UNCERTAINTY OF THE *IN VITRO* EXPERIMENTS IN THE CONSTRUCTION OF PREDICTIVE MODELS

Sabina Smusz[1], Wojciech Czarnecki[2], Dawid Warszycki[1], Andrzej J. Bojarski[1]

[1]Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland
[2]Faculty of Mathematics and Computer Sciences, Jagiellonian University, 6 Łojasiewicza Street, 30-342 Kraków, Poland
e-mail: smusz@if-pan.krakow.pl

Introduction

Molecular modeling methods, although abstract, are always drawing from experimental data, either as a very basis, upon which the model is constructed (e.g. pharmacophore models) or as a training/verification element (docking, machine learning) [1]. At some stage, they all use knowledge about the activity of a given group of compounds.

There is a number of databases providing quantitative information about the biological activity of chemical compounds, such as ChEMBL, PDSP and PubChem. However, due to the inconsistency of the results obtained in *in vitro* experiments, for some compounds there is more than one K_i (or equivalent) value provided (e.g. for cocaine, there are 815 different activity records (with differences occurring also within the same assays) stored in the ChEMBL database [2]).

This study introduces the uncertainty of the biological data as a parameter for the Support Vector Machine (SVM) in classification experiments.

Methods

Eight protein targets were selected for the study: serotonin receptors 5-HT_{2A}, 5-HT_{2C}, 5-HT_6 and 5-HT_7, muscarinic receptor M_1, histamine receptor H_1, HIV integrase (HIVi), and HIV protease (HIVp). Compounds with experimentally verified activity towards those were selected from the ChEMBL database. Only compounds, for which the activity was quantified in K_i or IC_{50} (it was assumed that $K_i = IC_{50}/2$) and which were tested in assays on human, rat cloned or native receptors, were taken into account (Figure 1A). Structures were considered active, when the median value of all K_i values provided for particular instance was lower than 100 nM and inactive when it was higher than 1000 nM. The numbers of active and inactive compounds for each target were gathered in Table 1. Four different fingerprints were used as representations, generated with the use of PaDEL-Descriptor (Table 2) [3].

Support Vector Machine (SVM) method was selected for machine learning experiments. The core concept behind SVM is to seek for the hyperplane (defined by its normal and distance from the origin), that separates the binary labeled data in such a way that the margin (sum of minimum distances from the hyperplane to the nearest data points of both classes) is maximized [4].

In order to incorporate the uncertainty measure to our problem, several weighting schemes were developed (Figure 1B):

a) standard classification

b) class weighting $c_i = 1 - \frac{N_j}{\sum_{j=1}^{2} N_j}$

c) weights linearly proportional to logarithm of K_i: $c_i = |\log_{10} K_i - 2.5|$

d) weights invertibly proportional to logarithm of K_i variance: $c_i = \frac{1}{\log_{10}(\operatorname{var}(K_i))+1}$

e) weights exponentially proportional to K_i variance: $c_i = \exp^{-\operatorname{var}(K_i)}$

f) – h) weighting schemes c) – e) combined with class weights (defined in b))

where:

c_i – weight assigned to particular example
N_j – number of compounds in particular class
K_i – median of K_i values provided for particular compound
$\operatorname{var}(K_i)$ – variance of K_i values provided for particular compound

Results

As expected, the introduction of information about the uncertainty of biological experiments affects the results of SVM classification (Figure 2). Out of 7 different weighting approaches, weights linearly proportional to K_i values turned out to provide the improvement of the results (both with and without inclusion of the class weights) for all tested compounds representations (the improvement in all cases was at the level of 1-2% in terms of MCC).

Simple class weighting, led to the enhancement of SVM performance in the majority of experiments only for KlekFP. The experiments have shown, that the exponential dependence on Ki variance does not affect the SVM experiments by any means, whereas the logarithmic dependence uplifted MCC only in some cases for MACCSFP and SubFP.

Conclusions

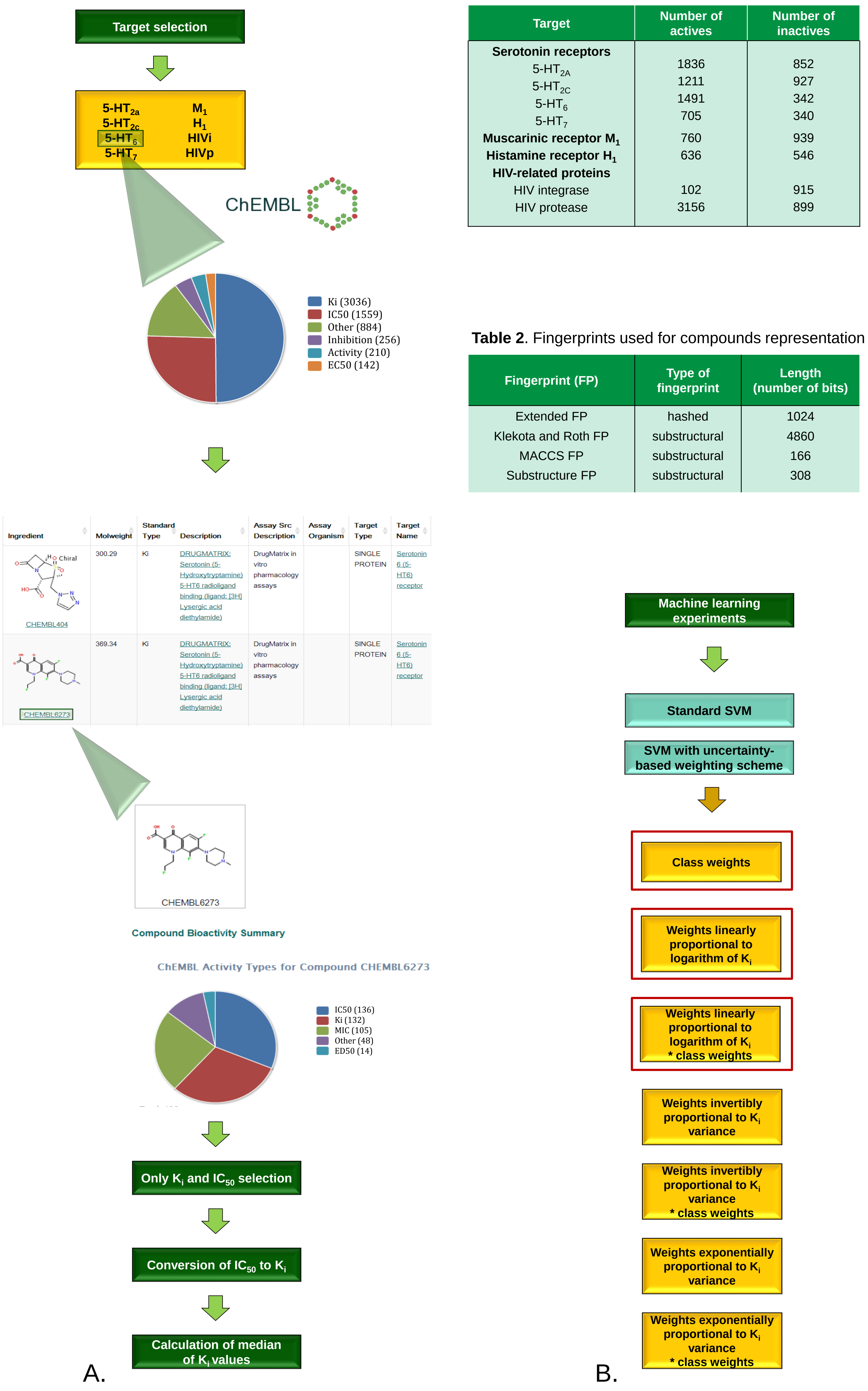
Uncertainty of the biological experiments is an important aspect of *in silico* research. However, as our experiments prove, the method of incorporating such knowledge also has an effect and should be carefully considered for any type of molecular modeling approach. Although the increase of MCC was not significant (1-2%), the effect is noticeable as so consitutes a valuable starting point for further optimization of the weighting protocol.

Table 1. Numbers of compounds used in the experiments

Target	Number of actives	Number of inactives
Serotonin receptors		
5-HT_{2A}	1836	852
5-HT_{2C}	1211	927
5-HT_6	1491	342
5-HT_7	705	340
Muscarinic receptor M_1	760	939
Histamine receptor H_1	636	546
HIV-related proteins		
HIV integrase	102	915
HIV protease	3156	899

Table 2. Fingerprints used for compounds representation

Fingerprint (FP)	Type of fingerprint	Length (number of bits)
Extended FP	hashed	1024
Klekota and Roth FP	substructural	4860
MACCS FP	substructural	166
Substructure FP	substructural	308


Target selection
5-HT_{2a}
5-HT_{2c}
5-HT_6
5-HT_7
M_1
H_1
HIVi
HIVp
ChEMBL
Ki (3036)
IC50 (1559)
Other (884)
Inhibition (256)
Activity (210)
EC50 (142)
CHEMBL6273
Compound Bioactivity Summary
ChEMBL Activity Types for Compound CHEMBL6273
IC50 (136)
Ki (132)
MIC (105)
Other (48)
ED50 (14)
Only K_i and IC_{50} selection
Conversion of IC_{50} to K_i
Calculation of median of K_i values
A.
Machine learning experiments
Standard SVM
SVM with uncertainty-based weighting scheme
Class weights
Weights linearly proportional to logarithm of K_i
Weights linearly proportional to logarithm of K_i * class weights
Weights invertibly proportional to K_i variance
Weights invertibly proportional to K_i variance * class weights
Weights exponentially proportional to K_i variance
Weights exponentially proportional to K_i variance * class weights
B.

Ingredient	Molweight	Standard Type	Description	Assay Src Description	Assay Organism	Target Type	Target Name
CHEMBL404	300.29	Ki	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT6 radioligand binding (ligand: [3H] Lysergic acid diethylamide)	DrugMatrix in vitro pharmacology assays		SINGLE PROTEIN	Serotonin 6 (5-HT6) receptor
CHEMBL6273	369.34	Ki	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT6 radioligand binding (ligand: [3H] Lysergic acid diethylamide)	DrugMatrix in vitro pharmacology assays		SINGLE PROTEIN	Serotonin 6 (5-HT6) receptor

Figure 1. Scheme of the work carried out within the study
A. Sets preparation
B. Machine learning experiments

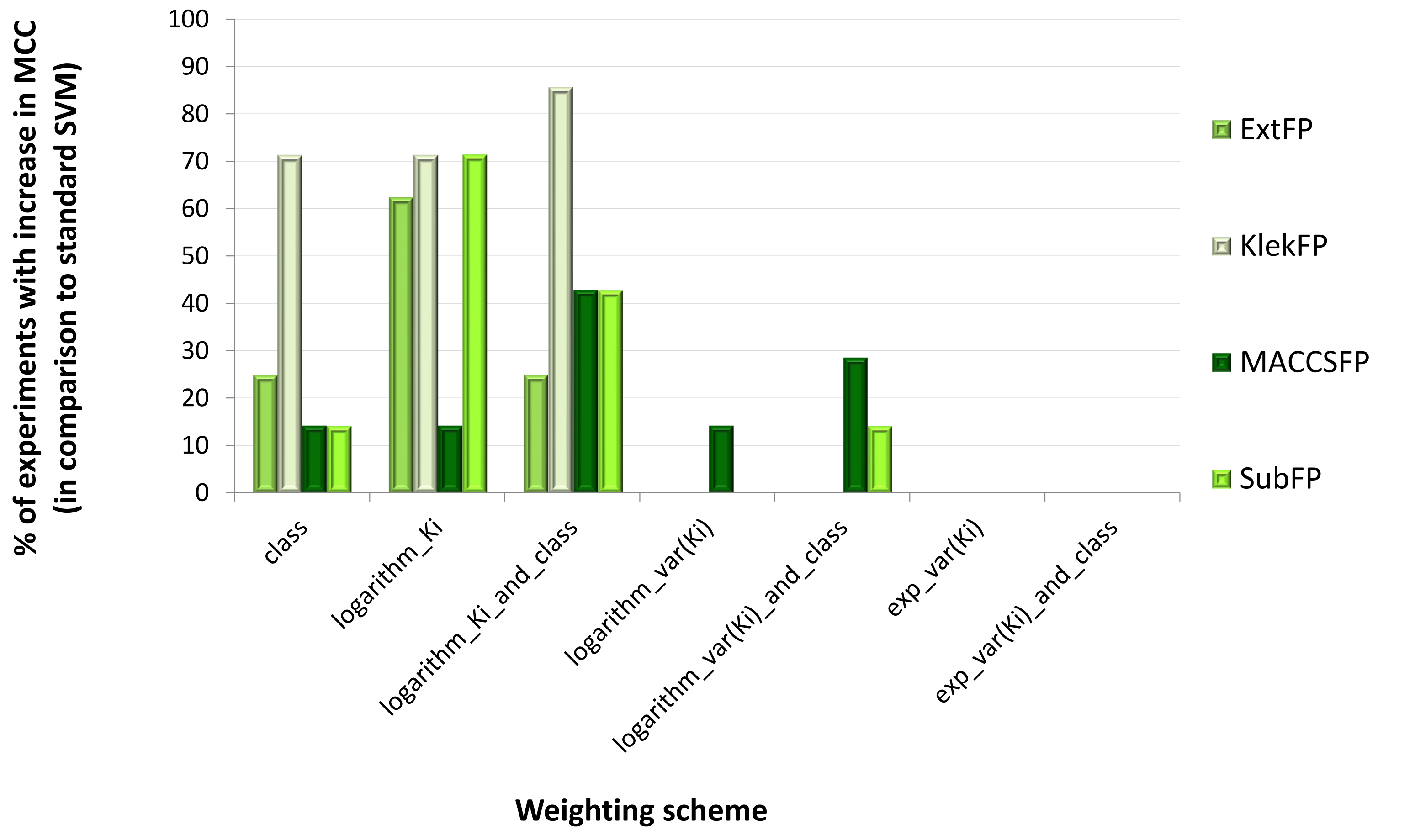


Figure 2. Comparison of the influence of adding information about the uncertainty of K_i values on the SVM performance.

References

[1] Reddy AS, Pati, SP, Kumar PP, Pradeep HN, Sastry GN Curr Prot Pept Sci, 2007, 8(4), 329–351.

[2] Gaulton A, Bellis LJ, Bento P, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al.-Lazikani B, Overington JP Nucleic Acids Res, 2011, 40(D1), D1100-D1107.

[3] Yap CWEI Journal of Computational Chemistry, 2010, 32(7), 1466-1474.

[4] Han LY, Ma XH, Lin HH, Jia J, Zhu F, Xue Y, Li ZR, Cao ZW, Ji ZL, Chen YZ J Mol Graph Model, 2008, 26(8), 1276–1286.

Acknowledgements

The study was partially supported by the Polish-Norwegian Research Programme operated by the National Centre for Research and Development under the Norwegian Financial Mechanism 2009-2014 in the frame of Project PLATFORMex (Pol-Nor/198887/73/2013 .