JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

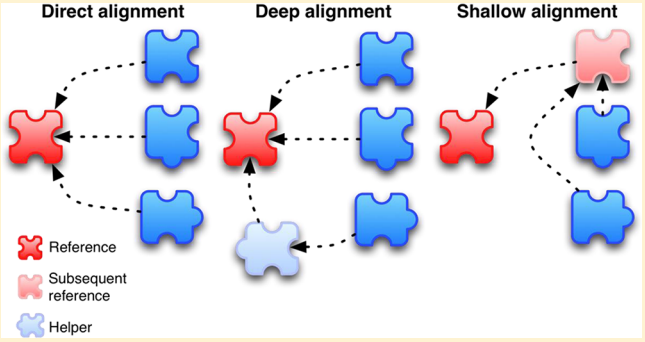
MARS: Computing Three-Dimensional Alignments for Multiple Ligands Using Pairwise Similarities

Thomas Klabunde,[†] Clemens Giegerich,[†] and Andreas Evers*[,†]

[†]Sanofi-Aventis Deutschland GmbH, R&D LGCR/Struct., Design & Informatics, 65926 Frankfurt am Main, Germany

Supporting Information

ABSTRACT: The three-dimensional (3D) superimposition of molecules of one biological target reflecting their relative bioactive orientation is key for several ligand-based drug design studies (e.g., QSAR studies, pharmacophore modeling). However, with the lack of sufficient ligand-protein complex structures, an experimental alignment is difficult or often impossible to obtain. Several computational 3D alignment tools have been developed by academic or commercial groups to address this challenge. Here, we present a new approach, MARS (Multiple Alignments by ROCS-based Similarity), that is based on the pairwise alignment of all molecules within the data set using the tool ROCS (Rapid Overlay of Chemical Structures). Each pairwise alignment is scored, and the results are captured in a score matrix. The ideal superimposition of the compounds in the set is then identified by the analysis of the score matrix building stepwise a superimposition of all molecules. The algorithm exploits similarities among all molecules in the data set to compute an optimal 3D alignment. This alignment tool presented here can be used for several applications, including pharmacophore model generation, 3D QSAR modeling, 3D clustering, identification of structural outliers, and addition of compounds to an already existing alignment. Case studies are shown, validating the 3D alignments for six different data sets.



INTRODUCTION

The generation of ligand alignments is generally an important task in computer-aided drug design. Visual inspection and computational analysis of molecular features which are common among compounds with different chemical scaffolds but similar with respect to chemical features allow for the derivation of a "common-pharmacophore" hypothesis. Analysis of the different features within one chemical series in 3D space gives insights into structure−activity relationships. Finally, a correct overlay of different ligands (showing a similar binding mode) allows to combine parts of two molecules into new hybrid molecules. Furthermore, a 3D overlay of different compounds of a target allows to identify compounds with similar or different binding modes, which can be used as a starting point for the generation of binding-mode specific pharmacophores.

For the generation of ligand alignments, several options are available. In the presence of protein−ligand cocrystal structures, the ligand alignment is inferred from the alignment of the complex structures. In the absence of cocrystal structures, the application of pharmacophore tools[1−5] is one option for the generation of ligand alignments and 3D pharmacophore models. Another option is the application of 3D molecular alignment techniques, which generate 3D alignments of ligands by maximizing their mutual similarities.

Different alignment techniques have been described in the literature. For a detailed overview, the reader is referred to refs 6 and 7. One of the first reported approaches for the generation of pairwise alignments is SEAL (steric and electrostatic alignment),[8] which provides an alignment of two rigid molecules by translating and rotating the structures with respect to each other while minimizing an alignment function containing electrostatic (partial charges) and steric (van der Waals radii) terms represented by smooth Gaussian functions. An extension of SEAL (TORSEAL) has been described by Klebe et al.[9] The alignment function was extended by additional physicochemical properties (hydrophobic fields and hydrogen-bonding properties). A preliminary alignment is generated by SEAL, followed by a postoptimization with the MIMUMBA conformer generator to consider local molecular flexibility. FlexS[10] superimposes two molecules by aligning a flexible test molecule onto a rigid template. The ligand structure is partitioned into fragments, which are reassembled during the search progress. Similarity is measured using hydrogen-bonding and steric overlap terms. The program ROCS[11−13] (Rapid Overlay of Chemical Structures) generates rigid superimpositions of molecule pairs, maximizing the shape overlap of ligand structures represented by smooth Gaussian functions. ROCS can be applied to multiple conformations for the test molecule, but it is necessary to provide the reference molecule in a given conformation. Since different conforma-

Received: January 18, 2012
Published: July 16, 2012

   dx.doi.org/10.1021/ci3000369 | J. Chem. Inf. Model. 2012, 52, 2022−2030

tions of the molecule to be aligned can be considered, flexibility of the ligand (to be aligned) is implicitly considered. Multiple ligand alignments can be obtained by superimposing several compounds onto one reference compound. A detailed evaluation was performed by Chen et al.,[14] who examined the dependence of molecular alignment accuracy (using ROCS and FlexS) on a variety of factors. They found that, among others, alignment accuracy depends critically on the choice of a given reference molecule. Compounds will only be correctly aligned, if they are similar enough to the reference, i.e., showing a similar binding mode and similar interactions to the protein. In a real-life scenario, the best reference molecule, which would provide the best global alignment, will not be known. As a further drawback, knowledge about the bioactive conformation of a reference compound is not always available.

Multiple ligand alignment approaches,[15−19] which consider the conformational space of all ligands and all pairwise ligand similarities, do not require assumptions about a reference compound or its conformer. Such a multiple ligand alignment tool, MULTISEAL,[15] was introduced by Feher et al., which applies the SEAL method to multiple conformations of multiple molecules. Another multiple alignment approach is the DIFGAPE[19] (DIstance geometry Focused Genetic Algorithm Pose Evaluator) approach. It creates multiple ligand alignments by combining the results from exhaustive pairwise ligand alignments generated by ROCS. DIFGAPE takes pairwise alignments of ligand conformations as input together with the scores for those alignments. It then selects conformers using a scoring function which maximizes pairwise scores and penalizes geometric inconsistencies through a distance geometry term. A consensus method is used to create an overlay from the pairwise alignments. In cases where no conformation for a ligand can be reasonably incorporated into the overlay, the ligand is dropped from the final alignment. Thus, as an advantage, the DIFGAPE approach implicitly provides the information whether certain compounds reveal different binding modes or are dissimilar from the major part of alignment molecules. A further approach, named pharmACOphore,[18] generates pairwise as well as multiple flexible alignments of ligands based on ant colony optimization. During the optimization, translational, rotational, and torsional degrees of freedom are varied, thus, molecular flexibility of all ligands is explicitly considered. An empirical scoring function is used, which describes ligand similarity by minimizing the distance of pharmacophoric features.

Most multiple ligand alignment tools are based on optimization of a global alignment scoring function in a way that ligands which exhibit a binding mode that is incompatible with the remaining data set have a negative impact on the result and quality of the final alignment. As described above, this issue has been implicitly addressed with the DIFGAPE approach, which drops compounds that cannot be incorporated into the overlay.

Approaches, which treat the molecules explicitly flexible in the multiple alignment generation, only offer limited possibilities to influence the generation of conformations or to consider only limited conformers of particular molecules. In particular, the option to update existing alignments seems not given, which might be relevant in different stages of a drug design project for the design of new compounds.

By separating the conformation generation process from the alignment process, we introduce the option to use different conformer generators and to consider particular molecules in defined conformations, which might have been derived from experimental studies.

Herein, we present MARS (Multiple Alignments by ROCS-based Similarity), which provides flexible alignments of multiple compounds based on OMEGA[20] conformers and pairwise superimpositions generated in ROCS. MARS uses three different mathematical approaches (the *direct*, *deep*, and *shallow* algorithm) for combining pairwise ROCS-based similarities into final multiple ligand alignments. In contrast to ROCS, which allows to align all molecules of a data set to one reference, while neglecting potential additional similarities among the molecules in the data set, MARS exploits the similarities of all molecule pairs in the data set (beyond the similarity of each aligned molecule to the reference) to get the final alignment. Due to the nature of the algorithms, MARS can be used to generate alignments of large data sets of one or more chemical series for the purpose of 3D QSAR studies, to perform a 3D clustering to group compound sets with (dis-)similar binding modes, and to derive pharmacophore-independent alignments for a smaller compound set of different chemotypes. MARS provides a total score for a multiple alignment and scores for each molecule within the alignment. Since the MARS approach is using ROCS-based similarities, the possibility to fine-tune the influence of specific scoring-terms is possible via selection and relative scaling of appropriate scoring terms available in ROCS. Thus, it is possible to judge the quality of an alignment and eliminate (structural) outliers from a multiple ligand alignment. The MARS approach allows to consistently add new compounds to an already existing ligand alignment. Furthermore, MARS offers the option to eliminate structural outliers from an alignment during alignment generation. On the other hand, it is possible to consider all compounds for alignment generation and perform a subsequent analysis about structural outliers or compounds with different binding modes using a 3D-clustering approach. MARS offers the possibility to start alignment generation without prior knowledge or in-depth analysis of a ligand data set, but it is also possible to influence the alignment generation, for example by selecting molecules to be used as references, by eliminating molecules from alignment generation below a user-defined similarity cutoff and by subsequent addition of molecules to existing 3D-alignments.

To test, whether MARS is able to generate relevant alignment modes, we evaluated the software on six test systems using ligands extracted from structure-based alignments of protein crystal structures.

MATERIALS AND METHODS

General Workflow. For the generation of MARS alignments, the following steps are performed subsequently (see also Figure 1): (1) conformer generation, (2) generation of pairwise ROCS alignments between all conformers of different molecules and generation of a matrix capturing all pairwise similarity values, and (3) analysis of the ROCS similarity matrix to identify multiple ligand alignments using three different algorithms (*direct*, *deep*, and *shallow*) and output of the three final alignments.

1. Generation of Conformer Libraries. In this study, we used OMEGA, version 2.1[20] for conformer generation. Except for the following parameters, default settings were used: The maximum number of conformers was set to 1000, and the energy window was set to 20 kcal/mol. The threshold for duplicate removal was set to 0.8 A rmsd. It is generally possible to use conformers generated from other conformer generation

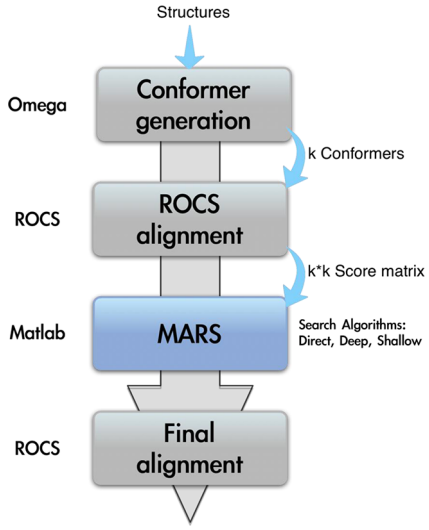


Figure 1. The default MARS workflow starts with conformer generation for all molecules to be aligned, followed by a pairwise ROCS alignment for each conformer pair between different molecules. Subsequently, the matrix is analyzed using three different algorithms (*direct*, *deep*, and *shallow*) with Matlab[28] by searching rows and columns. Finally, for each algorithm, an alignment is generated, resulting in three different multiple ligand alignments.

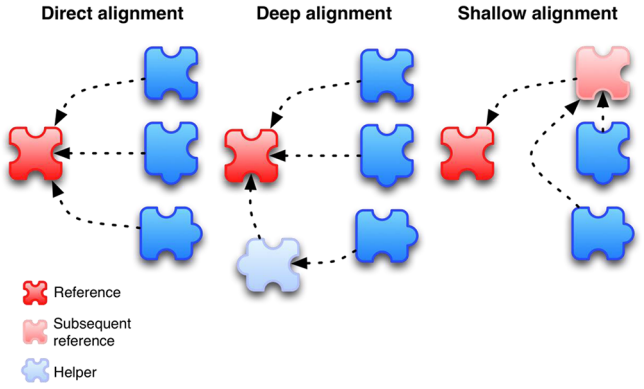


Figure 2. Illustration of the three different MARS algorithms. The ***direct*** algorithm starts with the identification of the optimal reference. All remaining compounds are aligned *directly* onto this reference. The ***deep*** algorithm evaluates for each compound whether the alignment to the reference can be improved by alignment via a helper molecule according to eq 1. The ***shallow*** algorithm starts with the reference from the *direct* algorithm and iteratively moves the most similar compound (= subsequent references) from the alignment set (to any compound from the reference set) into the reference set, until all compounds are transferred from the alignment to the reference set.

methods. In particular, it is possible — and recommended - to include experimentally derived conformations, for example from crystallographic or NMR studies.

2. Generation of Pairwise Alignments and Calculation of the ROCS Similarity Matrix. We used ROCS, version 2.4.1[11] for the calculation of all possible pairwise alignments for each conformer pair between different molecules. ROCS provides different similarity metrics. By default, the ComboScore is used to identify the best alignment between two molecule conformers, but it is possible to use any similarity metric (available in ROCS) for alignment scoring. The results from the pairwise ROCS alignments are stored in a matrix (see Figure 3a).

3. Generation of Multiple Ligand Alignments by Analysis of the Pairwise Similarity Score Matrix Using Three Different MARS Algorithms. In contrast to ROCS, which allows to align all molecules of a data set onto one reference molecule, while neglecting potential additional similarities among the molecules in the data set, MARS exploits the similarities of all molecule pairs in the data set (beyond the similarity of each aligned molecule to the reference) to get the optimal alignments. Thus, the runtime of MARS correlates with the size of the scoring matrix, i.e. grows quadratically with the number of conformers in the input data set.

MARS Algorithms. *Definition of Reference and Helper Molecule.* In order to stepwise generate a multiple alignment, three different MARS algorithms are applied (see also Figure 2 and Figure 3). For the search of optimal alignments from pairwise similarities, we apply the concept of *reference* and *helper* molecules. A *reference* is a molecule which is used as the reference — in a defined conformation — for an alignment. This means that other molecules or molecule conformers are aligned to this reference. It might be detrimental to align a molecule (conformer) directly onto a *reference* when the alignment score is low. In such cases, it is beneficial to use a so-called *helper* molecule which helps to find the correct mapping to the starting *reference*. The *helper* molecule is used when (1.) the mapping of the *target* molecule to the *helper* **and** (2.) the mapping of the *helper* to the *reference* is high enough.

Direct Algorithm. The *direct* algorithm generates a "classical" ROCS alignment, i.e., the best-scoring conformers are directly aligned onto a reference molecule, as demonstrated in Figure 2. The scores of the individual molecules are summarized, yielding an alignment score. This procedure is performed for each molecule conformer as possible "end" reference by calculations on matrix elements only as shown in Figure 3b. Finally, the alignment with the reference, which provides the highest alignment score, is provided as output alignment of the *direct* algorithm.

Deep Algorithm. Like the *direct* algorithm, the *deep* algorithm evaluates each molecule (conformer) as possible *reference* (see Figure 3c). In contrast to the *direct* algorithm, also indirect alignments using a *helper* reference are considered, as indicated in Figure 2. For each conformer to be aligned, the deep algorithm evaluates all remaining molecules as helper molecules which might improve the mapping to the *reference*, according to the following equation

$$score(conf \rightarrow ref) = \begin{cases} score(conf \rightarrow ref) & if\ helperscore(conf \rightarrow ref) \leq 1.1 \\ \max\begin{pmatrix} score(conf \rightarrow ref) \\ helperscore(conf \rightarrow ref) \end{pmatrix} & if\ helperscore(conf \rightarrow ref) > 1.1 \end{cases} \quad (1)$$

where

dx.doi.org/10.1021/ci3000369 | J. Chem. Inf. Model. 2012, 52, 2022−2030

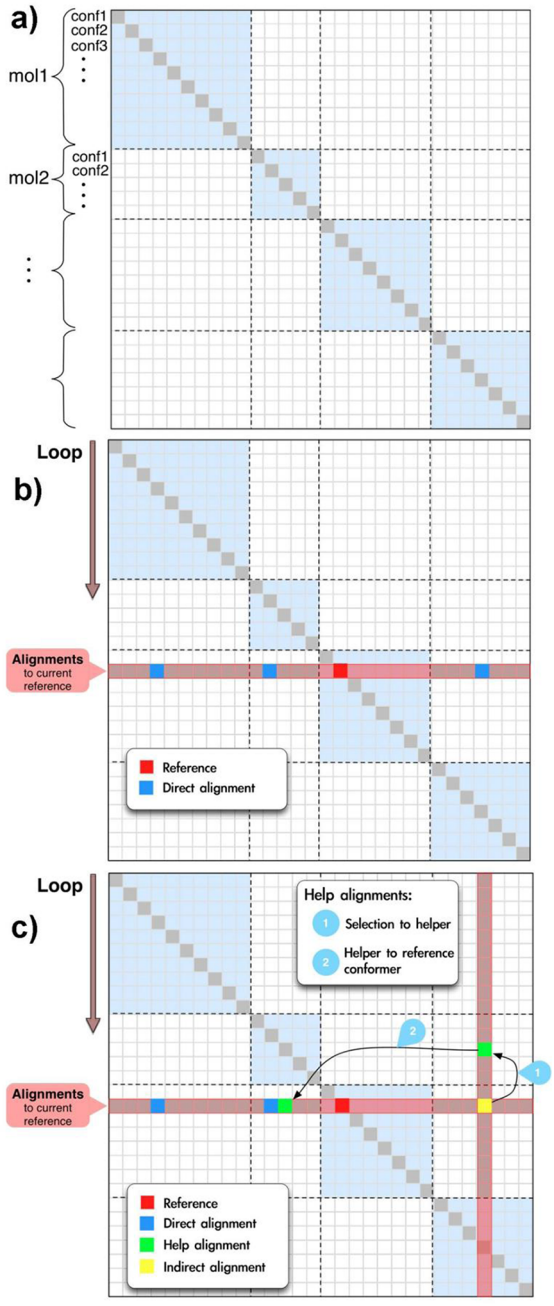


Figure 3. Illustration of matrix calculations applied by MARS. a) shows the similarity matrix between conformers of four different molecules. b) *direct* algorithm: each molecule conformer is considered as a possible reference (traversal along the rows). For the molecule conformer under evaluation (example highlighted in red), all conformers of the remaining molecules are first directly aligned to the reference (traversal along the columns). The best-scoring conformer of each molecule is selected, and the respective scores are added, thus, yielding an alignment score for each reference. Finally, the alignment with the highest alignment score is selected as *direct* alignment. c) *deep* algorithm. As for the *direct* algorithm, each molecule (conformer) is evaluated as a possible reference. The best-scoring conformer of each molecule is aligned to the current reference. If a molecule conformer shows an alignment score below a cutoff of 1.1 to the reference (highlighted in yellow), all conformers of the remaining molecules are evaluated as helper references (green). If the helperscore (according to eq 2) of a conformer to a reference is higher than the *direct* score, the indirect alignment via the helper reference is generated.

$$helperscore(conf \rightarrow ref) = \frac{score(conf \rightarrow helper) + score(helper \rightarrow ref)}{2} \quad (2)$$

If a molecule conformer shows a *helperscore* below a defined cutoff (e.g., ComboScore <1.1) to the reference, the algorithm uses the *direct* alignment to the reference. Otherwise, a comparison of the numeric values of the *direct* score and the *helperscore* determines whether a molecule conformer is aligned directly or via the helper molecule to the reference. All molecules are evaluated as *helper* molecules, and the alignment via the molecule yielding the highest *helperscore* is used. The multiple ligand alignment with the maximum total ROCS score, considering the *helperscores* as given in eq 1, is finally selected. The selection is performed by traversal of the matrix as shown in Figure 3c.

Shallow Algorithm. The *shallow* algorithm is illustrated in Figure 2. It uses the molecule which was identified by the *direct* algorithm as *reference*. It then selects the most similar molecule (conformer) as additional *subsequent reference*. In the following, the molecule (conformer) with the highest similarity to one of the references is selected as further subsequent reference. This procedure is repeated until all molecules are added to the set of subsequent references.

Elimination of Structural Outliers from a MARS Alignment. Using the default settings of MARS, all molecules appear in the final output alignment. This procedure might be inappropriate if some ligands bind into a different (sub)pocket, adopt different binding modes, or are dissimilar from the major part of the data set. Thus, it might not be possible to obtain a consistent alignment for all molecules. Based on experience from ROCS, an alignment with a ComboScore below 1.1 appears to be questionable. Due to this, we offer in MARS to exclude molecules below a certain similarity score - according to eq 1 - from the alignment (by default: ComboScore < 1.1).

Addition of Compounds to an Existing MARS Alignment. MARS not only allows to generate 3D alignments from a set of input molecules. It also allows to add compounds to existing computed or experimentally determined alignments. Here, compounds of an existing alignment represent the references. For the set of compounds to be aligned, 3D conformations are calculated. Following the procedure of the *shallow* algorithm (see above), the molecule (conformer) with the highest similarity to one of the references is selected as further (subsequent) reference. This procedure is repeated until all molecules to be aligned are added to the set of subsequent references.

3D Clustering. Given a MARS-generated 3D alignment, pairwise ROCS similarities of all aligned compounds in their relative spatial orientation are calculated (using the ROCS option *−scoreonly*) and stored in a matrix. A hierarchical clustering of this matrix provides clusters or singletons based on a specified ROCS similarity cutoff value. This approach supports the differentiation of compounds with different binding modes and identification of structural outliers.

Molecular Overlay Data Sets for Method Validation. We investigated whether MARS is able to reproduce the relative spatial orientations of compounds observed in protein crystal structures. For this purpose, we analyzed validation data sets, which have previously been used in other validation studies for ligand-based alignment protocols. We found that for different reasons, not all data sets represent realistic test scenarios for

dx.doi.org/10.1021/ci3000369 | J. Chem. Inf. Model. 2012, 52, 2022−2030

ligand-based alignment procedures. In particular when ligands are highly dissimilar, show different or multiple binding modes, inducing significant rearrangements of the protein binding site, or bind into different (sub)pockets, it is beyond the scope of a ligand-based alignment program to provide the correct answer. Furthermore, several data sets were not used for this study, because ligands reach out of the binding site, stick into the solvent or obviously interact with neighboring proteins from the crystal packing, or show multiple binding modes in the crystal structure. Another reason to skip data sets was due to the observation that several binding sites contained cofactors, organic solvent or buffer molecules next to a ligand, which should but can not be incorporated in the alignment procedure.

We finally selected the following data sets, which have previously been used in other validation studies[14,19] for ligand-based alignment protocols, for method validation: cyclin-dependent kinase 2 (cdk2), p38 mitogen-activated protein kinase (p38), estrogen receptor 1 (ESR1), dihydrofolate reductase (DHFR), and two factor Xa (fXa) data sets (see Table 1).

Table 1. Targets and Ligand Data sets

data set	target	protein family	ligand count
cdk2	cdk2	kinase	57
p38	p38	kinase	13
ESR1	ESR1	hormone receptor	13
DHFR	DHFR	oxidoreductase	12
Fxa_diverse	factor Xa	hydrolase	8
Fxa_focused	factor Xa	hydrolase	11

Generation of Experimental Alignments and Comparison between Computed and Experimental Alignments. For the six validation data sets, the protein crystal structures were extracted from the Protein Data Bank (PDB)[21] and aligned based on the C*alpha* atoms of the binding-site residues (i.e., all residues within a 6.0 Å distance from the ligand molecules) using MOE.[22] The interactions to the target protein were visually inspected to ensure that correct and consistent tautomers and protonation states are assigned to the ligand molecules. Subsequently, the ligands were extracted from the structure-based alignment to generate the experimental ligand alignment. For comparison between computed and experimental alignment, we superimposed the computed conformation of the reference molecule onto the corresponding experimentally determined conformation and applied this coordinate transformation to the computed conformations of the remaining molecules. The heavy atom positions of the computationally aligned molecules were compared to those in the experimental alignment. A molecule was considered to be correctly aligned if the rmsd to the experimental structure was ≤2 Å.

RESULTS AND DISCUSSION

We provide results for four different evaluation and application studies: (1) The first study illustrates the difference between a ROCS- and a MARS-generated alignment for a small data set of four inhibitors of cdk2. (2) The second study is similar to an assessment performed at Lilly:[14] We have used MARS for computation of 3D alignments for six different data sets (see Table 1) and compared these to experimentally derived spatial orientations by analyzing the number of correctly aligned molecules (rmsd ≤2.0 Å) in each data set and comparing the results with classical ROCS alignments. (3) The third study was inspired by an evaluation published by Arena.[19] Here, the same six data sets were used as in (2), but we eliminated structural outliers from the MARS alignment (as described in Materials and Methods). Here, an alignment has been considered as correct if a) 50% of the molecules could be aligned and b) an average rmsd of ≤2.0 Å for all molecules in the final alignment was obtained. (4) Finally, we will show the benefit of 3D clustering for a MARS alignment of a data set of 13 inhibitors of p38.

(1). Comparison of MARS- and ROCS-Generated Alignments of Four Inhibitors of cdk2. Figure 4 illustrates the difference between a MARS- and a ROCS-generated alignment. a) depicts the ligand alignment extracted from a structure-based superimposition of three protein complex structures of cdk2 (pdb codes: 1pxk, 1oit, 1jsv) with the respective ligand. On the left side, the alignment is shown; the right side shows the individual molecules in their aligned coordinates. b) shows an alignment generated by ROCS, which was obtained by superimposing OMEGA-generated conformers of ligands (1oit and 1jsv) onto the reference (1pxk), which was provided in the bioactive conformation. The molecule 1oit (ComboScore 1.18 to the reference) is in good agreement with the orientation observed in the X-ray structure (rmsd 1.46 Å). However, molecule 1jsv, which shows a ROCS score of only 0.88, does not show a near-native orientation (rmsd 4.13 Å). In c) 1oit (ComboScore 1.18, rmsd 1.46 Å) aligns well with the reference and can now serve as additional reference for aligning the next molecule, 1jsv. Since 1jsv shows a higher ROCS score to 1oit (ComboScore: 1.11), 1oit is utilized as a *helper* reference for aligning 1jsv onto 1pxk. Comparison of the MARS-generated pose is in good agreement with the bioactive conformation (rmsd 1.22 Å). This example shows that the alignment accuracy can be improved by using a helper reference, especially when the data set contains molecules that structurally "bridge" different chemical classes by offering chemical features of both.

(2). Comparison of MARS- and ROCS-Generated Alignments for Six Different Data Sets. Similar to an evaluation study performed at Lilly,[14] we compared the agreement of MARS- and ROCS-generated 3D alignments with experimentally determined alignments for the six different data sets listed in Table 1. Whereas the MARS algorithms automatically provide a selection of a reference molecule, the reference for a ROCS-generated alignment must be provided by the user. The results listed in Table 2 confirm a conclusion from the evaluation performed at Lilly, that ROCS-generated alignment results are very sensitive to the choice of the best reference molecule. In the rigid case (i.e., utilizing the molecules in the bioactive conformations and ignoring molecular flexibility), 82 out of 114 molecules from all six data sets (71.9%) are correctly aligned, when the optimal reference was selected (ROCS "max" column). Since the optimal reference is not known in real-life scenarios, the Mean value represents a more realistic benchmark: On average, only 56.8 out of 114 compounds (49.8%) are correctly aligned. With the MARS approach, where the reference compound is automatically selected, the number of correctly aligned structures is higher: The *direct* algorithm, which first provides an automatic selection of a reference molecule and then aligns the remaining molecules in a standard ROCS fashion onto this reference, correctly aligns 75 out of 114 compounds (65.8%). The *shallow* and *deep* algorithms provide the correct alignment

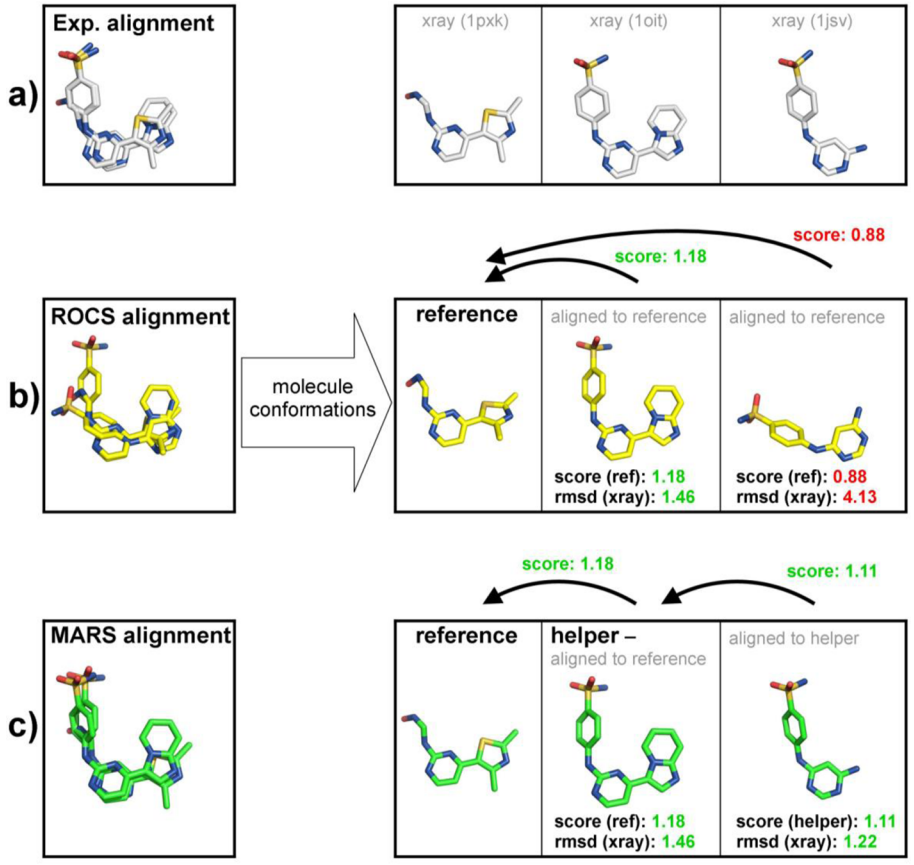


Figure 4. 3D alignment of four inhibitors of cdk2. a) depicts ligands extracted from a structure-based alignment of three protein crystal structures of cdk2 (1pxk, 1oit, 1jsv). b) shows the ROCS-generated 3D alignment, whereas the MARS-generated alignment is shown in c).

Table 2. Number of Correctly Aligned Ligands[a]

		rigid					implicit flexible				
		MARS			ROCS		MARS			ROCS	
data set	#	*shallow*	*deep*	*direct*	max	mean	*shallow*	*deep*	*direct*	max	mean
cdk2	57	39	41	34	36	21.0	30	27	26	27	15.1
p38	13	9	9	6	9	4.8	7	5	5	6	4.2
ESR1	13	7	7	7	7	6.4	5	8	5	6	4.7
DHFR	12	12	12	11	11	8.3	8	7	6	10	5.3
Fxa_diverse	8	7	6	6	8	5.6	5	3	3	4	3.0
Fxa_focused	11	11	11	11	11	10.7	8	9	9	8	5.2
∑	114	85	86	75	82	56.8	63	59	54	61	37.5

[a]A compound is considered to be correctly aligned if the rmsd to the experimentally determined conformation is ≤2 Å. Molecules have been used either in their bioactive conformation ("rigid", left side of table) or conformer libraries have been used as input ("implicit flexible", right side of table).

even in 85 (74.6%) and 86 (75.4%) out of 114 cases, which even outperforms the ROCS-generated alignments with the best-possible reference. The number of correctly aligned structures is reduced when the alignment is done based on a conformational library for the molecules in the data set (right part of the table). Nevertheless, also here all three MARS algorithms outperform the average alignment using ROCS, again reflecting that usage of pairwise similarities improves multiple alignments.

(3). Focusing the Alignment on Most Predictable Structures. Like the publication by Arena,[19] we evaluated the outcome of multiple ligand alignments as follows: Molecules below a ComboScore similarity cutoff of 1.1 are excluded from the final alignment. The multiple alignment is then considered to be correct if a) at least 50% of the molecules are present in the final alignment and b) the average rmsd of the final alignment to the experimental alignment is ≤2.0 Å. Utilizing the bioactive conformations as a starting point for a rigid MARS alignment, more than 50% of the compounds from each data set appeared in the final output in 5 out of 6 cases (fulfilling the criteria for a successful alignment) when using the *deep* and *shallow* algorithm (see Table 3). In contrast, the *direct* algorithm only provides correct alignments (i.e., ≥50% correctly aligned compounds) in 2 out of 6 cases. The number of correct alignments is slightly reduced when the alignment is done based on an a conformational library for the molecules in

dx.doi.org/10.1021/ci3000369 | J. Chem. Inf. Model. 2012, 52, 2022−2030

Table 3. MARS Alignment Results Using Molecules in Their Bioactive Conformations - with a Cutoff Similarity Score >1.1[a]

		rigid %output			rigid rmsd			rigid "passed"		
Data set	#	*Shallow*	*Deep*	*Direct*	*Shallow*	*Deep*	*Direct*	*Shallow*	*Deep*	*Direct*
cdk2	57	86	58	23	1.4	2.0	0.8	1	1	0
p38	13	31	31	31	0.6	0.6	0.6	0	0	0
ESR1	13	54	54	46	0.4	0.4	0.4	1	1	0
DHFR	12	58	58	58	0.3	0.5	0.5	1	1	1
Fxa_diverse	8	50	50	38	0.7	0.7	0.3	1	1	0
Fxa_focused	11	91	91	91	0.4	0.4	0.4	1	1	1

[a]Å compound is considered to be correctly aligned if the average rmsd over all molecules of a data set to the experimentally determined conformation is ≤2 Å.

Table 4. MARS Alignment Results for Molecules Using Conformational Libraries for Each Molecule Generated with OMEGA (Implicit Flexible) − with a Cutoff Score >1.1 Å[a]

		implicit flexible %output			implicit flexible rmsd			implicit flexible "passed"		
Data set	#	*Shallow*	*Deep*	*Direct*	*Shallow*	*Deep*	*Direct*	*Shallow*	*Deep*	*Direct*
cdk2	57	88	70	30	4.0	5.3	5.1	0	0	0
p38	13	31	38	31	0.7	0.9	0.7	0	0	0
ESR1	13	77	54	46	3.7	3.0	4.9	0	0	0
DHFR	12	58	92	58	1.8	1.6	2.2	1	1	0
Fxa_diverse	8	50	50	25	1.3	1.4	0.8	1	1	0
Fxa_focused	11	91	100	91	1.3	1.7	1.2	1	1	1

[a]A compound is considered to be correctly aligned if the average rmsd over all molecules of a data set to the experimentally determined conformation is ≤2 Å.

the data set (see Table 4). The *shallow* and *deep* algorithms provide correct alignments in 3 out of 6 cases. Notably, for the p38 data set, all compounds which appeared in the output alignment were correctly aligned (i.e., average rmsd ≤2 Å). However, in both the rigid and flexible case, less than 50% of the ligands appeared in the output alignment. We will show in the following application study (*3D clustering of the p38 data set*) that this is not due to weakness of the MARS algorithms but due to the diversity of the p38 data set. In summary, the fact that the *deep* and *shallow* algorithms provide better alignments than the *direct* algorithm again demonstrates that usage of pairwise similarities to guide the superimposition of all molecules in the data set improves the quality of multiple alignments.

(4). 3D Clustering of the p38 Data Set. An additional benefit of the MARS algorithms is that a matrix calculated by MARS can be used for clustering a set of superimposed compounds for the identification of outliers or compounds with different binding modes. The results can be used for analyzing the general homogenity or diversity of a data set. As an application, we inspected the p38 data set, for which MARS was not able to generate a correct alignment (for ≥50% of the molecules) in the previous validation study (*focusing the alignment on most predictable structures*), even if the ligands were aligned rigidly in their bioactive conformations. As mentioned above, the fact that only 31% (i.e., 4 out of 13 p38 ligands) appeared in the final output alignments of the *shallow* algorithm indicates a high diversity of the p38 data set. To check this hypothesis, we performed a hierarchical clustering (using a

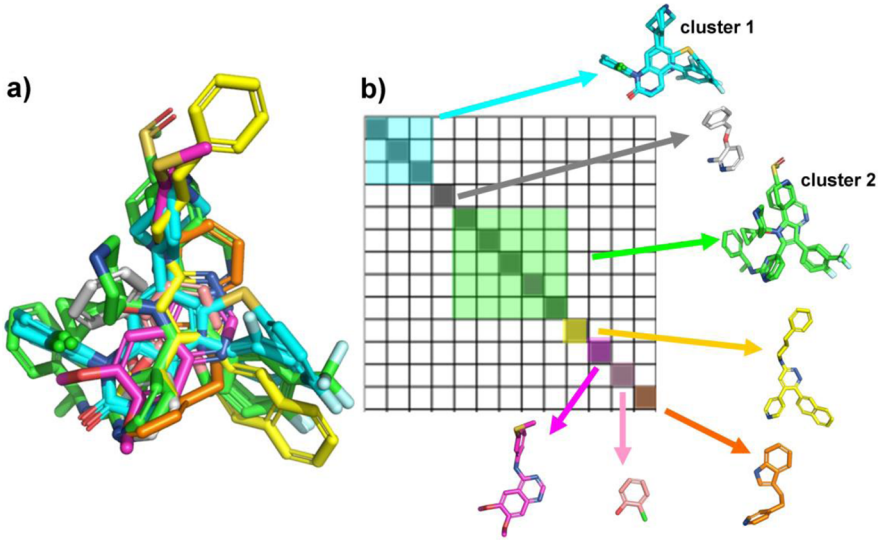


Figure 5. a) MARS alignment of 13 inhibitors of p38, which was generated by the *shallow* algorithm. b) pairwise ROCS similarity matrix of the molecules based on their aligned coordinates, after hierarchical clustering. Molecules are colored according to their cluster affiliations.

ComboScore cutoff value of 0.9) on the matrix of pairwise ROCS similarities for the 3D alignment generated by the *shallow* algorithm (where the ligands were treated as flexible). Figure 5 shows this 3D alignment and the correlation matrix, which is colored according to the ligands cluster affiliations. Analysis of these clusters confirms that the p38 data set is relatively diverse: although we applied a loose ComboScore cutoff value of 0.9, the data set was divided into two clusters (Figure 5: cluster 1 highlighted blue with three compounds and cluster 2 highlighted green with five compounds) and 5 singletons. Thus, with this diverse 3D similarity distribution, it is not possible to provide a correct alignment for ≥50% of the compounds. Consequently, the failure of MARS is not due to limitations in the algorithms but due to the large diversity in the ROCS 3D similarity descriptor space for these compounds. Encouragingly, visual inspection of the chemical structures in Figure 5b confirms that the 3D clustering approach provides (1) a reasonable identification of structural outliers (shown in gray, orange, yellow, pink, and magenta) and (2) provides consistent superimpositions within different clusters (clusters 1 and 2), i.e. compounds with different binding modes. In fact, we compared the alignments of the individual clusters with the corresponding experimental alignments and found that for both clusters, the average rmsd values were ≤2.0 Å (cluster 1: 1.32 Å, cluster 2: 1.26 Å). This example shows that the 3D clustering approach is useful for finding similar and dissimilar molecule sets. This information can be used for the identification of compounds with (dis-)similar binding modes or as a starting point for the generation of specific pharmacophore or 3D QSAR models.

CONCLUSION

We have presented MARS (Multiple Alignments by ROCS-based Similarity), an approach which generates multiple ligand alignments based on the pairwise comparison of all molecules within the data set using ROCS. Each pairwise alignment is scored, and the results are captured in a matrix. The ideal superimposition of the compounds in the set are then identified by the analysis of the score matrix building stepwise a superimposition of all molecules. The evaluation of MARS using six diverse molecule sets reveals that taking into account 3D similarities between all molecule pairs results in better alignments than considering the similarities to only one reference molecule.

In many (multiple) ligand-based alignment approaches, the user has to provide a reference molecule in a defined 3D conformation. By default, this is not necessary in MARS, since it automatically identifies the best reference, which provides the best-scoring overall alignment. Thus, MARS can be used for hypothesis generation without prior knowledge or in-depth analysis of a ligand data set. As option, it is also possible to define a reference molecule, which might be beneficial, if a bioactive conformation is known, for example from experimental studies, such as X-ray crystallography or NMR studies.[23]

Alignments generated by MARS can be used for different aspects of computer-aided drug design, for example the generation of *3D pharmacophore* models[1−5,7] for molecules with different scaffolds or *3D QSAR* models,[24] such as CoMFA[25] or CoMSIA[26] for molecules from one chemical series. Since an extension of the *shallow* algorithm allows the addition of new molecules to an already existing alignment, it is possible to regularly update existing ligand alignments and 3D QSAR models with new compounds at different stages of a drug discovery project, e.g. resulting from the next synthesis cycle. Furthermore, it is possible to align virtual molecules (e.g., synthesis proposals) onto an existing alignment and evaluate them by an available model for the priorization of chemical synthesis. In another study,[27] we generated a MARS alignment of 63 P2Y12 antagonists, using the conformer of the initial lead structure taken from a receptor−ligand model as structural reference. Based on the multiple alignment of the data set generated by the *shallow* algorithm, we were able to generate statistically good 3D QSAR models. These 3D QSAR models were used to reveal essential interactions for activity and to guide further synthesis.

Finally, generation of MARS alignments for a diverse set of molecules, followed by a 3D clustering has proven to be useful to detect relationships between compounds which are not obvious from the 2D-analysis. For example, a 3D similarity can be detected between an in-house compound and competitor compounds with available SAR on the target of interest. This

dx.doi.org/10.1021/ci3000369 | J. Chem. Inf. Model. 2012, 52, 2022−2030

can directly support the chemical optimization of the in-house series by providing novel synthesis ideas. Furthermore, it was shown that this approach can be used for the identification of different binding modes. These could be used as starting point for the generation of "binding-mode"-specific pharmacophores.

In summary, the broad applicability of MARS (e.g., for pharmacophore modeling, 3D QSAR, 3D clustering, alignment generation for rescaffolding) has made MARS an important tool in ligand-based drug design approaches for the expert and nonexpert within our company.

■ ASSOCIATED CONTENT

Supporting Information

Data sets and alignment results. SDF files containing the crystallographic ligand superimpositions and MARS alignment results according to Table 2. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49-69-305-12636. Fax: +49-69-331399. E-mail: Andreas.Evers@sanofi.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are thankful to Karl-Heinz Baringhaus for several helpful discussions.

■ REFERENCES

(1) *Catalyst*; Accelrys Software Inc.: San Diego, CA, 92121.

(2) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. 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.* **2006**, *20*, 647−671.

(3) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: a novel approach to pharmacophore modeling and 3D database searching. *Chem. Biol. Drug. Des.* **2006**, *67*, 370−372.

(4) 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.

(5) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(6) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(7) Poptodorov, K.; Luu, T.; Hoffmann, R. D. Pharmacophore Model Generation Software Tools. In *Pharmacophores and Pharmacophore Searches*; Langer, T., Hoffmann, R., Eds.; Wiley-VCH: Weinheim, Germany, 2006; pp 17−47.

(8) Kearsley, S.; Smith, G. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.

(9) Klebe, G.; Mietzner, T.; Weber, F. Different approaches toward an automatic structural alignment of drug molecules: applications to sterol mimics, thrombin and thermolysin inhibitors. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 751−778.

(10) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.

(11) *ROCS*, version 2.4.1; OpenEye Scientific Software: Santa Fe, NM.

(12) 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.

(13) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(14) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric accuracy of three-dimensional molecular overlays. *J. Chem. Inf. Model.* **2006**, *46*, 1996−2002.

(15) Feher, M.; Schmidt, J. M. Multiple flexible alignment with SEAL: a study of molecules acting on the colchicine binding site. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 495−502.

(16) Cho, S. J.; Sun, Y. FLAME: A Program to Flexibly Align Molecules. *J. Chem. Inf. Model.* **2005**, *46*, 298−306.

(17) Anghelescu, A. V.; DeLisle, R. K.; Lowrie, J. F.; Klon, A. E.; Xie, X.; Diller, D. J. Technique for Generating Three-Dimensional Alignments of Multiple Ligands from One-Dimensional Alignments. *J. Chem. Inf. Model.* **2008**, *48*, 1041−1054.

(18) Korb, O.; Monecke, P.; Hessler, G.; Stützle, T.; Exner, T. E. pharmACOphore: multiple flexible ligand alignment based on ant colony optimization. *J. Chem. Inf. Model.* **2010**, *50*, 1669−1681.

(19) Jones, G.; Gao, Y.; Sage, C. R. Elucidating molecular overlays from pairwise alignments using a genetic algorithm. *J. Chem. Inf. Model.* **2009**, *49*, 1847−1855.

(20) *OMEGA*, version 2.1; OpenEye Scientific Software: Santa Fe, NM.

(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(22) *Molecular Operating Environment (MOE)*, 2010.10; Chemical Computing Group Inc.: Montreal, Canada.

(23) Orts, J.; Tuma, J.; Reese, M.; Grimm, S. K.; Monecke, P.; Bartoschek, S.; Schiffer, A.; Wendt, K. U.; Griesinger, C.; Carlomagno, T. Crystallography-independent determination of ligand binding modes. *Angew. Chem., Int. Ed. Engl.* **2008**, *47*, 7736−7740.

(24) Kubinyi, H.; Folkers, G.; Martin, Y. C. *3D Qsar in Drug Design: Recent Advances*; Springer: Berlin, Heidelberg, Germany, 1998.

(25) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **2011**, *110*, 5959−5967.

(26) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(27) Zech, G.; Hessler, G.; Evers, A.; Weiss, T.; Florian, P.; Just, M.; Czech, J.; Czechtitzky, W.; Görlitzer, J.; Ruf, S.; Kohlmann, M; Nazare, M. manuscript has been submitted.

(28) *MATLAB*, version 7; The MathWorks Inc.: Natick, MA.