templates per conformation)

[c] Time (s) required to accomplish a single alignment operation on a single template, computed as the average over all alignment operations performed during the multi-conformation validation suite

than 3 hours on a 2.4 GHz 8-core Intel Xeon E5620 workstation (0.02 s conformation^{-1}). As might be expected, the greater accuracy of the atom-based algorithm implies a roughly 5-fold increase in CPU time compared to the pharmacophore-based one; the reason why the atom-based algorithm appears to perform better in the single-conformation benchmark is due to the I/O bottleneck connected with calls to Pharao, which reduces the efficiency of the parallelization on short computations. On the same machine, running 1,000 5-ps QMD runs followed by conformational sorting requires about 4–6 min per molecule, depending on its size.

Conclusions

An open-source tool was described capable of carrying out conformational searches and multi-conformational, unsupervised rigid-body alignment of 3D molecular structures. Multiple alignment paradigms were implemented (atom-based, pharmacophore-based, mixed); in general, the mixed and atom-based superposition algorithms are those giving rise to the most consistent and well-ordered alignments, which are particularly suited for 3D-QSAR techniques, where the quality of the model critically depends on the consistency of the underlying alignment. In this respect, Open3DALIGN constitutes an ideal complement to Open3DGRID [12] and Open3DQSAR [13], our recently released open-source tools focused on MIF computation and 3D-QSAR model building, since it shares with the former the same file formats and user interface. High computational performance, the unsupervised nature of the alignment algorithms and its scriptable interface make Open3DALIGN an ideal component of automated cheminformatics workflows.

Open3DALIGN is available for download free of charge under the terms of the GNU GPLv3 at the URL http://open3dalign.org/, both as source code and as precompiled binaries for the mainstream operating systems; full documentation as well as the files necessary to run the validation suite described in this paper are also included.

Acknowledgments We are grateful to the developers of OpenBabel, Pharao and TINKER, on which Open3DALIGN depends to do its job, and to an anonymous reviewer who contributed to improve this manuscript. We acknowledge the support of Chemical Computing Group. Part of this work was carried out by P.T. at the University of Copenhagen under a visiting scientist grant supported by the Drug Research Academy (DRA). T.B. was supported by grants from the Lundbeck Foundation. Part of this work was carried out by F.S. at the University of Turin under a visiting scientist grant.

References

1. Chan SL, Labute P (2010) Training a scoring function for the alignment of small molecules. J Chem Inf Model 50:1724–1735
2. Lemmen C, Lengauer T (2000) Computational methods for the structural alignment of molecules. J Comput Aided Mol Des 14:215–232
3. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) Three-dimensional pharmacophore methods in drug discovery. J Med Chem 53:539–558
4. Rönkkö T, Tervo AJ, Parkkinen J, Poso A (2006) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization. J Comput Aided Mol Des 20:227–236
5. Vainio MJ, Santeri Puranen J, Johnson MS (2009) ShaEP: molecular overlay based on shape and electrostatic potential. J Chem Inf Model 49:492–502
6. Güner O, Clement O, Kurogi Y (2004) Pharmacophore modeling and three dimensional database searching for drug design using Catalyst: recent advances. Curr Med Chem 11:2991–3005
7. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J Comput Aided Mol Des 20:647–671
8. Richmond NJ, Abrams CA, Wolohan PRN, Abrahamian E, Willett P, Clark RD (2006) GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. J Comput Aided Mol Des 20:567–587

9. Van Drie JH (2010) History of 3D pharmacophore searching: commercial, academic and open-source tools. Drug Disc Today 7:e255–e262
10. Taminau J, Thijs G, De Winter H (2008) Pharao: pharmacophore alignment and optimization. J Mol Graph Model 27:161–169
11. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. Curr Pharm Des 12:2111–2120
12. Tosco P, Balle T (2011) Open3DGRID: an open-source software aimed at high-throughput generation of molecular interaction fields (MIFs); http://open3dgrid.org/. Accessed 13 June 2011
13. Tosco P, Balle T (2011) Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. J Mol Model 17: 201–208; http://open3dqsar.org/. Accessed 13 June 2011
14. O'Connor SD, Smith PE, al-Obeidi F, Pettitt BM (1992) Quenched molecular dynamics simulations of tuftsin and proposed cyclic analogs. J Med Chem 35:2870–2881
15. TINKER—Software tools for molecular design, version 5.1; http://dasher.wustl.edu/tinker/. Accessed 13 June 2011
16. Tosco P, Balle T, Shiri F (2011) SDF2XYZ2SDF: how to exploit TINKER power in chemoinformatics projects. J Mol Model. Doi: 10.1007/s00894-011-1111-7; http://sdf2xyz2sdf.sourceforge.net/. Accessed 13 June 2011
17. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner JK, Willighagen E (2006) The Blue Obelisk—Interoperability in chemical informatics. J Chem Inf Model 46:991–998
18. MOE 2010.10; Chemical Computing Group Inc., Montreal, Quebec, Canada, 2010
19. Kirchmair J, Wolber G, Laggner C, Langer T (2006) Comparative performance assessment of the conformational model generators Omega and Catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations. J Chem Inf Model 46:1848–1861
20. Richmond NJ, Willett P, Clark RD (2004) Alignment of three-dimensional molecules using an image recognition algorithm. J Mol Graph Model 23:199–209
21. Jonker R, Volgenant A (1987) A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing 38:325–340
22. Kearsley SK (1989) On the orthogonal transformation used for structural comparisons. Acta Cryst A45:208–210
23. Petitjean M (1998) Interactive maximal common 3D substructure searching with the combined SDM/RMS algorithm. Computers Chem 22:463–465
24. Sutherland JJ, O'Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure—activity relationships. J Med Chem 47:5541–5554