# Prediction of Solvents Suitable for Crystallization of Small Organic Molecules

Kozo HOSOKAWA,[a] Junichi GOTO,[a] and Noriaki HIRAYAMA*,[b]

[a] *Computational Science Department, Science & Technology Systems Division, Ryoka Systems Inc.; 1–5–2 Irifune, Urayasu, Chiba 279–0012, Japan: and* [b] *Tokai University School of Medicine, Boseidai; Isehara, Kanagawa 259–1193, Japan.* Received June 13, 2005; accepted July 13, 2005; published online July 15, 2005

**Selection of suitable solvent is essential for crystallization of pharmaceuticals. Based on chemical structures of 6397 compounds and 15 single solvents that were used to obtain their single crystals, correlations between the molecular characteristics and the solvents have been investigated by cheminformatics methods. Decision-tree and Bayesian-probability methods have been applied to make classification models. These two models are complementary in character in the present case. It has been proven that the prediction of the solvent rankings for particular compounds by use of the classification models is satisfactory from the practical point of view. The present study has demonstrated that cheminformatics methods could greatly help rational crystallization of small organic molecules such as pharmaceuticals.**

**Key words** crystallization; pharmaceutical; small organic molecule; cheminformatics; molecular descriptor

Crystallization is one of the oldest and most important unit operations to purify molecules. Although several useful chromatographic techniques are available now, crystallization is still widely applied in pharmaceutical industries to separate or purify compounds. For example, it is known that more than a half of chiral pharmaceuticals and their intermediates in market are produced by crystallization techniques.[1)] Since different polymorphs have usually quite different physical properties, it is also highly required to obtain a suitable polymorph for a drug compound to make best use of it.[2)]

However, it has long accepted the fact that nature governs crystallization and we have very little control on crystallization. Since crystallization has been the most serious problem for protein crystallography, quite a lot of efforts have been made to find out the ways to undertake crystallization in more systematic and rather rational ways. Therefore automation and high throughput techniques are now prevalent on the side of protein crystallization.[3)] On the side of crystallization of small organic molecules, however, such attention has not been paid and crystallization is still a matter of trial and error.

A tremendous number of compounds have been crystallized so far. It is highly desirable to utilize these accumulated experiences in order to improve crystallization efficacy. To our best knowledge, however, no study has been made to put these experiences to good use. A number of factors such as solvent, temperature, pH, concentration, vibration and others affect crystallization. Among these factors, solvent is the most easily controlled variable. Hence usually crystallization begins with selection of solvents for crystallization. Crystallographic data of small organic molecules are compiled into the Cambridge Structural Database (CSD).[4)] The information regarding solvents or solvent systems successfully used to obtain single crystals of pertinent molecules is also described in the database.

Molecular descriptors calculated from two-dimensional chemical structures (hereafter 2D descriptors)[5)] are easily calculated and can represent characteristics of molecules reasonably well. In this study, 2D descriptors have been employed to investigate the correlations between chemical characteristics of molecules and solvents used to obtain their single crystals. From the practical points of view, we must select a few most promising solvents to start crystallization experiments. We have applied classification methods to predict priority rank of solvents suitable for crystallization of particular molecules.

## Experimental

Unfortunately crystallization solvents are not described for all compounds registered in CSD. If we eliminate compounds crystallized from mixed solvent systems, only 6397 compounds crystallized from 15 single solvents are remained for our study. All these 15 solvents given in Table 1 are popular ones used for crystallization of small organic molecules. A training dataset was constructed using these 6397 compounds.

Crystallization is a very complex process. Hence our data and knowledge about the process is incomplete, indirect, and noisy. Even under these difficult situations, data mining techniques are useful to make reasoning and help our decision-making. Decision-tree (C4.5)[6)] and Bayesian-probability[7)] are common methods used in data mining. Decision-tree is a predictive model that is a mapping of observations about an item to conclusions about the item's target value. Bayesian probability method can provide a formal and consistent way to reasoning in the presence of uncertainty. Therefore decision-tree and Bayesian probability method are applied in this study.

Classification models based on the values of 2D descriptors of compounds in the training dataset were made in order to apply these methods. 15 solvents are used as target values for classification. The object of the present study is to predict the priority rankings of 15 solvents for a particular compound. To accomplish this object, classification models were made so as to discriminate a specific solvent from others. Therefore classification models were made for 15 different solvents.

2D descriptors are numerical properties that can be calculated from the

Table 1. Fifteen Organic Solvents Used in Crystallization of 6397 Compounds

| | Solvent | Number of compounds | | Solvent | Number of compounds |
|---|---|---|---|---|---|
| 1 | Ethanol | 1328 | 9 | Chloroform | 342 |
| 2 | Methanol | 1030 | 10 | Toluene | 304 |
| 3 | Hexane | 821 | 11 | Benzene | 174 |
| 4 | Ethyl acetate | 573 | 12 | Water | 140 |
| 5 | Dichloromethane | 402 | 13 | DMF | 79 |
| 6 | Acetone | 394 | 14 | Cyclohexane | 56 |
| 7 | Acetonitrile | 379 | 15 | DMSO | 29 |
| 8 | Diethyl ether | 346 | | | |

* To whom correspondence should be addressed. e-mail: hirayama@is.icc.u-tokai.ac.jp

connection table representation of a molecule. Therefore 2D descriptor is not dependent on the conformation of a molecule and suitable for this study. A software package MOE (Molecular Operating Environment)[8] prepares various 2D descriptors to express physical properties, subdivided surface areas, atom counts, bond counts, Kier and Hall connectivity indices, kappa shape indices, adjacency and distance matrices, pharmacophore features, and partial charge. We have calculated the values of these 133 2D descriptors available in MOE for the compounds in the training dataset.

To handle and analyze high volume data effectively and flexibly we have chosen a software system KDE (Knowledge Discovery Environment)[9] as a data-mining tool.

The decision-tree classification models were created by KDE using default parameter set, and no cross-validation is followed as manual optimization and pruning of molecular descriptors are required. Bayesian-probability models were created by MOE. For each model, an optimal set of descriptors were selected to maximize prediction and cross-validation accuracy.

Ranks of confidence values for predicted solvents were used to express priority ranks. In decision-tree model, confidence values were re-calculated according to the overall population of true positive, false positive, true negative, and false negative. The confidence value for the predicted solvent was multiplied by TP/(TP+FP), and that for predicted otherwise was calculated as (1-confidence)*TN/(TN+FN). TP, FP, TN, and FN are the count of true positives, false positives, true negatives and false negatives, respectively. In Bayesian-probability model, the confidence value was calculated as a probability that the compound is crystallized from the specified solvent. Before ranking solvents, confidence values from both models were normalized to have unit sum for each target compound.

## Results and Discussion

Predicted priority ranks of 15 solvents are illustrated in Fig. 1. If predictions are perfect, the predicted ranks for 15 solvents should be 1. Predicted ranks in this figure, however, distribute in certain ranges because the predictions are not accurate enough. In the case of ethanol the ranks from decision-tree model distribute from 1 to 11, but the average rank is almost 1. It means that ethanol is almost perfectly predicted as a crystallizing solvent from the chemical structures of compounds. The results by two classification models for cyclohexane, DMF and DMSO are significantly different. The numbers of compounds crystallized from these solvents are relatively small. For other solvents, however, these two models gave similar ranking results. It is obvious that well-trained solvents are predicted precisely. Although the decision-tree model predicts generally better than the Bayesian probability model, the latter model seems to be good at prediction of solvents such as cyclohexane, DMF and DMSO. It indicates that the two models have rather complementary characteristics.

The enrichment curve for acetone is given in Fig. 2. If all compounds crystallized from acetone are predicted perfectly, the sampled percentage of 100% can be attained at the count number of 394 that is the number of compounds crystallized from acetone. In reality, however, the decision-tree model indicates that the rankings of about 1500 compounds should be considered to cover 80% of compounds that were crystallized from acetone. In this case both models predict similarly. The shape of this curve is typical for solvents with moderate number of data. The decision-tree model is generally superior to the Bayesian model. The curve of decision-tree model approaches the best prediction curve for the solvents with many data. For the solvents with fewer data, however, the Bayesian model approaches the best prediction curve.

Ranking histograms obtained by the two classification models are shown in Fig. 3. The abscissa axis gives the predicted ranking of solvents. A histogram bar at a certain rank gives the number of compounds whose crystallization solvent was correctly determined from the predicted solvent with the ranking number. For example, solvents for nearly 4500 compounds were correctly determined from the first-ranking solvent predicted by decision-tree model. This figure shows that suitable crystallization solvents for most compounds can be obtained from a few high-ranking solvents indicated by prediction. The decision-tree model gives generally higher count, namely better prediction, than the Bayesian model. For nearly 70% compounds, the solvents ranked first

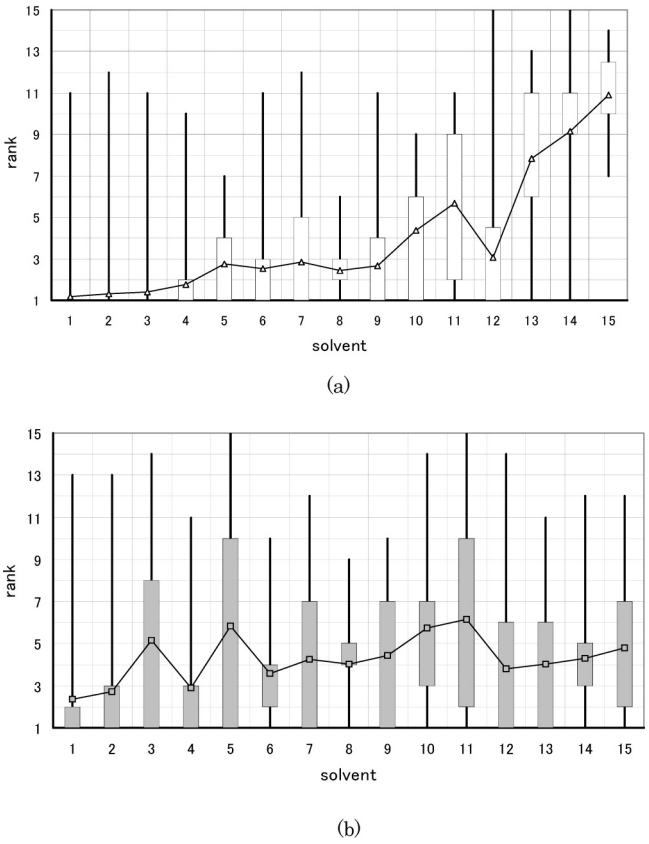


Fig. 1. Ranking Distribution for Each Solvent Predicted by Decision-Tree Model (a) and Bayesian-Probability Model (b)

The numbers of solvents correspond to those in Table 1. The vertical lines designate the ranges of minimum and maximum ranks, the boxes the ranges between 25 and 75 percentiles. The small triangular and rectangular marks designate their averages.

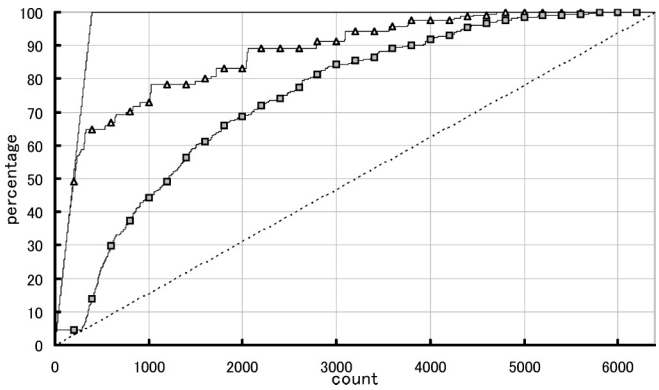


Fig. 2. Enrichment Curves for Acetone

The vertical axis is a sampled percentage of total entries of acetone. Solid and dotted lines show the cases of ideal and random samplings, respectively. The lines connected by small triangular and rectangular marks show predicted results by decision-tree and Bayesian-probability methods, respectively.

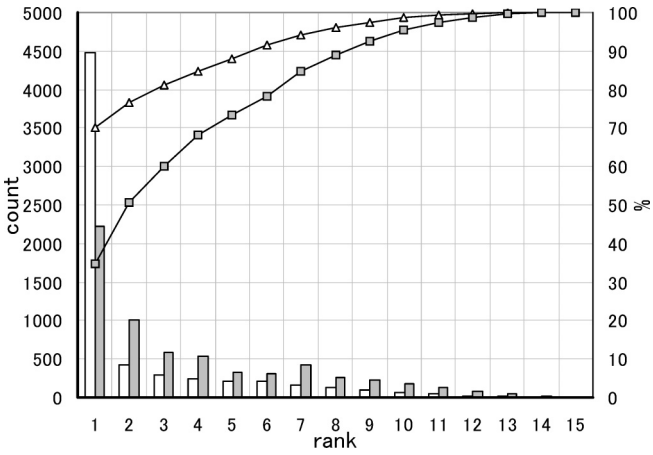


Fig. 3. Histogram of Counts (Bar) and Accumulated Percentages (Line) of Predicted Solvents

The results obtained by decision-tree and Bayesian-probability models are shown in white and grey, respectively.

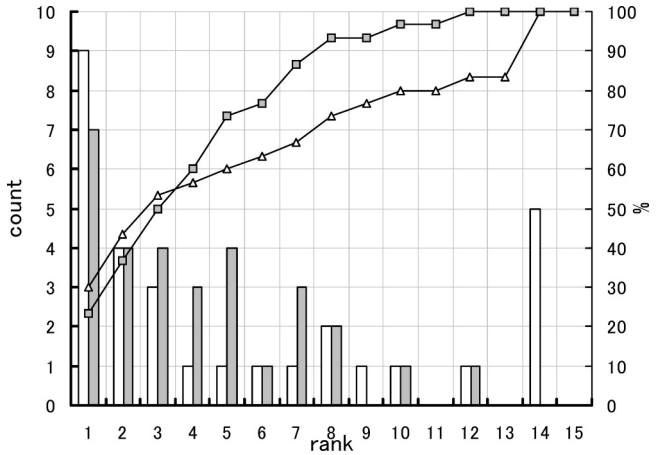


Fig. 4. Histogram of Counts (Bar) and Accumulated Percentages (Line) of Predicted Solvents for 30 Test Compounds

The results obtained by decision-tree and Bayesian-probability models are shown in white and grey, respectively.

Table 2. Systematic Chemical Names of 30 Randomly Selected Molecules

| Compound number | Systematic name |
|---|---|
| 1 | Ethyl 6-amino-2-methoxypyridine-3-carboxylate |
| 2 | Di-*n*-butyl pyridine-2,6-dicarboxylate |
| 3 | 6b,8a-Dihydrocyclobut(a)acenaphthylene-7,8-dicarbonitrile |
| 4 | (4-(Dimethylammonio)butyl)-bis(2-methyllactato-*O*,*O*′)-silicate |
| 5 | (1*R**,5*S**,7*S**)-1-(4-Nitrophenyl)-7-phenyl-7-trimethylsiloxy-2-oxabicyclo(3.2.0)heptane-3,4-dione |
| 6 | Tris(3,5-di-*tert*-butyl-2-(cyanomethoxy)phenyl)methane |
| 7 | 1,2-Bis(bis(2-methylthiophenyl)phosphino)ethane |
| 8 | 2,4-Bis-(2,4,6-tri-*t*-butylphenyl)-1-trimethylsilyl-2,4-diphosphabicyclo(1.1.0)butane |
| 9 | 1-(2-Deoxy-$\beta$-D-erythro-pentofuranosyl)-4-methylbenzimidazole monohydrate |
| 10 | *N*,*N*′-Bis(2-methoxyphenyl)-4,6-dibenzyloxybenzene-1,3-dicarboxamide |
| 11 | 1,5,6,7,8,12,13,14,15,15-Decachloro-16,16-dimethoxypentacyclo(10.2.1.1$^{5,8}$.0$^{2,11}$.0$^{4,9}$)hexadeca-6,13-diene |
| 12 | Tetraethyl 2,2′-(1,4-phenylene)-bis(1,3-dioxane)-5,5,5′,5′-tetracarboxylate |
| 13 | 4-(Dimethylamino)-3-cyanobiphenyl |
| 14 | *E*-2-(2-Nitrophenylhydrazono)-3-oxobutanenitrile |
| 15 | 2,4′-Diphenyl-3,3′-diisopropoxy-4,2′-dihydroxy-4,4′-bicyclobutenone |
| 16 | 3$\beta$-Acetoxy nor-31-lanostene-7,11-dione |
| 17 | Bis($\mu_2$-hydrido bis(dimethyl(*t*-butyl)silyl)silyl)-bis(tetrahydrofuran)-di-lithium |
| 18 | *N*-(3-(1,1,1-Triethoxysilyl)propyl)-(*E*)-3-phenyl-2-propenamide |
| 19 | 5,6-(1,2-Dicarba-closo-dodecaboranylene)-1,1,4,4-tetramethyl-2,3-diphenyl-1,4-disilacyclohex-2-ene |
| 20 | Dibenzyl-*cis*-[4a]-cisoid-[4a,4b]-*cis*-[4b]-dodecahydro-9,11-dimethyl-2,4,6,8-tetraoxocyclobuta[1,2-d; 3,4-d′]dipyrimidine-1,5-diacetate |
| 21 | 7-Methyl-2,7-diaza-11-oxatetracyclo(6.6.2.0$^{4,16}$.0$^{12,15}$)hexadeca-1(14),3,12,15-tetraene |
| 22 | 6-Methoxy-1,6-diphenyl-4-thioxo-3,4,5,6-tetrahydro-2,3,5-triazine |
| 23 | Rac-(*S**,*S**)-2,2-diphenyl-4-(phenoxy)-6-(anilino)-4,6-(3,6,9-trioxaundecane-1,11-dioxy)cyclotriphosphazene |
| 24 | *N*,*N*′-Bis(cyanoborane)-1,4-diammoniobutane |
| 25 | Methyl 2,3-di-*O*-pivaloyl-4-(*O*-3-methylpent-4-enoyl)-6-deoxy-6-iodo-$\alpha$-D-glucopyranoside |
| 26 | 2,2-Tetramethylene-1,2,3,4-tetrahydroquinazolin-4-one |
| 27 | *N*,*N*-Bis(diphenylphosphino)-*N*-((*S*)-a-methylbenzyl)amine |
| 28 | (2c,4ar,8ac)-2-Methyl-4a,5,6,7,8,8a-hexahydro-2*H*,4*H*-1,3-benzodithiine |
| 29 | Bis((diphenyl(piperidinomethyl)silyl)methyl)-magnesium |
| 30 | *cis*,*cis*-1,3,5-Cyclohexanetricarboxylic acid tris(urea) |

by the decision-tree model accord with the very solvents used in actual crystallization. Accuracy rate achieved by the decision-tree model reaches to 80% if the three highest ranking solvents are considered. By use of the Bayesian model, however, the same accuracy rate is attained if the seven highest ranking solvents are considered. Although these accuracy rates alone indicate that the decision-tree model is much better than the Bayesian, both models complement each other.

We have checked the performance of the relevant methods by predicting the solvent rankings for randomly selected 30 compounds from the training set. These 30 compounds are given in Table 2. A large variety of compounds are contained in the test dataset. Now classification models were obtained from the remaining 6367 compounds. The results are given in Fig. 4 that shows the ranking histograms of predicted solvents. In Table 3, predicted top three solvents are given. The results obtained by the decision-tree model indicate that the accuracy rate of 80% can be attained if we consider the ten highest ranking solvents. By the Bayesian-probability model, however, the same accuracy rate can be attained with the top

Table 3. Predicted Top Three Solvents for 30 Randomly Selected Compounds

| Compound number | CSD code | Used solvent | Predicted solvents | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Decision-tree | | | Bayesian | | |
| | | | 1st | 2nd | 3rd | 1st | 2nd | 3rd |
| 1 | SAQZEO | 6 | 1 | 2 | 6 | 1 | 2 | 4 |
| 2 | TIFRUV | 6 | 1 | 6 | 8 | 1 | 2 | 4 |
| 3 | VAQVAK | 6 | 2 | 8 | 5 | 1 | 4 | 14 |
| 4 | HOQVUE | 7 | 7 | 1 | 6 | 7 | 1 | 10 |
| 5 | NINJOJ | 11 | 2 | 8 | 5 | 1 | 4 | 2 |
| 6 | XILKEI | 9 | 1 | 11 | 8 | 3 | 11 | 10 |
| 7 | NESWOX | 5 | 10 | 8 | 4 | 5 | 3 | 7 |
| 8 | CIQQOI | 13 | 2 | 11 | 8 | 3 | 11 | 10 |
| 9 | BAGVEK | 1 | 2 | 4 | 6 | 1 | 4 | 2 |
| 10 | LUDXAJ | 1 | 9 | 8 | 1 | 7 | 1 | 2 |
| 11 | NUYPIG | 1 | 6 | 8 | 5 | 6 | 7 | 9 |
| 12 | SAQMEC | 1 | 8 | 5 | 3 | 9 | 2 | 10 |
| 13 | XEMWER | 1 | 6 | 4 | 2 | 1 | 2 | 4 |
| 14 | GUFLUO | 4 | 6 | 4 | 2 | 1 | 2 | 4 |
| 15 | XIWPAU | 4 | 4 | 2 | 6 | 1 | 2 | 4 |
| 16 | CABZAH | 3 | 2 | 6 | 1 | 2 | 1 | 4 |
| 17 | LUGFIC | 3 | 3 | 8 | 5 | 6 | 7 | 9 |
| 18 | TACNIV | 3 | 3 | 1 | 6 | 1 | 2 | 4 |
| 19 | XEKLII | 3 | 3 | 2 | 8 | 3 | 10 | 11 |
| 20 | CUPHAW | 2 | 2 | 9 | 1 | 7 | 6 | 2 |
| 21 | FOXQOY | 2 | 2 | 6 | 8 | 2 | 8 | 1 |
| 22 | GIZNUY | 2 | 1 | 2 | 8 | 1 | 2 | 4 |
| 23 | GUZQAT | 2 | 2 | 10 | 8 | 6 | 7 | 9 |
| 24 | LOYTAU | 2 | 3 | 8 | 9 | 6 | 7 | 9 |
| 25 | QEKFOB | 2 | 4 | 6 | 1 | 3 | 4 | 2 |
| 26 | QELCEP | 2 | 6 | 1 | 8 | 1 | 2 | 4 |
| 27 | XAGMOH | 2 | 2 | 8 | 4 | 10 | 2 | 8 |
| 28 | XOYCOD | 2 | 1 | 3 | 2 | 1 | 6 | 4 |
| 29 | XORVUV | 10 | 11 | 7 | 6 | 3 | 5 | 9 |
| 30 | XORMUM | 12 | 2 | 12 | 1 | 12 | 2 | 9 |

The numbers of solvents correspond to those in Table 1.

seven solvents. From the practical point of view, these results are not necessarily satisfactory, but encouraging and useful. As the two models are complementary to each other as already mentioned, it is worthwhile to take both results into account. If the results are simply combined, the experimentally used solvents are correctly predicted for 21 compounds by use of the top three solvents predicted from the two methods. It means that up to six solvents must be considered. The accuracy rate is 70% in this case. The results are now almost satisfactory from a practical standpoint.

Since crystallization conditions of small organic molecules have not been screened systematically so far, the solvents used to get crystals for structure determinations compiled in CSD are not necessarily the optimum ones. The solvents described in CSD should be considered as those that gave single crystals sufficient for diffraction work. It is highly possible that other solvents can give better crystals. In addition, crystallization process is governed by various conditions and solvent is an important but just one factor that affects crystallization process. In spite of these difficult circumstances, reasonably good prediction results have been obtained. It strongly suggests that the method applied in the present study is sound in principle, and it is expected that the accuracy rate will be greatly improved if we could use the data obtained by more systematic experiments.

## Conclusion

Crystallization is a very important operation in pharmaceutical industries. The operation for a small organic molecule is, however, still a matter of trial and error. In order to carry out crystallization in more rational way, we made use of cheminformatics methods to rank the solvents suitable for crystallization. A large number of crystallization data compiled in CSD were employed. We applied decision-tree and Bayesian-probability methods to rank solvents that can be used for crystallization of a particular compound. Both methods are found to be complementary to each other in the present case. Classifications were undertaken using 2D descriptors calculated for 6397 compounds in CSD. Suitable crystallization solvents for test compounds that were randomly chosen from the training dataset were predicted satisfactorily. The present study has demonstrated that cheminformatics methods can be applied to a complex problem such as prediction of solvents suitable for crystallization of small organic molecules including various pharmaceuticals.

**Acknowledgements** We thank Ms. Rumiko Tanaka for her technical assistance. One of the authors (N.H.) is grateful to the Research and Study Program of Tokai University Educational System General Research Organization for financial support.

**References**

1) Rouhi A. M., *Chem. Eng. News*, **81**, 45—55 (2003).
2) Brittain H. H., "Polymorphism in Pharmaceutical Solids," Marcel Dekker, Inc., New York, 1999.
3) Stock D., Perisic O., Löwe J., *Prog. Biophys. Mol. Biol.*, **88**, 311—327 (2005).
4) Allen F. H., *Acta Crystallogr.*, *Sect B*, **58**, 380—388 (2002).
5) Leach A. R., Gillet V. J., "An Introduction to Chemoinformatics," Kluwer Academic Publishing, Dordrecht, 2003, pp. 54—64.
6) Quinlan J. R., "C4.5: Programs for Machine Learning," Morgan Kaufman, San Mateo, California, 1993.
7) Gelman A., Carlin J. B., Stern H. S., Rubin D. B., "Bayesian Data Analysis," 2nd ed., Chapman & Hall/CRC, London, 2004.
8) MOE (Molecular Operating Environment), Version 2004.04, Chemical Computing Group Inc., Montreal, Quebec, 2004.
9) KDE (Knowledge Discovery Environment), Version 1.9, InforSense Ltd., London, 2004.