Classifying the World Anti-Doping Agency's 2005 Prohibited List Using the Chemistry Development Kit Fingerprint

Edward O. Cannon and John B. O. Mitchell*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

* Tel: +44-1223-762983; Fax: +44-1223-763076, e-mail: jbom1@cam.ac.uk

Abstract. We used the freely available Chemistry Development Kit (CDK) fingerprint to classify 5235 representative molecules taken from ten banned classes in the 2005 World Anti-Doping Agency's (WADA) prohibited list, including molecules taken from the corresponding activity classes in the MDL Drug Data Report (MDDR). We used both Random Forest and *k*-Nearest Neighbours (*k*NN) algorithms to generate classifiers. The *k*NN classifiers with *k* = 1 gave a very slightly better Matthews Correlation Coefficient than the Random Forest classifiers; the latter, however, predicted fewer false positives. The performance of *k*NN classifiers tended to decline with increasing *k*. The performance of the CDK fingerprint is essentially equivalent to that of Unity 2D. Our results suggest that it will be possible to use freely available chemoinformatics tools to aid the fight against drugs in sport, while minimising the risk of wrongfully penalising innocent athletes.

1. Introduction

Doping comes from the Dutch word "doop", meaning a thick liquid or sauce and originally a South African drink, drunk to help make an individual work harder. Here, we discuss illegal doping in sport, the objective of which is to enhance athletic performance, with little thought as to either the consequences for athlete's health or the integrity of competition. The issue of doping in sport is further complicated by a minefield of legal, political, and ethical questions. The urgency and importance of the battle against drugs in sport was underlined when several of the world's leading cyclists were forcibly withdrawn on the eve of the 2006 Tour de France, following an investigation by Spanish police.

The WADA[1] prohibited list contains 11 different classes of substance: one of these, alcohol (P1), has just one member and is not considered further. Anabolic agents (S1) are artificial synthetic analogues of the male sex hormone testosterone. They are used to promote growth of the skeletal muscles and red blood cells; particularly useful in events such as weightlifting or the 100m sprint, whereby these substances increase muscle size and strength allowing the athlete to train harder. Hormones and related substances (S2) include: erythropoietin, growth hormones, gonadotrophins, insulin and corticotrophins. These substances are taken by athletes to stimulate cell growth and red blood cell production and to increase sugar levels in the blood to avoid fatigue.

The primary medical use of beta-2 agonists (S3) is to treat asthmatic patients during an asthma attack. The drugs are used to open up the airways in the lungs which become restricted following an asthma attack. They are now being used in sport because if injected into the bloodstream they have a powerful anabolic effect that can cause muscle mass to increase and body fat to drop. Anti-estrogenic agents (S4) are substances that prevent the full expression of estrogen. Examples of anti-estrogens include tamoxifen and clomiphene.

Diuretics (S5), normally used to treat heart failure or high blood pressure, have been abused in sport for weight loss and elimination of drugs from the system. Diuretics work by increasing urine production in the kidneys. Sports where diuretics might be abused for promoting weight loss include boxing and lightweight rowing, and indeed any sports where competitors are required to reduce their body weight to below a specified level. Diuretics have been abused as masking agents to dilute the concentration of substances in the urine and avoid detection of other performance-enhancing drugs. Stimulants (S6) increase the activity of the sympathetic nervous system. Examples of stimulants include cocaine, amphetamine and modafinil; caffeine has recently been removed from the WADA prohibited list. These substances make the user feel more alert, energetic and able to concentrate. Narcotics (S7) enhance performance in sport by acting as pain killers. Narcotics allow an injured athlete to continue to train and compete by relieving pain. Examples of narcotics banned in sport include heroin, morphine and fentanyl.

Cannabinoids (S8) have been used to treat pain, migraine, insomnia, nausea and high blood pressure. They are used in sport to relax an athlete before competition. Glucocorticosteroids (S9) are now used as anti-inflammatory agents to treat arthritis and dermatitis. Examples of glucocorticosteroids include hydrocortisone and fludrocortisone acetate. Beta blockers (P2) act as performance-enhancing drugs by lowering the human heart rate and blood pressure, particularly useful in Olympic sports such as archery or shooting where the beta blockers provide more time for the athlete to aim in between heart beats. Examples of beta blockers include acebutolol, alprenolol, nadolol and atenolol.

The repertoire of substances used as doping agents in sport is continually evolving. This leads to an "arms race" between cheats and testers. The former are engaged in the design and synthesis of novel drugs, exemplified by "designer steroids"[2,3] such as tetrahydrogestrinone (THG), which has recently gained notoriety in track and field athletics. The WADA list of prohibited substances uses the phrase "and other substances with a similar chemical structure or similar biological effect(s)" to prohibit analogues of known performance-enhancing molecules. This is a very delicate area legally and ethically, since the authorities run the risk of criminalising athletes who ingest substances which are in some way "similar", without any hard evidence of bioactivity.

Prior to our work, interest in chemoinformatics approaches to drugs in sport appears to have been limited to the single study of Kontaxakis and Christodoulou,[4] devoted to

the prediction of chromatographic retention times of prohibited substances using an artificial neural network. Nonetheless, chemoinformatics may have an important role to play, since much of the discipline is built around, firstly, quantifying chemical similarity and, secondly, predicting bioactivity – exactly the two issues that are most relevant in the present context. In recent work,[5] we have built classifiers which can be used to predict whether a given molecule is likely to exhibit the bioactivity specific to any particular class of prohibited substances.

Our approach has a number of advantages, not least of which is putting the definition of chemical similarity on a quantitative (algorithmic) footing, which should be less vulnerable to legal challenge than a purely qualitative definition. It can also identify molecules unlikely to be bioactive and hence reduce the likelihood of athletes being unjustifiably penalised. We anticipate that in practice such classifiers would be used to complement, rather than replace, experimental methods such as assays.[3] Experimental methods would allow confirmation of the bioactivities suggested by chemoinformatics. The use of classifiers such as ours on large databases or libraries of molecules can help the authorities predict where in chemical space their opponents are likely to be sourcing the next (or even current) generation of designer drugs. This would be highly beneficial, since it seems almost certain that much drug abuse in sport involves bioactive substances that are not currently known to, and hence not specifically looked for by, the drug testing regime.

In this paper, we will demonstrate that the freely available CDK Fingerprint[6] can be used to generate excellent classifiers. This is part of the Chemistry Development Kit, described as "a freely available open-source Java library for Structural Chemo- and Bioinformatics".[7] This decouples the classifiers from the commercial fingerprints such as Unity 2D[8] and MACCS,[9] which had been the basis of the successful classifiers in our previous work.[5] We will show that Random Forest is particularly suitable for minimising false positives. For *k*NN classifiers, we will find that $k = 1$ is most successful. We will also consider the class-specific predictive abilities of our classifiers, which exhibit a fairly consistent pattern. We believe that our work facilitates the use of chemoinformatics in the fight against doping in sport.

2. Methods

2.1 Datasets

All methods were applied to a dataset of 5235 molecules, some derived directly from the prohibited list and others taken from activity classes in the MDDR database (Version 2003.1) corresponding to each WADA prohibited class of substance.[9] The use of MDDR molecules of the corresponding bioactivities was necessary since the number of explicitly named molecules in the WADA list is relatively low, and justified by the "similar chemical structure or similar biological effect(s)" criterion. Our dataset contained: 47 anabolic agents (S1), 272 hormones and related substances (S2), 367 beta-2 agonists (S3), 928 anti-estrogenic agents (S4), 995 diuretics and masking agents (S5), 804 stimulants (S6), 195 narcotics (S7), 995 cannabinoids (S8), 26 glucocorticosteroids (S9), 239 beta-blockers (P2) and 367 explicitly allowed substances.

2.2 Fingerprints

This work considers two fingerprints, the Chemical Development Kit (CDK) fingerprint and the Unity 2D fingerprint. The CDK fingerprint used in this work is modelled on the Daylight[10] fingerprint. It operates by running a breadth-first search starting at each atom in the molecule and produces a string representation of paths up to six atoms in length. The software is written in Java and uses the Java hashing function in combination with a pseudorandom number generator with a default range of 0-1023. The number indicates a position in a fingerprint of length 1024 bits that is set to 1, based on the paths computed for the molecule.

The Unity 2D fingerprint is composed of 992 feature bits. It is also similar to the Daylight fingerprint, the difference being that the Unity fingerprint segregates different path lengths into different regions of the fingerprint.[11] Unity was the best performing fingerprint in our recent work,[5] hence Unity provides an important benchmark.

This work is underpinned by the "Similar Property Principle", that molecules close together in the chemical space defined by our descriptors are likely to share similar properties (in this case bioactivities).

2.3 Classification

The two machine learning algorithms used in this work are *k*-Nearest Neighbours and Random Forest. These algorithms were run using R software.[12] In all cases the classification was performed in a binary fashion, such that a query molecule was either predicted to be part of a prohibited class under question or was not.

In our *k*-Nearest Neighbour (*k*NN) classifiers, the class of a query molecule is determined by the majority vote of the class labels (member or non-member) of its *k* nearest neighbours, according to Euclidean distance in descriptor space, with tied votes resolved randomly.

Random Forest[13] generates a forest of decision trees. At each node of each tree, a descriptor is chosen for branch splitting; this is not selected from the full set of available descriptors, but from a random subset of candidates. The parameter *mtry* indicates how many descriptors will be randomly selected as candidates at each node in the tree. Its default value was used in this work, defined as the square root of the number of bits in the fingerprint (rounded down to an integer). Hence for the 1024 bit CDK fingerprint *mtry* is taken as 32, and for the 992 bit Unity 2D fingerprint the default *mtry* is 31. For each tree, branches continue to be subdivided while the minimum number of observations in each leaf is no less than a pre-determined *nodesize* value. Branches are not pruned back. The Random Forest algorithm produces one output per molecule per tree. Each output classifies the molecule into either the category of member or that of non-member of a particular prohibited class. The outputs of the trees are aggregated using majority voting. We used 500 trees per Random Forest (*ntree* = 500).

We used fivefold cross-validation everywhere. This means that results for the Random Forest classifiers are based on five runs, each using a different 20% of the dataset as an independent test set, with the results being aggregated. Each molecule thus appears in exactly one of the five test sets (and exactly four of the five training sets). A similar procedure was used in the *k*NN work, with 20% of the molecules being predicted based on their nearest neighbours in the remaining 80%. This nearest neighbour prediction test was repeated five times on mutually exclusive test sets. Thus each molecule was predicted once, and the results aggregated.

2.4 Performance Measures

For each of the classifiers operating on each of the 10 prohibited classes, a 2×2 confusion matrix was generated, giving the numbers of:

- True positives (t_p), correctly classified members of the class;
- True negatives (t_n), correctly classified non-members;
- False positives (f_p), non-members misclassified as members;
- False negatives (f_n), members misclassified as non-members.

Since each classifier was run separately against each of the 10 WADA classes, a false positive could arise in two different ways. One is that a molecule from an incorrect class is predicted as positive, for instance a member of S1 being labelled as a member of P2. The other is that an explicitly allowed substance is predicted as a member of the WADA class under test. A given test molecule could be classified by our methods as belonging to any combination of the 10 classes (or none). The "correct" labels of our 5235 molecules are, however, unique with each molecule being assigned membership of either zero or one WADA class.

Using the numbers of true and false positives and negatives, we calculated a version of the Matthews Correlation Coefficient with a slight modification, which we introduced in recent work:[5]

$$MCC^* = \frac{t_p t_n - f_p f_n}{MAX[1, \sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}]} \ .$$

The modification involves the *MAX* function in the denominator and ensures that MCC^* is defined even if one of the four sums inside the square root is zero, a situation which may occur if no positives are predicted (and thus $MCC^* = 0$). Baldi *et al.*[14] have shown that the limiting value of the unmodified *MCC* as $(t_p + f_p)$ tends to zero is, as expected, zero. This may be considered by some a more mathematically elegant way of ensuring that the coefficient is defined; nonetheless our introduction of MCC^* provides a pragmatic solution. The possible range of MCC^* values is from -1 (perfect anticorrelation), through 0 (random performance) to +1 (perfect correlation).

3. Results and Discussion

The Random Forest results are illustrated in Fig. 1. The levels of performance obtained with CDK and Unity are almost identical, both overall and across the individual classes. Unity does a little better on S1, but conversely CDK is superior in classifying S9. Comparison of the first two lines of Table 1 shows that the overall MCC^*s, with t_p, t_n, f_p and f_n aggregated over the ten classes, of the two Random Forest classifiers are virtually identical (MCC^* is 0.8143 for Unity and 0.8136 for CDK). The principal purpose of Unity's inclusion here is comparison with the new results for the freely available CDK fingerprint; the results for Unity are naturally very close to those we obtained in previous work[5] on an extremely similar dataset. In that work, Unity was shown to perform better than four other fingerprint definitions in classifying prohibited substances.

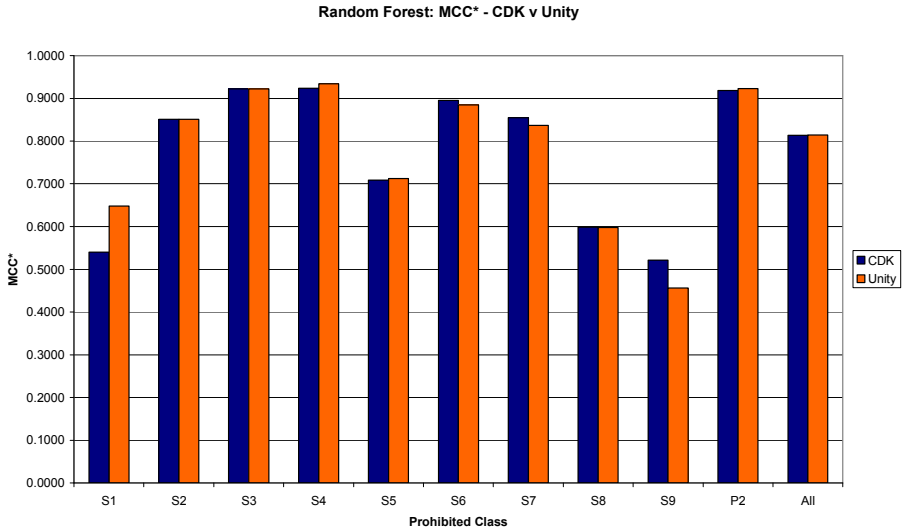

Fig. 1. MCC^* values obtained for each prohibited class using Random Forest classifiers based on the CDK (left hand side of each pair of bars) and Unity (right) fingerprints.

The Random Forest classifiers have the very useful property of predicting very few false positives. For Unity, only 147 false positives are predicted; this amounts to only 0.3% of the individual class assignments that ought correctly to be negatives (0.35% for CDK). Even considering that there are ten possible banned classes that a molecule could be assigned to, these figures suggest that the overall probability of an inactive molecule being wrongly classified as a positive by this Random Forest classifier is approximately 3% for Unity (approximately 3.5% for CDK). The false positive rate would be further reduced by combining the chemoinformatics approach with suitable assays.[3] Minimising false positives is important for legal reasons, and for the credibility and integrity of the anti-doping process; wrongful disqualification of athletes is to be avoided so far as possible.

Table 1. Performance of the classifiers aggregated across all ten prohibited classes.

FP	Method	t_p	t_n	f_p	f_n	MCC^*
CDK	RF	3493	47318	164	1375	0.8136
Unity	RF	3482	47335	147	1386	0.8143
CDK	1NN	4091	46763	719	777	0.8297
Unity	1NN	4124	46766	716	744	0.8342
CDK	3NN	3813	46792	690	1055	0.7962
Unity	3NN	3833	46808	674	1035	0.8005
CDK	5NN	3592	46799	683	1276	0.7673
Unity	5NN	3605	46792	690	1263	0.7683
CDK	10NN	3098	46876	606	1770	0.7063
Unity	10NN	3193	46783	699	1675	0.7098
CDK	20NN	2665	46972	510	2203	0.6530
Unity	20NN	2688	46895	587	2180	0.6474

Fig. 2 shows the performance of the *k*NN classifiers with $k = 1$, which we shall call 1NN classifiers (those with $k = 3$ are called 3NN classifiers *etc.*). The performance of CDK is very similar to that of Unity, except that it fares less well on class S1. The overall MCC^* values, shown in Table 1, are very similar for 1NN and Random Forest classifiers. In fact, 1NN achieves a slightly higher value than Random Forest in each case. Unity does very marginally better than CDK. An important difference is that, despite the very similar MCC^* values, the 1NN classifiers predict many more false positives, but fewer false negatives, than Random Forest (Table 1). As a consequence, 1NN gives a higher recall but lower precision for positives. This is true for both CDK and Unity fingerprints.

This illustrates the point that *k*NN generates models which are local in nature, with the class membership of a test molecule being predicted based on a very small number of its neighbours. This is especially true for the 1NN models. We believe that this makes the *k*NN method especially suitable for identifying members of those classes which correspond to several different clusters in chemical space. This is likely to occur when interaction with any one of a plurality of receptors can give rise to the specified bioactivity.

The four classifiers generated from Unity and CDK, Random Forest and 1NN (illustrated in Fig. 1 and Fig. 2) are in excellent agreement about the relative degrees of difficulty of predicting the ten prohibited classes. The six independent correlation coefficients between their sets of class-specific *MCC** values are all in the range r = 0.9154 (Unity-RF *vs* CDK-1NN) to r = 0.9906 (Unity-RF *vs* Unity-1NN). Although the

smallest classes, S1 and S9, are amongst the hardest to predict, there is only a weak relationship between class size and *MCC**. The overall consensus ranking of the classes, in decreasing order of prediction quality, is:

S3 ≈ S4 ≈ P2 > S2 ≈ S6 ≈ S7 > S5 > S1 ≈ S8 ≈ S9.

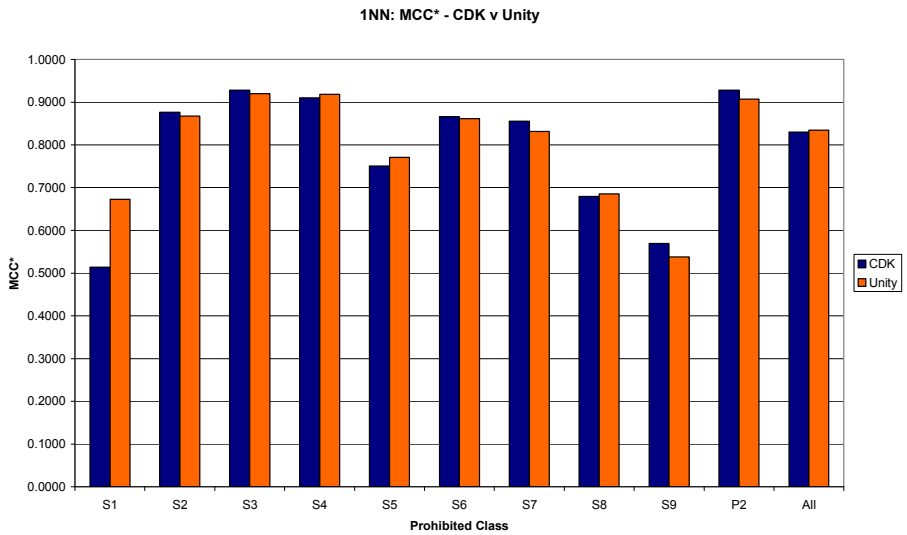

Fig. 2. *MCC** values obtained for each prohibited class using 1NN classifiers based on the CDK (left hand side of each pair of bars) and Unity (right) fingerprints**.**

We have also evaluated (Fig. 3 and Table 1) the performance of the *k*NN classifiers for higher values of *k*, for both the CDK and Unity fingerprints. There are two salient features of these results. Firstly, the *MCC** values tend to deteriorate for higher values of *k*. Secondly, the differences in performance between the two fingerprints are tiny. The fall-off with increasing *k* reinforces the local nature of the successful *k*NN models. For these data, at least, inclusion of additional neighbours generally reduces the *MCC** obtained. This indicates that the potential benefit of having information from more molecules is outweighed by the fact that these extra molecules are further away from that being classified. Fig. 3 contains some information additional to that in Table 1, in particular the inclusion of $k = 2$, $k = 4$ and $k = 15$. The slight recovery between $k = 2$ and $k = 3$ may be related to the random resolution of ties in the $k = 2$ case. This mirrors the observation, in a rather different field, by Lam and Suen[15] that augmenting an odd number of classifiers by an additional one can have a deleterious effect on overall prediction quality. Having an odd number of voters for a binary classification problem is an obvious way of avoiding problems with tied votes.

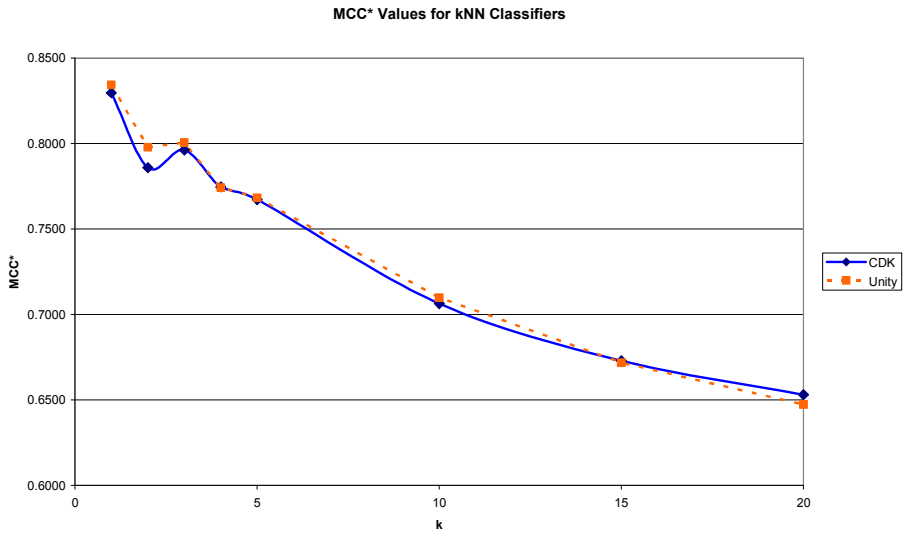

Fig. 3. *MCC** values obtained by *k*NN classifiers as a function of *k* for CDK (solid line) and Unity (broken line) fingerprints; based on results aggregated over all ten classes.

4. Conclusions

We have successfully categorised molecules into WADA prohibited classes using both Random Forest and *k*-Nearest Neighbours algorithms, with Matthews Correlation Coefficients above 0.8. In addition, we have shown that the freely available CDK fingerprint performs almost exactly as well as Unity 2D, which we previously demonstrated to be the best of five commercial fingerprints for this purpose. Although the 1NN algorithm (*k*NN with $k = 1$) gives the slightly higher MCC^*, the Random Forest classifier produces fewer false positives. Our best Random Forest models have a false positive rate, aggregated over all classes, of around 3%. The relative prediction accuracies of the different prohibited classes are very similar for the four different classifiers comprising Random Forest and 1NN algorithms with Unity and CDK fingerprints.

We find that 1NN is clearly the best *k*NN model for both fingerprints. The use of 2NN models is problematic due to the occurrence of tied votes, which are then resolved randomly. We favour the use of odd numbers of votes in classification problems of this kind. We also argue that the highly local nature of our 1NN models makes them particularly suitable for assigning molecules to classes of prohibited substances which comprise more than one cluster in chemical space.

These results suggest that it will be possible to create chemoinformatics-based classifiers, using freely available software, to determine whether novel molecules should be assigned to WADA prohibited classes. This will be especially powerful in combina-

tion with complementary experimental methods. Such tools will aid the fight against drug abuse in sport, while protecting competitors against unjustified sanctions.

Acknowledgements

We thank Unilever plc and the EPSRC for funding.

References

[1] World Anti-Doping Agency (WADA), Stock Exchange Tower, 800 Place Victoria, (Suite 1700), P.O. Box 120, Montreal, Quebec, H4Z 1B7, Canada; http://www.wada-ama.org/

[2] Handelsman, D. J., Designer Androgens in Sport: When too Much is Never Enough. *Sci. STKE* **2004**, *Issue 244,* pp. pe41.

[3] Death, A. K.; McGrath, K. C. Y.; Kazlauskas, R.; Handelsman, D. J., Tetrahydrogestrinone is a Potent Androgen and Progestin. *J. Clin. Endocrinol. Metab*. **2004**, *89*, 2498-2500.

[4] Kontaxakis, S. G.; Christodoulou, M. A., A Neural Network System for Doping Detection in Athletes. *Proceedings 4th International Conference on Technology and Automation,* Thessaloniki, Greece, October **2002**.

[5] Cannon, E. O.; Bender, A.; Palmer, D. S.; Mitchell, J. B. O., Chemoinformatics-based Classification of Prohibited Substances Employed for Doping in Sport. *J. Chem. Inf. Model.,* **submitted**.

[6] http://cdk.sourceforge.net/api/

[7] Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E., The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 493-500.

[8] Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319, USA; http://www.tripos.com

[9] Elsevier MDL, 2440 Camino Ramon, San Ramon, CA 94583, USA; http://www.mdli.com

[10] Daylight Chemical Information Systems, Inc. 120 Vantis - Suite 550 - Aliso Viejo, CA 92656, USA; http://www.daylight.com/

[11] Wild, D.; Blankley, C. J., Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf. Comput. Sci.,* **2000***, 40*, 155 -162.

[12] R Development Core Team (2005). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; ISBN 3-900051-07-0; http://www.R-project.org.

[13] Breiman, L., Random Forests. *Machine Learning*, **2001**, *45*, 5-32.

[14] Baldi, P; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H., Assessing the Accuracy of Prediction Algorithms for Classification: an overview. *Bioinformatics* **2000**, *16*, 412-424.

[15] Lam, L.; Suen, C. Y., Application of Majority Voting to Pattern Recognition: An Analysis of its Behavior and Performance. *IEEE Trans Systems, Man and Cybernetics* **1997**, *27*, 553-567.