



High Performance Discovery Software

SysChem Scientific White Paper

Lynn Settle
Chief Software Architect

Robert Bunn
Director of R&D

SysChem Inc.

Version 1.4 (1B)
Revision of July 12, 2006

1. Overview

This White Paper supercedes and replaces all earlier editions.

It reflects several important changes in SysChem's corporate operating strategy and in the development of SysChem's proprietary software product, SystematiChem©.

Most important among the changes, **SysChem has reconfigured its business model** so that it functions as a service bureau, rather than a seller of software.

SysChem now retains complete control of SystematiChem©. This means SysChem can better control the quality of the solutions generated. It also means only SysChem personnel with the highest security authorization have access to critical programs, significantly increasing the level of security SysChem can guarantee.

The paper anticipates **an upgraded Solution Viewer** that will enable the customer to effectively manipulate the output of SystematiChem© without accessing SysChem's proprietary programs.

The White Paper also incorporates the most recent improvements to Version 1 of its SystematiChem© software. Version 1B – as this version has been designated – will continue to experience minor fixes. Major new developments, however, will be incorporated in Version 2.

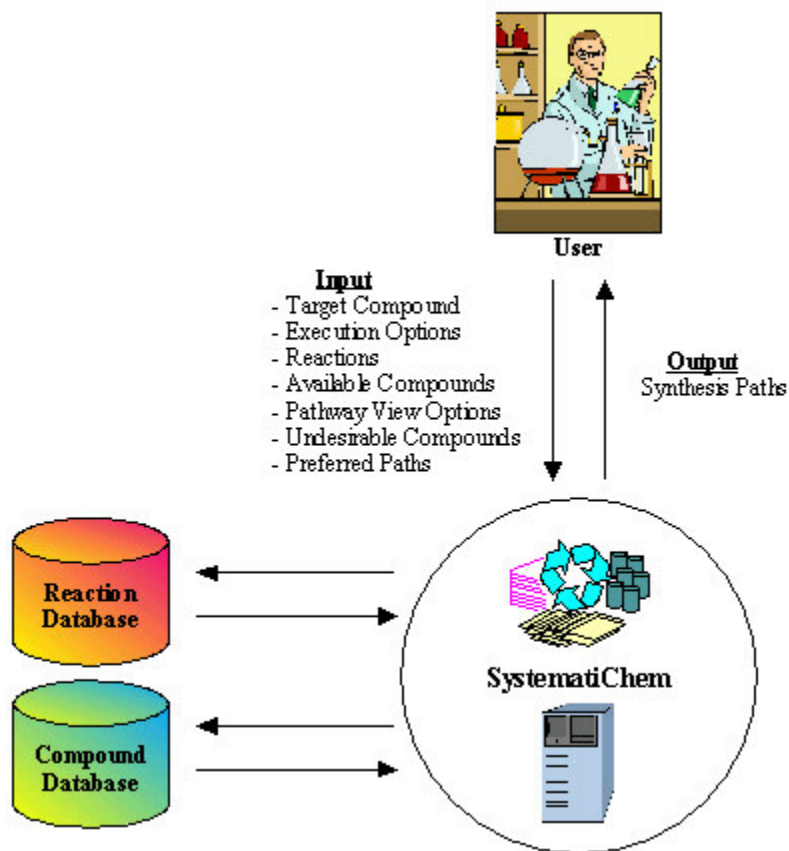
As a consequence, the “wish list” of enhancements to Version 1 – found in earlier White Papers – has been purged. This paper deals only with what is currently in place or can be easily accomplished through minor modifications to Version 1.

The pages that follow describe the design, functionality and capabilities of Version 1B of SystematiChem©.

View from the top:

SystematiChem© is a high-performance discovery software that, when applied to organic chemistry, has the potential to greatly accelerate industrial, governmental, and academic research. SystematiChem© is the first product to fully exploit computational organic chemistry and computer science advances to solve real world commercial applications.

From a very high level view, the user provides the target compound. The user receives the candidate solution pathways to evaluate for viability, using SysChem's proprietary Solution Viewer, which is distributed to customers and others with a legitimate interest in the product at no cost.



The remainder of this document will dissect the above diagram into its various components and explain its application. The system's application will be discussed with a general explanation of the database design organization and content. From there the focus will move logically through the SystematiChem© software design itself.

2. SystematiChem© Functionality and Data

2.1 SystematiChem© database

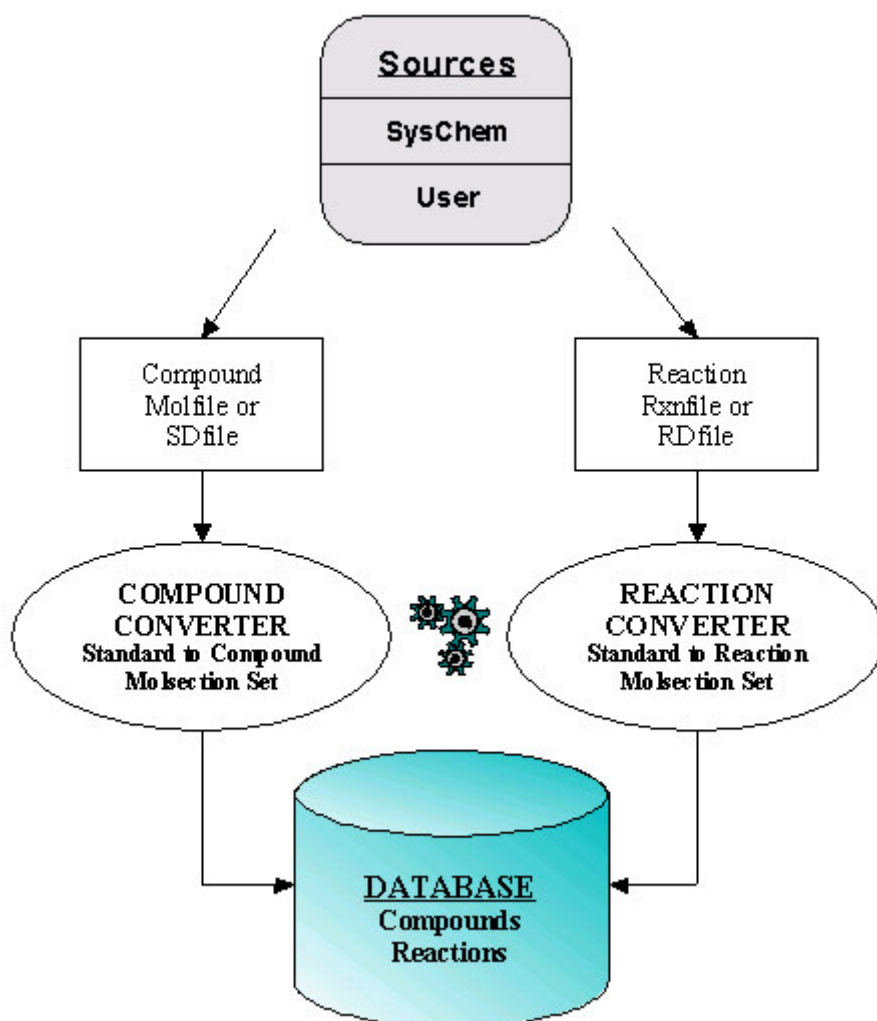
Flexible: Prior to running a particular target compound, SysChem's chemists select the reaction and compound source files that will produce the ideal results for that type of compound. SystematiChem© is organized to allow the maximum possible number of varying sets of available compounds and reactions to be utilized from SysChem's database. This allows SysChem to target specific types of compounds to significantly and substantially improve the candidate synthesis routes generated

Extensible: SystematiChem© can incorporate a customer's proprietary reactions into SysChem's database for secure, exclusive operations upon the submitted compounds.

Adaptable: The design enables SysChem's chemists to improve the pathway generation by defining initial reactions that are not feasible and/or desirable. For example, depending upon the chemist's goals with the particular compound, syntheses using more expensive reactions may be completely filtered from the generated results.

2.2 Database conversion

Before compounds and reactions can be loaded into the software's database, they must first be individually converted into "molsection sets". The software database retains the original compound molfile and reaction rxnfile formats for reference by outside drawing tools such as MDL Chime© or JMol©.

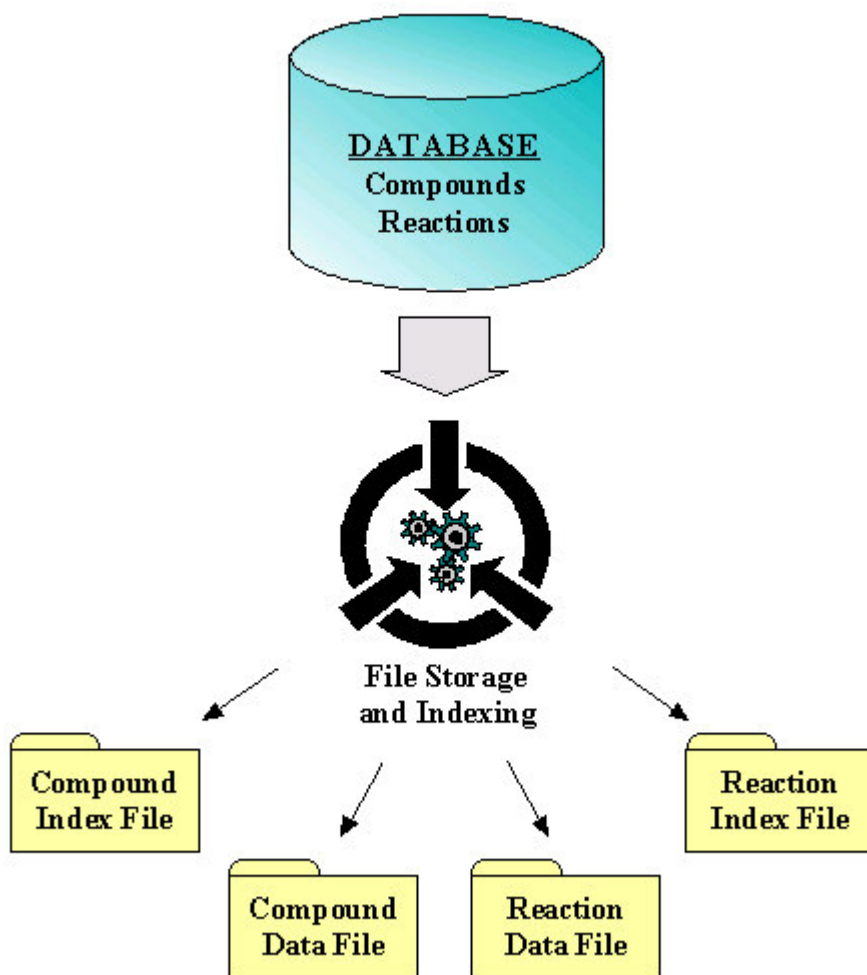


2.3 Discovery process data files

An SQL database is capable of storing large amounts of compound and reaction data. Unfortunately the SQL format has two serious shortcomings that make it undesirable for direct use by the automated discovery process.

- 1) Slow data access. An SQL database requires access to the hard drive. Fast processing absolutely requires 100% RAM data access.
- 2) No compatible indexing. Complex data like the compound and reaction molsection sets cannot be indexed in an SQL database.

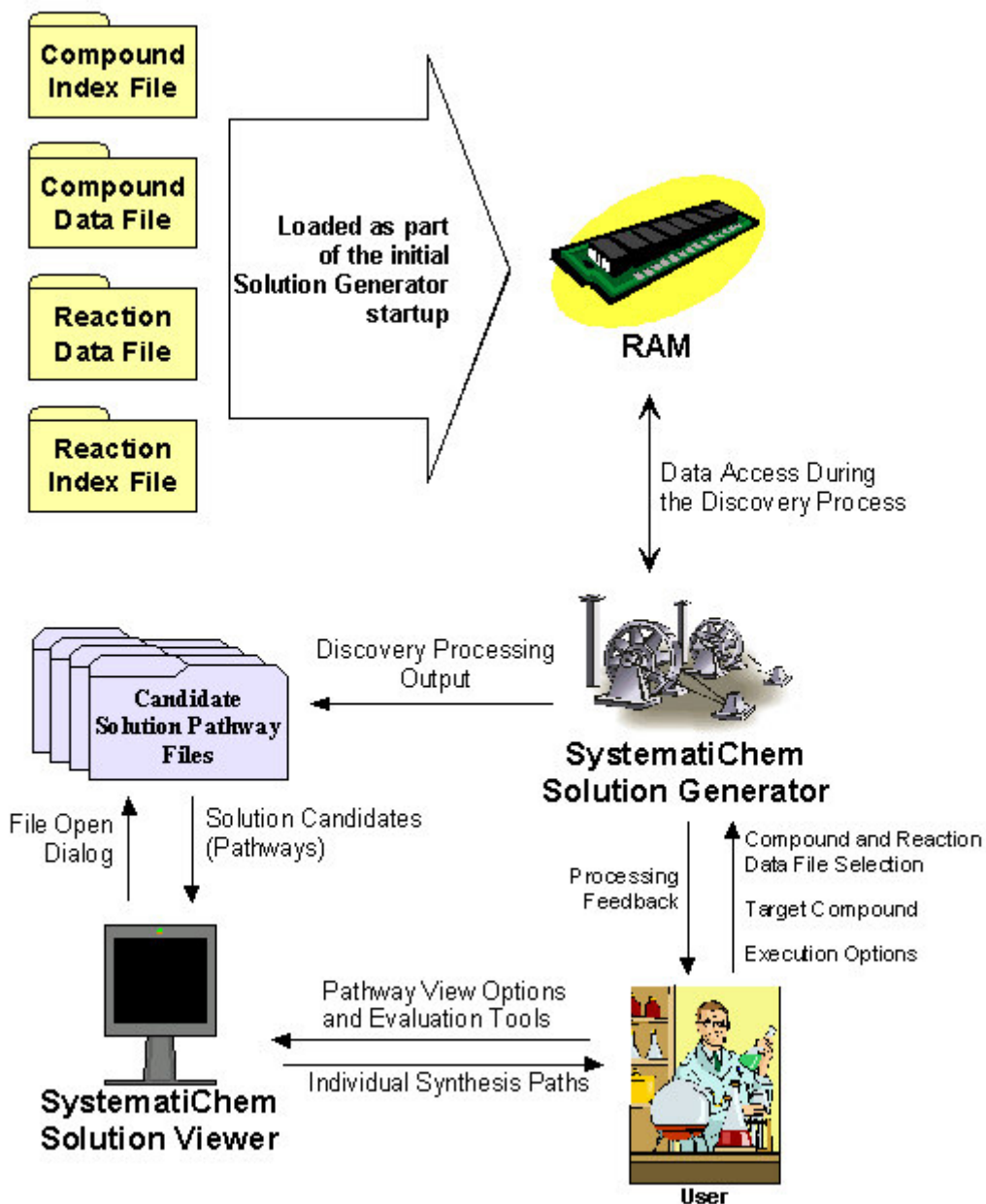
To resolve these problems, a separate program called the “Data Loader” extracts from the SQL database the desired compound and reaction data. SysChem identifies various combinations of reaction and compound categories to be extracted. This allows the creation of specialized file sets.



The files containing the exported data are then ready for use by the discovery process.

2.4 Discovery processing

With the desired data files selected, SysChem's in-house operators start the SysChem Solution Generator. The data files are loaded into RAM. All compounds and reactions now reside in memory and are quickly accessed through the indexing information.



The Solution Generator is given a target compound and begins the discovery process. The generated solution candidates are stored in a separate file. The Solution Viewer allows SysChem's chemists and customers to open these solution files to evaluate the candidate pathways. Filtering tools assist the user in finding desired synthesis paths.

2.5 SystematiChem© solution generator

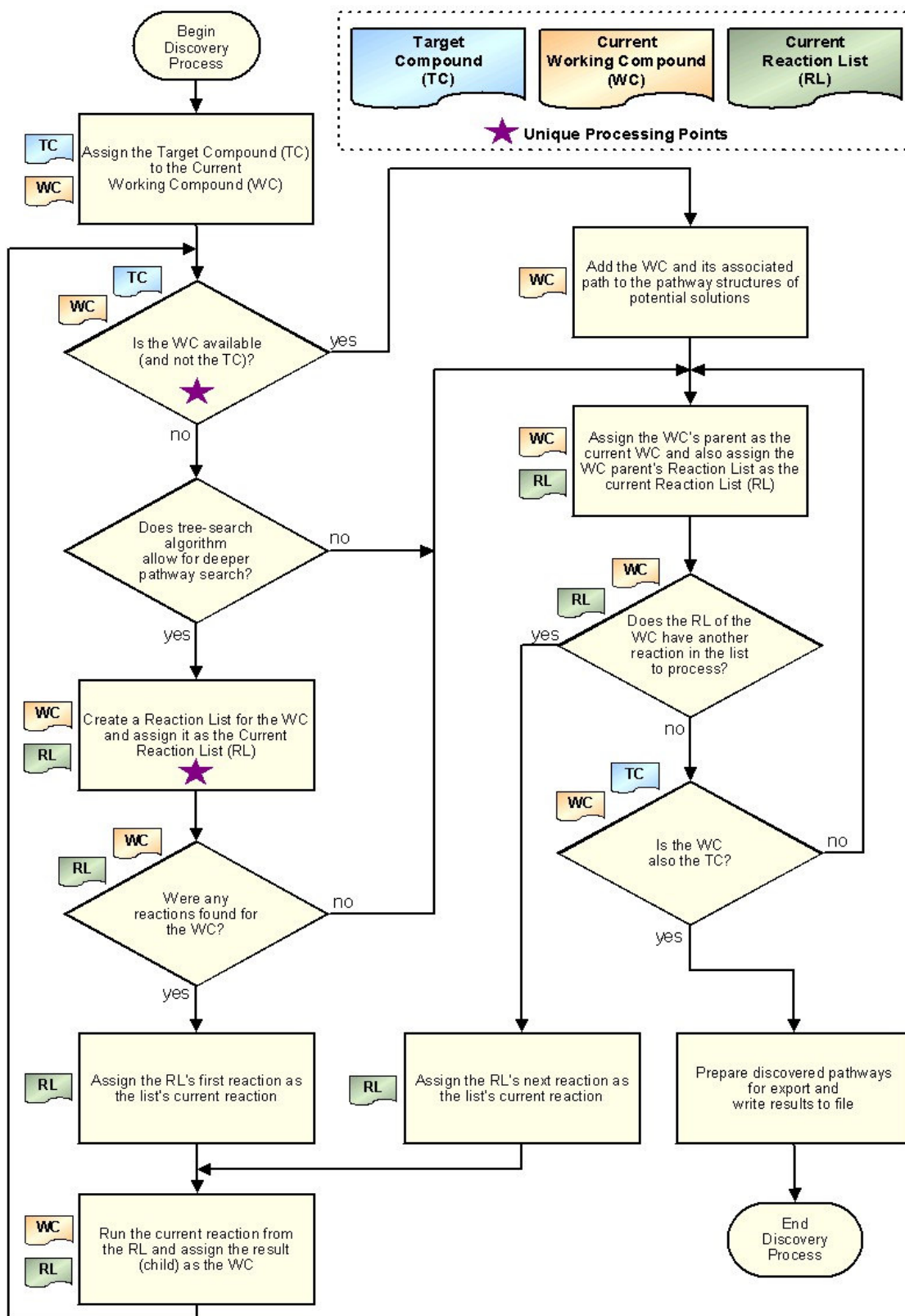
The Solution Viewer is exactly that – a viewer. While no processing takes place within the viewer, the Solution Viewer allows the user to view and manipulate the proposed synthesis routes.

The synthesis pathway results are displayed in the Solution Viewer. The Solution Viewer provides methods to assist the user in evaluating the candidate synthesis paths. These pathways are prepared and stored into a desktop database file by the Solution Generator.

The Solution Generator is the heart of the discovery mechanism and is where the actual discovery process takes place.

The Solution Generator Process Flow diagram (below) provides a high-level, generalized portrayal of the essential processing that takes place within the Solution Generator. It does not detail any of the complexities associated with the optimized synthesis path tree-search algorithms, handling multiple reactants, manipulation of data structures, memory caching, and many of the other important components of the process.

Solution Generator: Process Flow



The diagram's purpose is to identify the two Unique Processing Points (UPP's) that will be covered in greater detail. These points are identified by the two flowchart symbols that contain a star.

- **UPP #1:** Is the Working Compound Available?
- **UPP #2:** Create a Reaction List for the Working Compound and assign it as the Current Reaction List.

Two fundamental questions immediately arise:

Question #1: What makes these processing points unique?

Question #2: Why are none of the many other processing areas identified as unique?

Question #2 is the easiest to explain and will be answered first.

The other processing areas all involve programming techniques that have been repeatedly employed and improved upon for many years. Extensive literature is available that discusses multitudes of tree searching techniques. Data modeling design, memory management routines, and optimized coding techniques are all well documented and have been programmed innumerable times into many different software projects. Thousands of hours have gone into the design and development of these processes in the current Solution Generator. These processes are both substantial and valuable, but not unique functionalities in and of themselves.

The answer to Question #1 requires a more detailed discussion of UPP #1 and UPP #2, which follows in sections 3.1 and 3.2.

3. SysChem Capabilities

3.1 UPP #1: Determining working compound availability

The challenge: Determining working compound availability requires a search into a database of tens of thousands of organic compounds. A compound is primarily defined as a set of elements with specified bond relationships. Due to the nature of this complex data representation, this search will be relatively slow and tedious. Some vendors have created fast compound searches using chemical names, CAS numbers, molecular formulas, and molecular weights. Unfortunately such lookups are useless for an automated retrosynthesis program, which must search for and match the compound structure itself.

For example: Consider an incredibly optimized compound search mechanism that requires as little as 10 seconds to search 100,000 compounds for a structure match.

Assume that all compounds average 50 qualifying reactions and result in a total of only 60 reactants (allowing for a few multiple reactants).

The processing time for just the first step of the compound search would require (10×60) seconds, or 10 minutes.

The second step would involve 3600 compounds (60×60) . This level would require (3600×10) seconds, or 10 hours.

The third step becomes a serious challenge, with 216,000 compounds $(60 \times 60 \times 60)$, requiring $(216,000 \times 10)$ seconds, or 25 days.

The fourth step must be measured in years, as many compounds can easily have over 200 reactants.

The above example only includes the compound structure searches. The example does not account for the extensive reaction logic and other processing required for a truly effective automated retrosynthesis discovery product.

The problem attendant with compound structure search speed is one of the two key obstacles for an automated and essentially brute-force approach to overcome the combinatorial explosion problem of retrosynthesis.

SysChem's Solution: What if an indexing algorithm existed that could accurately search the many thousands of available compounds in a fraction of a second? What if the average search times were measured not just in milliseconds, but by *microseconds* on a standard desktop computer?

Benchmark tests: SysChem's compound searches typically average approximately one microsecond on a dual core desktop workstation.

Unique Processing Point conclusion: Regarding the handling of issues associated with the combinatorial explosion problem, the SysChem design approach addresses the requirement for an efficient search of available compound structures.

3.2 UPP #2: Creating a reaction list for the working compound and assigning it as the current reaction list

The challenge: This requires a search into a database of thousands of reactions. The search must determine those reactions that qualify for execution upon the target compound and intermediary compounds within the pathways being analyzed.

For a variety of reasons a retrosynthetic structure search of valid reactions for a compound is much more complex than the compound structure search. The most significant difference is that where compound searches are looking for “exact matches”, the products of valid reactions represent a subset of the compound. An additional difference is that the retrosynthetic structure search may be required to return dozens of reactions, and unlike compounds, the search must accommodate potential reaction groups as they might retrosynthetically apply to the working compound.

The SysChem technique: Efficiently finding valid reactions in a retrosynthetic direction is resolved using a derivative of the same approach as SystematiChem’s© compound search mechanism. The distinctive requirements necessary to locate the reaction subsets to the compound are resolved in the search logic that wraps the indexing mechanism.

Benchmark test restraints: The reaction search mechanism cannot be specifically benchmarked by itself. Its logic is closely interwoven within the synthesis path tree-search algorithms, the reaction execution routines, and the compound search mechanism. Benchmark tests have been conducted on this overall process to help evaluate the speed of the reaction search and reaction execution routines.

Benchmark tests: The benchmarked processes include identifying the compound’s qualifying reactions, running these reactions to create their associated precursors, and identifying those precursors which are available compounds. The benchmarks indicate that this set of tasks is performed in an average of fewer than 5 milliseconds for synthesis path termination compounds and an average of fewer than 10 milliseconds for synthesis path continuation compounds.

Unique Processing Point conclusion: The benchmark tests indicate expeditious performance levels for the reaction search and execution algorithms. The SystematiChem© design approach for the working compound reaction search portion of the combinatorial explosion problem is efficiently addressed.

4.0 Recent Enhancements and Future Releases

SysChem has been continually improving its current set of Version 1 software programs.

SysChem recently completed Version 1-B, which included a wide variety of improvements. The most significant new feature is the display of full secondary and duplicate reaction information within the solution viewer, as well as duplicate compound information.

As powerful and valuable as the SystematiChem© Version 1 releases are today, SysChem accepts the challenge of producing the next breakthrough in automated retrosynthetic software.

The development of Version 2 is already well underway. This is not a mere upgrade, but a ground-up rewrite that takes full advantage of the Version 1 experience.

SysChem expects Version 2 to be no less than the next quantum leap forward in chemistry syntheses software.