Indexing molecules for their hERG liability

Anwar Rayan a,*, Mizied Falah c, Jamal Rayin a, b, Beny Da’adoosh d, Sleman Kadan a, Hilal Zaid a, Amiram Goldblum d, a

a Drug Discovery Informatics Lab, QRC-Qasemi Research Center, Israel
b Computer Science Department, Al-Qasemi Academic College, P. O. Box 124, Baka El-Garbiah 30100, Israel
c The Institute of Applied Research, The Galilee Society, P. O. Box 330, Shefa-Amr 20200, Israel
d Drug Molecular Modeling and Drug Design Laboratory and the Alex Grass Center for Drug Design and Synthesis, Institute of Drug Research, School of Pharmacy of the Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 91120, Israel

ARTICLE INFO

Article history:
Received 29 March 2012
Received in revised form 25 April 2013
Accepted 27 April 2013
Available online 13 May 2013

Keywords:
Computational chemistry
Structure activity relations
hERG
Cardio-toxicity
Iterative Stochastic Elimination (ISE)

ABSTRACT

The human Ether-a-go-go-Related-Gene (hERG) potassium (K+) channel is liable to drug-inducing blockage that prolongs the QT interval of the cardiac action potential, triggers arrhythmia and possibly causes sudden cardiac death. Early prediction of drug liability to hERG K+ channel is therefore highly important and preferably obligatory at earlier stages of any drug discovery process. In vitro assessment of drug binding affinity to hERG K+ channel involves substantial expenses, time, and labor; and therefore computational models for predicting liabilities of drug candidates for hERG toxicity is of much importance. In the present study, we apply the iterative stochastic elimination (ISE) algorithm to construct a large number of rule-based models (filters) and exploit their combination for developing the concept of hERG Toxicity Index (ETI). ETI estimates the molecular risk to be a blocker of hERG potassium channel. The area under the curve (AUC) of the attained model is 0.94. The averaged ETI of hERG binders, drugs from CMC, clinical-MDDR, endogenous molecules, ACD and ZINC, were found to be 9.17, 2.53, 3.3, 1.98, −2.49 and −3.86 respectively. Applying the proposed hERG Toxicity Index Model on external test set composed of more than 1300 hERG blockers picked from chEMBL shows excellent performance (Matthews Correlation Coefficient of 0.89). The proposed strategy could be implemented for the evaluation of chemicals in the hit/lead optimization stages of the drug discovery process, improve the selection of drug candidates as well as the development of safe pharmaceutical products.

© 2013 Elsevier Masson SAS. All rights reserved.

1. Introduction

The human Ether-a-go-go Related Gene (hERG) potassium (K+) channel has a central role in regulating rhythm of the heart [1–4]. Unfortunately, this integral membrane protein that constitutes the potassium channel pore, responsible for the outward potassium (K+) current and cardiac myocyte repolarization has a promiscuous nature in its binding to drugs [2,4–9]. A wide variety of drugs of different structures used in cardiac and non-cardiac diseases can bind to hERG in cardiac tissue, lead to potassium blockade at the cellular level, and to a risk of sudden death [10,11]. The outcome of interactions between certain drugs and hERG is a prolongation of the QT interval in the action potential (AP) giving rise to the well-known phenomenon – acquired long QT syndrome (LQTS) [8,9,12,13]. Drugs prolonging QT interval has been found to predominantly (>80%) block the hERG K+ channel and thus inhibit potassium conduction [14,15]. This cardiac potassium channel is voltage-activated. It has a relatively large inner cavity that allows drug molecules to enter and bind tightly to the channel pore [3,8,16]. Aromatic amino acid residues arranged in two rings of four amino acids each within the central cavity enhance interactions with aromatic groups of drugs and thus may lead to channel blockage, prevent potassium conduction, and eventually cause the drug-acquired LQTS [17–21]. Patients with congenital LQTS, who have mutations in the hERG gene [22–25] exhibit similar inhibition in potassium conduction. Their cardiac hERG K+ channel inhibition is not triggered by drugs but rather by exercises or excitement leading to an abnormal heart rhythm (arrhythmia), fainting or loss of consciousness (syncope) or even sudden cardiac death known as

Abbreviations: hERG, human ether-a-go-go-related gene; ISE, Iterative Stochastic Elimination; MCC, Matthews’ Correlation Coefficient; EF, enrichment factor; HTS, high-throughput screening; ETI, hERG Toxicity Index; CMC, comprehensive medicinal chemistry database; MDDR, MDL drug data report; ACD, available chemicals directory.

* Corresponding authors. E-mail addresses: a_rayan@qsm.ac.il, anwar@genearrest.com (A. Rayan), amiram@vms.huji.ac.il (A. Goldblum).
Torsade de points (TdP) [23, 24, 26]. Thus, TdP phenomenon associated with the hERG failure and QT interval prolongation known as a distinctive polymorphic ventricular tachycardia or ventricular fibrillation [10, 21, 27, 28] could either be congenital due to hERG gene mutation, or acquired as a result of hERG potassium channel blockade due to strong binding of drugs [23].

The binding of hERG to drugs/blockers is usually weak and may not cause sustained arrhythmia; as such the heart can spontaneously revert to its normal rhythm. However, in some cases hERG-drug/blocker binding could frequently recur and may degenerate into TdP even in patients with structurally normal hearts [29, 30]. Hence, drugs targeting hERG channel and mainly hERG blockers are of great concern to the field of pharmaceutics. It is thus one of the main problems in drug design and development along with toxicity and ADME (Absorption, Distribution, Metabolism and Excretion) [31]. Since the blockage of hERG channels is a common mechanism to most drugs or compounds with reported QT-prolongation, hERG screening tests should be performed at a very early stage in drug development. hERG potassium channel blockers should be abandoned even if they are promising drugs for some disease condition.

For this reason, drug binding to hERG has been considered in many studies as a surrogate marker of proarrhythmic potency [32]. As such, drug safety agencies as well as regulatory authorities always call for assays of hERG binding affinity to new chemical entities [33, 34]. At present, IC_{50} for hERG K+ channel inhibition is considered a primary test for cardio-safety of any new drug candidate that can support go/no go decision making at early stages of drug development [35]. In addition, studies in silico have revolutionized the field of drug discovery as well as greatly contributed to the investigation and scrutiny of drug toxicity through its development [20]. Meanwhile, drugs that may block hERG, inducing cardiac arrhythmia and TdP are encountered more frequently and become a major concern for patients, physicians, and safety regulation authorities as well as to those who develop novel pharmaceuticals [9, 10]. Thus, it is important to ensure that these chemical entities display high therapeutic efficacy and have little or no toxicity. An inappropriate toxicity profile in humans is one of the major causes of failure of drug candidates in late clinical stages of drug development. Combinatorial chemistry techniques coupled with high throughput screening protocols have been developed to help pharmaceutical companies cope with an unprecedented number of chemicals. As a consequence, screening for undesirable toxicity properties in the early stages of drug development is now considered an essential paradigm [36]. Even though the early stage in vitro toxicity experiments reduce the probability of failure at the development stage, it is still time-consuming and resource-intensive. For this reason it became obvious to develop in silico methods for assessment of the potential for cardio-toxicity of new chemicals.

The promiscuity of hERG binding has been expressed in a decade of research reports, in which experimental information and models contributed to some confusion vis-à-vis the chemical characteristics of blockers. Prior to in silico modeling, the need for hERG blockers to have Van der Waals hydrophobic surface contact with Phe-656 of blockers. Prior to the early stage [37]. However, a paper by Aronov [38] criticized the “classic hERG motif” that appeared in several models, of “a basic nitrogen flanked by aromatic or hydrophobic groups attached with flexible linkers”, a model which was in full accord with the experiments mentioned above. Aronov suggested to consider the commonality of structure between various uncharged hERG blockers, and used a Vertex Pharmaceuticals in-house database of nearly 200 such molecules. Their pharmacophore is spanned by hydrophobic/ aromatic interactions and by a H-bond acceptor. The resulting recommendations to design drug candidates that would avoid hERG binding were to make them sufficiently polar (ClogP < 1) and to reduce their hydrophobicity or eliminate H-bond acceptors. Contrary to that, the addition of carboxylic groups to hERG binders was shown to diminish affinity by one to three log units, including candidates with a basic nitrogen. In a paper that combined authors from several pharmaceutical companies, Zhu et al. [39]. The authors noted that “it is remarkable that by simply tethering a carboxyl group on molecules regardless of where it is tethered, it generally suppresses their hERG binding affinity by more than two orders of magnitude”.

An interesting application of the Support Vector Machine Method with a pharmacophore type (3D) set of descriptors has shown success up to 72–73% in accurately predicting hERG blockers in an external test set [40] and the field of hERG toxicity models was reviewed in 2009 [41].

The previous two years have been marked by an increase in the publications that contribute further to the need to predict, in the preclinical phase, the potential for hERG toxicity. Bender and coworkers [42] used a very large data set of ligands (more than 2500) by applying support vector machines and linear discriminant analysis, to achieve a prediction accuracy of 80% in an external test set, and subsequently they picked 50 molecules from the Cambridge database predicted to be blockers of the channel and 10 that were predicted to be non-blockers. Out of the 50 expected blockers, 18 showed inhibition (IC_{50}) at 10 μM, and 4 of those showed low micromolar and even low nanomolar activity. All ten predicted inactives were indeed shown to be so. They propose their combined models to be used for pre-filtering of databases, and found very large numbers of potential hERG blockers in a few of the commercially available databases.

Other methods used to identify hERG toxicity were examined – ISIDA property labeling [43]; 4D Fingerprints from a conformational set of each molecule were used in addition to 2D and 3D descriptors for constructing binary classification QSAR models [44] and achieved high accuracy in several of the proposed models, although QSAR equations had relatively low correlation of up to 60%. However, mechanistic interpretation was enabled by the 2D descriptors, concluding that the presence of a polar negative group at a distance of 6–8 Å from a hydrogen bond donor in the compound is predicted to be a specific pharmacophore that increases blocking. That structural character should thus include a carboxylic acid moiety at that distance from an amine or other proton donor, and this is contrary to previous demonstrations of the overall effects of any carboxylic acid on potential hERG blockers. Finally, a very recent suggestion by Pastor and coworkers [45] combined docking simulations in two potassium channels and introduced a more elaborate electrophysiological model of cardiomyocytes and ventricular tissue, thus predicting directly the cardiotoxic effect of several compounds on electrocardiogram simulations.

The issue of predicting hERG toxicity is therefore not yet fully settled, and as we developed over the last few years a generic method that begins with classification and ends with numerical indexing of probabilities (described below) we applied this method to quantify the ability to predict molecular binding potentials.

Distinction between hERG blockers and non-blockers may be determined by using databases and statistical methods. Analysis of such databases may reveal characteristics of hERG blockers and facilitate prediction of the hERG affinity of other molecules. Physicochemical characters that are common to hERG liability may be discovered and used in analysis and predictions. However, the large number of potential physical descriptors of molecular structure requires special technologies that can deal with such complex problems. Clearly therefore, an improved method is required which
would overcome the limitations of the combinatorial nature of the problem of hERG toxicity evaluation in order to provide more accurate in silico testing of potential hERG blockers.

In the reported study herein, the Iterative Stochastic Elimination Approach (ISE) has been employed to distinguish between hERG blockers and non-blockers. ISE is already documented as a method for solving highly complex combinatorial problems including those of chemoinformatics [46]. The usefulness of this approach is due to the fact that a molecule’s (blocker/non-blocker) affinity to the hERG K⁺ channel is due to its chemical composition and depends on various descriptors, such as solubility, molecular weight, types of atoms and so forth. Our proposed approach for indexing molecules for their hERG liability exploits the combination of many rules-based models that are predicted by ISE, those allows us to develop the concept of a hERG Toxicity Index (ETI). High ETI for a molecule would indicate that it has a greater chance to be a blocker of hERG. Thus the potential risk of a molecule to be a blocker of hERG potassium channel is given by ETI. Such index could be attached to a single molecule and/or be used to differentiate between hERG blockers/non-blockers and/or be chosen for prioritization of molecules in combinatorial databases for drug design and development.

2. Methods

Literature was surveyed to collect three hundred molecules (see Supplementary material) that exhibit activity on hERG potassium channel with IC₅₀ less than 10 μM [47–52]. This low concentration reflects the need to identify toxicity as the concentrations of currently introduced drugs are in the micromolar range. Nine thousands molecules were selected randomly from the ZINC database to represent a set of presumably inactive molecules at hERG. The selected molecules from ZINC are highly diverse and were examined for their structural distance from known hERG blockers (Tanimoto index <0.2). We applied the Tanimoto Index (using structural keys) also to the set of hERG blockers and kept molecules that had a diversity <0.9.

In the current study, we describe a novel approach that could provide a discriminative model for verifying hERG liability and which may be employed for designing drug candidates without undesirable activity on hERG (potential cardio-safe drug candidates). For that purpose we utilized 2D-descriptors which are based on properties such as molecular weight, H-bond donors and acceptors, log P, solubility, types of atoms, charge distributions, surface area of hydrophilic and lipophilic nature and so forth (http://www.chemcomp.com/journal/descr.htm). The descriptors, defined as a function of the two-dimensional structure, were calculated by MOE2009.10 [53].

To assess the predictability and applicability of the proposed models, the data sets of blockers/non-blockers has been split 3 times into 66.7% (200/6000 hERG blockers/non-blockers) for the training set and 33.3% (100/3000 hERG blockers/non-blockers) for the test set, which corresponds to a leave-one third-out procedure. The training and test sets were generated by random picking.

2.1. Prediction modeling

We applied the iterative Stochastic Elimination algorithm [46,54] for constructing a model that is the basis for indexing hERG liability of molecules, by transforming the problem into a combinatorial one with many variables, each having many values. The optimal distinction between active and inactive molecules is achieved by searching, in multivariable space, the best sets of variables (“descriptors”) and the best ranges of each descriptor that “separate” actives from inactives. Since descriptors are generally interacting with each other, changes in the range of one descriptor could affect best range of another descriptor and the optimization process requires to consider all variables at a time for getting best filters. If each descriptor is independent of the others, then optimization may be performed for each descriptor separately. However, if there is a linkage between descriptors, so that they are not “orthogonal” to each other, optimization requires to consider more variables at a time. Applications of ISE for discriminations in structural biology [55–57] and for prediction of oral bioavailability druglikeness [46,58] were previously described. A flowchart of the optimization is shown in Fig. 1. Following are the main steps in the application of ISE:

1) Select sets of hERG blockers and of assumed non-blockers and divide each into training and test sets.
2) Calculate the values of the 2D descriptors of interest for all the molecules in the two sets, using MOE.
3) Histograms are constructed for hERG blockers and non-blockers, in order to perform the subsequent division of variable values into two subsets: a lower limit subset and an upper limit subset of each descriptor, with both subsets overlapping to avoid “extreme end effects”. Each of the subsets is divided into bins in order to have discrete values to choose from.
4) Two values are randomly picked for each molecular descriptor, one out of the lower limit subset, the other out of the upper limit subset. The “range” for a descriptor is between the two randomly chosen values. A “filter” is the combination of all randomly picked ranges for all descriptors chosen for examination. The filter is thus composed of descriptors’ ranges that are assumed to favor actives and to disfavor inactives. In other words, active molecules should have properties that fall within the ranges of all descriptors in a filter. The number of descriptor ranges in a filter can be pre-determined but is not necessarily limited.
5) Testing of the actives and inactives: molecules that belong to the known hERG blockers should pass all values of an optimal filter. Successful passage of an active blocker makes them true positives. If a known active fails in passing a single range (i.e., if log P is −2 while the current filter includes the range 0–10 for log P) the molecule is considered to be a non-blocker. That is, it is considered to be a false negative if it originates from the training set of hERG blockers. If a non-blocker fails to pass that filter, it is simply a true negative, while if it passes that filter, it is considered to be a false positive. The Matthews Correlation Coefficient (MCC), equation (1) [59], is used as the scoring function to measure the ability to differentiate between hERG blockers and non-blockers in the training set, based on the current filter.

Matthews Correlation Coefficient (MCC) equation:

\[
\text{MCC} = \frac{(PN) - (P_f N_f)}{\sqrt{(N + N_f) (N + P_f) (P + N_f) (P + P_f)}}
\]

where P, N, P_f and N_f are the percentage of true positive, true negative, false positive, false negative predictions, respectively. Values of MCC range from −1 and 1. A perfect prediction gives a correlation coefficient of 1 while a random performance (resulting in P = P_f and N = N_f) gives a correlation coefficient of 0. A negative value of MCC is worse than random performance and a completely erroneous prediction gives MCC = −1.

The main strength of MCC is in giving proper weights to the 4 possible types of results: true and false positives and negatives, in
calculating the scoring for databases with known positives and negatives. For hERG blockers and non-blockers, it is based on their database origin and on the current filter. Each filter can produce a different MCC, and each filter sets a different rule to predict if a molecule belongs to one database or to the other, while a better filter produces a higher value of MCC.

6) After random selection of many filters, in the order of \( \sim 10^5 \), a virtual histogram is constructed to examine this large sample. The histogram is arranged according to the number of occurrences of each MCC score. In that histogram, only the lowest and highest values of MCC values are examined for decision making. In this examination, each and every value of each and every variable is examined, in order to determine how many times it appears among low MCC values and among high MCC values.

Table 1
Filters composed of the same three descriptors but different ranges.a

<table>
<thead>
<tr>
<th>Filter</th>
<th>MCC</th>
<th>TP</th>
<th>TN</th>
<th>GCUT_SlogP_3</th>
<th>PEOE_PC</th>
<th>ChiDv_C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.782</td>
<td>82</td>
<td>95.5</td>
<td>2.63–3.68</td>
<td>1.92–7.45</td>
<td>9.57–30.12</td>
</tr>
<tr>
<td>2</td>
<td>0.771</td>
<td>84.4</td>
<td>92.5</td>
<td>2.655–3.68</td>
<td>1.92–7.45</td>
<td>8.91–30.12</td>
</tr>
<tr>
<td>3</td>
<td>0.766</td>
<td>86.4</td>
<td>90.2</td>
<td>2.63–3.68</td>
<td>1.29–7.45</td>
<td>9.57–30.12</td>
</tr>
<tr>
<td>4</td>
<td>0.765</td>
<td>85</td>
<td>91.4</td>
<td>2.59–3.68</td>
<td>1.92–7.45</td>
<td>0–30.12</td>
</tr>
<tr>
<td>5</td>
<td>0.763</td>
<td>84</td>
<td>92.1</td>
<td>2.63–3.68</td>
<td>1.66–7.45</td>
<td>8.91–30.12</td>
</tr>
<tr>
<td>6</td>
<td>0.762</td>
<td>83.7</td>
<td>92.2</td>
<td>2.55–3.68</td>
<td>1.76–7.45</td>
<td>9.57–30.12</td>
</tr>
</tbody>
</table>

a A chemical is considered active (positive) if it obeys all rules in the filter. Clustering criterion required that, for retaining a filter, at least 3% of total molecules change sign from hERG blockers to non-blockers or vice versa upon examination of each filter.

Fig. 1. Flowchart of hERG toxicity indexing by Iterative Stochastic Elimination.
values. These numbers are compared to the expected average, which is the total number of samples divided by the total (remaining, if a subsequent iteration, see below) number of values of a variable. The expected number of occurrences of each value in the low (worst) and high (best) regions of MCC (each containing about 1% of the total samples, \( \sim 10^3 \) filters in each) is thus compared to its actual number of occurrences.

7) The appearance of each value is counted in the best 1% of sample and in the worst 1% of the sample. Values (but not the variables) will be eliminated if they appear consistently more in the worst filters (those with lowest MCC values) than in the best filters (those with highest MCC values). Rules for the exact criteria may vary, depending on the system examined. In many cases, values may be eliminated if they appear certain times (e.g. twice or more) in the worst than in the best results. In other cases, standard deviations from the mean may be employed to make such decisions.

8) After the elimination of values from the set of values of a variable, due to the persistent association with worst results, the MCC values in subsequent samplings increase while the number of possible combinations of variable values is considerably reduced. This is borne out in subsequent iterations of sampling and repetitions of steps 4–7 above. Each iteration ends with a better top filter (by MCC value) and with a higher MCC average than the previous iteration. Eliminations in each iteration continue until a predetermined number of combinations, such as a set of some \( 10^6–7 \) combinations, is reached, from the initial \( 10^{10–30} \) for sets of 4–9 descriptors in the case of hERG. These large numbers are produced because if a variable range of values is divided into a total of 200 values, 100 of them in the low range and 100 of them in the high range, then 10,000 combinations of ranges for that variable alone are possible. 4 variables can thus reach \( 10^{16} \) options, and 9 variables could reach \( 10^{36} \).

9) The process is switched from stochastic to an exhaustive calculation of all remaining combinations, and their MCC values are sorted. The optimal k “best filters” are ranked according to their MCC score and clustered. Differences in MCC values between filters exist but are not large, while the filters do differ, following the clustering process, in the descriptors and/or in the values of descriptor ranges.

ISE thus forms a large set of “good filters” that are alternatives to the optimal solution, each of them being somewhat less successful than the optimal.

Finally, the large set of filters is now employed to increase the efficiency of discrimination by employing a “combined filters approach”, which results in constructing the ETI (hERG Toxicity Index). It is based on the assumption that “hERG blockers” would pass more of the “filters”, while “non-blockers” would pass a smaller number of filters. This assumption is the basis for constructing the hERG Toxicity Index (ETI), which is composed of the weighted contribution of the number of filters passed by a molecule and filters that the molecule failed to pass (equation (2))

\[
\text{ETI} = \frac{\sum_{i=1}^{n} P_{ni} - \delta_{ni} N_{ni}}{n}
\]

In equation (2), the hERG Toxicity Index (ETI) for a molecule is determined on the basis of a set of \( n \) filters (\( i = 1, n \)) and is constructed from all 4 numbers in two pairs – positives \( P_i \) (\%P, see Table 1 for each filter) and false positives \( P_{ni} \) (which is \( 100 - \%N \)), as well as false negatives \( N_{ni} \) (100 – \%P) and true negatives \( N_i \) (\%N).

### Table 2

<table>
<thead>
<tr>
<th>Filter</th>
<th>MCC</th>
<th>TP%</th>
<th>TN%</th>
<th>Descriptor-1 (ranges)</th>
<th>Descriptor-2 (ranges)</th>
<th>Descriptor-3 (ranges)</th>
<th>Descriptor-4 (ranges)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.782</td>
<td>82</td>
<td>95.5</td>
<td>GCUT_SlogP_3: 2.63–3.68</td>
<td>PEDE_PC: 1.92–7.45</td>
<td>Chidv_C: 9.57–30.12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.781</td>
<td>81</td>
<td>96.2</td>
<td>a_hyd: 15–37</td>
<td>GCUT_SlogP_3: 2.63–3.67</td>
<td>PEDE_PC: 1.92–7.45</td>
<td>SlogP_VSA5: 0–194</td>
</tr>
<tr>
<td>3</td>
<td>0.78</td>
<td>82</td>
<td>95.2</td>
<td>apol: 44.6–137.8</td>
<td>a_hyd: 15–37</td>
<td>PEDE_PC: 1.92–7.45</td>
<td>Chidv_C: 9.48–30.12</td>
</tr>
<tr>
<td>4</td>
<td>0.779</td>
<td>85.7</td>
<td>92</td>
<td>apol: 44.6–137.8</td>
<td>GCUT_SlogP_3: 2.63–3.68</td>
<td>Q_RPC: 0–0.35</td>
<td>a_hyd: 15–37</td>
</tr>
<tr>
<td>5</td>
<td>0.777</td>
<td>84</td>
<td>93.4</td>
<td>GCUT_PEDE_3: 2.59–3.689</td>
<td>GCUT_SlogP_3: 2.63–3.68</td>
<td>a_hyd: 15–37</td>
<td>Vdw_area: 378.5–1143.7</td>
</tr>
</tbody>
</table>

Detailed description of descriptors meaning and method for calculating its value could be found in the link http://www.chemcomp.com/journal/descr.htm. 

a_hyd: number of hydrophobic atoms.

VdW_area: area of van der Waals surface calculated using a connection table approximation.

GCUT_SlogP_3: log P GCUT (3/3).

GCUT_PEDE_3: PEDE partial charge GCUT (3/3).

PEDE_PC: total positive partial charge.

Chidv_C: carbon valence connectivity index (order 0).

Q_RPC: relative positive partial charge.

SlogP_VSA5: sum of \( v_i \) such that \( L_i \) is in [0.15,0.20].

A chemical is considered active (positive) if it obeys all rules in the filter. Clustering criterion required that, for retaining a filter, at least 3% of total molecules change sign from hERG blockers to non-blockers or vice versa upon examination of each filter.
The number \( n \) could range from dozens to thousands. The value of the delta function \( \delta_{NH} \) is 0 (Zero) if the molecule is a non-blocker according to the currently calculated filter \( i \), and 1 if it is a blocker according to that filter. Similarly, the value of the delta function \( \delta_{NH} \) is 1 if it is a non-blocker according to the currently calculated filter, and 0 if it is a blocker according to that filter. \( P_i \) is the percentage of hERG blockers that are predicted to be blockers according to filter \( i \) (“True positives”), while \( P_h \) is the percentage of false positives, i.e., non-blockers that are predicted to be blockers according to filter \( i \). \( N_h \) is the percent of blockers identified to be non-blockers according to the current filter (“False negatives”), and \( N_{Nh} \) is the percent of non-blockers identified as such by the current filter, i.e., “True negatives”. The quotient \( P_h/N_{Nh} \) may be regarded as an “efficiency factor” of filter \( i \) for the hERG blockers, while the quotient \( N_{Nh}/N_h \) is an “efficiency factor” for misidentifying non-blockers.

To validate the models’ predictive performance, we subjected the systems to a series of leave-1/3-out runs. The leave-1/3-out tests were designed as follows. One hundred blockers and three thousands non-blockers from the original data set were randomly picked to represent the external test set, while the rest of the data constituted the training set for this particular data partition. The model was built based on the training set and applied to predict class attributes of the compounds in the test set. Reported MCC is for performance of the model on the test set.

The ETI model was run on external databases of chemicals and drugs: 1) Comprehensive Medicinal Chemistry database (CMC) composed of 4289 drugs that obey Lipinski RO5, 2) clinical-MDDR composed of 4637 compounds in clinical trials that obey Lipinski RO5, 3) nine thousands molecules that were selected randomly from the Available Chemicals Directory (ACD) database and obey Lipinski RO5, 4) eight thousands nine hundred and twenty seven molecules that obey the Lipinski RO5 were chosen randomly from ZINC database and 46 endogenous compounds. Averaged ETI is calculated and reported.

3. Results and discussion

ISE was utilized to establish an in silico prediction system for detecting chemicals that might be active hERG blockers. This study was based on a set of 300 compounds with a known hERG blocking activity. The measured IC₅₀ of these compounds is less than 10 \( \mu \)M (“blockers” class), while 9000 compounds were selected randomly from ZINC as the “non-blockers” class. It is worth to note that few out of the 9000 compounds could be hERG blockers but the effect of that assumption on the quality of the prediction model is negligible. In order to guarantee that our active class would not be biased by similar structures, we checked the diversity between all hERG blockers and found that they are very highly diverse (see Fig. 2). About 250 compounds out of the 300 hERG blockers have a Tanimoto Similarity index <0.7 and more than 160 molecules have Tanimoto index <0.3. Also, the non-blockers from the ZINC database are even more diverse than the set of blockers and 100% of the molecules have a Tanimoto index of similarity <0.7.

It is interesting to note that in the current study we have found that 89.5% of the 300 hERG blockers obey Lipinski’s rule of 5 (ROF). As Lipinski’s ROF is frequently considered to be a measure of drug-likeness, it must imply to many that ROF filters toxic molecules. Our result should help to avoid such confusion, most approved drugs are orally bioavailable [60–63] but could still be hERG blockers. This emphasizes the utmost need for producing and applying a filters model for the potential of molecules to block hERG, combined with the oral bioavailability filter, early in the drug discovery process in order to reduce attrition at much later stages of research.

Our models thus consist of sets of filters that may be subsequently applied to molecules of the test set and for examining any other molecular candidates for drug activities. Table 1 lists the performance of several filters with different ranges for the same set of 3 descriptors (while Table 2 lists some filters composed of different sets of descriptors). It turned out that there are many unique and simple rules-based models that have a high predictive ability for discriminating between hERG blockers/non-blockers. Among the huge number of proposed filters, we choose a simple one to help medicinal chemists to select hERG safe compounds. The proposed filter is composed of 3 discriminative descriptors (total charge, number of aromatic atoms and log \( S \) — solubility in water, see histograms in Fig. 3). According to the 3-rules based filter, hERG-Toxic molecules are more likely to have total charge \( \leq 0 \) and \( \leq 3 \), number of aromatic atoms \( \geq 6 \) and \( \leq 28 \) and log \( S \) \( \leq 0 \). Table 4 summarizes the efficiency of such simple rules-based filter in evaluating the hERG liability of chemicals from different databases.

Fig. 3. Histograms of 3 discriminative molecular properties of the hERG blockers: a) total charge, b) number of aromatic atoms and c) calculated log \( S \).
Structures for few chemicals picked randomly from the hERG binders' database are shown on Fig. 4 and reveal these preferences. This simple rule could be utilized by medicinal chemists for the design of cardio-safer drug candidates by simply counting chemical features of their chemicals.

This top single filter has a Matthews Correlation Coefficient of 0.78 (Table 2) and nearly 82% of the hERG blockers (true positives) were successfully identified by it, while only 4.5% of the ZINC molecules (assumed not to be hERG blockers) were misclassified (thus, are false positives).

Fig. 5 depicts the enrichment plot and the receiver operating characteristic (ROC) plot of our model. The prediction model effectiveness is evaluated in terms of the enrichment factor (EF) as a function of the percentage $x$ of the completely ranked database that is selected for screening:

$$EF(x) = \frac{f_{\text{blocker}}(x)}{x}$$

Where $f_{\text{blocker}}$ is the percentage of the true blockers found at the top $x\%$ of the ranked database.

The enrichment plot (Fig. 5) illustrates how the hERG blockers could be identified if compounds were sorted according to the model prediction rather than on the basis of random selection. An enrichment plot close to the perfect model means a high prioritization power of the proposed model. In our ISE-based model, two-thirds of the hERG blockers could be captured in the top 3.1% of the evaluated compounds compared with finding them in 2.2% of the total compounds in the perfect model and in only 66.7% in the random model. More than 80% of the blockers are found in less than 5% of the total. However in searching for hERG blockers it is required to discover ALL molecules that have a substantial potential for that dangerous activity. Therefore we need to set criteria for large screenings for discovering all the toxic molecules.

If we pick molecules with ETI above 12.5, the ISE-based model and the perfect model totally overlap. Thus, it seems that the proposed model is highly discriminative and exhibits a very good performance for the classification of hERG blockers/non-blockers. The attained area under the curve (AUC) is 0.94 which indicates that the model is very good.

Among known drugs, there are some 62 molecules that were found to be active on hERG. In Table 3 we present their ETI values as calculated by the ISE model. Fifty seven molecules out of the sixty two drugs have positive ETI index.

The ETI model was applied to external databases of chemicals and drugs. Table 5 summarizes the averaged ETI. We found that the average ETI of hERG binders (9.17) > clinical-MDDR (3.3) > drugs from CMC (2.53) > endogenous molecules (−1.98) > ACD (−2.49) > ZINC (−3.86). Forty six known endogenous molecules were indexed by the ETI model and found to be less liable for hERG than marketed drugs. Fig. 6 shows 11 of those molecules and their ETI while Fig. 7 shows six potential hERG toxic endogenous molecules based on the proposed model.

![Fig. 4. hERG binders.](image-url)
of molecules to be rejected, even if one of the more stable models with the aim of minimizing the risks, thus maximizing the number described in our introduction, not in order to focus on overlaps but the predicted hERG liability index of 62 hERG blocking drugs.

Table 3
The predicted hERG liability index of 62 hERG blocking drugs.

<table>
<thead>
<tr>
<th>Drug name</th>
<th>ETI</th>
<th>Drug name</th>
<th>ETI</th>
<th>Drug name</th>
<th>ETI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfuzosin</td>
<td>1.99</td>
<td>Flecainide</td>
<td>3.36</td>
<td>Quetiapine</td>
<td>3.36</td>
</tr>
<tr>
<td>Amiodarone</td>
<td>3.17</td>
<td>(S)-Fluoxetine</td>
<td>1.51</td>
<td>Quinidine</td>
<td>3.36</td>
</tr>
<tr>
<td>Amitryptiline</td>
<td>2.99</td>
<td>Gatifloxacin</td>
<td>6.94</td>
<td>Risperidone</td>
<td>3.36</td>
</tr>
<tr>
<td>Bepridil</td>
<td>3.36</td>
<td>Granisetron</td>
<td>3.36</td>
<td>Roxithromycin</td>
<td>3.17</td>
</tr>
<tr>
<td>Chloroquine</td>
<td>3.36</td>
<td>(S)-Halofantrine</td>
<td>3.82</td>
<td>Sertindole</td>
<td>3.36</td>
</tr>
<tr>
<td>Chlorpromazine</td>
<td>1.51</td>
<td>Haloperidol</td>
<td>3.36</td>
<td>Sertaline</td>
<td>-0.06</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>1.33</td>
<td>Ibutilide</td>
<td>3.36</td>
<td>Sibutramine</td>
<td>3.13</td>
</tr>
<tr>
<td>Cisapride</td>
<td>3.82</td>
<td>Imipramine</td>
<td>3.36</td>
<td>Telithromycin</td>
<td>3.17</td>
</tr>
<tr>
<td>(S)-Citalopram</td>
<td>3.36</td>
<td>Isradipine</td>
<td>3.08</td>
<td>Thioridazine</td>
<td>3.36</td>
</tr>
<tr>
<td>Clarithromycin</td>
<td>3.17</td>
<td>Ketoconazole</td>
<td>2.87</td>
<td>Tolterodine</td>
<td>3.36</td>
</tr>
<tr>
<td>Clofibrate</td>
<td>3.36</td>
<td>Levofloxacin</td>
<td>6.94</td>
<td>Trimethoprim</td>
<td>-4.97</td>
</tr>
<tr>
<td>Clozapine</td>
<td>2.46</td>
<td>Levomethadyl</td>
<td>3.36</td>
<td>Ziprasidone</td>
<td>3.36</td>
</tr>
<tr>
<td>Cocaine</td>
<td>0.05</td>
<td>Mefloquine</td>
<td>2.46</td>
<td>Diphenhydramine</td>
<td>0.29</td>
</tr>
<tr>
<td>Desipramine</td>
<td>2.46</td>
<td>Mesoridazine</td>
<td>3.36</td>
<td>Verapamil</td>
<td>3.17</td>
</tr>
<tr>
<td>Disopyramide</td>
<td>3.36</td>
<td>Methadone</td>
<td>6.94</td>
<td>Terfenadine</td>
<td>3.17</td>
</tr>
<tr>
<td>Doxepin</td>
<td>2.46</td>
<td>Mexiletine</td>
<td>-5.13</td>
<td>Amlodipine</td>
<td>3.17</td>
</tr>
<tr>
<td>Dolasetron</td>
<td>2.99</td>
<td>Olanzapine</td>
<td>2.46</td>
<td>Amlodipine</td>
<td>3.36</td>
</tr>
<tr>
<td>Domperidone</td>
<td>3.36</td>
<td>Ondansetron</td>
<td>-3.07</td>
<td>Loratadine</td>
<td>1.4</td>
</tr>
<tr>
<td>Doxepin</td>
<td>2.46</td>
<td>Pimozide</td>
<td>3.17</td>
<td>Mibolerone</td>
<td>3.36</td>
</tr>
<tr>
<td>Droperidol</td>
<td>3.36</td>
<td>Procarbazine</td>
<td>-3.84</td>
<td>Vesnarinone</td>
<td>3.36</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>3.17</td>
<td>Protriptyline</td>
<td>1.77</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Our ETI model may be used in conjunction with other models described in our introduction, not in order to focus on overlaps but the aim of minimizing the risks, thus maximizing the number of molecules to be rejected, even if one of the more stable models predicts a potential for toxicity. This suggestion may result in many false negatives, but will minimize the false positives. Our suggestion is also based on our belief and our experience, that there are many molecules out there in the accessible chemical space that could serve as alternatives for any rejected molecular entity.

4. External test set

We searched the chEMBL database (June 2012) and found that it contains 1405 chemicals that exhibit blocking of hERG potassium channel with IC_{50} up to 10 μM. Out of the 1405 hERG blockers, 98 molecules were used for our previous model and thus were discarded. Fig. 8 shows similarity ranges compared to previous set composed of 300 blockers of hERG. Most of the new blockers are different from the previous ones used for the model and have a tanimoto index less than 0.3. Their 2D descriptors were calculated by MOE2009.10 and indexed for their ETI. The performance of the ETI model on the chEMBL database gave better results than for the previous set (see Fig. 9). Assigning ETI = +1.75 as a threshold for discriminating active blockers from non-actives we got MCC = 0.89, specificity = 0.94, sensitivity = 0.95. The sensitivity could reach 0.997 by selecting as a threshold ETI = −5 while specificity could reach 0.996 by selecting as a threshold ETI = +12.75. These results reveal that the proposed model is robust and highly efficient for screening external databases of chemicals and/or drug candidates.

5. ISE versus other approaches

The WEKA (Waikato Environment for Knowledge Analysis) package includes many machine learning techniques for classification purposes (http://www.cs.waikato.ac.nz/ml/weka). Four methods have been used to model hERG liability. The employed algorithms are the state-of-the-art learning techniques for classification: neural networks, support vector machines, nearest-neighbor method and decision trees. We applied the default parameters for all tested algorithms. Table 6 compares the performance of Iterative Stochastic Elimination approach to the other four approaches for constructing a hERG liability prediction model using active blockers/non-actives data sets coded by 184 MOE 2D descriptors (The same data sets utilized before for training and testing prediction models reported for ISE). The efficacy of each prediction

Table 4
Percentage of chemicals that obey the 3-rules based filter for hERG liability.a

<table>
<thead>
<tr>
<th>Database</th>
<th>No. molecules</th>
<th>Chemicals% that obey the filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>hERG binders</td>
<td>300</td>
<td>91.3%</td>
</tr>
<tr>
<td>CMC</td>
<td>4289</td>
<td>35.5%</td>
</tr>
<tr>
<td>Clinical-MDDR</td>
<td>4637</td>
<td>33.1%</td>
</tr>
<tr>
<td>ACD</td>
<td>9000</td>
<td>43.4%</td>
</tr>
<tr>
<td>ZINC</td>
<td>8927</td>
<td>34.4%</td>
</tr>
</tbody>
</table>

*a hERG-Toxic molecules are more likely to have total charge between 0 and 3, number of aromatic atoms between 6 and 28, log S less/equal to 0. 

b All selected chemicals pass Lipinski ROS for oral-bioavailability.
The model is measured by its MCC, sensitivity (SE) and specificity (SP) values. By the proposed indexing technique presented here we can get high specificity and/or sensitivity values while moving along the axis of hERG liability index. The comparison between performance of WEKA algorithms and Iterative Stochastic Elimination algorithm is not trivial since the former outputs 1/0 predictions while the latter outputs continuous index. In the proposed ETI model the index values range between $-5.5$ and $13.3$. In order to perform the comparison we have to determine certain ETI value as a threshold and to transform the ETI model output to 1/0 predictions. Chemicals with ETI above the threshold will be predicted as positives (1) while chemicals having ETI less than the threshold will be predicted as negatives (0). Models output 1/0 predictions could be considered as special case of models output continuous predictions. One other big advantage of the ETI model is that the filters are composed of sets of descriptors’ ranges and could be easily understood and utilized by medicinal chemists for the design of safer drug candidates.

Fig. 6. Endogenous chemicals indexed for their potential hERG toxicity.

Fig. 7. Potential hERG toxic endogenous molecules based on the proposed model.

Fig. 8. Similarity ranges of chemicals from the external test set (extracted from chEMBL database) compared to previous set composed of 300 blockers of hERG.
The performance of the ETI prediction model, external test set versus internal test set.

**Table 6**

<table>
<thead>
<tr>
<th>Method</th>
<th>MCC</th>
<th>SE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.865</td>
<td>0.967</td>
<td>0.898</td>
</tr>
<tr>
<td>KNN method (Kstar)</td>
<td>0.86</td>
<td>0.921</td>
<td>0.938</td>
</tr>
<tr>
<td>Artificial neural network (multi layer perceptron)</td>
<td>0.861</td>
<td>0.947</td>
<td>0.913</td>
</tr>
<tr>
<td>Support vector classifier (SMO)</td>
<td>0.881</td>
<td>0.958</td>
<td>0.923</td>
</tr>
<tr>
<td>Decision tree (J48)</td>
<td>0.82</td>
<td>0.901</td>
<td>0.918</td>
</tr>
<tr>
<td>ETI (SE)</td>
<td>0.89*</td>
<td>0.997*</td>
<td>0.996*</td>
</tr>
</tbody>
</table>

*ETI threshold of 0.750.
*ETI threshold of 0.50.
*ETI threshold of 0.1275.

Equations for sensitivity (SE) and specificity (SP):

\[
SE = \frac{TP}{TP + FN}
\]
\[
SP = \frac{TN}{TN + FP}
\]

where: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

It is worth mentioning that the speed of screening by this novel indexing technique is very fast. The computation time for 1000 compounds takes a few seconds with the 2D-descriptors input of the compounds.

6. Conclusions

By applying the ISE algorithm to the scoring of the molecular potential for hERG toxicity we provide a tool that may be employed for prior filtering out of drug candidates with a potential for toxicity. The assessment requires transformation of molecular structures into properties, and thus the molecular 2D descriptors employed in this study are easily calculated and have shown a clear relevance to hERG binding affinity. A highly discriminative and robust model has been constructed. Such an approach proposed here could become an important utility for predicting and differentiating between hERG blockers and non-blockers. Utilizing such technique in the hit and/or lead optimization stages of the drug discovery process will undoubtedly offer valuable information concerning the rational design of safer drug candidates that should be exempted of undesirable activities on hERG. The prediction of hERG Toxicity Index based on the proposed model is fast since the employed 2D descriptors are easily and quickly calculated from the 2D structure of a molecule.

Acknowledgments

We gratefully acknowledge RAND Biotechnologies LTD for providing us the database of hERG blockers. As well, we would like to thank the reviewers for their valuable comments which helped to considerably improve the quality of the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at


References


B. Darpo, T. Nebout, P.T. Sager, Clinical evaluation of QT/QTc prolongation and


P. Hoffmann, B. Warner, Are hERG channel inhibition and QT interval pro-

L. Crotti, G. Celano, F. Dagradi, P.J. Schwartz, Congenital long QT syndrome,

A.M. Aronov, Common pharmacophores for uncharged human ether-a-go-go-

Q. Li, F.S. Jorgensen, T. Oprea, S. Brunak, O. Taboureau, hERG classi-


A. Rayan, A. Goldblum, Stochastic Method to Determine, In Silico, the Drug-


J.S. Mitcheson, hERG potassium channels and the structural basis of drug-


C. Obsil-Pardo, J. Comis-Tena, F. Sanz, J. Saiz, M. Pastor, A multiscale simu-


Y.J. Kim, H.K. Hong, H.S. Lee, S.H. Moh, J.C. Park, S.H. Jo, H. Choe, Papaverine, a vasodilator, blocks the pore of the HERG channel at submicromolar concentra-

S.H. Jo, H.K. Hong, S.H. Chong, K.S. Jung, H. Choe, Clozapine, a novel hy-

A. Rayan, A. Goldblum, Stochastic Method to Determine, In Silico, the Drug-

B. W. Matthews, Comparison of the predicted and observed secondary struc-


B.W. Matthews, Comparison of the predicted and observed secondary struc-

