

Comparative Chemometric and QSAR/SAR Study of Structurally Unrelated Substrates of a MATE Efflux Pump VmrA from *V. parahaemolyticus*: Prediction of Multidrug Resistance

Rudolf Kiralj* and Márcia M. C. Ferreira

Laboratório de Quimiometria Teórica e Aplicada, Instituto de Química, Universidade Estadual de Campinas, Campinas, SP 13084-971, Brazil, E-mail: rudolf@iqm.unicamp.br

Keywords: Agent–receptor interactions, Molecular descriptors, Multidrug resistance, Regression models, Semiempirical methods

Received: December 9, 2006; Accepted: June 8, 2007

DOI: 10.1002/qsar.200630164

Abstract

Structures of 12 structurally unrelated agents (training set) and 19 diverse agents (prediction set) were modeled at the PM3 level and the molecular descriptors obtained were correlated with pMIC ($\text{pMIC} = -\log(\text{MIC}/\text{mol}/\text{dm}^3)$) values for *E. coli* strains KAM32 (without VmrA efflux pump) and KAM32/pVCJ6 (with VmrA). Partial Least Squares (PLS) and Principal Component Regression (PCR) models were built and used to predict pMICs. Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA) were performed for training and training + prediction sets. Agent–receptor interaction descriptors were calculated for complexes with these or the most similar agents, as retrieved from the Protein Data Bank (PDB). Good substrates of VmrA from *Vibrio parahaemolyticus* AQ3334 are nonlinear, rigid, and condensed heteroaromatics, rather hydrophobic, involved in aromatic–aromatic, hydrophobic–hydrophobic, hydrophobic–hydrophilic, and hydrogen bond contacts with VmrA. Qualitative and quantitative structure–activity relationships [(Q)SARs] predicted well the resistance/sensitivity of VmrA to 19 agents. This can have practical application in assays with *V. parahaemolyticus* and other noncholera *Vibrio* spp. Agent–receptor interaction descriptors correlate with experimental and predicted pMICs, giving a new insight into possible VmrA-mediated agent efflux mechanisms.

1 Introduction

Vibrio parahaemolyticus is a slightly halophilic bacterium that shares its habitat with other noncholera *Vibrio* species [1–4]. It occurs in most of the marine animals (mammals, fish, shellfish, crustaceans, and plankton) in coastal waters of North America, Mediterranean, Asia, and Pacific. Pathogenic strains of *V. parahaemolyticus* (Kanagawa phenomenon-positive) and other *Vibrio* spp. cause seafood poisoning (associated gastroenteritis), wound and soft tissue infections, septicemia, and other infections. Several thousands of people get infected by these bacteria worldwide, and 3000 persons in the US *per year* [2, 4]. Severe disease occurs in more than 10% of the infections in immunocompromised persons and may result in death (in the US: average of 7 deaths *per year*, 46 and 33 deaths in the 1997–1998 and 1999 outbreaks, respectively [2]). These morbidity and mortality statistics are a strong reason to investigate chemical means in combating noncho-

lera *Vibrio* infections, especially that of the most common species, *V. parahaemolyticus*.

Vibrio spp. need salt for growth: Na^+ electrochemical potential of across the cell membrane is one of the major driving forces for energy-dependent membrane processes [4, 5]. *V. parahaemolyticus* exhibits Multidrug Resistance (MDR) to antimicrobial agents [5] *via* multidrug efflux pumps such as NorM [6, 7] and VmrA [5]. The genome of *V. parahaemolyticus* strain RIMD2210633 [8, 9] contains VmrA and 11 other efflux pumps from the Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) superfamily of secondary transporters [10], one of the commonest super-

List of crystal structures retrieved from the CSD and PDB, additional supporting figures and chemometric analyses including Y-randomization tests. This material is available free of charge *via* the internet at <http://www3.interscience.wiley.com/>.

families in this bacterium. VmrA is a Na⁺/drug antiporter from the Multidrug And Toxic compound Extrusion (MATE) family [10–12]. It has no 3D structure determined yet. Its functional monomeric form consists of 447 residues in strain AQ3334 and 448 residues in strain RIMD221063/serotype O3:K6, and has 12 hydrophobic transmembrane segments embedded in the inner cell membrane.

VmrA and other members of the MATE family are effective against several structurally unrelated drugs, organic dyes, detergents, and xenobiotics [5, 10]. VmrA substrates share some common characteristics responsible for their efflux from the cytoplasm into periplasm in *V. parahaemolyticus* cells. This work intends to correlate the efflux activity of VmrA with molecular properties of 12 diverse agents (Figure 1) [5] at Quantitative Structure–Activity Relationship (QSAR) and Structure–Activity Relationship (SAR) levels. The prediction of the MDR character of VmrA with respect to 19 diverse substrates (Figure 1) is performed at both QSAR and SAR levels. This analysis can be of practical applicability in predicting whether VmrA is resistant or sensitive to a particular agent.

2 Methods

2.1 Biological Activities

To observe the efflux potency of the VmrA pump in *V. parahaemolyticus*, Chen *et al.* [5] have constructed two *E. coli* strains: KAM32 without the major efflux system AcrAB [13] and Na⁺/drug antiporter YdhE [6], and KAM32/pVCJ6 obtained as KAM32 transformant with introduced ethidium bromide-resistant plasmid pVCJ6 from *V. parahaemolyticus* AQ3334. They measured the Minimal Inhibitory Concentration (MIC) of 12 structurally unrelated compounds (Figure 1) as extruded by KAM32 (without VmrA) and KAM32/pVCJ6 (with VmrA): DAPI (4',6-diamino-2-phenylindole) (**1**), TPPCl (tetraphenylphosphonium chloride) (**2**), acriflavine chloride (**3**), ethidium bromide (**4**), chloramphenicol (**5**), norfloxacin (**6**), rhodamine 6G chloride (**7**), tetracycline (**8**), erythromycin (**9**), streptomycin (**10**), sodium deoxycholate (**11**), and sodium dodecyl sulfate (**12**). **1–12** make the training set in this work, and the corresponding activities $\text{pMIC} = -\log(\text{MIC}/\text{mol}/\text{dm}^3)$ are $\text{pMIC}(\text{KAM})$ and $\text{pMIC}(\text{pVCJ6})$ for KAM32 and KAM32/pVCJ6, respectively (Table 1).

2.2 Modeling Agent Structures

Agents (organic molecules/ions of **1–12**) were modeled by Titan [14] from structures for these or similar agents (Table 1) from the Cambridge Structural Database (CSD) [15, 16] retrieved by ConQuest v. 1.7 [17]. The ionic states at neutral pH were considered as in Figure 1.

The prediction set (Figure 1) was constructed in order to verify QSAR relationships with respect to agents known to be effective against *V. parahaemolyticus*, and to predict MDR character of VmrA with respect to other different agents that could be effective or inefficient against this pump. Drugs (**13–22**) for the treatment of severe *V. parahaemolyticus* infections [2, 4] are tetracyclines, ampicillin, ciprofloxacin, doxycycline, cefotaxime, piperacillin + tazobactam, and ticarcillin + clavulanate. *V. parahaemolyticus* also shows high sensitivity [3] to butylated hydroxyanisoles (**23**) and sorbic acid (**24**). The prediction set was extended to nucleic acid intercalators to which VmrA might be resistant (**25–27**) as MDAPY (4-methyl-4,9-diazapyrenium hydrogen sulfate) [18], proflavine and ellipticine, to organic dyes (**28, 29**), and an amino acid (**30**). Table 2 contains detailed information about biological activities of these agents. Modeling of molecular/ionic species at neutral pH was performed in the same way as for **1–12**, using experimental structures from the CSD (Table 2).

The geometries of modeled neutral/ionic species (Figure 1) were optimized at the PM3 semiempirical level in Titan. Several molecular descriptors were calculated by Titan, MOPAC 6 [23] and Chem3D [24], and from 2D chemical formulas. Hydrogen-depleted bonds in the molecules CH₃–CH₃, CH₃–NH₂, CH₃–PH₂, HS–OH, CH₃–F, CH₃–OH, H₂N–OH, CH₃–Cl, CH₃–SH, and H₂N–NH₂ were modeled by Titan at PM3 level to obtain parameters for standard bond length descriptors for **1–30**.

2.3 Chemometrics and QSAR

Molecular descriptor data were autoscaled prior to chemometric analysis. Cut-off in descriptor–pMIC correlation coefficients was 0.50. Partial Least Squares (PLS) and Principal Component Regression (PCR) models [25–28] for pMIC(pVCJ6) and pMIC(KAM) for the training set **1–12** were built by manual variable selection and validated by leave-one-out crossvalidation. The robustness of the models was tested by performing Y-randomization according to Wold and Eriksson [25]. Then biological activities pMICs were calculated for the prediction set **13–30**.

Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) [25–28] were performed on the autoscaled data previously used in the regression modeling of pMIC(pVCJ6) and pMIC(KAM). All chemometric analyses were made using Matlab [29] and Pirouette [30].

In order to predict whether VmrA is resistant or sensitive to **13–30**, the QSAR approach was used [the absolute difference between predicted pMIC(pVCJ6) and pMIC(KAM)] as well as two SAR approaches. One SAR consisted of PCA–HCA analysis of the two descriptor data matrices for the training + prediction set, already used in QSAR. Comparison of the clustering of **1–12** and of **1–30** could point out if VmrA would be resistant or sensitive to **13–30**. The same idea was used in another SAR

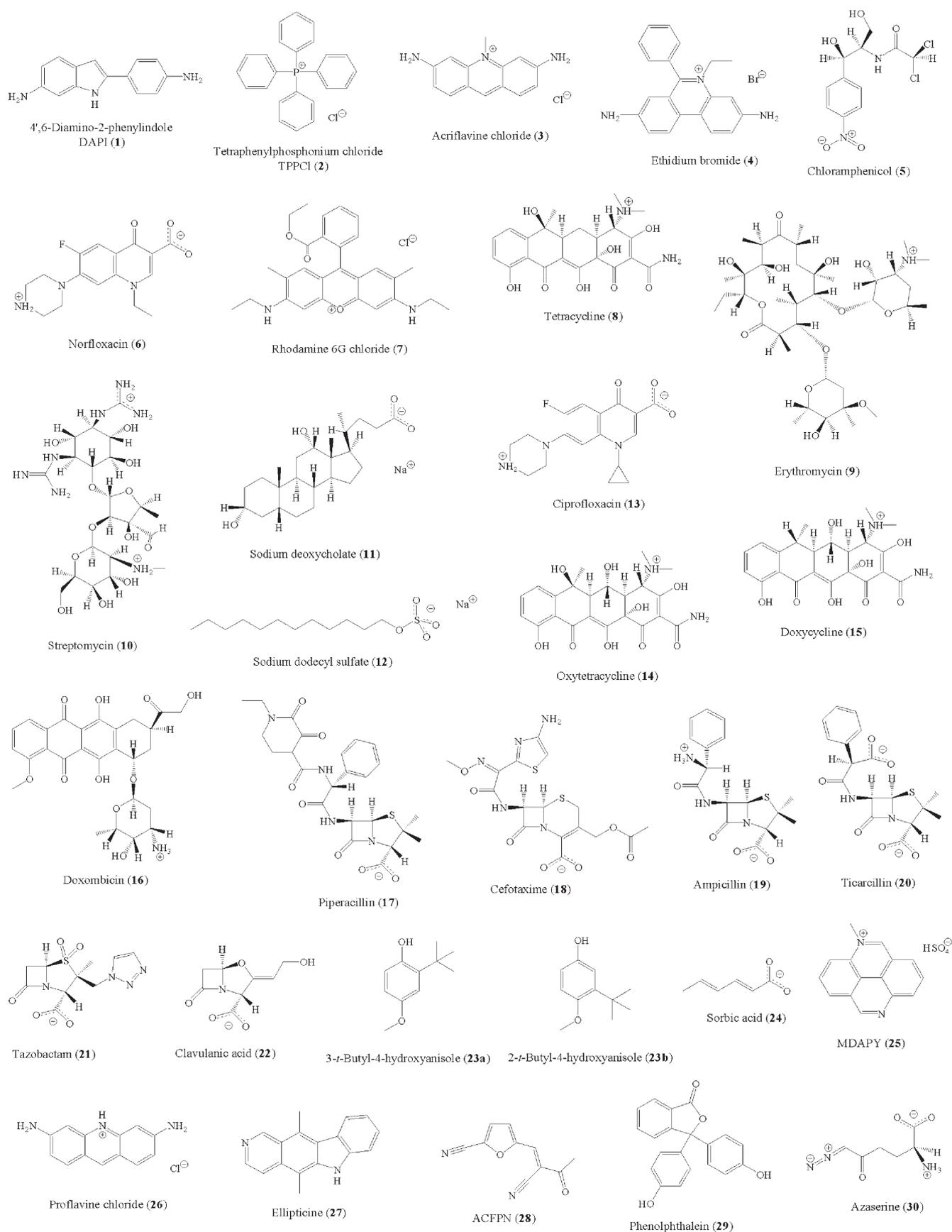


Figure 1. Structures of the agents from the training (1–12) and prediction (13–30) sets.

Table 1. Experimental data for agents **1–12**.

No.	CSD source ^a	Formula ^b	pMIC(pVCJ6) ^c	pMIC(KAM) ^d	pMICΔ ^e
1	ASUMEG	C ₁₄ H ₁₃ N ₃	3.242	5.961	2.719
2	DURDAV01	[C ₂₄ H ₂₀ P] ⁺	3.468	4.672	1.204
3	TIJZOB	[C ₁₄ H ₁₄ N ₃] ⁺	3.909	5.114	1.205
4	ETHIDB	[C ₂₁ H ₂₀ N ₃] ⁺	4.392	4.994	0.602
5	CLMPCL02	[C ₁₁ H ₁₂ Cl ₂ N ₂ O ₃] [±]	5.810	5.810	0
6	XAYGEJ	C ₁₆ H ₁₈ FN ₃ O ₃	7.027	7.027	0
7	QIMMEE	[C ₂₈ H ₃₁ N ₂ O ₃] ⁺	4.777	4.777	0
8	TETCYH10	[C ₂₂ H ₂₅ N ₂ O ₈] ⁺	5.949	5.949	0
9	NAVTEJ	[C ₃₇ H ₆₈ O ₁₃ N] ⁺	5.264	5.264	0
10	STOSEH10	[C ₂₁ H ₄₁ O ₁₂ N ₇] ⁺	5.464	5.464	0
11	GOLWIN	[C ₂₄ H ₃₉ O ₄] ⁻	5.318	5.318	0
12	SATLUU	[C ₁₂ H ₂₅ O ₄ S] ⁻	6.461	6.461	0

^a CSD codes for the structures retrieved from the CSD database. Complete or partial structures were used in molecular modeling.

^b Formula for the agent organic component in neutral, cationic/protonated (+), anionic (-), or zwitterionic (±) state as applied in molecular modeling.

^c Efflux activity pMIC of the *E. coli* strain KAM32/pVCJ6.

^d Efflux activity pMIC of the *E. coli* strain KAM32.

^e Absolute difference between the two efflux activities.

Table 2. Experimental data for agents **13–30** and their known or potential bioactive profiles in relation to *V. parahaemolyticus*.

No.	CSD source ^a	Formula ^b	Compound-type	Relevant biological activity
13	UJAGUH	[C ₁₇ H ₁₈ FN ₃ O ₃] [±]	Fluoroquinolone antibiotic	Potential activity against <i>V. parahaemolyticus</i> ^c
14	OXYTET	[C ₂₂ H ₂₄ N ₂ O ₉] ⁺	Tetracycline antibiotic	Potential activity against <i>V. parahaemolyticus</i> ^d
15	OXYTET	[C ₂₂ H ₂₄ N ₂ O ₈] ⁺	Tetracycline antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections [4]
16	CARMYC11	[C ₂₇ H ₃₀ NO ₁₀] ⁺	Tetracycline antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections [4]
17	PIPCIL	[C ₂₄ H ₂₇ N ₄ O ₇ S] ⁻	β-Lactam antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections in combination with 21 [4]
18	BODKOU	[C ₁₆ H ₁₅ N ₅ O ₇ S ₂] ⁻	β-Lactam antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections [4]
19	AMCILL	[C ₁₆ H ₁₉ N ₃ O ₄ S] ⁻	β-Lactam antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections [2]
20	BEHWIU	[C ₁₇ H ₁₆ N ₂ O ₆ S] ⁻	β-Lactam antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections in combination with 22 [4]
21	KOFNIC	[C ₁₀ H ₁₁ N ₄ O ₅ S] ⁻	Bactam antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections in combination with 17 [4]
22	CLAVBB10	[C ₈ H ₈ NO ₅] ⁻	Bactam-like antibiotic	Treatments of severe <i>V. parahaemolyticus</i> infections in combination with 20 [4]
23a,b	ESALUF01	C ₁₁ H ₁₆ O ₂	Disubstituted phenol, a preservative	Inhibitor of <i>V. parahaemolyticus</i> growth [3]
24	LEZHUT	[C ₆ H ₇ O ₂] ⁻	Dienoic carboxylic acid, a preservative	Inhibitor of <i>V. parahaemolyticus</i> growth [3, 19]
25	SUWHOHO	[C ₁₅ H ₁₁ N ₂] ⁺	Diaza derivative of polycyclic aromatic hydrocarbon	Potential DNA intercalator [18, 20], probably not effective against <i>V. parahaemolyticus</i> ^e
26	PROFLS	[C ₁₃ H ₁₂ N ₃] ⁺	Acridine antibiotic DNA intercalator)	Probably not effective against <i>V. parahaemolyticus</i> ^f
27	ELLIPT	C ₁₇ H ₁₃ N ₂	Antibiotic (DNA intercalator)	Probably not effective against <i>V. parahaemolyticus</i> ^g
28	RUWCAN	C ₁₀ H ₁₆ N ₂ O ₂	Potential organic dye [21]	Perhaps effective against <i>V. parahaemolyticus</i>
29	NIMDAO	C ₂₀ H ₁₄ O ₄	Organic dye	Probably not effective against <i>V. parahaemolyticus</i> ^h
30	AZASER11	[C ₆ H ₉ N ₃ O ₃] [±]	Modified amino acid	Antibiotic and teratogenic activity [22]

^a CSD codes for the structures retrieved from the CSD database. Complete or partial structures were used in molecular modeling.

^b Formula for the agent organic component in neutral, cationic/protonated (+), anionic (-), or zwitterionic (±) state as applied in molecular modeling.

^c Ciprofloxacin **13** is used in treatments of severe *V. parahaemolyticus* infections [2, 4].

^d Tetracycline **14** [2], doxycycline **15**, and other tetracyclines [4] are used in treatments of severe *V. parahaemolyticus* infections.

^e *V. parahaemolyticus* is probably resistant to this agent, due to its molecular similarity with **3** and **4** which are not effective against this bacterium [5].

^f *V. parahaemolyticus* is probably resistant to this agent, due to its molecular similarity with **2** and **4** which are not effective against this bacterium [5].

^g *V. parahaemolyticus* is probably resistant to this agent, due to its molecular similarity with **1** and **3** which are not effective against this bacterium [5].

^h *V. parahaemolyticus* is probably resistant to this agent, due to its molecular similarity with **2** and **4** which are not effective against this bacterium [5].

approach, where the data matrix consisted of descriptors that showed high discrimination between agents to which

VmrA was resistant and sensitive (discriminatory descriptors).

2.4 Agent–Receptor Interaction Statistics

Search for **1–30** or similar agents in the Protein Data Bank (PDB; [31]) was carried out in order to rationalize QSAR results at the molecular level. Most retrieved complexes had no agent–receptor positional disorder and could be treated by the online Ligand–Protein Contacts (LPCs) software [32, 33] to obtain several agent–receptor interaction descriptors. The receptors were of diverse nature (Table 3), ranging from MDR efflux pumps and transcriptional repressors to β -lactamases and DNA oligomers, from diverse biological kingdoms (bacteria and archaea, animals including human, fungi/metazoa) and synthetic constructs. The descriptors were the total number of agent–environment contacts; number and corresponding agent surface areas involved in specific interaction or contact types (hydrogen bonds and hydrophobic, aromatic–aromatic, hydrophilic–hydrophobic, acceptor–acceptor, receptor–receptor, and other contacts); the number, mean distance, and corresponding agent surface areas of agent–O, N, C (receptor) contacts. All molecules/ions in interaction with a studied molecule were considered as the receptor. The descriptors for **1–12** were correlated with experimental pMIC(KAM) and pMIC(pVCJ6), and those for **1–30** with PLS predicted pMIC(KAM), pMIC(pVCJ6), and $\text{pMIC}\Delta = |\text{pMIC}(\text{pVCJ6}) - \text{pMIC}(\text{KAM})|$.

3 Results and Discussion

3.1 Molecular Descriptors and Correlation Analysis

A total of 120 molecular descriptors for **1–12** were calculated: steric, electronic, hydrophobicity, hydrogen bonding, topological, compositional, and complex descriptors (some as rational functions of one or two simple descriptors). Considerable correlations of 57 descriptors with the activities (abs. corr. coefficients >0.500 , bold in Table 4) were found: 26 with pMIC(KAM), 46 with pMIC(pVCJ6), and 15 with both pMICs. The absolute difference between the biological activities pMIC Δ (Table 1) is nonzero only for N, P-containing heteroaromatic species **1–4**. Chen *et al.* [5] considered that the VmrA-containing strain KAM32/pVCJ6 was resistant to **1–4**, while some intrinsic resistance to the other drugs came from the parent strain KAM32. Eight discriminatory descriptors clearly distinguish **1–4** from **5–12** via extremely low or high values for **1–4** (Figure A in Supplementary Material). Low values of bond length RD and number of polar atoms Np account for elevated resistance to **1–4**. High values of the other six descriptors show the same trend: a bond length descriptor RRD, the square Np², the surface density of ring atoms sigr, and the number fractions of the ring atoms wr, aromatic carbon atoms wa, and the nonhydrogen atoms in planar fragments wl. These descriptors show noticeable differences in correlation coefficients with the two activi-

ties, from 0.25 to 0.43. In other words, MDR of VmrA is directed against agents with shorter bonds, a few polar atoms, elevated contents of rings, planar structures, and aromatic carbon atoms. This is generally a rigid aromatic structure with exo- and/or endocyclic heteroatoms.

3.2 Regression Models

PLS and PCR regression models for pMIC(KAM) and pMIC(pVCJ6) included one principal component (PC) and five molecular descriptors (Tables 5–7). pMIC(pVCJ6) was modeled from these descriptors: FF (a dipole moment function), wr, sighyd (surface density of hydrophobic carbon atoms), Np², and Mrefn2 (a normalized molecular refractivity function). Only sighyd is common for both activities. Other descriptors in modeling pMIC(KAM) were: Nh (the number of hydrophobic carbon atoms), E4 (average polarizability), HOMO (energy of the HOMO orbital), and wh2 (square of the hydrophobic carbon fraction). There is no essential difference between the PLS and PCR statistics (Table 6) or between predicted pMICs (Table 7) for the two *E. coli* strains. Satisfactory prediction power statistics $Q^2 > 0.5$ and $R^2 > 0.6$ for QSAR models [34, 35] can be observed in Table 6 for all the models. The Y-randomization tests are represented by Q^2 versus R^2 plots in Figure B in Supplementary Material. The low R^2 and Q^2 values indicate that the good results in the original models are not due to a chance correlation or structural dependency of the training set.

It would be reasonable to expect that the major efflux system in KAM32/pVCJ6 is VmrA and that other efflux pumps, possibly not from the MATE family, are not responsible for the efflux of **1–12** from KAM32. This would justify the different efflux mechanisms coming from corresponding QSAR/chemometrics results that describe the efflux powers of the two strains. Consequently, SEV, SEP, and Δ are smaller for models for pMIC(KAM) than those for pMIC(pVCJ6). Particular deviations (Table 7) follow the same trend (Figure C in Supplementary Material): there is only one agent (**6**) with error over 10% for pMIC(KAM), but there are four such agents (**1**, **2**, **6**, and **9**) for pMIC(pVCJ6). Moreover, correlation coefficients Q , Q^2 , R , and R^2 are higher for pMIC(pVCJ6) than those for pMIC(KAM). PC1 for pMIC(pVCJ6) contains much more total variance than that for pMIC(KAM). A remarkable difference between the two strains is visible in the variable selection (Table 6) and descriptor–pMIC correlation coefficients R(KAM) and R(pVCJ6). All this reflects the absence and presence of VmrA in strains KAM32/pVCJ6 and KAM32, respectively. According to the regression vectors (Table 6), KAM32 extrudes agents **1–12** which are characterized by three lipophilicity descriptors (Nh, sighyd, and wh2) and two electronic descriptors (E4 and HOMO). Good substrates of the efflux systems in this strain (substrates with low pMIC) have high contents of hydrophobic atoms and they are less polarizable than poor

Table 3. Experimental agent–receptor structures retrieved from the PDB and treated by the LPC software

No.	Substrate ^a	Formula ^a	Receptor (organism)	PDB ^b	Mols ^c	
1 ^d	6-Amidine-2-(4-amidinophenyl)indole	C ₁₅ H ₁₅ N ₅	Double-stranded DNA dodecamer (synthetic construct)	1D30	1	
			Double-stranded DNA decamer (synthetic construct)	432D	1	
2	Tetraphenylphosphonium	[C ₂₄ H ₂₀ P] ⁺	Transcription activator Bmrr (<i>B. subtilis</i>)	2BOW	1	
3 ^d	Proflavine	[C ₁₃ H ₁₂ N ₃] ⁺	Transcriptional repressor QacR (<i>S. aureus</i>)	1QVU	1	
			Transcriptional repressor QacR (<i>S. aureus</i>)	1QVT	1	
4	Ethidium	[C ₂₁ H ₂₀ N ₃] ⁺	Transcriptional repressor QacR (<i>S. aureus</i>)	1JTY	1	
			Transcriptional repressor QacR (<i>S. aureus</i>)	1QVU	1	
5	Chloramphenicol	C ₁₁ H ₁₂ N ₂ O ₅ Cl ₂	Transmembrane efflux pump AcrB (<i>E. coli</i>)	1OY9	1	
			Chloramphenicol acetyltransferase type III (<i>E. coli</i>)	1CLA	1	
			Chloramphenicol acetyltransferase type III (<i>E. coli</i>)	3CLA	1	
			Chloramphenicol acetyltransferase type III (<i>E. coli</i>)	4CLA	1	
			23S rRNA polymer (<i>D. radiodurans</i>)	1K01	1	
			Xenobiotic acetyltransferase (<i>P. aeruginosa</i>)	2XAT	1	
6 ^d	Ciprofloxacin	C ₁₇ H ₁₈ N ₃ O ₃ F	Transmembrane efflux pump AcrB (<i>E. coli</i>)	1OYE	1	
			Chloramphenicol phosphotransferase (<i>S. venezuelae</i>)	1QHY	2	
7	Rhodamine 6G	[C ₂₈ H ₃₁ N ₂ O ₃] ⁺	Transcriptional repressor QacR (<i>S. aureus</i>)	1JUS	1	
8	Tetracycline	[C ₂₂ H ₂₅ N ₂ O ₈] ⁺	Transmembrane efflux pump AcrB (<i>E. coli</i>)	1OY8	1	
			16S rRNA polymer (<i>T. thermophilus</i>)	1HNW	1	
9 ^d	Telithromycin	[C ₄₃ H ₆₅ N ₅ O ₁₀] ⁺	16S rRNA polymer (<i>T. thermophilus</i>)	1I97	6	
			Tetracycline repressor TetR (<i>E. coli</i>)	2TRT	1	
10	Streptomycin	[C ₄₁ H ₆₇ NO ₁₅] ⁺	23S rRNA polymer (<i>D. radiodurans</i>)	1P9X	1	
			23S rRNA polymer (<i>D. radiodurans</i>)	1OND	1	
11	Deoxycholate	[C ₂₄ H ₃₉ O ₄] ⁻	Double-stranded RNA aptamer (synthetic construct)	1NTA	1	
			Double-stranded RNA aptamer (synthetic construct)	1NTB	1	
			Ketosteroid isomerase (<i>P. putida</i>)	1E3V	1	
			Translation elongation factor Selb (<i>M. maripaludis</i>)	1WB1	7	
			Translation elongation factor Selb (<i>M. maripaludis</i>)	1WB2	7	
18 ^d	Cefotaxime group	[C ₁₆ H ₁₆ N ₅ O ₇ S ₂] ⁻	Translation elongation factor Selb (<i>M. maripaludis</i>)	1WB3	7	
			Conjugated bile acid hydrolase (<i>C. perfringens</i>)	2BJF	1	
19	Ampicillin	[C ₁₆ H ₁₉ N ₃ O ₄ S] [±]	D-Ala-D-Ala-peptidase (<i>Streptomyces</i> sp.)	1CEF	1	
			Toho-1 β-lactamase (<i>E. coli</i>)	1IYO	1	
20	667-Coumate	C ₁₄ H ₁₅ NO ₅ S	α-Amino acid ester hydrolase (<i>A. pasteurianus</i>)	1NX9	4	
			Deacetoxycephalosporin C synthase (<i>S. clavuligerus</i>)	1W2N	1	
23a,b ^d	Loracabef	[C ₁₆ H ₁₆ N ₃ O ₄ Cl] [±]	Antiampicillin antibody (Mouse e 010090)	1H8S	1	
			Carbonic anhydrase II (<i>H. sapiens</i>)	1TTM	1	
23a,b ^d	5-Methyl-2-(1-methylethyl)phenol N-acetyl-D-glucosamine	C ₁₀ H ₁₄ O	β-Lactamase (<i>E. coli</i>)	1FCN	2	
			C ₈ H ₁₅ NO ₆	Porcine odorant binding protein (<i>S. scrofa</i>)	1E06	2
				Acetylcholinesterase (<i>T. californica</i>)	1GQS	2
			C ₉ H ₁₀ O ₂	Acetylcholinesterase (<i>T. californica</i>)	1QQR	1
				Vanillyl-alcohol oxidase (<i>P. simplicissimum</i>)	1DZN	2
				Vanillyl-alcohol oxidase (<i>P. simplicissimum</i>)	1W1M	2
				Vanillyl-alcohol oxidase (<i>P. simplicissimum</i>)	1W1L	2
				Vanillyl-alcohol oxidase (<i>P. simplicissimum</i>)	1W1K	2
				Vanillyl-alcohol oxidase (<i>P. simplicissimum</i>)	1W1J	2
			24 ^d	3-(Prop-2-ene-1-sulfinyl)-propene-1-thiol	C ₆ H ₁₀ OS ₂	Vanillyl-alcohol oxidase (<i>P. simplicissimum</i>)
Human glutathione reductase (<i>H. sapiens</i>)	1BWC	1				
25 ^d	6-Aminohexanoic acid	C ₆ H ₁₃ NO ₂	Tissue-type plasminogen activator (<i>H. sapiens</i>)	1PK2	1	
			1-Pyrenyl ^e	[C ₁₉ H ₉]	Double-stranded DNA dodecamer with 1-pyrenyl nucleobase (synthetic construct)	1FZL
27 ^d	Cryptolepine	[C ₁₆ H ₁₂ N ₂] ⁺	Double-stranded DNA dodecamer with 1-pyrenyl nucleobase (synthetic construct)	1FZS	8	
			Double-stranded DNA hexamer (synthetic construct)	1K9G	2	

^a The name and formula of the substrate retrieved from the PDB.^b PDB reference code.^c Number of symmetrically independent substrate molecules.^d Agents that were not found in the PDB as complexed with a receptor. Whenever possible, the most similar substrates complexed to a macromolecular receptor were retrieved from the PDB, even in cases when the agent in question was found in the PDB.^e Pyrene–ribose bond in the deoxyribofuranosylpyrene nucleoside was broken before the LPC analysis.

Table 4. Molecular descriptors above the cut-off (0.500) in correlation with the biological activities pMIC.

No.	Symbol	Definition ^a	R(KAM) ^b	R(pVCJ6) ^c
8	Iy	2nd principal moment of inertia	-0.262	-0.621
9	Iz	3rd principal moment of inertia	-0.216	-0.654
10	FF	log(Dip + 1), Dip is molecular dipole moment (Titan)	0.431	0.721
14	Nh	No. hydrophobic carbon atoms	-0.601	-0.200
15	Na	No. aromatic carbon atoms	-0.469	-0.645
17	Np	No. polar (not hydrophobic) atoms (non-H)	0.203	0.517
26	Nr	No. ring atoms (non-H)	-0.570	-0.410
27	wh	Nh/Nt, number fraction of hydrophobic atoms; Nt is total no. non-H atoms	-0.566	-0.726
28	wa	Na/Nt, number fraction of aromatic atoms	-0.387	-0.754
29	wb	Nb/Nt, number fraction of hydrogen bonding non-H atoms; Nb is no. HB donors/acceptors	0.562	0.579
30	wp	Np/Nt, number fraction of polar atoms	0.566	0.726
31	wl	Nl/Nt, number fraction of planar atoms; Nl is no. non-H atoms in all planar fragments	-0.177	-0.553
40	wON	NON/Nt, number fraction of O/N atoms; NON is no. oxygen and nitrogen atoms	0.454	0.558
45	Bat	B/Nt, no. bonds per atom; B is no. bonds (non-H)	-0.415	-0.603
46	wr	Nr/Nt, number fraction of ring atoms	-0.425	-0.713
50	DM	molecular dipole moment (E4 method, MOPAC)	0.694	0.788
54	E4	average polarizability (E4 method, MOPAC)	-0.611	-0.227
55	BT	β hyperpolarizability along the dipole moment (E4 method, MOPAC)	-0.663	-0.456
59	GM	Absolute average γ hyperpolarizability (E4 method, MOPAC)	-0.108	-0.508
63	Wn3	W/(Nt) ³ , normalized Wiener index; W is H-depleted Wiener index	0.510	0.221
64	HOMO-1	Energy of HOMO-1 orbital	0.584	0.375
65	HOMO	Energy of HOMO orbital	0.588	0.376
68	Q-	The most negative ESP atomic charge (non-H)	-0.672	-0.675
69	Q+	The most positive ESP atomic charge (non-H)	0.426	0.541
74	Qdf	(Q+) - (Q-), the largest ESP charge difference	-0.539	-0.636
76	%hcpk	hcpk/S, surface area fraction of hydrophobic C/H atoms; hcpk and Scpk are CPK surface areas of hydrophobic C/H atoms and the molecule, respectively	-0.508	-0.508
78	Mrefn	Mref/N, molecular refractivity per atom; Mref is molecular refractivity (C log P method, CHEM3D); N is no. all atoms	-0.259	-0.606
79	Enh1	(HOMO-1)/N, HOMO-1 orbital energy per atom	0.233	0.538
80	Enh	(HOMO)/N, HOMO orbital energy per atom	0.232	0.529
83	Enh-1h	[(HOMO-1) + (HOMO)]/N, frontier orbital energy sum (HOMO-1) + (HOMO) per atom	0.232	0.534
84	Enh-1l	[(HOMO-1) + (LUMO)]/N, frontier orbital energy sum (HOMO-1) + (LUMO) per atom	0.328	0.501
87	DMn	DM/N, DM per atom	0.819	0.788
88	BTn	BT/N, hyperpolarizability β per atom	-0.743	-0.551
89	Bpa	B/Nt, no. bonds per atom; B is no. bonds (non-H)	-0.465	-0.633
90	L	No. non-H atoms along the longest bond chain	0.215	0.540
93	Xmin	Minimum X coordinate	-0.398	-0.528
97	Ymax	Maximum Y coordinate	-0.638	-0.138
98	Zmax	Maximum Z coordinate	-0.577	-0.045
100	DY	Ymax - Ymin, molecular box width	-0.543	-0.074
101	DZ	Zmax - Zmin, molecular box height	-0.538	-0.015
104	sighb	Nb/Sm, HB donors/acceptors surface density; Sm is molecular surface area	0.496	0.533
105	sighyd	Nh/Sm, hydrophobic carbon surface density	-0.660	-0.752
106	sig	Na/Sm, aromatic carbon surface density	-0.347	-0.686
107	sigp	Np/Sm, polar atom surface density	0.494	0.669
108	sigr	Nr/Sm, ring atom surface density	-0.426	-0.671
109	sigon	ON/Sm, ON atom surface density	0.375	0.536
110	ntb	Nt/B, non-H atoms per bond	0.469	0.629
111	Np2	(Np-12) ² , square function of Np	-0.527	-0.787
112	ON2	(ON-10) ² , square function of ON	-0.463	-0.593
113	wh2	(wh-0.6) ² , square function of wh	-0.682	-0.784
114	wa2	(wa-0.3) ² , square function of wa	-0.468	-0.627
115	wp2	(wp-0.4) ² , square function of wp	-0.682	-0.784
116	Mrefn2	(Mrefn-0.19) ² , square function of Mrefn	-0.485	-0.745
117	sigp2	(sig-0.022) ² ; sig = Nve/Sm, valence electron surface density; Nve is no. valence electrons	-0.609	-0.674
118	RD	Ratio of actual and standard bond lengths sum ^d	0.232	0.662
119	RD2	(RD-0.98) ² , square function of RD	-0.583	-0.787
120	RRD	1/RD, inverse of RD	-0.243	-0.673

^a Abbreviation "No." stands for "the number" or "the number of".^b Correlation coefficient with activity pMIC(KAM).^c Correlation coefficient with activity pMIC(pVCJ6).^d The bond lengths sum excludes H atoms. Actual bond lengths are from agents **1-12** and standard bond lengths are analog single bond lengths from the simplest molecules: C-C (CH₃-CH₃), C-N (CH₃-NH₂), C-P (CH₃-PH₂), S-O (HS-OH), C-F (CH₃-F), C-O (CH₃-OH), N-O (H₂N-OH), C-Cl (CH₃-Cl), C-S (CH₃-SH), N-N (H₂N-NH₂).

Table 5. Selected molecular descriptors^a for agents **1–30**.

No.	FF	Nh	wr	E4/a.u.	HOMO/eV	sighyd/Å ⁻²	Np2	wh2	Mrefn2/cm ⁶ /mol ²
1	0.4236	14	0.8824	163.649	-7.837	0.0510	81	0.0500	0.204304
2	0.0004	24	0.9600	215.097	-13.003	0.0622	121	0.1296	0.385641
3	0.4307	14	0.8235	203.297	-11.666	0.0527	81	0.0500	0.163216
4	0.5931	21	0.8333	241.318	-11.251	0.0580	81	0.0756	0.160000
5	0.8224	10	0.3000	129.924	-10.264	0.0303	4	0.0100	0.170569
6	1.6850	14	0.6957	175.499	-6.345	0.0396	9	0.0001	0.015625
7	0.6896	27	0.6061	356.642	-11.168	0.0509	36	0.0476	0.037636
8	1.1533	19	0.5625	219.420	-11.366	0.0435	1	0.0000	0.002601
9	2.3617	35	0.5098	316.371	-11.867	0.0466	16	0.0074	0.095481
10	1.7979	20	0.4250	249.845	-13.862	0.0329	64	0.0100	0.076176
11	1.5021	23	0.6071	176.495	-4.333	0.0526	49	0.0490	0.078961
12	1.6083	12	0.0000	110.602	-6.123	0.0324	49	0.0112	0.063504
13	1.6795	15	0.7917	181.952	-6.382	0.0404	9	0.0600	0.000237
14	0.9847	19	0.5455	223.595	-11.831	0.0446	4	0.0600	0.000061
15	1.0390	19	0.5625	219.762	-11.625	0.0450	1	0.0000	0.000074
16	1.2843	24	0.6316	262.929	-11.066	0.0464	4	0.1000	0.000033
17	1.5743	18	0.5278	248.055	-4.857	0.0323	36	0.0100	0.004361
18	1.2869	12	0.4333	223.456	-5.099	0.0253	36	0.0400	0.003017
19	1.5679	13	0.5417	172.012	-6.874	0.0333	1	0.3400	0.000020
20	1.1681	13	0.5000	177.386	-2.837	0.0324	1	0.0100	0.001189
21	1.1157	8	0.6000	132.976	-5.275	0.0266	0	0.0400	0.000409
22	1.1321	6	0.5000	85.768	-4.742	0.0260	16	0.0295	0.000075
23a	0.4598	11	0.4615	94.985	-8.637	0.0460	100	0.0606	0.000046
23b	0.4777	11	0.4615	94.575	-8.711	0.0463	100	0.0606	0.000046
24	1.2735	5	0.0000	63.570	-4.328	0.0297	81	0.0600	0.000100
25	0.8000	15	0.9412	158.033	-12.588	0.0608	100	0.0797	0.003543
26	0.4350	13	0.8750	196.349	-11.709	0.0516	81	0.0452	0.002060
27	0.5812	15	0.8824	174.146	-8.298	0.0536	64	0.0797	0.003230
28	0.9236	7	0.3571	112.417	-9.684	0.0299	25	0.0100	0.001005
29	0.5990	19	0.8750	179.072	-9.211	0.0550	49	0.0367	0.002088
30	0.9827	3	0.0000	77.944	-9.332	0.0147	9	0.1225	0.006467

^a Descriptors are defined in Table 4.

substrates. The regression vectors for KAM32/pVCJ6 have contributions of three electronic descriptors (FF, Np2, and Mrefn2), a steric (wr) and a hydrophobic (sighyd) descriptors accounting for agent efflux by VmrA. Np2 and wr are important to distinguish **1–4** from **5–12** (Table 4). Good VmrA substrates have more ring and hydrophobic atoms, smaller dipole moments, fewer polar atoms, and lower molar refractivity than poor ones. Pronounced agent hydrophobicity relative to strain KAM32 and agent electronic/steric properties relative to KAM32/pVCJ6 can indicate differences in the predominant types of agent–pump interactions. pMIC Δ (Table 7) can be rounded to 1 or 0, correctly showing VmrA resistance to **1–4** and sensitivity to **6–8, 10, and 12**. Incorrectly characterized are **9** (false resistance) and **5 and 11** (resistance not clear).

3.3 Exploratory Analysis for the Training Set

The autoscaled data matrix 12 \times 5 already used in the PLS and PCR modeling of pMIC(pVCJ6), *i.e.*, based on descriptors FF, wr, sighyd, Np2, and Mrefn2 was analyzed by means of PCA and HCA. The autoscaled data matrix 12 \times 5 from the regression modeling of pMIC(KAM), *i.e.*, built

from descriptors Nh, E4, HOMO, sighyd, and wh2 was also treated by PCA and HCA.

In case of the dataset for strain KAM32/pVCJ6, PCA shows that PC1–PC3 contain 92.7% of the original information. PC1 alone has 70.0% of the variance, the same as in the regression analyses (Table 6). The PCA scores plot (Figure 2, top) clearly distinguishes good VmrA substrates **1–4** (group G with pMICs from 3.2 to 4.4) from partially mixed moderately good **7, 9–11** (group M with pMICs from 4.7 to 5.5) and poor substrates **5 and 12** (group P1), and poor substrates **6 and 8** (group P2; pMICs for P1 and P2 are from 5.8 to 7.1). Structural distinction can also be observed: G – heteroaromatic species, M – macrocycles/polycycles (**7** – exception, a heteroaromatic system), P1 – extended structures with the main chain, P2 – extended structures built from condensed and linked rings. Main cluster discrimination comes from PC1, which is highly correlated with the biological activity, correlation coefficient is -0.889. PC2 is related to molecular rigidity, flatness, and compactness: more rigid and flatter species lie at high positive PC2, while more flexible and branched species are situated at highly negative PC2. Transform (PC2 + 0.2)² is highly correlated with the pMIC (correlation coef-

Table 6. PLS and PCR regression vectors^a and statistics^b.

Parameter	PLS (pVCJ6)	PCR (pVCJ6)	PLS (KAM)	PCR (KAM)
PCs (%)	1 (70.0)	1 (70.0)	1 (52.8)	1 (53.6)
SEV	0.721	0.683	0.452	0.443
SEP	0.563	0.565	0.368	0.389
Q	0.803	0.818	0.762	0.763
R	0.889	0.888	0.866	0.849
Q ²	0.587	0.630	0.548	0.566
R ²	0.791	0.789	0.750	0.721
Δ	0.410	0.410	0.288	0.270
FF	0.206	0.209		
Nh			−0.228	−0.260
wr	−0.204	−0.212		
E4			−0.231	−0.261
HOMO			0.223	0.190
sighyd	−0.215	−0.217	−0.250	−0.258
Np2	−0.225	−0.215		
wh2			−0.258	−0.250
Mrefn2	−0.213	−0.210		

^a Regression vectors and statistical parameters are given for PLS and PCR models for the prediction of the efflux activity of *E. coli* strains KAM32/pVCJ6 (pVCJ6) and KAM32 (KAM).

^b Parameters: PCs (%) – the number of principal components and % of the total variance that they contain; SEV – standard error of validation; SEP – standard error of prediction; Q – linear correlation coefficient of leave-one-out crossvalidation; R – linear correlation coefficient of prediction; Q² – correlation coefficient of leave-one-out crossvalidation; R² – correlation coefficient of prediction; Δ – average absolute deviation of predicted from experimental activity values.

Table 7. Predicted^a efflux activities (pMICs) with absolute and relative deviations^b and difference^c.

No.	pMIC(pVCJ6)	Δ(pVCJ6)	Δ(pVCJ6)	pMIC(KAM)	Δ(KAM)	Δ(KAM)	pMICΔ
1	4.076/4.076	0.834/0.834	25.7/25.7	5.763/5.796	0.198/0.165	3.3/2.8	1.687/1.720
2	2.884/2.895	0.584/0.573	16.8/16.5	4.594/4.716	0.078/0.044	1.7/0.9	1.710/1.821
3	4.187/4.186	0.278/0.277	7.1/7.1	5.446/5.497	0.332/0.383	6.5/7.5	1.259/1.311
4	4.113/4.112	0.279/0.280	6.4/6.4	5.016/5.063	0.022/0.069	0.4/1.4	0.903/0.951
5	5.844/5.844	0.034/0.034	0.6/0.6	6.338/6.356	0.532/0.546	9.1/9.4	0.454/0.512
6	5.899/5.883	1.128/1.144	16.1/16.3	6.239/6.196	1.128/0.831	16.1/11.8	0.340/0.313
7	5.122/5.112	0.345/0.335	7.2/7.0	4.876/4.835	0.099/0.058	2.1/1.2	0.246/0.277
8	5.826/5.809	0.123/0.140	2.1/2.4	5.703/5.671	0.246/0.278	4.1/4.7	0.123/0.138
9	5.887/5.884	0.623/0.620	11.8/11.8	5.015/4.918	0.249/0.346	4.7/6.6	0.872/0.968
10	5.810/5.824	0.346/0.360	6.3/6.6	5.614/5.594	0.150/0.130	2.7/2.4	0.196/0.230
11	5.170/5.169	0.148/0.149	2.8/2.8	5.693/5.668	0.375/0.350	7.1/6.6	0.523/0.499
12	6.265/6.289	0.196/0.172	3.0/2.7	6.512/6.500	0.051/0.039	0.8/0.6	0.247/0.211

^a Left and right values are from PLS and PCR models, respectively.

^b Efflux activities pMIC, their absolute (Δ) and relative (%Δ) deviations from experimental activities for *E. coli* strains KAM32/pVCJ6 and KAM32.

^c Absolute difference between the two predicted efflux activities.

efficient is 0.876) and PC1 (correlation coefficient is −0.731). Because of this, PC2 does not bring useful information in PLS and PCR. The loadings plot (Figure 2, bottom) demonstrates that good VmrA substrates have small dipole moment (FF at negative PC1) and refractivity (Mrefn2 at positive PC1), low content of polar atoms (Np2 at positive PC1), and high content of hydrophobic and ring atoms (wr and sighyd at positive PC1). The opposite is valid for poor substrates.

In the case of the dataset for the strain KAM32, PCA shows that PC1 contains over 50% of the total variance, similar to the regression models (Table 6). The PCA scores plot (Figure 3, top) exhibits two groups of agents with no

sharp boundary: one of more elongated, linear, or chain-like species (group E: **1, 3, 5, 6, 11, 12**), and the other of more branched, cyclic, or spherical species (group B: **2, 4, 7–10**). This distinction applies to both PC1 and PC2. Along PC1, there is a difference between B and E in terms of higher and lower pMIC values, respectively: pMIC ranges from 5.114 to 7.027 in B, and from 4.672 to 5.949 in E. Branching/sphericity increases and elongation/linearity decreases along both PC1 and PC2, placing the most branched and spherical **2** into the top right corner of the scores plot, while more elongated/linear **5, 6, 8, 10, and 12** are concentrated around the bottom of the left quadrant. Similar trends can be observed for the activity. The best

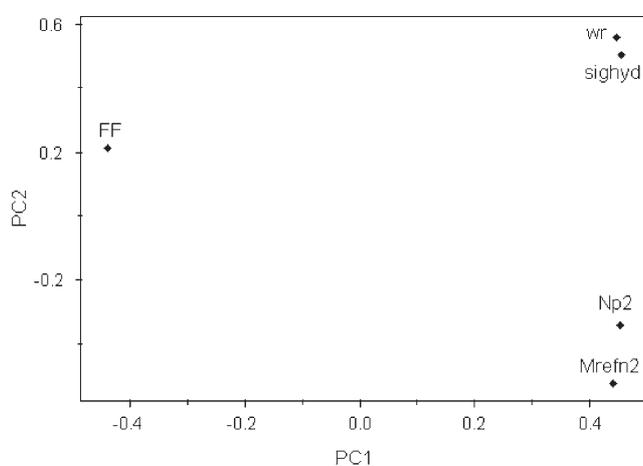
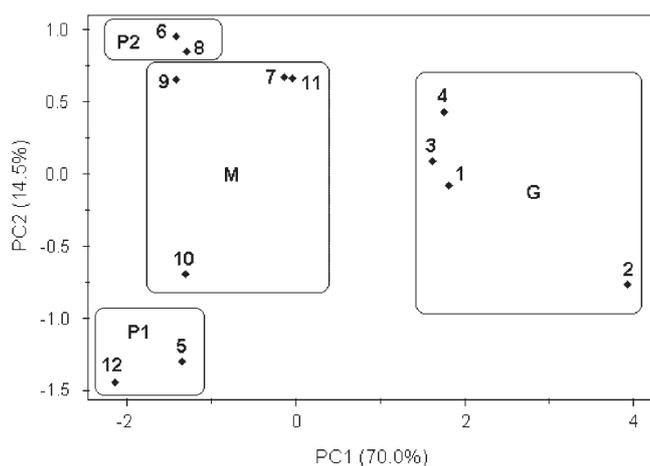


Figure 2. PCA scores (top) and loadings (bottom) plots for the training set, related to the efflux power of the VmrA-containing *E. coli* strain KAM32/pVCJ6.

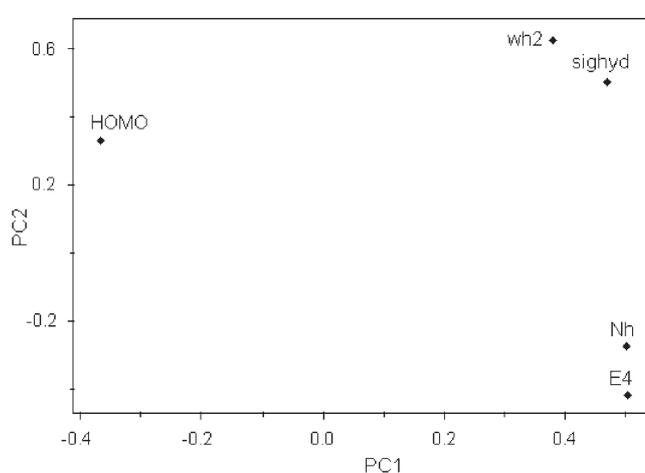
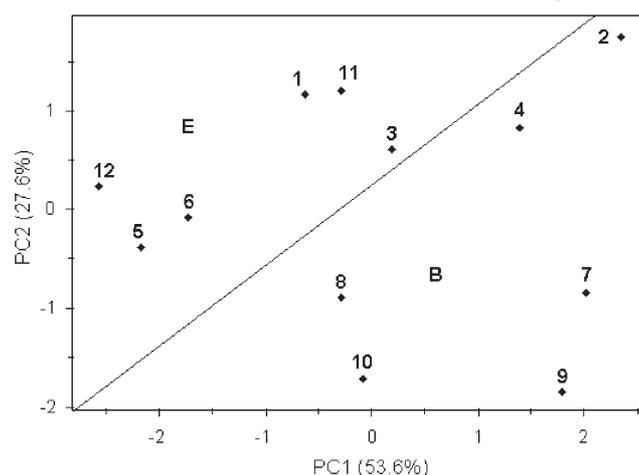


Figure 3. PCA scores (top) and loadings (bottom) plots for the training set, related to the efflux power of *E. coli* strain KAM32.

substrates **2**, **4**, and **7** are at positive PC1 and PC2, while the worst substrates **5**, **6**, **8**, and **12** concentrate in the opposite region. PC1 is well correlated with the activity (correlation coefficient -0.849). $(PC2)^2$ has a moderate correlation with PC1 and the biological activity (correlation coefficients are 0.610 and -0.486 , respectively), which partially explains why this PC does not contribute to regression models (Table 6). The maximum variation in pMIC(KAM) is 2.345, smaller than in pMIC(pVCJ6), 3.785. That is why well-defined clusters of good, moderately good, and poor substrates of KAM32 strain are not visible in the scores (Figure 3). Loadings (Figure 3, bottom) agree well with regression and correlation analyses: good substrates have more negative HOMO, higher content of hydrophobic atoms, and weaker polarizability, while poor substrates have the opposite characteristics.

HCA dendrograms with samples relating to the two datasets for **1–12** (Figure 4) agree with the scores plots (Figures 2 and 3). The dendrogram related to KAM32/pVCJ6 separates the G (good substrates) cluster from mixed M

(moderately good substrates), and P1, P2 (poor substrates). Two-membered subclusters have structurally similar species: heteroaromatics with linear arrangement of the rings (**1**, **3**), complex linear ring systems (**6**, **8**), complex nonlinear ring systems (**7**, **11**), and extremely elongated systems (**10**, **12**). The dendrogram related to KAM32 separates E (more elongated substrates) from B (more branched substrates) clusters. Two-membered subclusters have structurally similar species: linear heteroaromatic systems (**1**, **3**), branched heteroaromatic systems (**2**, **4**), general linear systems (**6**, **12**), complex linear ring systems (**8**, **10**), and very complex ring systems (**7**, **9**).

1–12 represent 12 distinct compound classes, including organic dyes (**1–4**, **7**), a detergent (**12**), a tetracycline (**8**), fluoroquinolone (**6**) and macrocyclic (**9**) antibiotics, and other agents (**5**, **10**, **11**). There are also some common structural features related to efflux from KAM32 and KAM32/pVCJ6 strains. In general, agent efflux is not determined by specific functional groups due to the strictly

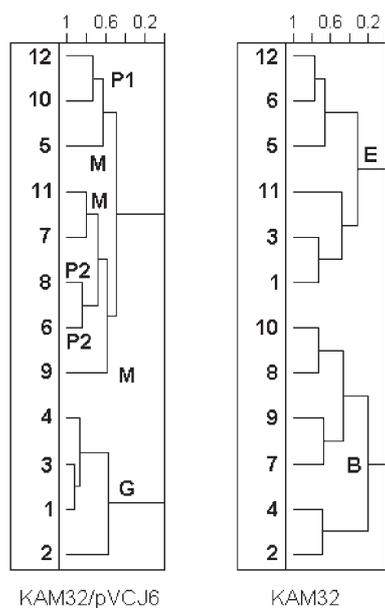


Figure 4. HCA dendrogram of agents with complete linkage for the training sets that characterize the efflux power of the two *E. coli* strains.

nonbonding nature of pump-mediated MDR. Exploratory analysis related to the two strains points out that poor substrates are usually more linear, elongated, chain-like systems, relatively poor in hydrophobic groups and rich in polar groups, easy to polarize, and good dipoles. Good substrates are more branched, cyclic and spherical systems, weak dipoles, hard to polarize, poor in polar and rich in hydrophobic groups. The strains differ in their preferences for good substrates: KAM32 binds rather hydrophobic species, while KAM32/pVCJ6 attracts heteroaromatics. Thus, aromatic–aromatic and hydrophobic–hydrophobic VmrA-substrate interactions seem to be crucial.

3.4 QSAR/SAR Prediction of the MDR Character of VmrA with Respect to the Prediction Set

Descriptors for **13–30** are in Table 5, and predicted activities pMIC(pVCJ6), pMIC(KAM), and pMIC Δ in Table 8. Predicted efflux activities for a particular *E. coli* strain, from PLS and PCR models, are very similar (Table 7). However, the differences between pMIC Δ obtained by PLS and PCR for the training set reached 0.111 in log units (average difference 0.047). The analog differences for the prediction set reached 0.166 (average difference 0.056). This shows an elevated error accumulation in pMIC Δ for the prediction set, in comparison with the training set.

To which agent will VmrA be sensitive or resistant? Is it possible to predict how much VmrA will be resistant to a particular compound? According to the literature, *V. parahaemolyticus* is sensitive to **13–24**, meaning that pMIC Δ should be small for these agents. On the other hand, exper-

Table 8. Predicted^a pMICs for **13–30** and the differences pMIC Δ ^b.

No.	pMIC(pVCJ6)	pMIC(KAM)	pMIC Δ
13	5.774/5.755	6.185/6.137	0.411/0.382
14	5.727/5.711	5.648/5.620	0.079/0.091
15	5.739/5.721	5.664/5.634	0.075/0.087
16	5.717/5.699	5.579/5.395	0.138/0.304
17	5.002/5.015	6.139/6.058	1.137/1.043
18	5.473/5.486	6.293/6.269	0.820/0.783
19	6.234/6.222	6.333/6.297	0.099/0.075
20	5.879/5.870	6.514/6.453	0.635/0.583
21	6.105/6.091	6.556/6.573	0.451/0.482
22	6.180/6.173	6.794/6.816	0.614/0.643
23a	4.919/4.933	5.978/6.056	1.059/1.123
23b	4.918/4.932	5.970/6.049	1.052/1.117
24	6.118/6.149	6.962/6.959	0.844/0.810
25	3.429/3.437	5.203/5.309	1.774/1.872
26	4.068/4.068	5.523/5.574	1.455/1.506
27	3.906/3.904	5.510/5.575	1.604/1.671
28	5.857/5.860	6.482/6.509	0.625/0.649
29	4.260/4.249	5.541/5.551	1.281/1.302
30	5.379/5.411	6.395/6.560	1.016/1.149

^a Left and right values are from PLS and PCR models, respectively. Predicted activities pMIC are for *E. coli* strains KAM32/pVCJ6 and KAM32.

^b Absolute difference between the two predicted efflux activities.

imental data (Table 1) and QSAR/chemometric analyses for the **1–12** point out that VmrA is resistant to **1–4** and sensitive to **5–12**. Similar agents should behave similarly: VmrA should be resistant to heteroaromatics **25–27** and **29**. This approach to predict VmrA resistance is qualitative, and is only possible because of the literature. Without having literature data for β -lactams and their analogs **17–22**, it is impossible to predict VmrA character toward these agents and also for **28** and **30**. There are three systematic ways to predict if VmrA would be resistant or sensitive to **13–30**: QSARs (Table 8), exploratory analysis for training + prediction set using data from Table 5, and employing discriminatory descriptors (Np, wa, wl, wr, sigr, Np2, RD, RRD; see Figure A).

A limit of 0.5 in log units was applied for experimental pMIC Δ to distinguish agents to which VmrA was resistant or sensitive (**1–4** or **5–12**, respectively, Table 1). When this limit is applied to predicted pMIC Δ for the training set (Table 7), 9 out of 12 agents displayed correctly assigned VmrA resistance/sensitivity. Due to error accumulation in regression models and pMIC Δ , the limit should be 0.89 (PLS models) or 0.96 (PCR models). When the experimental limit of 0.5 is applied to the predicted pMIC Δ for the prediction set (Table 8), the VmrA resistance/sensitivity is clearly assigned to 10 out of 19 agents. As one would expect, VmrA is probably resistant to heteroaromatics **25–27** and **29** (very high pMIC Δ), and sensitive to agents **13–16**, **19**, and **21** (very low pMIC Δ). The other nine agents have intermediate pMIC Δ values. When a limit of 1.20 is used, VmrA is resistant only to **25–27** and **29**,

and sensitive to all other agents. The ambiguity regarding which limit should be used for the predicted pMIC Δ can be resolved by exploratory analysis. What exploratory analysis cannot provide and QSAR models can, is the relative efflux power of VmrA for **13–30**. Very low efflux power (high sensitivity) is clearly predicted for agents **13–16** and **19**. Lower VmrA sensitivity accounts for the others: β -lactam-like systems (**17, 18, 20–22**), xenobiotics (**23, 24**), and other agents (**28, 30**).

PCA and HCA were performed for the training + prediction set relative to KAM32/pVCJ6 (Figure 5). PC1–PC3 describes 88.4% of the total variance. PC1 and PC2 are sufficient to distinguish good (G) from moderately good or poor (M, P) substrates, *i.e.*, to which VmrA will be resistant or sensitive, respectively. In fact, the agents from the prediction set (Figure 5, left) are concentrated around structurally similar agents from the training set (see Figure 2, top). This situation results in simple extension of previous clusters G and mixed P, M. The dendrogram (Figure 5, right) is even clearer in terms of the resistance/sensitivity classification.

Exploratory analysis for the prediction set relative to strain KAM32 (Figure 3, top) could not recognize clearly the agents to which VmrA would be resistant: two groups of more elongated (E) and more branched (B) agents were observed. In the new analysis for the training + prediction set (Figure D in Supplementary Material), this E–B discrimination can be perceived with a more detailed clustering that may aid in VmrA resistance/sensitivity assignments for these agents.

Eight discriminatory descriptors (Np, wa, wl, wr, sigr, Np2, RD, RRD) for the training and training + prediction sets were used in PCA and HCA (Figure 6). Three PCs contain 91.9 and 91.6% of the total variance for the training and training + prediction sets, respectively. The descriptors distinguish well (Figure A) agents to which VmrA is resistant (G group) from those to which VmrA is sensitive (mixed P, M group). This trend is seen in the scores plot and HCA dendrogram (Figure 6, top). In case of the training + prediction set, this discrimination is well defined in the scores plot (Figure 6, left bottom). The HCA dendrogram also shows that the G cluster is separated from the big P, M cluster, and is far from the small isolated group (**24, 28, 30**) because of a low similarity index (<0.2). Therefore, the exploratory analyses and QSAR results for pMIC Δ , all presented in this section, can be used for the prediction of VmrA resistance/sensitivity to general drugs and xenobiotics.

3.5 Agent–Receptor Interaction Statistics

1, 4, 5, 7, 8, 10, and 11 were found as substrates of protein or nucleic acid receptors in the PDB (Table 3). Instead of **2, 3, 6, and 9**, the most similar species as substrates of diverse macromolecules were retrieved from the PDB. In total, 56 crystallographically different substrates were retrieved and more than 30 agent–receptor interaction descriptors (interaction counts, surface areas, and their fractions) were calculated. Only **12** could not be included since no appropriate substrate for it was found. Some PDB structures had MDR efflux pumps as receptors (Table 3).

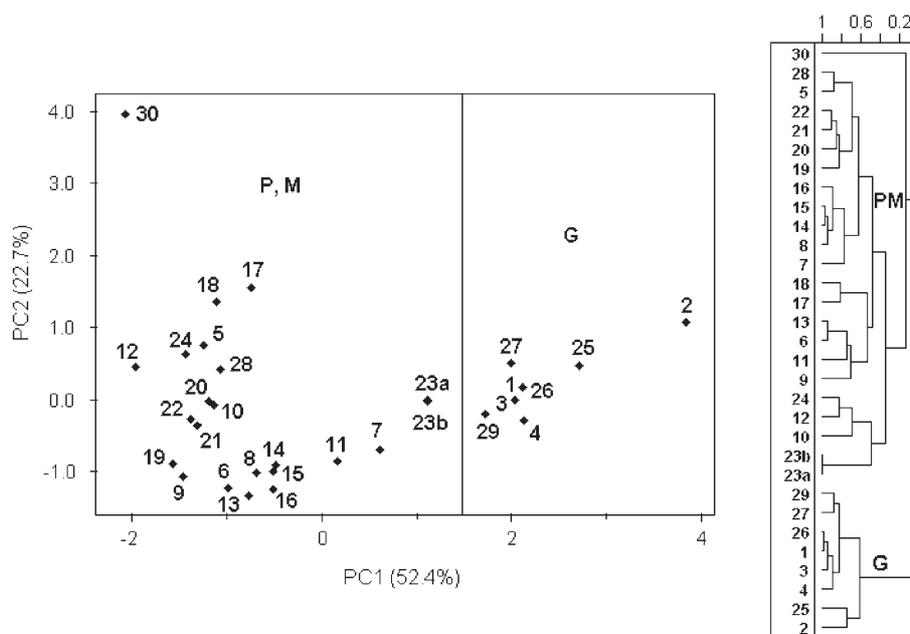


Figure 5. PCA scores plot (left) and HCA dendrogram of samples with complete linkage (right) related to the efflux power of *E. coli* strain KAM32/pVCJ6 with respect to the training + prediction set **1–30**.

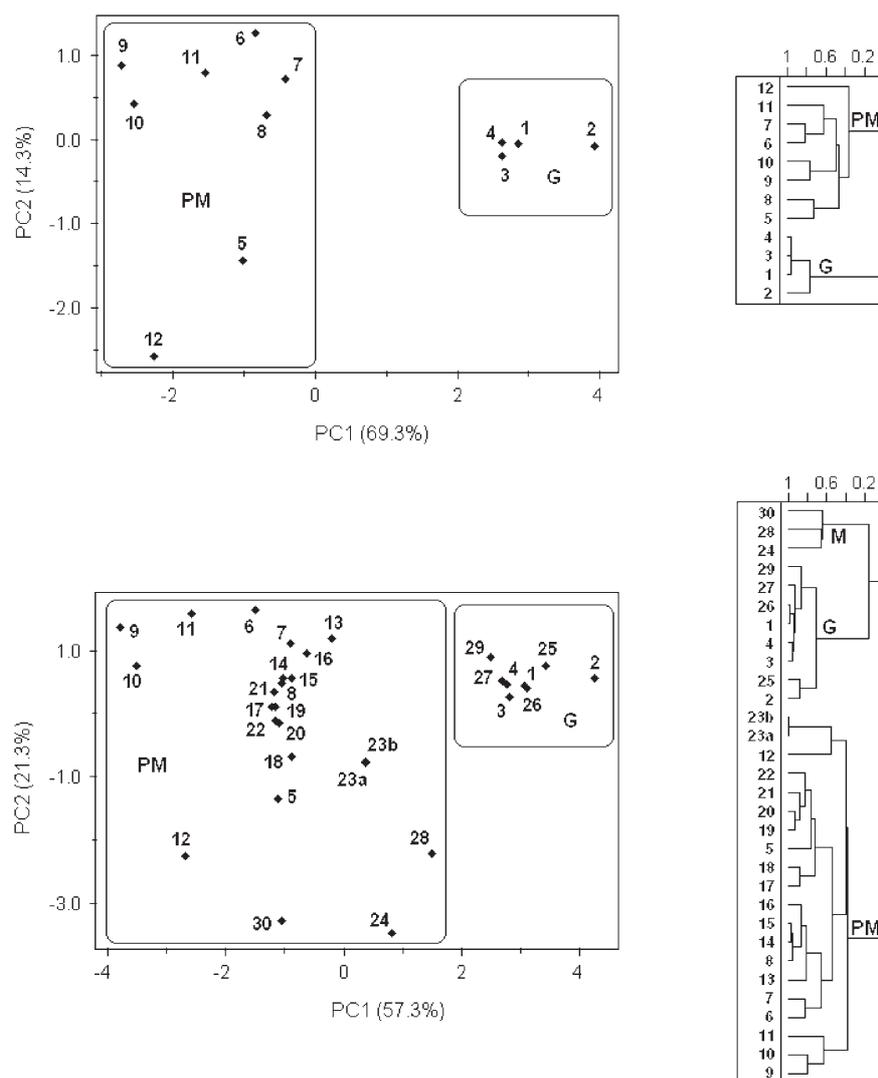


Figure 6. PCA scores plots (left) and HCA dendrograms of samples with complete linkage (right) for the training (top) and training + prediction (bottom) datasets of eight discriminatory descriptors.

In general, agent–receptor interaction descriptors depend on agent molecular properties, receptor molecular properties, agent position and binding mode relative to the receptor, and the presence of other molecular/ionic species. Supposing that agent properties may substantially affect agent–receptor interactions, it was expected that variations in receptors and agent binding modes/positions would have a secondary importance. Correlations of the activities pMIC(KAM) and pMIC(pVCJ6) with the agent–receptor interaction descriptors were studied in two modes: using descriptors for all 56 substrates, and using descriptors averaged for each of the substrates **1–11**.

Correlations for 56 (Figure 7) and 11 (Figure 8) substrates show the role of aromatic–aromatic, hydrophobic–hydrophobic, and hydrophobic–hydrophilic (destabilizing) interactions or contacts, hydrogen bonds and other contacts, and agent contacts with N and C atoms from the receptors.

Aromatic–aromatic contacts (Figure 7, top) are important for the activity pMIC(pVCJ6) and not for pMIC(KAM): the number (AA), agent surface area (SSAR), and the number fraction of these contacts (AA/TC where TC = number of total contacts). This fact agrees with other observations already presented in this work that heteroaromatic species are good VmrA substrates. On the other hand, 13% residues in VmrA from *V. parahaemolyticus* AQ3334 are aromatic or heteroaromatic (His, Trp, Tyr, and Phe), and 15% are rings (Pro, His, Trp, Tyr, and Phe). As a transmembrane protein, VmrA has elevated content of highly hydrophobic residues (36%: Phe, Val, Leu, Ile), which agrees with the noticeable correlation between pMICs and the number fraction of hydrophobic–hydrophobic contacts (HH/TC). Elevated agent hydrophobicity was already noticed as an essential property of good VmrA substrates in this work. It is interesting to note the parabolic correlation between pMICs and the number

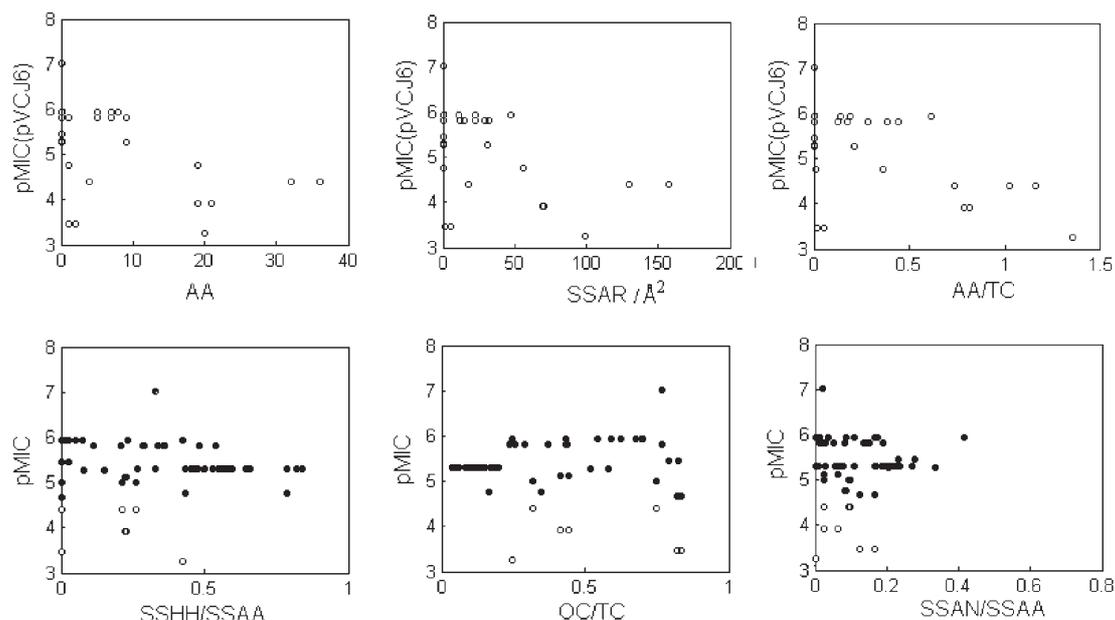


Figure 7. Complete agent–receptor interaction correlations. Biological activities are correlated with: (a) top figures: the number (AA) and the agent surface area (SSAR) of aromatic–aromatic contacts and their number fraction (AA/TC; TC is the total number of agent–environment contacts); (b) bottom figures: the number fraction of hydrophobic–hydrophobic (HH/TC) and of other contacts (OC/TC), and the agent surface area fraction of contacts with receptor nitrogen atoms (SSAN/SSAA; SSAN – the agent surface area of contacts with receptor nitrogen atoms, SSAA – the agent surface area of all contacts). pMIC represents both pMIC(KAM) and pMIC(pVCJ6). White circles in the bottom plots stand for 1–4 with respect to pMIC(pVCJ6), and black circles represent 1–12 with respect to pMIC(KAM) as well as 5–12 with respect to pMIC(pVCJ6).

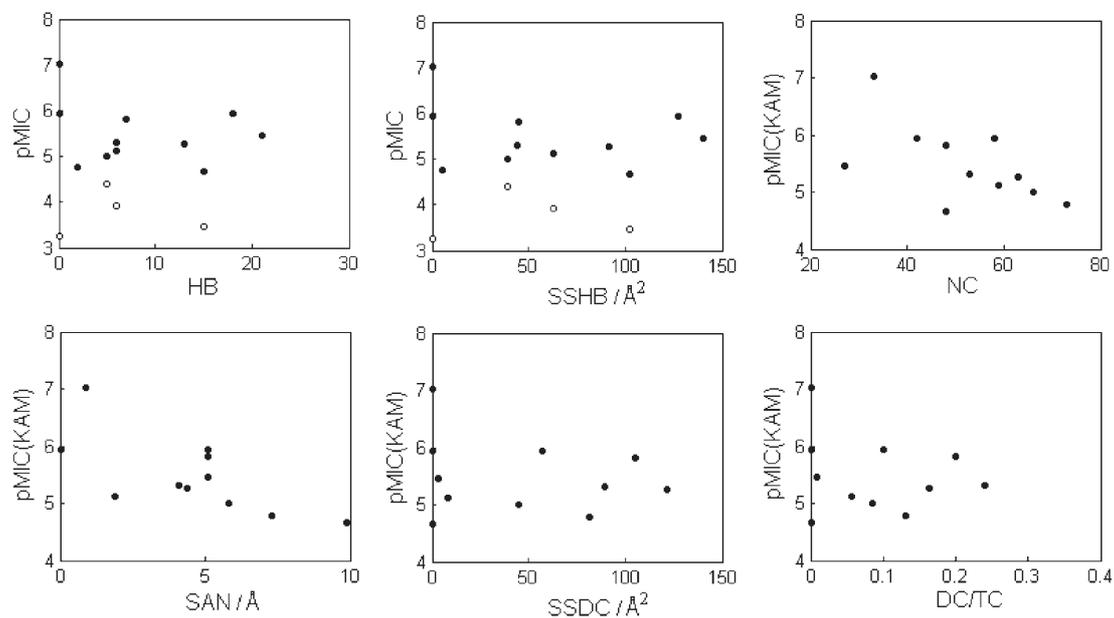


Figure 8. Average agent–receptor interaction correlations. Biological activities are correlated with: (a) top figures: the number of hydrogen bond contacts (HB), the agent surface area of the hydrogen bond contacts (SSHB), and the number of receptor carbon atoms (NC) in contact with the agent; (b) bottom figures: the agent surface area of contacts with receptor’s nitrogen atoms (SAN), the agent surface area of destabilizing contacts, and the number fraction of the destabilizing contacts (DC/TC). pMIC represents both pMIC(KAM) and pMIC(pVCJ6). White circles in the upper plots stand for 1–4 with respect to pMIC(pVCJ6), and black circles represent 1–12 with respect to pMIC(KAM) as well as 5–12 with respect to pMIC(pVCJ6).

fraction of other contacts (OC/TC, Figure 7, bottom). The poor substrates are involved in 50–60% contacts which are of OC contact-type. The surface area fraction of agent contacts with all receptor N atoms shows some regularity. Good substrates are usually in contact with various N atoms (heteroaromatic, hydrogen bonding, charged).

When the averaged descriptors are correlated with the pMICs, correlations become clearer for some other contact types. The number of hydrogen bonding contacts (HB) and agent surface area of these contacts (SSHB) exhibit a parabolic correlation with pMICs, with a minimum of around 13 contacts, and 75–90 Å² surface area (Figure 8, top). In other words, good substrates of VmrA and other pumps in *E. coli* strains establish an optimum HB network with the pumps. pMIC(KAM) shows a clearer correlation with the number of agent contacts with receptor C atoms (NC, Figure 8, top) than pMIC(pVCJ6). The same can be said for the agent surface area of contacts with the receptor N atoms (SSAN, Figure 8, bottom), the agent surface area of destabilizing contacts (SSDC, Figure 8, bottom), and the number fraction of these contacts (DC/TC, Figure 8, bottom).

Correlograms (Figures 7 and 8) regarding HB, destabilizing and other contacts exhibited rather nonlinear (parabolic) relationship with pMICs. The increase or decrease in the number or surface area of HB and destabilizing contacts causes too strong or too weak interactions and, consequently, worse agent efflux. In the case of the other contacts, their increase or decrease leads to a better agent efflux.

Elevated resistance of KAM32/pVCJ6 with respect to KAM32 is detected experimentally as $\text{pMIC(KAM)} \geq \text{pMIC(pVCJ6)}$ (Table 1). This inequality is satisfied for all PLS/PCR predicted pMICs, except for agents 7–10 (Table 7) and 14–16 (Table 8). The difference between $\text{pMIC}\Delta$ and $\text{pMIC(KAM)} - \text{pMIC(pVCJ6)}$ is nonzero for these samples, being small and not essential for 7, 8, 10, 14–16, and significant for 9. This is why 9 was excluded from further analysis. In total, 46 crystallographically different substrates were retrieved from the PDB (Table 3), with structures identical or similar to 18–20, 23–25, and 27. It means that the new agent–receptor descriptors contain in total 100 structures or 18 averaged structures for the training+prediction set (Figure 9). At first, $\text{pMIC}\Delta$ decreases linearly with the agent surface area of all contacts (SSAA), number of destabilizing contacts (NDC), and average distance agent–carbon (receptor) (DC). Therefore, VmrA resistance (high $\text{pMIC}\Delta$) is associated with relatively small molecules (contact area < 200 Å²), the absence of destabilizing interactions, and the presence of strong aromatic, hydrophobic, and other interactions (agent–C contacts, 3.3–3.8 Å as in $\pi\cdots\pi$ stacking interactions [20, 36]). The importance of aromatic–aromatic interactions is reconfirmed for pMIC(pVCJ6) (see example in Figure 9, bottom left). The underpredicted pMIC(pVCJ6) for 2 is the cause why this agent is an outlier in some correlograms (Figure 9). Two more correlograms exhibit the role of destabilizing and other contacts, expressed as the agent surface area fractions, for activities of the two *E. coli* strains (Figure 9; bottom, middle, and

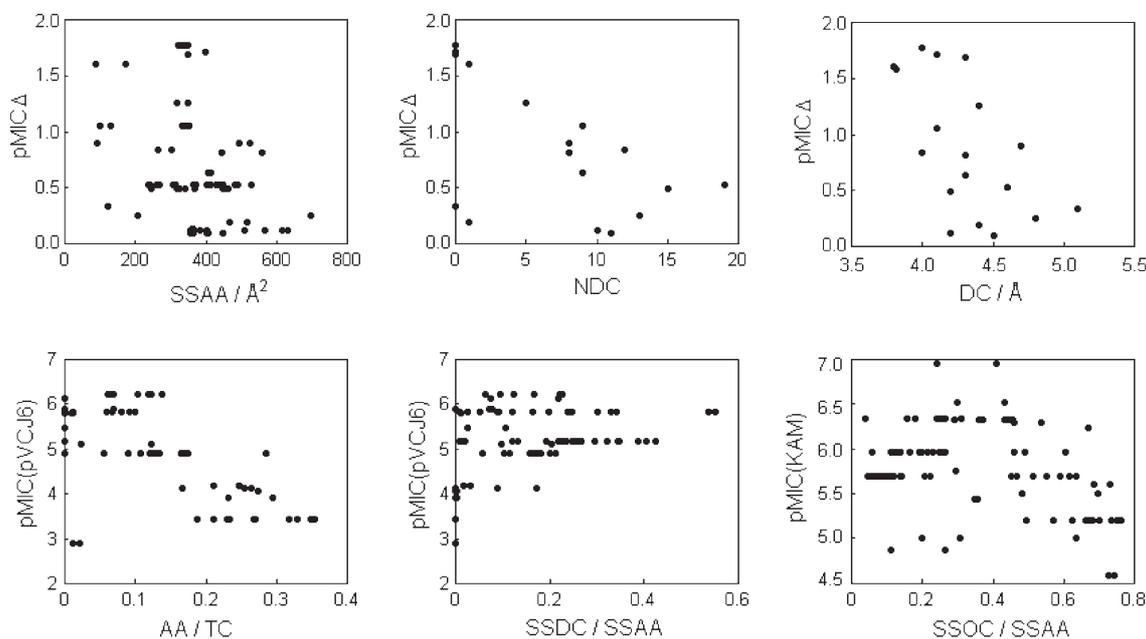


Figure 9. Complete (top left, bottom middle and right) and average (top middle and right, bottom left) agent–receptor interaction correlations. $\text{pMIC}\Delta$ is correlated with (top figures): the agent surface area of all contacts (SSAA), the number of destabilizing (hydrophobic–hydrophilic) contacts (NDC), and the average agent–carbon (receptor) distance (DC). pMIC(pVCJ6) is correlated with (bottom): the number fraction of aromatic–aromatic interactions (AA/TC), and the agent surface area fraction of destabilizing contacts (SSDC/SSAA). pMIC(KAM) is correlated (bottom) with the agent surface area fraction of other contacts (SSOC/SSAA).

right). The other contacts fraction (extreme around 0.4) reconfirms the parabolic nonlinearity in Figure 7 (OC/TC vs. pMICs). Thus linear and nonlinear relations (Figure 9) consist of parallel curves. This is a consequence of the multivariate character of pMICs.

4 Conclusions

Good substrates of efflux systems in *E. coli* strains KAM32 and KAM32/pVCJ6 are neither linear nor extended, preferably are ring structures with some branching, rather hydrophobic, weak dipoles, difficult to polarize, possess some polar groups, and can establish hydrogen bonds with receptors. Besides, good substrates of VmrA efflux pump from *V. parahaemolyticus* strain AQ3334 must also be rather rigid and condensed heteroaromatic systems, with very few or no flexible side-chains, with exo- or endocyclic heteroatoms, and establish aromatic–aromatic contacts with VmrA.

Parsimonious PLS and PCR regression models were established to predict the efflux power of the two strains for the training and prediction sets. The MDR character of VmrA with respect to diverse agents from the two sets was predicted at qualitative (SAR models: VmrA resistance/sensitivity assignments) and quantitative (QSAR models) levels. QSAR-SAR analysis provided a correct distinction of agents to which VmrA is resistant or sensitive, indicating chemometric methods as very useful in practice. Agent–receptor interaction descriptors, based on PDB data, correlate reasonably with biological activities, further rationalizing the VmrA–agent interactions.

Acknowledgements

The authors acknowledge the State of São Paulo Funding Agency (FAPESP) for financial support and Dr. Carol H. Collins for English revision.

References

[1] FDA/CFSAN Bad Bug Book: *Vibrio parahaemolyticus*. <http://vm.cfsan.fda.gov/~mow/chap9.html> (last access 7/9/2005).

[2] CDC/DBMD Disease Information: *Vibrio parahaemolyticus*. <http://www.cdc.gov/neidod/dbmd/diseaseinfo/vibrioparahaemolyticu-g.htm> (last access 7/9/2005).

[3] NZFSA Data Sheet: *Vibrio parahaemolyticus*. http://www.nzfsa.govt.nz/science/data_sheets/vibrio-parahaemolyticus.pdf (last access 8/9/2005).

[4] H. Ho, H. Amin, H.-R. Lee. e-Medicine.com. *Vibrio* Infections. <http://www.emedicine.com/med/topic2375.htm> (last access 8/9/2005).

[5] J. Chen, Y. Morita, M. N. Huda, T. Kuroda, T. Mizushima, T. Tsuchiya, *J. Bacteriol.* **2002**, *184*, 572–576.

[6] Y. Morita, K. Kodama, S. Shiota, T. Mine, A. Kataoka, T. Mizushima, T. Tsuchiya, *Antimicrob. Agents Chemother.* **1998**, *42*, 1778–1782.

[7] M. H. Brown, I. T. Paulsen, R. A. Skurray, *Mol. Microbiol.* **1999**, *31*, 394–395.

[8] Genome composition of *V. parahaemolyticus* RIND2210633. <http://www.membranetransport.org> (last access 12/7/2005).

[9] K. Makin, K. Oshima, K. Kurokawa, K. Yokoyama, T. Uda, K. Tagomori, Y. Iijima, M. Najima, M. Nakano, A. Yamashita, Y. Kubota, S. Kimura, T. Yasunaga, T. Honda, H. Shinagawa, M. Hattori, T. Iida, *Lancet* **2003**, *361*, 743–749.

[10] R. N. Hvorup, B. Winnen, A. B. Chang, Y. Jiang, X.-F. Zhou, M. H. Saier, Jr., *Eur. J. Biochem.* **2003**, *270*, 799–813.

[11] K. Poole, *Antimicrob. Agents Chemother.* **2000**, *44*, 2233–2241.

[12] V. Cattoir, *Pathol. Biol.* **2004**, *52*, 607–616.

[13] D. Ma, D. N. Cook, M. Alberti, N. G. Pon, H. Nikaido, J. E. Hearst, *Mol. Microbiol.* **1995**, *16*, 45–55.

[14] *Titan 1.0.8*, Wavefunction, Inc., Irvine, CA **2001**.

[15] F. H. Allen, *Acta Cryst. B* **2002**, *38*, 380–388.

[16] *Cambridge Structural Database 5.26*, Release November 2004 – Update February 2005, The Cambridge Structural Data Centre, University of Cambridge, Cambridge, UK **2004**.

[17] I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, R. Taylor, *Acta Cryst. B* **2002**, *58*, 389–397.

[18] I. Piantanida, V. Tomisic, M. Zinic, *J. Chem. Soc. Perkin Trans.* **2000**, 375–383.

[19] C.-P. Hsiao, K. J. Siebert, *Int. J. Food Microbiol.* **1999**, *47*, 189–201.

[20] R. Kiralj, B. Kojic-Prodic, I. Piantanida, M. Zinic, *Acta Cryst. B* **1999**, *55*, 55–69.

[21] V. Vrabel, J. Sivy, S. Marchalin, *Acta Cryst. C* **1997**, *53*, 1926–1927.

[22] A. Fitzgerald, L. H. Jensen, *Acta Cryst. B* **1978**, *34*, 828–836.

[23] V. S. Lobanov, *MOPAC 6.0 for Microsoft Windows*, University of Florida **1996**.

[24] *Chem3Dultra 6.0*, CambridgeSoft.Com, Cambridge, MA **2000**.

[25] S. Wold, L. Eriksson, Statistical Validation of QSAR Results, in: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim **1995**, pp. 309–318.

[26] K. R. Beebe, R. Pell, M. B. Seasholtz, *Chemometrics: A Practical Guide*, Wiley, New York **1998**.

[27] M. M. C. Ferreira, *J. Braz. Chem. Soc.* **2002**, *13*, 742–753.

[28] R. Kiralj, M. M. C. Ferreira, *Croat. Chem. Acta* **2005**, *78*, 541–549.

[29] *Matlab 5.2*, Mathworks, Inc., Natick, MA **2001**.

[30] *Pirouette 3.02*, Infometrix, Woodinville, WA **2001**.

[31] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucl. Acid. Res.* **2000**, *28*, 235–242.

[32] *LPC/CSU: Ligand-Protein Contacts & Contacts of Structural Units*, SPACE servers at the Weizmann Institute of Science, Rehovot, Israel. <http://atlantis.weizmann.ac.il/space/servers/> (last access 05/01/2006)

[33] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, M. Edelman, *Bioinformatics* **1999**, *15*, 327–332.

[34] A. Galbraikh, A. Tropsha, *J. Mol. Graph. Mod.* **2002**, *20*, 269–276.

[35] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69–77.

[36] R. Kiralj, Structural studies of 4,9-diazapyrene derivatives. PhD Thesis, University of Zagreb, Zagreb, Croatia **1999**.