Identifying “Known Unknowns” Using ChemSpider and Automated MS/MS Structure Correlation

James Little, Eastman Chemical Company, Kingsport, TN
Frank Kuhlmann, Xiangdong Li, Jerry Zweigenbaum, Agilent Technologies, Inc., Santa Clara, CA
Antony Williams, Valery Tkachenko, ChemSpider/Royal Society of Chemistry (RSC)

Tuesday Poster Session TP28 Slot: 332
2012 ASMS Conference
60th ASMS Conference
Vancouver, Canada

EASTMAN

Agilent Technologies

ChemSpider
The free chemical database

RSC | Advancing the Chemical Sciences

Overview

- "Known Unknowns" [1-3] can be routinely identified using large "spectraless" databases such as ChemSpider with accurate mass MS and MS/MS data

- Known in ChemSpider or other databases, but unknown in the sample to the investigator

- Prototype version of Agilent MassHunter Molecular Structure Correlator (MSC), which is based on a systematic bond-breaking approach,[4] imports associated references in addition to structures when querying ChemSpider by elemental composition

- MSC fragmentation score and assigned substructural information is used in combination with associated references to obtain most probable candidate structures for an unknown

Introduction

A compound known in the chemical literature is often an unknown to an investigator when being analyzed in a sample. Thus, the origin of the term “known unknown” for this type of non-targeted species.[1-3] Large “spectraless” databases and accurate mass LC-MS data are routinely utilized to identify unknowns prioritized by the number of associated references using these approaches. ChemSpider is a particularly valuable resource since it contains >27 million compounds and is provided as a free resource to the community via the internet.

In previous work, the candidate structures from elemental composition searches were further evaluated by ***manually*** assigning the observed fragment ions to substructures using a drawing program. The Agilent MassHunter Molecular Structure Correlator (MSC) software ***automatically*** scores and visually compares the observed MS/MS fragmentation to that of all the candidate structures utilizing a “systematic bond breaking approach.”[4] A unique feature of the MSC program is that it correlates an accurate mass MS/MS spectrum for each candidate structure and assigns a comparative scoring.

In the current work, a prototype MSC program was used which considers the number of associated ChemSpider references together with each structural correlation score to rank possible candidate structures. The resulting structures can then be sorted either by the references or by the structure correlation score. This approach for the identification of unknowns was evaluated with a group of test compounds.

Methods

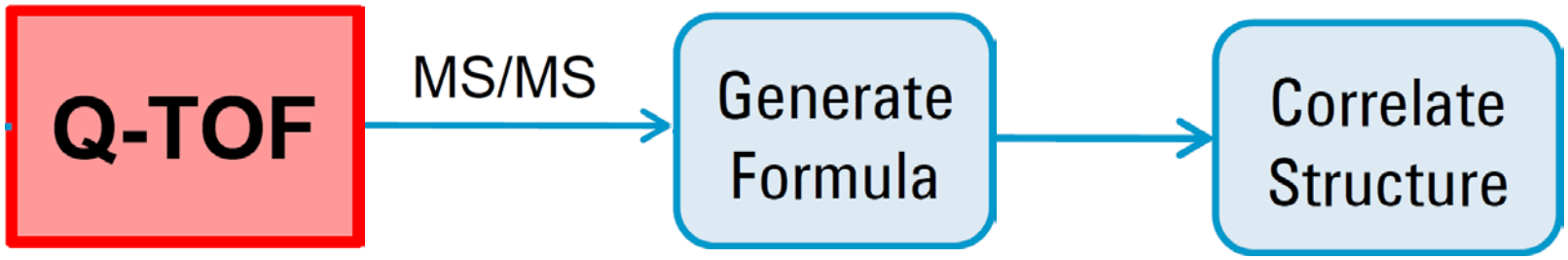
The accurate mass data was obtained for 38 components in Agilent's Toxicology and Environmental Test mixtures. These known compounds were used to evaluate the effectiveness of the approach for the identification of unknowns. The positive ion electrospray accurate mass data was obtained on an Agilent 6540 Q-TOF LC/MS system using Auto MS/MS acquisition.

MassHunter Qualitative Analysis software was used to automatically detect compounds in the test samples using an untargeted approach via the Molecular Feature Extractor (MFE). The latest version of MFE also automatically extracts MS/MS spectra if they exist. For each compound the accurate mass MS and MS/MS information was then exported to a compound exchange file (*.CEF) and imported into the Agilent Molecular Structure Correlator (MSC) software.

The ***current*** release version of the MSC software calculates the most probable molecular formula from the accurate mass MS and MS/MS data of an unknown compound and then queries the ChemSpider database via an applications programming interface (API) to automatically retrieve all structures consistent with that elemental composition. The MS/MS spectrum of the unknown compound is then correlated with each candidate structure and an overall correlation score is assigned based on the fraction of explainable fragment ion intensity, their mass accuracy, and the energy needed to create each fragment ion.

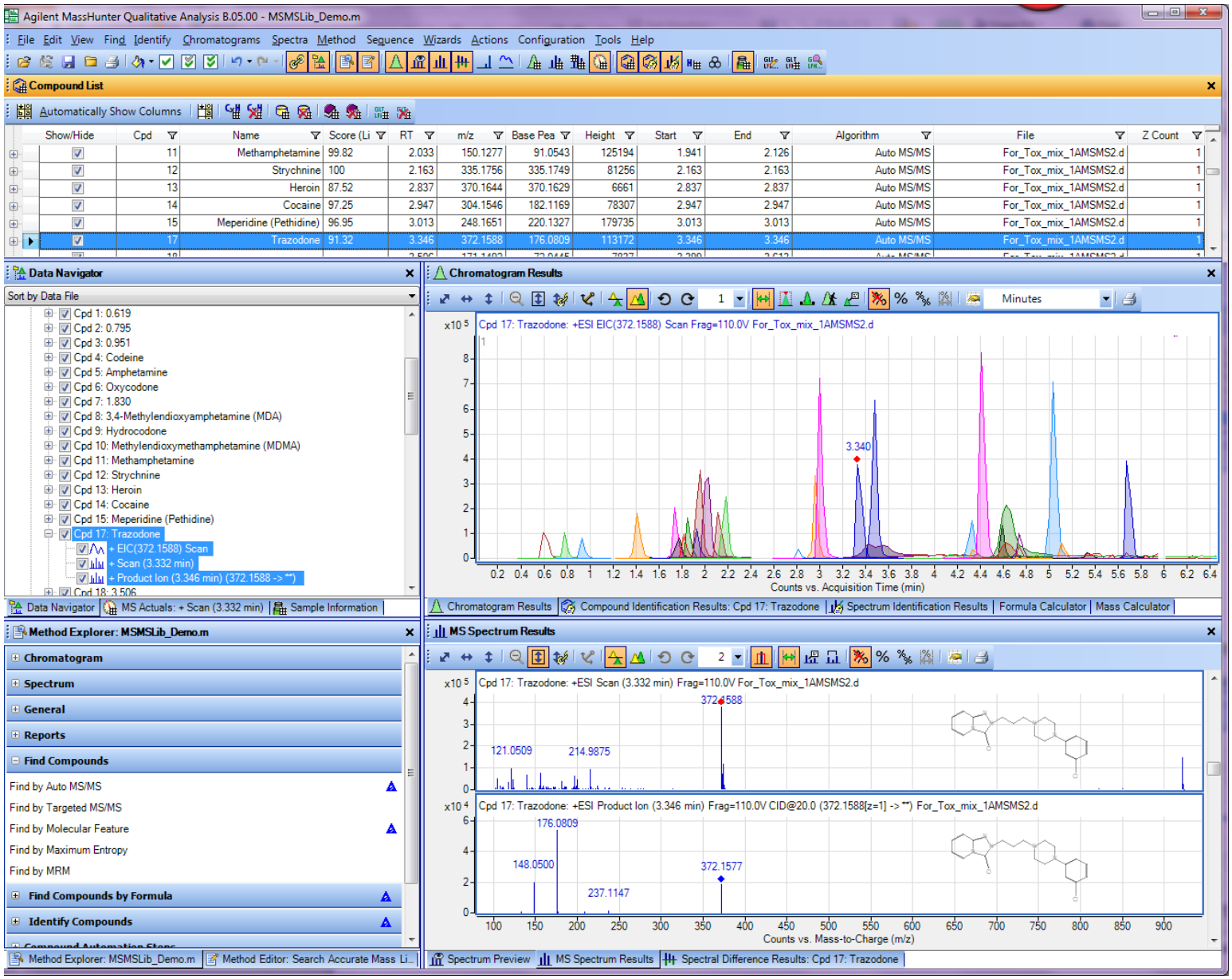
The ***prototype*** MSC program contains additional capabilities to query and process the ChemSpider compound database for the number of associated references. Reference fields and the overall correlation score can then be sorted for evaluation.

Approach

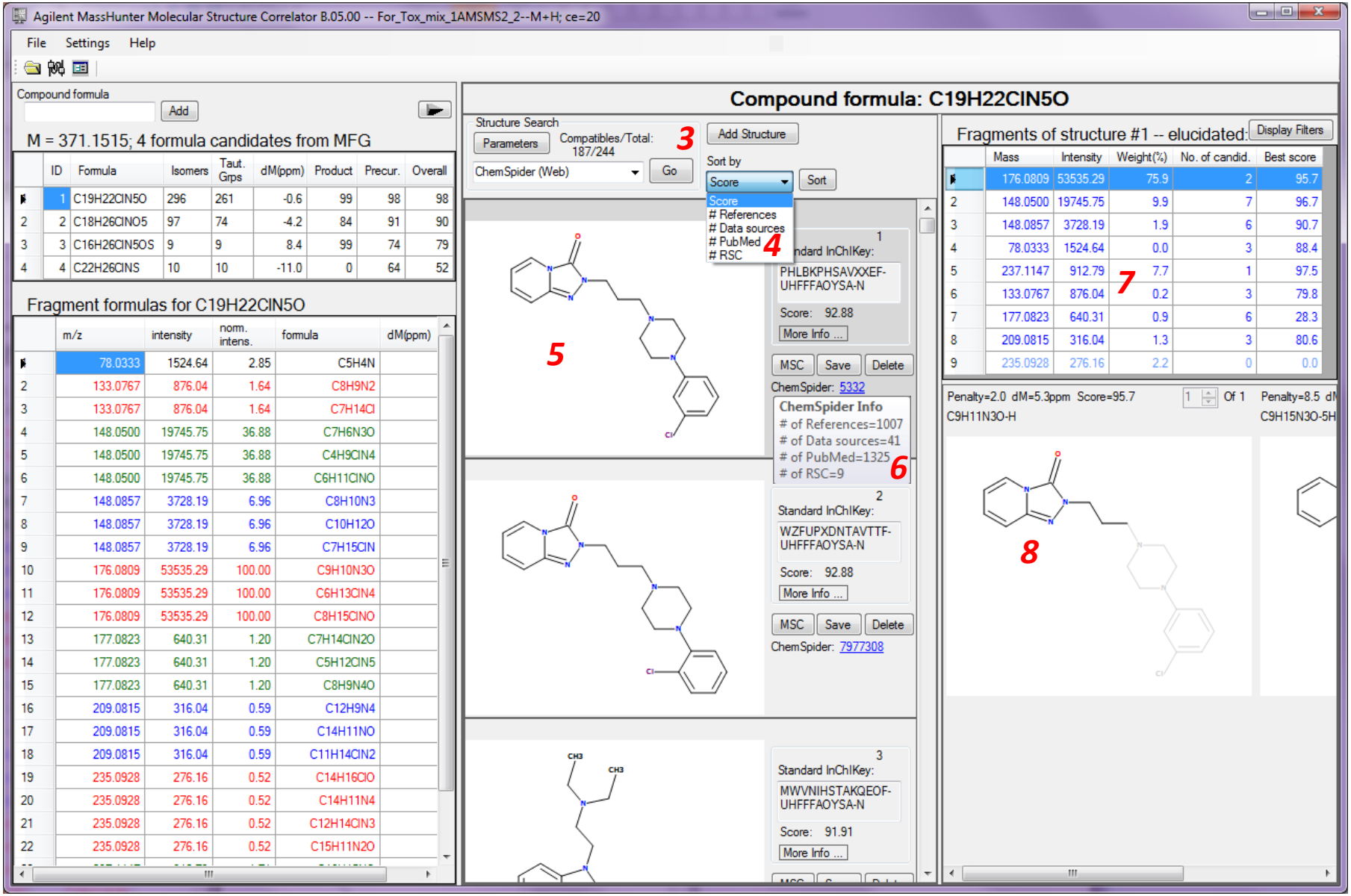


- Accurate mass data acquired by LC/MS in data-dependent MS/MS mode
- Agilent MassHunter Qualitative Analysis software used to find compounds and generate molecular formulas using monoisotopic mass, isotope abundance, and isotope spacing, as well as look for matching formulas for each fragment ion and its neutral loss from the precursor
- Prototype Agilent MassHunter Molecular Structure Correlator software uses ChemSpider interface to obtain candidate structures for a target molecular formula and associated number of references for candidate structures
- Calculation of correlation scores to explain MS/MS fragmentation pattern for each candidate structure using a “systematic bond-breaking” approach
- Results sorted by number of references and/or correlation score to evaluate structure candidates for identification of components

Data Processing: MassHunter



MSC *Prototype* Interface



Windows in MSC Prototype

1. Molecular formulas with MS score (mass position, isotope abundance, isotope spacing), MS/MS score and combined score

2. Molecular formulas for fragment ions with *m/z* error, part of MS/MS score

3. Number of structures retrieved for formula/number of structures qualified in structure correlation

4. Fields used to sort the results including reference categories from ChemSpider and overall structure correlation score calculated by MSC program

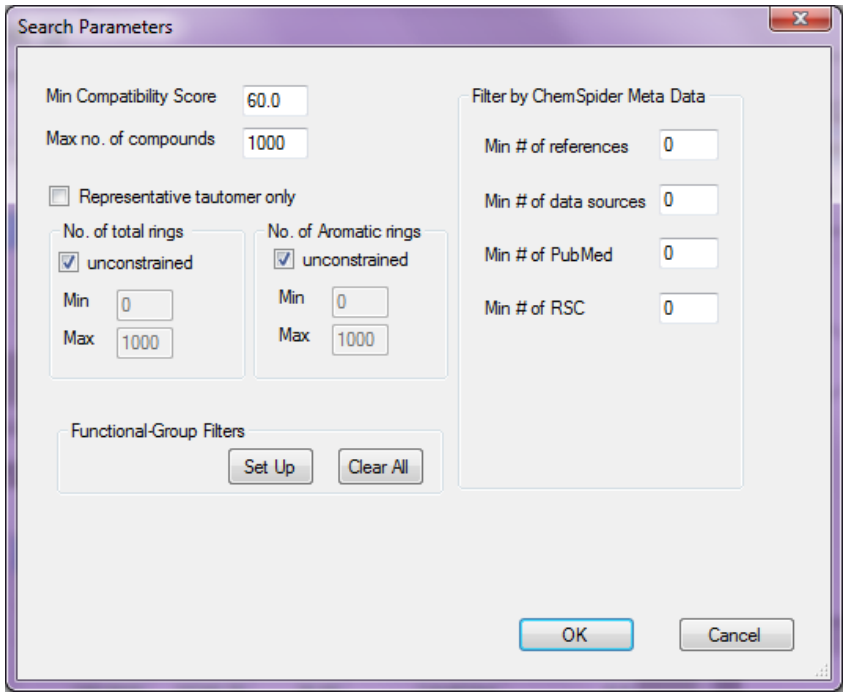
5. Structure selected for showing substructural information in sections ***7*** and ***8***

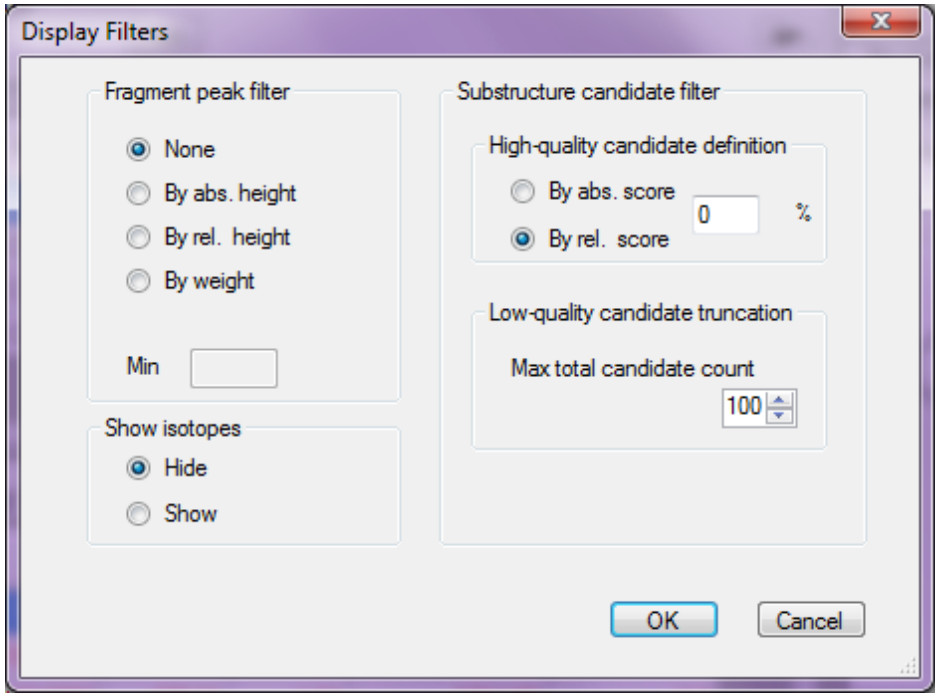
6. Number of references retrieved from ChemSpider for each category

7. Tabulated information for substructures including best individual fragment score

8. Substructures for selected structure with score, mass error, and penalty

Candidate Structure Filters in MSC

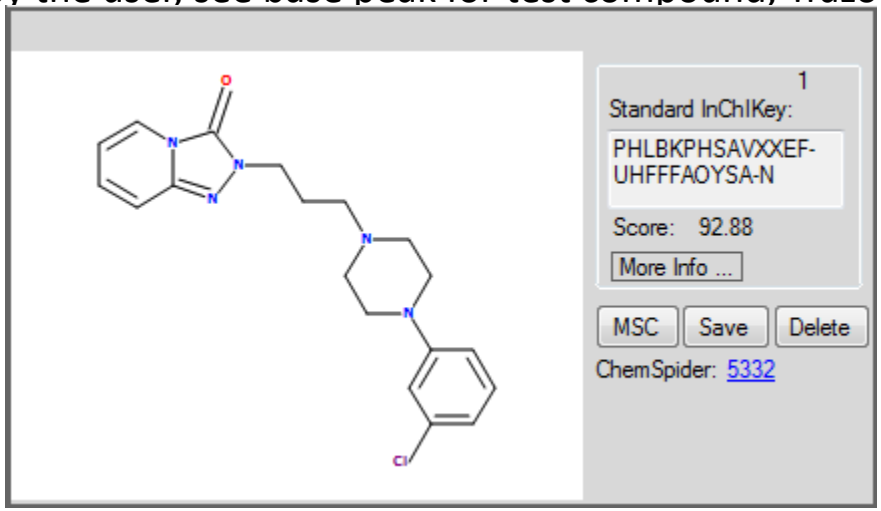




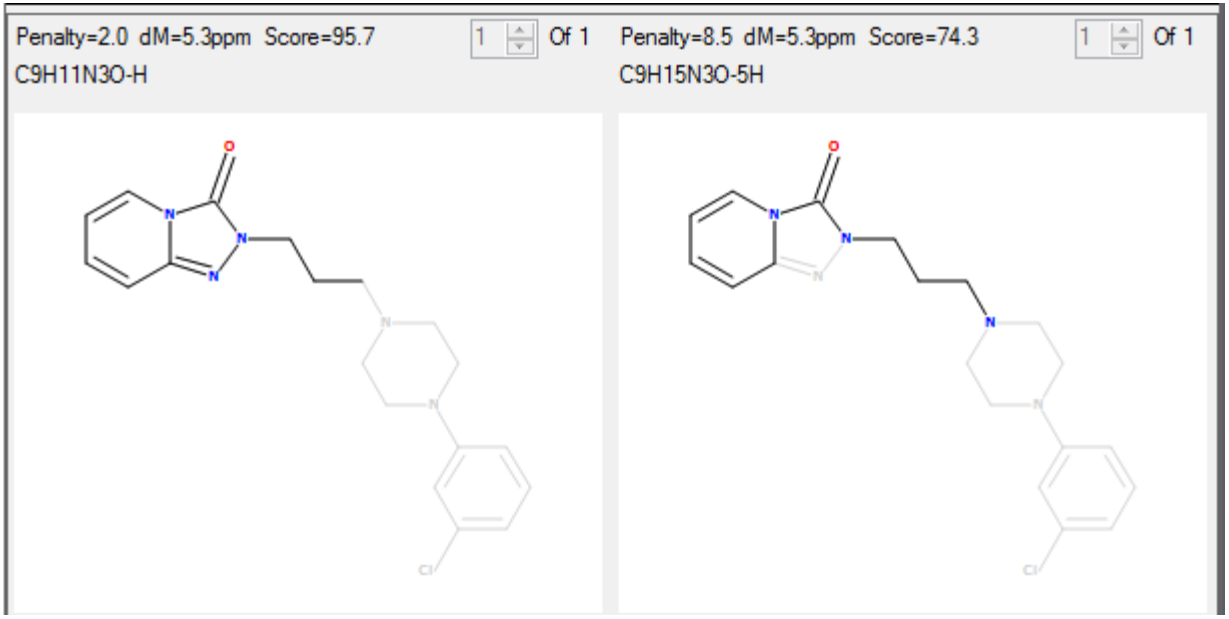
Easy Review of Substructures for Fragment Ions

The proposed substructures and associated individual fragment ion score are easily reviewed by the user, see base peak for test compound, Trazodone, in study

5



8



Partial Results of Searches

Compound	No. Structure Candidates	Structure Score	Structure Rank	No. References	Reference Rank
ecgonine methyl ester	743	87.21	7	156	1
codeine	2769	73.6	104	2246	1
amphetamine	277	94.5	3	10829	1
oxycodone	2371	87.07	94	611	1
3,4-methylenedioxyamphetamine	1494	92.26	32	272	2
hydrocodone	2769	66.33	212	201	2
methamphetamine	446	95.91	26	3559	1
strychnine	2256	14.58	195	2468	1
cocaine	1417	95.44	9	15414	1
meperidine	1953	97.46	27	2959	1
trazodone	244	93.17	1	956	1
phencyclidine	274	94.01	12	2524	1
verapamil	210	93.96	4	13379	1
levomethadone	758	89.93	13	5713	1
nitrazepam	568	86.07	49	765	1
alprazolam xanax	104	81.71	16	1229	1
clonazepam	210	87.62	18	1975	1
proadifen	594	92.39	8	272	1
temazepam	786	94.14	27	1164	1
diazepam	494	89.03	10	12751	1
diethyl phthalate	1264	95.35	16	179	1
aminocarb	1163	97.37	7	97	3
thiabendazole	75	95.52	18	1422	1
imazapyr	1040	94.61	5	88	1
dimethoate	2	94.89	1	660	1
imazapic	1322	88.82	16	30	1
metoxuron	205	88.76	7	67	1
enilconazole	104	81.42	14	171	1
atrazine	22	97.24	2	1776	1
metosulam	11	43.93	7	25	1
metazachlor	440	96.42	1	46	1
pyraclostrobin	840	85.51	1	19	1
monocrotaline	303	39.42	184	556	1
simazine	15	82.16	7	522	1
molinate	99	92.38	6	153	1
malathion	8	91.68	4	1794	1
heroin	2680	85.94	4	147	1

Tabulated Results

No. Candidates Average	No. Candidates Median	Range of Candidates	
849	531	22-769	
Average Structure Score	**Median Structure Score**	**Range Structure Score**	
85	91	15-97	
Average Structure Rank	**Median Structure Rank**	**Range of Structure Rank**	
32	11	1-212	
Average No. of References	**Median No. of References**	**Range of References**	
2319	713	13-15000	
Average Reference Rank	**Median Reference Rank**	**Range of Rank**	**Entries with No Rank (0 Entries)**
1	1	1-3	0
Average PubMed References	**Median PubMed References**	**Range of PubMed References**	
2328	24	0-26,000	
Average PubMed Rank	**Median PubMed Rank**	**Range of PubMed Rank**	**Entries with No Rank (0 Entries)**
2	1	1-7	9
Average RSC References	**Median RSC References**	**Range of RSC References**	
245	4	0-1737	
Average RSC Rank	**Median RSC Rank**	**Range of RSC Rank**	**Entries with No Rank (0 Entries)**
2	1	1-7	11
Average Data Sources	**Median Data Sources**	**Range of Data Sources**	
27	24	12-63	
Average RSC Rank	**Median RSC Rank**	**Range of RSC Rank**	**Entries with No Rank (0 Entries)**
3	1	1-7	0

Conclusions

- Initial results with the limited data set indicate that the MassHunter MSC program is a very useful tool for the identification of unknown compounds in mixtures especially with the addition of the associated reference fields

- Hundreds of candidate structures can be returned from a large compound database like ChemSpider for one elemental composition; thus, there can be many hits with a similar correlation score (and similar chemical structure) which makes it difficult to pinpoint the correct structure for the unknown compound

- Sorting of the candidate structures by “number of references” is most useful for bringing the proper identification to the top of the hit list

- Sorting by PubMed or RSC references and data sources also proves to increase the discrimination between the hits

- The combination of the MSC correlation score in combination with the number of references results in getting the correct hits to the top of the results list

- Efficient substructure viewing in MSC greatly helps in validating the *most likely* candidate after first sorting by “number of references” field and then the correlation score

Possible Future Work

- Implement the ability to import and sort by references into the next version of MSC supplied with MassHunter and provide the meta information about the number of references in the ChemSpider application programming interface

- Provide a weighted combined scoring between the correlation score and the different no. of references criteria to increase rank of correct structure for known test compounds

- Refine MSC interface for efficient sorting/displaying of ChemSpider reference fields

Acknowledgments

Many thanks to Curt Cleven, Jean Coffman, Mike Ramsey, Stacy Brown, Alexey Pshenichnov, Mike Scott, Bill Tindall, and Kent Morrill for their contributions to the various approaches for identifying “known unknowns.”

References

1. Identification of “Known Unknowns” Utilizing Accurate Mass Data and Chemical Abstracts Service, J. Little, C. Cleven, and S. Brown, JASMS, Vol 22, No. 2 (2011), 348-359.

2. Identification of “Known Unknowns" Utilizing Accurate Mass Data and ChemSpider, J. Little, A. Williams, A. Pshenichnov, V. Tkachenko, JASMS, Vol 23, No. (2012), 179-185.

3. Littlemsandsailing.wordpress.com

4. Automated Assignment of High-Resolution Collisionally Activated Dissociation Mass Spectra Using a Systematic Bond Disconnection Approach, A. W. Hill, R. J. Mortishire-Smith, Rapid Communications in Mass Spectrometry, Vol 19, Issue 21 (2005) 3111-3118.