

EXACT MATCH SEARCHING

INTRODUCTION

Institutions find it valuable to maintain a registry of all the chemical substances that have prepared. The purpose of this registry is to determine whether a new material is novel, or another example of a previously registered material. The properties, be they physical – melting point, boiling point, viscosity, crystal form (polymorph) - or biological – toxicity, activity in an assay – also need to be stored and referenced to the relevant chemical structure.

Registry and retrieval systems require exact structural matches. Such matches have to be performed frequently to confirm novelty for a structure, or assign as a new batch (sample) of an existing substance. Accuracy is of course important for exact structural matches, but the process also needs to be fast. Reliability and speed of searching is greatly simplified if a canonical¹ name can be generated for a structure, and the process is further simplified if that can be machine generated.

CHEMICAL STRUCTURE GRAPHS

In chemistry a graph makes a natural model for a structure, where vertices represent atoms and edges bonds. Graph theory² is an extensive area of mathematics, and reducing the chemical structure to a graph enables it to be handled efficiently using developed mathematical principles.

It is then possible to process the input chemical structure by applying a set of rules that will organize the input graph into a canonical form. This canonical graph is then processed to produce a name.

In this context the name is not like the names that we are familiar with. Firstly it is unique, secondly it is probably impractical for a human to comprehend or even remember.

-
- ¹ In computer science, canonicalization is a process for converting data that has more than one possible representation into a “standard” canonical representation. This can be done to compare different representations for equivalence, to count the number of distinct data structures, to improve the efficiency of various algorithms by eliminating repeated calculations, or to make it possible to impose a meaningful sorting order. (<http://en.wikipedia.org/wiki/Canonicalization>, accessed July 22, 2011; mathematical structures used to model pairwise relations between objects from a certain collection. A “graph” in this context refers to a collection of vertices or ‘nodes’ and a collection of edges that connect pairs of vertices. (http://en.wikipedia.org/wiki/Graph_theory, accessed July 22, 2011).
 - ² W. J. Wiswesser, *Comput. Automat.*, 19, 2 (1970); E. G. Smith, “The Wiswesser Line-Formula Chemical Notation,” McGraw-Hill, New York, NY, 1968.

THE MORGAN ALGORITHM

Many structure naming systems have been developed including Wiswesser Line Notation (WLN)³, and DENDRAL⁴, but the Morgan Algorithm⁵, named after its inventor H L Morgan, has proved to be the most enduring, and it underpins many of the approaches in use today. The Morgan algorithm proved to be an excellent approach for computer applications because it is readily converted into a more traditional connection table representation of the structure, but it does not handle stereoisomers.

The Morgan name became the basis of the CAS registry system.

SEMA

Wipke⁶ and Dyott extended the Morgan algorithm to handle stereoisomerism. The resulting Stereochemically Extended Morgan Algorithm (SEMA) was adopted by Molecular Design Limited (now Accelrys Technologies inc.) and became a major feature of MACCS (Molecular ACCess System).

HASH FUNCTIONS

The final step required to produce an efficient exact structure searching system is to convert the name to a hash code⁷. The hash codes are used to build a hash table. The input structure is named, the name hashed and the hash code is search in the hash table. The result of the search is a pointer to an entry in the database if the structure is already registered, or a null return if the structure is novel.

Hash codes may not be unique, and occasionally one code can represent more than one structure. In this case the hash table points to more than one structure, and it is a relatively fast process to compare the name of the input structure with the names of the registered structures.

The SEMA hashing algorithm is described by Wipke, Krishnan, and Ouchi.⁸

3 J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *J. Amer. Chem. Soc.*, 91,2973 (1969).

4 H. L. Morgan, *J. Chem. Doc.*, 5, 107 (1965)

5 W Todd Wipke, with Stuart Marston, and Stephen Peacock, was a founder of Molecular Design Limited: http://en.wikipedia.org/wiki/MDL_Information_Systems (accessed July 22, 2011)

6 W. T. Wipke and T. M. Dyott, *J. Amer. Chem. Soc.*, 96, 4825, (1974).

7 A **hash function** is a reproducible method of turning some kind of data into a (relatively) small number that may serve as a digital "fingerprint" of the data. The algorithm "chops and mixes" (i.e., substitutes or transposes) the data to create such fingerprints. The fingerprints are called **hash sums**, **hash values**, **hash codes** or simply **hashes**. (Note that hashes can also mean the hash functions.) Hash sums are commonly used as indices into hash tables or hash files: http://en.wikipedia.org/wiki/Hash_code (accessed July 22, 2011)

8 W. T. Wipke, S. Krishnan, and G. I. Ouchi, *J. Chem. Inf Comput. Sci.*, 18, 32, 1978

The characteristic of hash table methods is that the search time is independent of the number of records in the file⁹. This is because it is possible to derive an index from the hash code that points to the position in the hash table where the structure will be if already registered.

DEVELOPMENT OF SEMA

Accelrys extended its chemical representation to include polymers¹⁰, generic structures¹¹, and relative and mixed stereogenic centers¹². These enhancements required extensions to the SEMA name.

In 2006 Accelrys added support for rotationally restricted biaryls, and allenes to Accelrys Direct. The need to support these types of stereogenic centers, and allow extensions into the higher orders of stereogenic centers found in organometallics and transition metal complexes demanded a major revision to SEMA.

NEMA

In the past 30 years, efforts have been made to improve the SEMA program. However, several known defects still exist. Furthermore, SEMA also lacks some important features, such as non-tetrahedral stereochemistry perception. This has led to the development of a new method for stereochemistry perception, the NEMA (Newly Enhanced Morgan Algorithm) method.

The NEMA method was developed to address all the deficiencies that had been identified in the SEMA algorithm. It calculates equivalence classes for a given molecular structure, to perceive tetrahedral and geometric stereogenic centers, to verify non-tetrahedral stereogenic centers, to perform absolute stereocollection demotion. NEMA supports both 2D and 3D stereochemistry perception.

NEMA consists of the following procedures: Constitutional equivalence class calculation, stereochemistry perception, and stereo collection type modification. In the first stage, constitutional equivalence classes are calculated from the topology of the molecular structure and its properties (such as atom type, atom change, etc.) using an improved Morgan algorithm. In the second stage, the stereogenic centers are perceived based on the so-called Morgan-parity values, which are calculated from the EC (Extended Connectivity) values. The EC values are originally generated in the first stage and may be updated several times in the second stage. In the last stage, the demotion of an absolute stereo collection to a racemic stereo collection is performed, if needed. However, it should be noticed that this last step is only necessary for query structures during the structure exact matching process.

9 (a) W. D. Maurer and T. G. Lewis, "Hash Table Methods", *Comput. Surveys*, 7, 5-19 (1975); (b) D. E. Knuth, "Sorting and Searching", *Art Comput. Programming*, 3, 506-549 (1973); (c) D. E. Knuth, "Algorithms", *Sci. Am.*, 236, 63-80 (April 1977).

10 A. J. Gushurst, J. G. Nourse, W. D. Hounshell, B. A. Leland, and D. G. Raich, *J. Chem. Inf. Comput. Sci.*, 31, 447, 1991

11 B. A. Leland, B. D. Christie, J. G. Nourse, D. L. Grier, R. E. Carhart, T. Maffett, S. M. Welford, and D. H. Smith, *J. Chem. Inf. Comput. Sci.*, 37, 62, 1997

12 See whitepaper on Accelrys' Enhanced Stereochemical Representation.

NEMA PERCEPTION OF STEREOGENIC CENTERS IN ALLENES

NEMA correctly determines that the allenic stereogenic center in structure 1 is not valid, but that in structure 2 is valid. SEMA cannot solve such problems. Figures 1 and 2 show the structures with their arbitrary atom numbers displayed. Figures 3 and 4 show the structures with the calculated stereochemically aware Extended Connectivity (*atStereoEC*) values that are calculated by NEMA.

NEMA determined that the allene non-tetrahedral stereogenic center in structure 1 is invalid because in this structure atoms 40 and 44, 50 and 55 are equivalent, respectively, as indicated by the *atStereoEC* values of 839 (for atoms 40 and 44), 836 (for atoms 50 and 55). NEMA determines, however, that the allene non-tetrahedral stereogenic center in structure 2 is a valid one. This is because atoms 10 and 14 of this structure are not equivalent, they have the different *atStereoEC* values of 247 and 246, respectively; atoms 22 and 28 are not equivalent either, they have the different *atStereoEC* values of 246 and 247, respectively.

Structure 1

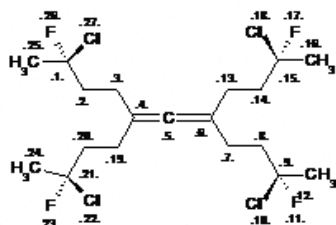


Figure 1

Structure 2

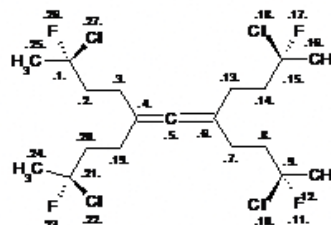


Figure 2

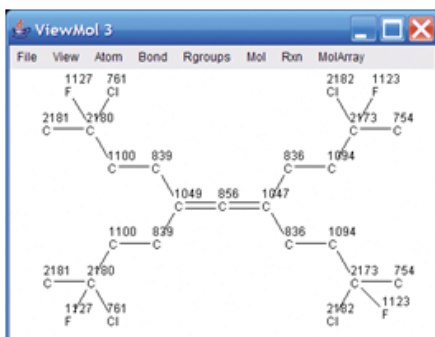


Figure 3

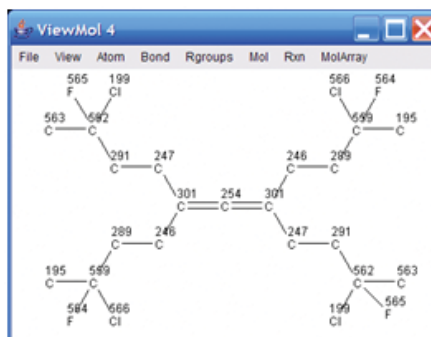


Figure 4

NEMA PERCEPTION OF STEREOGENIC CENTERS IN BIARYLS

Similarly, NEMA detects that the biphenyl non-tetrahedral stereogenic center in structure 3 is invalid, and that in 4 is valid. Verify these determinations by examining the *atStereoEC* values for atom pairs 2, and 4 and 8, and 11 in the two structures.

Structure 3

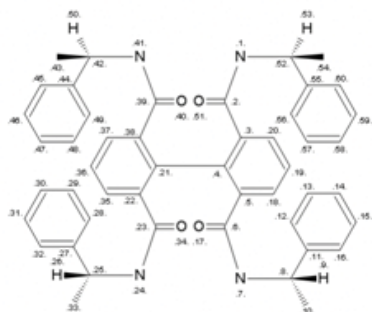


Figure 5

Structure 4

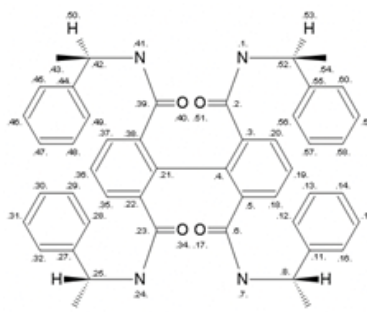


Figure 6

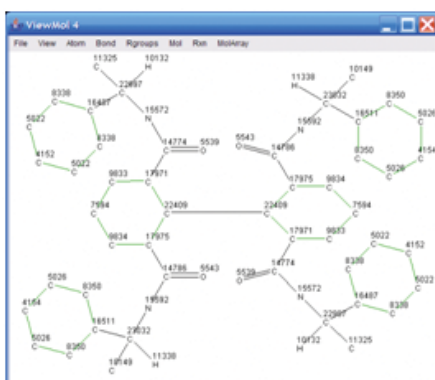


Figure 7

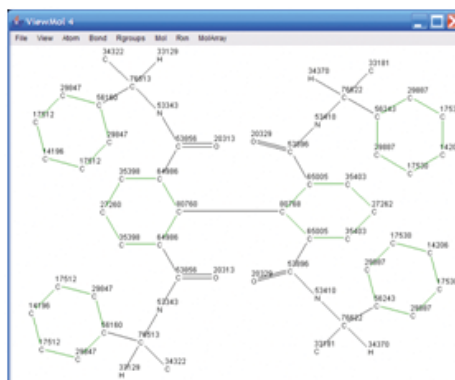


Figure 8

EXACT AND PARTIAL MATCHING

Wipke¹⁰ et al noted that there is value in doing partial matching of stereoisomers for query situations. Confirmation of uniqueness requires an exact match, but chemists are often interested in all the stereoisomers of the query. Their approach was to calculate a full SEMA key, and a constitution key derived from the SEMA name by ignoring the information about the stereogenic centers. NEMA follows this approach and stereo and constitution keys are generated. The keys derived from Lipitor (5) and its enantiomer (6) are shown in the following table.

Structure 5

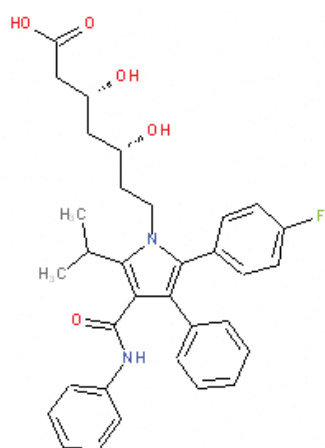


Figure 9

Constitutional Key N5D5UGJE96QG YM2WY8CZHUF8WTA4W9
Stereo Key Q3NBHYFBCZG BPVU4QFM11HN9YYDK7P

Structure 6

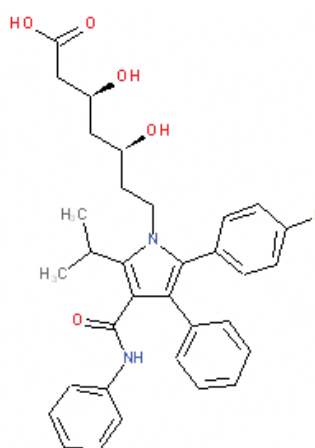


Figure 10

Constitutional Key N5D5UGJE96QG YM2WY8CZHUF8WTA4W9
Stereo Key ST54DVA3FFCEMNSSEN4NV2M49HV1ZQ

The isomers generate the same constitutional key, allowing a query to retrieve both structures, but the stereo keys are unique, and will not find other isomers.

FAST DEDUPLICATION

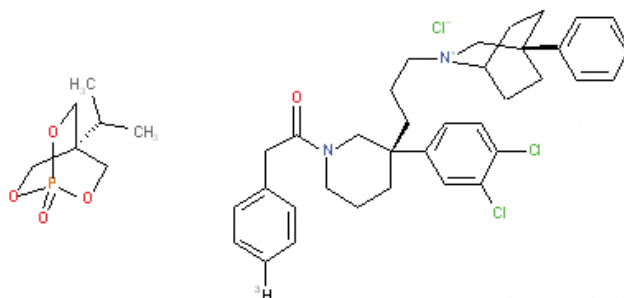
The NEMA key provides a very efficient method for duplicate checking, and the de-duplication of lists of structures. NEMA key technology is incorporated in Accelrys Direct, Accelrys Draw, Pipeline Pilot, and Accelrys Cheshire.

FURTHER APPLICATIONS OF NEMA

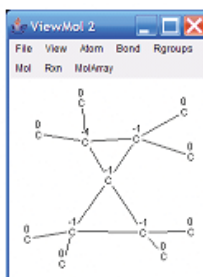
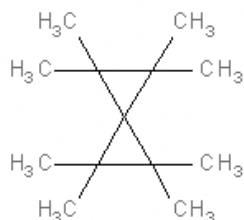
Further investigations are under way to determine what other useful capabilities can be developed using the NEMA technologies. NEMA can be used to create a new structure similarity search method that allows the retrieval of structures in which an atom has a specified number of equivalent attachments. The new similarity search has the following characteristics:

- The “center” atom can be either a stereogenic center, or not a stereocenter, or any atom.
- The equivalence of the attachments can be either constitutional or stereochemical.
- The attachments can be specified as either all ring members, or all chains, or any.

This type of structure similarity search features is not available in the traditional structure search approaches. Examples of this type of similarity search follow:



Query: Find out structures in which a marked stereogenic atom has exactly 3 constitutionally equivalent attachments that are all ring members.



Query: Find out structures in which a marked stereogenic atom has exactly 4 constitutionally equivalent attachments that are all ring members.

In this example the atom values calculated by NEMA are shown.

NEMA KEYS FOR DEALING WITH BIOPOLYMERS

Most recently, a new type of the NEMA keys called the sequence NEMA key was developed specifically for fast exact matching and deduplication of biopolymers. The new technology is discussed in detail in the paper "Self-Contained Sequence Representation: Bridge the Gap between Bioinformatics and Cheminformatics"¹³.

Acknowledgements

Accelrys's chemistry representation team (CHRP) in particular Burt Leland and Jim Nourse, and Lingran Chen, the author of NEMA, who provided the much of the background information and the graphical examples.

13 W.L. Chen, B.A. Leland, J.L. Durant, D.L. Grier, B.D. Christie, J.G. Nourse, K.T. Taylor. Self-Contained Sequence Representation: Bridge the Gap between Bioinformatics and Cheminformatics J. **Chem. Inf Mod.** Submitted for publication. 2011.

APPENDIX A: SHORTCOMINGS OF THE SEMA METHOD ADDRESSED BY NEMA

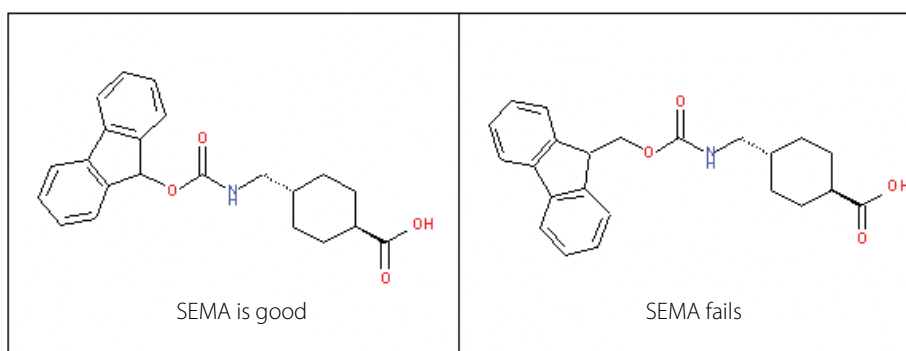
The following are some examples problems in the SEMA method. Note that the representations follow the IUPAC recommendations and absolute centers are not marked.

SEMA gives inconsistent results for symmetrical structures with minor structure differences.

Structures 7 and 8 are very similar and both differ by a CH₂ group in the chain that links the two ring systems. This substituent does not affect their stereochemistry. It should also be noticed that both structures were marked as chiral, which is incorrect. SEMA correctly perceives structure 7, but gives an incorrect result for structure 8. NEMA correctly perceives both structures.

Structure 7

Structure 8



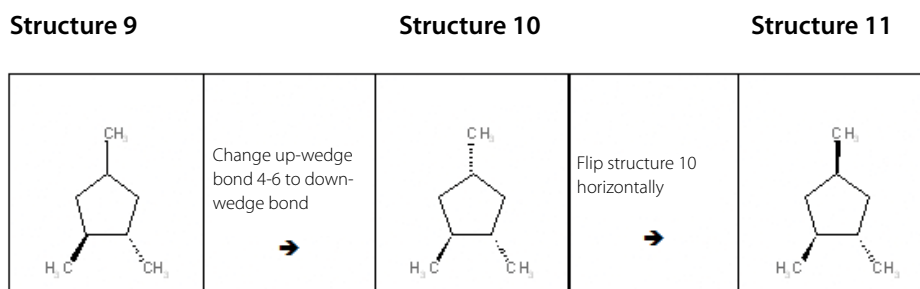
SEMA has difficulties in perceiving the correct tetrahedral stereocenters for certain types of cyclic compounds.

First it is necessary to understand that in Accelrys' Flexmatch technology, a marked chiral center is treated rigorously, thus it is essential that only stereogenic centers that are structurally differentiating be marked. In structure 9, the stereogenic center at atom 10 is degenerate, due to symmetry in the structure. If it is marked as up, and the structure in the database is marked as down, then the query will not match the database entry. For this reason Accelrys' chemistry perception needs to remove any marking on such an atom during the registration process, and similarly it must remove degenerate marks on exact match queries. If this is not done structures may not find themselves, and duplicate created unintentionally.

This example executing the following process converts structure 9 to structure 10, and then rotation converts structure 10 to structure 11, which is identical with structure 9:

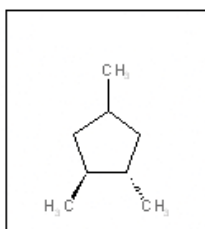
1. Change the up-wedge bond 4-6 of structure 9 into a down-wedge bond, leading to structure 10.
2. Flip structure 10 horizontally, leading to structure 11.

Comparing structure 11 with structure 9, it can be easily seen that these two structures are identical except the atom numbering. Since the flipping of a structure horizontally does not modify the structure itself at all, structure 10 is identical with structure 11, and thus structure 10 must also be identical with structure 9. This means that changing the up-wedge bond 4-6 to the down-wedge bond in structure 9 does not modify its structure at all. Therefore it is necessary to demote it to a plain bond.



Structure 9 needs to be modified to structure 9a on registration to ensure that it can be found by similarly flattened queries.

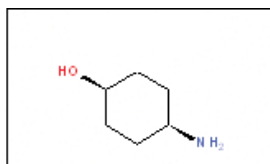
Structure 9a



SEMA has difficulties in supporting the absolute stereocollection demotion.

For example, currently, the SEMA based stereo perception method can demote an absolute stereo collection to a racemic stereo collection for some simple meso-structures like structure 12. This method, however, is unable to handle more complicated structures like structure 14.

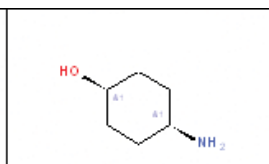
Structure 12



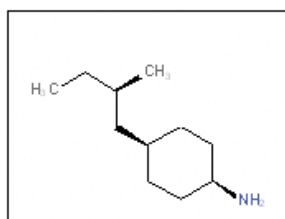
Absolute to racemic demotion

SEMA is OK.

Structure 13

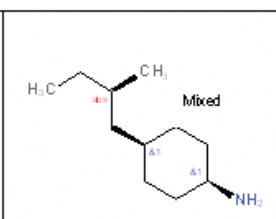


Structure 14

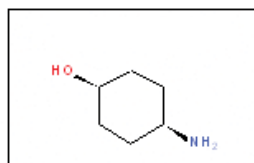


Partial absolute to racemic demotion

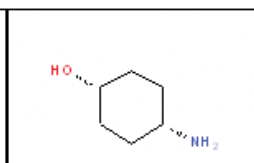
Structure 15



Structure 12



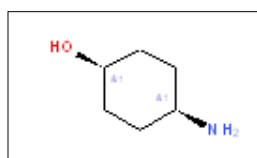
Structure 12a



In these examples, structures (12) and (12a) are both valid representations of the same structure. But in an exact structure search (12) would not match (12a) (and vice versa). It is, therefore, essential that the absolute stereogenic centers be demoted to AND (racemic) centers on registration, structure (13).

A similar process is undertaken for queries that contain this motif. This insures that no matter how the query is drawn it will find itself. NEMA handles this demotion for structures like (12) and (14).

Structure 13



SEMA does not support non-tetrahedral stereochemistry perception

This is described in the main body of this whitepaper.