Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 14 (2006) 5325–5332

Bioorganic & Medicinal Chemistry

Determination of hERG channel blockers using a decision tree

Michael M. Gepp and Michael C. Hutter*

Center for Bioinformatics, Saarland University, Building C7 1, PO Box 15 11 50, D-66041 Saarbruecken, Germany

Received 29 November 2005; revised 20 March 2006; accepted 24 March 2006
Available online 17 April 2006

Abstract—A decision tree approach for the in silico prediction of Torsade de Pointes (TdP)-causing drugs is presented. As TdP is frequently associated with QT-interval prolongation due to inhibition of the rapid activating delayed rectifier potassium channel in the heart (hERG channel), the properties of such blockers were investigated by molecular modeling and semi-empirical AM1 molecular orbital calculations. In addition, we derived a pharmacophoric SMARTS string using structural information from high affinity compounds. A corresponding search in the PubChem database identified several compounds that exhibit QT-interval prolonging activity that were not among our data set. This SMARTS string furthermore showed to be the most significant descriptor in the decision tree approach from which guidelines for the design of safe compounds are suggested.

1. Introduction

Cardiac arrhythmias due to induced QT-interval prolongation have received increased awareness as side effect of pharmaceutical treatment and have led to the withdrawal of several widely used drugs, particularly COX-2 inhibitors, between 2004 and 2005. Blockade of the delayed rectifier potassium current (I_{Kr}) is, however, a well-known mechanism of class III anti-arrhythmic agents leading to a desired moderate prolongation of the QT-interval in the electrocardiogram. Further extension of this time span beyond 440 ms is, however, likely to cause so-called Torsade de pointes (TdP) with the life-threatening risk of ventricular fibrillation.[1–4] Unlike the congenital long-QT syndrome, acquired forms have been associated almost exclusively with the human ether-á-go-go related gene (hERG) that encodes the pore-forming α-subunit of the rapid component of the delayed rectifier current.[5,6] Like other potassium channels this particular hERG channel (K_v11.1) has six transmembrane helices (denoted S1–S6) but exhibits a larger cavity that is able to accommodate even large molecules. These may enter in the open state and subsequently block the channel by preventing it to adopt the closed conformation. The responsible interaction is attributed to binding of the substances to specific residues at the end and the middle of the S6 helix.

Especially Phe656 and Tyr652 at the base of the pore helix have been identified by mutagenesis studies to be invoked in binding of high affinity compounds such as cisapride and terfenadine.[7,8] Further residues that were suggested to be involved are Val625, Thr623, and Gly648.[4] The results of computational docking techniques into homology models of the hERG channel also agree with the experimental findings.[9–13] Hydrophobic interaction with Phe656, for example, by π–π stacking of phenyl rings, is frequently observed. Similar interactions are also reported with Tyr652 where alternatively cation–π interactions with protonated ternary or positively charged quaternary nitrogen atoms are also possible (e.g., astemizole and clofilium, respectively).[12,14] Further appropriate compounds have been identified in pharmacophore models that usually comprise three hydrophobic and one protonable features.[4,15–21] On the other hand, it is obvious that a large number of drugs match these pharmacophoric features but do not induce QT-interval prolongation. Nevertheless, hypotheses from the various pharmacophore models are able to estimate binding affinity to the hERG channel.

Further and combined approaches including conventional QSAR as well as other statistical learning techniques that directly predict affinities for molecules have been reported by several groups.[22–29] For example, Yap

Abbreviations: AM1, Austin method 1; COX-2, cyclooxygenase 2; hERG, human ether-á-go-go related gene; QSAR, quantitative structure–activity relationship; SVM, support vector machine; TdP, Torsade de pointes.

Keywords: ADMETox; In silico prediction; Drug design; Molecular modeling; Medicinal chemistry.

* Corresponding author. Tel.: +49 681 302 64178; fax: +49 681 302 64180; e-mail: michael.hutter@bioinformatik.uni-saarland.de

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.03.043

and co-workers applied a support vector machine approach to distinguish between TdP-causing and non-causing molecules.[25] Recently, Rajamani et al.[26] showed that binding affinities are correlated with corresponding electrostatic and van der Waals energies from docking results.

To allow in silico screening of compounds, for example, as filtering tool in lead discovery and optimization to detect substances with possible implications of the interconnected issues of TdP, QT-interval prolongation, and hERG channel blocking, a fast method is desirable that is based only on information of the ligands. Therefore, we derived molecular properties and structural information from a series of compounds for use in a descriptor based decision tree approach.

2. Results and discussion

To obtain common structural binding motives of experimentally known hERG channel blockers, we initially derived a homology model based on the MthK channel from *Methanobacterium thermoautotrophicum* (PDB code 1LNQ)[30] for the purpose of molecular docking studies. Corresponding results (data not shown) did, however, not provide unambiguous hints of structural features that are indispensable for high affinity binding to the hERG channel. Thus, we searched for substructures present in known inhibitors. Several of such molecular fragments were suggested earlier, for example, by Tobita et al. [22,27] as well as by Roche and co-workers. Recently Song and Clark identified a large series of fragments to quantitatively express hERG binding affinity.[29] The reported substructures, however, comprise either substituents of low complexity that are common in many drugs, or very specific groups that are present only in a small fraction of TdP-causing agents.

The majority of hERG channel blockers contain a ternary nitrogen that can be part of a six-membered ring and have varying substituents. In the proximity of this nitrogen hydrophobic moieties like aromatic rings or aliphatic chains are found. Hydrophobic substituents of ternary nitrogens have been suggested as pharmacophores frequently.[2,4,16,17,19,20,28] We observed that the aliphatic chain length strongly varies, while the distance between the nitrogen and the first aromatic carbon ranges from 2 to 4 bonds. Roche et al. also noted that specific alkyl-benzene motives were present in 49% of their investigated hERG blockers.[22] Including these features we yielded the pharmacophoric substructure shown in Figure. 1. Note that the aromatic moiety includes any kind of aromatic ring system. To exclude a series of non-TdP-causing drugs containing peptide bonds, the presence of a vicinal oxygen was furthermore ruled out. Applying the SMARTS notation[31] this pharmacophoric substructure can be expressed as N([H,C])(C)(C[!O]* ~ * ~ c). This string is hereafter referred to as PHARM$. It contains the scaffold suggested by Roche et al. as well as two of the patterns used by Tobita et al. as substructures.[22,27] For the amide fragment (C–C(=O)–N) that is ruled out by PHARM$, Song and Clark found a negative contribution (−0.352) in a linear support vector regression equation for hERG binding affinity.[29]

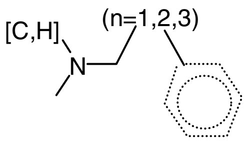

Figure 1. Graphical representation of the SMARTS string N([H,C])(C)(C[!O]* ~ * ~ c) termed PHARM$ that was derived from common structural features of hERG channel blockers.

Using this SMARTS string for substructure search in the PubChem database[51] we obtained among other well-known hERG channel blockers clobutinol, amodiaquine, fluphenazine, fentanyl, and melperone (see Fig. 2). All of these agents were neither part of the training nor the test set used in this study. The common antitussive clobutinol has been shown to induce TdP in a case of congenital long QT syndrome, whilst an IC_{50} value of 2.9 μM was measured in African green monkey kidney cells.[32] Similar effects were also observed during anesthesia with fentanyl.[33] For the anti-psychotics fluphenazine and melperone as well as for the anti-malarial amodiaquine various levels of heart rate changes and QT-interval prolongation have been reported recently.[34,35]

To further test the significance of PHARM$ for the prediction of TdP-causing substances, a decision tree algorithm was applied. This recursive partitioning scheme generates rules based on the numerical data of the available descriptors for each molecule. At each branching point the respective descriptor is chosen that is able to achieve the best separation possible. In this case, a classification of TdP-causing and non-TdP-causing drugs was desired. The obtained topology of the decision tree is shown in Figure 3, where the respective descriptor is denoted with an alphanumerical abbreviation that refers to Table 1. We generated two decision trees with a different number of available descriptors. For the generation of tree 1 a total of 123 descriptors with non-zero variance were available, whereas for tree 2 the additional descriptor SIMAST was applicable. This variable expresses the similarity of each substance based on its molecular fingerprint compared to the reference compound astemizole (see Section 4 for details). As can be seen in Table 1, SIMAST is used in favor of MR and LOGP two times, respectively. In this context SIMAST incorporates a higher information content than either LOGP or MR.

The pharmacophoric string PHARM$ is chosen as the very first descriptor allowing correct classification of already more than 71% of all compounds in both decision trees. Therefore, it is the most significant descriptor to separate TdP-causing from non-TdP-causing agents among the available variables. Furthermore, it can be assumed that occurrence of the corresponding substructure renders a substance as potential hERG channel inhibitor and likewise as TdP causing. Based on match-

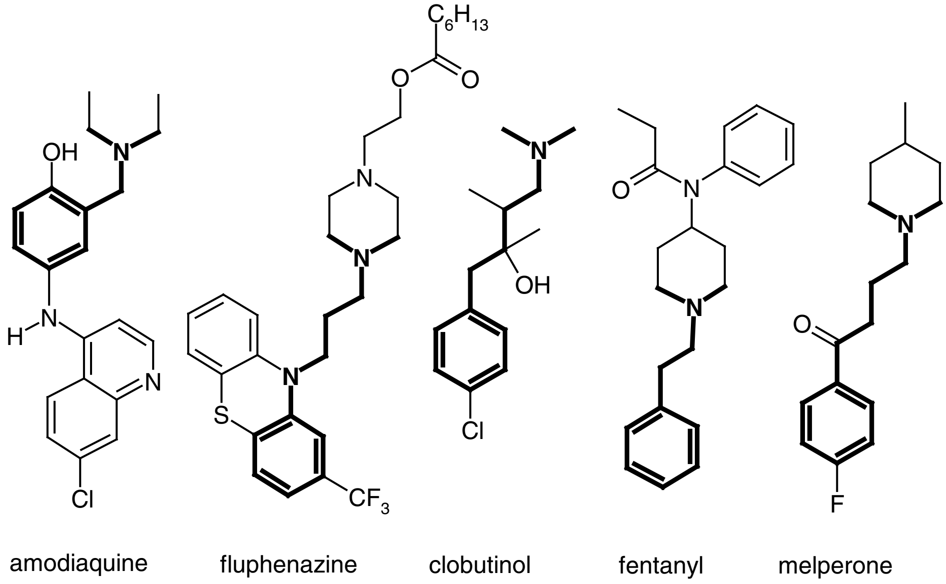

Figure 2. Using the SMARTS string given in Figure 1 for a search in the PubChem database yielded these additional drugs that turned out to be associated with QT-interval prolongation. The matching substructures are marked bold.

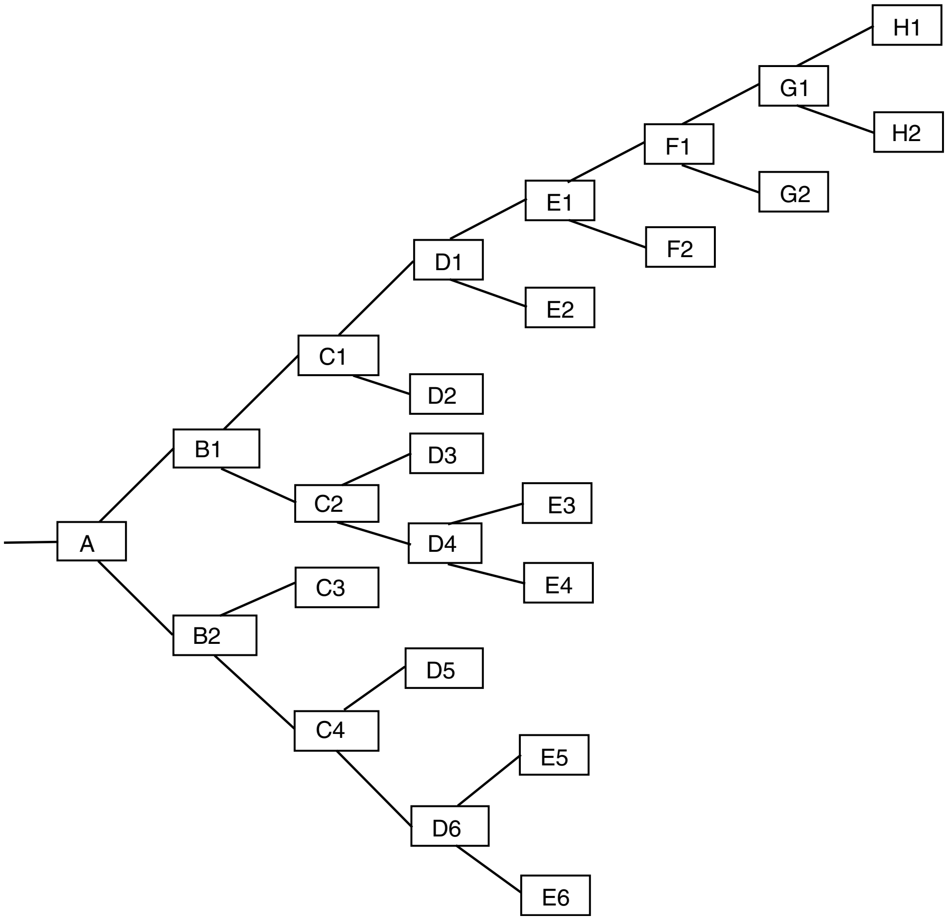

Figure 3. Topology of the decision trees for the prediction of TdP-causing activity (+/−). Descriptors used are shown in boxes and listed in Table 2 together with the corresponding margins.

ing of PHARM$ alone, only 13 compounds from the training and test set (3.8%) were misclassified as false positives (TdP+). These comprise butenafine, cefoperazone, clopidogrel, dicloxacillin, ethotoin, glimepiride, levamisole, piperacillin, raloxifene, ritonavir, vincristine, vinorelbine, and zileuton. A closer inspection of these

Table 1. Comparison of the derived decision trees[a]

Property	Tree 1		Tree 2	
	Descriptor	Margin[b]	Descriptor	Margin[b]
A	PHARM$	1	PHARM$	1
B1	HACSUR	0.05967	HACSUR	0.05967
B2	T1E	27.074	T1E	27.074
C1	HY	64	HY	64
C2	SGECA	17.0816	SGECA	17.0816
C3	DIPDENS	0.00385	DIPDENS	0.00385
C4	T2E	2.783	T2E	2.783
D1	HLSURF	0.33763	SIMAST	0.55556
D2	MDE23	17.0804	MDE23	17.0804
D3	MR	169.72	SIMAST	0.36842
D4	MR	43.65	MR	43.65
D5	MR	104.33	SIMAST	0.45455
D6	MR	134.09	MR	134.09
E1	CHBBA	16.09	HLSURF	0.33763
E2	MR	29.25	QSUM−	−4.42
E3	LOGP	3.45	SIMAST	0.23810
E4	QSUMN	−1.58	QSUMN	−1.58
E5	LOGP	4.04	SIMAST	0.57143
E6	QSUMN	−0.25	QSUMN	−0.25
F1	MGHBD	12.258	CHBBA	16.09
F2	MR	18.5	MR	29.25
G1	MR	268.85	MGHBD	12.258
G2	MR	98.7	MR	18.50
H1	—	—	MR	268.85
H2	—	—	MR	98.70
Accuracy				
Training set false positives	0.0%		0.0%	
Training set false negatives	8.3%		6.8%	
Training set total	91.7%		93.2%	
Test set false positives	2.7%		2.7%	
Test set false negatives	21.3%		17.3%	
Test set total	76.0%		80.0%	

[a] Descriptors used at branching points of the decision trees in Figure 3. All descriptors are listed in Table 2.

[b] Values below this number cause branching to the upward direction, otherwise downwards.

molecules revealed distinct structural motives: lactones (e.g., penicillin derivatives) and multiple (sometimes overlapping) occurrence of amide patterns. PHARM$ matches furthermore substances where nitrogen is in bridgehead position of two or more fused rings (e.g., vincristine and vinorelbine). Except for ethotoin and zileuton all others were correctly predicted as non-TdP causing at later branching points in the decision tree, most of them due to higher values of DIPDENS at node C3 (see below).

Following this first separation, the two subsequent layers of both decision trees contain identical descriptors (B1 to C4) accounting for molecular surface properties. The ability of a molecule to accept hydrogen bonds can be expressed by its surface portion of appropriate atoms such as nitrogen, oxygen, and sulfur. A ratio of less than 0.05967 for HACSUR indicates agents with a more hydrophobic surface as potential TdP causing. Heteroatoms in general cause a more unequal charge distribution in the molecule compared to hydrocarbons and a higher dipole moment. The corresponding atomic charges enter into the topological electronic descriptors T1E and T2E that are, however, less easy to interpret, but also determine the dipolar density (DIPDENS). Computed values above the margin of 0.00385 are found to be typically for polar non-TdP-causing drugs. Consequently, more hydrophilic substances are less likely to be TdP causing. This is also apparent from descriptors such as LOGP and the molar refractivity (MR). According to the first decision tree, substances are likely to be TdP causing for log *P* values higher than 3.45.

From the topological descriptors SGECA and HY, the latter indicates that non-TdP-causing molecules should possess less than 64 hydrogen atoms. A different choice of descriptors in the decision trees is seen from the third layer on (starting at D1), where SIMAST replaces HLSURF and particularly MR and LOGP several times. These variables are related to the hydrophobicity as expressed either by the calculated log *P* or due to the halogen content (HLSURF), and likewise the molar refractivity (MR). These indicate those drugs as TdP causing that exhibit a higher similarity to astemizole and/or possess a very high content of halogens (i.e., chloralhydrate). As the molar refractivity increases with growing size of the molecule it is difficult to assign a common margin. Therefore, we found values ranging from 29 up to 269. For typical blockers of the hERG channel, Buyck et al. suggested an upper barrier of 176 for MR.[23]

Descriptors retained at identical positions in the decision tree account for specific features, that is, the branching topology of the carbon skeleton (MDE23) and the sum of the atomic charges on the nitrogen atoms (QSUMN). Put together, these two variables resemble the frequently found pharmacophore models of hERG channel binders, for example, a protonable nitrogen atom surrounded by hydrophobic carbon atoms.[4] This is reflected by low values for the sum of the atomic charges on the nitrogen atoms (QSUMN) at E4 and E6. Except for SIMAST, the only additional variable appearing in tree 2 is QSUM− replacing MR at position E2. Seemingly, the sum of the negative atomic charges that predominately arises from heteroatoms shows a lower variance here, than does the molar refractivity. Interestingly, three descriptors common to both decision trees are found at different branching points: HLSURF, CHBBA, and MGHBD. SIMAST is used in favor of HLSURF at position D1, that itself replaces CHBBA at E1. CHBBA again supersedes MGHBD that reappears at the later branching point G1. This is not surprising, because MGHBD contains topological data only (the minimal Euclidean distance between two hydrogen-bond donor atoms) and CHHBA expresses the covalent hydrogen-bond basicity (a quantum chemical property), while SIMAST incorporates the fingerprint similarity. It can therefore be concluded that the information content decreases along the sequence SIMAST, HLSURF, MGHBD at least in this context. A similar criterion to MGHBD (the minimal geometric distance between two hydrogen-bond donor atoms) was recognized by Roche and co-workers who reported that 30% of the blockers in their study possessed pairs of hydrogen-bond donors separated by six bonds.[22] As MGHBD appears in a rather late stage in the decision tree, this distance seems to be of minor relevance.

The introduction of SIMAST as available descriptor improves the accuracy in predicting TdP-causing agents in tree 2. A closer investigation of the compounds found in terminating leaves showed that most of the improvement was due to the descriptor QSUM− at position E2 that was able to predict another four compounds (gatifloxacin, granisetron, moxifloxacin, and tizanidine) correctly as TdP causing that were misclassified in tree 1. This is due to the less negative partial charges on the heteroatoms in the vicinity of the protonable nitrogen compared to irinotecan and tolazamide that are non-TdP-causing compounds. Otherwise, the partitioning is rather similar in both trees. Most of the TdP-causing drugs appear in the terminal leaf after E6, whereas the vast majority of non-TdP substances are present in the terminal leaves of G1 and H1, respectively. All of the falsely predicted compounds are also found in these leaves (22 in tree 1 and 18 in tree 2, respectively). These comprise predominately drugs of low molecular weight such as epinephrine, dopamine, isoproterenol, and norepinephrine that possess only few chemical features. Obviously there are no descriptors present that would allow a further conclusive classification of those remaining substances in these leaves. Nevertheless, we observe a very low percentage of false positive predicted compounds (below 3%) for both the training and the test set (see Table 1). This indicates that drugs classified as TdP causing by this algorithm are highly likely to bind to the hERG channel.

In a comparable decision tree approach of Buyck et al., hERG binding compounds were indicated based on sufficient lipophilicity, a range of the molar refractivity, and a protonable nitrogen atom.[23] The applied descriptors comprised $C\log P$, MR, and the pK_a of the most basic nitrogen. Again the agreement with the published pharmacophore models is striking. While similarly derived $\log P$ and MR data appear in our two decision trees as well, the protonable nitrogen is described by two variables. First, the occurrence of an appropriate substructure according to the PHARM$ string and second, the sum of the atomic charges on the nitrogen atoms (QSUMN), whereas the first criterion is not stringent. Yap and co-workers applied several machine learning algorithms to classify TdP-causing agents.[25] Using the C4.5 decision tree program they yielded a lower overall accuracy for their test set (65.4%) than in our approach (80.0%). This can be attributed at least in part to the restriction to solely five descriptors, namely the LFER descriptors.[36,37] These have shown to be applicable to phase transfer processes since they incorporate hydrogen-bonding, lipophilicity, polarity, and molecular size properties. The best results (97.4% of TdP causing and 84.6% of non-TdP-causing compounds) were, however, achieved by using a support vector machine (SVM) approach. SVM algorithms generate a hyperplane that separates the two different classes in the underlying multidimensional space spanned by the descriptors. Although the prediction accuracy may benefit from this approach, the interpretation is rather complicated. In contrast to conventional QSAR equations, the effect upon changing the value of a specific descriptor cannot be predicted unequivocal. Likewise, the combined use of descriptors to partition compounds, as it can easily be done by a decision tree, is not possible and thus complicates the transfer of results to the design of new compounds with desired properties. Similarly, the use of descriptors reflecting the presence of certain chemical substituents (e.g., halogen atoms) and molecular substructures (such as SMARTS strings) is attractive for medicinal chemists.

According to our results, a compound is likely to bind to the hERG channel and cause TdP if it contains the substructure as given by SMARTS string PHARM$. The likelihood for TdP increases furthermore if the calculated $\log P$ is higher than 3.5 and an appropriate protonable basic nitrogen is present as indicated by a value for QSUMN lower than −0.25. Corresponding atomic charges may alternatively be derived from other charge fitting procedures than the electrostatic potential fit applied here. For compounds that do not match PHARM$ potential TdP risk is likely in the case of a predominately hydrophobic surface (HACSUR < 0.05967) and the appearance of 64 or more hydrogen atoms in the molecule, due to extensive hydrocarbon skeletons, few heteroatoms (N, O, and S), and very high halogen content. To design non-TdP-causing drugs, thus following substituents of ternary nitrogens should be avoided: unsubstituted phenyl rings and aromatic ring systems that contain only hydrophobic substituents such as fluorine, as well as unsubstituted hydrocarbon chains as linkers. Conversely, the appearance of multiple amide fragments may render a compound less likely to induce TdP.

3. Conclusion

From structural information of drugs with high binding affinity to the hERG potassium we derived a pharmacophoric SMARTS string. Its ability for in silico filtering of potential Torsade de Pointes-causing substances was tested by performing a substructure search in the PubChem database, that identified several drugs related to QT-interval prolongation that were not in our set of data before. Based on this SMARTS string alone, 71% of all compounds were classified correctly according to their TdP-causing potential. This renders the SMARTS string the most relevant descriptor in our decision tree approach. The presence of further variables supports the assumption of a protonable nitrogen atom surrounded by hydrophobic moieties as typical features of substances likely to bind to the hERG channel and/or cause TdP.

4. Method

4.1. Decision tree

The algorithm for deriving binary decision trees was previously developed in our laboratory and is described in full detail elsewhere.[45] It is based on recursive partitioning and creates an iterative branching topology in which the branch taken at each intersection is determined by a rule related to a descriptor of the molecule. Finally, each terminating leaf of the tree is assigned to a class.[46] Therefore, easy assumptions about the effect of each descriptor at a branching point are possible. To avoid excessive partitioning, here the maximum branching depth is limited to 8.

4.2. Drug data set

The chemical substances used in this study have been related to either blocking of the hERG channel, QT-prolongation, or TdP, for example, belonging to one of the classes 1–4.[22,47,48] Class 1 comprises drugs with a risk of TdP such as amiodarone and agents that already have been withdrawn, for example, astemizole, cisapride, grepafloxacin, and sertindole. Class 2 contains drugs with a possible risk of TdP that have been associated with TdP and/or QT-prolongation in some reports. Drugs to be avoided by congenital long QT patients are found in class 3, for example, adrenergic agonists such as ephedrine and isoproterenol. Finally, class 4 collects those compounds that have been weakly associated with TdP and/or QT-prolongation in some cases, but are unlikely to be a risk when used in usual recommended dosages and in patients without other risk factors. Substances in this class are predominately anti-depressants, anti-fungals, and antibiotics.[48]

The compounds of the training and test set used for the decision tree approach consist of a total of 339 molecules. The selection of TdP- and non-TdP-causing agents is essentially identical to those compiled by Yap et al.[25] but excluding following compounds found to be problematic: anakinra (large polypeptide), colestipol and cholestyramine (polymeric anion-exchange resins), trimetaphan camsilate (negative counterion), as well as fluconazole, miconazole, and troleandromycin. These three anti-infectives are reported by B. Fermini[3] to cause QT-interval prolongation but have been assigned as non-TdP-causing agents by Yap et al.[25] Thus in the context of the decision tree approach, molecules belonging to one of the classes 1, 2, or 3 are considered as TdP causing. Yap et al. have excluded most of the agents in class 4 due to the unclear association except for ampicillin, sulfamethoxazole, and trimethoprim that they classified as non-TdP causing. This gives rise to a ratio of TdP-causing to non-TdP-causing compounds of 67 to 197 in our training set and 37 to 38 in the test set, respectively.

It should be kept in mind that TdP is frequently associated with QT-prolongation, as TdP-causing agents are expected to act through hERG channel blockade and all drugs exhibiting high affinity to the hERG channel cause QT-prolongation.[49] The converse reasoning is certainly not true (e.g., procainamide and disopyramide cause TdP but are not potent inhibitors of the hERG channel).[50] Consequently these issues cannot be handled fully independently, at least in the context of in silico approaches.

4.3. Calculation of molecular descriptors

2D structures of the investigated compounds were obtained from the PubChem database, manually converted into 3D coordinates, and subsequently structurally optimized using the MM+ force field as implemented in HYPERCHEM for further use.[51,52] The obtained conformation was visually inspected for errors before generating the descriptors. A total of 155 descriptors and numerical properties were computed for each substance. Of those, 21 were discarded as they showed zero variance among their data range. The full list of all computed descriptors is given in the supplementary material. The majority of the variables were obtained from quantum chemical calculations using a modified version of the semi-empirical program package VAMP applying the AM1 Hamiltonian.[53,54] Compounds were energetically optimized to a gradient norm below $0.25\,\mathrm{kcal\,mol^{-1}\,\AA^{-1}}$ using the default eigenvector following algorithm.[55] Values for log P and the molar refractivity (MR) were computed with HYPERCHEM according to the approach of Viswanadhan et al.[56] Descriptors were generated from the respective output files using PERL scripts to obtain appropriate input data for the decision tree algorithm.

Table 2. Descriptors used in the decision tree

Variable	Descriptor definition	Reference
PHARM$	SMARTS string N([H,C])(C)(C[!O]* ~ * ~ c)	This study[a]
HACSUR	Ratio of surface of hydrogen-bond acceptor atoms (N, O, and S) to total surface	This study[a]
T1E	Topological electronic index using the number of non-hydrogen atoms	[38]
HY	Number of hydrogen atoms	—
SGECA	Sum of Kier and Hall E-states on carbon atoms based on geometrical distance	[39]
DIPDENS	Dipolar density (=dipole moment/molecular volume)	[40]
T2E	Topological electronic index using the number of bonds between non-hydrogen atoms	[38]
HLSURF	Ratio of surface on halogen atoms to total surface	This study[a]
SIMAST	Fingerprint similarity compared to astemizole	This study[a]
MDE23	Molecular distance-edge vector λ_{23}	[41]
MR	Molar refractivity	[56]
CHBBA	Covalent hydrogen-bond basicity	[42]
QSUM−	Sum of negative ESP charges	[43]
LOGP	Calculated water/*n*-octanol partition coefficient	[56]
QSUMN	Sum of atomic charges on nitrogen atoms	[44]
MGHBD	Minimal geometric distance between two hydrogen-bond donor atoms	This study[a]

[a] See Section 4 for details. A full list of all computed descriptors available to the decision tree algorithm is provided as supplementary material.

Table 3. SMARTS strings used for deriving the molecular fingerprints

ON(C)C	[!C;H]AA[CH2]A
OC(O)O	[!C;H]AAA[CH2]A
C#C	OC(N)C
NC(O)O	[!C][CH3]
NO	NAAAO
C[!C](C)(C)A	C=C
[!C][F,Cl,Br,I]	A[CH2]N
[S;R]	aa(a)a
NC(C)N	[!C]A([!C])[!C]
OS(O)N	AA(A)(A)A
SAN	O[!a]a
NN	[CH3]AA[CH2]A
[!C;H1]AA[A;H1][!C]	[N;R]
[!C;H1]AA[!C;H]	OCO
OSO	A[CH2]AA[CH2]A
ON(O)C	OA[CH2]A
a-a	N[!a]a
a-AS	[!C;R]
cn	[OH]
CC(C)(C)A	aaO
[!C;H][!C;H]	A[!a]@A[!a]
OAAO	A[CH2][CH2]A
[CH3]A[CH3]	A[!C](A)A
aaS	[NH1]
NAN	OC(C)O
NAAAN	C-N
A[CH2][!C;H]	[NH2]
[NH2]	[CH2][!C][CH2]
OAAAO	

Those descriptors that are used at branching points are listed in Table 1 together with the assigned margins. Table 2 shows a brief description of these variables and references that contain a more detailed description. Specific descriptors used are defined as follows: MGHBD is the minimal geometric distance between two atoms that possess hydrogen-bond donor properties. The molecular van der Waals surface was partitioned into special areas, where HLSURF is the ratio of the surface area on all halogen atoms to the total surface area and HACSUR is the ratio of the surface area belonging to atoms with hydrogen-bond acceptor characteristic (nitrogen, oxygen, and sulfur) to the total surface area, respectively.

To account for the presence of specific molecular substructures, the SMARTS strings notation was used.[31] For this purpose corresponding SMILES[57] for each compound were generated and checked for matching SMARTS strings (see Table 3) using the 'obgrep' command of Open Babel.[58] This gave rise to a 57 bit long fingerprint reflecting the occurrence of corresponding substructure fragments. These can be regarded as a subset of the MDL key set.[59] A similar subset was used by Ajay and co-workers to derive fingerprints.[60] For each compound the Tanimoto coefficient[61] of its fingerprint compared to that of the reference molecule astemizole was computed. Astemizole was chosen as it is one of the most potent blocking agents of the hERG channel and that with the highest binding affinity among our compound set. This similarity index was therefore termed SIMAST. Based on common structural features of hERG channel blockers (see Section 2) we derived the 'pharmacophoric' SMARTS string PHARM$ (see Table 2 and Figure 1.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2006.03.043.

References and notes

1. Vandenberg, J. I.; Walker, B.; Campbell, T. J. *Trends Pharm. Sci.* **2001**, *22*, 240–246.
2. Pearlstein, R.; Vaz, R.; Rampe, D. *J. Med. Chem.* **2003**, *46*, 2017–2022.
3. Fermini, B.; Fossa, A. A. *Nat. Rev. Drug Disc.* **2003**, *2*, 439–447.
4. Recanatini, M.; Poluzzi, E.; Masetti, M.; Cavalli, A.; De Ponti, F. *Med. Res. Rev.* **2005**, *25*, 133–166.
5. Splawski, I.; Shen, J.; Timothy, K. W.; Lehmann, M. H.; Priori, S.; Robinson, J. L.; Moss, A. J.; Schwartz, P. J.; Towbin, J. A.; Vincent, G. M.; Keating, M. T. *Circulation* **2000**, *102*, 1178–1185.

6. Sanguinetti, M. C.; Jinag, C.; Curran, M. E.; Keating, M. T. *Cell* **1995**, *81*, 299–307.
7. Mitcheson, J. S.; Chen, J.; Lin, M.; Culberson, C.; Sanguinetti, M. C. *Proc. Nat. Acad. Sci. U.S.A.* **2000**, *97*, 12329–12333.
8. Chen, J.; Seebohm, G.; Sanguinetti, M. C. *Proc. Nat. Acad. Sci. U.S.A.* **2002**, *99*, 12461–12466.
9. Sánchez-Chapula, J. A.; Navarro-Polanco, R. A.; Culberson, C.; Chen, J.; Sanguinetti, M. C. *J. Biol. Chem.* **2002**, *277*, 23587–23595.
10. Perry, M.; de Groot, M. J.; Helliwell, R.; Leishman, D.; Tristani-Firouzi, M.; Sanguinetti, M. C.; Mitcheson, J. *Mol. Pharm.* **2004**, *66*, 240–249.
11. Witchel, H. J.; Dempsey, C. E.; Sessions, R. B.; Perry, M.; Milnes, J. T.; Hancox, J. C.; Mitcheson, J. S. *Mol. Pharm.* **2004**, *66*, 1201–1212.
12. Sanguinetti, M. C.; Mitcheson, J. S. *Trends Pharm. Sci.* **2005**, *26*, 119–124.
13. Österberg, F.; Åqvist, J. *FEBS Lett.* **2005**, *579*, 2939–2944.
14. Fernandez, D.; Ghanta, A.; Kauffman, G. W.; Sanguinetti, M. C. *J. Biol. Chem.* **2004**, *279*, 10120–10127.
15. Mátyus, P.; Borosy, A. P.; Varró, A.; Papp, J. G.; Barlocco, D.; Cignarella, G. *Int. J. Quant. Chem.* **1998**, *69*, 21–30.
16. Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. *J. Med. Chem.* **2002**, *45*, 3844–3853.
17. Ekins, S.; Crumb, W. J.; Sarazan, D.; Wikel, J. H.; Wrighton, S. A. *J. Pharm. Exp. Therap.* **2002**, *301*, 427–434.
18. Pearlstein, R. A.; Vaz, R. J.; Kang, J.; Chen, X.-L.; Preobrazhenskaya, M.; Shchekotikhin, A. E.; Korolev, A. M.; Lysenkova, L.; Miroshnikova, O. V.; Hendric, J.; Rampe, D. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829–1835.
19. Du, L.-P.; Tsai, K.-C.; Li, M.-Y.; You, Q.-D.; Xia, L. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4771–4777.
20. Aronov, A. M.; Goldman, B. B. *Bioorg. Med. Chem.* **2004**, *12*, 2307–2315.
21. Cianchetta, G.; Li, Y.; Kang, J.; Rampe, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3637–3642.
22. Roche, O.; Trube, G.; Zuegge, J.; Pflimlin, P.; Alanine, A.; Schneider, G. *ChemBioChem* **2002**, *3*, 455–459.
23. Buyck, C.; Tollenaere, J.; Engels, M.; De Clerck, F. An in silico Model for Detecting Potential hERG Blocking. *EuroQSAR 2002, Designing Drugs and Crop Protectants: Processes, Problems, and Solutions*, 8–13 September 2002; Bournemouth, UK, pp 86–89.
24. Keserü, G. M. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773–2775.
25. Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. *Toxicol. Sci.* **2004**, *79*, 170–177.
26. Rajamani, R.; Tounge, B. A.; Li, J.; Reynolds, C. H. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1737–1741.
27. Tobita, M.; Nishikawa, T.; Nagashima, R. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886–2890.
28. Aronov, A. M. *Drug Disc. Today* **2005**, *10*, 149–155.
29. Song, M.; Clark, M. *J. Chem. Inf. Model.* **2006**, *46*, 392–400.
30. Jiang, Y.; Lee, A.; Chen, J.; Cadene, M.; Chait, B. T.; MacKinnon, R. *Nature* **2002**, *417*, 515–522.
31. Daylight Chemical Information Systems Inc., Suite 550, Aliso Viejo, CA 92656, see http://www.daylight.com for full details of SMILES and SMARTS.
32. Bellocq, C.; Wilders, R.; Schott, J.-J.; Louéreat-Oriou, B.; Boisseau, P.; Le Marec, H.; Escande, D.; Baró, I. *Mol. Pharmacol.* **2004**, *66*, 1093–1102.
33. Katz, R. I.; Quijano, I.; Barcelon, N.; Biancaniello, T. *Can. J. Anaesth.* **2003**, *50*, 398–403.
34. Stollberger, C.; Huber, J. O.; Finsterer, J. *Int. Clin. Psychopharmacol.* **2005**, *20*, 243–251.
35. Traebert, M.; Dumotier, B. *Expert Opin. Drug Saf.* **2005**, *4*, 421–431.
36. Abraham, M. H. *Chem. Soc. Rev.* **1993**, *22*, 73–83.
37. Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.
38. Osmialowski, K.; Halkiewicz, J.; Kaliszan, R. *J. Chromatogr. A* **1986**, *361*, 63–69.
39. Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. *J. Med. Chem.* **2002**, *45*, 3345–3355.
40. Mu, L.; Drago, R. S.; Richardson, D. E. *J. Chem. Soc., Perkin Trans. 2* **1998**, 159–167.
41. Liu, S.; Cao, C.; Li, Z. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387–394.
42. Cronce, D. T.; Famini, G. R.; De Soto, J. A.; Wilson, L. Y. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293–1301.
43. Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. *J. Mol. Model.* **1997**, *3*, 142–155.
44. Beck, B.; Glen, R. C.; Clark, T. *J. Comput. Chem.* **1997**, *18*, 744–756.
45. Andres, C.; Hutter, M. C. *QSAR Comb. Sci.* **2006**, *25*, 305–309.
46. Quinlan, J. R. *Machine Learning* **1986**, *1*, 81–106.
47. Fenichel, R. R., http://www.fenichel.net/pages/Professional/subpages/QT/qt.htm.
48. QT Drug Lists. The University of Arizona, Health Science Center, as of June, 30th 2005, http://www.qtdrugs.org.
49. Malik, M.; Camm, A. J. *Drug Saf.* **2001**, *24*, 323–351.
50. Muzikant, A. L.; Penland, R. C. *Curr. Opin. Drug Discov. Devel.* **2002**, *5*, 127–135.
51. The PubChem Database at the National Center for Biotechnology Information, http://pubchem.ncbi.nlm.nih.gov/.
52. HYPERCHEM, 6.02, Hypercube Inc., Gainsville, FL, 1999.
53. Rauhut, G.; Alex, A.; Chandrasekhar, J.; Steinke, T.; Sauer, W.; Beck, B.; Hutter, M.; Gedeck, P.; Clark, T., VAMP Version 6.5, Oxford Molecular, Erlangen, 1997.
54. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
55. Baker, J. *J. Comput. Chem.* **1986**, *7*, 385–395.
56. Viswanadhan, V. N.; Ghose, A. K.; Rebankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
57. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
58. Banck, M.; Bresciani, F.; Bréfort, J.; Clark, A.; Corkery, J.; Favre-Nicolin, V.; Fontaine, F.; Gillies, M.; Gillilan, R.; Goldman, B.; Hassinen, T.; Herger, B.; Hutchison, G.; Kebekus, S.; Kruus, E.; Leitl, E.; Mathog, D.; Morley, C.; Murray-Rust, P.; Nicholls, A.; Patchkovskii, S.; Reith, S.; Richard, L.; Sayle, R.; Shah, A.; Stahl, M.; Tolbert, B.; Walters, P.; Wolinski, P.; Wegner, J., Open Babel, Version 1.100.2, http://openbabel.sourceforge.net, 2005.
59. ISIS keys, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
60. Ajay, A.; Walters, W. P.; Murcko, M. A. *J. Med. Chem.* **1998**, *41*, 3314–3324.
61. Willet, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.