JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Jointly Handling Potency and Toxicity of Antimicrobial Peptidomimetics by Simple Rules from Desirability Theory and Chemoinformatics

Maykel Cruz-Monteagudo,[*,†,‡] Fernanda Borges,[†] and M. Natália D. S. Cordeiro[*,§]

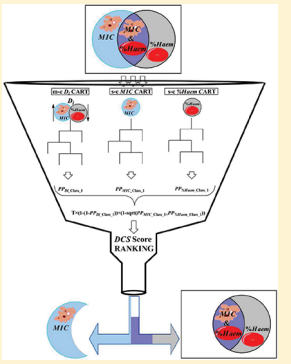
[†]CIQ, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

[‡]Applied Chemistry Research Center - Faculty of Chemistry and Pharmacy, Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830, Cuba

[§]REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

Supporting Information

ABSTRACT: Today, emerging and increasing resistance to antibiotics has become a threat to public health worldwide. Antimicrobial peptides have unique action mechanisms making them an attractive therapeutic prospect to be applied against resistant bacteria. However, the major drawback is related with their high hemolytic activity which cancels out the safety requirements for a human antibiotic. Therefore, additional efforts are needed to develop new antimicrobial peptides that possess a greater potency for bacterial cells and less or no toxicity over erythrocytes. In this paper, we introduce a practical approach to simultaneously deal with these two conflicting properties. The convergence of machine learning techniques and desirability theory allowed us to derive a simple, predictive, and interpretable multicriteria classification rule for simultaneously handling the antibacterial and hemolytic properties of a set of cyclic β-hairpin cationic peptidomimetics (Cβ-HCPs). The multicriteria classification rule exhibited a prediction accuracy of about 80% on training and external validation sets. Results from an additional concordance test have shown an excellent agreement between the multicriteria classification rule predictions and the predictions from independent classifiers for complementary antibacterial and hemolytic activities, respectively, evidencing the reliability of the multicriteria classification rule. The rule was also consistent with the general mode of action of cationic peptides pointing out its biophysical relevance. We also propose a multicriteria virtual screening strategy based on the joint use of the multicriteria classification rule, desirability, similarity, and chemometrics concepts. The ability of such a virtual screening strategy to prioritize selective (nonhemolytic) antibacterial Cβ-HCPs was assessed and challenged for their predictivity regarding the training, validation, and overall data. In doing so, we were able to rank a selective antibacterial Cβ-HCP earlier than a biologically inactive or nonselective antibacterial Cβ-HCP with a probability of ca. 0.9. Our results thus indicate that promising chemoinformatics tools were obtained by considering both the multicriteria classification rule and the virtual screening strategy, which could, for instance, be used to aid the discovery and development of potent and nontoxic antimicrobial peptides.



INTRODUCTION

At the beginning of the new century, the problem of microbial drug resistance has achieved a global dimension and an alarming magnitude, being one of the leading unresolved problems in public health.[1] The most alarming resistance trends are those observed for *Enterobacteriaceae* and the Gram-negative nonfermenters, with a generalized increase in rates of resistance to the most important anti-Gram-negative agents (β-lactams and fluoroquinolones) and the cocirculation of multidrug-resistant strains.[2,3] The relentless evolution of resistance, in face of the decrease in the development of new antimicrobial agents active against resistant pathogens, has led to an increasing number of cases in which the pathogen is resistant to most, or even all, drugs available for clinical use (the so-called pandrug resistance phenotypes).[1] This situation has stimulated the search for new classes of antimicrobial agents from natural sources.[4]

Antimicrobial peptides possess an unique mechanism of action that involves a different and faster killing process, making peptide antibiotics an attractive therapeutic option to be applied against resistant bacteria.[5] This discovery has stimulated the drug development process of new active peptide antibiotics.[4] Antimicrobial peptides are effector molecules that initiate the fight against bacterial infection by perturbing the phospholipid bilayer of the cell membrane, interfering with metabolism, or by targeting cytoplasmic components.[6] Specifically, naturally occurring cationic antimicrobial peptides represent interesting leads for peptidomimetic research. These natural products have been discovered in animal kingdoms where they are now recognized as key mediators of the innate immunity to bacterial infections.[7] Generally, they are peptides of less than 50 amino acids, with a

Received: May 16, 2011
Published: November 25, 2011

net positive charge due to multiple lysine and/or arginine residues and with multiple hydrophobic residues. The *minimum inhibitory concentration* for cationic antimicrobial peptides is typically in the range 0.25−16 μg/mL against Gram-positive and -negative bacteria bacteria, fungi, and protozoa.[8] However, potential cytotoxicity to the host cells remains a major unsolved challenge.[9] Specifically, their high hemolytic activity lacks the selectivity required for a human antibiotic, a major drawback regarding drug safety requirements.[10] Therefore, additional efforts are needed to develop new cationic antimicrobial peptides that possess greater selectivity for bacterial cells over erythrocytes, leading to a potent antimicrobial peptide candidate with less or no toxicity over erythrocytes.

There are different chemoinformatics techniques that can be used to assist peptidomimetic research.[11] The most widely used are quantitative structure−activity relationship (QSAR) techniques, which seek to uncover the correlation of biological properties with molecular structure by numerical analysis. The QSAR techniques have been frequently applied to the pharmaceutical drug research and, presently, to antimicrobial peptides design and discovery.[12] The recent efforts done by Wang et al.[13] reflect the future direction for developing practically useful predictors[14] by implementing an user friendly and publicly accessible web server for predicting novel antimicrobial peptides. However, save for a few exceptions,[15] most of the chemoinformatics efforts have been focused on antibacterial activity, overlooking the equally determinant hemolytic counterpart.[11] Noteworthy, similar QSAR studies considering simultaneously the bioactivity and toxicity of drug candidates toward diverse therapeutic applications have been recently reported.[16,17]

The efforts to combat multidrug-resistant microorganisms during the past decade have been mainly focused on Gram-positive bacteria and the development of novel antimicrobial agents to combat this type of bacteria. In that way, several compounds with novel mechanisms of action, e.g., linezolid and daptomycin, have recently been launched in the market.[18] Unfortunately, no new class of antibiotics have been developed specifically for multidrug-resistant Gram-negative bacilli.[19] The return to the preantibiotic era has become a reality in many parts of the world, making novel Gram-negative antimicrobial agents particularly needed in the fight against multidrug-resistant microorganisms.[20]

The purpose of this paper is to test the efficiency and ability of a multicriteria classification QSAR to identify antimicrobial peptides with selective permeation of Gram-negative bacteria membranes. Toward that goal, our study resorts to experimental data on Gram-negative antibacterial and hemolytic properties of cyclic β-hairpin peptidomimetics based on the cationic antimicrobial peptide protegrin I coming from an extensive structure−activity relationship (SAR) study reported by Robinson et al.[8] Based on the proper application of machine learning tools[21] and desirability theory[22] concepts, a simple, predictive, and biophysically meaningful rule was derived enabling the identification of selective antimicrobial peptidomimetics, i.e., with a proper balance of antibacterial and hemolytic profiles. Additionally, we propose a virtual screening strategy based on the use of a multicriteria classification rule combined with desirability, similarity, and chemometrics concepts. The applicability and the reliability of such a virtual screening strategy to identify potentially selective (non hemolytic) cyclic β-hairpin Gram-negative antibacterial peptidomimetics were assessed and challenged on training, test, and overall data sets.

METHODS

According to a recent comprehensive review,[23] to establish a really useful statistical predictor for a protein or peptide system, one needs to consider the following procedures: (i) construct or select a valid benchmark data set to train and test the predictor; (ii) formulate an effective mathematical expression that can truly reflect the intrinsic correlation between the protein or peptide samples with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (v) establish a user friendly web server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps.

Bioactivity Data. The minimal inhibitory concentration against the Gram-negative bacteria strain *Pseudomona aeruginosa* expressed in μg/mL (MIC) and the percentage of hemolysis of human red blood cells at the concentration of 100 μg/mL (%Hem) of the 136 cyclic β-hairpin cationic peptidomimetics (Cβ-HCPs) considered were taken from Robinson et al.[8] This benchmark data set consists of libraries of analogues of the Cβ-HCP lead used in the work of the latter.[8] The analogues were synthesized by substituting a variety of different amino acids or amino acid analogues at 12 different positions around the β-hairpin. As a consequence, analogues in the data set show a high degree of homology. The full list of sequences and the corresponding identification can be consulted on Tables S1 and S2 of the Supporting Information.

To avoid homology bias and remove the redundant sequences from the benchmark data set, a cutoff threshold of 25% was recommended[24,25] to exclude those proteins from the benchmark data sets that have equal to or greater than 25% sequence identity to any other. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to reach statistical significance.

Desirability Theory. The desirability function approach is a well-known multicriteria decision-making method originally proposed by Harrington[26] and later modified by Derringer and Suich.[22] This approach has been extensively employed in several fields.[27−31] However, despite perfectly fitting with the drug development problem, a trend for its regular application to computational medicinal chemistry problems has emerged just recently.[16,32−37]

This multicriteria approach, is based on the definition of a desirability function for each end point property Y_i in order to transform their values to the same scale. Each property is independently transformed into a desirability value d_i by an arbitrary function. Let L_i, U_i, and T_i be the lower, upper, and target values, respectively, that are desired for the property Y_i, with $L_i \le T_i \le U_i$. Depending on whether a particular property is to be maximized, minimized, or assigned a target value, different desirability functions can be used.

First, if a property is of the target best kind, then its individual desirability function is defined as

$$d_i = \begin{cases} \left[\dfrac{Y_i - L_i}{T_i - L_i}\right]^s & \text{if } L_i \le Y_i \le T_i \\ \left[\dfrac{Y_i - U_i}{T_i - U_i}\right]^t & \text{if } T_i < Y_i \le U_i \\ 0 & \text{if } \hat{Y}_i < L_i \text{ or } Y_i > U_i \end{cases} \quad (1)$$

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

If a property is to be maximized instead, its individual desirability function is defined as

$$d_i = \begin{cases} 0 & \text{if } Y_i \le L_i \\ \left[\dfrac{Y_i - L_i}{T_i - L_i}\right]^s & \text{if } L_i < Y_i < T_i \\ 1 & \text{if } Y_i \ge T_i = U_i \end{cases} \quad (2)$$

In this case, T_i is interpreted as a large enough value for the property, which can be U_i.

Finally, if one wants to minimize a property, one might use

$$d_i = \begin{cases} 1 & \text{if } Y_i \le T_i = L_i \\ \left[\dfrac{Y_i - U_i}{T_i - U_i}\right]^s & \text{if } U_i < Y_i < T_i \\ 0 & \text{if } Y_i \ge U_i \end{cases} \quad (3)$$

Here, T_i denotes a small enough value for the property, which can be L_i. Moreover, the exponents s and t determine how important is to hit the target value T_i. For $s = t = 1$, the desirability function increases linearly toward T_i. Large values for s and t should be selected if it is very desirable that the value of Y_i be close to T_i or increase rapidly above L_i. On the other hand, small values of s and t should be chosen if almost any value of Y_i above L_i, and below U_i is acceptable or if having values of Y_i considerably above L_i are not of critical importance.[22]

Once the kind of function for each property is defined, the overall desirability D_i of each candidate can be evaluated as the geometric mean of the individual desirability values d_i, as expressed in eq 4. This single value of D_i gives the overall assessment of the desirability of the combined property levels. The range of D_i will fall in the interval [0, 1] and will increase as the balance of the properties becomes more favorable, being 0 if at least one of the properties takes a value of $d_i = 0$.

$$D_i = (d_1 \times d_2 \times \ldots \times d_n)^{1/n} \quad (4)$$

Desirability-Based Class Assignment. The desirability theory[26] concepts described above were applied for a binary class assignment (X_Class_1/0) in order to setup one class including selective antibacterial Cβ-HCPs D_i_Class_1: those peptidomimetics with a proper balance between antibacterial and hemolytic properties, and a second class represented by inactive or nonselective antibacterial Cβ-HCPs D_i_Class_0: those peptidomimetics with no antibacterial potency nor a favorable hemolytic profile or those exhibiting good antibacterial potency but a negative hemolytic profile or vice versa. A graphic representation of this class assignment is shown in Figure 1. Since a selective antimicrobial peptide should exhibit a high antibacterial potency while inducing minimal or no hemolysis (as low as possible values of both MIC and %Hem), a minimization Derringer desirability function (eq 3) was applied to each property Y_i in such a way that a Cβ-HCP with the lowest/highest values of the corresponding Y_i is defined as the most desirable/undesirable ($d_i = 1/d_i = 0$).

Lower L_i and upper U_i bounds for both properties (MIC and %Hem) were selected by considering previous information on the antibacterial and hemolytic profiles of the presently tackled β-hairpin cationic antimicrobial peptides. So, taking into account that cationic antimicrobial peptides typically show minimum inhibitory concentration in the range of 0.25−16 μg/mL against Gram-positive and -negative bacteria, fungi, and protozoa,[8] the upper bound (U_i) for MIC was set to 16 μg/mL. The lower bound (L_i) was set to 3.1 μg/mL, the MIC value exhibited by the most potent peptidomimetic in the data set. Like this, one ensures including in the class of selective peptidomimetics (D_i_Class_1), only those Cβ-HCPs in the range of minimum inhibitory concentration of the present cationic antimicrobial peptides. On the other hand, the U_i for %Hem was set to 37% at 100 μg/mL. This value of %Hem coincides with the percentage of lysis of human red blood cells induced by 100 μg/mL of protegrin-1.[8] The L_i was set to 0.1% at 100 μg/mL, the %Hem induced by the less hemolytic Cβ-HCP in the data set. In this way, one ensures that only Cβ-HCPs less hemolytic than protegrin-1 will conform to the D_i_Class_1.

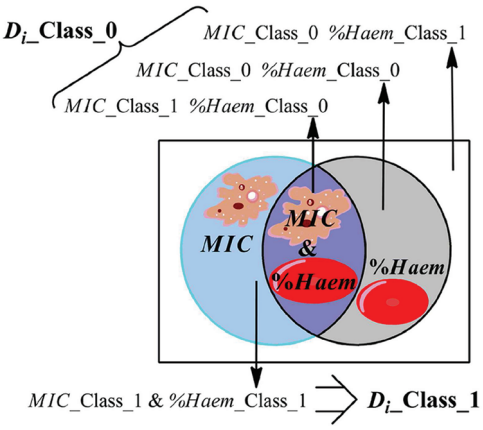


Figure 1. Venńs diagram representing the binary class assignment applied to the set of Cβ-HCPs. D_i_Class_1 compounds are selective antibacterial Cβ-HCPs including those with a proper balance between antibacterial and hemolytic properties (MIC_Class_1 and %Hem_Class_1). D_i_Class_0 are inactive or nonselective antibacterial Cβ-HCPs including those with no antibacterial potency or a favorable hemolytic profile (MIC_Class_0 and %Hem_Class_0) or those exhibiting good antibacterial potency but a negative hemolytic profile (MIC_Class_1 and %Hem_Class_0) or vice versa (MIC_Class_0 and %Hem_Class_1).

Once the D_i values for the full set of Cβ-HCPs were estimated, a threshold value of $D_i = 0.6$ was used to assign each peptidomimetic to the corresponding class. If $D_i \ge 0.6$, then peptidomimetic i was set to the D_i_Class_1, otherwise it was set to the D_i_Class_0. So, Cβ-HCPs with a selectivity profile, that is to say, the balance between antibacterial potency and hemolysis encoded by D_i, greater or equal than 60% will conform to the class of selective peptidomimetics (D_i_Class_1). Although a different threshold could be used, 0.6 represents an adequate choice for both biophysical sense, compounds with a 60% of desirability or higher are considered adequate for the end point, and statistical significance, proper distribution of cases among classes for classification analysis.[38] Thus, 32 Cβ-HCPs were included on the D_i_Class_1, while the remaining 104 were set to the D_i_Class_0. The D_i-based class assignment, the MIC and % Hem values and their respective desirability values (d_i) as well as the corresponding overall desirability (D_i) values for the full set of Cβ-HCPs are depicted in Table 1.

Peptide Structure Codification. Simple (0−2D) molecular descriptors implemented in DRAGON 6.0[39] were selected to codify the molecular structure of the 136 peptidomimetics. Specifically, 3737 DRAGON molecular descriptors were computed. After eliminating constant and near constant variables as

Table 1. Identification, Antibacterial and Hemolytic Properties, Individual and Overall Desirabilities, X_Class_1/0 Assignment, Training/Validation Splitting, and Subset of Molecular Descriptors Used for the CART Analysis of the Full Set of Cβ-HCPs

	properties		desirabilities			X_Class_1/0				molecular descriptors					
Cβ-HCP ID	MIC	%Hem	d_{MIC}	$d_{\%Hem}$	$D_{MIC-\%Hem}$	D_i	MIC	%Hem	T/V	RBF	Alog P	P_VSA_MR_2	D/Dtr10	SpMAD_B(s)	SpMAD_EA(bo)
000_Cβ-HCP lead	6.2	1.4	0.760	0.965	0.856	1	1	1	T	0.143	−0.072	901.875	0.000	1.942	1.555
001_Leu1xPhe	9.4	1.1	0.512	0.973	0.706	1	1	1	T	0.142	0.252	901.875	0.000	1.944	1.589
002_Leu1xCha	25	7.1	0.000	0.810	0.000	0	0	1	T	0.139	0.925	901.875	0.000	1.933	1.563
003_Leu1xTyr	25	0.6	0.000	0.986	0.000	0	0	1	T	0.141	−0.016	944.558	0.000	1.958	1.595
004_Leu1xTrp	50	2.2	0.000	0.943	0.000	0	0	1	T	0.139	0.545	901.875	0.000	1.939	1.615
005_Leu1xHfe	12.5	6.8	0.271	0.818	0.471	0	0	1	T	0.144	0.708	901.875	0.000	1.940	1.594
006_Leu1 × 1-Nal	12.5	13.2	0.271	0.645	0.418	0	0	1	V	0.139	1.160	901.875	576.424	1.939	1.615
007_Leu1 × 2-Nal	12.5	7.7	0.271	0.794	0.464	0	0	1	T	0.139	1.160	901.875	591.913	1.938	1.615
008_Leu1xCIF	3.1	11.5	1.000	0.691	0.831	1	1	1	T	0.142	0.916	901.875	0.000	1.945	1.595
009_Leu1xHis	100	0.2	0.000	0.997	0.000	0	0	1	T	0.143	−1.719	878.333	0.000	1.946	1.586
010_Leu1xBip	6.2	17.6	0.760	0.526	0.632	1	1	1	V	0.140	1.770	901.875	0.000	1.933	1.616
011_Leu1xOrn	100	0.7	0.000	0.984	0.000	0	0	1	T	0.147	−1.840	925.417	0.000	1.950	1.561
012_Arg2xTyr	12.5	1.4	0.271	0.965	0.512	0	0	1	T	0.133	1.722	921.016	0.000	1.949	1.603
013_Arg2xTrp	18	0.2	0.000	0.997	0.000	0	0	1	T	0.131	2.283	878.333	0.000	1.928	1.623
014_Arg2xLeu	12.5	1.6	0.271	0.959	0.510	0	0	1	T	0.134	1.665	878.333	0.000	1.932	1.574
015_Arg2xHis	12.5	1.8	0.271	0.954	0.509	0	0	1	T	0.135	0.019	854.790	0.000	1.936	1.593
016_Arg2xThr	12.5	1.3	0.271	0.967	0.512	0	0	1	T	0.133	−0.077	890.104	0.000	1.957	1.576
017_Arg2xGln	12.5	1.1	0.271	0.973	0.514	0	0	1	T	0.139	−0.492	901.868	0.000	1.965	1.572
018_Leu3xPhe	25	2.6	0.000	0.932	0.000	0	0	1	T	0.142	0.252	901.875	0.000	1.944	1.589
019_Leu3xCha	12.5	3.1	0.271	0.919	0.499	0	0	1	V	0.139	0.925	901.875	0.000	1.933	1.563
020_Leu3xTrp	12.5	4.5	0.271	0.881	0.489	0	0	1	T	0.139	0.545	901.875	0.000	1.938	1.615
021_Leu3xHfe	6.2	3.8	0.760	0.900	0.827	1	1	1	V	0.144	0.708	901.875	0.000	1.940	1.594
022_Leu3 × 1-Nal	12.5	5.2	0.271	0.862	0.484	0	0	1	T	0.139	1.160	901.875	552.378	1.939	1.615
023_Leu3 × 2-Nal	25	5.8	0.000	0.846	0.000	0	0	1	T	0.139	1.160	901.875	567.411	1.938	1.615
024_Leu3xVal	6.2	1.1	0.760	0.973	0.860	1	1	1	T	0.141	−0.461	890.104	0.000	1.947	1.569
025_Leu3xCIF	50	13.1	0.000	0.648	0.000	0	0	1	T	0.142	0.916	901.875	0.000	1.945	1.595
026_Leu3xHis	100	1.2	0.000	0.970	0.000	0	0	1	T	0.143	−1.719	878.333	0.000	1.946	1.586
027_Leu3xBip	12.5	7.7	0.271	0.794	0.464	0	0	1	T	0.140	1.770	901.875	0.000	1.933	1.616
028_Leu3xOrn	100	1.5	0.000	0.962	0.000	0	0	1	V	0.147	−1.840	925.417	0.000	1.950	1.561
029_Lys4xArg	6.2	1.2	0.760	0.970	0.859	1	1	1	V	0.145	−0.500	901.875	0.000	1.948	1.567
030_Lys4xTrp	100	4	0.000	0.894	0.000	0	0	1	T	0.134	1.856	878.333	0.000	1.935	1.622
031_Lys4xLeu	25	0.5	0.000	0.989	0.000	0	0	1	T	0.137	1.238	878.333	0.000	1.939	1.573
032_Lys4xHis	12.5	0.4	0.271	0.992	0.519	0	0	1	T	0.137	−0.408	854.790	0.000	1.943	1.592
033_Lys4xThr	25	0.7	0.000	0.984	0.000	0	0	1	T	0.136	−0.504	890.104	0.000	1.962	1.576
034_Lys4xGln	25	1	0.000	0.976	0.000	0	0	1	V	0.141	−0.919	901.868	0.000	1.970	1.571
035_Lys5xArg	6.2	2.4	0.760	0.938	0.844	1	1	1	T	0.145	−0.500	901.875	0.000	1.948	1.579

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

Table 1. Continued

	properties		desirabilities			X_Class_1/0				molecular descriptors					
Cβ-HCP ID	MIC	%Hem	d_{MIC}	$d_{\%Hem}$	$D_{MIC-\%Hem}$	D_i	MIC	%Hem	T/V	RBF	Alog P	P_VSA_MR_2	D/Dtr10	SpMAD_B(s)	SpMAD_EA(bo)
036_Lys5xTrp	100	6	0.000	0.840	0.000	0	0	1	T	0.134	1.856	878.333	0.000	1.935	1.633
037_Lys5xLeu	25	27.4	0.000	0.260	0.000	0	0	0	V	0.137	1.238	878.333	0.000	1.939	1.586
038_Lys5xHis	25	4.1	0.000	0.892	0.000	0	0	1	T	0.137	−0.408	854.790	0.000	1.943	1.604
039_Lys5xThr	12.5	1.3	0.271	0.967	0.512	0	0	1	T	0.136	−0.504	890.104	0.000	1.962	1.588
040_Lys5xGln	12.5	1.1	0.271	0.973	0.514	0	0	1	T	0.141	−0.919	901.868	0.000	1.970	1.583
041_Arg6xTyr	25	1.8	0.000	0.954	0.000	0	0	1	V	0.133	1.722	921.016	0.000	1.949	1.602
042_Arg6xTrp	25	9.4	0.000	0.748	0.000	0	0	1	T	0.131	2.283	878.333	0.000	1.928	1.622
043_Arg6xLeu	12.5	2.6	0.271	0.932	0.503	0	0	1	T	0.134	1.665	878.333	0.000	1.932	1.573
044_Arg6xHis	12.5	0.3	0.271	0.995	0.519	0	0	1	V	0.135	0.019	854.790	0.000	1.936	1.592
045_Arg6xThr	25	2.5	0.000	0.935	0.000	0	0	1	V	0.133	−0.077	890.104	0.000	1.957	1.575
046_Arg6xGln	25	0.6	0.000	0.986	0.000	0	0	1	T	0.139	−0.492	901.868	0.000	1.965	1.571
047_Arg7xTyr	12.5	4.6	0.271	0.878	0.488	0	0	1	T	0.133	1.722	921.016	0.000	1.949	1.625
048_Arg7xTrp	6.2	1.3	0.760	0.967	0.857	1	1	1	T	0.131	2.283	878.333	0.000	1.928	1.644
049_Arg7xLeu	25	2.5	0.000	0.935	0.000	0	0	1	V	0.134	1.665	878.333	0.000	1.932	1.597
050_Arg7xHis	12.5	6.7	0.271	0.821	0.472	0	0	1	T	0.135	0.019	854.790	0.000	1.936	1.616
051_Arg7xGln	12.5	2.5	0.271	0.935	0.504	0	0	1	T	0.139	−0.492	901.868	0.000	1.965	1.595
052_Arg7xThr	6.2	2.7	0.760	0.930	0.840	1	1	1	V	0.133	−0.077	890.104	0.000	1.957	1.599
053_Trp8xPhe	12.5	2.1	0.271	0.946	0.507	0	0	1	T	0.145	−0.366	901.875	0.000	1.949	1.537
054_Trp8xTyr	25	1.7	0.000	0.957	0.000	0	0	1	T	0.145	−0.633	944.558	0.000	1.964	1.545
055_Trp8xY(B)	6.2	4.5	0.760	0.881	0.818	1	1	1	V	0.148	1.201	912.875	0.000	1.932	1.564
056_Trp8xHfe	9.4	3.1	0.512	0.919	0.686	1	1	1	T	0.147	0.090	901.875	0.000	1.944	1.537
057_Trp8 × 1-Nal	6.2	2.7	0.760	0.930	0.840	1	1	1	V	0.142	0.542	901.875	485.381	1.943	1.568
058_Trp8 × 2-Nal	6.2	3.5	0.760	0.908	0.830	1	1	1	T	0.142	0.542	901.875	499.201	1.942	1.567
059_Trp8xVal	12.5	2	0.271	0.949	0.507	0	0	1	T	0.144	−1.079	890.104	0.000	1.952	1.514
060_Trp8xCIF	6.2	3.3	0.760	0.913	0.833	1	1	1	T	0.145	0.298	901.875	0.000	1.950	1.545
061_Trp8xLeu	12.5	1.3	0.271	0.967	0.512	0	0	1	T	0.146	−0.690	901.875	0.000	1.948	1.513
062_Trp8xIle	12.5	2.3	0.271	0.940	0.505	0	0	1	V	0.146	−0.623	890.104	0.000	1.946	1.510
063_Trp8xOrn	50	3.1	0.000	0.919	0.000	0	0	1	T	0.151	−2.457	925.417	0.000	1.955	1.506
064_Lys9xTyr	18.8	2.3	0.000	0.940	0.000	0	0	1	T	0.135	1.295	921.016	0.000	1.955	1.613
065_Lys9xTrp	18.8	1.6	0.000	0.959	0.000	0	0	1	T	0.134	1.856	878.333	0.000	1.935	1.632
066_Lys9xLeu	12.5	3.5	0.271	0.908	0.496	0	0	1	T	0.137	1.238	878.333	0.000	1.939	1.585
067_Lys9xThr	12.5	2.7	0.271	0.930	0.502	0	0	1	V	0.136	−0.504	890.104	0.000	1.962	1.587
068_Lys9xGln	12.5	3.7	0.271	0.902	0.495	0	0	1	T	0.141	−0.919	901.868	0.000	1.970	1.583
069_Lys9xHis	25	1.4	0.000	0.965	0.000	0	0	1	T	0.137	−0.408	854.790	0.000	1.943	1.604
070_Tyr10xPhe	12.5	30.1	0.271	0.187	0.225	0	0	0	T	0.143	0.195	859.192	0.000	1.928	1.571
071_Tyr10xCha	50	57.6	0.000	0.000	0.000	0	0	0	T	0.140	0.868	859.192	0.000	1.916	1.544
072_Tyr10xTrp	12.5	3.9	0.271	0.897	0.493	0	0	1	V	0.141	0.488	859.192	0.000	1.922	1.599
073_Tyr10xHfe	12.5	31.8	0.271	0.141	0.196	0	0	0	T	0.145	0.651	859.192	0.000	1.923	1.577

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

Table 1. Continued

	properties		desirabilities			X_Class_1/0				molecular descriptors					
Cβ-HCP ID	MIC	%Hem	d_{MIC}	$d_{\%Hem}$	$D_{MIC-\%Hem}$	D_i	MIC	%Hem	T/V	RBF	Alog P	P_VSA_MR_2	D/Dtr10	SpMAD_B(s)	SpMAD_EA(bo)
074_Tyr10 × 1-Nal	12.5	29.3	0.271	0.209	0.238	0	0	0	T	0.140	1.103	859.192	517.057	1.922	1.599
075_Tyr10 × 2-Nal	50	39.6	0.000	0.000	0.000	0	0	0	V	0.140	1.103	859.192	531.554	1.921	1.599
076_Tyr10xVal	50	4.9	0.000	0.870	0.000	0	0	1	T	0.142	−0.518	847.420	0.000	1.930	1.550
077_Tyr10xCIF	9.4	34.8	0.512	0.060	0.175	0	1	0	T	0.143	0.859	859.192	0.000	1.929	1.578
078_Tyr10xLeu	9.4	26.6	0.512	0.282	0.380	0	1	0	V	0.144	−0.129	859.192	0.000	1.925	1.548
079_Tyr10xBip	6.2	33	0.760	0.108	0.287	0	1	0	T	0.141	1.713	859.192	0.000	1.916	1.600
080_Tyr10xOrn	50	0.1	0.000	1.000	0.000	0	0	1	T	0.148	−1.896	882.734	0.000	1.933	1.542
081_Arg11xTyr	12.5	4.4	0.271	0.883	0.490	0	0	1	T	0.133	1.722	921.016	0.000	1.949	1.602
082_Arg11xTrp	12.5	9.9	0.271	0.734	0.446	0	0	1	T	0.131	2.283	878.333	0.000	1.928	1.622
083_Arg11xLeu	25	5.9	0.000	0.843	0.000	0	0	1	T	0.134	1.665	878.333	0.000	1.932	1.573
084_Arg11xOrn	6.2	2.2	0.760	0.943	0.846	1	1	1	T	0.140	0.355	901.875	0.000	1.936	1.566
085_Arg11xCitt	25	1.4	0.000	0.965	0.000	0	0	1	T	0.141	−0.036	901.868	0.000	1.961	1.578
086_Val12xPhe	12.5	10.1	0.271	0.729	0.445	0	0	1	V	0.144	0.640	913.646	0.000	1.940	1.577
087_Val12xCha	12.5	35.2	0.271	0.049	0.115	0	0	0	V	0.141	1.313	913.646	0.000	1.929	1.551
088_Val12xTrp	6.2	10.5	0.760	0.718	0.739	1	1	1	T	0.141	0.934	913.646	0.000	1.934	1.603
089_Val12xHfe	12.5	21.7	0.271	0.415	0.335	0	0	0	T	0.146	1.097	913.646	0.000	1.935	1.582
090_Val12 × 1-Nal	12.5	3.7	0.271	0.902	0.495	0	0	1	T	0.140	1.549	913.646	554.885	1.934	1.604
091_Val12 × 2-Nal	12.5	23.3	0.271	0.371	0.317	0	0	0	T	0.140	1.549	913.646	570.198	1.934	1.603
092_Val12xTyr	25	6	0.000	0.840	0.000	0	0	1	T	0.143	0.373	956.329	0.000	1.954	1.583
093_Val12xCIF	6.2	6.4	0.760	0.829	0.794	1	1	1	T	0.144	1.305	913.646	0.000	1.941	1.583
094_Val12xLeu	12.5	4.5	0.271	0.881	0.489	0	0	1	T	0.145	0.316	913.646	0.000	1.938	1.554
095_Val12xBip	6.2	18.4	0.760	0.504	0.619	1	1	1	T	0.142	2.159	913.646	0.000	1.929	1.604
096_Val12xOrn	25	0.3	0.000	0.995	0.000	0	0	1	T	0.149	−1.451	937.188	0.000	1.945	1.548
097_Pro14xGly	25	1.1	0.000	0.973	0.000	0	0	1	T	0.147	−0.768	854.790	0.000	1.953	1.561
098_Pro14xArg	12.5	1.7	0.271	0.957	0.509	0	0	1	T	0.156	−0.897	901.875	0.000	1.945	1.556
099_Pro14xTyr	25	1	0.000	0.976	0.000	0	0	1	V	0.146	0.898	921.016	0.000	1.951	1.591
100_Pro14xPhe	6.2	0.7	0.760	0.984	0.864	1	1	1	T	0.147	1.165	878.333	0.000	1.937	1.584
101_Pro14xTrp	12.5	4.1	0.271	0.892	0.492	0	0	1	T	0.144	1.459	878.333	0.000	1.932	1.611
102_Pro14xLeu	12.5	2.7	0.271	0.930	0.502	0	0	1	T	0.149	1.328	925.417	0.000	1.938	1.562
103_Pro14xIle	25	2	0.000	0.949	0.000	0	0	1	T	0.148	0.909	866.562	0.000	1.935	1.559
104_Pro14xVal	12.5	1	0.271	0.976	0.514	0	0	1	T	0.146	0.452	866.562	0.000	1.940	1.564
105_Pro14xGln	50	1.2	0.000	0.970	0.000	0	0	1	T	0.152	−1.316	901.868	0.000	1.966	1.560
106_Pro14xCha	6.2	11.7	0.760	0.686	0.722	1	1	1	T	0.144	1.838	878.333	0.000	1.926	1.558
108_Pro14xHfe	12.5	5.9	0.271	0.843	0.478	0	0	1	T	0.149	1.622	878.333	0.000	1.933	1.590
109_Pro14 × 2-Nal	12.5	18.2	0.271	0.509	0.372	0	0	1	V	0.143	2.074	878.333	620.071	1.931	1.611
110_Pro14 × 1-Nal	12.5	13.1	0.271	0.648	0.419	0	0	1	T	0.143	2.074	878.333	603.703	1.932	1.611
112_Pro14xBip	6.2	13.1	0.760	0.648	0.701	1	1	1	V	0.145	2.684	878.333	0.000	1.926	1.612
113_Pro14xClF	6.2	4.2	0.760	0.889	0.822	1	1	1	T	0.147	1.830	878.333	0.000	1.938	1.591

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

Table 1. Continued

	properties		desirabilities			X_Class_1/0				molecular descriptors					
Cβ-HCP ID	MIC	%Hem	d_{MIC}	$d_{\%Hem}$	$D_{MIC-\%Hem}$	D_i	MIC	%Hem	T/V	RBF	AlogP	P_VSA_MR_2	D/Dtr10	SpMAD_B(s)	SpMAD_EA(bo)
114_Pro14xS(B)	6.2	19.7	0.760	0.469	0.597	0	1	0	T	0.152	0.713	854.790	0.000	1.934	1.580
115_Pro14xOrn	25	0.8	0.000	0.981	0.000	0	0	1	T	0.152	−0.926	901.875	0.000	1.943	1.556
116_Pro14xhCha	6.2	13.8	0.760	0.629	0.691	1	1	1	T	0.146	2.294	878.333	0.000	1.922	1.557
117_Pro14 × 14A	3.1	2.7	1.000	0.930	0.964	1	1	1	T	0.150	0.604	901.868	0.000	1.938	1.546
118_Pro14 × 14B	6.2	19.6	0.760	0.472	0.599	0	1	0	V	0.154	1.516	901.868	0.000	1.930	1.542
119_Pro14 × 14C	6.2	23.6	0.760	0.363	0.525	0	1	0	T	0.142	1.786	901.868	0.000	1.932	1.601
120_Pro14 × 14D	6.2	18	0.760	0.515	0.625	1	1	1	T	0.142	1.762	901.868	0.000	1.933	1.598
121_Pro14 × 14E	25	0.5	0.000	0.989	0.000	0	0	1	T	0.149	−2.725	925.410	0.000	1.948	1.556
122_Pro14 × 14F	100	1.9	0.000	0.951	0.000	0	0	1	V	0.153	−1.812	925.410	0.000	1.940	1.553
123_Pro14 × 14G	50	0.6	0.000	0.986	0.000	0	0	1	T	0.143	−1.431	901.868	0.000	1.952	1.565
124_Pro14 × 14H	25	0.5	0.000	0.989	0.000	0	0	1	V	0.145	−0.764	901.868	0.000	1.950	1.560
125_Pro14 × 14I	12.5	1.7	0.271	0.957	0.509	0	0	1	T	0.143	−0.302	901.868	0.000	1.946	1.563
126_Pro14 × 14J	6.2	2.7	0.760	0.930	0.840	1	1	1	T	0.145	−0.057	901.868	0.000	1.941	1.562
127_Pro14 × 14K	25	0.9	0.000	0.978	0.000	0	0	1	T	0.147	−0.308	901.868	0.000	1.946	1.559
128_Pro14 × 14 L	6.2	3.4	0.760	0.911	0.832	1	1	1	T	0.140	0.695	901.868	0.000	1.936	1.560
129_Pro14 × 14M	12.5	3.9	0.271	0.897	0.493	0	0	1	T	0.141	0.940	901.868	0.000	1.932	1.558
130_Pro14 × 14N	12.5	0.8	0.271	0.981	0.516	0	0	1	T	0.149	0.148	901.868	0.000	1.942	1.556
131_Pro14 × 14O	6.2	2.2	0.760	0.943	0.846	1	1	1	T	0.142	0.233	901.868	0.000	1.947	1.586
132_Pro14 × 14P	12.5	1	0.271	0.976	0.514	0	0	1	T	0.144	0.268	901.868	0.000	1.943	1.590
133_Pro14 × 14Q	9.4	0.2	0.512	0.997	0.714	1	1	1	T	0.142	0.561	901.868	0.000	1.937	1.608
134_Pro14 × 14R	25	0.7	0.000	0.984	0.000	0	0	1	T	0.144	1.017	901.868	0.000	1.934	1.612
135_Pro14 × 14S	6.2	3	0.760	0.921	0.837	1	1	1	T	0.141	1.751	901.868	0.000	1.936	1.613
136_Pro14 × 14T	6.2	3.4	0.760	0.911	0.832	1	1	1	T	0.160	3.311	901.875	0.000	1.903	1.537
137_Pro14 × 14U	12.5	11	0.271	0.705	0.437	0	0	1	T	0.146	1.493	878.333	0.000	1.929	1.582

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

Table 2. Classification Performance of the Multicriteria D_i CART Classifier

IF RBF > 0.1395 AND ALOGP > -0.509 AND 872.4475 < P_VSA_MR_2 ≤ 919.5315
THEN, D_i_Class_1; OTHERWISE D_i_Class_0

TRAINING SET					VALIDATION SET			
		OBSERVED					OBSERVED	
		0	1	*Confusion Matrix*			0	1
PREDICTED	0	**62**	1		PREDICTED	0	**17**	1
	1	21	**24**			1	4	**6**
				CLASSIFICATION PERFORMANCE[a]				
79.63				*Accuracy (%)*	82.14			
96.00				*Sensitivity/TP Rate (%)*	85.71			
74.70				*Specificity/TN Rate (%)*	80.95			
4.00				*FN Rate (%)*	14.29			
25.30				*FP Rate (%)*	19.05			
60.48				*MCC (%)*	60.24			
68.57				*F-Measure* for D_i_Class_1 *(%)*	70.59			
84.93				*F-Measure* for D_i_Class_0 *(%)*	87.18			

[a] TP, TN, FN, FP and MCC stand for true positive, true negative, false negative, false positive, and Matthews correlation coefficient, respectively.

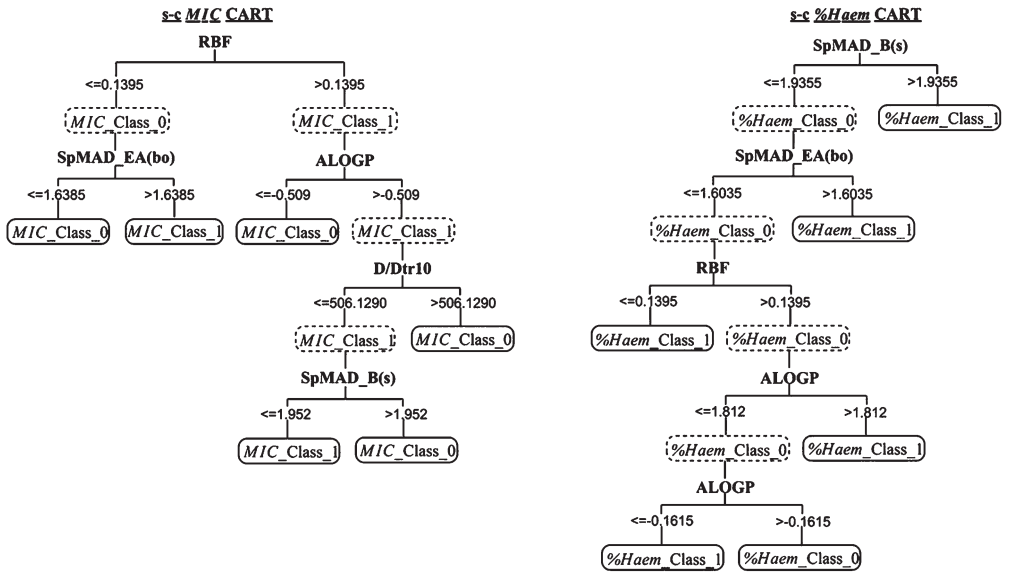


Figure 2. Tree graphs corresponding to the respective single-criterion CART classifiers for MIC and %Hem.

well as variables with standard deviation less than 0.0001, a total of 1724 molecular descriptors remained for further analyses. Details and definitions of each molecular descriptor included in the respective 18 blocks can be found in ref 40, a particularly comprehensive review of the available molecular descriptors and how they are calculated. The 18 blocks and the number of molecular descriptors in the respective block are depicted in Table S3 of the Supporting Information. Exact data of the 1724 molecular descriptors employed in the modeling stage are also provided as Supporting Information of this paper.

The choice on how to codify the chemical structure of Cβ-HCPs used in this work relies on four main elements as follows:

(i) Numerous studies with linear, cyclic, and diastereomeric antimicrobial peptides have strongly supported the hypothesis that their physicochemical properties, rather than any specific amino acid sequence, are responsible for their microbiological activities.[41] So, global molecular descriptors are preferred over amino acid specific descriptors.

(ii) Enantiomers of PG-I and its cyclic peptidomimetics show a comparable antimicrobial activity, suggesting that their mechanism of action does not involve binding to a chiral receptor but rather reflects a nonenantiospecific interaction with membrane lipids.[8] Therefore, three-dimensional (3D) chiral codification is not essential.

(iii) β-hairpin conformation plays a determinant role in the antimicrobial activity and the D-Pro-L-Pro template used by Robinson et al. is known to adopt a stable type-II β-turn [42] and so is ideal to nucleate β-hairpin conformations.[43,44] Thus, as specific 3D information, it is assumed by considering a predefined β-hairpin bioactive conformation ensured by the D-Pro-L-Pro template, 0−2D structural information is preferred to unravel unidentified aspects of the antimicrobial peptide ligand−target(s) interaction.

(iv) Computational approaches for virtual screening, besides a proper enrichment performance, must possess an adequate computational efficiency, so simple and fast approaches are favored over computationally intensive methods.[45]

From (i−iii), it is apparent that 3D information is not crucial to codify the SAR encoded in this family of compounds. On the other hand, the calculation of 3D molecular descriptors besides

Table 3. Classification Performance of the Respective Single-Criterion MIC and %Hem CART Classifiers[a]

Single-criterion CART Classifier for MIC						*Single-criterion CART Classifier for %Hem.*					
TRAINING				*VALIDATION*		*TRAINING*				*VALIDATION*	
OBS.				OBS.		OBS.				OBS.	
	0	1			0	1					0	1
PRED. 0	**57**	0	*Confusion Matrix*	PRED. 0	**15**	1	PRED. 0	**10**	6	*Confusion Matrix*	PRED. 0	**4**	2
1	22	**29**		1	4	**8**	1	0	**92**		1	1	**21**
79.63			*Acc* (%)	82.14			94.44			*Acc* (%)	89.29		
100.00			*Se* (%)	88.89			93.88			*Se* (%)	91.30		
72.15			*Sp* (%)	78.95			100.00			*Sp* (%)	80.00		
0.00			*FN* (%)	11.11			6.12			*FN* (%)	8.70		
27.85			*FP* (%)	21.05			0.00			*FP* (%)	20.00		
64.05			*MCC* (%)	64.02			76.60			*MCC* (%)	66.55		
*MIC*_Class_1: 72.50			*F* (%)	*MIC*_Class_1: 76.19			*%Hem.*_Class_1: 96.84			*F* (%)	*%Hem.*_Class_1: 93.33		
*MIC*_Class_0: 83.82				*MIC*_Class_0: 85.71			*%Hem.*_Class_0: 76.92				*%Hem.*_Class_0: 72.73		

[a] Acc = accuracy; Se = sensitivity or TPs rate; Sp = specificity or TNs rate; FN = false negatives rate; FP = false positives rate; MCC = Mathews correlation coefficient, *F* = *F*-measure.

more computationally complex than lower (0−2D) dimensional molecular descriptors requires a computationally demanding conformational optimization preprocessing, which leads us to consider (iv). Even so, it is important to note that successful peptide QSAR can be also obtained with descriptors based on the physicochemical properties of amino acids, such as the Z1-Z5 Scale by Wold,[46] the Amino Acid Scales from AA Index Database,[47] or the E1-E5 Scale[48] used for allergenic proteins by Ivanciuc et al.[49−51] Other amino acids scales for peptides have been compared in a recent work by Ivanciuc.[52]

Training/Validation Data Splitting. The process of data splitting on training (T) and validation (V) subsets was carried out by the sequential application of principal component analysis and joining tree/*k*-means cluster analysis,[53] as implemented in the respective modules of Principal Components and Classification Analysis and Cluster Analysis of STATISTICA 8.0.[54] Thus, the molecular structure of each peptidomimetic was represented as a vector of 57 variables: 20 constitutional indices, 20 ring descriptors, and 17 molecular properties. A principal component analysis was independently applied to the members of each class (D_i_Class_1: 32; D_i_Class_0: 104). Like this, D_i_Class_1/0 members are now described as a smaller vector of 6/15 principal components accounting for 96/99% of the variance, respectively. Then, a joining tree cluster analysis using the complete linkage as amalgamation rule was independently applied to the members of each class. In this way it is possible to obtain a graphical picture of the distribution and clustering of members of each class according to the principal components-based structural representation provided.

Subsequently, a *k*-means cluster analysis using the Euclidean distance as a structural proximity (actually, dissimilarity) measure was independently applied to the members of each class. The number of clusters (5/10 clusters for D_i_Class_1/D_i_Class_0) was set in such a way that members of each cluster agree as much as possible with the distribution obtained in the previous joining tree cluster analysis. Finally, 20% of the respective D_i_Class_1/0 members were reserved for validation so that each cluster was represented on both training and validation subsets. As a result, 108 (25/83 D_i_Class_1/D_i_Class_0) out of 136 peptidomimetics were used for training, while the remaining 28 (7/21 D_i_Class_1/D_i_Class_0) were reserved for model validation and never used for training. Distribution of training and validation cases can be accessed by checking Table 1. This procedure ensures that both training and validation subsets are uniformly populated from the molecular structure point of view and that each structure pattern in the validation subset is represented in the training subset. The goal here is to guarantee that predictions of new compounds based on models derived from such a training subset will be based on interpolations, avoiding the lack of reliability associated to extrapolations.[38,55]

Table 4. Results of the Concordance Test for Training, Validation, and Full Sets

	agree	disagree	% agreement
	Training Set		
D_i_Class_1	22 out of 25	3 out of 25	88.00
D_i_Class_0	72 out of 83	11 out of 83	86.75
total	94 out of 108	14 out of 108	87.04
	Validation Set		
D_i_Class_1	6 out of 7	1 out of 7	85.71
D_i_Class_0	17 out of 21	4 out of 21	80.95
total	23 out of 28	5 out of 28	82.14
	Overall		
D_i_Class_1	28 out of 32	4 out of 32	87.50
D_i_Class_0	89 out of 104	15 out of 104	85.58
total	117 out of 136	19 out of 136	86.03

Data Modeling and Validation. Classification and regression trees (CART)[56] was the machine learning technique applied to our classification problem. We specifically used CART implemented in the Data Mining module of STATISTICA 8.0. The classification tree training was specified by using equal a priori probabilities and misclassification costs for both classes. Prune on misclassification error was the rule applied for stopping tree splitting. The predictive accuracy of the classification tree was tested on the validation sample.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical applications: the independent data set test, the

subsampling test (usually 5-fold or 10-fold cross-validation), and the jackknife test.[57] However, as elucidated in the work by Chou[23] and demonstrated by eqs (28−32) therein, among the three cross-validation methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark data set, being thus increasingly used and recognized by researchers to examine the accuracy of various predictors (see, e.g., refs 24 and 58−65). However, to reduce computational time, in this study we have adopted the independent data set for cross-validation as a compromise, as many researchers did when using SVM as a prediction engine.

Both the learning and the predictive ability of the classification tree models were assessed by checking overall and class-specific performance measures on training and external validation sets, respectively.[66] Accuracy, F-measure,[67] and Matthews correlation coefficient[68] were used to quantify the overall predictive ability of the classifier. Sensitivity and specificity were computed to quantify the class-specific predictive performance of the classifier since they encode the ability of the classifier to identify positive and negative cases, respectively. False positive and false negative rates were also computed to measure the respective misclassification costs associated to the specificity and sensitivity of the classifier. Definitions for all of these classification performance metrics are provided in the Supporting Information.

Enrichment Assessment. Several enrichment metrics have been proposed in the literature to measure the enrichment ability of a virtual screening protocol.[69] In this work, we use some of the most extended metrics. Based on the analysis of the receiver operating characteristic (ROC) curve,[70] it is possible to derive the area under the ROC curve (ROC metric)[69] as well as the ratio of true positive (TP) and false positive (FP) cases found at the operating point of the ROC curve ($TP/FP_{ROC\text{-}OP}$).[66] From the accumulation curve, one can deduce enrichment from the area under the curve (AUAC),[69] from the yield of actives at certain filtered fractions ($Y_{a_{5\%}}$, $Y_{a_{10\%}}$, $Y_{a_{20\%}}$, and $Y_{a_{50\%}}$) as well as from the fraction of the database that has to be screened in order to retrieve a certain percentage of TP cases (screening percentage: $\chi_{25\%}$, $\chi_{50\%}$, $\chi_{75\%}$, and $\chi_{100\%}$). Finally, the enrichment factor (EF) takes into account the improvement of the hit rate by a virtual screening protocol compared to a random selection. Details on definition and interpretation of the above-mentioned enrichment metrics can be found in the Supporting Information.

RESULTS AND DISCUSSION

Identification of Selective Antibacterial Cβ-HCPs using a CART-Derived Multicriteria Classification Rule. In order to identify Cβ-HCPs with selective permeation of Gram-negative bacteria membranes over human red blood cells (i.e., potent and nonhemolytic antimicrobial Cβ-HCPs), a multicriteria classification QSAR study was carried out. As depicted previously, we took advantage of machine learning tools to solve the above-mentioned multicriteria classification problem. Specifically, the CART algorithm was used to find a tree able to efficiently separate both classes, D_i_Class_1 and D_i_Class_0, of Cβ-HCPs. CART was selected over many other alternative techniques due to the atypical convergence of nonlinearity and interpretability of this technique. As a result, complex nonlinear relationships between predictor and dependent variables can be uncovered by usually a few logical if−then rules.[71]

The architecture of the simplest best performing classification tree found can be consulted in the Supporting Information (see Figure S1). Predictions are made based on three simple molecular descriptors: the rotatable bond fraction (RBF), the Ghose−Crippen−Viswanadhan octanol−water partition coefficient (Alog P), and the amount of van der Waals surface area occupied by atoms with molar refractivity at bin no. 2 (P_VSA_MR_2).[72] Although already simple, the classification tree can be translated into a simpler rule for the identification of selective antibacterial Cβ-HCPs:

```
IF RBF > 0.1395
AND ALOGP > -0.509
AND 872.4475 < P_VSA_MR_2 ≤ 919.5315
THEN, Classify Cβ-HCP i as Di_Class_1;
OTHERWISE Classify Cβ-HCP i as Di_Class_0
```

The application of this rule to the training set was characterized by high levels of accuracy and specificity, a particularly high sensitivity (96%), and an acceptable FP rate. However, the predictive accuracy of the rule evaluated on the external validation test set was slightly superior, reaching balanced levels of sensitivity and specificity around 80% (see Table 2 for details on the classification performance). Additionally, Figure S2 in the Supporting Information allows one to visually assess the classification performance of the rule by examining the 3D chemical space defined by RBF, Alog P, and P_VSA_MR_2. Based on the level of complexity of this multicriteria classification problem, the predictive accuracy reached on the external test set, and the simplicity of the rule as well as the simplicity and full interpretability of the molecular descriptors selected, the simple multicriteria classification rule proposed can be regarded as predictive, interpretable, and reliable.

Concordance Test. In order to challenge the reliability of the multicriteria classification rule for the identification of selective cyclic β-hairpin cationic antimicrobial peptides, a further concordance test was applied. This concordance test is based on the assumption that accurate predictions of a composite property (that is, D_i used as a measure of the balance between MIC and %Hem)

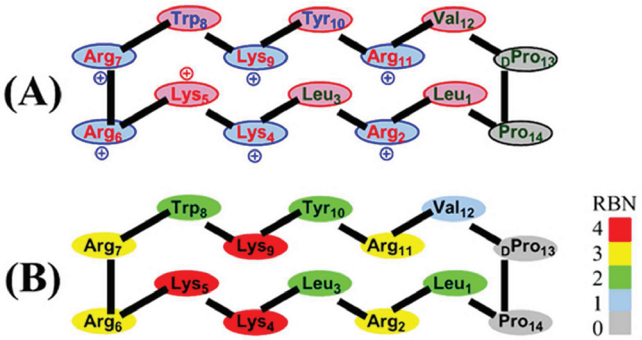


Figure 3. Cβ-HCP lead of Robinson et al.[8] (A) Residues are represented by the respective three-letter code and colored according to their classification based on their hydrophobic (green), polar (blue), or charged (red) character. Blue and red spheres were used to represent the respective polar and nonpolar faces of the Cβ-HCP lead. The β-hairpin-like secondary structure of the Cβ-HCP lead allows to formally distinguish between cationic residues and the β-hairpin stabilizing D-Pro_{13}-L-Pro_{14} template forming the polar face (Arg_2, Lys_4, Arg_6, Arg_7, Lys_9, Arg_{11}, Pro_{13}, and Pro_{14}) and residues with a lipophilic/aromatic character (Leu_1, Leu_3, Lys_5, Trp_8, Tyr_{10}, and Val_{12}) forming the nonpolar face of the Cβ-HCP. Lys_5, Pro_{13}, and Pro_{14} represent a deviation from an ideal cyclic amphipathic motif. Nevertheless, due to the antiparallel β-sheet structure of the backbone, the side chains of the residues occupy the space roughly on opposite sides (faces) of the backbone scaffold. (B) Residues are represented using a flexibility scale.

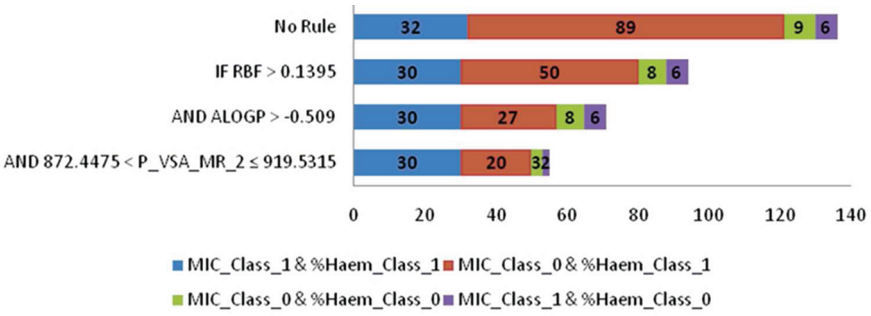


Figure 4. Graphic representation of the sequential application of the multicriteria classification rule. Each bar fragment is labeled with the number of Cβ-HCPs present in the subset and colored according to the corresponding class: D_i_Class_1 Cβ-HCPs: MIC_Class_1 and %Hem_Class_1 (blue); D_i_Class_0 Cβ-HCPs: MIC_Class_0 and %Hem_Class_1 (red), MIC_Class_0 and %Hem_Class_0 (green), or MIC_Class_1 and %Hem_Class_0 (magenta).

Table 5. Classification performance of the multicriteria D_i CART classifier for concordant predictions

TRAINING SET		VALIDATION SET
OBSERVED		OBSERVED
0 1	*Confusion Matrix*	0 1
PREDICTED 0 **60** 0		PREDICTED 0 **16** 1
1 12 **22**		1 1 **5**
	CLASSIFICATION PERFORMANCE[a]	
87.23	***Accuracy (%)***	91.30
100.00	***Sensitivity/TP Rate (%)***	83.33
83.33	***Specificity/TN Rate (%)***	94.12
0.00	***FN Rate (%)***	16.67
16.67	***FP Rate (%)***	5.88
73.43	***MCC (%)***	77.45
78.57	***F-Measure* for D_i_Class_1 *(%)***	83.33
90.91	***F-Measure* for D_i_Class_0 *(%)***	94.12

[a] TP, TN, FN, FP, and MCC stand for true positive, true negative, false negative, false positive, and Matthews correlation coefficient, respectively.

for a compound *i* must agree with accurate predictions of each independent property (MIC and %Hem). Thus, independent classification models for each property determining D_i can act as a feedback for the composite classifier. Specifically, for our particular case the following conditions should be satisfied for a reliable separation of the D_i_Class_1 from the D_i_Class_0 of Cβ-HCPs:

Condition I (D_i_Class_1 => MIC_Class_1 + %Hem_Class_1). Compound *i* is reliably predicted as belonging to the D_i_Class_1 if it is also assigned to both the class of antimicrobial peptides (MIC_Class_1) and the class of nonhemolytic peptides (%Hem_Class_1) by independent classification models for the respective properties; otherwise the prediction cannot be regarded as concordant.

Condition II (D_i_Class_0 => MIC_Class_0 + %Hem_Class_0 or MIC_Class_1 + %Hem_Class_0 ir MIC_Class_0 + %Hem_Class_1). Compound *i* is reliably predicted as belonging to the D_i_Class_0 if it is also assigned to both the class of nonantimicrobial peptides (MIC_Class_0) and the class of hemolytic peptides (%Hem_Class_0), or MIC_Class_1 and %Hem_Class_0, or MIC_Class_0 and %Hem_Class_1 by independent classification models for the respective properties; otherwise the prediction cannot be regarded as concordant.

Thresholds to assign compounds to the MIC_Class_1/0 or %Hem_Class_1/0 were selected as the respective highest values of MIC (9.4 μg/mL) and %Hem (18.4% at 100 μg/mL) exhibited by the subset of compounds belonging to the D_i_Class_1 ($D_i \geq 0.6$). In doing so, we ensure that the above-mentioned concordance conditions (I and II) hold true between the actual classes set by the respective, D_i, MIC, and %Hem thresholds. The same scheme applied for the D_i_Class_1/0 training/validation splitting was used for the new classes based on the MIC and %Hem thresholds. So, MIC_Class_1 includes 38 cases (29/9 for training/validation) while MIC_Class_0 includes 98 cases (79T/19 V). On the other hand %Hem_Class_1 includes 121 cases (98T/23 V) while %Hem_Class_0 includes 15 cases (10T/5 V). Details on the distribution of classes are given in Table 1.

A CART classification analysis was applied to both MIC and %Hem problems. The best performing classification tree found for each property is shown in Figure 2. Names and definitions of variables included on the respective classifiers are depicted in the Supporting Information (see Table S4). While the single-criterion CART classifier for MIC shows a classification performance comparable to the multicriteria CART classifier for D_i in both training and validation sets, the single-criterion CART classifier for %Hem exhibits a superior predictive accuracy on both training and validation sets (see details in Table 3).

An elevated level of overall agreement (86%) between the multicriteria D_i CART classifier and both the single-criterion CART classifiers for MIC and %Hem was observed after checking the fulfilling of conditions I and II. An elevated degree of concordance, always higher than 80%, was also achieved for both D_i_Class_1/0 classes regarding the training, validation, and overall sets (see Table 4 for details). The high level of concordance found can be considered an important measure of the reliability of the composite predictions made by the multicriteria classification rule proposed. Based on the sensitivity (TP rate) exhibited by the multicriteria classification rule on training

and validation sets (79.6% and 82.1%, respectively) as well as the 87.5% of agreement achieved between the multicriteria classification rule and the single-criterion MIC CART and %Hem CART classifiers for compounds included on the class of selective antimicrobial peptides (D_i_Class_1), we can realistically expect that by using the rule, the probability of identifying a selective antibacterial cyclic cationic β-hairpin peptidomimetic is about 80% and that such a prediction can be regarded as true with a level of reliability of about 87%.

Biophysical Relevance and Inference Making. Whenever using high-dimensional QSAR data, if the predictive ability of the model is not properly validated and is based on variables lacking some structural or biophysical sense, there is a high probability of generating a statistically significant but unreliable QSAR model. As cautioned by Unger and Hansch,[73] one is apt to generate "statistical unicorns", beasts that exist in papers but not in reality. So, checking the biophysical coherence of the model as well as the convergence of the inferences suggested by the model and experimental findings on the problem under study, it is a very important sign of reliability.

It has been proposed that cationic peptides initially bind to negatively charged lipopolysaccharides and phospholipids of the outer leaflet of bacterial membrane, accumulate and aggregate on the membrane surface, and finally permeabilize/disintegrate the bacterial membrane through various possible mechanisms including pore formation.[74] Thus, a biophysically coherent rule should be consistent with this general mode of action. To check this, each of the three subrules making up the multicriteria classification rule was sequentially applied and analyzed.

First Sub-Rule: RBF > 0.1395. If we apply this subrule, one finds that approximately 90% of cases uncovered by the RBF subrule share a common pattern related to substitutions of cationic residues. The pattern observed indicates that it is accounting for the deleterious effect over the antimicrobial properties of substitutions on cationic residues oriented to the polar face observed in the work by Robinson et al.[8] (see Figure 3A for a schematic representation of the general molecular structure of Cβ-HCPs[15]). Similar observations relating the antimicrobial effect of cationic peptides to the molecular charge have been previously published by several laboratories.[75−79]

A flexibility descriptor accounting for charge properties does not seem to be very coherent. However, in these substitutions (i.e., on cases uncovered by the subrule) the loss in the net charge of the molecule was also accompanied by a significant reduction of the flexibility of the polar face. This trend suggests that a certain degree of flexibility is required in addition to a net positive charge for binding to negatively charged lipopolysaccharides and phospholipids of the outer leaflet of the bacterial membrane. The probable influence over peptide antibacterial activity of other or additional biophysical parameters rather than just cationicity, amphipaticity, and lipophilicity have already been suggested by other authors.[11,80,81] Moreover, as defined in ref 72, the numbers of rotatable bonds of the Lys, Arg, Trp, Tyr, Leu, Val, and Pro residues are 4, 3, 2, 2, 2, 1, and 0, respectively. If we use this "scale of flexibility" to represent the Cβ-HCP lead, as in Figure 3A, we will realize that the polar face is obviously also the charged face but is, in addition, the flexible face (see Figure 3B).

Second Subrule: Alog P > −0.509. When analyzing the consistency of the Alog P subrule, (Alog P > −0.509), we must expect that after applying it, most of the D_i_Class_1/D_i_Class_0 Cβ-HCPs should fall into the subset covered/uncovered by the subrule, which is what actually happens. The same pattern was observed for the MIC_Class_1/0 which suggests that Alog P, an hydrophobicity/lipophilicity descriptor, like RBF, is accounting for the antibacterial properties of the Cβ-HCPs. Specifically, 20 out of the 26 Cβ-HCPs uncovered by the rule were characterized by substitutions on the hydrophobic (Leu_1, Leu_3, Val_{12}, and Pro_{14}) and polar (Trp_8 and Tyr_{10}) residues, which in turn induce a significant reduction of their lipophilicity, with the resulting loss of their amphipatic character. The role of amphipaticity for aggregation on the bacterial membrane surface and lipophilicity for permeation into lipophilic membrane interior has been reported by several authors[74,82−84] as critical steps for antimicrobial peptides bacterial membrane disruption. These properties have been also used to successfully explain antimicrobial potency of some cationic peptides,[15,75−79] which thus supports the consistency of this subrule.

Third Subrule: 872.4475 < P_VSA_MR_2 ≤ 919.5315. For our specific problem, P_VSA_MR_2 codifies the approximate van der Waals surface area occupied by atoms in the Cβ-HCP molecule having a value of molar refractivity in the range [0.9, 1.5]. These atom types are atom centered fragments (ACF)[85] of the type H-052: hydrogen atoms attached to sp^3 carbon atoms with one electronegative atom (O, N, S, P, Se, halogens) attached to the next carbon atom (molar refractivity = 0.9215); O-057: fenol/enol/carboxyl OH groups (molar refractivity = 1.4778); and O-058: carbonyl oxygens (molar refractivity = 1.4429).[39,72] A detailed definition of P_VSA_MR_2-type descriptors can be found in the work by Labute.[86] Excluding the carbonyl oxygens in the cyclic peptide backbone and the H-052 atom types in the sp^3 carbon atom directly attached to the Cα in the peptide backbone, which are constantly present on all the residues, the only residues

Table 6. Enrichment Performance of the Multicriteria Virtual Screening Strategy Regarding the Training, Validation, and Overall Sets of Cβ-HCPs[a]

training set			validation set
0.914	ROC metric		0.908
0.320/0.012	$TP/FP_{ROC\text{-}OP}$		0.714/0.048
0.818	AUAC		0.806
0.380	$\chi_{100\%}$		0.607
4.320	EF_{Max}		4.000
OVERALL			
ROC metric	0.899	AUAC	0.805
$\chi_{25\%}$	0.066	$Y_{a5\%}$	0.857
$\chi_{50\%}$	0.184	$Y_{a10\%}$	0.786
$\chi_{75\%}$	0.265	$Y_{a20\%}$	0.667
$\chi_{100\%}$	0.890	$Y_{a50\%}$	0.456

$TP/FP_{ROC\text{-}OP}$	0.313/0.010		
$EF_{5\%}$	3.643		
$EF_{10\%}$	3.339		
$EF_{20\%}$	2.833		
EF_{max}	4.250		

[a] ROC metric: area under the ROC curve; AUAC: area under the accumulation curve; $TP/FP_{ROC\text{-}OP}$: ratio of true positive (TP) cases and false positive (FP) cases found at the operating point of the ROC curve; $\chi_{25\%/50\%/75\%/100\%}$: fraction of the database that has to be screened in order to retrieve a 25%/50%/75%/100% of true positive cases; $Ya_{5\%/10\%/20\%/50\%}$: yield of actives after filtering the top 5%/10%/20%/50% of the database; $EF_{5\%/10\%/20\%/50\%}$: enrichment factor achieved after filtering the top 5%/10%/20%/50% of the database; and EF_{max}: maximum enrichment factor achieved.

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

contributing significantly to P_VSA_MR_2 are Lys and Arg (two H-052 atoms each) and Tyr (a phenol OH). So, only Lys, Arg, and Tyr residues are determinant for P_VSA_MR_2.

It is interesting to note that when one applies the last subrule to the subset of Cβ-HCPs remaining after the sequential application of RBF and Alog P subrules, 9 out of the 11 substitutions made on Tyr_{10} were then eliminated. This finding is consistent with the conclusions taken from the SAR study of Robinson et al.[8] regarding hemolytic properties of Cβ-HCPs, i.e., substitutions on Tyr_{10} can cause large increases in hemolytic activity. The role of the P_VSA_MR_2 subrule for the codification of the hemolytic properties of Cβ-HCPs can be confirmed in Figure 4, which shows a graphic representation of the sequential application of the multicriteria classification rule. From this figure, it is apparent that RBF and Alog P only account for the antibacterial properties since it is not until the application of the P_VSA_MR_2 subrule that Cβ-HCPs with poor hemolytic profiles (%Hem_Class_0) is eliminated. The role of Tyr_{10} on determining a selective antibacterial activity could be related with a reduction of the overall lipophilicity of Cβ-HCPs induced by this residue that could be still acceptable to permeate into the lipophilic bacterial membrane but not to permeate red blood cells membranes. Finally, it is possible to assert that the multicriteria classification rule derived is consistent with the general mode of action of cationic peptides.

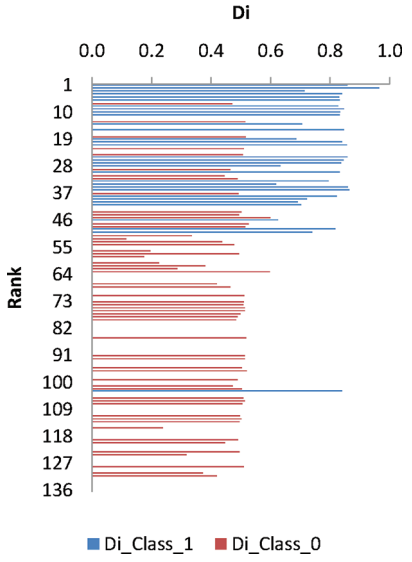


Figure 5. DCS score-based ranking of the overall set of Cβ-HCPs.

Multicriteria Virtual Screening Strategy for Selective Antibacterial Cβ-HCPs Prioritization. Hitherto, the multicriteria classification rule has proved to be a predictive, biophysically consistent, and reliable classifier. Therefore, its use as virtual screening tool could provide a practical solution to the problem of identifying selective nonhemolytic antibacterial Cβ-HCPs. A high-quality virtual screening tool should render an ordered list of candidates, where the promising ones are placed at the top of the list, while the noninteresting candidates are relegated to the bottom of the list. Toward that end, the predictive algorithm on which it is based, the virtual screening tool must be characterized by a good predictive performance. Nevertheless, for virtual screening the most important features are a particularly high/low TP/FP rate, looking for maximizing/minimizing the number of actual positive/negative cases regarded as positive by the predictive algorithm. The other key feature is to provide a measure for the quantitative scoring of the target property, so that using it as ranking criterion, the resultant ordered list resembles as much as possible the actual levels of the target property.

Our multicriteria classification rule shows an attractively high TP rate (96%/86% on training/validation sets) but an acceptable and slightly high FP rate (25%/19% on training/validation sets). This means that, based on its validation performance, a subset of Cβ-HCPs classified as D_i_Class_1 by using the multicriteria classification rule will contain 86% of the D_i_Class_1 Cβ-HCPs screened but 19% of D_i_Class_0 Cβ-HCPs, which is not a bad performance, but not what we are seeking for in a virtual screening campaign. The other problem of the CART-derived multicriteria classification rule is related to the quantitative measure used for ranking. Specifically, the only available option for library ranking is using posterior probabilities assigned by the multicriteria CART classifier for D_i_Class_1 ($PP_{D_i\_Class\_1}$). The problem here is the low variability of $PP_{D_i\_Class\_1}$ values coming from the multicriteria CART classifier (in general from CARTs), causing many cases to share equal $PP_{D_i\_Class\_1}$ values, thus avoiding its use as a ranking criterion. Even so, the idea of a single virtual screening tool able to filter Cβ-HCPs with antibacterial and hemolytic profiles simultaneously improved is still highly attractive. So, we focused on finding a solution to use the CART-derived multicriteria classification rule as the basis of a virtual screening tool for selective nonhemolytic antibacterial Cβ-HCPs prioritization.

A solution to the first limitation, a nonoptimal FP rate, was found on the high concordance shown by the CART-derived multicriteria classification rule and the individual single-criterion CART classifiers for MIC and %Hem, respectively. If cases with

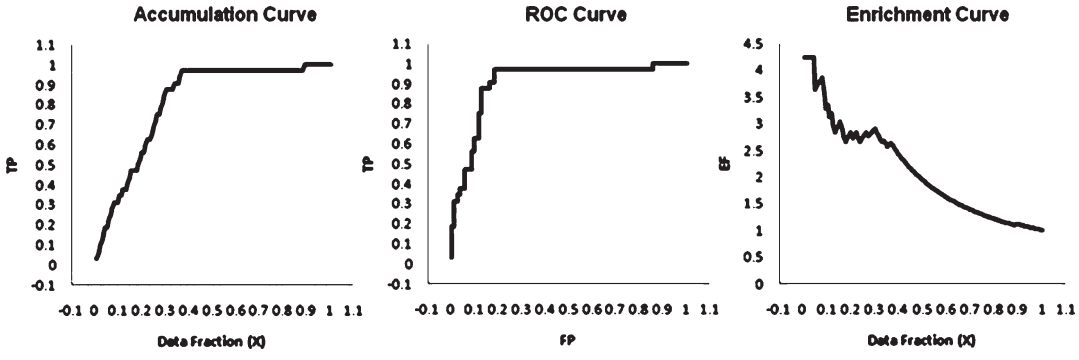


Figure 6. Accumulation, ROC, and enrichment curves for the DCS score-based rank of the overall set of Cβ-HCPs.

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

nonconcordant predictions are removed and the prediction performance is re-evaluated, a noteworthy improvement for the training and validation sets is observed. Specifically, the FP rate was significantly improved from 25%/19% to 17%/6% on the training/validation sets (see details in Table 5). It is apparent that the aid of both individual single-criterion CART classifiers can enhance the predictive performance, specifically those features required for virtual screening. Hence, their use as part of the virtual screening strategy should be considered.

The second drawback (i.e., the limited variability of $PP_{D_i\_Class\_1}$ values) was tackled by the use of similarity concepts. Specifically, a similarity measure, the Tanimoto coefficient (tan)[87] was used to modify $PP_{D_i\_Class\_1}$ and attain the variability required for library ranking. By using as query for similarity measuring the Cβ-HCPs with the best balance between antibacterial and hemolytic properties (117_Pro14 × 14A: D_i = 0.964; MIC = 3.1 μg/mL; % Hem = 2.7%) one also ensures that tan will act as a weighting factor favoring cases structurally close to the query and consequently pushing up to the top of the ordered list those Cβ-HCPs similar to 117_Pro14 × 14A.

The molecular similarity of the full set of Cβ-HCPs to the 117_Pro14 × 14A query was quantified by using the respective three-variable vector comprising the molecular descriptors included on the multicriteria classification rule (RBF, Alog P, and P_VSA_MR_2). The Tanimoto similarity coefficient for quantitative features is defined by

$$\tan = \frac{\sum_{j=1}^{n} X_{j\mathrm{I}} \times X_{j\mathrm{Q}}}{\sum_{j=1}^{n} X_{j\mathrm{I}} \times X_{j\mathrm{I}} + \sum_{j=1}^{n} X_{j\mathrm{Q}} \times X_{j\mathrm{Q}} + \sum_{j=1}^{n} X_{j\mathrm{I}} \times X_{j\mathrm{Q}}} \quad (5)$$

where $X_{j\mathrm{I}}$ and $X_{j\mathrm{Q}}$ represent the value of the molecular descriptor j for a molecule I, and the molecule Q used as query, respectively.

The respective posterior probabilities of belonging to Class_1 assigned by the multicriteria CART classifier for D_i and the single-criterion CART classifiers for MIC and %Hem were used jointly along with the Tanimoto similarity values to derivate a quantitative feature suitable for library ranking. The whole process involves the encoding of joint probabilities for MIC and %Hem, the credibility encoding of concurrent probabilities, and the derivation of a desirability-credibility-similarity score. The main steps of this process are summarized:

(i) Joint probability desirability encoding: The geometric mean of the n desirability values for n properties encodes the level of desirability of the balance of the n properties. Both, desirability values and probabilities are included in the range [0,1], and for both the target is 1. Therefore, the same concept can be applied to probability values. Based on this observation, it is possible to obtain a joint value encoding the probability of a Cβ-HCP to exhibit a desirable balance between antibacterial and hemolytic properties ($P_{\mathrm{MIC+\%Hem}}$):

$$P_{\mathrm{MIC+\%Hem}} = \sqrt[2]{PP_{\mathrm{MIC_{Class\,1}}} \times PP_{\mathrm{\%Hem_{Class\,1}}}} \quad (6)$$

where $PP_{MIC\_Class\_1}$ and $PP_{\%Hem\_Class\_1}$ are the posterior probabilities of belonging to Class_1 for each Cβ-HCP as assigned by the single-criterion CART classifiers for MIC and %Hem, respectively

(ii) Credibility encoding of concurrent probabilities: Once the previous step is performed, one finds that two probability measures derived from independent approaches, $P_{\mathrm{MIC+\%Hem}}$ and $PP_{D_i\_Class\_1}$, encode similar information: the probability of a Cβ-HCP to exhibit a desirable balance between antibacterial and hemolytic properties. It is reasonable to assume that the higher is the similarity between the $P_{\mathrm{MIC+\%Hem}}$ and $PP_{D_i\_Class\_1}$ values, the higher should be their reliability and vice versa. Clearly, the results will depend on the prediction performance of each classifier. In addition, the degree of uncertainty of classifiers with different sets of molecular descriptors will be diverse. Consequently, a framework is required to allow the fusion of results from different approaches in order to access the reliability of predictions from several approaches with different degrees of uncertainty. Here, we selected the Dempster−Shafer theory[88,89] (also known as belief theory) to achieve that goal.[90] Belief theory is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence.[90] For our particular case, we have used the rule for concurrent testimony.[88,89]
The rule in its original context says that if a report is concurrently attested to by n reporters, each with credibility p, then the credibility of the report is $1 - (1 - p)^n$; where $0 \leq p \leq 1$. Thus, the credibility of a report is strengthened by the concurrence of reporters.[88,89] If we make a simple analogy of this situation to our problem, one may notice that the belief theory, specifically the Hospers's rule for combining concurrent evidence,[88,89] is fully applicable. One needs only to use $P_{\mathrm{MIC+\%Hem}}$ and $PP_{D_i\_Class\_1}$ as the respective values of "credibility p" for each classification approach and to replace "report" with "prediction" and "reporter" with "classification approach", showing in turn that the previous paragraph will almost literally describe our problem. On the other hand, developing a probability assignment is the basic function in belief theory and is an expression of the level of confidence that can be ascribed to a particular measurement. So, $P_{\mathrm{MIC+\%Hem}}$ and $PP_{D_i\_Class\_1}$ can be used to encode the reliability of the probability estimated for a Cβ-HCP to exhibit a desirable balance between antibacterial and hemolytic properties by means of two independent but complementary classification approaches ($B_{D_i\mathrm{andMIC+\%Hem}}$):

$$B_{D_i\mathrm{andMIC+\%Hem}} = 1 - (1 - PP_{D_i class_1}) \times (1 - P_{\mathrm{MIC+\%Hem}}) \quad (7)$$

(iii) Desirability−credibility−similarity scoring: As detailed above, $B_{D_i\mathrm{andMIC+\%Hem}}$ encodes information on the desirability (selective antibacterial profile) of the Cβ-HCP as well as the reliability of such desirability derived from two independent but complementary classification approaches ($PP_{D_i\_Class\_1}$ from multicriteria CART classifier for D_i and, joint $P_{\mathrm{MIC+\%Hem}}$ from the respective single-criterion CART classifiers for MIC and %Hem). Finally, as previously detailed, the use of the Tanimoto similarity coefficient is used as a weighting factor favoring cases structurally close to the query (117_Pro14 × 14A) and conferring to $B_{D_i\mathrm{andMIC+\%Hem}}$, the required

dx.doi.org/10.1021/ci2002186 |J. Chem. Inf. Model. 2011, 51, 3060–3077

variability of a ranking criterion. Like this, it is possible to obtain a quantitative feature encoding information related to the desirability, the degree of credibility ascribed to this desirability, and the similarity of a Cβ-HCP to a highly desirable Cβ-HCP query, which can be used as ranking criterion in a virtual screening strategy, the desirability−credibility−similarity (DCS) score:

$$\begin{aligned}\text{DSC score} &= \tan \times (1 - (1 - \text{PP}_{D_i\text{Class}_1}) \\ &\times (1 - \sqrt[2]{\text{PP}_{\text{MIC}_{\text{Class}_1}} \times \text{PPP}_{\%\text{Hem}_{\text{Class}_1}}})) \\ &= \tan \times B_{D_i\text{andMIC} + \%\text{Hem}}\end{aligned} \quad (8)$$

Assessing the Enrichment Ability of the Multicriteria Virtual Screening Strategy. The main goal of a virtual screening campaign is to filter the fragment of the library containing the most promising candidates to propose those for synthesis and biological assessment. With this regard, we decided to test the ability of the multicriteria virtual screening strategy to prioritize selective antibacterial Cβ-HCPs ($D_i \geq 0.6$) dispersed in a data set of biologically inactive or nonselective antibacterial Cβ-HCPs. Toward that goal, the respective training, validation, and overall sets of Cβ-HCPs were decreasingly ranked according to DCS score, and the enrichment ability of this strategy was finally assessed according to the enrichment metrics previously detailed and now depicted in Table 6. The ranked list of the overall set of Cβ-HCPs based on the DCS score can be consulted in the Supporting Information (see Table S5). Nevertheless, a graphic picture of the DCS score-based ranking is shown in Figure 5.

The positive and similar performance according to the enrichment metrics applied to training and validation sets supports the consistency of assessing the enrichment performance of our multicriteria virtual screening strategy based on the overall set of Cβ-HCPs. The respective values of AUAC and ROC metric obtained suggest that the method is able to rank a selective antibacterial Cβ-HCP earlier than a biologically inactive or nonselective antibacterial Cβ-HCP with a probability around 0.9. At the same time, $\text{TP/FP}_{\text{ROC-OP}}$ informs that to obtain the best performance, it is necessary to filter 8.1% of the overall set, in turn leading to find 31.3% of the TP cases at a minimal cost of less than 1% of FP cases, which represents a maximum enrichment factor of 4.25. Furthermore, although all the selective antibacterial Cβ-HCPs can be found only after filtering 89% of the overall set, 75% are found after filtering only the first 26.5% of compounds. On the other hand, 85.7% of the Cβ-HCPs retrieved after filtering the top 5% of the library were selective antibacterial Cβ-HCPs ($Y_{a5\%} = 0.857$), which represents an enrichment factor of 3.643, being 4.25 the maximum possible value of enrichment factor for this data fraction. The respective ROC, accumulation, and enrichment curves can be checked in the Figure 6. So, considering the previous results, one may well expect that large (real or virtual) libraries of Cβ-HCPs could be efficiently filtered by using the multicriteria virtual screening strategy proposed in this work.

CONCLUSIONS

In this paper, a practical approach that permits a simultaneous analysis of two conflicting biological activities of antimicrobial peptides was introduced. A simple, predictive, interpretable, and biophysically relevant multicriteria classification rule was derived with the aid of the convergence of machine learning techniques and desirability theory. The multicriteria classification rule also showed a high degree of concordance with predictions made by independent CART classifiers for complementary activities (MIC and %Hem), which further supports the reliability of its predictions. Moreover, the sequential application of the multicriteria classification rule evidenced its consistency with the general mode of action of cationic peptides in agreement with previous experimental and theoretical findings on the domain of antimicrobial peptides, clearly indicating in turn their biophysical relevance. Additionally, the combined use of the multicriteria classification rule, single-criterion CART classifiers for MIC and %Hem, desirability and belief theories, and similarity search allowed the implementation of a multicriteria virtual screening strategy showing promising results for the prioritization of potent and nonhemolytic antibacterial Cβ-HCPs. Finally, the multicriteria tool depicted in this work has proved to be a promising chemoinformatic solution to the serious problem of the hemolytic side effect of potent cationic antimicrobial peptides, which hampers them from entering into the drug development pipeline. Since user friendly and publicly accessible web servers represent the future direction for developing practically more useful models, simulated methods, or predictors,[14] we shall make efforts in our future work to provide a web server for the method presented in this paper.

ASSOCIATED CONTENT

Supporting Information. Identification and aminoacids sequence of the full set of Cβ-HCPs. Exact data of the 1724 molecular descriptors employed in the modeling stage. Tree graph for the multicriteria classification tree derived for overall desirability class assigment. Names and definitions of variables included on the respective CART classifiers. Ranked list of the overall set of Cβ-HCPs based on the DCS score. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: maikelcm@uclv.edu.cu or gmailkelcm@yahoo.es; ncordeir@fc.up.pt.

ACKNOWLEDGMENT

The authors acknowledge the University of Porto and Santander Bank for financial support (project PP06-2010) and the Portuguese Fundação para a Ciência e a Tecnologia (FCT). The three reviewers are sincerely acknowledged by their useful comments and suggestions.

REFERENCES

(1) Rossolini, G. M.; Mantengoli, E. Antimicrobial Resistance in Europe and Its Potential Impact on Empirical Therapy. *Clin. Microbiol. Infect.* **2008**, *14*, 2–8.

(2) Goossens, H.; Grabein, B. Prevalence and Antimicrobial Susceptibility Data for Extended-Spectrum Betalactamase and Ampc-Producing Enterobacteriaceae from the Mystic Program in Europe and the United States (1997−2004). *Diagn. Microbiol. Infect. Dis.* **2005**, *53*, 257–264.

(3) Pitout, J. D.; Nordmann, P.; Laupland, K. B.; Poirel, L. Emergence of Enterobacteriaceae Producing Extended-Spectrum Beta-Lactamases (Esbls) in the Community. *J. Antimicrob. Chemother.* **2005**, *56*, 52–59.

(4) Hancock, R. E.; Sahl, H. G. Antimicrobial and Host-Defense Peptides as New Anti-Infective Therapeutic Strategies. *Nat. Biotechnol.* **2006**, *24*, 1551–1557.
(5) Toke, O. Antimicrobial Peptides: New Candidates in the Fight against Bacterial Infections. *Biopolymers* **2005**, *80*, 717–735.
(6) Simmaco, M.; Mignogna, G.; Barra, D. Antimicrobial Peptides from Amphibian Skin: What Do They Tell Us? *Biopolymers* **1998**, *47*, 435–450
(7) Bulet, P.; Stocklin, R.; Menin, L. Anti-Microbial Peptides: From Invertebrates to Vertebrates. *Immunol. Rev.* **2004**, *198*, 169–184.
(8) Robinson, J. A.; Shankaramma, S. C.; Jetter, P.; Kienzl, U.; Schwendener, R. A.; Vrijbloed, J. W.; Obrecht, D. Properties and Structure-Activity Studies of Cyclic Beta-Hairpin Peptidomimetics Based on the Cationic Antimicrobial Peptide Protegrin I. *Bioorg. Med. Chem.* **2005**, *13*, 2055–2064.
(9) Marr, A. K.; Gooderham, W. J.; Hancock, R. E. Antibacterial Peptides for Therapeutic Use: Obstacles and Realistic Outlook. *Curr. Opin. Pharmacol.* **2006**, *6*, 468–472.
(10) Kondejewski, L. H.; Jelokhani-Niaraki, M.; Farmer, S. W.; Lix, B.; Kay, C. M.; Sykes, B. D.; Hancock, R. E.; Hodges, R. S. Dissociation of Antimicrobial and Hemolytic Activities in Cyclic Peptide Diastereomers by Systematic Alterations in Amphipathicity. *J. Biol. Chem.* **1999**, *274*, 13181–13192.
(11) Jenssen, H. Descriptors for Antimicrobial Peptides. *Expert Opin. Drug Discovery* **2011**, *6*, 171–184.
(12) Fjell, C. D.; Hancock, R. E. W.; Jenssen, H. Computer-Aided Design of Antimicrobial Peptides. *Curr. Pharm. Anal.* **2010**, *6*, 66–75.
(13) Wang, P.; Hu, L.; Liu, G.; Jiang, N.; Chen, X.; Xu, J.; Zheng, W.; Li, L.; Tan, M.; Chen, Z.; Song, H.; Cai, Y. D.; Chou, K. C. Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods. *PLoS One* **2011**, *6*, e18476.
(14) Chou, K. C.; Shen, H. B. Review: Recent Advances in Developing Web-Servers for Predicting Protein Attributes. *Nat. Sci.* **2009**, *2*, 63–92.
(15) Frecer, V. Qsar Analysis of Antimicrobial and Haemolytic Effects of Cyclic Cationic Antimicrobial Peptides Derived from Protegrin-1. *Bioorg. Med. Chem.* **2006**, *14*, 6065–6074.
(16) Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M. N. D. S.; Cagide Fajin, J. L.; Morell, C.; Molina Ruiz, R.; Cañizares-Carmenate, Y.; Rosa Dominguez, E. Desirability-Based Methods of Multiobjective Optimization and Ranking for Global Qsar Studies. Filtering Safe and Potent Drug Candidates from Combinatorial Libraries. *J. Comb. Chem.* **2008**, *10*, 897–913.
(17) Nicolotti, O.; Giangreco, I.; Miscioscia, T. F.; Carotti, A. Improving Quantitative Structure-Activity Relationships through Multiobjective Optimization. *J. Chem. Inf. Model.* **2009**, *49*, 2290–2302.
(18) Schito, G. C. The Importance of the Development of Antibiotic Resistance in Staphylococcus Aureus. *Clin. Microbiol. Infect.* **2006**, *12*, 3–8.
(19) Falagas, M. E.; Bliziotis, I. A.; Kasiakou, S. K.; Samonis, G.; Athanassopoulou, P.; Michalopoulos, A. Outcome of Infections Due to Pandrugresistant (Pdr) Gram-Negative Bacteria. *BMC Infect. Dis.* **2005**, *5*, 24.
(20) Paterson, D. L.; Bonomo, R. A.; Extended-Spectrum Beta-Lactamases, A Clinical Update. *Clin. Microbiol. Rev.* **2005**, *18*, 657–686.
(21) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.
(22) Derringer, G.; Suich, R. Simultaneous Optimization of Several Response Variables. *J. Quality Technol.* **1980**, *12*, 214–219.
(23) Chou, K. C. Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236–247.
(24) Chou, K. C.; Wu, Z. C.; Xiao, X.; Iloc-Euk, A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. *PLoS One* **2011**, *6*, e18258.
(25) Chou, K. C.; Shen, H. B. Review: Recent Progresses in Protein Subcellular Location Prediction. *Anal. Biochem.* **2007**, *370*, 1–16.
(26) Harrington, E. C. The Desirability Function. *Ind. Quality Control* **1965**, *21*, 494–498.
(27) Outinen, K.; Haario, H.; Vuorela, P.; Nyman, M.; Ukkonen, E.; Vuorela, H. Optimization of Selectivity in High-Performance Liquid Chromatography Using Desirability Functions and Mixture Designs According to Prisma. *Eur. J. Pharm. Sci.* **1998**, *6*, 197–205.
(28) Shih, M.; Gennings, C.; Chinchilli, V. M.; Carter, W. H., Jr. Titrating and Evaluating Multi-Drug Regimens within Subjects. *Stat. Med.* **2003**, *22*, 2257–2279.
(29) Kording, K. P.; Fukunaga, I.; Howard, I. S.; Ingram, J. N.; Wolpert, D. M. A Neuroeconomics Approach to Inferring Utility Functions in Sensorimotor Control. *PLoS Biol.* **2004**, *2*, e330.
(30) Cojocaru, C.; Khayet, M.; Zakrzewska-Trznadel, G.; Jaworska, A. Modeling and Multi-Response Optimization of Pervaporation of Organic Aqueous Solutions Using Desirability Function Approach. *J. Hazard. Mater.* **2009**, *167*, 52–63.
(31) Jancic-Stojanovic, B.; Malenovic, A.; Ivanovic, D.; Rakic, T.; Medenica, M. Chemometrical Evaluation of Ropinirole and Its Impurity's Chromatographic Behavior. *J. Chromatogr., A* **2009**, *1216*, 1263–1269.
(32) Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M. N. Desirability-Based Multiobjective Optimization for Global Qsar Studies: Application to the Design of Novel Nsaids with Improved Analgesic, Antiinflammatory, and Ulcerogenic Profiles. *J. Comput. Chem.* **2008**, *29*, 2445–2459.
(33) Ekins, S.; Honeycutt, J. D.; Metz, J. T. Evolving Molecules Using Multi-Objective Optimization: Applying to Adme/Tox. *Drug Discovery Today* **2010**, *15*, 451–460.
(34) Cruz-Monteagudo, M.; The, H. P.; Cordeiro, M. N. D. S.; Borges, F. Prioritizing Hits with Appropriate Trade-Offs between Hiv-1 Reverse Transcriptase Inhibitory Efficacy and Mt4 Blood Cells Toxicity through Desirability-Based Multi-Objective Optimization and Ranking. *Mol. Inf.* **2010**, *29*, 303–321.
(35) Machado, A.; Tejera, E.; Cruz-Monteagudo, M.; Rebelo, I. Application of Desirability-Based Multi(Bi)-Objective Optimization in the Design of Selective Arylpiperazine Derivates for the 5-Ht1a Serotonin Receptor. *Eur. J. Med. Chem.* **2009**, *44*, 5045–5054.
(36) Cruz-Monteagudo, M.; Cordeiro, M. N. D. S.; Teijeira, M.; González, M. P.; Borges, F. Multidimensional Drug Design: Simultaneous Analysis of Binding and Relative Efficacy Profiles of N6-Substituted-4́-Thioadenosines A3 Adenosine Receptor Agonists. *Chem. Biol. Drug. Des.* **2010**, *75*, 607–618.
(37) Manoharan, P.; Vijayan, R. S. K.; Ghoshal, N. Rationalizing Fragment Based Drug Discovery for Bace1: Insights from Fb-Qsar, Fb-Qssr, Multi Objective (Mo-Qspr) and Mif Studies. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 843–864.
(38) Tropsha, A. Best Practices for Qsar Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.
(39) *Dragon*, version 6.0; (Software for Molecular Descriptor Calculation); Talete srl: Milano, Italy, 2010.
(40) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH Verlag GmbH & Co: Weinheim, Germany, 2009; Vol. 1&2.
(41) Findlay, B.; Zhanel, G. G.; Schweizer, F. Cationic Amphiphiles, a New Generation of Antimicrobials Inspired by the Natural Antimicrobial Peptide Scaffold. *Antimicrob. Agents Chemother.* **2010**, *54*, 4049–4058.
(42) Nair, C. M.; Vijayan, M.; Venkatachalapathi, Y. V.; Balaram, P. X-Ray Crystal Structure of Pivaloyl-D-Pro-L-Pro-L-Ala-N-Methylamide; Observation of a Consecutive B-Turn Conformation. *J. Chem. Soc., Chem. Commun.* **1979**, 1183–1184.
(43) Bean, J. W.; Kopple, K. D.; Peishoff, C. E. Conformational Analysis of Cyclic Hexapeptides Containing the D-Pro-L-Pro Sequence to Fix B-Turn Positions. *J. Am. Chem. Soc.* **1992**, *114*, 5328–5334.
(44) Chalmers, D. K.; Marshall, G. R. Pro-D-Nme-Amino Acid and D-Pro-Nme-Amino Acid: Simple, Efficient Reverse Turn Constraints. *J. Am. Chem. Soc.* **1995**, *117*, 5927–5937.
(45) Tropsha, A. Integrated Chemo- and Bioinformatics Approaches to Virtual Screening. In *Chemoinformatics Approaches to Virtual Screening*;

Varnek, A.; Tropsha, A., Eds. Royal Society of Chemistry: Cambridge, U.K., 2008; pp 295−325.

(46) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.

(47) Kawashima, S.; Ogata, H.; Kanehisa, M. Aaindex: Amino Acid Index Database. *Nucleic Acids Res.* **1999**, *27*, 368–369.

(48) Venkatarajan, M. S.; Braun, W. New Quantitative Descriptors of Amino Acids Based on Multidimensional Scaling of a Large Number of Physical-Chemical Properties. *J. Mol. Model.* **2001**, *7*, 445–453.

(49) Ivanciuc, O.; Midoro-Horiuti, T.; Schein, C. H.; Xie, L.; Hillman, G. R.; Goldblum, R. M.; W., B. The Property Distance Index Pd Predicts Peptides That Cross-React with Ige Antibodies. *Mol. Immunol.* **2009**, *46*, 873–883.

(50) Ivanciuc, O.; Schein, C. H.; Braun, W. Sdap: Database and Computational Tools for Allergenic Proteins. *Nucleic Acids Res.* **2003**, *31*, 359–362.

(51) Ivanciuc, O.; Schein, C. H.; Braun, W. Data Mining of Sequences and 3d Structures of Allergenic Proteins. *Bioinformatics* **2002**, *18*, 1358–1364.

(52) Ivanciuc, O. Machine Learning Quantitative Structure-Activity Relationships (Qsar) for Peptides Binding to the Human Amphiphysin-1 Sh3 Domain. *Curr. Proteomics* **2009**, *6*, 289–302.

(53) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of Automatic Relevance Determination in Qsar Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423–1430.

(54) *Statistica*, version 8.0; (Data analysis software system); StatSoft Inc.: Tulsa, OK, 2007.

(55) Kubinyi, H. Virtual Screening - the Road to Success. In Proceedings of the XIX International Symposium on Medicinal Chemistry, Istanbul, Turkey, August 29−September 2, 2006, 2006; http://kubinyi.de/istanbul-09-06.pdf (accessed April 12, 2011).

(56) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*. Wadsworth: Monterey, CA, 1984.

(57) Chou, K. C.; Zhang, C. T. Review: Prediction of Protein Structural Classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.

(58) Hayat, M.; Khan, A. Predicting Membrane Protein Types by Fusing Composite Protein Sequence Features into Pseudo Amino Acid Composition. *J. Theor. Biol.* **2011**, *271*, 10–17.

(59) Chou, K. C. Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition. *Proteins* **2001**, *43*, 246–255.

(60) Zhou, X. B.; Chen, C.; Li, Z. C.; Zou, X. Y. Using Chou's Amphiphilic Pseudo-Amino Acid Composition and Support Vector Machine for Prediction of Enzyme Subfamily Classes. *J. Theor. Biol.* **2007**, *248*, 546–551.

(61) Kandaswamy, K. K.; Chou, K. C.; Martinetz, T.; Moller, S.; Suganthan, P. N.; Sridharan, S.; Pugalenthi, G.; Afp-Pred, A Random Forest Approach for Predicting Antifreeze Proteins from Sequence-Derived Properties. *J. Theor. Biol.* **2011**, *270*, 56–62.

(62) Zakeri, P.; Moshiri, B.; Sadeghi, M. Prediction of Protein Submitochondria Locations Based on Data Fusion of Various Features of Sequences. *J. Theor. Biol.* **2011**, *269*, 208–216.

(63) Mohabatkar, H. Prediction of Cyclin Proteins Using Chou's Pseudo Amino Acid Composition. *Protein Pept. Lett.* **2010**, *17*, 1207–1214.

(64) J., G.; Rao, N.; Liu, G.; Yang, Y.; Wang, G. Predicting Protein Folding Rates Using the Concept of Chou's Pseudo Amino Acid Composition. *J. Comput. Chem.* **2011**, *32*, 1612–1617.

(65) Gu, Q.; Ding, Y. S.; Zhang, T. L. Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. *Protein Pept. Lett* **2010**, *17*, 559–567.

(66) Witten, I. H.; Frank, E. Chapter 5: Credibility: Evaluating What's Been Learned. In *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Gray, J., Ed.; Morgan Kaufman: San Francisco, CA, 2005; pp 143−186.

(67) Cannon, E. O.; Amini, A.; Bender, A.; Sternberg, M. J.; Muggleton, S. H.; Glen, R. C.; Mitchell, J. B. Support Vector Inductive Logic Programming Outperforms the Naive Bayes Classifier and Inductive Logic Programming for the Classification of Bioactive Chemical Compounds. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 269–280.

(68) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(69) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(70) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the Performance of 3d Virtual Screening Protocols: Rmsd Comparisons, Enrichment Assessments, and Decoy Selection--What Can We Learn from Earlier Mistakes? *J Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.

(71) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary Qsar Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.

(72) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*. Wiley-VCH: Weinheim, Germany, 2009; Vol. 1, pp 608−612.

(73) Unger, S. H.; Hansch, C. On Model Building in Structure-Activity Relationships. A Reexamination of Adrenergic Blocking Activity of Beta-Halo-Beta-Arylalkylamines. *J. Med. Chem.* **1973**, *16*, 745–749.

(74) Huang, H. W. Action of Antimicrobial Peptides: Two-State Model. *Biochemistry* **2000**, *39*, 8347–8352.

(75) Tam, J. P.; Wu, C.; Yang, J. L. Membranolytic Selectivity of Cystine-Stabilized Cyclic Protegrins. *Eur. J. Biochem.* **2000**, *267*, 3289–3300.

(76) Strom, M. B.; Haug, B. E.; Skar, M. L.; Stensen, W.; Stiberg, T.; Svendsen, J. S. The Pharmacophore of Short Cationic Antibacterial Peptides. *J. Med. Chem.* **2003**, *46*, 1567–1570.

(77) Lejon, T.; Stiberg, T.; Strom, M. B.; Svendsen, J. S. Prediction of Antibiotic Activity and Synthesis of New Pentadecapeptides Based on Lactoferricins. *J. Pept. Sci.* **2004**, *10*, 329–335.

(78) Frecer, V.; Ho, B.; Ding, J. L. De Novo Design of Potent Antimicrobial Peptides. *Antimicrob. Agents Chemother.* **2004**, *48*, 3349–3357.

(79) Ostberg, N.; Kaznessis, Y. Protegrin Structure-Activity Relationships: Using Homology Models of Synthetic Sequences to Determine Structural Characteristics Important for Activity. *Peptides* **2005**, *26*, 197–206.

(80) Cherkasov, A.; Hilpert, K.; Jenssen, H.; Fjell, C. D.; Waldbrook, M.; Mullaly, S. C.; Volkmer, R.; Hancock, R. E. Use of Artificial Intelligence in the Design of Small Peptide Antibiotics Effective against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs. *ACS Chem. Biol.* **2009**, *4*, 65–74.

(81) Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Pante, N.; Hancock, R. E.; Cherkasov, A. Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. *J. Med. Chem.* **2009**, *52*, 2006–2015.

(82) Hancock, R. E. Host Defence (Cationic) Peptides: What Is Their Future Clinical Potential? *Drugs* **1999**, *57*, 469–473

(83) Hwang, P. M.; Vogel, H. J. Structure-Function Relationships of Antimicrobial Peptides. *Biochem. Cell. Biol.* **1998**, *76*, 235–246.

(84) Oren, Z.; Shai, Y. Mode of Action of Linear Amphipathic Alpha-Helical Antimicrobial Peptides. *Biopolymers* **1998**, *47*, 451–463.

(85) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

(86) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.

(87) Kochev, N.; Monev, V.; Bangov, I. Searching Chemical Structures. In *Chemoinformatics: A Textbook*, Gasteiger, J., Engel, T., Eds.; Wiley-VCH Verlag GmbH & Co. KgaA: Weinheim, Germany, 2003; pp 291−318.

(88) Hooper, G. A Calculation of the Credibility of Human Testimony. *Phil. Trans. Royal Soc.* **1699**, *21*, 359–365.
(89) Shafer, G. The Combination of Evidence. *Int. J. Intell. Syst.* **1986**, *1*, 155–179.
(90) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.