



MathSpec, Inc.
1-847-840-4994
www.mathspec.com

***Rational Numbers*TM on the SunTM Grid Compute Utility**

Introduction

The relative ease of obtaining accurate-mass fragmentation data on modern LCMS instruments; faster computers; and the availability of large molecular structure databases have recently made it possible to change the “art” of interpreting mass spectral data into a systematic computational process. However, this brute force computational approach is best suited to scientists who have vast computational power at their fingertips. That is why MathSpec, Inc. recently teamed up with Sun Microsystems to deliver *Rational Numbers*TM on the SunTM Grid Compute Utility. For the very reasonable price of \$1/CPU hour¹, mass spectrometrists can analyze their mass spectral data on the Sun Grid Compute Utility. The Sun Grid can now do much of the tedious work, freeing the mass spectrometrist to spend more time on higher level tasks.

What is *Rational Numbers*?

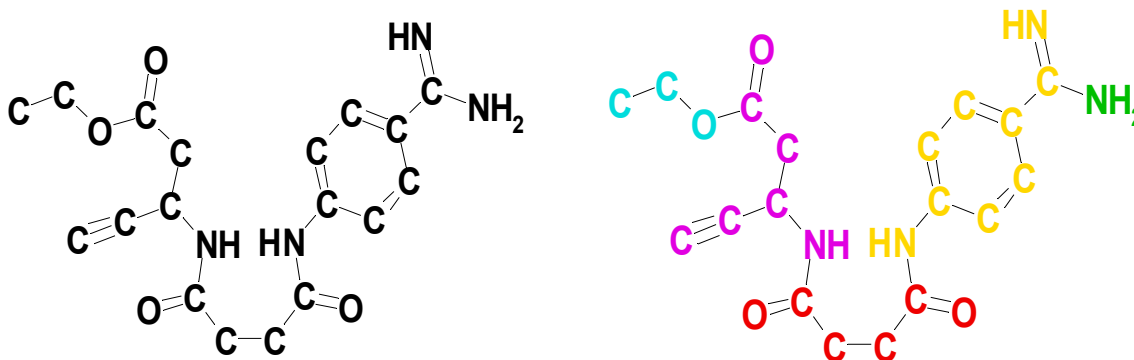
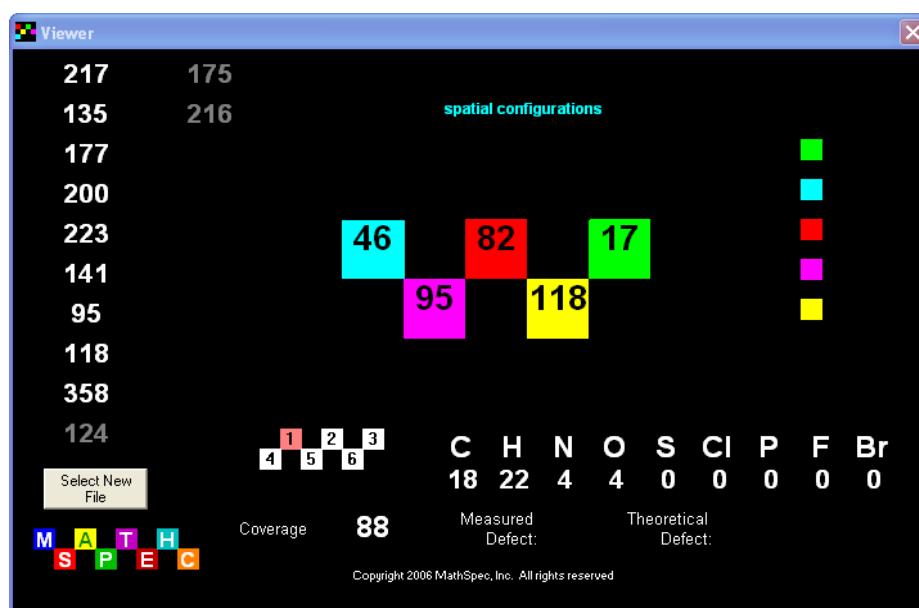
Completely complementary to traditional interpretation and spectral library searching, *Rational Numbers* is a software application by MathSpec, Inc. designed to assist scientists in identifying small organic compounds from mass spectral data. The overall objective is to draw a rough picture of molecules that could yield a particular set of numbers, and, where applicable, to search through molecular structure databases to find matching compounds.

Rational Numbers software treats small molecules as mathematical partitions and then summarizes the mass spectral data in a visual format (subfragments and modular structures). The program has the unique capability of combining up to eight spectral files from different types of mass spectrometers in both positive and negative ion mode, using both accurate-mass data and integral-mass data, and analyzing all of the data from one compound as a single data set. The software strategy is not based on fragmentation mechanisms, but rather on the numbers derived from the mass spectra, atomic properties, and logic tables.

This software tool can do three tasks. If the compound of interest is known, *Rational Numbers* can assign fragment ions in its spectrum (Assign). If the compound is unknown but previously reported, the program can search accurate-mass fragmentation data directly against molecular structure databases (Search); traditional mass spectral libraries are no longer needed. If the compound is novel, *Rational Numbers* can compute and generate modular structures which closely approximate the molecular structure (Partition).

Assign

If the compound of interest is known, *Rational Numbers* can assign many of the fragment ions in its spectrum. Assigning fragments is accomplished by comparing the heavy atom distributions of modular structures to the heavy atom distributions of a single known molecular structure. *Rational Numbers* Assign color-codes black-and-white structures in tgf format to correspond to color-coded modular structures that match the known compound. The color-coding reveals which heavy atoms in the structure are found in which sub-fragment. Below the color-coded molecular structure of xemilofiban is compared to a modular structure. The Assign program requires that the mass spectrometrists have access to ISIS/Draw™ or ChemDraw™ Pro to create the black-and-white structures in tgf format.



Search

Searching is accomplished by comparing the heavy atom distributions of modular structures to the heavy atom distributions of molecular structures found in databases such as PubChem™ (<http://pubchem.ncbi.nlm.nih.gov>). While the modular structures obtained from the program do not have the detail of a molecular structure, the modular structures provide considerable structural information. A modular structure is very similar to the molecular structure; the ordering of the atoms in the sub-fragments cannot be determined from mass spectral data alone.

Rational Numbers can search vast molecular structure databases (e.g. PubChem) for compounds with molecular structures that are consistent with that mass spectral data. Determining the accurate mass of an unknown compound, even getting a single elemental composition, does not necessarily lead to a direct identification. For example, PubChem currently has 1621 compounds with the formula $C_{18}H_{17}N_3O_3S_1$ and 1271 compounds with the formula $C_{15}H_{16}N_2O_2$. By comparing the heavy atom distributions of compounds in a database to heavy atom distributions of modular structures derived from the mass spectral data, the software is able to rapidly zero in on the matching structure. Furthermore, by directly searching molecular structures against the mass spectral data, MSMS libraries are no longer needed. The Search algorithm is very CPU intensive.

Partition

If the compound is novel, *Rational Numbers* can compute and generate modular structures which closely approximate the molecular structure (Partition). While lacking the fine detail of a molecular structure, these modular structures provide structural information that can be combined with background information or information from alternative techniques such as NMR, to help elucidate the correct structure.

The Sun Grid Compute Utility (www.network.com)

Sun is providing easy and affordable access to an enormous computing resource for the predictable price of \$1/CPU-hr. The application catalog feature allows independent software vendors and developers the ability to host their applications on the Sun Grid Compute Utility and deliver them as a service to end users. Sun Grid Customers are just a click away from using applications in an on-demand environment.

The Sun Grid Compute Utility was built using industry best practices and incorporates comprehensive security to protect sensitive data. Based on Sun Fire™ x64 servers, the Solaris™ 10 Operating System, Solaris Containers technology, and Sun N1™ Grid Engine software, the Sun Grid Compute Utility provides a secure, high performance environment for batch workloads.

***Rational Numbers* on the Sun Grid**

Rational Numbers is one of the first commercial scientific applications offered on the Sun Grid. For MathSpec and its customers, the Sun Grid has two big advantages.

First, the Sun Grid offers the ability to run CPU intensive algorithms, such as Search, in parallel. This means that the mass spectrometrist can get results faster, while paying for the same amount of CPU time. In addition, each customer has virtually unlimited computer capacity when large numbers of datasets need to be processed.

Secondly, the local Mac mini implementation requires that mass spectrometrists periodically perform software or database updates. By offering its software on the Sun Grid, MathSpec can update its software and databases without involving its customers; its customers can focus on doing what they do best - mass spectrometry.

Running *Rational Numbers* Search on the Sun Grid

The dataset input for *Rational Numbers* Search consists of: 1) up to eight mass spectral listings which can include MS, MSMS, and MSⁿ data, both accurate-mass and integral mass, and both positive and negative ion modes; 2) a parameter file basically consisting of sample information (e.g. sample name) and constraints such as the mass error window, whether odd-electron sub-fragments are allowed, etc; 3) a compound.txt file which is an abbreviated text file corresponding to a slice of an sdf database (e.g. PubChem) around an exact mass wherein that exact mass slice should include compounds corresponding in exact mass to the unknown compound. In the examples here, the exact mass slice was +/- 7.5 ppm of the experimentally determined molecular weight (MH⁺ - proton). Typically, all of these files are zipped together as one dataset.

To start the run, the spectrometrist first uploads the dataset as a resource, selects the program *Rational Numbers* Search from the catalog and the DET (digital entitlement token²) as an additional resource, enters the total number of compounds in the exact mass slice, and then starts the analysis.

Complex Datasets Run Faster on the Sun Grid with Parallel Processing

The parallel version of the Search program divides the exact mass slice into individual compounds and then generates a separate process for each. Each process could potentially use a separate CPU. At the end of the analysis, the results are summarized and then sorted into one search report, Results.doc.

Five datasets were analyzed on the Sun Grid version 1.00. Each was analyzed 5 times using the parallel version of the program, and 2 times using essentially the same program on 1 CPU as baseline. Times shown in the table are the average of the replicates in units of seconds. Identical search reports were obtained for all data sets whether run on 1 CPU or in parallel.

Search CPU time increases with the molecular weight and the number of compounds in the molecular weight slice with the same formula. The most dramatic decrease in elapsed time (11.4X) occurred with xemilofiban, which had 671 compounds in the exact mass slice with the same formula. Malathion, which had only six compounds with the same elemental composition in its exact mass slice, actually took 144 seconds longer to run in parallel. Running in parallel is preferred because it is usually not possible to predict *a priori* how much CPU time will be required for an analysis.

Compound	MW	Num of Compounds in PubChem slice	Compounds with Same Elemental Composition	Search Time (sec)1CPU	Search Time (sec) parallel	Elapsed Time 1CPU (sec)	Elapsed Time Parallel (sec)	Reduction Factor 1CPU/ parallel
Xemilofiban ³	358	840	671	4536	4378	4597	403	11.4
LeuEnk ³	555	122	18	986	1001	1015	306	3.3
compound D ³	499	42	5	180	180	227	158	1.4
compound B ³	361	396	112	875	864	914	205	4.5
Malathion ⁴	330	312	6	14	18	29	173	increased

Bottom Line: *Rational Numbers* on the Sun Grid saves you time

Suppose a mass spectrometrists was identifying what eventually would prove to be leucine enkephalin from its mass spectral data by comparing that data to all 122 structures in PubChem that had an exact mass within 7.5 ppm of the experimentally determined value.

Without *Rational Numbers*, let's assume that it would take require about 2 minutes to check each structure against the dataset. Assuming no breaks or interruptions, it would take 4 hours to complete this rather tedious task.

That same job using *Rational Numbers* on the Sun Grid required 5 minutes of elapsed computer time. Assuming that it would take an additional 5 minutes to upload the dataset and download the results, the total elapsed time would be about 10 minutes, and for half of that time, the Sun Grid would be doing all of the work.

Using the Sun Grid allows the mass spectrometrists to meet tight project timelines without constantly working overtime.

Move ahead with MathSpec and Sun

MathSpec, Inc. develops innovative, leading edge software that enables users to identify compounds faster. For over 20 years, Sun has continued to create innovative high performance computing solutions, like the Sun Grid Compute Utility, that help organizations run larger numbers of compute tasks faster. Together, MathSpec and Sun enable organizations to speed their research and move projects to market faster.

Acknowledgments

MathSpec would like to thank our partners at Sun Microsystems for their assistance on this project.

-
- 1) License fees also usually apply.
 - 2) The DET is essentially a digital verification of a valid license.
 - 3) Sweeney, D. L., Anal. Chem. 2003, 75(20), 5362-5373
 - 4) Thurman et. al., Anal. Chem. 2006, 78(19), 6703-6708).

Rational Numbers is a trademark of MathSpec, Inc.
PubChem is a trademark of the National Library of Medicine, NIH
ChemDraw is a trademark of CambridgeSoft Corporation
ISIS/Draw is a trademark of Elsevier MDL
Sun, Sun Microsystems, the Sun logo, N1, Solaris, and Sun Fire are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and other countries.

Information subject to change without notice.