J Comput Aided Mol Des (2011) 25:931–945
DOI 10.1007/s10822-011-9473-6

De novo design by pharmacophore-based searches in fragment spaces

Tobias Lippert · Tanja Schulz-Gasch · Olivier Roche · Wolfgang Guba · Matthias Rarey

Received: 18 April 2011 / Accepted: 5 September 2011 / Published online: 16 September 2011

Abstract De novo ligand design supports the search for novel molecular scaffolds in medicinal chemistry projects. This search can either be based on structural information of the targeted active site (structure-based approach) or on similarity to known binders (ligand-based approach). In the absence of structural information on the target, pharmacophores provide a way to find topologically novel scaffolds. Fragment spaces have proven to be a valuable source for molecular structures in de novo design that are both diverse and synthetically accessible. They also offer a simple way to formulate custom chemical spaces. We have implemented a new method which stochastically constructs new molecules from fragment spaces under consideration of a three dimensional pharmacophore. The program has been tested on several published pharmacophores and is shown to be able to reproduce scaffold hops from the literature, which resulted in new chemical entities.

Keywords De novo design · Pharmacophores · Fragment-based design

Introduction

De novo design has been covered extensively in many recent reviews [1–3]. The existing methods can be categorized roughly by their methodological approaches, especially by how molecues are generated and how they are assessed. The molecule generation can either be atom-based or fragment-based. Early programs grew molecules one atom at a time. This approach enables programs to generate all possible molecules, which however includes many molecules which are synthetically not accessible or have other undesired properties [4]. More recent methods use molecular fragments and connect them by defined rules. This greatly narrows the number of possible molecules down to a synthetically more feasible set. A careful selection of the utilized fragments also allows to influence the properties of the resulting molecules in a beneficial way. Structural motifs which are undesired can be excluded, while motifs which represent common building blocks of drugs can be included. Furthermore, there is evidence that the resulting molecules are more drug-like, if the fragments have been obtained by cleaving drug molecules [5–7]. By encoding standard synthetic reactions in the connection rules, the generated molecules may also be easier to synthesize [8]. Another strength of this approach is the ability to create problem-specific fragment spaces very easily. Among others, the programs GenStar [9] and CONCEPTS [10] implement the atom-based strategy, and LUDI [11], SkelGen [12], FlexNovo [13], BREED [14] and E-Novo [15] use the fragment-based approach.

The basis for the assessment of the generated molecules is another characteristic of the programs. The generated molecules can either be assessed with structure-based or ligand-based methods. Structure-based methods try to make predictions on the affinity of a molecule by interaction-based scoring functions. Published methods that pursue this approach include LUDI [11], SkelGen [12] and FlexNovo [13]. In contrast to these, ligand-based methods do not rely on protein structures directly, but use similarity

T. Lippert · M. Rarey (✉)
Center for Bioinformatics, University of Hamburg,
Bundesstr. 43, 20146 Hamburg, Germany
e-mail: rarey@zbh.uni-hamburg.de

T. Schulz-Gasch · O. Roche · W. Guba
Pharmaceutical Division, F. Hoffmann-La Roche Ltd,
4070 Basel, Switzerland
e-mail: tanja.schulz-gasch@roche.com

to one or more known active ligands as a target function instead. Usually, the known actives are used as templates to which the resulting molecules are compared. The comparison may either be based on 2D- or 3D-properties of the molecules.

2D properties are based on a comparison solely on the topological description of the molecules. They are easier and faster to calculate since they can neglect the conformational freedom and 3D-alignment of the molecules. Thus, they excel in their speed and can cover a much larger portion of the search space in the same time than their 3D-considering counterparts. 2D similarity is employed by programs such as Flux [16], which uses the proprietary Daylight fingerprint along with the Ghose and Crippen descriptor [17], and F-Trees-FS [18], which maps molecular fragments onto each other and scores their similarity.

3D properties are based on the spatial arrangement of the electrostatic properties and shape of the molecules. In order to compare molecules with these measures, their conformational freedom has to be considered, which makes the calculations computationally much more expensive. The electrostatic properties can be described by a large variety of different approaches. For example, Tripos's LeapFrog program[1] scores ligands with a molecular field analysis, and NovoFLAP [19], employs the external program ROCS[2] to score molecules with respect to the shape overlap and electronic similarity. But there are also other approaches [21] which use distance-geometry to define molecular similarity such as the CATS3D-descriptor [20].

A special case for 3D-based similarity measures is the use of a pharmacophore. Pharmacophores abstract the properties which are needed for activity from the provided template molecules. In contrast to the majority of the published similarity-based de novo methods which employ the comparison to a single template molecule, this approach allows to avoid the depencency on concrete molecules, and thus makes the identification of new topologies more feasible.

In this article we discuss a novel pharmacophore-based de novo design approach named Qsearch. The input consists of a three dimensional pharmacophore and a fragment space. Qsearch then uses a simulated annealing algorithm to search ligands in the input fragment space which fulfill the pharmacophore. The novelty of this de novo design approach is the usage of three dimensional pharmacophores in a purely ligand-based mode combined with fragment spaces as a source for novel molecules. It also features an innovative undecoration step, which extends the set of possible molecules in a meaningful way. We put much emphasis on the speed of the method, which—depending on the complexity of the pharmacophore—allows to test virtual molecules in the order of million entities over night on a single workstation.

Comparable approaches

Several comparable approaches have been described in the literature. The two most similar are SkelGen [12] and NovoFLAP [19]. SkelGen creates molecular scaffolds by connecting carbon-skeletons and decorating them afterwards with hetero-atoms such that they are favorable with respect to the scoring function. The search is driven by a simulated annealing procedure, while the objective function can either be calculated on basis of a receptor structure [22] or with a pharmacophore-like description of active ligands [23]. The main difference to Qsearch is the generation of molecular structures which does not mimic standard-reactions as the RECAP-based approaches do, and hence will probably yield chemically less valid structures.

NovoFLAP uses more sophisticated proprietary connection rules than SkelGen to generate novel molecules from fragments. The molecule generation is delegated to the external Program EA-Inventor[3] which uses a population-based evolutionary algorithm to search for well-scoring molecules. The fitness value for a molecule is calculated in a two-step algorithm. In the first step, the external programs Omega and ROCS (both OpenEye) are employed to generate an ensemble of conformations for a molecule and to superimpose them onto a reference ligand. Then, the best superposition in terms of volume overlap is used to calculate a fitness value. The value comprises of the volume overlap [24] and the chemical similarity the the reference ligand as calculated by the proprietary program Triphic. The main difference of Qsearch to NovoFLAP is the usage of a user-supplied pharmacophore as scoring function instead of automatically generated scores.

Methods

Overview

Qsearch navigates the space of possible molecules by simulated annealing [25], which is an established method to deal with large and rugged energy landscapes [26]. The overall workflow is depicted in Fig. 1. It starts out with creating random molecules from the provided fragments

[1] LeapFrog, Tripos L.P., St.Louis, MO (USA), http://www.tripos.com.

[2] ROCS, OpenEye Scientific Software Inc., Santa Fe, NM (USA), http://www.eyesopen.com/rocs.

[3] EA-Inventor, Tripos L.P., St.Louis, MO (USA), http://www.tripos.com.

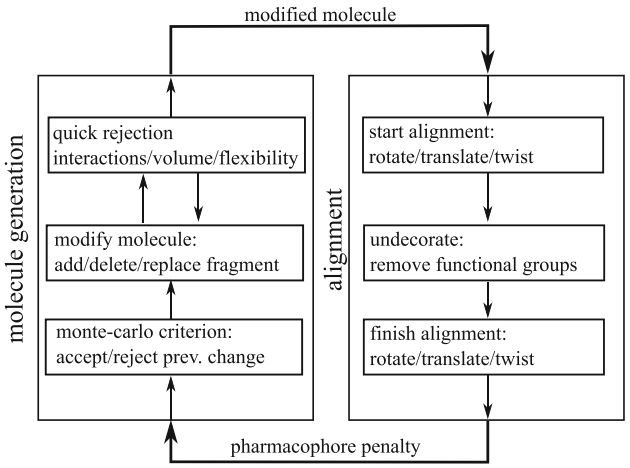

Fig. 1 Workflow of Qsearch

until a molecule is encountered which meets a set of pharmacophore-dependent minimal criteria. (cf. Molecule generation) Then, Qsearch enters a loop of continuous molecular modifications. At every step, a new modification is introduced into the current molecule and the result is tested against the pharmacophore (cf. Pharmacophore penalties). Depending on the score difference, the modification is either accepted or rejected. If it is accepted, the next step is conducted with the modified molecule, else its predecessor is taken. If a molecule which satisfies all constraints of the pharmacophore, is encountered, it is added to the set of solutions. In the following, a detailed view on the individual components of Qsearch will be given.

Fragment spaces

A fragment space consists of a set of molecular fragments and a set of rules which defines how the fragments may be connected. The attachment points of the fragments are denoted with special link atoms. Fragment spaces have been introduced as a way to restrict the huge chemical universe to a more relevant part which is synthetically more accessible and exhibits more drug-like properties. They are either created by cleaving molecule libraries with desired properties by means of retrosynthetic decomposition rules or by merging combinatorial libraries. A corresponding set of rules is then used to reconnect the fragments and is usually modeled analog to common synthetic pathways. The RECAP rules [8] were the first published set of such retrosynthetic decomposition rules, but they have been used and enhanced by many other groups to fit their specific needs. Custom fragment spaces have been created by cleaving the COBRA library [16], the World Drug Index 2004 and Medchem03 databases [7], and the Cambridge Structural Database (CSD) [27].

Molecule generation

The molecular generation process in Qsearch is based on the fragment space concept. New molecules are generated by modifying an existing molecule by either adding a fragment at a free link atom, removing a terminal fragment or replacing a fragment with another one which has compatible links. Whether the modifications are accepted, depends on how well the resulting molecule satisfies the pharmacophore (cf. Pharmacophore penalties) and on the Metropolis criterion of the simulated annealing algorithm.

The fragment combination procedure is divided in three steps:

1. Merge the new fragment into the existing molecule. The ensuing bond type can be specified in the fragment space definition.
2. Recalculate the topology-dependent chemical properties (delocalized zones, functional groups, stereo centers, etc.) of the molecule.
3. Conduct a geometric correction. The newly attached fragment is moved to fit the formed bond. Also, a planar (E-)conformation is enforced, if the bond is part of a delocalized system.

Prior to the application of the regular simulated annealing acceptance criterion, a quick-rejection step immediately rejects states representing molecules which are too large (topology-based volume), too flexible (number of continuous rotatable bonds), or do not have enough interactions to fulfill the pharmacophore. If a molecule is found to be too small, the algorithm immediately adds a fragment to satisfy the volume constraints. The calculation of the quick-rejection criteria can be conducted on the basis of a molecule's topology and is conformation-independent. Hence these states may be skipped without going through the relatively time-consuming alignment to the pharmacophore.

Pharmacophore model

Most implementations of the pharmacophore model share a consensus on the types of electronic features [28], but lack a common model in their geometric placement and the precise assignment to atoms and functional groups [29]. Qsearch implements the annotation scheme and geometric properties of the frequently applied PCHD model from the MOE software suite.[4] The pharmacophore features and their molecular equivalences are listed in Table 1. All pharmacophore features are modeled with spheres which are annotated with the type of interaction they represent. A

[4] MOE software suite, Chemical Computing Group, Montreal (Canada), http://www.chemcomp.com

Table 1 Pharmacophore features and their molecular correspondences

MOE name	Molecular correspondence
Acc	Polar atom with free electron pairs
Acc2	Free electron pair of "Acc" atom
Aro	A fully conjugated ring
Aro2	Above and below ring
Ani	"Acc" and negative charge
Cat	"Don" and positive charge
Don	Polar atom with hydrogens
Don2	Hydrogen atom of "Don"
Hyd	Any atom except "Acc" and "Don"
ML	As "Acc"
ML2	As "Acc2"

pharmacophore feature is considered to be fulfilled if a corresponding molecular motif lies within that sphere. Additional constraints may be posed to enforce that a set of features is fulfilled by a single atom or functional group. As shown in Fig. 2, in case no constraints are used, two features which are intended to model directionality may be satisfied by independent atoms or groups thus leading to counter-intuitive results.

Spatial constraints are used to define the shape that the ligands may have. They are modeled with a set of spheres which are usually placed on the template's heavy atoms.

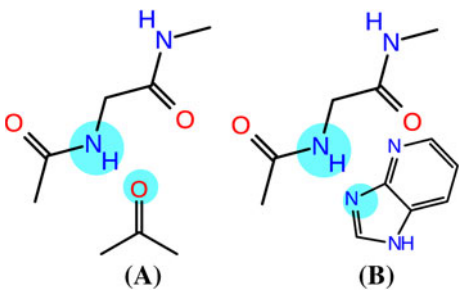

Fig. 2 A pharmacophore for an interaction with a protein backbone N–H. The pharmacophore describes the position of an acceptor atom and the direction of a free electron pair. **a**, **b** Show two molecular fragments which fit the pharmacophore. **a** Is an expected fragment, **b** is a fragment whose match is unintended since the two pharmacophore features are satisfied by different atoms of the molecule. To unambiguously describe the first option, both pharmacophore features can be linked, so that they must be satisfied by one single atom

The spatial constraints are fulfilled if no heavy atom of a molecule lies outside of them.

Interactions

The interaction model of Qsearch is based on a point description of hydrogen bond donors, acceptors, and aromatic groups. The interaction directions are based on the VSEPR [30] geometries of the atoms.

For hydrogen bond donors, one interaction is created for each hydrogen atom that is connected to a polar atom. The start point of the interaction is placed on the polar atom, and its end point on the hydrogen atom. Hydrogen bond acceptor interactions are assigned to each free electron pair of a polar heavy atom. Their start points are placed on the polar atom as well, while the end points are placed on the positions of the free electron pairs. Also, for each aromatic ring, two interactions are created. The start point of both lies on the geometric center of the ring, and end points are created above and below the ring center (Fig. 3).

Alignment to the pharmacophore

Every molecule which is generated and passes the quick rejection criteria (cf. Molecule generation), is subject to an analysis of its conformational space to determine how well it is able to fulfill the pharmacophore. Again, simulated annealing is used to optimize the molecule's conformation and orientation. Qsearch starts the search process with a conformation close to that of the previously generated molecule. This is done by preserving the positions of the unchanged fragments and recalculating coordinates only for the new atoms. In case a central fragment is changed and the molecule becomes disconnected, the largest component keeps its coordinates, and the coordinates for the remaining components are recalculated. After a start conformation has been generated, the conformational space of the molecule is explored iteratively with a simulated annealing optimization. The workflow of the algorithm is

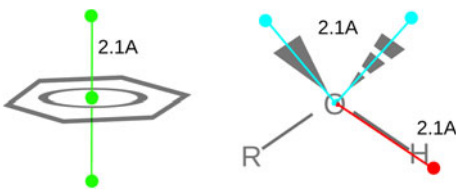

Fig. 3 Placement of interaction points. *Left*: Aromatic points (*green*) are placed in the plane of the ring and along its normals. *Right*: Donor points (*red*) are placed on the heavy atom and in direction of the hydrogen atoms. Acceptor points (*cyan*) are placed on the heavy atom and in direction of the free electron pairs. All distances are 2.1A to provide compatibility with the MOE interaction model

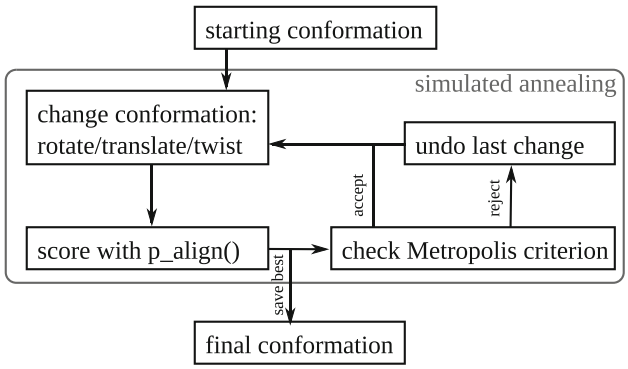

Fig. 4 Workflow of the molecular alignment routine. Qsearch uses simulated annealing to explore the conformational space of the generated molecules. The algorithm terminates after a defined number of cycles in the workflow

exemplified in Fig. 4. Qsearch runs the optimization four times for each molecule and chooses the one with the best score.[5]

In course of the optimization, the torsion angles and the overall molecule transformation are randomly adjusted within user-defined constraints. We use a default maximum angle change of 20° for changes to the torsion angles and the orientation and a maximum translation of 0.4 Å. For the simulated annealing procedure, we use a start temperature of 7.0, a cooling factor of 0.988 and 1,000 steps.

The penalty for the conformation at each step is calculated with $p_{\text{align}}(M, P)$ which is defined as follows.

$$p_{\text{align}}(M,P) = p_{\text{pharm}}(M,P) + p_{\text{clash}}(M) \tag{1}$$

p_{pharm} (M, P) determines how well the molecule's conformation obeys to the pharmacophore and $p_{\text{clash}}(M)$ prevents the molecule from collapsing:

$$p_{\text{clash}}(M) = w \cdot \sum_{a,b \in M, a \neq b} f_{\text{clash}}(a,b) \tag{2}$$

$$f_{\text{clash}}(a,b) = \begin{cases} 0.0 & \text{if } d(a,b) > r(a) + r(b) \\ \left(\frac{r(a)+r(b)}{d(a,b)}\right)^2 & \text{else} \end{cases} \tag{3}$$

Atom pairs a, b which are separated by only one or two bonds are ignored as well as pairs which belong to the same ringsystem. $r(a)$ returns the van-der-Waals radius of an atom a, and $d(a, b)$ returns the distance between the centers of two atoms a and b. w is a weight factor which is used to scale the clash penalty against the other terms.

Pharmacophore penalties

Qsearch uses a numerical function $p_{\text{pharm}}(M, P)$ which penalizes geometric deviations of a molecule to a pharmacophore to guide the simulated annealing optimization towards well-scoring conformations. Analogous to the pharmacophore definition, the penalty formula consists of a steric and an electronic part:

$$p_{\text{pharm}}(M,P) = p_{\text{st}}(M, st(P)) + p_{\text{el}}(ia(M), el(P)) \tag{4}$$

$st(P)$ returns the steric constraints of a pharmacophore P as a set of spheres S, $ia(M)$ returns the interactions of a molecule M and $el(P)$ returns the electronic features of the pharmacophore. All atoms whose center lies outside of the shape contribute to the steric penalty as follows

$$p_{\text{st}}(M,S) = \sum_{a \in M} sdist^2(a,S) \tag{5}$$

where $sdist^2(a,S)$ denotes the squared distance of the center of an atom a to the shape S.

Electronic features contribute to the penalty of the pharmacophore with the function $p_{\text{el}}(I, E)$ which takes the interactions of a molecule and the electronic features of the pharmacophore as parameters. The minimal penalty of every electronic feature with any of the molecule's interactions is summed and results in the overall penalty for the electronic part.

$$p_{\text{el}}(I,E) = \sum_{e \in E} \min_{i \in I} \{p_{\text{ia}}(i,e)\} \tag{6}$$

The actual penalty is then calculated as follows

$$p_{ia}(i,e) \begin{cases} \left(\frac{d(e,i)}{r(e)}\right)^2 & \text{if } comp(i,e) = 1 \\ \infty & \text{else} \end{cases} \tag{7}$$

where $d(e, i)$ is the distance from the electronic feature e to the center of the interaction i. $comp(p, e)$ is a boolean function which indicates whether a point of an interaction can potentially satisfy the constraint of an electronic feature. $r(e)$ results in the radius of an electronic feature. The square in the calculation is used for the sake of computational speed as well as ensuring an adequately steep energy landscape for the optimization.

Undecorating step

Most of the fragments which result from cleaving drug molecules are decorated with functional groups and other structural elements such as carboxylic acids or amines which have been added to enhance their physiochemical and pharmacokinetic properties. Since our method is supposed to suggest new scaffolds, the generated topologies are much more important than their complete decoration.

Qsearch removes functional groups from molecules if they are not needed to fulfill the pharmacophore. This is done after fifty percent of the alignment procedure to ensure that an initial alignment exists and the important functional groups are oriented such that they are near their

[5] Since these calculations are independent from each other, they can be easily parallelized and take advantage of multi-core CPUs.

corresponding pharmacophore features. This has the additional advantage of improving the molecule's ability to stay within the steric constraints. Also, hetero atoms in heterocycles are replaced by carbon if they stay aromatic.

Fragment growing

Another common application for de novo design is the partial modification of an existing molecule. This may become necessary to improve its physico-chemical properties or to circumvent patents [27]. Qsearch supports this process by a "fragment growing" mode. In this mode, a start fragment is supplied by the user, which is used as a seed for growing a complete molecule. The employed algorithm is a slight modification of the aforementioned one (cf. Molecule generation). The differences include:

- In the molecule modification steps, the user-provided start fragment must not be changed.
- In the molecular alignment routines, the movement of the start fragment from its original position is restricted. This is done by introducing a new term p_{RMS} into the penalty function. It is calculated by summing the squares of the distance to their original position for all heavy atoms of the start fragment.

The BRICS fragment space

For the following results, the BRICS cleavage rules and fragment spaces [6] were used. The cleavage rules are a modified superset of the RECAP rules. Modifications include a better differentiation between cyclic and acyclic environments and the explicit treatment of activated and inactivated heterocycles. Altogether, the fragment space provides 16 different linkers and 47 rules by which these may be combined (Fig. 5).

There are three sets of fragments which have been provided along with the publication and which are based on cleaving large sets of publicly available molecule catalogs. The sets differ in their size and have been named BRICS_4k, BRICS_8k and BRICS_20k accordingly. The selection of the included fragments has been optimized for their ability to reproduce known drugs from the World Drug Index.[6] Hence, many of the fragments feature the planar aromatic ringsystems which are commonly found in drug molecules. For example, in the BRICS_4k fragment space, 85% of all ringsystems belong to this group. Also, the resulting molecules are usually linear due to the fact that most fragments have one or two linking positions. This reduces the geometric diversity of the fragment space, but increases the drug-likeness of the results.

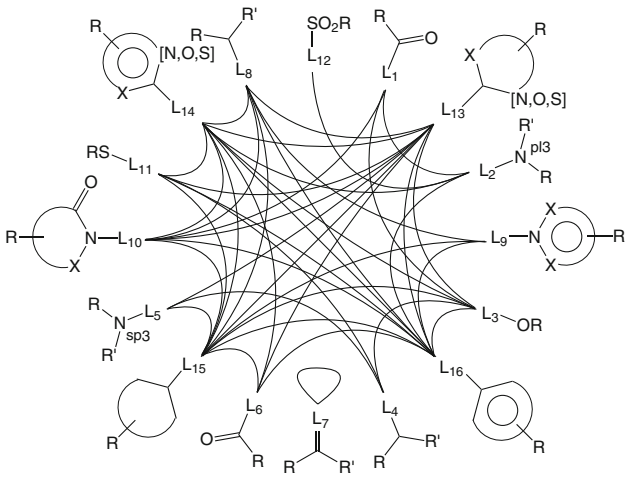

Fig. 5 Connection rules for the BRICS fragment space. L1–L16 are the linkers in their chemical environments. A *line* connects two linkers if they may be combined (Figure taken from [6].)

In order to compensate for some of the introduced bias and to increase the geometric and chemical variability of the BRICS_4k fragment space, it is enriched with a subselection of fragments from the BRICS_20k space. We included fragments in that subselection if they were non-planar and had more than two link atoms. A fragment is considered planar if more than 90% of the variance in its coordinates can be reproduced by two linear combinations of the x, y and z coordinates. This resulted in a total of 5913 fragments.

Parameters

Typical values for the size of the molecules are 70–120% of the query volume, and three continuous rotatable bonds. For the simulated annealing algorithm, we commonly use a start temperature of 4.0, a cooling factor of .982 and 700 fragment modifications. Depending on the machine and the input, a single run takes usually <3 min.

Results

The validation of a de novo program is confronted with several problems. Other virtual methods such as docking programs can easily be validated by comparing the results with existing experimental data. Since the results of de novo programs (ideally) consist of novel molecular structures, no experimental data exists which can be used for a similar validation. Using scoring functions to estimate the binding energy of the Qsearch solutions is another option, which is available if structural information on the target

[6] The World Drug Index contains marketed and development drugs and can be licensed from Thomson, Philadelphia, PA (USA).

exists. It is however unlikely that Qsearch solutions will show a high score, since the program produces novel scaffolds rather than optimized leads. To make this a useful approach, the resulting scaffolds would have to be decorated properly in order to optimize the interactions with the active site, since Qsearch bases its calculations only on a pharmacophore. Using and assessing only the scaffolds of the solutions is not feasible either, since there is no generally applicable definition of the term "scaffold" [31] and thus no easy, computationally accessible way to classify these. Hence, we rely on visual inspection and intuition of experienced scientists with a broad medicinal chemistry knowledge.

In spite of these limitations, the results will be compared to existing drugs, and scaffold hops which have been described in the literature will be highlighted to demonstrate the general applicability of the method. In the following, Qsearch is evaluated in typical drug design scenarios for two different targets with generic and custom fragment spaces.

COX-2 inhibitors

For the first example, we used a custom fragment space to limit the diversity of the input, which makes it easier to compare the solutions to known inhibitors. The fragment space was created by shredding 128 known actives of COX-2 [32] with the BRICS cleavage rules and contained 414 fragments altogether. Since a custom fragment space was used, the undecoration step was disabled to retain its properties. The pharmacophore query was set up by using PDB structure 6cox [33]. Since COX-2 is a predominantly lipophilic and buried cavity, only two directed aromatic moieties were used in the setup. Similar unspecific pharmacophore constraints for this target have also been defined in the evaluation study of the de novo design tool SkelGen [34].[7] For the shape constraints, spheres of radius 1.8 were put on every heavy atom of the ligand. With this pharmacophore setup (cf. Fig. 6) Qsearch was run for 8 min with default parameters on a common computer[8] and obtained 100 results, which are summarized in Fig. 7. The left columns contain general scaffold depictions of known COX-2 actives [34]. The scaffolds represent topological frameworks as defined by Bemis and Murcko [35] with slight modifications: numbers in rings indicate variable sizes, and aromaticity is included if it is essential to the scaffold class. The middle columns contain Qsearch solutions of the corresponding framework plus a related known active in the right columns.

In a similar setup, we used the same pharmacophore for a search within the enriched BRICS_4k fragment space. The number of possible molecules that can be generated from this fragment space is much higher than in the previous example. This is favorable with regard to the diversity and novelty of the results, but it is also harder for the method to find valid solutions or reproduce known actives. The program ran for 16 min and resulted in roughly 500 solutions. We have searched for representatives for each of the previous example's scaffold classes and compiled them in Fig. 8. The branching sp^2 carbon in (D) could not be reconstructed exactly. Instead, solutions were found to have a geometrically similar planar nitrogen at that position. The molecules which are listed in (F) all have an additional ring, which was added by Qsearch to fulfill the volume constraints which were posed.

Interpretation

Within a very limited time span, Qsearch is able to retrieve topologies of all major classes of known actives from the custom fragment space. For example, in the top right-hand scaffold depiction of Table Fig. 7, Qsearch reproduced or almost reproduced the two known actives in the first two rows. In the following rows, examples for the novelty which may be achieved even within a restricted set of fragments are shown: Qsearch proposes an analog to an open-chain precursor of the compound with the dihydrofuran-2-one fragment, and a molecule with a ring closure which may be considered as novel (highlighted in blue). The three-dimensional orientation of the solutions is shown with four selected representatives in Fig. 6.

In the second example, the enriched BRICS fragment space was used and Qsearch was run twice as long in order to get the desired diversity in the scaffolds. The number of results is much higher than in the previous example, which is a direct result of the size of the fragment space. When a fragment is replaced by another of the same structural class in the molecule generation process, the resulting molecules will also have the same scaffold and differ only in their decoration. Unlike the custom fragment space, the BRICS space contains many more representatives for each fragment scaffold class. This results in more solutions with the same scaffold and different decorations. Nevertheless, we were able to find representatives for all of the scaffold classes. As expected, the solutions were more dissimilar to known COX-2 inhibitors, since the fragments were obtained from several different drug classes. The results contained molecules with a cis-stilbene scaffold (Fig. 8f), which was not found in the first example.

[7] Stahl et al. placed one more lipophilic sphere than we did. This was done mainly to ensure that the cavity is filled appropriately and there is no unoccupied space. This is not needed here, since Qsearch automatically rejects molecules which are too small (cf. Methods).

[8] 2008 Intel Core2 notebook with two cores at 2 GHz

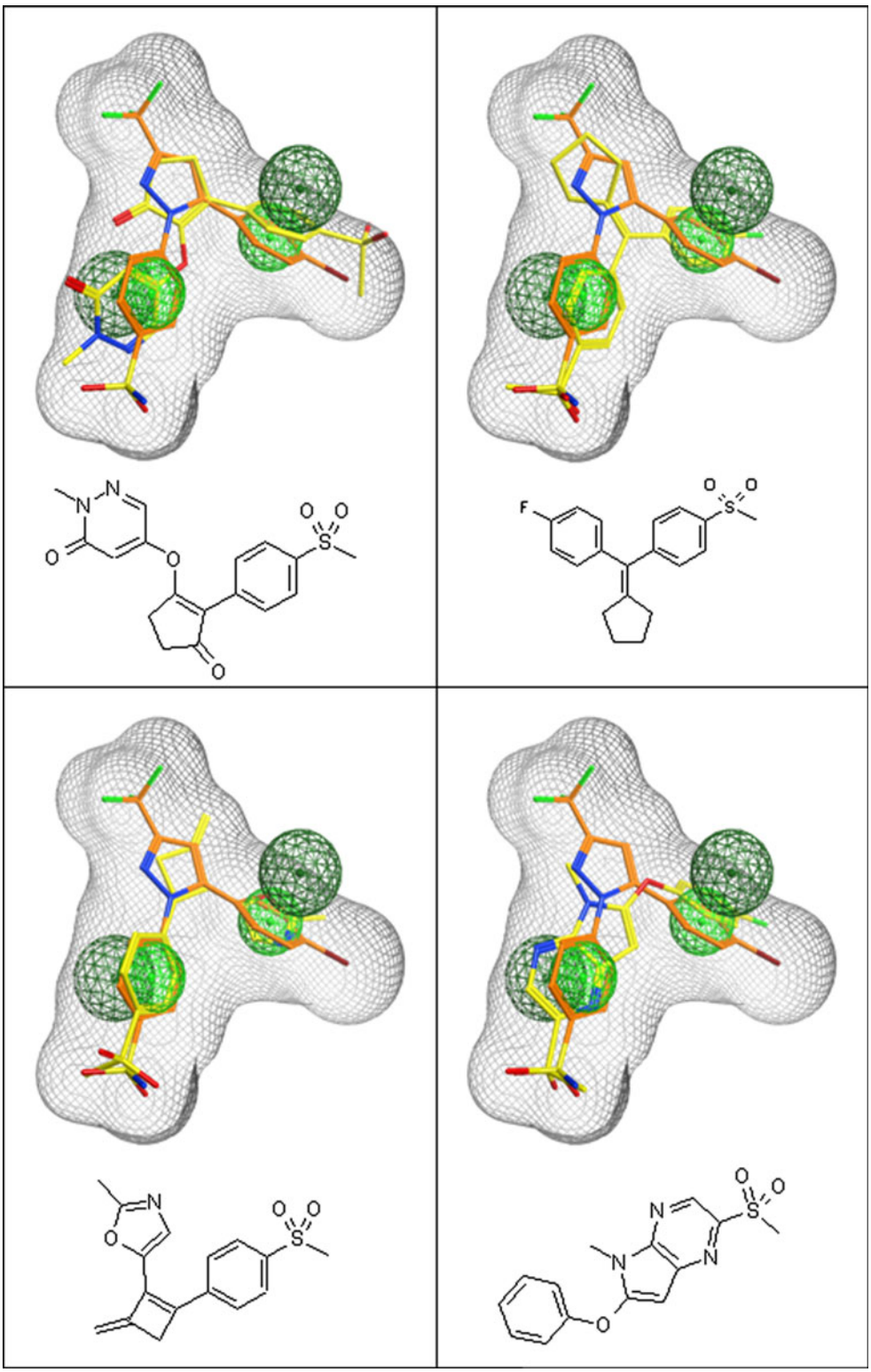

Fig. 6 Four selected Qsearch hits (*yellow*) from Fig. 7 superposed with the query pharmacophore definition and the original template molecule (*orange*) for setting up the pharmacophore from PDB entry 6cox

BCR-ABL tyrosine kinase inhibitors

In the second scenario, we investigated the ability of Qsearch to provide new ideas for BCR-ABL tyrosine kinase inhibitors in two different setups. The query was built analogous to a BCR-ABL kinase inhibitor pharmacophore from the literature, which was proven to be able to identify all imatinib molecules in a virtual PDB screening [36]. It consists of four aromatic features and two directed hydrogen bond acceptors, whose placement is sketched in

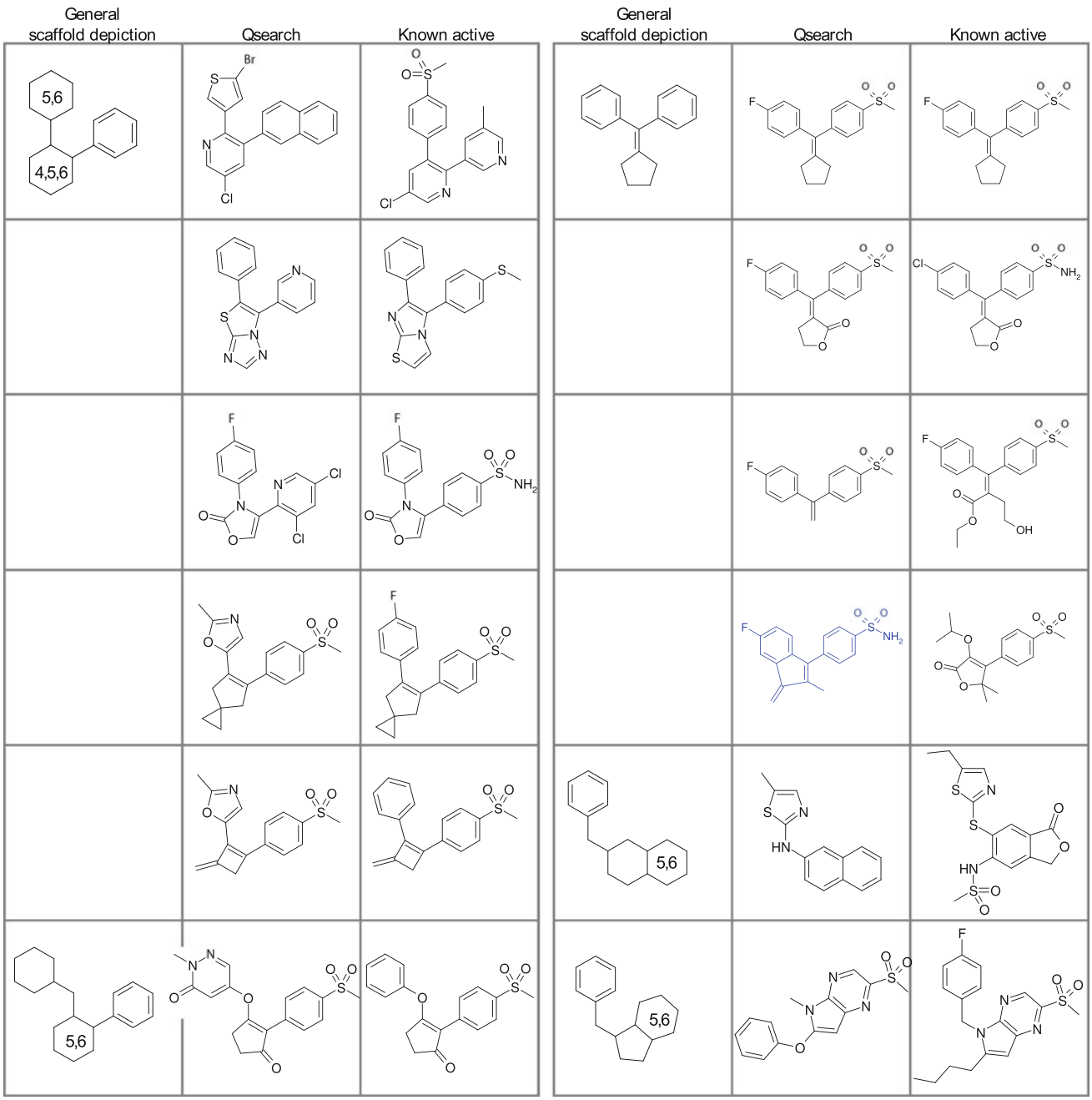

Fig. 7 Solutions for the COX-2 scenario. *Left columns*: General scaffold depictions of the major classes of known COX-2 inhibitors. *Middle columns*: Qsearch solution. *Right columns*: Closely related known active. A solution that features a novel ring-closure is highlighted in *blue*

Fig. 9. The pharmacophore features were placed on basis of the cocrystallized ligand in PDB structure 1opj [37]. The shape was defined by placing spheres of radius 1.8 on every heavy atom.

In order to account for the more complex pharmacophore query and the large fragment space with roughly 6,000 fragments, the calculations were conducted longer than in the previous examples. The program ran on four computers (each of which had an Intel Xeon with 4 cores at 2.5 GHz) with an accumulated running time of 24 h and resulted in 216 unique solutions. In the subsequent analysis, the results were checked for structural variations of other known actives as well as alternative scaffolds for imatinib.

In another experiment, the growing mode of Qsearch was used to find the scaffold change in imatinib which lead to ponatinib. The rationale for this scenario is an analysis of imatinib's interaction pattern in PDB structure 1opj. It is apparent that none of the two nitrogen atoms of the pyrimidine moiety are involved in hydrogen bonding, and

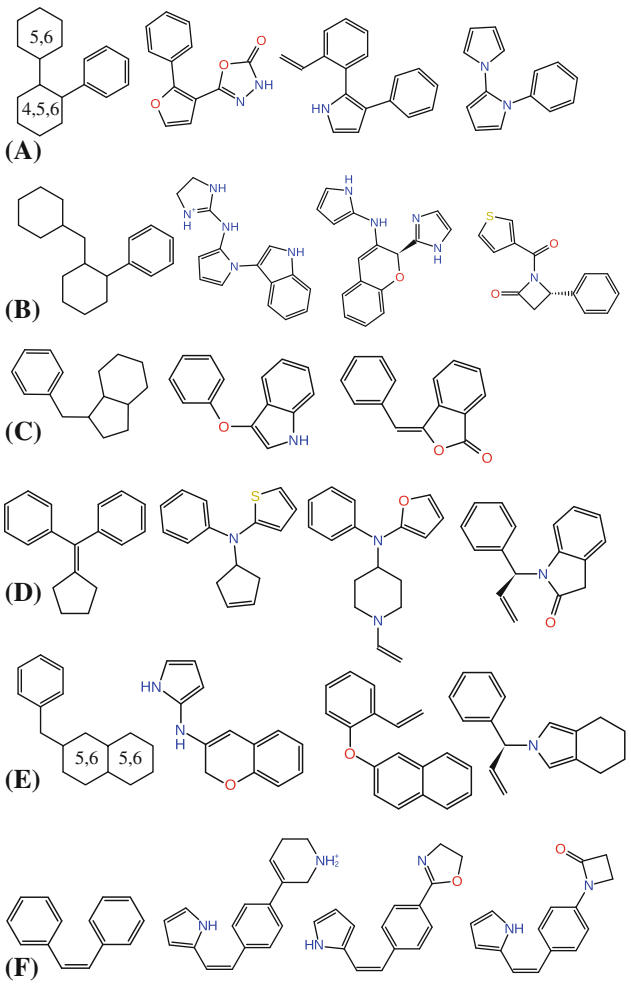

Fig. 8 COX-2 scaffolds and Qsearch solutions from the BRICS fragment space. The first structure in each row is a generalized scaffold depiction which was derived from known actives. The following structures are Qsearch solutions. The additional rings in (F) were added by Qsearch to fulfill the minimum volume constraint

that the pyrimidine may hence be eliminated. The pharmacophore contained only the acceptor and the aromatic feature for the terminal pyridine moiety of imatinib. We created a problem-specific fragment space for this problem by selecting fragments from the BRICS_4k fragment space. Fragments were included if they may act as linker fragment and are non-aromatic (541 of 1086 linker fragments) or if they are an aromatic terminal fragment with at least a single acceptor atom (717 of 2,566 terminal fragments). Qsearch was run for 9 h (Intel Xeon with 4 cores at 2.5 GHz) and delivered 2,152 solutions. The top-ranking solutions are depicted in Fig. 10.

Interpretation

In the first scenario, Qsearch was shown to reproduce scaffold changes which have been described in the literature and which have lead to new drugs. By its rigorous use of aromatic features, the pharmacophore ensured scaffolds which are relatively close to imatinib. In the following paragraphs, the solutions which were proposed by Qsearch are discussed along with their closest imatinib-analogues.

Nilotinib Nilotinib differs only slightly from imatinib. The most significant change in nilotinib is the inversion of the amide that connects the two phenyl rings. This inversion is a common bioisosteric substitution for amides in medicinal chemistry [38]. In both orientations, the oxygen atom is able to satisfy the acceptor feature of the pharmacophore (Fig. 11). Although Qsearch does not have a priori knowledge of bioisosteric substitutions, changes like the amide inversion are reproduced. This shows an important strength of combining fragment spaces and pharmacophore queries.

Sorafenib Compared to imatinib, there are several changes in the sorafenib structure. The changes that are compatible with the pharmacophore are the substitution of the central amide group with urea, and an exchange of an amine linker with an ether. The amine-ether substitution was found easily, since the motifs are geometrically nearly identical, but no urea group was found in the results. Instead many cyclic urea groups such as hydantoin were found (Fig. 9, third row). We investigated possible reasons for this behavior and found that the BRICS fragment space is generally very unlikely to construct urea residues. When the fragment space was created, the cleavage rules caused both amide bonds of all urea motifs to be cut. The resulting carbonyl group is included only once in the set of fragments. Cyclic urea on the other hand is included in 30 fragments in the BRICS_4k fragment space, every time in a different chemical context. It is much more likely that one of these cyclic groups is chosen, than that the only carbonyl group is combined with two correct fragments to result in an urea group. This example shows the impact that a well customized fragment space has on the obtained results. This specific problem can easily be solved by adding fragments that contain a urea group to the input. In general, the fragment space should be carefully selected for a design scenario.

Ponatinib This structure is chemically more different from imatinib than the previous examples, although it still has a linear scaffold like that of imatinib. Changes include the substitution of a pyrimidine ring by an ethinyl moiety and a replacement of the terminal pyridine residue with a fused heterocycle. Obviously none of our solutions contain the ethinyl scaffold replacement, since the pharmacophore demands an aromatic feature to be placed at that position.

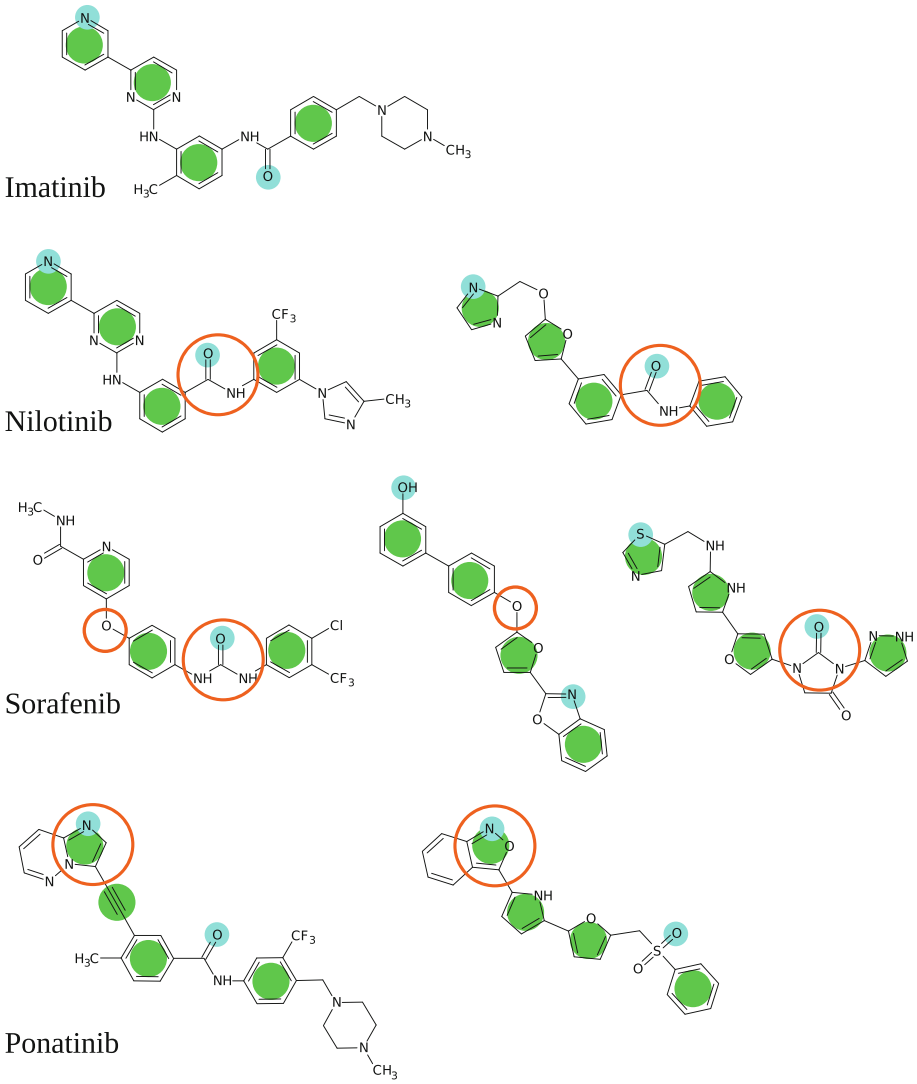

Fig. 9 Imatinib and several analogs with the pharmacophore by Wolber and Langer. Aromatic features are *green*, acceptor features are *cyan*. Directionality is omitted in this sketch. The first molecules in each row are drugs, each followed by a Qsearch solution which reproduces the changes that were introduced in the analogs (*orange circles*). For the urea in sorafenib, only an hydantoin substitution was found

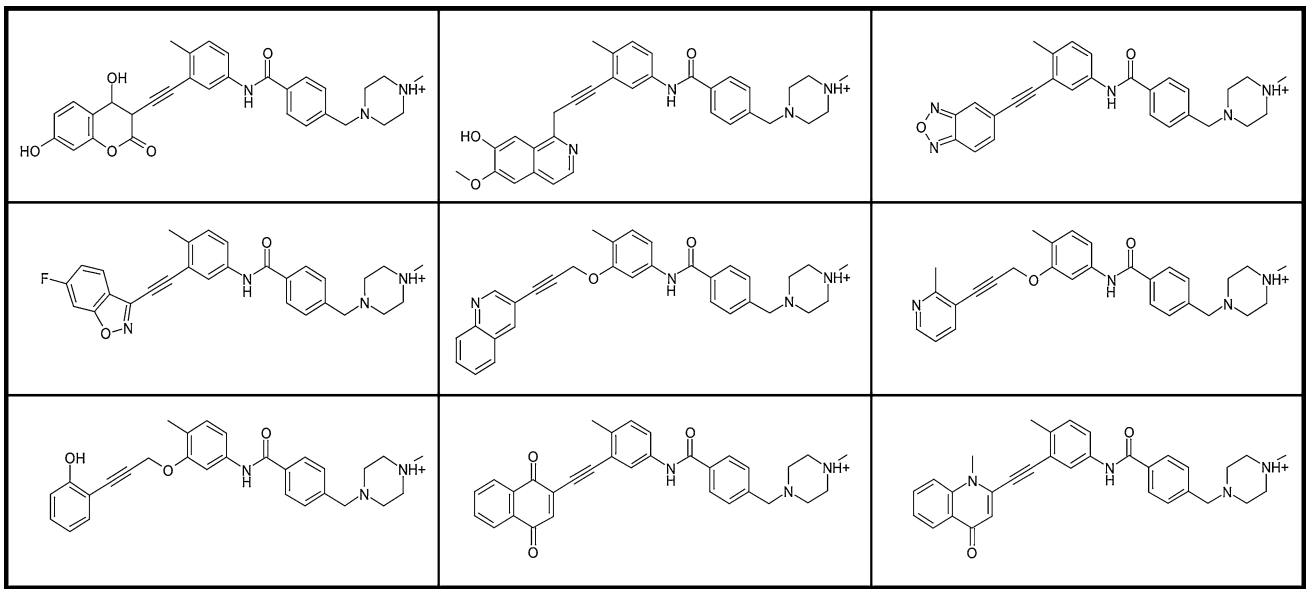

Fig. 10 2D depictions of the 9 highest scoring Qsearch solutions after placement into the active site of 1OPJ

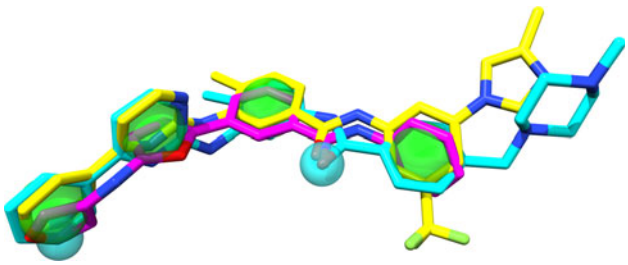

Fig. 11 Inverted amide in nilotinib (*orange*). Spheres represent pharmacophore constraints. Molecules are color-coded. *Cyan*: Imatinib, *yellow*: nilotinib, *magenta*: Qsearch solution. The amide in its normal as well in its inverted form is able to satisfy the acceptor feature

However, the pyridine residue was replaced with a benzooxazole, a heterocycle which is very similar to the one found in ponatinib.

Novel and non-linear scaffolds Additionally to the scaffold variations found in the linear topologies of known actives (cf. Fig. 13), non-linear, branched scaffolds were generated. Two of the non-linear examples are shown in Fig. 12. Both are clearly different from imatinib and other related linear scaffolds. The first scaffold contains a central sp^3 carbon atom for branching and the second a terminal bicyclic system with ortho-substitution. Evidently, not all known key interactions are fulfilled by these solutions, since they were not part of the pharmacophore. To optimize the interactions with the target structure, the solutions have to be manually modified or further processed with structure-based modeling approaches.

For example, several options for optimization are available in the sp^3-branched solution. The X-ray structure reveals that the central amide of imatinib forms a hydrogen bond with a backbone N-H. Hence, the triazolone of the solution could be replaced by an imidazole which has the donor and acceptor nitrogen atoms in the correct positions to develop this interaction pattern. Also, with the knowledge that the pyridine nitrogen of imatinib as well as its neighboring acidified C-H are involved in hydrogen

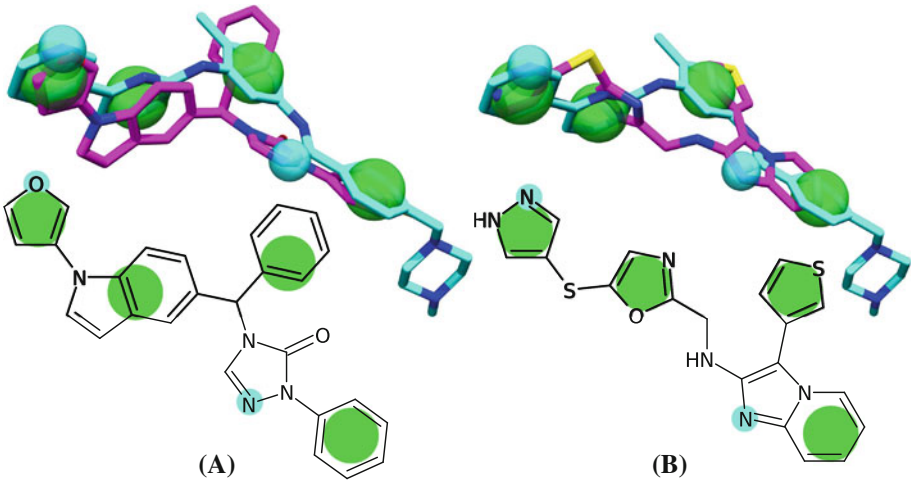

Fig. 12 Molecules found with the imatinib pharmacophore that have a topology which is different from imatinib. The pharmacophore is represented by *green* (aromatic) and *cyan* (acceptor) spheres. Acceptor directions have been omitted. In the 3D depiction, imatinib is displayed in blue and our solutions in magenta

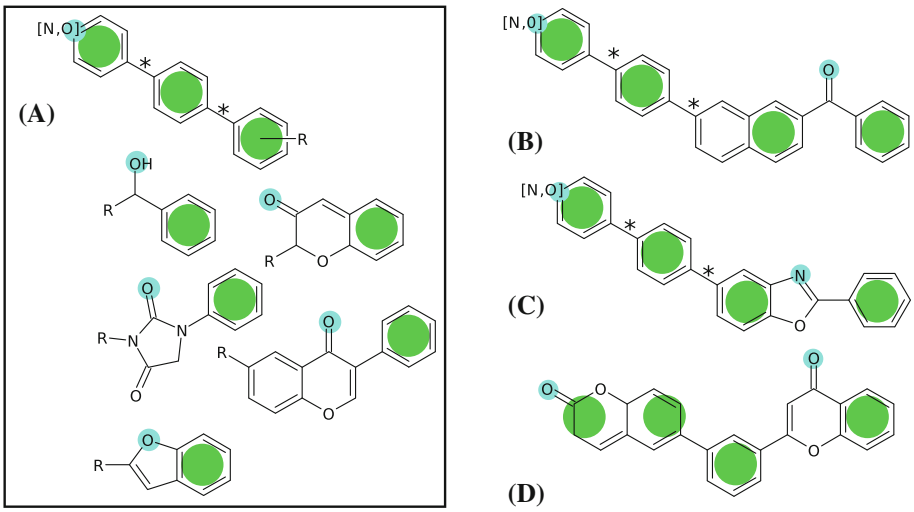

Fig. 13 Linear scaffolds found for in the BCR-ABL pharmacophore search. Heteroatoms are shown when they are essential for the pharmacophore or for aromaticity. *Green circles*: aromatic, *cyan circles*: acceptor. *Bonds* marked with an *asterisk* represent one to three bonds, depending on the ring connectivity and the geometry. All *benzene rings* represent five or six membered rings. **a** Scaffolds that belong to the same class as imatinib. **b–d** Scaffolds that show variations, but are still linear

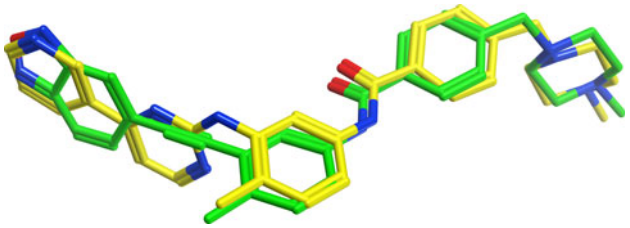

Fig. 14 Overlay of Imatinib (*yellow*) with a Qsearch solution resulting from fragment growing mode (*green*, ranked on position 5). The ethinyl linker replaces the meta-substituted pyrimidine linker of Imatinib and the terminal pyridine is replaced with a benzo[1,2,5]oxadiazole ensuring the hydrogen bonding of the pyridine nitrogen atom with a backbone N–H of the protein structure. Note that the start fragment was allowed to move slightly

bonding, the terminal furane ring of this molecule could be substituted by a 3-substituted pyrazole ring.

The solutions of the second experiment were ranked by placing them into the active site of 1opj and scoring them with MOE on basis of the developed interactions. The top 9 solutions are depicted in Fig. 10. All of them contain an ethinyl or extended ethinyl linker fragment, which is exactly the scaffold replacement which lead to ponatinib. Qsearch was thus able to identify the 8 ethinyl fragments (out of 541) as the ones which lead to the geometrically most satisfying results. The crystal structure of imatinib is superimposed with the rank 5 solution in Fig. 14. The molecule has a very good shape overlap with imatinib, contains the aromaticity of the pyridine ring and an acceptor atom in the correct position to interact with the target's backbone N-H.

Discussion

It is important to stress that the results of the de novo program Qsearch are not fully optimized lead structures but suggestions for new scaffolds, and should be read as such.

Qsearch creates intermediate solutions from fully decorated fragments which were obtained by cleaving libraries of known drug molecules. These solutions are undecorated by cleaving functional groups which are not needed for the pharmacophore (cf. Methods). This step was introduced after analyzing results of the first Qsearch runs because the solutions fulfilled the pharmacophore constraints but often did not exhibit obvious similarities to classes of other known actives. For COX-2, for example, mostly shape and hydrophobicity are responsible for molecular recognition. COX-2 solutions which were generated with the BRICS fragment space were too polar and contained too many heterocycles and acyclic polar substituents. Such abundance of spurious functional groups showed to be a general problem with the results, and is a direct consequence of using drug molecules as fragment source. The undecoration step successfully helped to lower the frequency of functional groups.

In spite of the undecoration step, the solutions sometimes show hetero atoms outside of pharmacophore features. This does not necessarily imply that these heteroatoms would be favorable within the target binding site. If, for example, the central scaffold of a Qsearch solution is a meta-substituted pyridine, and its nitrogen atom is not required by a pharmacophore feature, the relevant information is that a meta-substituted, 6-membered aromatic moiety, is appropriate as a central scaffold.

In the context of reducing the overall polarity of the solutions, we also explored the usage of hydrophobic features in the query setup, for example by placing hydrophobic features on top of aromatic features to eliminate all heterocyclic solutions. Since typical fragment spaces are highly biased towards diverse heterocycles, the likelihood of detecting a pure carbon aromatic system is rather low compared to all the present heterocyclic aromatics. The reduction of polarity through hydrophobic features was not successful, simply because most of the scaffolds that were generated had one or more polar atoms within the hydrophobic feature and were thus discarded during the search. The use of more than one hydrophobic feature often resulted in Qsearch runs without any solution. Since it is easy to replace the undesired hetero atoms in a post-processing step, we suggest to avoid using non-essential hydrophobic features, or to run Qsearch in growing mode.

The molecules which were created for the BCR-ABL pharmacophore very seldom extend to the space where the piperazine residue of imatinib lies (cf. Fig. 12). This results from the treatment of the pharmacophore's shape in the calculation of the penalty term. The penalty is only used to describe the volume which may be occupied, and there is no penalty for not covering the complete volume. Since there are no electronic pharmacophore features placed in that region, the optimization has no incentive to grow molecules beyond what is needed to fulfill the electronic features. Theoretically, one could increase the minimal volume for molecules in the quick-rejection step, to obtain larger molecules. However, there is little to no additional information which can be obtained from these larger molecules. They have simply non-functional additions to the (much more relevant) molecular scaffold which will actually clutter the relevant information and make an analysis harder. Hence, we advise to avoid regions in the pharmacophore with few pharmacophoric features for the use in de novo design, since they will not improve the results.

We have tried several parameter sets for the simulated annealing procedure which guides the molecule generation process. However, we were not able to base the choice of our parameters on a sound statistical basis, since the

number of molecules which the algorithm produces in a fixed period of time does not change significantly with different (meaningful) parameters. In contrast to this, the number and quality of the generated scaffolds does change. Unfortunately the analysis of the scaffold-diversity in a set of molecules can not yet be automated with satisfactory results. (cf. Results) A manual analysis is not an alternative due to the very large numbers of molecules which have to be considered. Instead, we obtained parameters by visually following the optimization procedure and making sensibe changes to the parameters to aid the algorithm with producing better results. What we could observe during our experiments is that many short runs with fast cooling schedules are favorable since they produce more diverse scaffolds. Longer runs will usually result in many variations of the found scaffolds, while slower cooling schedules lead to less scaffolds overall. We decided to use a parameter set which will result in a temperature close to zero after 12.5% of the optimization and conduct each run with 700 steps.

Using a start fragment in growing mode addresses two different issues. On the one hand, the start fragment may be placed at a position where only an unspecific feature, e.g. "aromatic" exists, for which many fragments fit. This avoids a multitude of solutions with slightly different aromatic ringsystems at that position, which does not confer any new information. On the other hand, a start fragment can be used if a very specific fragment is needed, e.g. a benzamidine moiety for thrombin inhibitors. The definition of a corresponding pharmacophore pattern would lead Qsearch to spend most of the computations in finding benzamidine. If a start fragment is provided instead, Qsearch will run faster and yield more meaningful results.

We have also demonstrated the advantages of customized problem-specific fragment spaces. For the COX-2 example a focused fragment space was used to ensure structural similarity to known COX-2 inhibitors. For the tyrosine kinase example customization of the BRICS fragment space proved to be an easy way to formulate a priori knowledge about linkers and terminal fragments. Instead of removing highly redundant and known linker fragments from the underlying fragment space one could also enrich the generic BRICS-space with urea fragments in order to obtain more sorafenib-like results. This shows the importance of an intelligent pre-selection of fragments in order to obtain meaningful results within optimal Qsearch run times.

Another important aspect of the Qsearch method is the pragmatic scoring approach with pharmacophores. Their second application is the ranking of the solution based on the geometric fit to the pharmacophore. The pharmacophore concept is a simple and intuitive measure which has been successfully applied in pharmacophore searches in molecular databases. More detailed scoring schemes, which try to predict binding affinities, and which are used in structure-based de novo design tools have the disadvantage of being less reliable [39], and—more importantly—of needing expert knowledge for a correct interpretation, which hinders interdisciplinary discussions [40]. Furthermore, they are much more costly to calculate than the simple pharmacophore terms which have been presented here.

Conclusions

In this paper, we presented Qsearch, a new ligand-based de novo design method, which uses pharmacophore constraints to search in fragment spaces. We have demonstrated Qsearch's ability to reproduce scaffolds of known actives but also to construct new molecules with novel scaffolds for various different design scenarios.

As always, choosing sensible input for the algorithm is of utmost importance for meaningful results. For the pharmacophore query, one has to find a good balance between leaving room for novelty and constraining the search to produce reasonable amounts of solutions. A good example was the presented BCR-ABL pharmacophore, which was quite strict, but still allowed for some alternative novel scaffolds to be produced. If knowledge exists for anchor fragments, it is advised to use a start fragment instead of using very specific pharmacophore constraints. Also, special attention need to be given on the composition of the fragment space which is used as input. For example, if certain topologies for linker fragments are highly represented in a fragment space, and are already known from actives, one should either remove them from the fragment space or enrich the fragment space with other novel linker fragments.

Last but not least, the results must be intelligently interpreted by the user. The molecules which are produced by the method must not be read as molecules, which can readily be synthesized and will definitely be active, but as ideas for new scaffolds which hold functional groups in place such that they fulfill the query pharmacophore. For example, one would read a pyridine whose nitrogen atom is not needed in the pharmacophore query, as a six-membered ring with the same exit vectors as the pyridine.

We are optimistic that Qsearch will help medicinal chemistry projects to come up with new ideas for scaffolds or replacements for parts of existing molecules. The fragment space approach offers a valuable source of molecules with beneficial properties, although it is well known from other de novo design approaches that it will not produce molecules which can instantly be used as lead structures, and need post-processing.

Acknowledgments TL: I would like to thank the team I worked with on the NAOMI project for the practical and emotional support: Sascha Urbaczek, Robert Fischer and Adrian Kolodzik. TSG,TL: We would like to express our gratitude for the good cooperation with the CAMM team at Roche. The fruitful discussions which we had were the foundations for the unique features in Qsearch. 3D images were created either with the UCSF Chimera package [41] or the MOE software suite.

References

1. Schneider G, Fechner U (2005) Nat Rev Drug Discov 4(8):649
2. Mauser H, Guba W (2008) Curr Opin Drug Discov Dev 11(3):365
3. Kutchukian P, Shakhnovich E (2010) Expert Opin Drug Discov 5(8):789
4. Dobson C (2004) Nature 432(7019):824
5. Schneider P, Schneider G (2003) QSAR Comb Sci 22(7):713
6. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M (2008) ChemMedChem 3(10):1503
7. Mauser H, Stahl M (2007) J Chem Inf Model 47(2):318
8. Lewell X, Judd D, Watson S, Hann M (1998) J Chem Inf Comput Sci 38(3):511
9. Rotstein S, Murcko M (1993) J Comput Aided Mol Des 7(1):23
10. Pearlman D, Murcko M (1993) J Comput Chem 14(10):1184
11. Boehm H (1992) J Comput Aided Mol Des 6(1):61
12. Todorov N, Dean P (1998) J Comput Aided Mol Des 12(4):335
13. Degen J, Rarey M (2006) ChemMedChem 1(8):854
14. Pierce A, Rao G, Bemis G (2004) J Med Chem 47(11):2768
15. Pearce B, Langley D, Kang J, Huang H, Kulkarni A (2009) J Chem Inf Model 49(7):1797
16. Fechner U, Schneider G (2006) J Chem Inf Model 46(2):699
17. Viswanadhan V, Ghose A, Revankar G, Robins R (1989) J Chem Inf Model 29(3):163
18. Rarey M, Stahl M (2001) J Comput Aided Mol Des 15(6):497
19. Damewood J, Lerman C, Masek B (2010) J Chem Inf Model 50(7):1296
20. Fechner U, Franke L, Renner S, Schneider P, Schneider G (2003) J Comput Aided Mol Des 17(10):687
21. Renner S, Hechenberger M, Noeske T, Bocker A, Jatzke C, Schmuker M, Parsons C, Weil T, Schneider G (2007) Angew Chem (International ed. in English) 46(28):5336
22. Todorov N, Dean P (1997) J Comput Aided Mol Des 11(2):175
23. Lloyd D, Buenemann C, Todorov N, Manallack D, Dean P (2004) J Med Chem 47(3):493
24. Grant J, Gallardo M, Pickup B (1996) J Comput Chem 17(14):1653
25. Kirkpatrick S, Gelatt C, Vecchi M (1983) Science (New York) 220(4598):671
26. Schneider G, Hartenfeller M, Reutlinger M, Tanrikulu Y, Proschak E, Schneider P (2009) Trends Biotechnol 27(1):18
27. Maass P, Schulz-Gasch T, Stahl M, Rarey M (2007) J Chem Inf Model 47(2):390
28. Leach A, Gillet V, Lewis R, Taylor R (2010) J Med Chem 53(2):539
29. Spitzer G, Heiss M, Mangold M, Markt P, Kirchmair J, Wolber G, Liedl K (2010) J Chem Inf Model 50(7):1241
30. Gmespie R (1970) J Chem Edu 47(1):18
31. Gillet V, Willett P, Bradshaw J (2003) J Chem Inf Comput Sci 43(2):338
32. Stahl M, Rarey M (2001) J Med Chem 44(7):1035
33. Kurumbail R, Stevens A, Gierse J, McDonald J, Stegeman R, Pak J, Gildehaus D, Miyashiro J, Penning T, Seibert K, Isakson P, Stallings W (1996) Nature 384(6610):644
34. Stahl M, Todorov N, James T, Mauser H, Boehm HJ, Dean P (2002) J Comput Aided Mol Des 16(7):459
35. Bemis G, Murcko M (1996) J Med Chem 39(15):2887
36. Wolber G, Langer T (2005) J Chem Inf Model 45(1):160
37. Nagar B, Hantschel O, Young M, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J (2003) Cell 112(6):859
38. Nogrady T, Weaver D (2005) Medicinal Chemistry. 3rd edn. Oxford University Press, New York
39. Shoichet B (2004) Nature 432(7019):862
40. Stahl M, Guba W, Kansy M (2006) Drug Discov Today 11(7–8):326
41. Pettersen E, Goddard T, Huang C, Couch G, Greenblatt D, Meng E, Ferrin T (2004) J Comput Chem 25(13):1605