CODEN IEJMAT *Internet Electronic Journal of Molecular Design* **2006**, *5*, 515–529 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

October 2006, Volume 5, Number 10, Pages 515–529

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Lemont B. Kier on the occasion of the 75th birthday

Artificial Immune System Prediction of the Human Intestinal Absorption of Drugs with AIRS (Artificial Immune Recognition System)

Ovidiu Ivanciuc [1]

[1] Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555–0857

Received: June 26, 2006; Revised: September 1, 2006; Accepted: September 19, 2006; Published: December 15, 2006

Citation of the article:
O. Ivanciuc, Artificial Immune System Prediction of the Human Intestinal Absorption of Drugs with AIRS (Artificial Immune Recognition System), *Internet Electron. J. Mol. Des.* **2006**, *5*, 515–529, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Artificial Immune System Prediction of the Human Intestinal Absorption of Drugs with AIRS (Artificial Immune Recognition System)

Ovidiu Ivanciuc [1,*]

[1] Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555–0857

Received: June 26, 2006; Revised: September 1, 2006; Accepted: September 19, 2006; Published: December 15, 2006

Internet Electron. J. Mol. Des. **2006**, *5* (10), 515–529

Abstract

Artificial immune systems (AIS) are machine learning procedures inspired by the structure and function of the biological immune system. In this paper we present the first application of the artificial immune recognition system (AIRS) to the modeling of the human intestinal absorption (HIA) of drugs. The learning task was to classify a dataset of 196 drugs into a subset of 131 that penetrate the human intestine and a subset of 65 drugs that do not penetrate the intestine. The chemical structure was encoded with 159 structural descriptors from five classes, namely constitutional, topological indices, electrotopological state indices, quantum descriptors, and geometrical indices. Eight user defined parameters influences the classification performance of the AIRS algorithm, namely affinity threshold scalar, clonal rate, hypermutation rate, number of nearest neighbors, initial memory cell pool size, number of instances to compute the affinity threshold, stimulation threshold, and total resources. In order to explore the AIRS sensitivity to these parameters, leave–20%–out (five–fold) cross–validation predictions were performed over a wide range of values for the AIRS parameters. The affinity threshold scalar has the highest influence on the prediction quality, whereas the remaining parameters have only a marginal effect. The best predictions of the AIRS algorithm (selectivity 0.794, specificity 0.615, accuracy 0.735, and Matthews correlation coefficient 0.406) surpass those obtained with several well–established machine learning methods. In a comparison with 13 machine learning algorithms, AIRS predictions were better in seven cases (Bayesian network, naïve Bayes classifier, updateable naïve Bayes classifier, logistic regression, Gaussian radial basis function network, decision tree with naïve Bayes classifiers at the leaves, and random tree), and worse in six cases (K* instance–based classifier, alternating decision tree, C4.5 decision tree, logistic model trees, random forest, and fast decision tree learner). The results obtained suggest that classifiers based on artificial immune systems may be successful in structure–activity relationships (SAR), quantitative structure–activity relationships (QSAR), drug design, and virtual screening of chemical libraries.

Keywords. Artificial immune system; AIS; artificial immune recognition system; AIRS; human intestinal absorption; HIA; quantitative structure–activity relationships; QSAR.

Abbreviations and notations

AIRS, artificial immune recognition system	IMPS, initial memory cell pool size
ATS, affinity threshold scalar	NIAT, number of instances to compute the affinity threshold
CR, clonal rate	ST, stimulation threshold
HR, hypermutation rate	TR, total resources
kNN, number of nearest neighbors	HIA, human intestinal absorption

Dedicated to Professor Lemont B. Kier on the occasion of the 75th birthday.

* Correspondence author; E–mail: ivanciuc@gmail.com.

1 INTRODUCTION

A large number of machine learning procedures were developed by encoding into an algorithm the mechanisms and functions of biological systems. Well–known examples of biologically inspired algorithms are DNA computing, ant colony optimization, particle swarm optimization, genetic algorithms, artificial neural networks, and artificial immune systems. Artificial immune systems (AIS) represent a family of machine learning procedures that emulate the learning and memory capabilities of the biological immune system [1–9]. AIS are used with success in pattern recognition, function optimization, classification, process control, computer network security, and intrusion detection [10–12]. Artificial immune systems were successfully applied to biological and medical problems, such as classification of gene expression data [13–15], breast cancer identification [16,17], classification of liver disorders [16], detection of heart diseases [18], and diagnosis of thyroid diseases [19].

Several principles and mechanisms of the immune system were used by Watkins, Timmis, and Boggess to assemble an efficient machine learning algorithm, the artificial immune recognition system (AIRS) [20–22]. The extensive experiments performed by Brownlee demonstrate the AIRS capacity to solve classification problems [23]. We recently published the first application of the AIRS algorithm in modeling structure–activity relationships for drug design [24]. The learning task was to discriminate between drugs that induce torsade de pointes and drugs that do not induce torsade de pointes. In the present study we demonstrate the first AIRS application to the prediction of the human intestinal absorption (HIA) of drugs. Using a dataset of 196 drugs and 159 structural descriptors [25], AIRS is trained to discriminate between a subset of 131 drugs that penetrate the human intestine and a subset of 65 drugs that do not penetrate the intestine.

2 THE ARTIFICIAL IMMUNE RECOGNITION SYSTEM

AIRS uses the biological immune system as a source of inspiration in modeling complex systems [2,20,21]. However, the scope of AIRS is not to simulate the function of the biological immune system, but to provide an efficient machine learning procedure with general application in classification problems. Pathogens that attack an organism are identified and killed by various cell types that constitute the immune system. Proteins and protein fragments from pathogens bind to receptors situated on the surface of the B–cells and T–cells, thus signaling that foreign cells are present in the bloodstream. The recognition mechanism encoded into an antibody may be improved upon the presentation of several antigens with similar characteristics.

In the AIRS classification algorithm, an antigen is represented as an n–dimensional vector $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ and an associated class $Y = \{+1, -1\}$. Each structural descriptor x_i is a real number ($x_i \in R$ for $i = 1, 2, \ldots, n$) representing a numerical feature of the antigen. For quantitative structure–

activity relationships (QSAR), the **X** vector contains the structural descriptors for a molecule, whereas in the class variable *Y*, +1 encodes the presence of a property and –1 encodes the absence of that property. In the present study, this property is the penetration of the human intestine by a drug. An identical {**X**, *Y*} encoding is used for antibodies (the solutions for the classification problem). A B–cell is represented in the AIRS procedure by an artificial recognition ball (ARB). An ARB contains an antibody, a number of resources, and a stimulation value. The stimulation value measures the similarity between an ARB and an antigen. Each AIRS model has a limited number of resources, and ARBs compete for their allocation. Resources are removed from the least stimulated ARBs, and ARBs without resources are eliminated from the cell population. The ARB population is trained during several cycles of competition for limited resources. In each cycle of ARB training, the best ARB classifiers generate mutated clones that enhance the antigen recognition process, whereas the ARBs with insufficient resources are removed from the population. After training, the top ARB classifiers are selected as memory cells. Finally, the memory cells are used to classify novel antigens (patterns).

Detailed descriptions of the artificial immune recognition system may be found in the literature [20–23]. We present here only the most important characteristics of the AIRS procedure, in order to highlight the parameters that control its classification ability. The steps of the AIRS algorithm are summarized below:

(1) Initialization. The training data are normalized between 0 and 1. The Euclidean distance is computed for all pairs of antigens, and then the affinity is determined as the ratio between the distance and the maximum distance. The affinity threshold is computed as the average affinity for all antigens in the training set. The memory cell pool is populated with randomly selected antigens. At the end of the AIRS algorithm, the memory cell pool represents the recognition ARBs used as classifiers.

(2) Train for all Antigens

(2.1) Antigen Presentation. Each training antigen is presented to the memory cell pool, and each memory cell receives a stimulation value, stimulation = 1 – affinity. The memory cells with the highest stimulation are selected, and a number of mutated clones are created and added to the ARB pool. The number of clones generated is computed with the formula:

$$\text{NumberClones} = \text{Stimulation} \times \text{ClonalRate} \times \text{HypermutationRate} \quad (1)$$

where ClonalRate and HypermutationRate are user defined parameters.

(2.2) Competition for Limited Resources. The scope of this process is to select those ARBs that have the best recognition capabilities, while optimally allocating the resources to the best ARBs.

(2.2.1) Perform Competition for Resources

(2.2.1.1) Stimulate the ARB Pool with Antigen

(2.2.1.2) Normalize the ARB Stimulation Values

(2.2.1.3) Allocate Limited Resources Based on Stimulation. The amount of resources allocated to each ARB is:

$$\text{Resources} = \text{NormalizedStimulation} \times \text{ClonalRate} \quad (2)$$

(2.2.1.4) Remove ARBs with Insufficient Resources

(2.2.2) Continue with (2.3) if the Stop Condition is Satisfied. The stop condition for the ARB refinement is met when the average normalized stimulation is higher than a user defined stimulation threshold.

(2.2.3) Generate Mutated Clones of Surviving ARBs. The number of clones generated is:

$$\text{NumberClones} = \text{Stimulation} \times \text{ClonalRate} \quad (3)$$

(2.2.4) Go to (2.2.1)

(2.3) Memory Cell Selection. In this step, new ARB classifiers are evaluated for inclusion in the memory cell pool. An ARB is inserted in the memory cell pool if its stimulation value is better than that of the existing best matching memory cell. The existing best matching memory cell is then removed if the affinity between the candidate ARB and the existing memory cell is less than a CutOff value:

$$\text{CutOff} = \text{AffinityThreshold} \times \text{AffinityThresholdScalar} \quad (4)$$

where the AffinityThreshold was computed during the Initialization phase, and the AffinityThresholdScalar is a user defined parameter.

(3) Classification. The memory cell pool represents the AIRS classifier. The classification is performed with a *k*–nearest neighbor method, in which the *k* best matches to a prediction pattern are identified and the predicted class is determined with a majority vote.

3 MATERIALS AND METHODS

A good intestinal absorption is a major requirement for oral drugs [26–28], and various computational models were proposed as fast, reliable, and inexpensive *in silico* methods to assess the intestinal permeability of a chemical compound before synthesis [29–33]. The oral absorption of a drug is influence by a large number of variables, such as drug formulation and stability, aqueous solubility, contents of the gastrointestinal tract, residence time in the intestine, intestinal metabolism, rate of passive intestinal permeability, carrier–mediated influx, and active efflux *via* transporters [27]. A wide range of structural descriptors [34–39] and machine learning QSAR procedures [40–43] were explored in order to obtain HIA models with good prediction power.

In the present investigation we apply the AIRS classifier to a dataset of 196 drugs and 159 structural descriptors [25], in order to discriminate between a subset of 131 drugs that penetrate the

human intestine (HIA+) and a subset of 65 drugs that do not penetrate the intestine (HIA–). The chemical structure was encoded with 159 structural descriptors from five classes, namely constitutional, topological indices, electrotopological state indices, quantum descriptors, and geometrical indices. Eight user defined parameters influences the classification performance of the AIRS algorithm, namely affinity threshold scalar, clonal rate, hypermutation rate, number of nearest neighbors, initial memory cell pool size, number of instances to compute the affinity threshold, stimulation threshold, and total resources. In order to explore the AIRS sensitivity to these parameters, leave–20%–out (five–fold) cross–validation predictions were performed over a wide range of values for all eight parameters. All computations were performed with the AIRS2 implementation of Brownlee [23] using Weka 3.5.4 [44].

4 RESULTS AND DISCUSSION

We investigated the classification performance of AIRS2 over a large range of the eight user defined parameters, and for each AIRS model we report the following statistical indices: TP_c, true positive in calibration (number of HIA+ drugs classified as HIA+); FN_c, false negative in calibration (number of HIA+ drugs classified as HIA–); TN_c, true negative in calibration (number of HIA– drugs classified as HIA–); FP_c, false positive in calibration (number of HIA– drugs classified as HIA+); Se_c, calibration selectivity; Sp_c, calibration specificity; Ac_c, calibration accuracy; MCC_c, calibration Matthews correlation coefficient [45]; TP_p, true positive in prediction; FN_p, false negative in prediction; TN_p, true negative in prediction; FP_p, false positive in prediction; Se_p, prediction selectivity; Sp_p, prediction specificity; Ac_p, prediction accuracy; MCC_p, prediction Matthews correlation coefficient.

Affinity Threshold Scalar (ATS). The affinity threshold scalar is used in Eq. (4) to compute a cut–off value for memory cell replacement, and takes values between 0 and 1. If the affinity between a candidate ARB and the best matching memory cell is lower that the threshold computed with Eq. (4), then the ARB replaces the memory cell. A high ATS value results in a high replacement rate, whereas a low ATS value results in a low replacement rate. In experiments **1**–**27** (Table 1) we varied the ATS value between 0.01 and 0.95 in order to identify the optimum replacement regimen. The initial values for the remaining parameters are: clonal rate = 10, hypermutation rate = 2, number of nearest neighbors = 3, initial memory cell pool size = 50, number of instances to compute the affinity threshold = all, stimulation threshold = 0.5, and total resources = 150. These parameters are optimized in the above order, and the optimum value is used in all subsequent experiments. The highest prediction MCC = 0.3506 is obtained for ATS = 0.09, indicating that for the HIA classification problem a low memory cell replacement rate is beneficial. The prediction statistics decrease significantly when ATS increases over 0.1, suggesting that a high memory cell replacement rate results in poor AIRS models.

Table 1. AIRS Calibration and Prediction Statistics for Various Values of ATS (Affinity Threshold Scalar)

Exp	ATS	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
1	0.01	117	14	44	21	0.8931	0.6769	0.8214	0.5880
2	0.02	117	14	44	21	0.8931	0.6769	0.8214	0.5880
3	0.03	117	14	44	21	0.8931	0.6769	0.8214	0.5880
4	0.04	117	14	44	21	0.8931	0.6769	0.8214	0.5880
5	0.05	117	14	42	23	0.8931	0.6462	0.8112	0.5620
6	0.06	117	14	43	22	0.8931	0.6615	0.8163	0.5750
7	0.07	117	14	43	22	0.8931	0.6615	0.8163	0.5750
8	0.08	117	14	43	22	0.8931	0.6615	0.8163	0.5750
9	0.09	117	14	42	23	0.8931	0.6462	0.8112	0.5620
10	0.10	117	14	41	24	0.8931	0.6308	0.8061	0.5490
11	0.15	115	16	45	20	0.8779	0.6923	0.8163	0.5798
12	0.20	107	24	48	17	0.8168	0.7385	0.7908	0.5423
13	0.25	105	26	42	23	0.8015	0.6462	0.7500	0.4428
14	0.30	102	29	45	20	0.7786	0.6923	0.7500	0.4574
15	0.35	108	23	41	24	0.8244	0.6308	0.7602	0.4570
16	0.40	108	23	41	24	0.8244	0.6308	0.7602	0.4570
17	0.45	108	23	41	24	0.8244	0.6308	0.7602	0.4570
18	0.50	108	23	41	24	0.8244	0.6308	0.7602	0.4570
19	0.55	108	23	41	24	0.8244	0.6308	0.7602	0.4570
20	0.60	108	23	41	24	0.8244	0.6308	0.7602	0.4570
21	0.65	108	23	41	24	0.8244	0.6308	0.7602	0.4570
22	0.70	108	23	41	24	0.8244	0.6308	0.7602	0.4570
23	0.75	108	23	41	24	0.8244	0.6308	0.7602	0.4570
24	0.80	108	23	41	24	0.8244	0.6308	0.7602	0.4570
25	0.85	108	23	41	24	0.8244	0.6308	0.7602	0.4570
26	0.90	108	23	41	24	0.8244	0.6308	0.7602	0.4570
27	0.95	108	23	41	24	0.8244	0.6308	0.7602	0.4570

Exp	ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
1	0.01	105	26	33	32	0.8015	0.5077	0.7041	0.3174
2	0.02	105	26	33	32	0.8015	0.5077	0.7041	0.3174
3	0.03	103	28	33	32	0.7863	0.5077	0.6939	0.2989
4	0.04	107	24	33	32	0.8168	0.5077	0.7143	0.3364
5	0.05	107	24	33	32	0.8168	0.5077	0.7143	0.3364
6	0.06	106	25	33	32	0.8092	0.5077	0.7092	0.3268
7	0.07	104	27	35	30	0.7939	0.5385	0.7092	0.3365
8	0.08	106	25	34	31	0.8092	0.5231	0.7143	0.3410
9	0.09	107	24	34	31	0.8168	0.5231	0.7194	0.3506
10	0.10	107	24	33	32	0.8168	0.5077	0.7143	0.3364
11	0.15	107	24	28	37	0.8168	0.4308	0.6888	0.2640
12	0.20	107	24	26	39	0.8168	0.4000	0.6786	0.2341
13	0.25	99	32	30	35	0.7557	0.4615	0.6582	0.2200
14	0.30	100	31	30	35	0.7634	0.4615	0.6633	0.2287
15	0.35	103	28	28	37	0.7863	0.4308	0.6684	0.2262
16	0.40	103	28	25	40	0.7863	0.3846	0.6531	0.1811
17	0.45	105	26	25	40	0.8015	0.3846	0.6633	0.1997
18	0.50	105	26	25	40	0.8015	0.3846	0.6633	0.1997
19	0.55	105	26	25	40	0.8015	0.3846	0.6633	0.1997
20	0.60	105	26	25	40	0.8015	0.3846	0.6633	0.1997
21	0.65	105	26	25	40	0.8015	0.3846	0.6633	0.1997
22	0.70	105	26	25	40	0.8015	0.3846	0.6633	0.1997
23	0.75	105	26	25	40	0.8015	0.3846	0.6633	0.1997
24	0.80	105	26	25	40	0.8015	0.3846	0.6633	0.1997
25	0.85	105	26	25	40	0.8015	0.3846	0.6633	0.1997
26	0.90	105	26	25	40	0.8015	0.3846	0.6633	0.1997
27	0.95	105	26	25	40	0.8015	0.3846	0.6633	0.1997

Clonal Rate (CR). The clonal rate takes integer values, and is used in ARB resource allocation and in controlling the clonal mutation for the memory cell population. In Eq (1), CR is used to determine the number of mutated clones generated from each memory cell and then added to the ARB pool. In Eq. (2), CR is multiplied with the normalized stimulation of an ARB to determine the number of resources allocated to that ARB. The number of resources allocated to each ARB is in the range [0, CR]. In Eq. (3), CR is involved in the computation of the number of clones generated from each ARB during the ARB refinement process. Therefore, the number of ARB clones generated is in the range [0, CR].

In Table 2 we show the prediction statistics of the AIRS models obtained when the clonal rate was varied between 3 and 20 (experiments **28**–**37**). There is no apparent trend for the MCC values when CR increases. The best results, MCC = 0.3506, are obtained with CR = 10 and CR = 12, with no improvement over the best value obtained in the ATS experiments. CR = 10 is selected for further experiments.

Table 2. AIRS Calibration and Prediction Statistics for Various Values of CR (Clonal Rate); (ATS = 0.09)

Exp	CR	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
28	3	115	16	40	25	0.8779	0.6154	0.7908	0.5140
29	5	114	17	45	20	0.8702	0.6923	0.8112	0.5695
30	8	117	14	44	21	0.8931	0.6769	0.8214	0.5880
31	9	117	14	44	21	0.8931	0.6769	0.8214	0.5880
32	10	117	14	42	23	0.8931	0.6462	0.8112	0.5620
33	11	117	14	45	20	0.8931	0.6923	0.8265	0.6009
34	12	116	15	44	21	0.8855	0.6769	0.8163	0.5773
35	15	116	15	43	22	0.8855	0.6615	0.8112	0.5642
36	17	117	14	43	22	0.8931	0.6615	0.8163	0.5750
37	20	117	14	43	22	0.8931	0.6615	0.8163	0.5750

Exp	CR	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
28	3	111	20	28	37	0.8473	0.4308	0.7092	0.3045
29	5	103	28	32	33	0.7863	0.4923	0.6888	0.2846
30	8	105	26	28	37	0.8015	0.4308	0.6786	0.2448
31	9	106	25	27	38	0.8092	0.4154	0.6786	0.2394
32	10	107	24	34	31	0.8168	0.5231	0.7194	0.3506
33	11	103	28	35	30	0.7863	0.5385	0.7041	0.3273
34	12	107	24	34	31	0.8168	0.5231	0.7194	0.3506
35	15	105	26	35	30	0.8015	0.5385	0.7143	0.3457
36	17	102	29	34	31	0.7786	0.5231	0.6939	0.3041
37	20	106	25	34	31	0.8092	0.5231	0.7143	0.3410

Hypermutation Rate (HR). The hypermutation rate has integer values and is used in Eq. (1) to determine the number of clones for each memory cell, which is in the range [0, CR×HR]. We investigated the HIA classification for values of the hypermutation rate between 1 and 10, as shown in experiments **38**–**47** (Table 3). The best predictions are obtained with HR = 8, when the prediction MCC = 0.3691, which is a small improvement over the values from the previous groups of experiments.

Table 3. AIRS Calibration and Prediction Statistics for Various Values of HR (Hypermutation Rate); (CR = 10)

Exp	HR	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
38	1	110	21	45	20	0.8397	0.6923	0.7908	0.5300
39	2	117	14	42	23	0.8931	0.6462	0.8112	0.5620
40	3	117	14	43	22	0.8931	0.6615	0.8163	0.5750
41	4	117	14	44	21	0.8931	0.6769	0.8214	0.5880
42	5	117	14	43	22	0.8931	0.6615	0.8163	0.5750
43	6	117	14	44	21	0.8931	0.6769	0.8214	0.5880
44	7	117	14	43	22	0.8931	0.6615	0.8163	0.5750
45	8	117	14	44	21	0.8931	0.6769	0.8214	0.5880
46	9	117	14	45	20	0.8931	0.6923	0.8265	0.6009
47	10	117	14	42	23	0.8931	0.6462	0.8112	0.5620

Exp	HR	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
38	1	107	24	29	36	0.8168	0.4462	0.6939	0.2787
39	2	107	24	34	31	0.8168	0.5231	0.7194	0.3506
40	3	103	28	32	33	0.7863	0.4923	0.6888	0.2846
41	4	106	25	32	33	0.8092	0.4923	0.7041	0.3125
42	5	101	30	33	32	0.7710	0.5077	0.6837	0.2809
43	6	103	28	35	30	0.7863	0.5385	0.7041	0.3273
44	7	103	28	33	32	0.7863	0.5077	0.6939	0.2989
45	8	106	25	36	29	0.8092	0.5538	0.7245	0.3691
46	9	105	26	32	33	0.8015	0.4923	0.6990	0.3031
47	10	105	26	35	30	0.8015	0.5385	0.7143	0.3457

Number of Nearest Neighbors (kNN). The number *k* of nearest neighbors is used in the classification process, in which the *k* most stimulated memory cells to a given antigen vote for the class (HIA+ or HIA–) of that antigen. In Table 4 we show the prediction statistics of the AIRS models obtained when kNN takes values between 1 and 19 (experiments **48**–**57**). The best prediction (MCC = 0.3691) is obtained for kNN = 3, with no improvement compared with HR tests.

Table 4. AIRS Calibration and Prediction Statistics for Various kNN (Number of Nearest Neighbors); (HR = 8)

Exp	kNN	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
48	1	117	14	45	20	0.8931	0.6923	0.8265	0.6009
49	3	117	14	44	21	0.8931	0.6769	0.8214	0.5880
50	5	116	15	43	22	0.8855	0.6615	0.8112	0.5642
51	7	113	18	42	23	0.8626	0.6462	0.7908	0.5197
52	9	107	24	44	21	0.8168	0.6769	0.7704	0.4883
53	11	104	27	41	24	0.7939	0.6308	0.7398	0.4200
54	13	107	24	41	24	0.8168	0.6308	0.7551	0.4476
55	15	110	21	41	24	0.8397	0.6308	0.7704	0.4763
56	17	117	14	31	34	0.8931	0.4769	0.7551	0.4142
57	19	119	12	25	40	0.9084	0.3846	0.7347	0.3525

Exp	kNN	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
48	1	100	31	38	27	0.7634	0.5846	0.7041	0.3430
49	3	106	25	36	29	0.8092	0.5538	0.7245	0.3691
50	5	112	19	29	36	0.8550	0.4462	0.7194	0.3297
51	7	111	20	28	37	0.8473	0.4308	0.7092	0.3045
52	9	113	18	25	40	0.8626	0.3846	0.7041	0.2812
53	11	116	15	23	42	0.8855	0.3538	0.7092	0.2850
54	13	116	15	23	42	0.8855	0.3538	0.7092	0.2850
55	15	115	16	23	42	0.8779	0.3538	0.7041	0.2732
56	17	115	16	22	43	0.8779	0.3385	0.6990	0.2576
57	19	118	13	20	45	0.9008	0.3077	0.7041	0.2623

Initial Memory Cell Pool Size (IMCPS). The number of initial memory cells has a significant influence on the AIRS classification performance. The IMCPS parameter was modified between 1 and 100 (experiments **58**–**68**, Table 5), and the classification results show that when IMCPS < 30 the prediction statistics decrease. The best prediction MCC is obtained for IMCPS = 50, with no improvement compared with HR and kNN experiments.

Table 5. AIRS Calibration and Prediction Statistics for Various IMCPS (Initial Memory Cell Pool Size); (kNN = 3)

Exp	IMCPS	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
58	1	56	75	31	34	0.4275	0.4769	0.4439	–0.0903
59	10	96	35	35	30	0.7328	0.5385	0.6684	0.2666
60	20	87	44	49	16	0.6641	0.7538	0.6939	0.3941
61	30	107	24	47	18	0.8168	0.7231	0.7857	0.5288
62	40	106	25	47	18	0.8092	0.7231	0.7806	0.5198
63	50	117	14	44	21	0.8931	0.6769	0.8214	0.5880
64	60	115	16	46	19	0.8779	0.7077	0.8214	0.5928
65	70	113	18	46	19	0.8626	0.7077	0.8112	0.5725
66	80	114	17	46	19	0.8702	0.7077	0.8163	0.5826
67	90	113	18	47	18	0.8626	0.7231	0.8163	0.5857
68	100	115	16	48	17	0.8779	0.7385	0.8316	0.6188

Exp	IMCPS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
58	1	81	50	33	32	0.6183	0.5077	0.5816	0.1201
59	10	101	30	25	40	0.7710	0.3846	0.6429	0.1631
60	20	107	24	24	41	0.8168	0.3692	0.6684	0.2037
61	30	106	25	31	34	0.8092	0.4769	0.6990	0.2981
62	40	106	25	32	33	0.8092	0.4923	0.7041	0.3125
63	50	106	25	36	29	0.8092	0.5538	0.7245	0.3691
64	60	103	28	34	31	0.7863	0.5231	0.6990	0.3132
65	70	101	30	35	30	0.7710	0.5385	0.6939	0.3095
66	80	104	27	34	31	0.7939	0.5231	0.7041	0.3223
67	90	107	24	32	33	0.8168	0.4923	0.7092	0.3221
68	100	104	27	36	29	0.7939	0.5538	0.7143	0.3506

Number of Instances to Compute the Affinity Threshold (NIAT). NIAT specifies the number of antigens used to compute the affinity threshold in the AIRS initialization phase. In experiments **69**–**78** (Table 6) NIAT takes values between 10 and all training samples. The prediction accuracy is identical for NIAT = 40 and NIAT = all, but MCC is slightly higher for NIAT = 40, and we selected this value for further experiments.

Table 6. AIRS Calibration and Prediction Statistics for Various Values of NIAT (Number of Instances to Compute the Affinity Threshold); (IMCPS = 50)

Exp	NIAT	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
69	10	109	22	39	26	0.8321	0.6000	0.7551	0.4393
70	20	109	22	39	26	0.8321	0.6000	0.7551	0.4393
71	30	109	22	39	26	0.8321	0.6000	0.7551	0.4393
72	40	109	22	39	26	0.8321	0.6000	0.7551	0.4393
73	50	109	22	39	26	0.8321	0.6000	0.7551	0.4393
74	60	109	22	40	25	0.8321	0.6154	0.7602	0.4530
75	80	109	22	40	25	0.8321	0.6154	0.7602	0.4530
76	100	109	22	39	26	0.8321	0.6000	0.7551	0.4393
77	125	109	22	40	25	0.8321	0.6154	0.7602	0.4530
78	all	117	14	44	21	0.8931	0.6769	0.8214	0.5880

Table 6. (Continued)

Exp	NIAT	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
69	10	103	28	36	29	0.7863	0.5538	0.7092	0.3415
70	20	104	27	35	30	0.7939	0.5385	0.7092	0.3365
71	30	105	26	36	29	0.8015	0.5538	0.7194	0.3598
72	40	104	27	38	27	0.7939	0.5846	0.7245	0.3785
73	50	105	26	36	29	0.8015	0.5538	0.7194	0.3598
74	60	105	26	36	29	0.8015	0.5538	0.7194	0.3598
75	80	105	26	36	29	0.8015	0.5538	0.7194	0.3598
76	100	104	27	37	28	0.7939	0.5692	0.7194	0.3646
77	125	104	27	37	28	0.7939	0.5692	0.7194	0.3646
78	all	106	25	36	29	0.8092	0.5538	0.7245	0.3691

Stimulation Threshold (ST). The stimulation threshold is a parameter in the range [0, 1] and is used to determine the stop condition for the process of refining the ARB pool for a specific antigen. The ARB refinement stops when the average normalized ARB stimulation is higher than ST. The stimulation threshold was modified from 0.1 to 0.9 (experiments **79**–**93**, Table 7), and the best predictions were obtained for ST = 0.53 and ST = 0.55, with a small improvement compared with the NIAT tests. For further experiments we selected ST = 0.55.

Table 7. AIRS Calibration and Prediction Statistics for Various Values of ST (Stimulation Threshold); (NIAT = 40)

Exp	ST	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
79	0.10	109	22	39	26	0.8321	0.6000	0.7551	0.4393
80	0.20	109	22	39	26	0.8321	0.6000	0.7551	0.4393
81	0.30	109	22	39	26	0.8321	0.6000	0.7551	0.4393
82	0.40	109	22	39	26	0.8321	0.6000	0.7551	0.4393
83	0.45	109	22	39	26	0.8321	0.6000	0.7551	0.4393
84	0.47	109	22	39	26	0.8321	0.6000	0.7551	0.4393
85	0.49	109	22	39	26	0.8321	0.6000	0.7551	0.4393
86	0.50	109	22	39	26	0.8321	0.6000	0.7551	0.4393
87	0.51	109	22	39	26	0.8321	0.6000	0.7551	0.4393
88	0.53	109	22	39	26	0.8321	0.6000	0.7551	0.4393
89	0.55	109	22	41	24	0.8321	0.6308	0.7653	0.4666
90	0.60	109	22	39	26	0.8321	0.6000	0.7551	0.4393
91	0.70	110	21	39	26	0.8397	0.6000	0.7602	0.4492
92	0.80	110	21	38	27	0.8397	0.5846	0.7551	0.4355
93	0.90	110	21	39	26	0.8397	0.6000	0.7602	0.4492

Exp	ST	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
79	0.10	104	27	37	28	0.7939	0.5692	0.7194	0.3646
80	0.20	104	27	37	28	0.7939	0.5692	0.7194	0.3646
81	0.30	104	27	37	28	0.7939	0.5692	0.7194	0.3646
82	0.40	104	27	37	28	0.7939	0.5692	0.7194	0.3646
83	0.45	104	27	37	28	0.7939	0.5692	0.7194	0.3646
84	0.47	104	27	37	28	0.7939	0.5692	0.7194	0.3646
85	0.49	104	27	38	27	0.7939	0.5846	0.7245	0.3785
86	0.50	104	27	38	27	0.7939	0.5846	0.7245	0.3785
87	0.51	104	27	38	27	0.7939	0.5846	0.7245	0.3785
88	0.53	104	27	40	25	0.7939	0.6154	0.7347	0.4062
89	0.55	104	27	40	25	0.7939	0.6154	0.7347	0.4062
90	0.60	105	26	34	31	0.8015	0.5231	0.7092	0.3316
91	0.70	104	27	39	26	0.7939	0.6000	0.7296	0.3924
92	0.80	103	28	36	29	0.7863	0.5538	0.7092	0.3415
93	0.90	106	25	36	29	0.8092	0.5538	0.7245	0.3691

Total Resources (TR). The number of total resources limits the number of B–cells from the ARB pool. The amount of resources assigned to an ARB is calculated with Eq. (2) as a number in the range [0, CR]. Resources are allocated to the ARBs with high stimulation values, and taken from those with small stimulation values. ARBs without resources are removed from the cell population. The results obtained (Table 8, experiments **94**–**99**) show that for TR between 25 and 125 all prediction statistics are identical, and a slight increase is obtained for TR = 150. No improvement is obtained, compared with the ST tests.

Table 8. AIRS Calibration and Prediction Statistics for Various Values of TR (Total Resources); (ST = 0.55)

Exp	TR	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
94	25	109	22	39	26	0.8321	0.6000	0.7551	0.4393
95	50	109	22	39	26	0.8321	0.6000	0.7551	0.4393
96	75	109	22	39	26	0.8321	0.6000	0.7551	0.4393
97	100	109	22	39	26	0.8321	0.6000	0.7551	0.4393
98	125	109	22	39	26	0.8321	0.6000	0.7551	0.4393
99	150	109	22	41	24	0.8321	0.6308	0.7653	0.4666

Exp	TR	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
94	25	104	27	37	28	0.7939	0.5692	0.7194	0.3646
95	50	104	27	37	28	0.7939	0.5692	0.7194	0.3646
96	75	104	27	37	28	0.7939	0.5692	0.7194	0.3646
97	100	104	27	37	28	0.7939	0.5692	0.7194	0.3646
98	125	104	27	37	28	0.7939	0.5692	0.7194	0.3646
99	150	104	27	40	25	0.7939	0.6154	0.7347	0.4062

We presented the AIRS predictions for a large number of experiments, and we showed that this classifier is robust and offers stable predictions over a large range of the user defined parameters. Similarly with previous investigations [24], the largest variation of the prediction MCC was observed in the optimization of the affinity threshold scalar ATS. Only marginal improvement was obtained in the subsequent rounds, because the default values for these parameters were (near) optimal. Compared with SVM models obtained for the same dataset (MCC_p = 0.48) [25], the AIRS classifier is less successful in predicting the human intestinal absorption of drugs.

Table 9. Calibration and Prediction Statistics of Several Machine Learning Models

Exp	Model	TP_c	FN_c	TN_c	FP_c	Se_c	Sp_c	Ac_c	MCC_c
100	BayesNet	121	10	34	31	0.9237	0.5231	0.7908	0.5041
101	NaiveBayes	119	12	37	28	0.9084	0.5692	0.7959	0.5193
102	NaiveBayesUpdateable	129	2	28	37	0.9847	0.4308	0.8010	0.5433
103	Logistic	131	0	65	0	1.0000	1.0000	1.0000	1.0000
104	RBFNetwork	102	29	51	14	0.7786	0.7846	0.7806	0.5395
105	KStar	131	0	65	0	1.0000	1.0000	1.0000	1.0000
106	ADTree	124	7	58	7	0.9466	0.8923	0.9286	0.8389
107	J48	130	1	63	2	0.9924	0.9692	0.9847	0.9654
108	LMT	121	10	49	16	0.9237	0.7538	0.8673	0.6954
109	NBTree	131	0	56	9	1.0000	0.8615	0.9541	0.8979
110	RandomForest	131	0	65	0	1.0000	1.0000	1.0000	1.0000
111	RandomTree	131	0	65	0	1.0000	1.0000	1.0000	1.0000
112	REPTree	117	14	57	8	0.8931	0.8769	0.8878	0.7543

Table 9. (Continued)

Exp	Model	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
100	BayesNet	112	19	28	37	0.8550	0.4308	0.7143	0.3151
101	NaiveBayes	112	19	29	36	0.8550	0.4462	0.7194	0.3297
102	NaiveBayesUpdateable	120	11	26	39	0.9160	0.4000	0.7449	0.3802
103	Logistic	96	35	38	27	0.7328	0.5846	0.6837	0.3091
104	RBFNetwork	94	37	42	23	0.7176	0.6462	0.6939	0.3491
105	KStar	116	15	35	30	0.8855	0.5385	0.7704	0.4579
106	ADTree	113	18	42	23	0.8626	0.6462	0.7908	0.5197
107	J48	106	25	40	25	0.8092	0.6154	0.7449	0.4245
108	LMT	117	14	32	33	0.8931	0.4923	0.7602	0.4282
109	NBTree	108	23	35	30	0.8244	0.5385	0.7296	0.3743
110	RandomForest	114	17	33	32	0.8702	0.5077	0.7500	0.4082
111	RandomTree	98	33	35	30	0.7481	0.5385	0.6786	0.2834
112	REPTree	106	25	43	22	0.8092	0.6615	0.7602	0.4656

Comparison with other Machine Learning Algorithms. The same HIA+/HIA– classification problem was investigated with 13 other machine learning algorithms (Table 9, experiments **100**–**112**), namely Bayesian network (BayesNet), naïve Bayes classifier (NaiveBayes), updateable naïve Bayes classifier with kernel estimator (NaiveBayesUpdateable), logistic regression with ridge estimator (Logistic), Gaussian radial basis function network (RBFNetwork), K* instance–based classifier (KStar), alternating decision tree (ADTree), C4.5 decision tree (J48), logistic model trees (LMT), decision tree with naïve Bayes classifiers at the leaves (NBTree), random forest (RandomForest), random tree (RandomTree), fast decision tree learner (REPTree). All calculations were performed with Weka 3.5.4 [44], using all descriptors.

Six machine learning algorithms give better predictions than AIRS, namely ADTree (MCC_p = 0.5197), REPTree (MCC_p = 0.4656), KStar (MCC_p = 0.4579), LMT (MCC_p = 0.4282), J48 (MCC_p = 0.4245), and RandomForest (MCC_p = 0.4082), whereas AIRS is better than the remaining seven algorithms, namely NaiveBayesUpdateable, NBTree, RBFNetwork, NaiveBayes, BayesNet, Logistic, and RandomTree. These comparative tests indicate that the artificial immune recognition system surpasses several well–established machine learning algorithms, and ay be applied with success in HIA structure–activity studies.

5 CONCLUSIONS

A large number of machine learning algorithms are inspired from biological processes and mechanisms, such as particle swarm optimization, ant colony optimization, bee colony optimization, artificial neural networks, genetic algorithms, and DNA computing. Artificial immune systems represent a new class of biologically inspired algorithms that simulate elements of the biological immune system, such as pattern recognition, memory, and optimization. In this paper we present the first application of the artificial immune recognition system, AIRS, [20–22] to the modeling of the human intestinal absorption of drugs.

The AIRS algorithm was applied to the classification of a dataset of 196 drugs into a subset of 131 that penetrate the human intestine and a subset of 65 drugs that do not penetrate the intestine. All classifiers were trained with 159 structural descriptors from five classes, namely constitutional, topological indices, electrotopological state indices, quantum descriptors, and geometrical indices [25]. The calculations were performed with the AIRS2 algorithm [23] implemented in Weka [44]. The classification performance of the AIRS2 algorithm was investigated in 99 experiments and for a wide range of values for the eight user defined parameters, namely affinity threshold scalar, clonal rate, hypermutation rate, number of nearest neighbors, initial memory cell pool size, number of instances to compute the affinity threshold, stimulation threshold, and total resources. The HIA prediction ability was estimated with the leave–20%–out (five–fold) cross–validation.

The best predictions of the AIRS algorithm (selectivity 0.794, specificity 0.615, accuracy 0.735, and Matthews correlation coefficient 0.406) surpass those obtained with several well–established machine learning methods. The affinity threshold scalar has the highest influence on the prediction quality, whereas the remaining parameters have only a marginal effect. In a comparison with 13 machine learning algorithms, AIRS predictions were better in seven cases (Bayesian network, naïve Bayes classifier, updateable naïve Bayes classifier, logistic regression, Gaussian radial basis function network, decision tree with naïve Bayes classifiers at the leaves, and random tree). However, six machine learning algorithms have better predictions that AIRS (K* instance–based classifier, alternating decision tree, C4.5 decision tree, logistic model trees, random forest, and fast decision tree learner). The results obtained suggest that classifiers based on artificial immune systems may be successful in structure–activity relationships (SAR), quantitative structure–activity relationships (QSAR), virtual screening of chemical libraries, and drug design.

6 REFERENCES

[1] J. E. Hunt and D. E. Cooke, Learning using an artificial immune system, *J. Netw. Comput. Appl.* **1996**, *19*, 189–212.
[2] L. N. de Castro and F. J. Von Zuben, Artificial immune systems: Part I – Basic theory and applications. FEEC/UNICAMP, Brazil, 1999.
[3] L. N. de Castro and F. J. Von Zuben, Artificial immune systems: Part II – A survey of applications. FEEC/UNICAMP, Brazil, 2000.
[4] J. Timmis, M. Neal, and J. Hunt, An artificial immune system for data analysis, *BioSystems* **2000**, *55*, 143–150.
[5] L. N. de Castro and J. I. Timmis, Artificial immune systems as a novel soft computing paradigm, *Soft Comput.* **2003**, *7*, 526–544.
[6] L. N. De Castro, Dynamics of an artificial immune network, *J. Exp. Theor. Artif. Intell.* **2004**, *16*, 19–39.
[7] L. N. de Castro and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer–Verlag, Berlin, 2002.
[8] A. O. Tarakanov, V. A. Skormin, and S. P. Sokolova, *Immunocomputing: Principles and Applications*, Springer–Verlag, Berlin, 2003.
[9] Y. Ishida, *Immunity–Based Systems*, Springer–Verlag, Berlin, 2004.
[10] D. Dasgupta (Ed.), *Artificial Immune Systems and Their Applications*, Springer–Verlag, Berlin, 1999.
[11] G. Nicosia, V. Cutello, P. J. Bentley, and J. I. Timmis (Eds.), *Artificial Immune Systems: Third International Conference, ICARIS 2004, Catania, Sicily, Italy, September 13–16, 2004*, *Lecture Notes in Computer Science*, *Vol. 3239*, Springer–Verlag, Berlin, 2004.

[12] C. Jacob, M. L. Pilat, P. J. Bentley, and J. Timmis (Eds.), *Artificial Immune Systems: 4th International Conference, ICARIS 2005, Banff, Alberta, Canada, August 14–17, 2005*, *Lecture Notes in Computer Science*, *Vol. 3627*, Springer–Verlag, Berlin, 2005.
[13] S. Ando and H. Iba, Classification of gene expression profile using combinatory method of evolutionary computation and machine learning, *Genet. Programm. Evolv. Mach.* **2004**, *5*, 145–156.
[14] G. B. Bezerra, G. M. A. Cançado, M. Menossi, L. N. de Castro, and F. J. Von Zuben, Recent advances in gene expression data clustering: A case study with comparative results, *Genet. Mol. Res.* **2005**, *4*, 514–524.
[15] D. Tsankova, V. Georgieva, and N. Kasabov, Artificial immune networks as a paradigm for classification and profiling of gene expression data, *J. Comput. Theor. Nanosci.* **2005**, *2*, 543–550.
[16] K. Polat, S. Şahan, H. Kodaz, and S. Güneş, Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism, *Expert Syst. Appl.* **2007**, *32*, 172–183.
[17] S. Şahan, K. Polat, H. Kodaz, and S. Güneş, A new hybrid method based on fuzzy–artificial immune system and *k*–nn algorithm for breast cancer diagnosis, *Comput. Biol. Med.* **2007**, *37*, 415–423.
[18] K. Polat, S. Şahan, and S. Güneş, Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and *k*–nn (nearest neighbour) based weighting preprocessing, *Expert Syst. Appl.* **2007**, *32*, 625–631.
[19] K. Polat, S. Şahan, and S. Güneş, A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre–processing for thyroid disease diagnosis, *Expert Syst. Appl.* **2007**, *32*, 1141–1147.
[20] A. Watkins, J. Timmis, and L. Boggess, Artificial immune recognition system (AIRS): An immune–inspired supervised learning algorithm, *Genet. Programm. Evolv. Mach.* **2004**, *5*, 291–317.
[21] A. B. Watkins, AIRS: A resource limited artificial immune classifier. Department of Computer Science, MS Thesis, Mississippi State University, 2001, pp. 81.
[22] A. B. Watkins, Exploiting immunological metaphors in the development of serial, parallel and distributed learning algorithms. PhD Thesis, University of Kent, Canterbury, UK, 2005, pp. 314.
[23] J. Brownlee, Artificial immune recognition system (AIRS). A review and analysis. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT), Victoria, Australia, 2005.
[24] O. Ivanciuc, Artificial immune system classification of drug–induced torsade de pointes with AIRS (artificial immune recognition system), *Internet Electron. J. Mol. Des.* **2006**, *5*, 488–502, http://www.biochempress.com.
[25] Y. Xue, Z. R. Li, C. W. Yap, L. Z. Sun, X. Chen, and Y. Z. Chen, Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630–1638.
[26] P. Stenberg, K. Luthman, and P. Artursson, Virtual screening of intestinal drug permeability, *J. Control. Release* **2000**, *65*, 231–243.
[27] W. J. Egan and G. Lauri, Prediction of intestinal permeability, *Adv. Drug Deliv. Rev.* **2002**, *54*, 273–289.
[28] Y. M. Ponce, M. A. C. Pérez, V. R. Zaldivar, M. B. Sanz, D. S. Mota, and F. Torrens, Prediction of intestinal epithelial transport of drug in (Caco–2) cell culture from molecular structure using *in silico* approaches during early drug discovery, *Internet Electron. J. Mol. Des.* **2005**, *4*, 124–150.
[29] M. D. Wessel, P. C. Jurs, J. W. Tolan, and S. M. Muskal, Prediction of human intestinal absorption of drug compounds from molecular structure, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726–735.
[30] F. Yoshida and J. G. Topliss, QSAR model for drug human oral bioavailability, *J. Med. Chem.* **2000**, *43*, 2575–2585.
[31] D. E. Clark, Prediction of intestinal absorption and blood–brain barrier penetration by computational methods, *Comb. Chem. High Throughput Screen.* **2001**, *4*, 477–496.
[32] G. Klopman, L. R. Stefan, and R. D. Saiakhov, ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans, *Eur. J. Pharm. Sci.* **2002**, *17*, 253–263.
[33] E. Deretey, M. Feher, and J. M. Schmidt, Rapid prediction of human intestinal absorption, *Quant. Struct.–Act. Relat.* **2002**, *21*, 493–506.
[34] S. Winiwarter, N. M. Bonham, F. Ax, A. Hallberg, H. Lennernäs, and A. Karlén, Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach, *J. Med. Chem.* **1998**, *41*, 4939–4949.
[35] O. A. Raevsky, K.–J. Schaper, P. Artursson, and J. W. McFarland, A novel approach for prediction of intestinal absorption of drugs in humans based on hydrogen bond descriptors and structural similarity, *Quant. Struct.–Act. Relat.* **2002**, *20*, 402–413.
[36] S. Winiwarter, F. Ax, H. Lennernäs, A. Hallberg, C. Pettersson, and A. Karlén, Hydrogen bonding descriptors in the prediction of human in vivo intestinal permeability, *J. Mol. Graph. Modell.* **2003**, *21*, 273–287.
[37] T. Sanghvi, N. Ni, M. Mayersohn, and S. H. Yalkowsky, Predicting passive intestinal absorption using a single parameter, *QSAR Comb. Sci.* **2003**, *22*, 247–257.

[38] J. Linnankoski, J. M. Makela, V. P. Ranta, A. Urtti, and M. Yliperttula, Computational prediction of oral drug absorption based on absorption rate constants in humans, *J. Med. Chem.* **2006**, *49*, 3674–3681.
[39] E. Deconinck, H. Ates, N. Callebaut, E. Van Gyseghem, and Y. Vander Heyden, Evaluation of chromatographic descriptors for the prediction of gastro–intestinal absorption of drugs, *J. Chromatogr. A* **2007**, *1138*, 190–202.
[40] E. Deconinck, Q. S. Xu, R. Put, D. Coomans, D. L. Massart, and Y. Vander Heyden, Prediction of gastro–intestinal absorption using multivariate adaptive regression splines, *J. Pharm. Biomed. Anal.* **2005**, *39*, 1021–1030.
[41] E. Deconinck, T. Hancock, D. Coomans, D. L. Massart, and Y. Vander Heyden, Classification of drugs in absorption classes using the classification and regression trees (CART) methodology, *J. Pharm. Biomed. Anal.* **2005**, *39*, 91–103.
[42] E. Deconinck, D. Coomans, and Y. Vander Heyden, Exploration of linear modelling techniques and their combination with multivariate adaptive regression splines to predict gastro–intestinal absorption of drugs, *J. Pharm. Biomed. Anal.* **2007**, *43*, 119–130.
[43] M. Iyer, Y. J. Tseng, C. L. Senese, J. Liu, and A. J. Hopfinger, Prediction and mechanistic interpretation of human oral drug absorption using MI–QSAR analysis, *Mol. Pharmaceutics* **2007**, *4*, 218–231.
[44] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2 edn., Morgan Kaufmann, San Francisco, 2005.
[45] B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta* **1975**, *405*, 442–451.