J Mol Model (2013) 19:1573–1582
DOI 10.1007/s00894-012-1722-7

ORIGINAL PAPER

Structural and energetic insights into sequence-specific interaction in DNA–drug recognition: development of affinity predictor and analysis of binding selectivity

Jingheng Ning · Weiwei Chen · Jiaojiao Li · Zaixi Peng · Jianhui Wang · Zhong Ni

Received: 26 September 2012 / Accepted: 3 December 2012 / Published online: 29 December 2012

Abstract Although the molecular mechanism and thermodynamic profile of a wide variety of chemical agents have been examined intensively in the past decades in terms of specific recognition of their protein receptors, to date the physicochemical nature of DNA–drug recognition and association still remains largely unexplored. The present study focused on understanding the structural basis, energetic landscape, and biological implications underlying the binding of small-molecule ligands to their cognate or noncognate DNA receptors. First, a new method to capture the structural features of DNA–drug complex architecture was proposed and then used to correlate the extracted features with binding affinity of the complexes. By employing this method, a statistical regression-based predictor was developed to quantitatively evaluate the interaction potency of drug compounds with DNA in a fast and reliable manner. Subsequently, we use the predictor to examine the binding behavior of a number of structure-available, affinity-known DNA–drug complexes as well as a large pool of randomly generated DNA decoys in complex with the same drugs. It was found that (1) as compared with protein–DNA recognition, small-molecule agents have relatively low specificity in selecting their cognate DNA targets from the background of numerous random decoys; (2) the abundance of A–T base pairs in the DNA core motif exhibits a significant positive correlation with the affinity of drug ligand binding to the DNA receptor; and (3) high affinity seems not to be closely related to high selectivity for a DNA-targeting drug, although high-affinity drug entities have a greater possibility of being ranked computationally as top binders. We hope that this work will provide a preliminary insight into the molecular origin of sequence-specific interactions in DNA–drug recognition.

Keywords DNA–drug recognition · Sequence-specific interaction · Affinity · Selectivity

Introduction

A significant fraction of therapeutic approaches currently employed for modulation of gene function rely on the interaction of low molecular weight chemical agents with DNA targets so as to alter gene expression and the biological synthesis of proteins [1, 2]. Over the past decades, a wide variety of DNA-targeting drugs have been developed for this purpose to treat various diseases such as cancer, malaria, AIDS and other viral, bacterial and fungal infections [3]. The molecular design of sequence-selective DNA binding agents permits recognition and targeting of this biopolymer, thereby creating the possibility of gene-directed chemotherapies [4, 5]. However, it should be borne in mind that designing drug entities for selective recognition of a specific DNA site amongst the whole repertoire of the genome is hugely challenging as compared with the design of drugs targeting proteins [6].

Traditionally, protein–drug binding is explained by either Fischer's lock and key theory or Koshland's induced-fit principle [7]. The extrapolation of such models to DNA–drug complexes is not straightforward since, unlike enzymes, DNA has no formal active sites. In addition, the

Electronic supplementary material The online version of this article (doi:10.1007/s00894-012-1722-7) contains supplementary material, which is available to authorized users.

J. Ning · W. Chen · J. Li · Z. Peng · J. Wang (✉)
School of Chemical and Biological Engineering, Changsha University of Science and Technology, Changsha 410004, China
e-mail: wangjh0909@163.com

Z. Ni (✉)
Institute of Life Sciences, Jiangsu University,
Zhenjiang 212013, China
e-mail: nizhong@ujs.edu.cn

chemical forces that govern the binding of drug ligands to protein and to DNA receptors are not consistent; protein–ligand complexes are stabilized by a variety of interactions such as hydrophobic, electrostatics, hydrogen bonds, etc., whereas DNA–ligand complexes are clearly dominated by electrostatic effects [8, 9]. Thus, the advancement of understanding sequence-specific DNA–drug recognition is fundamentally important for the rational design of improved new and safe drugs as we move toward personalized medicine [10]. In recent years, considerable efforts have been devoted to exploring the molecular mechanisms underlying specific DNA–drug recognition and its biological implications, which are considered to stem from a number of exquisite balances between diverse physicochemical factors, including shape complementarity [11], enthalpy-entropy compensations [12], electrostatic earnings and desolvation penalty [13], as well as direct and indirect readouts [14]. In this study, we aimed at a deeper understanding of the molecular origin of sequence-specific interaction in DNA–drug recognition, attempting to answer the questions like: (1) how well do drugs specifically interact with their cognate sites on DNA helices? (2) How does base compositional bias influence specific DNA–drug recognition? (3) Is enhancing affinity equal to improving specificity for DNA–drug binding?

To elucidate these issues, we required a fast and reliable DNA–drug affinity predictor in order to perform intensive structural analysis and energetic examination of the binding behavior of drug entities to their cognate DNA targets as well as to vast noncognate counterparts. While a wide number of protein–ligand binding analysis tools are available today [15], only very few methods have been exploited over the past years to facilitate quantitative affinity prediction for DNA–drug association [16–18]. Unfortunately, most of these methods appear unsuitable for this study because of either significant computational demands (i.e., the MD-based MM/PBSA analysis [19]) or their relatively low accuracy (i.e., the empirical scoring function [20]). Although some rigorous strategies to parse energetic components involved in the interaction of DNA with small-molecule compounds have been developed successfully, these usually involve complicated theoretical processes and/or consist of various energy decompositions that are not feasible for high-throughput applications [21]. In this respect, we herein propose a novel quantitative structure-affinity relationship (QSAR) approach with which to correlate structural features with the energetic profile of drug ligand binding to DNA receptors statistically, which could be regarded as a good compromise between computational accuracy and efficiency. Subsequently, the resultant predictor was applied to analyze the target-binding selectivity of a number of drug molecules in recognizing their cognate DNA partners from a vast array of randomly generated (noncognate) decoys. In the procedure, we also examined the structural basis and energetic mechanism of the affinity-specificity relationship associated with DNA–drug interactions, in order to elicit straightforward guidelines for the structure-based rational design of sequence-selective DNA-targeting drugs.

Materials and methods

DNA–drug complex structure and affinity data set

Small-molecule drug ligands bound noncovalently to DNA receptors can be categorized as one of three classical types: minor groove binder, major groove binder, and intercalator, of which the first is the most commonly found (Fig. 1) [22]. Here, we collected 48 minor groove binders in complex with their cognate DNA receptors, along with experimentally

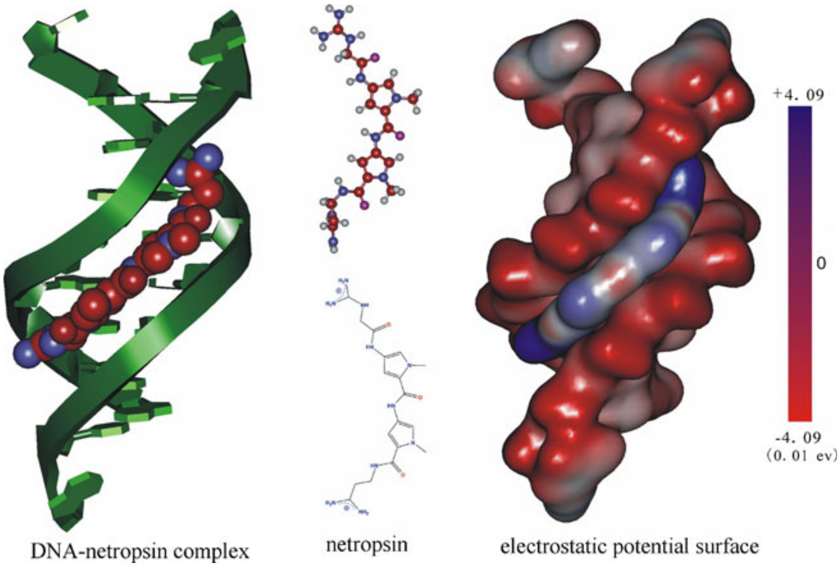


Fig. 1 An example of minor groove binder: the crystal structure of netropsin bound to a decamer d(CGCAATTGCG)2 (PDB: 261D), where the netropsin molecule is inlayed in and extended along the minor groove of the DNA double helix, forming intensive nonbonded interactions as electrostatic attractions, and van der Waals contacts, and desolvation effects between them

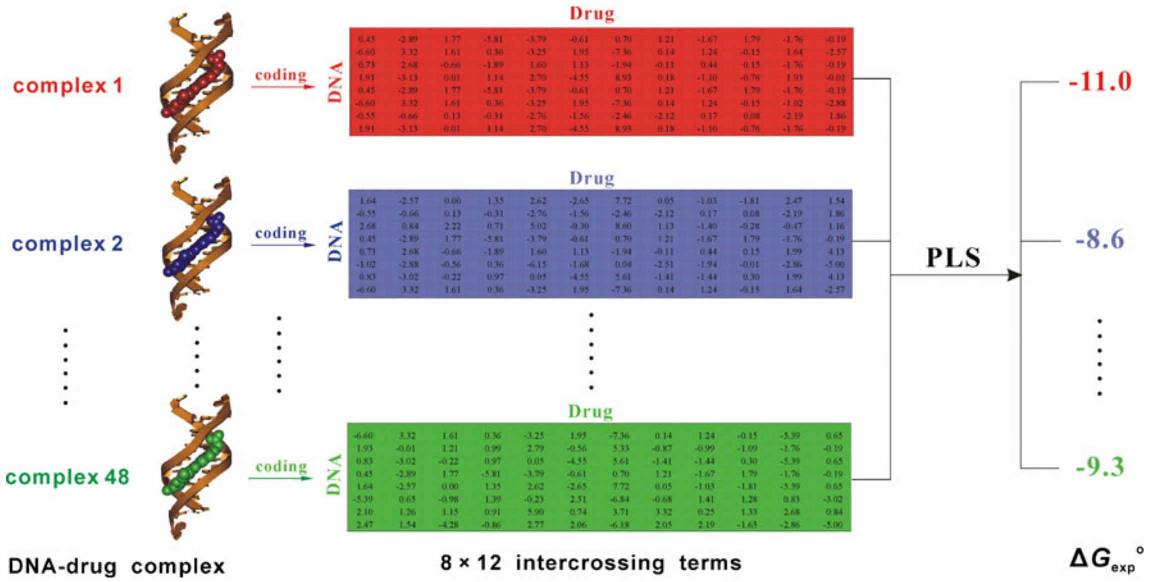
measured binding free energies $\Delta G_{\mathrm{exp}}^{\mathrm{o}}$ (collected from references [12, 23–39]). The three-dimensional (3D) structures of these complexes were either determined by X-ray crystallography at high resolution (<2.5 Å) or modeled by rigorous theoretical protocols. Information on the 48 DNA–drug complexes is tabulated in Table S1 in Supporting Information, in which the last 30 complex structures were modeled theoretically by Shaikh and Jayaram [40] and can be retrieved from the PDB FTP site [41]. In addition, the crystal structures of the remaining 18 complexes were treated as follows before use in subsequent analysis [40]: first, crystallized ions and water molecules were removed from, and hydrogen atoms added to, the complexes according to their structural information and ionized state. Thereafter, the systems were immersed into a 8 Å box of TIP3P waters to perform, in turn, 500-step hydrogen minimization, 5,000-step water minimization, 5,000-step all-atom minimization, and finally, 5,000-step free minimization with the AMBER9 force field [42] and GAFF parameterization [43]. In the minimization procedure, Na^+ counterions were added to neutralize the system. In this way, steric clashes and bond distortions involved in the crude crystal structures would be largely eliminated to achieve the nearest stable low-energy conformations.

Development of a structure-based QSAR predictor

Although numerous QSAR models have been proposed to predict the activity, toxicity and properties of various small-molecule compounds such as drugs, toxicants and surfactants, this widely used technique has only very limited applications to biomacromolecules. Here, we employed QSAR methodology to develop a quantitative predictor for the fast evaluation and reliable analysis of the binding affinity between DNA receptors and drug ligands based on 3D complex structure information (Fig. 2).

First, the heavy atoms contained in DNA and drug compounds were categorized roughly into 12 types in terms of chemical properties and hybridizion state [44], of which only 8 types were associated with DNA atoms (Table 1). According to the categorization scheme, at most 96 intercrossing terms between the 8 and 12 atom types of DNA and drug, respectively, can be generated to cover all the atom pairs involved in a DNA–drug complex. Subsequently, the pseudo potential of each atom pair in the complex was computed one-by-one using a distance-dependent Gaussian-type function modified from the classical CoMSIA method [45] (vide post), and then added to the corresponding one of the 96 intercrossing terms; the terms associated with missing atom types were always in zero.

A modified version of Gaussian-type function $U_{ij} = e^{-\alpha r_{ij}^2}$ was employed to describe the pseudo potential U_{ij} between atoms i and j, separated by distance r_{ij}, from DNA and drug, respectively. The attenuation factor α controls the sensitivity of U_{ij} to r_{ij} and was set to 0.3 according to the suggestion of Klebe et al. [45]. Unlike CoMSIA, here we did not consider the actual values of physicochemical properties of atoms, because we believed that the intrinsic peculiarity of the atoms were implicitly involved in their categorization. In other words, atom pairs with different characteristics would be separated into different terms and thus be distinguished in regression. In this way, the



0.45	-2.89	1.77	-5.81	-3.79	-0.61	0.70	1.21	-1.67	1.79	-1.76	-0.19
-6.60	3.32	1.61	0.36	-3.25	1.95	-7.36	0.14	1.24	-0.15	1.64	-2.57
0.73	2.68	-0.66	-1.89	1.60	1.13	-1.94	-0.11	0.44	0.15	-1.76	-0.19
1.91	-3.13	0.01	1.14	2.70	-4.55	8.93	0.18	-1.10	-0.76	1.93	-0.01
0.45	-2.89	1.77	-5.81	-3.79	-0.61	0.70	1.21	-1.67	1.79	-1.76	-0.19
-6.60	3.32	1.61	0.36	-3.25	1.95	-7.36	0.14	1.24	-0.15	-1.02	-2.88
-0.55	-0.66	0.13	-0.31	-2.76	-1.56	-2.46	-2.12	0.17	0.08	-2.19	1.86
1.91	-3.13	0.01	1.14	2.70	-4.55	8.93	0.18	-1.10	-0.76	-1.76	-0.19

1.64	-2.57	0.00	1.35	2.62	-2.65	7.72	0.05	-1.03	-1.81	2.47	1.54
-0.55	-0.66	0.13	-0.31	-2.76	-1.56	-2.46	-2.12	0.17	0.08	-2.19	1.86
2.68	0.84	2.22	0.71	5.02	-0.30	8.60	1.13	-1.40	-0.28	-0.47	1.16
0.45	-2.89	1.77	-5.81	-3.79	-0.61	0.70	1.21	-1.67	1.79	-1.76	-0.19
0.73	2.68	-0.66	-1.89	1.60	1.13	-1.94	-0.11	0.44	0.15	1.99	4.13
-1.02	-2.88	-0.56	0.36	-6.15	-1.68	0.04	-2.51	-1.94	-0.01	-2.86	-5.00
0.83	-3.02	-0.22	0.97	0.05	-4.55	5.61	-1.41	-1.44	0.30	1.99	4.13
-6.60	3.32	1.61	0.36	-3.25	1.95	-7.36	0.14	1.24	-0.15	1.64	-2.57

-6.60	3.32	1.61	0.36	-3.25	1.95	-7.36	0.14	1.24	-0.15	-5.39	0.65
1.93	-0.01	1.21	0.99	2.79	-0.56	5.33	-0.87	-0.99	-1.09	-1.76	-0.19
0.83	-3.02	-0.22	0.97	0.05	-4.55	5.61	-1.41	-1.44	0.30	-5.39	0.65
0.45	-2.89	1.77	-5.81	-3.79	-0.61	0.70	1.21	-1.67	1.79	-1.76	-0.19
1.64	-2.57	0.00	1.35	2.62	-2.65	7.72	0.05	-1.03	-1.81	-5.39	0.65
-5.39	0.65	-0.98	1.39	-0.23	2.51	-6.84	-0.68	1.41	1.28	0.83	-3.02
2.10	1.26	1.15	0.91	5.90	0.74	3.71	3.32	0.25	1.33	2.68	0.84
2.47	1.54	-4.28	-0.86	2.77	2.06	-6.18	2.05	2.19	-1.65	-2.86	-5.00

Fig. 2 The QSAR modeling procedure used to associate structural features with binding affinity for 48 DNA–drug complexes. The pseudo potential profiles of complexes were encoded one-by-one into 96 (8×12) intercrossing terms based on complex 3D structures, which were then correlated with experimentally measured binding affinity $\Delta G_{\mathrm{exp}}^{\mathrm{o}}$ using partial least squares (PLS) regression

Table 1 A coarse-grained categorization scheme of the atoms contained in DNA and drug molecules

Atom type	Description	Atom type	Description
C.3[a]	sp3 carbon	O.3[a]	sp3 oxygen
C.12	sp1 and sp2 carbon	O.2[a]	sp2 oxygen
C.ar [a]	Aromatic carbon	O.c[a]	Negatively charged oxygen
N.123[a]	sp1, sp2, and sp3 nitrogen	P.all[a]	All phosphorus atoms
N.c	Positively charged nitrogen	S.all	All sulfur atoms
N.ar [a]	Aromatic nitrogen	X.all	All halogen atoms

[a] Types associated with DNA atoms

modeling and prediction were largely simplified, and the exhaustive process traditionally used to pre-assign atomic parameters for the investigated DNA–drug complexes (and also the large number of decoy–drug complexes) was omitted.

By using the procedure described above, each DNA–drug complex can be generated with 96 descriptors parameterizing the structural properties and interaction behavior of each complex. The generated descriptors for all the 48 complexes come together to define an independent variable matrix X with size 48×96, which can be correlated statistically with a dependent variable vector y composed of 48 affinity values of these complexes by using the widely applied partial least squares (PLS) technique [46]. The obtained regression models were tested rigorously through 3-fold cross-validation, i.e., the 48 samples were randomly partitioned equally into three subgroups, and each subgroup was then used exactly once in a three-round testing procedure as validation data to assess the quality of the PLS model built upon other two subgroups.

In this study, PLS regression was implemented using the ChemoAC toolbox [47] running in MatLab platform; we modified this program in order to carry out variable selections.

Generation of drug-bound DNA decoys

In order to examine drug selectivity in targeting its cognate DNA receptor within the background of numerous noncognate sites, we need first to obtain large quantities of adducts of a drug molecule separately with vast DNA decoys in a swift manner. Therefore, a protocol that well considers the balance between accuracy and efficiency was introduced to generate and optimize the structures of decoy–drug adducts.

The minimized cognate DNA–drug complex crystal/model structure was then used as a template. The base pairs of DNA in the template were mutated randomly via a script executed within the 3DNA framework [48] to automatically generate noncognate adducts of DNA decoys with the drug. The complex structures were then optimized through the AMBER9 force field [42] in consideration of implicit GB/SA solvent effects [49], without any constraint and limited to 1,000 steps. A similar protocol has previously been used successfully to treat DNA–protein complexes, and hence we believed that this implicit solvation model-based minimization can be applied to DNA–drug systems as well. In fact, the strategy described above is rather fast, and can generate thousands of decoy–drug adducts from a template within several hours. This method is also effective since re-mutating several decoy–drug adducts back to their original state appeared to match well with corresponding templates, with small root-mean-square deviations (RMSD); for example, RMSD = 0.56 Å for the DNA–netropsin system (Fig. 3)

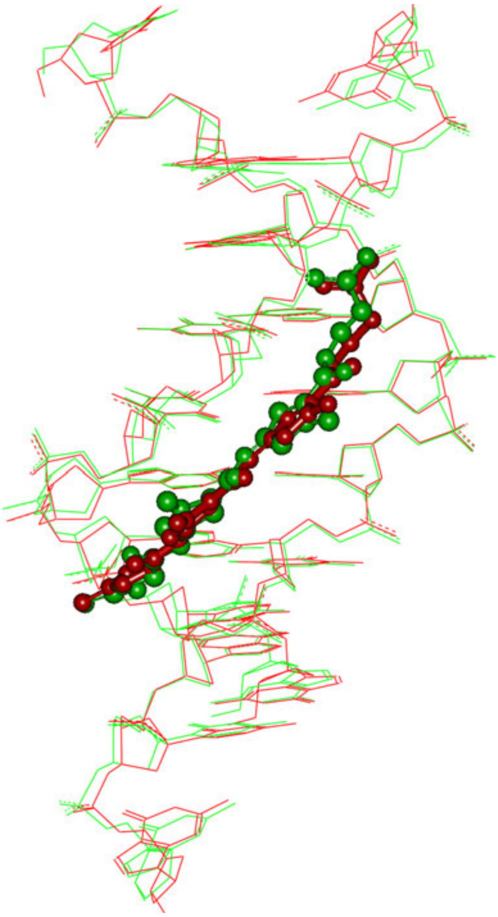


Fig. 3 Re-mutating decoy–netropsin adduct back to its original state (*green*) shows good superposition on corresponding template (*red*). The root-mean-square deviations (RMSD) value between two structures is 0.56 Å, mostly arising from the two ends of the DNA

of re-mutated structures superposed on templates, mostly contributed from the two ends of DNA.

Results and discussion

QSAR modeling and affinity predictor

As mentioned above, at most 96 (8×12) intercrossing terms can be generated for a DNA–drug complex. In fact, however, in most cases only a proportion of all terms associated with the atom types that really exist in studied systems are raised with actual values. According to our examination, 4 (O.c, P.all, S.all, and X.all, see Table 1) out of a total of 12 atom types are missing in currently studied drug molecules. Thus, we finally obtained 64 (8×8, separately from DNA and drug) valid intercrossing terms for each DNA–drug complex, which were then used as structural descriptors in subsequent QSAR modeling.

Based on the 48 complex samples, we conducted linear correlation between the structural descriptors and experimental affinity using PLS regression. The resultant QSAR model, although acceptable, was not very satisfactory, since it possessed a high performance in goodness-of-fit (coefficient of determination of fitting r^2=0.875) but only moderate predictability (coefficient of determination of 3-fold cross-validation q^2=0.540). As highlighted by Tropsha et al. [50], a predictable QSAR model should have (1) a relatively large q^2 value (>0.5) and (2) a small difference between r^2 and q^2 (r^2–q^2 close to or less than 0.2). It is evident that the built QSAR model satisfied only criterion 1 but skipped over criterion 2. Therefore, we attempted to further improve the model by performing variable selection—a sophisticated strategy that has been used widely in the QSAR community to enhance statistical quality for regression models [51].

Three variable selection methods, i.e., randomization (RD) [52], stepwise regression (SR) [53], and genetic algorithm (GA) [54], were adopted here. RD randomly creates thousands of variable subsets, establishes PLS models on these subsets, and selects the best from among these. SR performs forward selection to introduce significant (and simultaneously delete insignificant) variables one-by-one in terms of their contributions to a PLS model until the model's performance achieves its maximum. GA uses an evolutionary strategy to search variable space in a non-numerical manner to obtain near-optimal solutions for PLS variable combinations within an acceptable time-scale. The statistics, as well as the scatters of predicted against experimental affinity for different QSAR models are presented in Table 2 and Fig. 4. As can be seen, while the resulting r^2 is related directly to the number of variables engaged in models, the predictive power q^2 tells a different story, i.e., the effective prediction appears to associate with fewer variables, just as the M3 (28) and M4 (37) in Table 2. In fact, it is well known that too many variables can lead to overfitting problems to a regression model, viz. a high fitting ability on internal training samples but a low generalization capability to external unseen entities. Hence, a predictive q^2 should be considered as a gold standard with which to measure the quality of QSAR models. According to this criterion, M4 was finally selected to serve as the affinity predictor to perform subsequent analysis, which, apparently, obeyed the rule proposed by Tropsha et al. [50] (q^2=0.624>0.5 and r^2−q^2=0.188<0.2).

Table 2 Statistics of different quantitative structure-affinity relationship (QSAR) models

Model	Variable selection[a]	Number of variables	r^2 [b]	q_{rdm}^2 [c]	q_{clt}^2 [d]
M1	–	64	0.875	0.540	0.523
M2	RD	46	0.789	0.556	0.530
M3	SR	28	0.750	0.593	0.581
M4	GA	37	0.812	0.624	0.594

[a] *RD* Randomization; *SR* stepwise regression; *GA* genetic algorithm

[b] Coefficient of determination of fitting

[c] Coefficient of determination of randomized 3-fold cross-validation

[d] Coefficient of determination of cluster-based 5-fold cross-validation

Owing to the lack of availability of DNA–drug complexes with both known structure and determined affinity, our dataset involves significant redundancy across DNA-binding drugs. Therefore, we adopted SMILES strings to code drug molecules, and performed clustering on them using the threshold of the Tanimoto coefficient, 0.7. The Tanimoto coefficient is computed as the number of bits in common divided by the total number of bits. The Tanimoto coefficient can be expressed as: Tanimoto = BC/(B1 + B2 − BC). A Tanimoto of 1 indicates an identical molecule, while a Tanimoto of 0 will indicate that two molecules have nothing in common. Consequently, the drug ligands clustered into five groups. We found that the clustering result reflects mainly the size of the drug molecules. For example, bulky molecules such as Imidazole-Pyrrole Polyamide, Distamycin and Netropsin were clustered into one group, and the smaller Propamidine and Berenil into another group. We re-performed five-fold cross-validation on the five groups of clustering and found no significant difference with that performed previously with randomized three-fold cross-validation. For example, the model M1 had a q^2=0.540 with randomized three-fold cross-validation; this value was changed to 0.523 when performed with cluster-based five-fold cross-validation. The results arising from cluster-based five-fold cross-validation are also listed in Table 2 for comparison purposes.

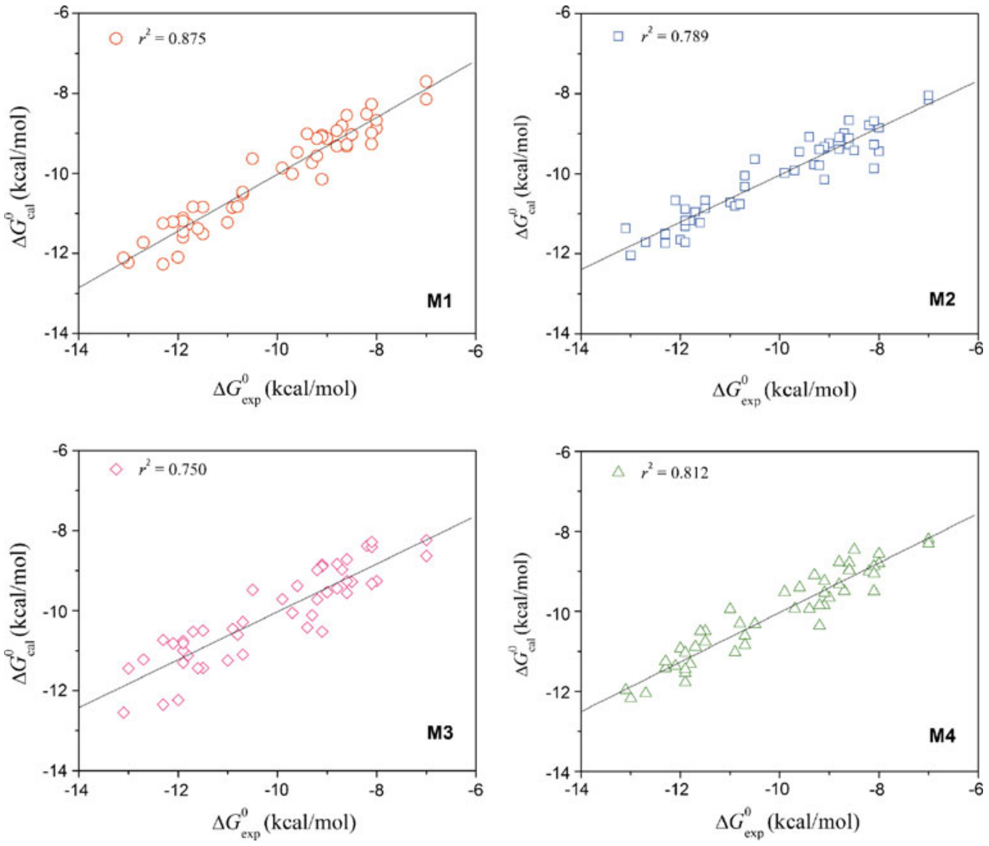


Fig. 4 Plots of calculated affinity against experimental values for the 48 DNA–drug complexes with models M1, M2, M3, and M4

Deeper analysis of affinity predictor M4

Here, we present a further examination of affinity predictor M4 to explore its structure and performance in interpreting and inferring the binding behavior of drug ligands, not only to their cognate DNA receptors but also to random decoys.

First, we analyzed the variable importance in the projection (VIP) of PLS [55], which is a direct measure of the relative contribution of GA-selected variables in M4 to DNA–drug binding affinity. The PLS VIP values of the 37 selected variables (plus 27 unselected variables) reverted into the 8×8 atom-type pairs between DNA and drug, and are illustrated as a heat map (Fig. 5), in which the hot red indicates a significant contribution of corresponding atom-type pairs to binding, while the pure black denotes variables of atom-type pairs not being selected by GA.

At a glance, it can be seen that the important pairs are mostly those associated with polar and charged atom types such as N.c, N.ar, O.2, O.c, P.all, etc., most of which locate at the upper right corner of the heap map. In contrast, the pairs involving nonpolar carbon atoms, such as C.3, C.12, and C.ar, were either ignored by GA or made only a modest contribution to binding, despite the fact that carbon is the most abundant element in DNA and drug compounds. The VIP heat map conveys clearly that polar and charged atoms play a central role in DNA–drug binding, and confer both stability and specificity to DNA–drug complex architecture by defining key nonbonded types at the interacting interface. In fact, the top five important atom-type pairs in M4 are all formed by O, N, and P atoms, i.e., O.c–N.c, O.2–N.ar, O.3–

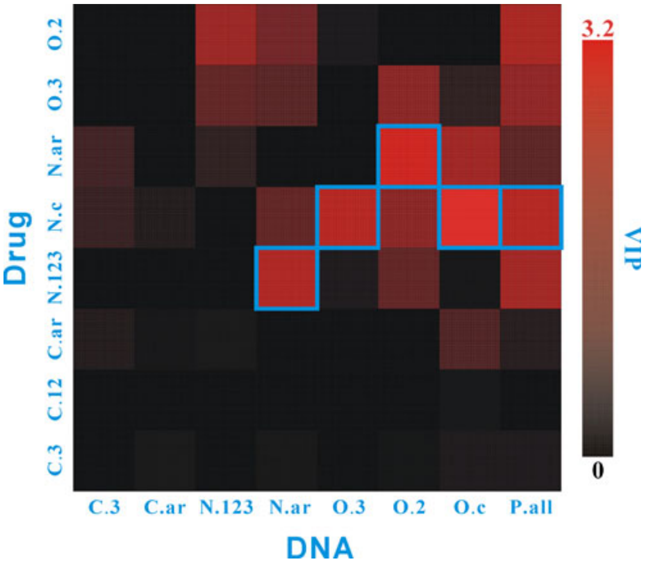


Fig. 5 The variable importance in the projection (VIP) heat map of variables in partial least squares(PLS) with variable selection by genetic algorithm (GA) (M4); each variable represents a atom-type pair between DNA and drug. *Hot red* Significant contribution of corresponding atom-type pairs to binding, *pure black* variables of atom-type pairs not being selected by GA. The top five important atom-type pairs are highlighted (*blue rim*)

N.c, N.ar–N.123, and P.all–N.c of DNA–drug (highlighted by blue rims in Fig. 5). It is expected that three out of the top five include charged atom types (i.e., O.c, N.c, and P.all), indicating the importance of electrostatic effects in dominating DNA–drug interactions [8]. In addition, the fact that most significant atom-type pairs include polar atoms implies that hydrogen bonding may also be a common and important phenomenon in DNA–drug binding. This is understandable considering that hydrogen bonding possesses a typical directionality that provides the molecular basis of sequence-specific DNA–drug recognition [56].

Second, we tested the capability of M4 in blind identification of cognate DNA–drug complexes from the background of numerous decoy–drug adducts. The 18 X-ray solved crystal structures of DNA–drug complexes were utilized to perform the test; each of the complexes was processed as follows: by using the protocol described in Materials and methods, the DNA in minimized complex structure was virtually mutated to 100 random decoys, followed by addressing structure optimization on the decoy–drug adducts. The binding affinities of the 100 noncognate adducts were then predicted using M4. Subsequently, we evaluated the relative (predicted) binding affinity of each cognate DNA–drug complex to 100 corresponding decoy–drug adducts, and the results obtained for the 18 samples are shown in Fig. 6. It can be seen that nearly half (8, pink bars) of the 18 samples were identified in the top 10 % highest affinities, 7 (green bars) of 18 were in the top 20 %, and remaining 3 (blue bars) in the top 50 %.

The test revealed some additional information about the predictor M4: (1) M4 is capable of properly estimating the binding affinity not only of cognate DNA–drug complexes but also of noncognate decoy–drug adducts; and (2) M4 can, from a statistical point of view, identify cognate complexes from a background of noncognate adducts reliably. Although the predictor M4 was preliminarily demonstrated to be effective in analyzing the binding behavior of drugs to both DNA targets and decoys, it is also worth noting that there were some cognate complexes that were not ranked as high-affinity binders as compared with corresponding decoy–drug adducts. The most typical of these is complex 10, which is formed by pentamidine [1,3-bis(4-amidinophenoxy)propane] bound to the dodecanucleotide d(CGCGAATTCGCG)2 and possesses the lowest binding affinity ($\Delta G_{\text{exp}}{}^{\text{o}}=-7.0$ kcal mol^{-1} in all 18 complexes. The other two low-ranked samples are complexes 9 and 14; both of these are also the weak binders ($\Delta G_{\text{exp}}{}^{\text{o}}=-8.2$ and -8.0 kcal mol^{-1} respectively). The findings suggested that either (1) our model appears to perform better on strong DNA–drug interactions than on weak ones, or (2) the low-affinity cognate complexes may actually not be the best choice for drug molecules to select their DNA targets. We discuss this point further in the next section.

Exploration of sequence-specific DNA–drug recognition

The selectivity of a drug molecule in sequence-specific recognition of its cognate DNA target can be defined as its capacity to pick up the target from the whole DNA repertoire it possibly sees in a cell. Here, we generated a large number of random DNA decoys to represent this repertoire, and calculated the differences in binding affinities of a drug entity to a cognate DNA target and to these noncognate decoys.

In order to gain more statistically significant conclusions, we decided to conduct exhaustive analyses for a few representative samples. Three cognate DNA–drug complexes, separately possessing high, moderate, and low affinities as well as distinct chemical structures of the drug molecules, were selected to carry out the examination of sequence-specific DNA–drug recognition, i.e., the drug ligands netropsin, berenil analogue [2,5-bis[4-(2-amidino)-phenyl] furan], and propamidine [1,3-bis(amidinophenoxy)propane] complexed with DNA receptors d(CGCAATTGCG)2, d(CGCGAATTCGCG)2, and d(CGCAAATTTGCG)2, respectively. For each cognate complex, 3,000 random DNA decoys bound with drug molecule were generated and the binding affinities of these noncognate decoy–drug adducts were then predicted using M4. The predicted affinity distributions of these adducts as well as the corresponding cognate complexes are shown in Fig. 7, which gives a straightforward insight into drug selectivity in sequence-

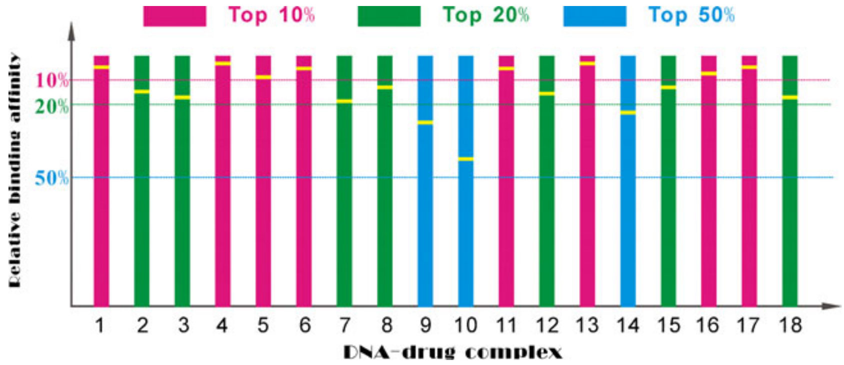


Fig. 6 Predicted affinities of 18 drug molecules binding to their cognate DNA targets relative to the values of the same drugs binding to 100 random decoys. The predicted affinities of cognate complexes ranked in top 10 %, top 20 %, and top 50 % are shown in *pink*, *green*, and *blue*, respectively

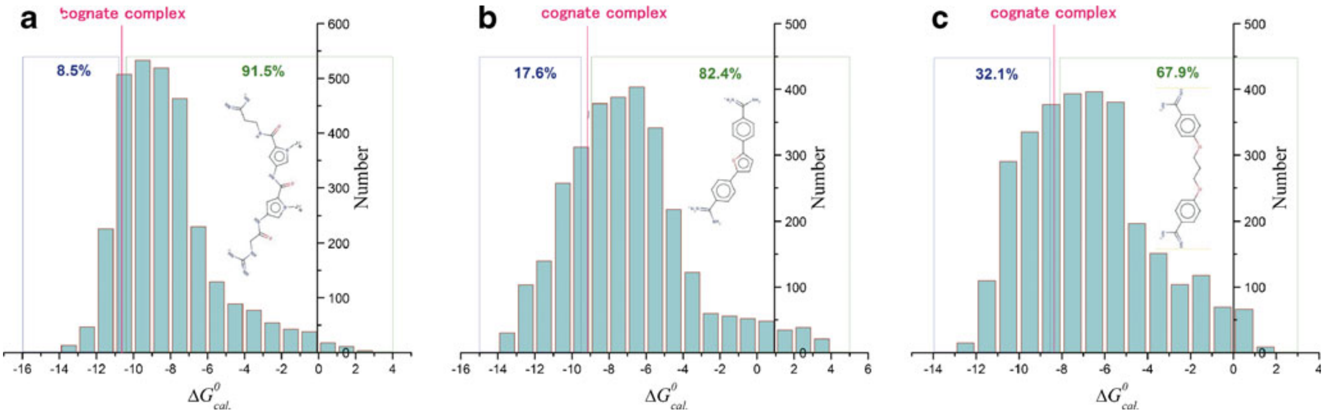


Fig. 7a–c Histogram distributions of the predicted binding affinities $\Delta G_{cal}{}^{o}$ (in 1 kcal mol^{-1} bins) of 3,000 randomly generated DNA decoys separately in complex with three drug ligands. **a** Netropsin, **b** 2,5-bis[4-(2-amidino)-phenyl]furan, **c** 1,3-bis(amidinophenoxy)propane

specific recognition of cognate DNA target from among a background of numerous noncognate decoys.

It can be seen from Fig. 7 that, although most noncognate adducts were predicted to have lower affinities than those of corresponding cognate complexes, there was also a substantial fraction of competitive decoys that can bind drugs with close or higher affinity as compared to cognate DNA targets. In fact, the differences between the affinity values of three cognate complexes and the averaged values of predicted affinities of 3,000 noncognate adducts are only 1.96, 2.23, and 2.17 kcal mol^{-1}, respectively, which are incapable of providing sufficient discrimination between the cognate and noncognate DNA–drug interactions. Thus, the specificity in DNA–drug recognition appears to be considerably lower than that of DNA–protein interactions; the latter usually undergo a significant decrease (3–10 kcal mol^{-1}) in binding affinity due to the slight change in native DNA sequence pattern [57]. In addition, selectivity seems not to be closely related to the affinity of drug ligands bound with DNA receptors. In other words, high affinity does not equal high specificity for a drug entity selecting its target. This is anticipated because the chemical forces that confer the majority of binding affinity for a drug compound are mostly those of nonspecific noncovalent interactions such as long-range electrostatic effects and hydrophobic potentials, which are not accurately coded by DNA sequence pattern.

Further, we surveyed the effect of base compositional bias on DNA–drug binding affinity. Here, only the core six residues of DNA that are in direct contact with drug ligands were considered in the survey. The base bias was quantified by the abundance of the base pair A–T in the core six-residue motif, of which the quantities can be enumerated as 0/6, 1/6, 2/6, 3/6, 4/6, 5/6, and 6/6, separately representing the ratio of A–T number to the total number of six base pairs in the motif. The scatter plots of the averaged affinity of 3,000 decoy–drug adducts against the abundance of base pair A–T in the core six-residue motif are shown in Fig. 8. As seen, a good linear correlation between the affinity and

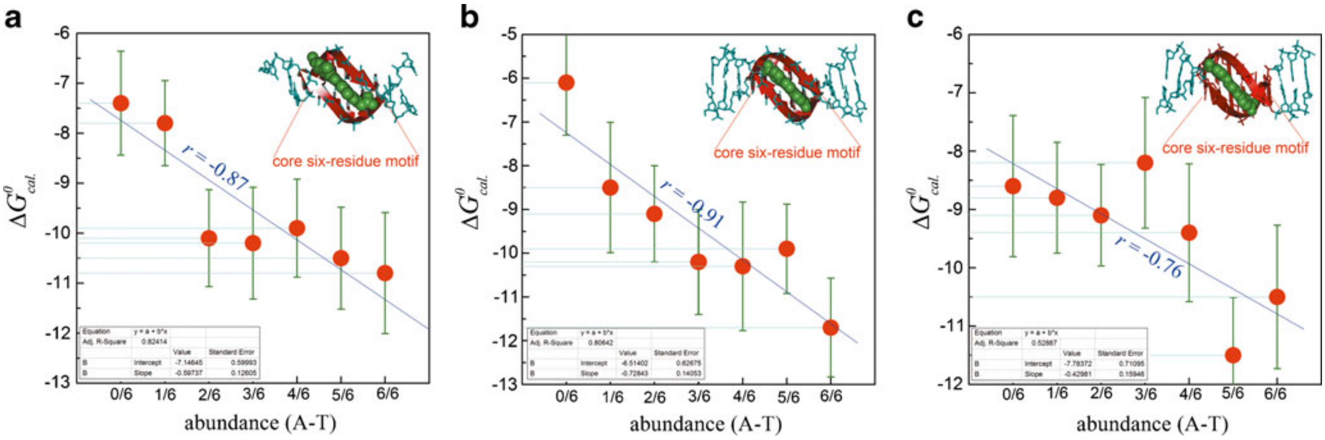


Fig. 8a–c Scatters of averaged binding affinity $\Delta G_{cal}{}^{o}$ against the abundance of the base pair A–T in the core six-residue motif for 3,000 randomly generated DNA decoys separately in complex with three drug ligands. **a** Netropsin, **b** 2,5-bis[4-(2-amidino)-phenyl]furan, **c** 1,3-bis(amidinophenoxy)propane. *Error bars* 95 % confidence intervals

A–T abundance emerged readily for the three investigated systems, with significant Pearson's correlation coefficients *r* of −0.87, −0.91, and −0.76; the negative value of *r* indicates a reverse relationship between the drug affinity and the A–T abundance of DNA; that is to say, a DNA helix with high A–T abundance is a promising candidate for drug ligands to target. This is expected since the A–T base pair does not introduce obvious steric hindrance to the DNA minor groove, which is used to accommodate drug ligands of the minor groove binder type, whereas G–C does [58]. Therefore, most minor groove binders discovered to date have A/T specificity. The noticeable dependence of drug affinity on A–T abundance is a marked feature of minor groove binders in selection of their favorable DNA fragments, which can be adopted as a coarse-grained rule to empirically exclude those DNA targeting candidates with low minor groove binding preference.

Conclusions

Understanding the structural basis and energetic mechanism of sequence-specific DNA–drug recognition is fundamentally important for the rational design of high-potency, low-toxicity and strong-selectivity DNA targeting agents against cancer, viral infection and other diseases. In order to achieve this goal, in the current study we attempted to evaluate the difference between the binding affinities of drug ligands to their cognate and noncognate DNA receptors. An efficient and reliable QSAR predictor was developed to estimate DNA–drug affinity based on complex 3D structure architecture, which was also validated rigorously via both statistical testing and the analysis of its biological implications. This predictor was employed to examine the binding behavior of a number of cognate DNA–drug complexes as well as large quantities of randomly generated DNA decoys in complex with the same drugs. By investigating the dependence of drug affinity on DNA sequence pattern and by comparing the interaction potencies of drug ligands with cognate DNA targets and with numerous random decoys, we found that DNA–drug recognition has lower sequence specificity as compared to that DNA recognized by proteins, and the specificity seems not to be closely related to the affinity of DNA targeting drugs. In addition, a significant correlation between the affinity of minor groove binders and the A–T abundance of DNA targets readily emerged. These findings may have implications for the physicochemical nature and molecular origin of sequence-specific DNA–drug recognition.

Acknowledgments This work was supported by the Hunan Provincial Natural Science Foundation (No. 11JJ6025), the Foundation for the Returned Overseas Chinese Scholar of State Education Ministry (39th), the National Natural Science Funds for Young Scholar of China (No. 31200602), and Jiangsu University Fund for Advanced Professionals (No. 1281330021).

References

1. Jarald E, Edwin S, Dubey P, Tiwari A, Thakre V (2004) Nucleic acid drugs: a novel approach. Afr J Biotechnol 3:662–666
2. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev 5:993–996
3. Bischoff G, Hoffmann S (2002) DNA-binding of drugs used in medicinal therapies. Curr Med Chem 9:321–348
4. Turner PR, Denny WA (2000) The genome as a drug target: sequence specific minor groove binding ligands. Curr Drug Targets 1:1–14
5. Fang YY, Morris VR, Lingani GM, Long EC, Southerland WM (2010) Genome-targeted drug design: understanding the netropsin–DNA interaction. Open Conf Proc J 1:157–163
6. Morávek Z, Neidle S, Schneider B (2002) Protein and drug interactions in the minor groove of DNA. Nucleic Acids Res 30:1182–1191
7. Koshland DE (1994) The key-lock theory and the induced fit theory. Angew Chem Int Ed Engl 33:2375–2378
8. Howerton SB, Nagpal A, Williams LD (2003) Surprising roles of electrostatic interactions in DNA–ligand complexes. Biopolymers 69:87–99
9. Zhou P, Huang J, Tian F (2012) Specific noncovalent interactions at protein–ligand interface: implications for rational drug design. Curr Med Chem 19:226–238
10. Evans WE, Relling MV (2004) Moving towards individualized medicine with pharmacogenomics. Nature 429:464–468
11. Boer DR, Canals A, Coll M (2009) DNA-binding drugs caught in action: the latest 3D pictures of drug–DNA complexes. Dalton Trans 3:99–414
12. Breslauer KJ, Remeta DP, Chou WY, Ferrante R, Curry J, Zaunczkowski D, Snyder JG, Marky LA (1987) Enthalpy-entropy compensations in drug–DNA binding studies. Proc Natl Acad Sci USA 84:8922–8926
13. Fenley MO, Harris RC, Jayaram B, Boschitsch AH (2010) Revisiting the association of cationic groove-binding drugs to DNA using a Poisson-Boltzmann approach. Biophys J 99:879–886
14. Arauzo-Bravo MJ, Sarai A (2008) Indirect readout in drug–DNA recognition: role of sequence-dependent DNA conformation. Nucleic Acids Res 36:376–386
15. Kirchmair J, Markt P, Distinto S, Schuster D, Spitzer GM, Liedl KR, Langer T, Wolber G (2008) The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery. J Med Chem 51:7021–7040
16. Rudnicki WR, Kurzepa M, Szczepanik T, Priebe W, Lesyng B (2000) A simple model for predicting the free energy of binding between anthracycline antibiotics and DNA. Acta Biochim Pol 47:1–9
17. Wang H, Laughton CA (2007) Molecular modelling methods for prediction of sequence-selectivity in DNA recognition. Methods 42:196–203
18. Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. Science 308:1424–1428
19. Shaikh SA, Ahmed SR, Jayaram B (2004) A molecular thermodynamic view of DNA–drug interactions: a case study of 25 minor-groove binders. Arch Biochem Biophys 429:81–99
20. Pérez-Montoto LG, Santana L, González-Díaza H (2009) Scoring function for DNA–drug docking of anticancer and antiparasitic compounds based on spectral moments of 2D lattice graphs for molecular dynamics trajectories. Eur J Med Chem 44:4461–4469

21. Haq I, Jenkins TC, Chowdhry BZ, Ren J, Chaires JB (2000) Parsing free energies of drug–DNA interactions. Methods Enzymol 323:373–405
22. Pindur U, Jansen M, Lemster T (2005) Advances in DNA–ligands with groove binding, intercalating and/or alkylating activity: chemistry, DNA-binding and biology. Curr Med Chem 12:2805–2847
23. Czarny A, Boykin DW, Wood AA, Nunn CM, Neidle S, Zhao M, Wilson WD (1995) Analysis of van der Waals and electrostatic contributions in the interactions of minor groove binding benzimidazoles with DNA. J Am Chem Soc 117:4716–4717
24. Haq I, Ladbury JE, Chowdhry BZ, Jenkins TC, Chaires JB (1997) Specific binding of Hoechst 33258 to the d(CGCAAATTTGCG)2 duplex: calorimetric and spectroscopic studies. J Mol Biol 271:244–257
25. Brown DG, Sanderson MR, Skelly JV, Jenkins TC, Brown T, Garman E, Stuart DI, Neidle S (1990) Crystal structure of a berenil-dodecanucleotide complex: the role of water in sequence-specific ligand binding. EMBO J 9:1329–1334
26. Haq I (2002) Thermodynamics of drug–DNA interactions. Arch Biochem Biophys 403:1–15
27. Brown DG, Sanderson MR, Garman E, Neidle S (1992) Crystal structure of a berenil-d(CGCAAATTTGCG) complex. An example of drug–DNA recognition based on sequence-dependent structural features. J Mol Biol 226:481–490
28. Rentzeperis D, Marky LA (1995) Interaction of minor groove ligands to an AAATT/AATTT site: correlation of thermodynamic characterization and solution structure. Biochemistry 34:2937–2945
29. Marky LA, Breslauer KJ (1987) Origins of netropsin binding affinity and specificity: correlations of thermodynamic and structural data. Proc Natl Acad Sci USA 84:4359–4363
30. Laughton CA, Tanious F, Nunn CM, Boykin DW, Wilson WD, Neidle S (1996) Structural origins of enhanced DNA-binding affinity. Biochemistry 35:5655–5661
31. Nunn CM, Jenkins TC, Neidle S (1993) Crystal structure of d (CGCGAATTCGCG) complexed with propamidine, a short-chain homologue of the drug pentamidine. Biochemistry 32:13838–13843
32. Edwards KJ, Jenkins TC, Neidle S (1992) Crystal structure of a pentamidine–oligonucleotide complex: implications for DNA-binding properties. Biochemistry 31:7104–7109
33. Sriram M, van der Marel GA, Roelen HL, van Boom JH, Wang AH (1992) Structural consequences of a carcinogenic alkylation lesion on DNA: effect of O6-ethylguanine on the molecular structure of the d(CGC[e6G]AATTCGCG)-netropsin complex. Biochemistry 31:11823–11834
34. Balendiran K, Rao ST, Sekharudu CY, Zon G, Sundaralingam M (1995) X-ray structures of the B-DNA dodecamer d (CGCGTTAACGCG) with an inverted central tetranucleotide and its netropsin complex. Acta Crystallogr D 51:190–198
35. Larsen TA, Goodsell DS, Cascio D, Grzeskowiak K, Dickerson RE (1989) The structure of DAPI bound to DNA. J Biomol Struct Dyn 7:477–491
36. Mazur S, Tanious FA, Ding D, Kumar A, Boykin DW, Simpson IJ, Neidle S, Wilson WD (2000) A thermodynamic and structural analysis of DNA minor-groove complex formation. J Mol Biol 300:321–337
37. Chaires JB (2006) A thermodynamic signature for drug–DNA binding mode. Arch Biochem Biophys 453:26–31
38. Lombardy RL, Tanious FA, Ramachandran K, Tidwell RR, Wilson WD (1996) Synthesis and DNA interactions of benzimidazole dications which have activity against opportunistic infections. J Med Chem 39:1452–1462
39. Pilch DS, Poklar N, Gelfand CA, Law SM, Breslauer KJ, Baird EE, Dervan PB (1996) Binding of a hairpin polyamide in the minor groove of DNA: sequence-specific enthalpic discrimination. Proc Natl Acad Sci USA 93:8306–8311
40. Shaikh SA, Jayaram B (2007) A swift all-atom energy-based computational protocol to predict DNA–ligand binding affinity and ΔT_{m}. J Med Chem 50:2240–2244
41. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542
42. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197
43. Wang J, Wolf RM, Caldwell JW, Kollman AP, Case DA (2004) Development and testing of a general Amber force field. J Comput Chem 25:1157–1174
44. Zhang C, Liu S, Zhu Q, Zhou Y (2005) A knowledge-based energy function for protein–ligand, protein–protein, and protein–DNA complexes. J Med Chem 48:2325–2335
45. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
46. Boulesteix AL, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 8:32–44
47. Vandeginste BGM, Smeyers-Verbeke J (2007) ChemoAC: its contribution to the advancement of chemometrics. J Chemometr 21:257–262
48. Lu XJ, Olson WK (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc 3:1213–1227
49. Qui D, Shenkin PS, Hollinger FP, Still WC (1997) The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate born radii. J Phys Chem A 101:3005–3014
50. Golbraikh A, Tropsha A (2002) Beware of q^2! J Mol Graph Model 20:269–276
51. Tsygankova IG (2008) Variable selection in QSAR models for drug design. Curr Comput Aid Drug Des 4:132–142
52. Katritzky AR, Dobchev DA, Slavov S, Karelson M (2008) Legitimate utilization of large descriptor pools for QSPR/QSAR models. J Chem Inf Model 48:2207–2213
53. Hocking RR (1976) The analysis and selection of variables in linear regression. Biometrics 32:1–49
54. Leardi R (2000) Application of genetic algorithm—PLS for feature selection in spectral data sets. J Chemometr 14:643–655
55. Wold S, Sjöström M, Eriksson L (2001) PLS regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130
56. Panigrahi SK, Desiraju GR (2007) Strong and weak hydrogen bonds in drug–DNA complexes: a statistical analysis. J Biosci 32:677–691
57. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS (2010) Origins of specificity in protein–DNA recognition. Annu Rev Biochem 79:233–269
58. Jenkins TC, Lane AN (1997) AT selectivity and DNA minor groove binding: modelling, NMR and structural studies of the interactions of propamidine and pentamidine with d(CGCGAATTCGCG)2. Biochim Biophys Acta 1350:189–204