JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Minor Groove Binders and Drugs Targeting Proteins Cover Complementary Regions in Chemical Shape Space

Julian E. Fuchs,[†] Gudrun M. Spitzer,[*,†] Ameera Javed,[†] Adam Biela,[‡] Christoph Kreutz,[†] Bernd Wellenzohn,[¶] and Klaus R. Liedl[†]

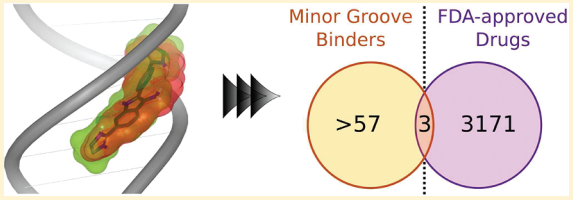
[†]Faculty of Chemistry and Pharmacy, University of Innsbruck, Innrain 52c, A-6020 Innsbruck, Austria

[‡]Department of Pharmaceutical Chemistry, Philipps University Marburg, Marbacher Weg 6, D-35032 Marburg, Germany

[¶]Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Strasse 65, D-88397 Biberach/Riss, Germany

Supporting Information

ABSTRACT: DNA minor groove binders (MGBs) are known to influence gene expression and are therefore widely studied to explore their therapeutic potential. We identified shape-based virtual screening with ROCS as a highly effective computational approach to enrich known MGBs in top-ranked molecules. Discovery of ten previously unknown MGBs by shape-based screening further confirmed the relevance of ligand shape for minor groove affinity. Based on experimental testing we propose three simple rules (at least two positive charges, four nitrogen atoms, and one aromatic ring) as filters to reach even better enrichment of true positives in ROCS hit lists. Interestingly, shape-based ranking of MGBs versus FDA-approved drugs again leads to high enrichment rates, indicating complementary coverage of chemical shape space and indicating minor groove affinity to be unfavorable for approval of drugs targeting proteins.



INTRODUCTION

Minor groove binders (MGBs) are known to prevent transcription factors from interacting with DNA. Thereby they influence up- or downregulation of gene expression, promising a general and effective approach for the treatment of diseases[1−7] and for the examination of biological processes.[8] Recent studies show that even MGBs with a weight exceeding 1000 Da are capable of accessing the nucleus in cell culture.[9] The number of genes affected by a MGB covering a reading frame of six base pairs was found to be comparable to an si-RNA experiment,[10] further encouraging the development of MGBs for therapeutic application.

The search for new MGBs is an active area of research with several different strategies pursued. Known ligands are modified synthetically,[11] building blocks are replaced,[12] and combinatorial libraries as well as experimental high-throughput affinity detection methods were established.[8,13,14] Reports even include high-throughput crystallization techniques[15] and isolation of compounds from plants.[16] With the diversity of molecules enumerated in virtual databases in mind it seems likely that an appropriate filter method could help to complement the experimental approaches. Conventional screening methods are designed for protein targets, making the selection of a screening tool appropriate for description of DNA interactions challenging.

Whereas pharmacophore models were found to be helpful for prediction of sequence preference,[17] they suffered from a lack of flexibility to describe ligands with different numbers of hydrogen bond donors. Moreover, they turned out to be highly susceptible to incompleteness of the set of conformers generated prior to virtual screening. Some of these conformers failed to match minor groove curvature. Statistical analysis of MGB-DNA complexes available from the PDB emphasized the importance of van der Waals contacts and therefore ligand shape complementarity.[18] Further motivation to use shape as a descriptor for MGBs resulted from recent reports on the influence of groove width and thereby groove shape on groove insertion of transcription factor side chains.[19,20] Grootenhuis et al. used a shape-based docking algorithm providing a proof of principle for the use of shape as a criterion in virtual screening for MGBs.[21]

Whereas we are interested in finding new MGB lead structures, other computational studies rather focus on drug−DNA complex geometry prediction and affinity prediction for lead optimization. Campbell et al.[22] present conditions for an in silico docking procedure that reliably reproduces the position of a minor groove binder observed by X-ray crystallography. Rohs et al.[23] propose an algorithm for drug−DNA docking accounting for molecular flexibility of both constituents. Combining docking with molecular dynamics simulations and quantum mechanical descriptors, Srivastava et al.[24] predict drug−DNA complex stability in order to support further MGB lead optimization. In a study also covering lead identification, Evans and Neidle[25] show that about 80% of MGB poses are predicted correctly by two docking programs (DOCK and Autodock). They achieve

Received: May 26, 2011
Published: August 07, 2011

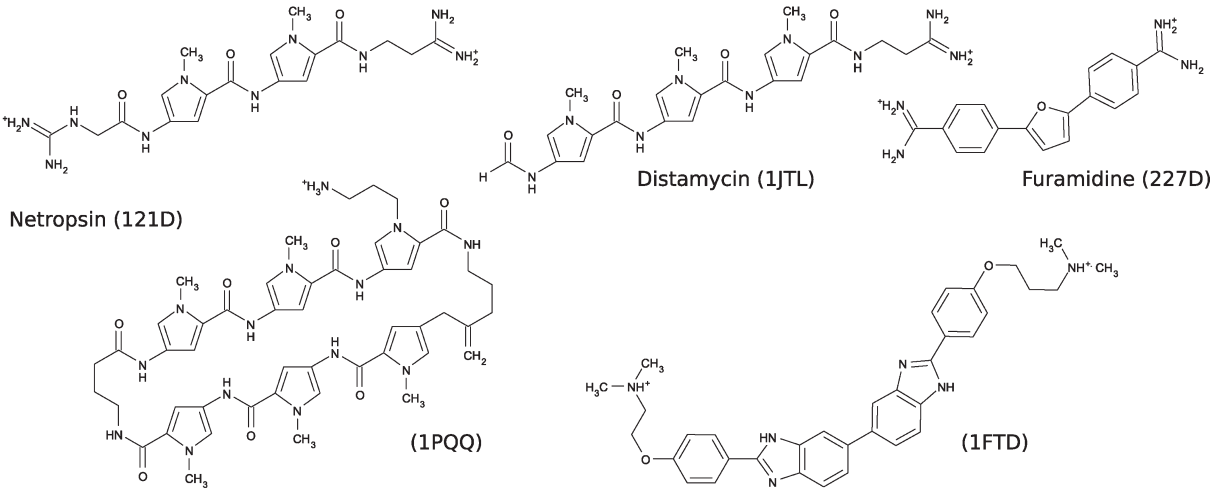


Figure 1. Ligands found in crystal structures (PDB codes indicated in brackets) used as query molecules for shape-based virtual screening.

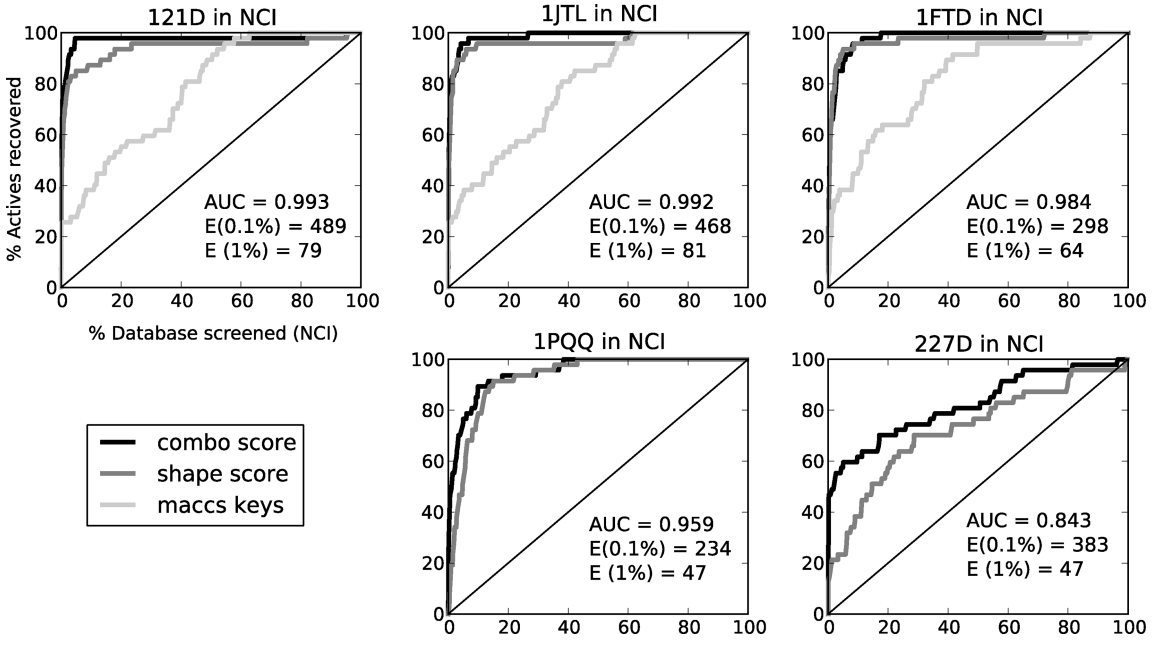


Figure 2. ROC-curves for query molecules described above show impressive retrieval rates for MGBs screening the NCI database. The combo-score setting (black) outperforms pure shape-score (dark gray). MACCS keys lead to worse retrieval of MGBs (light gray). AUC and early enrichment (E) at 0.1% and 1% of the hit list are indicated for screening with combo-score.

enrichment rates of known MGBs in a screen of randomly chosen decoy molecules close our results, but they use significantly smaller decoy sets. Shape-based screening is in many cases faster than docking[26] and therefore more appropriate for application to large-scale databases.

In the present work we use ROCS[27] for shape-based virtual screening. This approach works by describing each ligand atom by decaying Gaussians, allowing for very efficient calculation of molecular overlaps. The smoothness of ligand shape description provides access to structurally diverse hits (known as scaffold or lead hopping[28]), and it is the reason that ROCS is less susceptible to incompleteness of pregenerated conformers than pharmacophore models. ROCS does not use any structural information of DNA. It is a purely ligand based method that only relies on ligand similarity in terms of shape. In some cases this may lead to high ranking of molecules with good overall shape but small groups penetrating the DNA surface. This disadvantage is easily compensated by high enrichment rates and by avoiding problems typical for docking: lower screening speed, low cross-docking accuracy,[29] and overemphasis of molecules with high molecular weight.[25,30] The efficiency of shape screening using ROCS was proven by high enrichment factors in case of several protein targets.[31] After thorough assessment of screening efficiency for DNA minor groove binders we ordered highly ranked compounds found in the NCI database. Indeed, ten compounds were found to bind to DNA although they have not been reported in the literature to be minor groove binders yet. After this confirmation of the reliability of shape as a descriptor for minor groove affinity we screened compounds approved by the food and drug

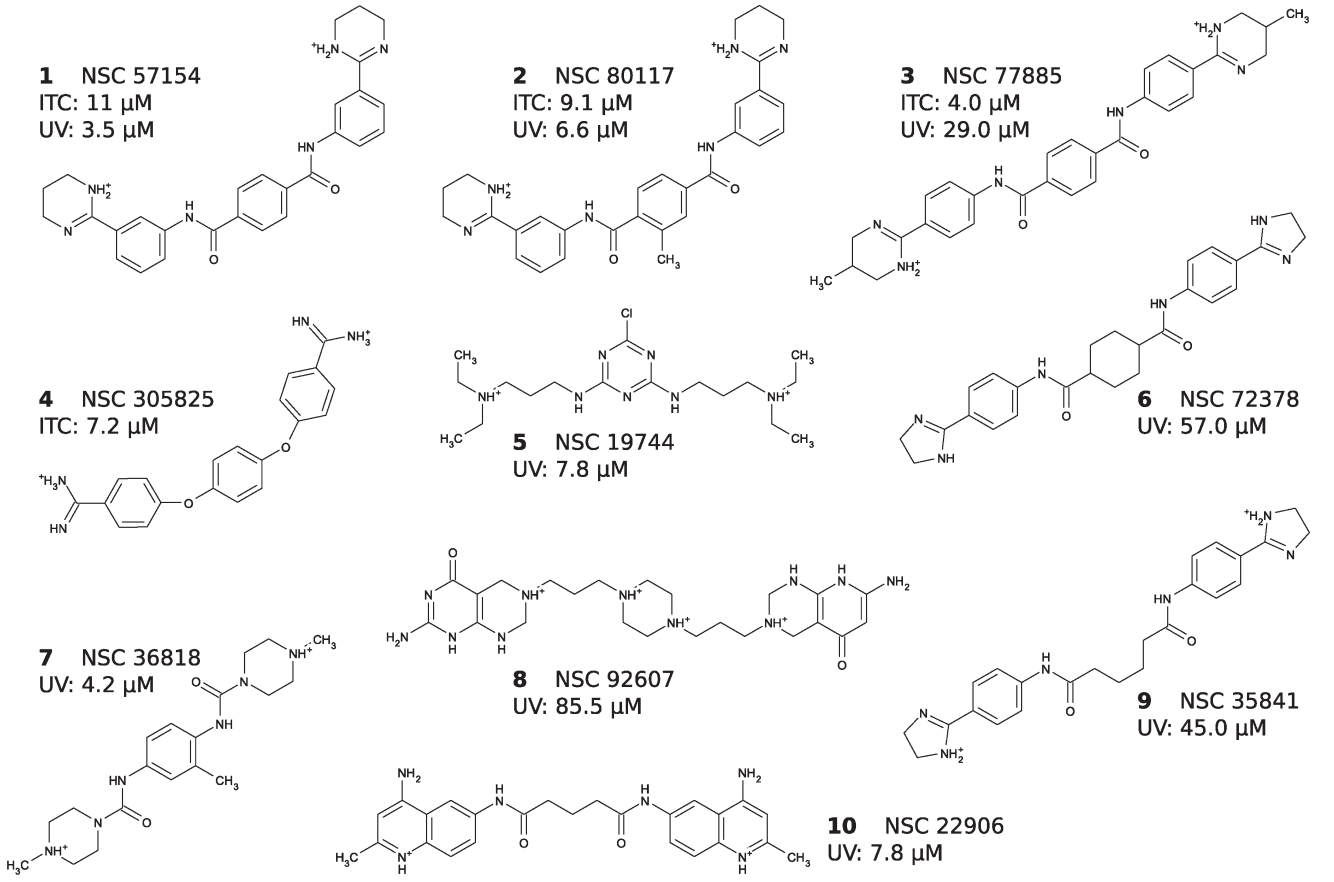


Figure 3. Ten ligands were found by ROCS shape-based screening in the NCI.

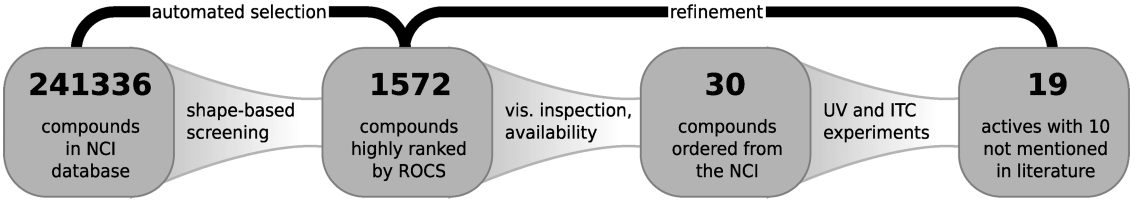


Figure 4. In a fully automated way ROCS easily reduces the NCI data set by a factor of 150 keeping MGB-like molecules in the hit list. The refinement step was intended to reduce the number of compounds for experimental investigation. Experiments showed that a more restrictive cutoff for accepting compounds in ROCS hit lists would have been even better and that the proportion of actives in the set of selected compounds is very high.

administration (FDA). We found FDA-approved drugs to cover regions in chemical space well separated from MGBs.

METHODS

Shape-Based Virtual Screening. For all databases and compound sets conformers were generated with Omega,[32] version 2.3.2 prior to virtual screening. Default settings were chosen allowing generation of up to 400 conformers covering an energy window of 10 kcal/mol per compound in the database. ROCS is an acronym for 'Rapid Overlay of Chemical Structures', a virtual screening application of OpenEye Scientific Software.[27,28] ROCS represents heavy atoms by Gaussians with parametrized decay constants according to the respective van der Waals radii. This representation of atoms allows a fast shape comparison of molecules due to the straightforward calculation of molecular overlaps providing sufficient speed for virtual screening of large databases. Besides shape description ('shape score') ROCS includes a color force field allowing for basic inclusion of chemical information in the screening process ('color score'). The Implicit Mills Dean Color Force Field[33] was used as recommended by OpenEye. Combination of both scores results in the so-called 'combo score' used for all screenings reported in this manuscript except for the shape-score run displayed in Figure 2. Besides the shape-score run, default settings were used for all screening runs. VIDA was used for visualization of superposed ligands and queries (see Figure 5), VIDA, build number 20090810, OpenEye-Scientific Software. MarvinSketch was used for displaying chemical structures in Figure 1 and Figure 3, MarvinSketch 5.2.0, 2009, ChemAxon (http://www.chemaxon.com).

Ligands from all available small molecule MGB-DNA complexes were extracted from the PDB to serve as a set of

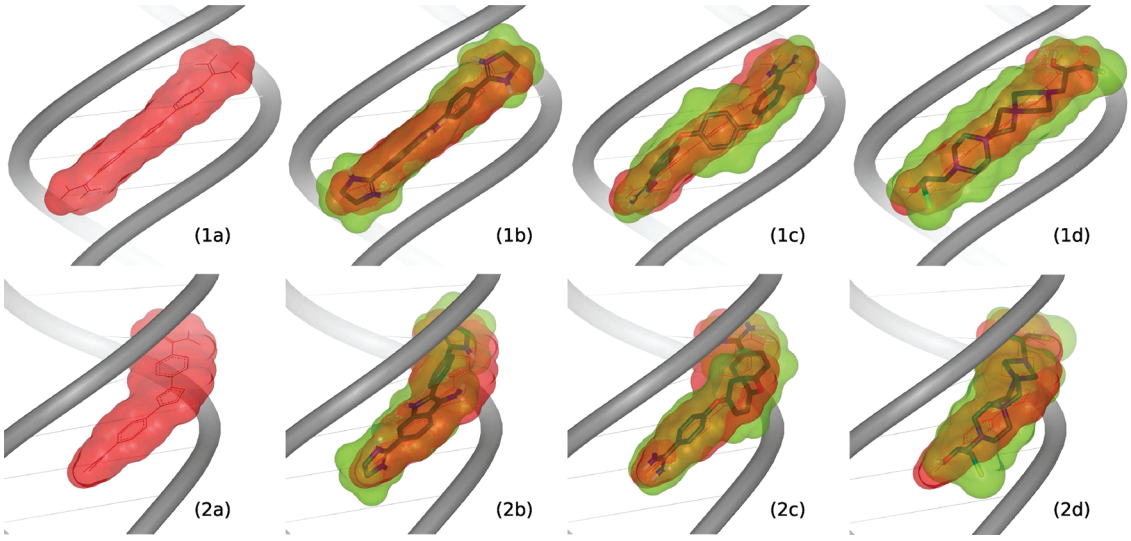


Figure 5. Row one and row two show different orientations of the same molecules. (a) Furamidine (red shape) is the ligand in crystal structure 227D and served as one of the query molecules in the shape-based screening process; (b) although the chemical scaffold of the well-known MGB[50] (green shape, ligand **24** from the known actives, see the Supporting Information) differs from the query molecule, it is easily found via its similar shape; perspective (1c) shows that the aligned ligand (compound **4**) does not perfectly match the query molecule, but from (2c) it is obvious why active compound **4** achieved a high overall shape score; column (d) shows inactive compound **40**; its 2D-structure resembles a MGB but especially orientation (1d) shows, why it is poorly ranked by ROCS shape-based screening.

active compounds. Since all MGBs beyond a molecular weight of 1000 Da possess the same polypyrrole scaffold and form hairpins in the minor groove, 1000 Da were used as a cutoff criterion for ligand selection. Selection of unique structures resulted in a test set of 47 different molecules for which conformers were computed with Omega according to the protocol described above. A list of respective PDB codes as well as structural formulas is given in the Supporting Information. Five reference molecules of diverse chemical scaffolds and lengths were chosen from this set for the actual screening protocol (PDB codes: 121D,[34] 1FTD,[35] 1JTL,[36] 1PQQ,[37] 227D[38]). See Figure 1 for query molecule structures. Query molecules were used in bioactive conformations as found in X-ray structures.

The database of the National Cancer Institute (NCI) was retrieved from http://129.43.27.140/ncidb2/download.html (access date: 23.10.2009). The latest version 'release 3, September 2003' contains 260,071 structures. Consistent with the cutoff chosen for active MGBs, we filtered the NCI database according to a maximum molecular weight of 1000 Da. Counterions and metal-containing fragments were removed in MOE, version 2009.10[39] (245,123 compounds retained) before Omega generated conformers with default settings. Omega further reduced the database by excluding molecules with unparametrized atoms like Se or As, as it has not been able to generate conformers for these molecules. 241,336 molecules in 17,278,451 conformations were finally scored by ROCS.

All other databases screened for this study (FDA-approved drugs, PubChem, InterBioScreen, Enamine) were downloaded from the ZINC-homepage (http://zinc.docking.org, access date: 2008-30-12) that provides a collection of commercially available compound sets prepared for virtual screening.[40] Due to tautomer enumeration the compound numbers in the ZINC databases are usually larger than in the originally published databases. In the versions used, the FDA-approved drugs comprise 3174 compounds, the PubChem consists of 121,207 compounds, and 945,771 compounds were screened in the Enamine and 569,699 in the InterBioScreen databases.

Enrichment factors, receiver operating characteristic (ROC) curves, and areas under the ROC-curves (AUC) were used for screening quality assessment. The enrichment factor indicates the improvement of the hit rate achieved by a virtual screening approach compared to a random selection.[41] It is calculated as the ratio of actives in the hit list at a given percentage of the database screened versus actives in the whole database. Enrichment factors at 1% and 0.1% were calculated for Figure 2 and Figure 8 to focus on the very early enrichment. The ROC-curve provides a graphical representation of the distribution of actives in a database of supposedly inactive compounds ranked by virtual screening.[42] The *x*-axis represents all molecules (actives as well as inactives) in a database rank-ordered by shape-based screening, scaled from 0 to 100%. For every active ligand in the rank-ordered list, the curve raises one step in the y-direction (the *y*-axis is also scaled to range from 0 to 100%). A steep slope at the beginning of the curve therefore indicates a high enrichment of active ligands among the highest ranked molecules and therefore high screening success. The diagonal represents random distribution. The AUC simply represents the area under the ROC curve and is used for easier comparability. See Kirchmair et al.[41] for a more detailed discussion of screening quality metrics.

Chemicals. DNA with the Dickerson-Drew-Dodecamer sequence (5′-CGCGAATTCGCG-3′) was purchased from Integrated DNA Technologies (IDT), BVBA, Leuven, Belgium. Compounds were ordered from the NCI (National Cancer Institute, Drug Synthesis and Chemistry Branch, Chemotherapeutic Agents Repository, operated by: Fisher BioServices, Rockville, MD).

UV Spectrophotometric Titrations. All experiments were carried out at room temperature in 0.01 M phosphate buffer containing 137 mM NaCl, 2.7 mM KCl, 10 mM sodium phosphate dibasic, 2 mM potassium phosphate monobasic at pH 7.4. 1% DMSO was added to the buffer to improve solubility

of all compounds except for 1 and 5. The absorption spectra were recorded on Perkin-Elmer lambda XLS+ UV−vis spectrophotometer and analyzed according to the Benesi−Hildebrand method.[43] The absorption titrations were generally performed twice by keeping ligand concentrations constant and varying DNA concentrations. The concentration of drug complexed to DNA was determined by subtracting the DNA absorbance. The binding constant was derived by plotting $A_0/(A_0\text{-}A)$ versus 1/[DNA]. A_0 and A are the absorbances of drug in the absence and presence of DNA. K_a corresponds to the intercept divided by the slope of the linearly fitted data.[44,45]

Isothermal Titration Calorimetry (ITC). ITC measurements were used to complement UV spectrophotometric titrations especially for compound **4** which has a wavelength of maximal absorbance (λ_{max}) too close to DNA's λ_{max}. Compounds **1**-**3** were investigated by ITC too, to confirm comparability of the two methods. ITC experiments were carried out on an ITC200 titration calorimeter (Microcal, Inc., Northampton, MA). Measurements were performed at 25 °C in buffer containing 100 mM Hepes, 200 mM NaCl, and 1 mM EDTA at pH 7.0. Compounds **2**, **3**, and **4** were dissolved in DMSO first and then diluted in buffer to achieve a final DMSO concentration of 5%. Compound **1** was easily soluble in pure buffer. Final ligand concentrations ranged from 0.3 mM to 2 mM. Salt and DMSO concentrations for the DNA solution were always chosen to be identical with the ligand solution. DNA concentrations ranged from 0.01 to 0.08 mM. Ligand solutions were titrated into the stirred cell (0.203 mL) containing DNA solution. The first injection had a volume of 0.5 μL to prevent artifacts arising from filling of the syringe (not used for data fitting). The volume for the following injections was chosen to be 3 μL for compounds **1**, **2**, and **3** and 2 μL for compound **4**.

Integration of raw data followed by correction for heats of dilution and mixing by subtracting the final small peaks was done within Origin 7.0 with its calorimetric models implemented by Microcal, Inc. A single-site binding isotherm was fitted to determine the association constant K_a, the reciprocal of the dissociation constant K_d.[46] Measurements were generally performed twice. No buffer dependence was found for DNA minor groove binders while repeating experiments using pyrophosphate instead of Hepes buffer.

RESULTS AND DISCUSSION

Method Validation. For evaluation of the virtual screening process a test set of known active molecules was generated by extracting ligands from all available small molecule minor groove binder-DNA complexes in the PDB as described in the Methods section. As a set of experimentally confirmed inactive compounds was not available, we used large databases as decoy sets to assess virtual screening efficiency. Results were equal for several large-scale databases (NCI, PubChem, Enamine, InterBioScreen) with regard to enrichment, shape of the receiver operating characteristic (ROC) curve, and area under the ROC-curve (AUC). As a representative we will discuss the results for screening in the NCI database in more detail, because of one important advantage: The NCI database contains known minor groove binders like Netropsin and Distamycin (see Figure 1). This ensures that the database is not subject to any kind of prefiltering that causes removal of MGB-like molecules and would lead to artificial enrichment in our screening experiments.

Five structurally diverse MGBs in their bioactive conformations served as query molecules (see Figure 1). The number of five query molecules was found to be optimal by Kirchmair et al.[31] We started our query selection with netropsin, a well characterized MGB with nanomolar affinity.[47] Netropsin (PDB code 121D[34]) turned out to be the most efficient query molecule. Even using pure 3D shape information of the ligand yielded a very high AUC value of 0.960 (see Figure 2), which is further increased to 0.993 by inclusion of chemical information using the color force field. For comparison, MACCS structural keys[48] indicating presence or absence of 166 substructure features of one to ten heavy atoms were calculated with MOE[39] using Tanimoto similarity. A similarity search based on these pure 2D fingerprints yielded a considerably lower AUC value of 0.789 for query 121D. Because of the considerably lower enrichment MACCS structural keys achieve for the queries most successful in ROCS screening (121D, 1JTL,[36] 1FTD[35]), they were not further considered.

The other four query molecules were selected to be most dissimilar from netropsin in terms of ROCS shape. To cover a range of MGB lengths, distamycin (1JTL) and a symmetric bisbenzimidazole-based MGB were selected (1FTD) in addition to 121D as queries for ROCS screening runs. Whereas both show again high enrichment and AUC, ligands shorter than netropsin (for example DAPI, extracted from PDB structure 432D[49]) are not comparably successful in retrieving MGBs. Furamidine (227D[38]) and the ligand in 1PQQ[37] are used as query molecules as they are most dissimilar from netropsin in a ranking of all active ligands. Whereas furamidine has a different chemical scaffold, 1PQQ is the most voluminous ligand with two strands similar to netropsin but connected to form a ring. Especially furamidine shows a worse AUC but still high early enrichment. Moreover it helps to recover small MGBs like the ligand Figure 5b.[50] An overview on ROC-curves, AUC values, and early enrichment is given in Figure 2. Using more than these five queries for database screening does not lead to additional diverse molecules among the top-ranked in the hit lists.

Identification of New Compounds. The NCI database was first used as a set of molecules for which inactivity was assumed to assess screening efficiency of several query molecules. After selection of five query molecules (see Figure 1) the first 400 hits (0.17% of the database) of each query were inspected visually to select compounds for experimental testing (altogether 2000, reduced to 1572 by selecting unique compounds, 0.65% of the database). Visual inspection was the method of choice to avoid loss of innovative structures due to further filtering by characteristics or substructures of known MGBs. We controlled the diversity of our selection by calculating MACCS similarities. The average Tanimoto similarity between a query molecule and a molecule in the hit list is 0.41. The average similarity between each query molecule and each selected compound slightly raised to 0.47, indicating a low bias by the author's previous knowledge on MGB structures. Visual inspection has been used to discard molecules with good overall shape complementarity but small groups still extending beyond the concave side of the query molecule. Whereas such small groups on the convex, solvent-exposed side can be accepted, they would lead to shape incompatibility with the minor groove on the concave side of the ligand. The appearance of such small groups at inaccessible regions is a consequence of ROCS' smooth shape description and is the price that is paid for the ability to achieve scaffold hopping. Unfortunately, our compound selection was limited

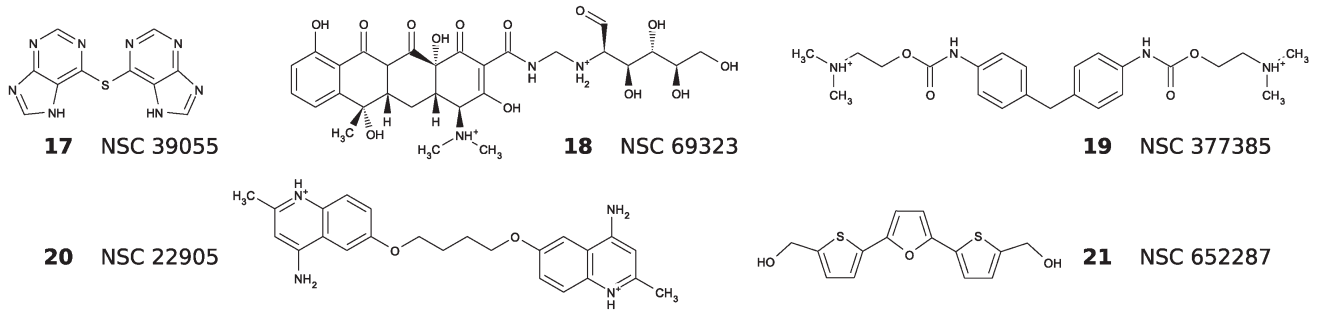


Figure 6. These five compounds were found by ROCS shape-based screening but turned out to show very low or no affinity at Dickerson-Drew dodecamer DNA. Protonation according to Wash-function in MOE.

by actual availability of compounds indicated in the NCI online order form.

Finally, 30 compounds were selected after shape-based screening from the NCI database. Due to insufficient solubility in buffer, six compounds could not be examined (see compounds **11**-**16** in the Supporting Information). Binding affinity to Dickerson-Drew-dodecamer (5′-CGCGAATTCGCG-3′) was determined by means of UV spectrophotometric titrations. Compound **4** could not be used in UV experiments because the wavelength of its maximal absorbance (λ_{max}) is too close to λ_{max} of DNA. Instead, its binding affinity was determined by isothermal titration calorimetry (ITC). Binding constants of compounds **1** to **3** were also determined by ITC to check comparability between both methods. Nine out of nineteen active compounds turned out to be known MGBs. Five compounds (**17**-**21**) were found to be inactive (see Figure 6). The remaining ten tested molecules were found to be active and were not reported to be MGBs according to SciFinder.[51] Compound **1** was found to be associated with tuberculostatic and cancerostatic effects,[52] compounds **4**[53,54] and **10**[55] were associated with trypanocidal activity, and compound **5** is a known dye. We found at least CAS numbers associated with compounds **2**, **3**, **6**, **8**, and **9** but no reports on their activity; compound **7** does not even have a CAS number assigned. See Figure 3 for structures and affinities and Figure 4 for an overview over the screening workflow.

As two representatives, compounds **3** and **4** were investigated in NMR experiments. For both ligands we observed the imino proton resonances of the central four base pairs to change with increasing amounts of ligand added to the DNA solution. This simultaneous change for central base pairs corresponds well with minor groove binding and excludes other mechanisms like intercalation. Both compounds cover the central four AT base pairs and contact the innermost GC base pairs. Due to the similarity of their shapes we expect the other ligands to occupy the same binding site in the minor groove of the DNA oligomers.

Since the set of newly identified MGBs shows affinities in the micromolar range whereas we found low nanomolar affinity for netropsin, it may be argued that shape-based screening was not successful in identifying highly active ligands. However, we found the affinity of berenil and pentamidine to be in the same range although both are well-known MGBs and FDA-approved drugs. Moreover, the ability to find ligands exhibiting lower affinity than netropsin indicates that shape-based screening with ROCS is highly sensitive in detecting MGBs.

Three of the unknown active compounds (**1**, **2**, and **3**) show similar chemical scaffolds derived from terephthalic acid, a moiety already known to be favorable for minor groove affinity.[56,57] Terminal regions of compounds **6** and **9** are similar to those of the terephthalic acid derivatives but differ in the central region where the aromatic ring is replaced by an aliphatic chain (ligand **9**) or an aliphatic ring (ligand **6**). Compound **4** contains a bend hydroquinone-diphenylether scaffold elongating the phenylether that was found to be an appropriate MGB building block by Tanious et al.[58] At first sight it seems surprising that two ether groups allow for a conformation compatible with the minor groove. Still, the compound is found by shape-based screening (see images (1c) and (2c) in Figure 5 for a superposition of ligand **4** and the query molecule), and it shows indeed affinity toward DNA oligomers. A piperazine core as found in compound **8** is reported to increase DNA binding activity and in vitro growth inhibition in a cancer cell line assay of interstrand DNA cross-linking molecules.[59] Despite the more voluminous shape and the requirement of a wider minor groove that characterize piperazine rings (in ligand **7** and **8**) in comparison with aromatic rings, they appear to impart a favorable binding interaction.[60,61]

Conclusions from Experiments on Screening Strategies. Initially, the first 400 hits in each query's hit list were inspected for selection of testing molecules. Later on, the experiments showed that nine out of ten compounds would have been found by considering only the best ranked 200 compounds. Only two of five inactives (compounds **19** and **21**) would have been selected with a cutoff after 200 best ranked compounds. This more restrictive cutoff would have led to an even more impressive enrichment of actives among the ordered compounds.

In addition to those well ranked by shape-based screening, ten arbitrarily selected compounds were ordered to test DNA affinity of molecules with any elongated structure but low shape overlap with query molecules (see **7** for structures). Three compounds (**31**-**33**) in this set are not soluble; the others did not show affinity to the DNA. Together with inactives found by shape-based screening they help to point out which characteristics are unfavorable for DNA affinity. Inactive ligands **17** and **21** (see Figure 6) lack a positive charge according to MOE's Wash-function which deprotonates strong acids and protonates strong bases.[39] There is only one molecule in the set of actives and PDB structures that is lacking a positive charge, too. It is the bis-benzimidazole ligand in PDB structure 453D. Although this ligand obviously binds to DNA, a positively charged derivative of the uncharged bis-benzimidazole is reported to show higher affinity,[62] suggesting that a positive charge may be not mandatory but highly favorable for minor groove interaction. Even the number of MGBs having only one charge is strictly limited to those containing the polyamide scaffold of distamycin or a bisbenzimidazole group. This leads to the conclusion that

ligands less well filling the minor groove require two charges to achieve sufficient affinity. In addition to two charges, none of the active ligands contains less than four nitrogen atoms, whereas their number is lower in inactive compounds **21**, **34**, **35**, and **36**. Besides forming amide groups that carry charges, nitrogen atoms may also donate hydrogen bonds toward the minor groove floor. Ligands 39 and 40 both fulfill the requirements of having two charges and four nitrogen atoms and even have elongated structures appropriate for DNA interaction but still do not show affinity. The right column in 5 shows a bad overlap between nonaromatic compound **40** with the aromatic shape query molecule. The absence of aromatic atoms in compound **40** explains bad shape overlap (see Figure 5d) as well as missing DNA affinity, since backbone sugar moieties may form stacking interactions with aromatic rings.[63]

Table 1. Comparison of The Effectiveness of Filter Critera

	NCI	FDA
no metal, weight <1000 Da	245007	
unfiltered, unique		3145
(1) number of positive charges (FCharge) > 1[a]	8549	191
(2) number of nitrogen atoms (a_nN) > 3[a]	4165	107
(3) number of aromatic atoms (a_aro) > 4[a]	3353	24
vdisteq ≥3.349[b]	2859	20
weinerpath ≥993[b]	3025	21
chi0 ≥ 14.983[b]	3044	21
chi1 ≥ 10.04[b]	3035	20
ROCS screening[b]	**462**	**4**

[a] Three filter criteria found by experimental investigation are applied consecutively, leading to a prominent reduction of the data set. [b] Five approaches are proposed as a last step to compare ROCS with simpler 2D descriptors. None of the 2D descriptors available in MOE is as effective as ROCS shape-based screening. Only the most successful four descriptors are reported.

Moreover, there is no example in the PDB ligand set lacking aromatic substructures. Too many aliphatic, freely rotatable bonds (compound **20**, **34**, **35**, **36**, and **38**) may hamper DNA binding if not compensated by highly favorable groups as benzamidinium in pentamidine (ligand with PDB code 1D64). Completely lacking hydrogen bond donor functions (**35**, **36**) and large annealed ring systems (compounds **18** and **37**) seem to be unfavorable too.

Simple Descriptors Cannot Replace Shape Description. We tested the discussed MGB characteristics for their efficiency as database filters. The requirement of at least two positive charges (according to descriptor 'FCharge' in MOE[39]), more than three nitrogen atoms and more than four aromatic atoms turned out to be most effective in further reducing hit lists after ROCS shape-based screening. The discussed three descriptors provide an easy to assess rule of thumb. Used without ROCS they already reduce the NCI data set from about 240,000 molecules to 3353, although these criteria lack the shape aspect. To test if ROCS could be replaced by a simpler 2D descriptor we applied a number of connectivity, topological, and shape indices provided by MOE[64] to the prefiltered data set. Cutoffs were chosen to keep active molecules of the PDB set. The most effective descriptors are reported in Table 1 (see the Supporting Information for complete list of tested 2D descriptors). Table 1shows that no other descriptor is as effective as ROCS that further reduces the NCI set by a factor of about seven.

Comparison with FDA-Approved Drugs. After successfully identifying ROCS shape-based screening as a reliable tool to discover MGBs, we were interested in the abundance of MGB-like molecules in approved drugs as therapeutic application of MGBs is widely discussed.[65,66] Strikingly, the enrichment and AUC values characterizing shape-based screening of MGBs in FDA-approved molecules are as excellent as already seen for screening the NCI database (see Figure 8). To test if this enrichment of MGBs in the FDA set is an artifact of the highly specific shape description, we

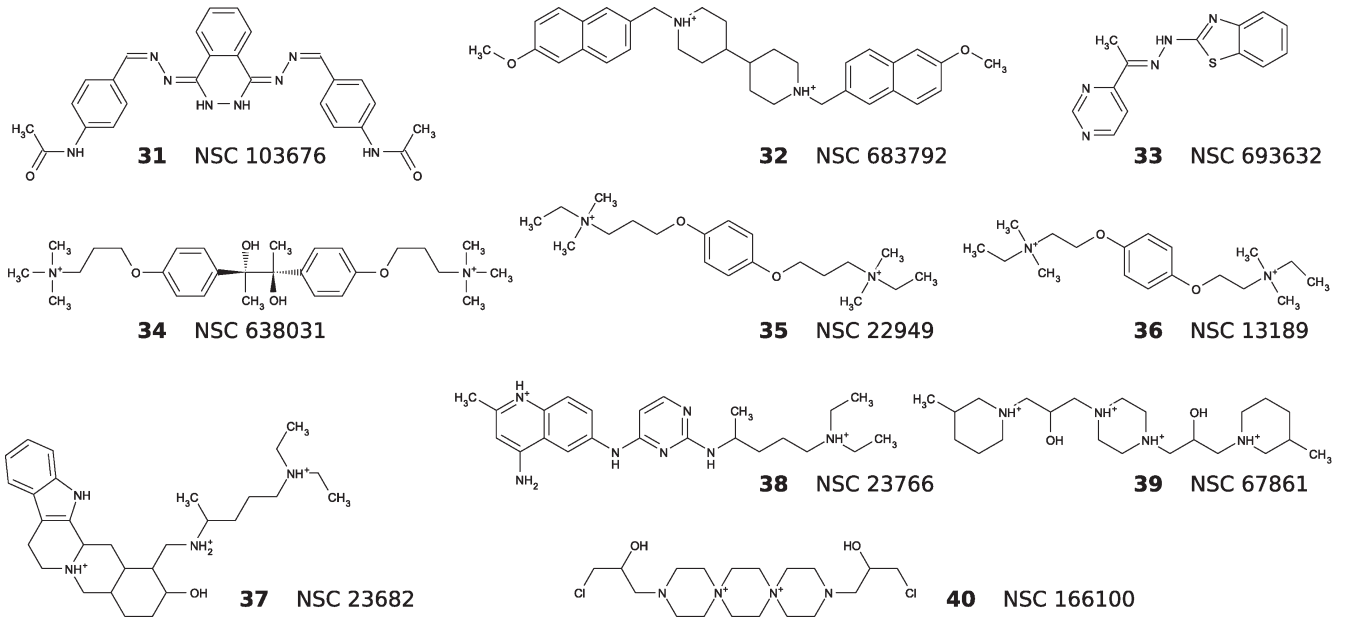


Figure 7. Ten compounds were ordered from the NCI although they did not achieve good ranking positions by shape-based screening. They were intended as a negative test for the screening approach and indeed turned out not to interact with DNA. Protonation according to Wash-function in MOE.

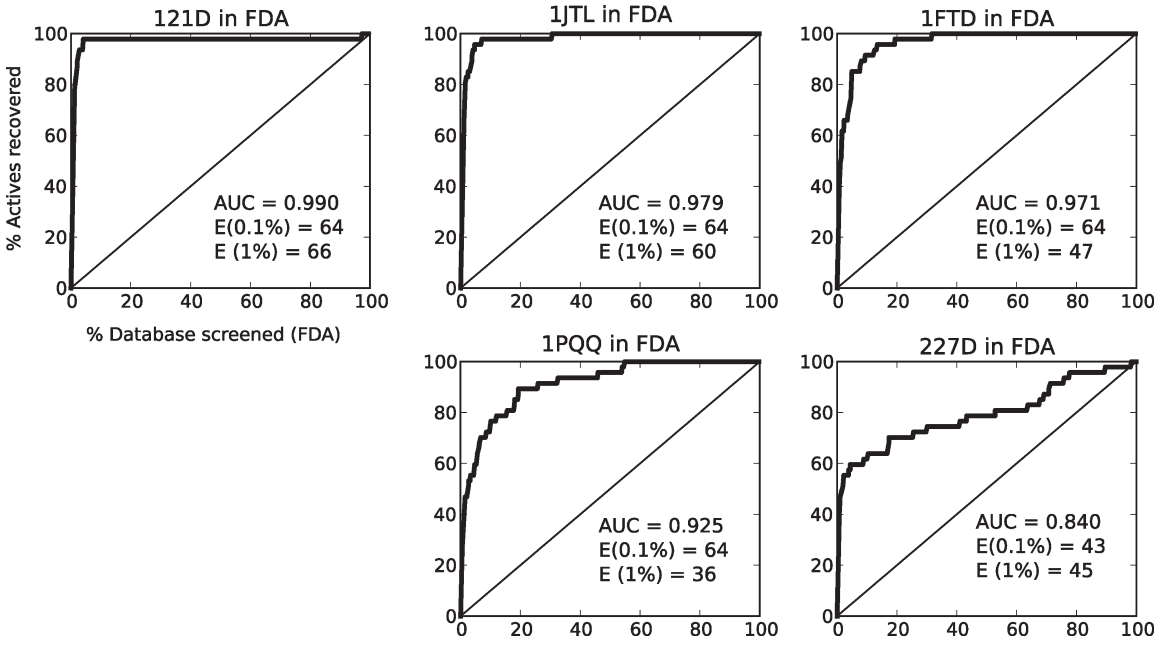


Figure 8. Similar to the screening success in the NCI database, ROCS shape-based screening leads to clear retrieval of MGBs from FDA-approved molecules (3174 entries).

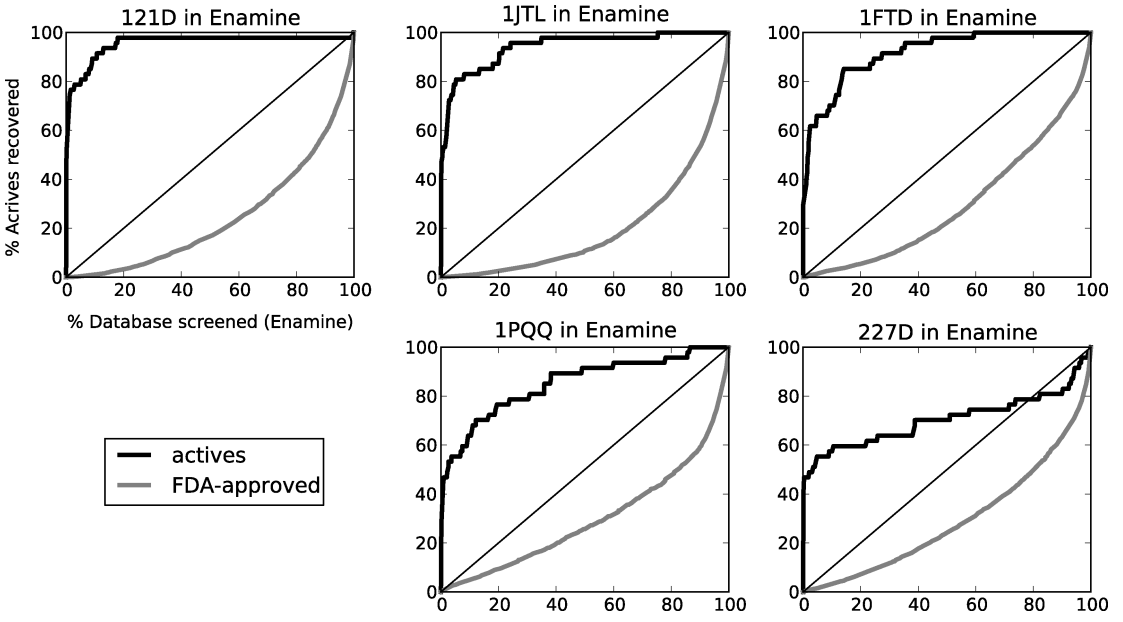


Figure 9. MGBs as well as FDA-approved drugs screened in the Enamine database (945,771 entries) as an example of a large vendor database leads to a good enrichment of MGBs but to a pronounced derichment of FDA-approved drugs.

compared shapes of MGBs as well as FDA-approved drugs with large vendor databases intended for drug discovery, expecting even distribution of FDA-approved drugs. Instead, in terms of ROCS shape the FDA-set is more dissimilar from MGBs than the average of for example the Enamine database (945,771 entries), indicated by the concave, gray curves in 9 (see the Supporting Information for similar results for screening in PubChem as well as InterBioScreen databases). This is an additional hint that there is a clear difference between MGBs and FDA-approved drugs in terms of ROCS shape.

Combination of 100 highest ranked FDA-approved compounds of each query results in a set of 321 unique molecules. Application of the three filter criteria found as explained above (charge higher than one, more then three nitrogen atoms, more than four aromatic atoms) leads to a set of only four compounds (see Table 1 for the screening process) with one bearing a system of four annealed rings unfavorable for groove binding. The remaining compounds are known MGBs Pentamidine, Diminazene (also called Berenil), and Hydroxystilbamidine (see Figure 10). They are active against protozoa and fungi with a long history of use in the treatment of African trypanosomiasis

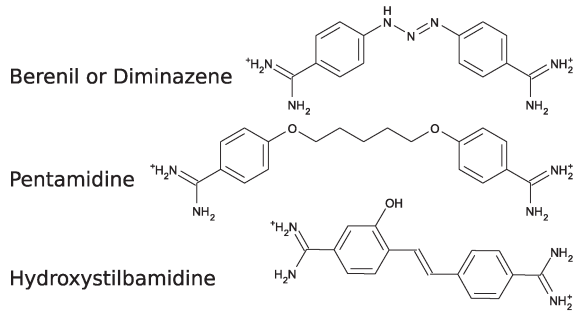


Figure 10. Only three FDA-approved molecules fulfill all criteria elaborated for MGB-likeness: High shape similarity with known MGBs, two positively charged groups, at least four nitrogen atoms, at least four aromatic atoms.

and leishmaniasis. Their mechanism of action is not completely understood, with several mechanisms being discussed controversially.[67−69]

There is no reason to expect that MGBs are not able to bind proteins. Perilo et al.[70] recently published structures of Pentamidine and Diminazene cocrystallized with bovine trypsin as model system for the serine protease class of enzymes. Still, approved drugs can easily be separated from MGBs, suggesting that minor groove affinity may be unfavorable for approval of drug molecules intended for targeting proteins. The underlying reasons can only be surmised. Every cell contains DNA with widespread AT-rich regions that are preferred targets of typical MGBs[71] and may absorb drug molecules preventing them from reaching their intended target. Influence on gene expression leading to side effects would be another plausible explanation. An analysis of molecules failed to reach FDA-approval could be helpful but hard to work out because of nondisclosure of the data.

CONCLUSION

We find shape-based screening with ROCS to be highly reliable in identifying MGBs in large compound databases. This is shown by assessing screening quality parameters like a very high early enrichment rate or excellent AUC-values as well as proven by experimental testing of compounds found in the NCI database. Nineteen compounds out of 30 found by ROCS shape-based screening were found to be indeed active. Ten of the active compounds are novel MGBs according to a substance search in SciFinder. Interested in the drug-likeness of MGBs we used the same shape-based screening approach to rank known MGBs together with FDA-approved drugs and found MGBs to be well separated. We therefore suggest that MGB-likeness may be disadvantageous for approval of compounds intended to target proteins. We define MGB-likeness as high shape similarity to the query molecules presented and propose three simple descriptors (at least two charged groups, four nitrogen atoms and one aromatic ring) to complement shape-based screening or serve as rough filters for fast assessment of MGB-likeness.

ASSOCIATED CONTENT

Supporting Information. A list of PDB codes used for assembling the set of active MGBs, structures and rankings of insoluble or inactive compounds, results of screening experiments in PubChem and InterBioScreen databases and a list of 2D-descriptors considered for fast database filtering. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: Gudrun.Spitzer@uibk.ac.at (G.M.S.).

ACKNOWLEDGMENT

This work was supported by the Austrian Science Fund (project P19756). We thank Gerhard Klebe for providing the ITC device to determine the binding constants. We thank Uli Schmitz for fruitful discussions.

REFERENCES

(1) Gottesfeld, J. M.; Neely, L.; Trauger, J. W.; Baird, E. E.; Dervan, P. B. *Nature* **1997**, *387*, 202–205.
(2) Neidle, S. *Nat. Prod. Rep.* **2001**, *18*, 291–309.
(3) Stafford, R. L.; Arndt, H.-D.; Brezinski, M. L.; Ansari, A. Z.; Dervan, P. B. *J. Am. Chem. Soc.* **2007**, *129*, 2660–2668.
(4) Uil, T. G.; Haisma, H. J.; Rots, M. G. *Nucleic Acids Res.* **2003**, *31*, 6064–6078.
(5) Dervan, P. B.; Edelson, B. S. *Curr. Opin. Struct. Biol.* **2003**, *13*, 264–299.
(6) Dervan, P. B.; Doss, R. M.; Marques, M. A. *Curr. Med. Chem. Anticancer Agents* **2005**, *5*, 373–387.
(7) Melander, C.; Burnett, R.; Gottesfeld, J. M. *J. Biotechnol.* **2004**, *112*, 195–220.
(8) Stover, J. S.; Shi, J.; Jin, W.; Vogt, P. K.; Boger, D. L. *J. Am. Chem. Soc.* **2009**, *131*, 3342–3348.
(9) Chenoweth, D. M.; Harki, D. A.; Phillips, J. W.; Dose, C.; Dervan, P. B. *J. Am. Chem. Soc.* **2009**, *131*, 7182–7188.
(10) Nickols, N. G.; Jacobs, C. S.; Farkas, M. E.; Dervan, P. B. *ACS Chem. Biol.* **2007**, *2*, 561–571.
(11) Anthony, N. G.; Breen, D.; Clarke, J.; Donoghue, G.; Drummond, A. J.; Ellis, E. M.; Gemmell, C. G.; Helesbeux, J.-J.; Hunter, I. S.; Khalaf, A. I.; Mackay, S. P.; Parkinson, J. A.; Suckling, C. J.; Waigh, R. D. *J. Med. Chem.* **2007**, *50*, 6116–6125.
(12) Reddy, B.; Sondhi, S.; Lown, J. *Pharmacol. Ther.* **1999**, *84*, 1–111.
(13) Boger, D. L.; Dechantsreiter, M. A.; Ishii, T.; Fink, B. E.; Hedrick, M. P. *Bioorg. Med. Chem.* **2000**, *8*, 2049–2057.
(14) Boger, D. L.; Fink, B. E.; Hedrick, M. P. *J. Am. Chem. Soc.* **2000**, *122*, 6382–6394.
(15) Goodwin, K. D.; Lewis, M. A.; Tanious, F. A.; Tidwell, R. R.; Wilson, W. D.; Georgiadis, M. M.; Long, E. C. *J. Am. Chem. Soc.* **2006**, *128*, 7846–7854.
(16) Rinehart, K. L.; Holt, T. G.; Fregeau, N. L.; Keifer, P. A.; Wilson, G. R.; Perun, T. J.; Sakai, R.; Thompson, A. G.; Stroh, J. G.; Shield, L. S. *J. Nat. Prod.* **1990**, *53*, 771–792.
(17) Spitzer, G. M.; Wellenzohn, B.; Laggner, C.; Langer, T.; Liedl, K. R. *J. Chem. Inf. Model.* **2007**, *47*, 1580–1589.
(18) Spitzer, G. M.; Wellenzohn, B.; Markt, P.; Kirchmair, J.; Langer, T.; Liedl, K. R. *J. Chem. Inf. Model.* **2009**, *49*, 1063–1069.
(19) Rohs, R.; West, S. M.; Sosinsky, A.; Liu, P.; Mann, R. S.; Honig, B. *Nature* **2009**, *461*, 1248–1253.
(20) Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S. *Annu. Rev. Biochem.* **2010**, *79*, 233–269.
(21) Grootenhuis, P.; Roe, D.; Kollman, P.; Kuntz, I. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 731–750.
(22) Campbell, N. H.; Evans, D. A.; Lee, M. P.; Parkinson, G. N.; Neidle, S. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 15–19.
(23) Rohs, R.; Bloch, I.; Sklenar, H.; Shakked, Z. *Nucleic Acids Res.* **2005**, *33*, 7048–7057.
(24) Srivastava, H. K.; Chourasia, M.; Kumar, D.; Sastry, G. N. *J. Chem. Inf. Model.* **2011**, *51*, 558–571.

(25) Evans, D. A.; Neidle, S. *J. Med. Chem.* **2006**, *49*, 4232–4238.
(26) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. *J. Chem. Inf. Model.* **2007**, *47*, 2182–2196.
(27) *ROCS*, version 2.4.2; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2009.
(28) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. *J. Med. Chem.* **2005**, *48*, 1489–1495.
(29) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. *J. Chem. Inf. Model.* **2007**, *47*, 2293–2302.
(30) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.
(31) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.
(32) *Omega*, version 2.3.2; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2008.
(33) Mills, J. E. J.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 607–622.
(34) Tabernero, L.; Verdaguer, N.; Coll, M.; Fita, I.; van der Marel, G. A.; van Boom, J. H.; Rich, A.; Aymami, J. *Biochemistry* **1993**, *32*, 8403–8410.
(35) Mann, J.; Baron, A.; Opoku-Boahen, Y.; Johansson, E.; Parkinson, G.; Kelland, L. R.; Neidle, S. *J. Med. Chem.* **2001**, *44*, 138–144.
(36) Uytterhoeven, K.; Sponer, J.; Van Meervelt, L. *Eur. J. Biochem.* **2002**, *269*, 2868–2877.
(37) Zhang, Q.; Dwyer, T. J.; Tsui, V.; Case, D. A.; Cho, J.; Dervan, P. B.; Wemmer, D. E. *J. Am. Chem. Soc.* **2004**, *126*, 7958–7966.
(38) Laughton, C.; Tanious, F.; Nunn, C.; Boykin, D.; Wilson, W.; Neidle, S. *Biochemistry* **1996**, *35*, 5655–5661.
(39) *Molecular Operating Environment (MOE)*, version 2009.10; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2009.
(40) Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(41) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.
(42) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. *J. Med. Chem.* **2005**, *48*, 2534–2547.
(43) Benesi, H. A.; Hildebrand, J. H. *J. Am. Chem. Soc.* **1949**, *71*, 2703–2707.
(44) Stephanos, J. J. *J. Inorg. Biochem.* **1996**, *62*, 155–169.
(45) Zhong, W.; Wang, Y.; Yu, J.-S.; Liang, Y.; Ni, K.; Tu, S. *J. Pharm. Sci.* **2004**, *93*, 1039–1046.
(46) Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L. N. *Anal. Biochem.* **1989**, *179*, 131–137.
(47) Lah, J.; Vesnaver, G. *J. Mol. Biol.* **2004**, *342*, 73–89.
(48) *MACCS Keys*; MDL Information Systems, Inc.: San Leandro, CA, 2002.
(49) Vlieghe, D.; Sponer, J.; Meervelt, L. V. *Biochemistry* **1999**, *38*, 16443–16451.
(50) Anne, J.; De Clercq, E.; Eyssen, H.; Dann, O. *Antimicrob. Agents Chemother.* **1980**, *18*, 231–239.
(51) *SciFinder*. https://scifinder.cas.org (accessed Dec 10, 2010).
(52) Hirt, R.; Fischer, R. *Patentschrift (Switz.)* 1968, CH 459172 19680913.
(53) Schauhan, P. M.; Iyer, R. N.; Bhakuni, D. S.; Shankhdhar, V.; Guru, P. Y.; Sen, A. B. *Indian J. Chem.* **1988**, *27B*, 38–42.
(54) Dann, O.; Fick, H.; Pietzner, B.; Walkenhorst, E.; Fernbach, R.; Zeh, D. *Liebigs Ann. Chem.* **1975**, *1*, 160–9.
(55) Goble, F. C. *J. Pharmacol. Exp. Ther.* **1950**, *98*, 49–61.
(56) Squire, C.; Clark, G.; Denny, W. *Nucleic Acids Res.* **1997**, *25*, 4072–4078.
(57) Dyatkina, N. B.; Roberts, C. D.; Keicher, J. D.; Dai, Y.; Nadherny, J. P.; Zhang, W.; Schmitz, U.; Kongpachith, A.; Fung, K.; Novikov, A. A.; Lou, L.; Velligan, M.; Khorlin, A. A.; Chen, M. S. *J. Med. Chem.* **2002**, *45*, 805–817.
(58) Tanious, F. A.; Laine, W.; Peixoto, P.; Bailly, C.; Goodwin, K. D.; Lewis, M. A.; Long, E. C.; Georgiadis, M. M.; Tidwell, R. R.; Wilson, W. D. *Biochemistry* **2007**, *46*, 6944–6956.
(59) Kamal, A.; Murali Mohan Reddy, P.; Rajasekhar Reddy, D.; Laxman, E. *Bioorg. Med. Chem.* **2006**, *14*, 385–394.
(60) Bostock-Smith, C. E.; Searle, M. S. *Nucleic Acids Res.* **1999**, *27*, 1619–1624.
(61) Carrondo, M. A. A. F. d. C. T.; Coll, M.; Aymami, J.; Wang, A. H. J.; Van der Marel, G. A.; Van Boom, J. H.; Rich, A. *Biochemistry* **1989**, *28*, 7849–7859.
(62) Neidle, S.; L. Rayner, E.; J. Simpson, I.; J. Smith, N.; Mann, J.; Baron, A.; Opoku-Boahen, Y.; R. Fox, K.; A. Hartley, J.; R. Kelland, L. *Chem. Commun.* **1999**, 929–930.
(63) Chen, X.; Ramakrishnan, B.; Sundaralingam, M. *J. Mol. Biol.* **1997**, *267*, 1157–1170.
(64) Labute, P. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
(65) Suckling, C. J. *Expert Opin. Ther. Patents* **2004**, *14*, 1693–1724.
(66) Peixoto, P.; Liu, Y.; Depauw, S.; Hildebrand, M.-P.; Boykin, D. W.; Bailly, C.; Wilson, W. D.; David-Cordonnier, M.-H. *Nucleic Acids Res.* **2008**, *36*, 3341–3353.
(67) Bray, P. G.; Barrett, M. P.; Ward, S. A.; de Koning, H. P. *Trends Parasitol.* **2003**, *19*, 232–239.
(68) Mathis, A.; Holman, J.; Sturk, L.; Ismail, M.; Boykin, D.; Tidwell, R.; Hall, J. *Antimicrob. Agents Chemother.* **2006**, *50*, 2185–2191.
(69) Vooturi, S. K.; Cheung, C. M.; Rybak, M. J.; Firestine, S. M. *J. Med. Chem.* **2009**, *52*, 5020–5031.
(70) Perilo, C. S.; Pereira, M. T.; Santoro, M. M.; Nagem, R. A. P. *Int. J. Biol. Macromol.* **2010**, *46*, 502–511.
(71) Haran, T. E.; Mohanty, U. *Q. Rev. Biophys.* **2009**, *42*, 41–81.