Computer-Assisted Structure Identification: Predictive Models and Automation

A. Monge[1], A. Knorr[2], M. Stueber[2], D. Arndt[2], A. Stratmann[2], E. Martin[3], M. C. Peitsch[3], N. V. Ivanov[3], and P. Pospisil[3]

[1] *blue-infinity, Geneva, Switzerland*
[2] *Philip Morris International R&D, Philip Morris Research Laboratories GmbH, Köln, Germany*
[3] *Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland*

Introduction

The CASI (Computer-Assisted Structure Identification) approach is used at Philip Morris International R&D to identify small molecules in complex matrices analyzed with GCxGC-TOF. In the CASI approach, structure candidates and associated match factors for a mass spectrum are obtained using NIST MS Search. In order to refine the results of NIST MS Search, we developed quantitative structure−property relationship models to predict values of the two retention times of a GCxGC-TOF-MS instrument. Two models that are specific for the GCxGC-TOF-MS instrument, non-polar (1st dimension) x polar (2nd dimension), have been built: a Kovats Indices (KI) model was built for the 1st dimension and a 2D retention time (2Drt) model was built for the 2nd dimension. The models can be adapted for different column combinations.

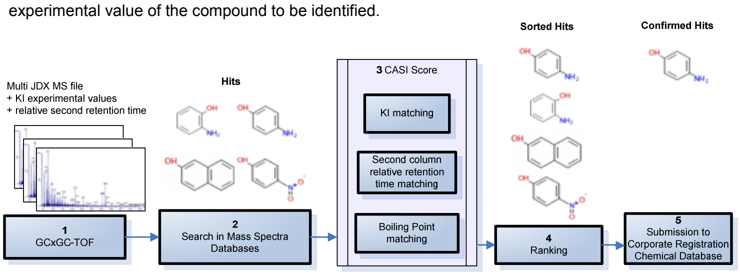
The web interface of the platform proposes a list of the best matched structure candidates allowing users to easily check and correct structure assignments. CASI also enables the authorized user to easily add new instruments, analytical columns, and retention models to the platform.

Concepts

Modeling of Kovats indices was already used to improve the results of GC-MS library search [1]. We developed an enhanced process based on GCxGC-TOF. The automated process is designed to propose chemical structures for the compounds to identify them without intervention of the user.

Automated Identification

The automated process of compound identification by CASI is described in Fig. 1. The process was automated in Java and is available as a web service. The descriptors for prediction models were calculated using software Dragon. RapidMiner was used to apply predictive retention models. Analytical scientists provide mass spectra files, KIs, and 2D relative retention times to the software (see oral presentation of Knorr et al. [2]). First, each mass spectra of the compound to be identified is searched in various mass spectra databases using NIST MS Search, and the first 100 hits are returned. Structures are standardized and structural duplicates are removed using Pipeline Pilot 8. For each hit, KI, relative retention time for the 2nd dimension, and boiling point (BP) are calculated using predictive models (see below). Final CASI Score is calculated using a function, taking into account the match factor of NIST MS Search and the difference between each predicted and experimental value of the compound to be identified.



Multi JDX MS file
+ KI experimental values
+ relative second retention time

1 GCxGC-TOF → 2 Search in Mass Spectra Databases → 3 CASI Score → 4 Ranking → 5 Submission to Corporate Registration Chemical Database

Hits

3 CASI Score: KI matching; Second column relative retention time matching; Boiling Point matching

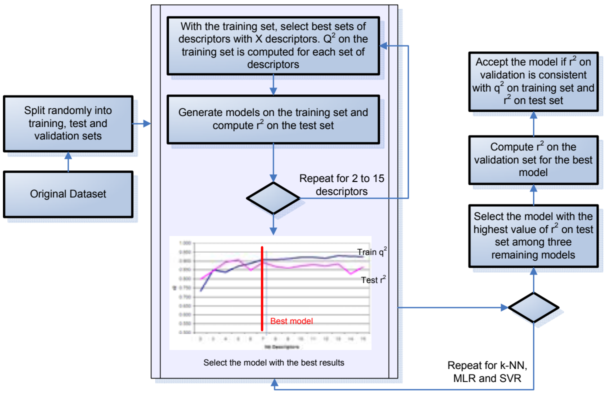
Sorted Hits

Confirmed Hits

Figure 1. Process run automatically by the CASI platform.

Retention Models

Three predictive retention models are key elements of our automated process, as they are used in *step 3 CASI Score*. We used a set of 160 non-polar compounds split as follows: training set (90), test set (35), and validation set (35). For each value to be predicted (KI and 2nd dimension relative retention time) predictive models were built using three learning algorithms: k-Nearest Neighbors (k-NN), Multi Linear Regression (MLR) and Support Vector Regression (SVR). For each learning algorithm, best sets of descriptors in the range of 2 to 15 descriptors were generated. At the end, the best model is kept for each value to be predicted. This process is described in Fig. 2.



Original Dataset

Split randomly into training, test and validation sets

With the training set, select best sets of descriptors with X descriptors. Q^2 on the training set is computed for each set of descriptors

Generate models on the training set and compute r^2 on the test set

Repeat for 2 to 15 descriptors

Select the model with the best results

Select the model with the highest value of r^2 on test set among three remaining models

Compute r^2 on the validation set for the best model

Accept the model if r^2 on validation is consistent with q^2 on training set and r^2 on test set

Repeat for k-NN, MLR and SVR

Figure 2. Process used to build the Kovats index and second dimension relative retention time models.

For KI, the best results were obtained with MLR algorithms with 15 descriptors and the r^2 on the validation set was 0.985. For 2D relative retention time, the best model used SVR algorithm with 7 descriptors and r^2 on validation set was 0.849 (see Table 1).

		GA - MLR	GA - kNN	GA-epsilon SVR (linear kernel)	
KI	Q2	**0.988**	0.972	0.979	C = 1.9 x 10^{-3}
	R2 (test set)	**0.982**	0.956	0.957	
2DRT	Q2	0.861	0.841	**0.840**	C = 3.8
	R2 (test set)	0.750	0.673	**0.827**	

Table 1. Result of the best models for KI and second dimension relative retention time with multi linear regression, k-nearest neighbors and support vector machine regression. Q2 values were obtained with leave-one-out cross validation for MLR and 10 folds cross validation for kNN and 5 folds cross validation for SVR. Results shown in bold are the one selected for each parameter.

Boiling Point

Boiling points were used as additional information for identification of chemical structures. We used ACD/PhysChem Batch to compute the boiling points from the chemical structures of hits found by NIST MS Search. It is known that boiling points are highly correlated to retention times. As a consequence, we built a linear equation (Eq. 1) with the experimental Kovats Indices of our training set (r^2 = 0.955) and checked this equation to compute boiling points for the compound in the test set (r^2 = 0.910) and validation set (r2 = 0.934).

$$BP = 0.1468 \times KI + 47.402 \qquad \text{Eq. 1}$$

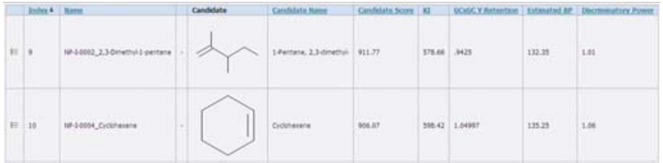
Ranking

For each query, the proposed hits are ranked according to decreasing CASI scores. CASI score is calculated according to Eq. 2. The hit with the highest value is selected by default through the CASI platform. The basis of the score is the NIST Match Factor (*NIST MF*), which is corrected by predicted values of *KI*, 2nd dimension relative retention times (*2DRT*), and Boiling Point (*BP*). Correction is done by multiplying *NIST MF* by hyperbolic (e.g., hyp_{KI}), which gives weights to each component according to the Standard Error of Prediction on the training set (e.g., SEP_{KI}), the experimental value for the query (e.g., KI_{query}) and the predicted value for the hit (e.g., KI_{pred}). The more accurate the prediction, the higher the contribution of the predicted parameter.

$$\begin{aligned} CASI\ Score &= NIST\ MF \times hyp_{KI}\left(SEP_{KI\ train}, KI_{query}, KI_{pred}\right) \qquad \text{Eq. 2} \\ &\times hyp_{2DRT}\left(SEP_{2DRT\ train}, 2DRT_{query}, 2DRT_{pred}\right) \times hyp_{BP}\left(SEP_{BP\ train}, BP_{query}, BP_{pred}\right) \end{aligned}$$

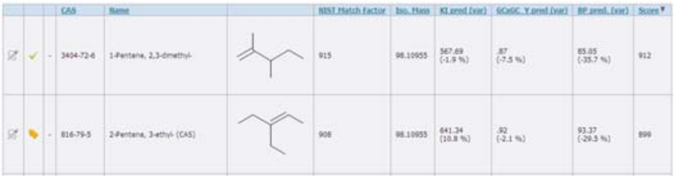
User Interface

For a given analysis, all compounds to be identified are presented, with the structure candidate having the best score (Fig. 3). Structure candidates can be browsed and selection can be changed (Fig. 4).



	Index	Name		Candidate	Candidate Name	Candidate Score	KI	GCxGC Y Retention	Estimated BP	Discriminatory Power
	9	NP-1-0002_2,3-Dimethyl-1-pentene	-		1-Pentene, 2,3-dimethyl-	911.77	578.66	.9425	132.35	1.01
	10	NP-1-0004_Cyclohexene	-		Cyclohexene	906.07	598.42	1.04997	135.25	1.06

Figure 3. For each structure to be identified (Query), the structure candidate with the highest score is selected by default.



			CAS	Name		NIST Match Factor	Iso. Mass	KI pred (var)	GCxGC Y pred (var)	BP pred. (var)	Score
	✓	-	3404-72-6	1-Pentene, 2,3-dimethyl-		915	98.10955	567.69 (-1.9 %)	.87 (-7.5 %)	85.05 (-35.7 %)	912
		-	816-79-5	2-Pentene, 3-ethyl- (CAS)		908	98.10955	641.34 (10.8 %)	.92 (-2.1 %)	93.37 (-29.5 %)	899

Figure 4. Each structure candidate (Hits) for the compound to be identified (Query, in this case 1-Pentene, 2,3-dimethyl) are listed with predicted properties. The one with the best score is selected by default. Users can change the selection and add comments, which will be inserted with the selected structure into the chemical registration system.

Validation Results

We used a set of 71 experimentally confirmed molecules. Some of these molecules are present in the validation set used to validate the models, but none of them are present in the training and test sets. CASI score ranking (51 correct hits ranked first and 14 correct hits ranked in second position) gives better results than NIST Match Factor (50 correct hits ranked first and 9 correct hits sorted in second position). The better example of the advantage of the CASI score is the Hentriacontane, which is sorted in 20th position with NIST MF but sorted in 2nd position with CASI score, because of the accurate prediction of the KI.

Position of correct hits	1	2	3	4	5	6	7	10	20
Frequency with CASI score	**51**	**14**	**3**	**2**		**1**			
Frequency NIST Match Factor	50	9	4	2	2	1	1	1	1

Table 2. Comparison of the position of correct hits using ranking based on CASI score and ranking based on NIST Match Factor.

Summary and Conclusion

We built an automated platform which assists to analyze the results of GCxGC-TOF by:

- ✓ improving the speed of the analysis
- ✓ suggesting hit names and their 2D structures with increased confidence (due to prediction of physical parameters (Kovats Indices, second dimension relative retention times, and Boiling Points)
- ✓ and the ranking of correct hits.

We have shown that the platform improves the identification in comparison to NIST Match Factor. Further components, such as accurate mass, are planned to be added in order to improve the quality of the prediction.

References

1. Mihaleva V.V. et al. *Automated procedure for candidate compound selection in GC-MS metabolomics based on prediction of Kovats retention index*. Bioinformatics; 2009; 25 (6); 787-94.
2. A. Knorr et al., *Computer-assisted structure identification (CASI) - A mass spectrometry-based automated platform for high-throughput identification of small molecules by two-dimensional gas chromatography-mass spectrometry*, oral presentation, metabolomics2011, Cairns, Australia.

PMI RESEARCH & DEVELOPMENT

Metabolomics 2011
Cairns, Australia
27 - 30 June

Philip Morris International Research & Development, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
T: +41 58 242 21 11, **F**: +41 58 242 28 11, **W**: www.philipmorrisinternational.com