JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

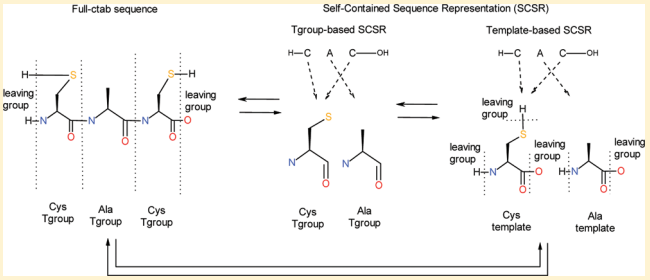
Self-Contained Sequence Representation: Bridging the Gap between Bioinformatics and Cheminformatics

William L. Chen,*[,†] Burton A. Leland,[†] Joseph L. Durant,[†,‡] David L. Grier,[†] Bradley D. Christie,[†] James G. Nourse,[†] and Keith T. Taylor[†]

[†]Accelrys, Incorporated, 2440 Camino Ramon, Suite 300, San Ramon, California 94583, United States

Supporting Information

ABSTRACT: The wide application of next-generation sequencing has presented a new hurdle to bioinformatics for managing the fast-growing sequence data. The management of biomacromolecules at the chemistry level imposes an even greater challenge in cheminformatics because of the lack of a good chemical representation of biopolymers. Here we introduce the self-contained sequence representation (SCSR). SCSR combines the best features of bioinformatics and cheminformatics notations. SCSR is the first general, extensible, and comprehensive representation of biopolymers in a compressed format that retains chemistry detail. The SCSR-based high-performance exact structure and substructure searching methods (NEMA key and SSS) offer new ways to search biopolymers that complement bioinformatics approaches. The widely used chemical structure file format (molfile) has been enhanced to support SCSR. SCSR offers a solid framework for future development of new methods and systems for managing and handling sequences at the chemistry level. SCSR lays the foundation for the integration of bioinformatics and cheminformatics.



INTRODUCTION

From the reporting of the first entire human genome to the wide application of next-generation sequencing technology, in the past decade we have witnessed breakthroughs in sequence discovery and generation. Nowadays, obtaining sequence data is no longer the major obstacle[1] in many applications, such as drug discovery. Rather, managing the fast-growing volumes of sequence data presents a new hurdle in bioinformatics.

Biomolecules stored in bioinformatics databases are text strings. Chemical modifications and cross-links are usually handled as annotations (Figure 1) and retrieved using text search approaches. The bioinformatics sequence representation based on one-letter residue names has two inherent limitations. First, there are only 26 alphabetic letters and 20 of them have already been used to represent 20 natural amino acids. Moreover, there are in theory an unlimited number of unnatural amino acids that can be incorporated into a protein sequence. Second, bioinformatics sequence representation lacks detailed chemistry. Up until now, most sequences are naturally occurring biopolymers and are stored and managed mainly in bioinformatics databases. The current bioinformatics research still primarily focuses on the development of new tools for studying naturally occurring sequences. For example, recently Sboner et al.[2] developed a tool for finding gene fusions.

In this new decade, biopolymer-related research focus has shifted from generating the sequence data to interpreting the sequence data and understanding the functions of biomolecules and to modifying properties of biopolymers and even to creating new biomolecules with desired properties. For instance, Neumann et al.[3] introduced a new way to encode multiple unnatural amino acids. Scientists now use computers and laboratory chemicals to design and create organisms that do new things, such as building a protein to detect light and producing biologics for use as drugs. The recent creation of a bacterial cell that is controlled by a chemically synthesized genome[4] is a milestone toward redesigning the building blocks of life and challenging clear-cut distinctions of natural versus artificial life. It can be expected that in this decade more new breakthroughs will occur in this area. Scientists working in this area need deeper knowledge of chemistry and better chemistry tools. Bioinformatics tools and systems that rely on the representation and comparison of sequences at the residue name level are inadequate for the needs of biochemists. Unfortunately, there are currently few cheminformatics tools and database systems that are designed to deal with biopolymers. This is because the management and handling of sequence data at the chemistry level imposes an even greater challenge in cheminformatics.[5,6] The development bottleneck in this area is the lack of a good chemical representation of biopolymers. In cheminformatics, a molecular structure is described at the atom level and thus contains the full detail of a chemical structure. This representation allows chemists to perform sophisticated structure and substructure searches (SSS)[7] in chemical databases, to manipulate molecular structures, and to calculate physicochemical

Received: May 4, 2011
Published: July 31, 2011

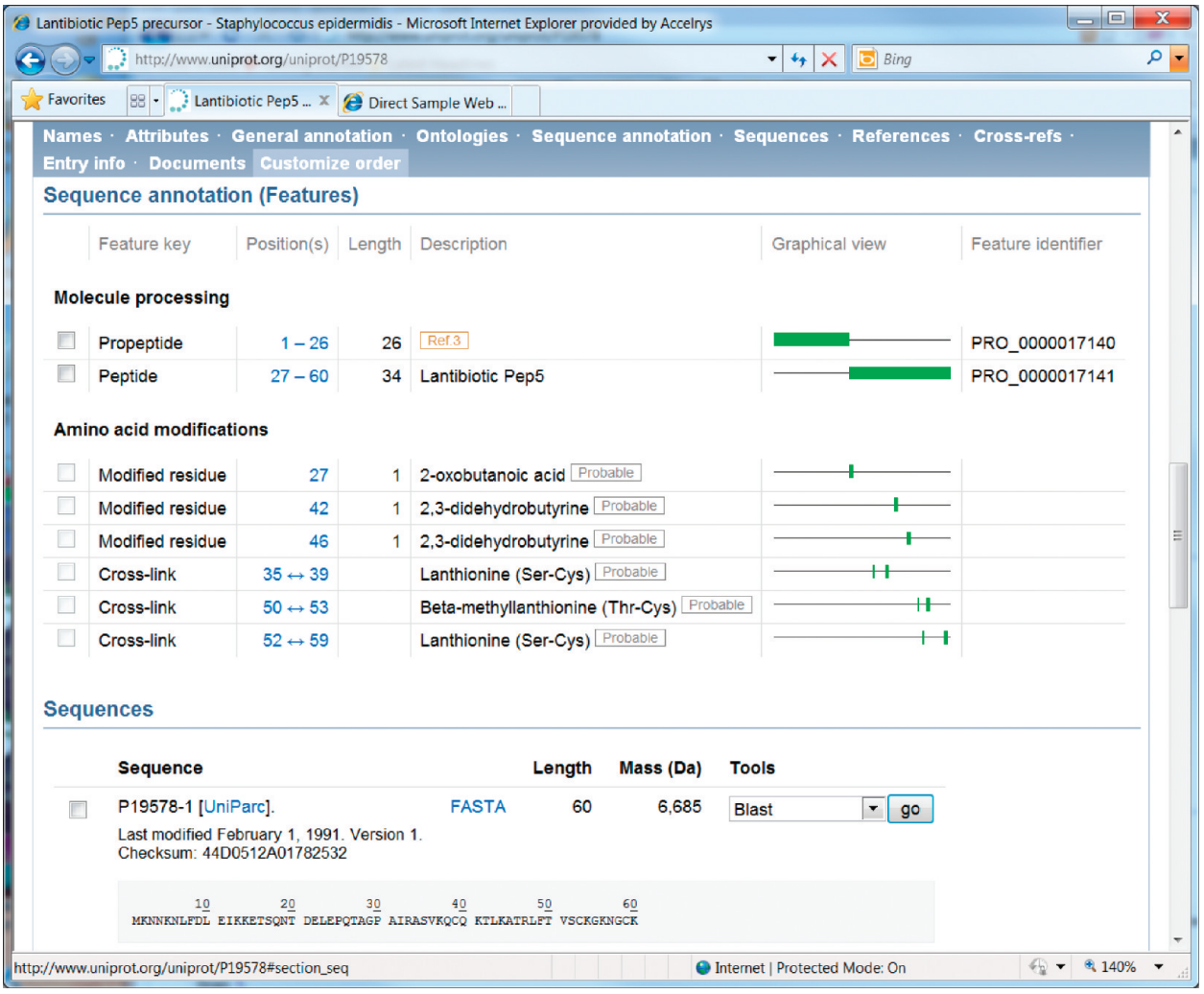


Figure 1. Sequence annotation and sequence of the lantibiotic Pep5 protein in UniProtKB.

properties. Although modern cheminformatics data management systems are efficient for managing molecular data with tens of millions of small chemical structures, they are not suitable for storing and handling large biomolecules.

To address the above performance challenge, we developed pseudoatom- and *atom (star atom)-based condensed representations in which each biopolymer residue is represented as a single atom:[8]

(a) Pseudoatoms alone: The Accelrys' extended periodic table (extended "Ptable") contains pseudoatoms, which are atom symbols that do not correspond to any of the chemical elements. The extended Ptable contains pseudoatoms for the 20 canonical natural amino acids, where the pseudoatom symbol corresponds to the amino acid's three-letter abbreviation. Pseudoatoms consist of a single "atom" that cannot be expanded. The major advantage of the pseudoatom representation is that it provides significant size reduction: About 7× for proteins, 10× for saccharides, and 20× for nucleotides. Thus, the pseudoatom representation converts a large biomolecule into a "small molecule", leading to more efficient storage and searching. Although the default extended Ptable can be customized to include more pseudoatom symbols, it is limited to a total of 200 entries, including the 103 natural elements, the 20 canonical amino acids, and a few specialized pseudoatoms, such as D for deuterium and T for tritium. To overcome limitations in the design of the pseudoatom approach, we introduced a more general representation based on a wildcard atom—the *atom (star atom).

(b) *Atoms alone: In the *atom representation of biopolymers, each residue consists of a single *atom with attached data that distinguishes chemically different residues. The *atom representation can also offer much better performance in the searching and the registration of large biopolymers than the full ctab representation. However, searching performance for structures that are represented as pseudoatoms is slightly better than for structures that are represented as *atoms. This is because in a search query, a *atom hits any atom or group of atoms at that position. To distinguish chemically different *atoms in searching, the data that are attached to the *atom (called attached data) must be checked. The major advantage of the *atom representation over the pseudoatom representation is in that there is no limit on the number of *atoms that can be used for describing residues of biomolecules.

dx.doi.org/10.1021/ci2001988 |J. Chem. Inf. Model. 2011, 51, 2186–2208

(c) Pseudoatom and *atom combination: This is a combined representation of the pseudoatom and *atom representations. It offers several advantages: (i) improved performance in searching and registration; (ii) no need to change the existing customized Ptable; and (iii) if your customized Ptable is full, then you can use *atoms to represent additional residues.

Problems of the Existing Condensed Representations. The above three existing condensed representations of biological polymers offer some advantages over the full ctab approach, requiring much less storage space and improving search performance. However, the limitation common to all three condensed representations is the loss of chemically significant data. This loss has consequences. For example:

(a) What is the residue definition?
  (i) Does the cysteine pseudoatom include the sulfur atom?
(b) What is the connectivity?
  (i) Should a cyclic peptide be read clockwise or counter clockwise?
  (ii) Where is the phosphate bound?

Consider the following example: Two different types of cysteine templates can be created: (a) put the sulfur atom inside the core structure of cysteine (Accelrys' convention) and (b) put the sulfur atom outside of the core structure of cysteine (Novo Nordisk convention),[9] as shown in Figure 2.

We use a sequence structure drawing tool (such as Accelrys Draw) to draw the sequence ACDC (Figure 3a) based on the template Figure 2b, and then add an S atom to each Cys and draw a bond between two sulfur atoms to create a disulfide bridge. The final sequence is shown in Figure 3b and saved in a molfile called sequence.mol. The fully expanded form of the sequence is shown in Figure 3c. Now, exit Accelrys Draw, replace the cysteine template Figure 2b with another cysteine template Figure 2a, which contains a sulfur atom inside the template, and then restart Accelrys Draw, load the sequence.mol file in Accelrys Draw. The contracted sequence looks exactly the same as before (Figure 3b). However, using the cysteine template, Figure 2a, to expand the sequence led to the strange structure shown in Figure 3d. In Figure 3d the disulfide bridge S−S has now become S−S−S−S.

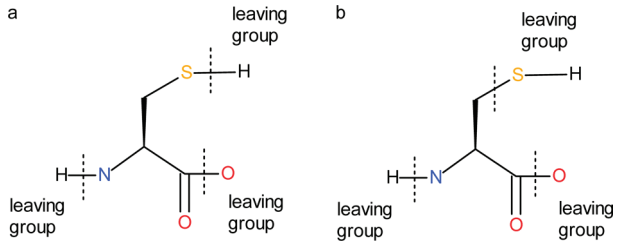


Figure 2. Two different types of cysteine templates. The cysteine template can take two different forms: (a) The cysteine template includes the sulfur atom inside the core structure of cysteine. (b) The cysteine template excludes the sulfur atom from the core structure of cysteine and treats −SH as a leaving group.

This example demonstrates the problem of the existing condensed sequence representations: The actual chemical content of the sequence is lost and becomes dependent on residue templates. In general, if one pseudoatom name is associated with multiple residue templates, then there will be ambiguity, and the easier it is for the user to modify residue templates, the greater the likelihood for confusion. Some examples are listed below:

(a) No way to check the consistency of residue templates during sequence registration. The formats of residue templates might be changed from time to time or might be different in different institutes (for example, the cysteine template examples shown in Figure 3). Unnaturally occurring residues (such as modified natural amino acids and artificial amino acids) have been playing increasingly important roles in the modern drug discovery. Different organizations can create different residue templates for those unnaturally occurring residues with the same pseudoatom names or give different names for the same artificial residue templates. However, there is no way to impose the business rules on sequence registration because no residue templates are centrally recorded in the database.
(b) The fully expanded sequence structure must be used as the exchange format between different applications and users. This is not suitable for large sequences because it can lead to a performance bottleneck.
(c) No support for substructure search. The substructure search (SSS) might not be important for searching databases of sequences that consist of only 20 natural amino acids. However, modified residues and unnatural residues have become common in modern drug discovery.

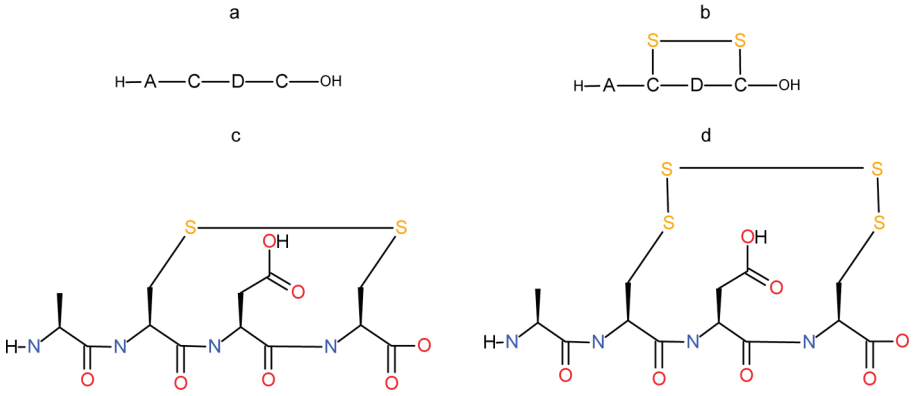


Figure 3. Example showing the problem of the pseudoatom-based sequence representation. (a) using Accelrys Draw to draw the sequence, ACDC. (b) Add a S atom to each Cys and draw a bond between two sulfur atoms to create a disulfide bridge. (c) Fully expand sequence (b) using template Figure 2b, and the disulfide bridge is shown correctly. (d) Fully expand sequence (b) using template Figure 2a and the disulfide bridge becomes −S−S−S−S−.

Thus, SSS on those structural fragments is important for medicinal chemists and protein biochemists. If all sequences stored in a database are in the fully contracted format, then SSS might return nothing unless a special "any atom" (A) query is used. The "A" query will match any atom, and thus the hits are not useful.

Recently, Jensen et al.[9] extended the Accelrys' pseudoatom representation to build the Novo Nordisk's protein database with some success. However, this approach has shortcomings. For example, pure cyclic peptides cannot be drawn using residue atoms alone,[9] and their protein structural identity method cannot handle disulfide bridges and may cause false positives.[9]

Major Requirements of a Good Chemical Representation of Biopolymers. A good chemical representation of large biological sequences must meet the following requirements:

(a) Small size to achieve high performance:
  (i) Encode large biopolymers into a compact format using the predefined monomers (such as amino acids and nucleotides) to significantly reduce the size of biomolecules for efficient storage and transfer.

(b) Wide coverage:
  (i) Can represent different types of biopolymers:
    - protein, DNA, RNA, etc. and their combinations
    - natural and synthetic biological polymers and their combinations
    - modified and unmodified biopolymers
  (ii) Can uniquely represent pure cyclic biopolymers without explicit chemistry.
  (iii) Can accommodate any modified, disulfide-bridged, cross-linked, and unnatural residues without using predefined monomers.

(c) Complete chemistry:
  (i) Should not lose any chemical structure information (including stereochemistry), so that the partial or full chemical structure of a sequence can be derived, if needed.

(d) Good integrity and integration:
  (i) Enable the development of methods for enforcing the integrity of predefined monomers and the uniqueness of biopolymers in a database to eliminate ambiguity.

(e) Good flexibility:
  (i) Impose no limitations on residue names.
  (ii) Offer flexibility of what, how, and when the chemistry detail can be used.
  (iii) Allow the storage of biopolymers and small molecules in a single repository.

(f) Good portability:
  (i) Provide a fully published and well-supported file format to allow data transfer and independent development of biopolymer database systems and analysis and annotation tools.
  (ii) Allow bidirectional conversion between bioinformatics sequence representation and cheminformatics sequence representation.

(g) Searching and matching:
  (i) Allow a biopolymer to be searched by a (sub)sequence or modification substructure.
  (ii) Allow two sequences to be compared for structural equivalence.

Here we present a new chemical representation of biopolymers called the self-contained sequence representation (SCSR).[10] SCSR can meet the above requirements for a good chemical representation of biopolymers. SCSR combines the best features of bioinformatics and cheminformatics notations and is the first general, comprehensive representation of biomolecular sequences in a compressed format without losing chemistry detail. It significantly reduces the size of biomolecules for efficient storage and information transfer. Unlike the bioinformatics sequence representation and our existing methods that lose detailed structural information, SCSR retains full chemistry, so that the partial or full chemical structure of a sequence can be easily derived, if needed. SCSR is general and can represent different types of biopolymers. It can accommodate modified residues, disulfide bridges, and cross-links without using predefined monomers. SCSR can represent pure cyclic biopolymers without explicit chemistry. The SCSR-based sequence NEMA[11] key provides a fast and reliable tool for exact sequence search and sequence duplicate checking. The SCSR-based SSS finds chemically identical regions of sequences and complements existing sequence searching tools, such as BLAST.[12]

We have enhanced the V3000 molfile format[13] to support SCSR. This allows the independent development of analysis and annotation tools and biopolymer data management systems.

METHODS

Implementation Details. SCSR has become an important new foundation of a suite of Accelrys' products for biopolymer representation and handling. SCSR was first implemented in Symyx Direct 7.0 (now Accelrys Direct) chemistry data cartridge.[14] Later, SCSR was also implemented in other Accelrys' products, such as Accelrys Draw 4.0.[15] Accelrys Draw is free for personal and academic use and is available at http://accelrys.com/.[16] The Accelrys CTfile formats[13] have also been extended to support SCSR. The latest version of Accelrys CTfile formats is available at http://accelrys.com/.[17]

SCSR. SCSR[10] consists of a contracted sequence in which each residue is encoded as a single atom (called the template atom) plus a set of unique residue templates associated with the template atoms (Figure 4a).

The templates that are embedded into the sequence representation are called template-based groups, or Tgroups, for short. For example, if a sequence consists of 22 residues: 10 alanines (for simplicity, the amino acids mentioned in this paper are the L-form amino acids unless otherwise explicitly stated), 5 cysteines, and 7 phenylalanines, then there are only three unique Tgroups attached to the sequence: 1 alanine, 1 cysteine, and 1 phenylalanine. The term "self-contained" means that a sequence in the SCSR format contains exactly the same chemical structure information as that in the full chemical structure format (full-ctab). The link between a template atom and a Tgroup is through the template identifier, which consists of the template atom name and the template class. The template class (such as AA, DNA, and RNA) allows different types of residues to coexist in the same structure. The SCSR also contains information about the directions of bonds between residues.

The SCSR allows any structural fragment, such as modified residues, to be kept as explicit chemistry. In one extreme, a SCSR consists of only the template atoms and their unique Tgroups, with no uncompressible substructures of any kind except the leaving groups. In the other extreme, a SCSR consists of only uncompressible substructures, with no template atoms and no

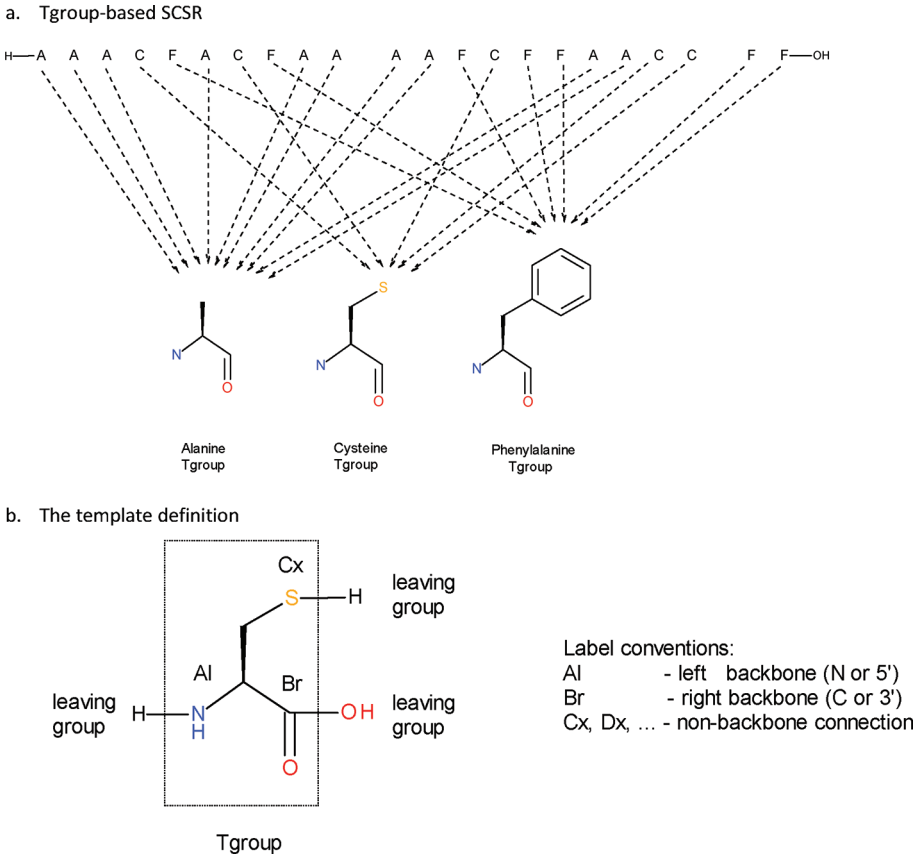


Figure 4. (a) Tgroup-based SCSR. (b) Template definition: a template consists of its core structure (Tgroup) and all the leaving groups that are attached to the attachment point atoms of the Tgroup.

Tgroups at all. Therefore, the full-ctab format of a molecular structure can be regarded as a special case of SCSR. From here, we can reach a conclusion: Biopolymers in the SCSR format and small molecules in the full-ctab format can be stored together in one database. SCSR does not impose limitations on the types of biopolymers and the length of residue names and residue classes.

The above sequence representation is called the Tgroup-based SCSR. It is the fundamental form of SCSR.

Template. A template defines the chemical structure of a functional group or a residue. To support the SCSR, an enhanced template format was developed that provides two or more attachment atoms (Figure 4b). These attachment atoms have special attributes (such as priority, direction, and leaving group) that define the ways that enhanced templates (called templates for short) attach to each other. The template contains no attachment bonds, and it carries the "template class", such as AA, DNA, and RNA. To allow the implementation of the business rules for the creation and the searching of biopolymers, a set of residue templates to be used by all persons within an organization can be created and stored in a global template file. Global templates are a way of allowing a registrar to define a standard set of template definitions. In a sense, this is an implementation detail. However, "global" templates offer several advantages, such as: (a) allowing multiple implementations of SCSR to share a collective and vetted definition of "standard" template definitions; (b) can provide a way for a registrar to enforce structures allowed into a repository by requiring only "standard" templates to be present; (c) can allow further "compression" of information exchange by omitting "standard" template definitions for client/server (which

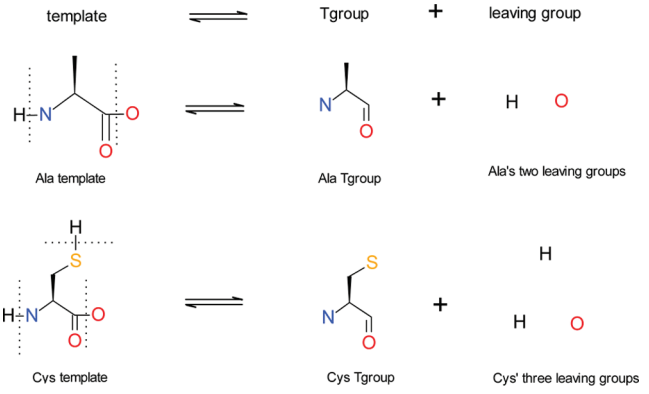


Figure 5. Relationship between the template and its Tgroup.

can synchronize global templates) structure exchanges; and (d) can decrease packed structure sizes by omitting the packing of "standard" templates definitions. Our global template definition file is simply a "no-structure" root with the template definitions per the TEMPLATE block. A sample global template definition file (sample_global_templates.mol) that contains two residues (alanine and cysteine) templates is available in the Supporting Information.

Tgroup. A Tgroup is the core structure of a template. It also contains the attachment point information of the corresponding template. A Tgroup can have one and only one member. The relationship between the Tgroup and the template is shown in Figure 4b (see also Figure 5). Converting a template

dx.doi.org/10.1021/ci2001988 |J. Chem. Inf. Model. 2011, 51, 2186–2208

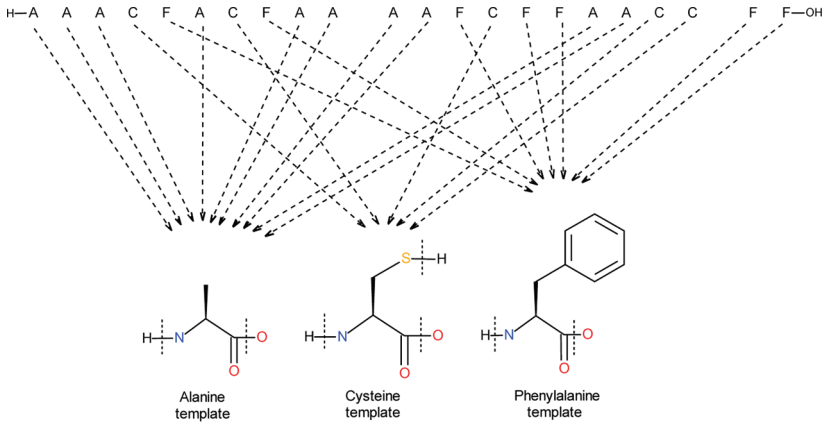


Figure 6. Template-based SCSR.

to a Tgroup requires the removal of all the leaving groups from the template. Each attachment point atom in the Tgroup bears free valences formed after the bonds to the leaving groups are cut off.

Alternate Form of SCSR. As the first example of the implementation of the original SCSR described above, here we will consider a slight modification to this Tgroup-based SCSR definition. Based on the relationship between the Tgroup and the template (Figure 5), if all the Tgroups in the Tgroup-based SCSR in Figure 4a are replaced with the corresponding templates, we obtain the alternate form of SCSR—the template-based SCSR (Figure 6).The relationship between the Tgroup-based SCSR and the template-based SCSR is given in Figure 7. These two forms of SCSR are identical except that the template-based one carries additional information on leaving groups for the attachment point atoms of the Tgroups. Therefore, the task of conversion between the Tgroup-based SCSR and the template-based SCSR is the conversion between the Tgroups and the corresponding templates.

Classification Hierarchy of Sequence Representations. As discussed above, to tackle the challenges of the chemical representation of biomolecular sequences, we have developed a series of sequence representations. The classification hierarchy of sequence representations, including the bioinformatics sequence representation, is given in Figure 8. All the sequence representations can be grouped into three categories. At the left is the standard bioinformatics representation of biopolymer sequences. It is text string based and thus is the most condensed format. At the right is the traditional cheminformatics representation of biomolecules. It contains an atom-by-atom and bond-by-bond detailed description of the chemical structure and thus is the least compressed format. In the middle is the biocheminformatics representation of sequences. Like the bioinformatics representation, standard residues are collapsed into residue atom names. Like the cheminformatics representation, atoms, both residue atoms and real chemical element atoms, are connected to each other through chemical bonds. In this category, although the pseudoatom and *atom representation are more compressed than SCSR, they lack chemical structure information of residues. The template-based SCSR is the least compressed representation in this category. It contains additional leaving groups that may not exist in the full-ctab format. The Tgroup-based SCSR is the only condensed format of sequence representations that contains chemical structure information identical with that of the full-ctab format.

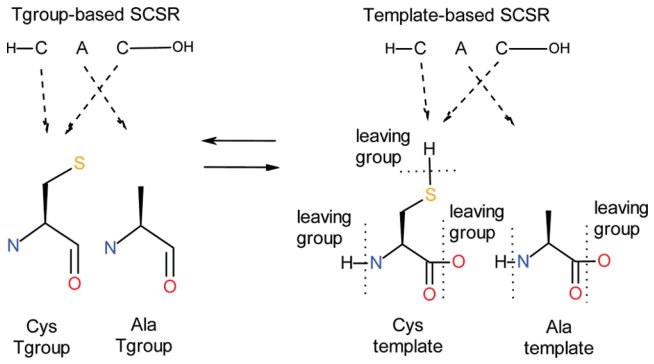


Figure 7. Relationship between the Tgroup-based SCSR and the template-based SCSR.

From Figure 8, it can be seen that SCSR has several advantages over the bioinformatics sequence representation as well as the pseudoatom and *atom condensed sequence representations. The single most important advantage of SCSR is that there is no loss of chemically significant information. This makes the following tasks possible:

(a) Expansion and contraction by both database server and client without using external templates.
(b) More efficient data exchanges between different applications and different parties.
(c) Support sequence and subsequence searches.
(d) Support exact structure and substructure searches (SSS).
(e) Support sequence verification during sequence registration.

The bottom part of Figure 8 shows the overall trend in two different directions. From right to left, the compactness of sequence increases, but its chemical detail decreases. From left to right, the compactness of sequence decreases, but its chemical detail increases. The only exception is that although SCSR is much more concise than the full-ctab format for large biopolymers, it contains the same chemical structure information as the latter.

Contraction and Expansion of Sequences. Contraction and expansion of sequences are two major sequence manipulation processes. They can be performed on only some of the residues of a biopolymer. For example, if a scientist wants to look at the structure of a single residue, then only that particular residue template atom needs to be expanded. After the scientist finishes the investigation, the residue structure can be contracted back to

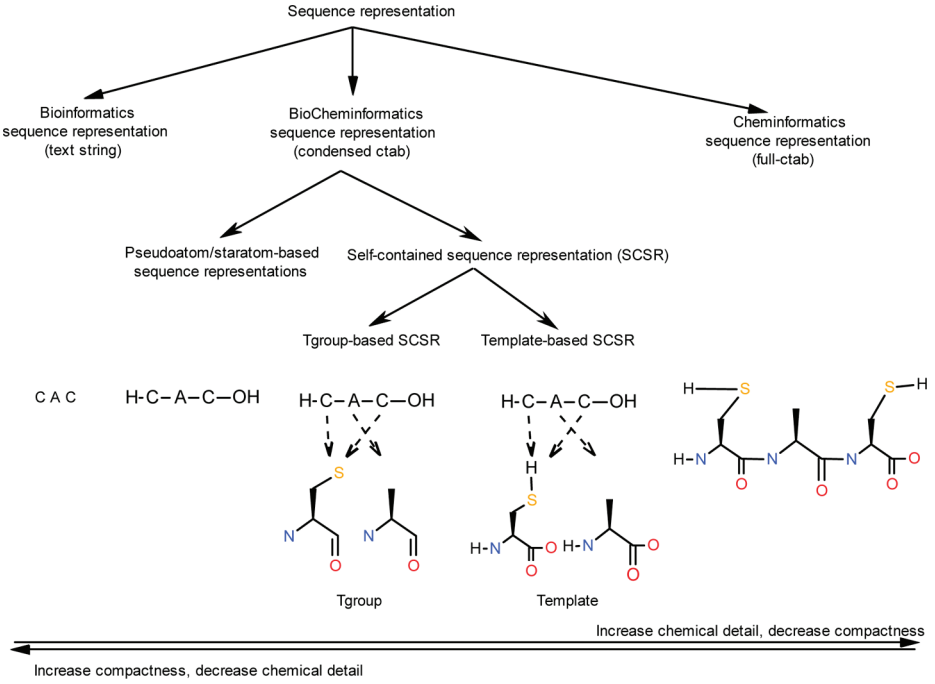


Figure 8. Classification hierarchy of sequence representations.

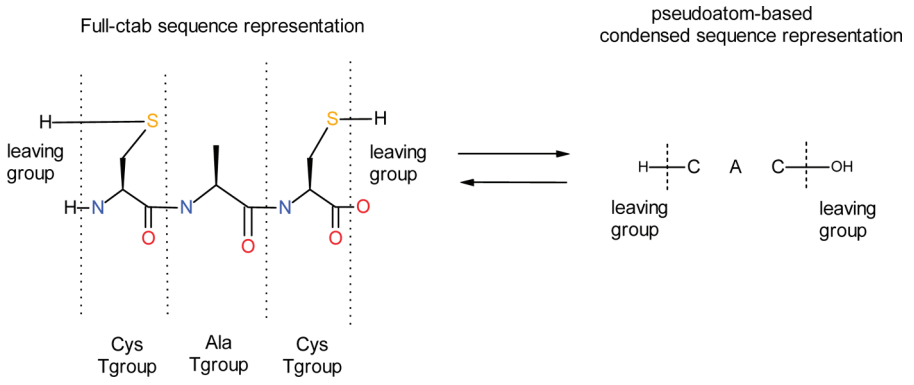


Figure 9. Conversion between the full-ctab sequence format and the pseudoatom-based condensed representation. Note that because no Tgroups are stored in the pseudoatom-based sequence representation, the chemical structure information of all residues is lost.

the template atom. Alternatively, the contraction and expansion can also be done for the whole sequence. This is equivalent to the conversion between a full-ctab sequence and the condensed format.

First, we look at the conversion between the full-ctab sequence format and the pseudoatom-based condensed representation (Note: the same procedure can be applied to *atom-based sequence representation), as shown in Figure 9. The conversion begins with replaceing all Tgroups (that is, the core structures of the residues) in the sequence with the corresponding residue pseudoatom names and throwing away all the Tgroups. This process leads to the loss of the detailed chemical structure information of the residues. The reverse process—converting the pseudoatom-based sequence into the full-ctab format—seems easy, which is to simply replace all pseudoatom names with the corresponding Tgroups. Practically, there is a problem here. The pseudoatom representation does not contain any Tgroup, so where do those Tgroups come from? The solution is to use external templates. From here we can better understand why the exact meaning of the existing condensed sequence representations, pseudoatom- and *atom-based representations, are template dependent.

Next, consider the conversion between the full-ctab sequence and the Tgroup-based SCSR. Converting a full-ctab sequence into the Tgroup-based SCSR involves the following steps: (a) replace all the Tgroups in the full-ctab sequence with the corresponding residue template atom names; (b) collect the unique Tgroups in the Tgroup list; and (c) link each template atom name with its corresponding Tgroup (Figure 10). Expanding the SCSR to the full-ctab sequence is also straightforward, replacing each template atom name with the linked Tgroup.

There are two advantages of the above sequence conversion: (a) The conversions in both directions can be carried out without help of any external templates; and (b) the chemical structure contents of the two formats are identical. There is neither data loss nor data gain during the conversions.

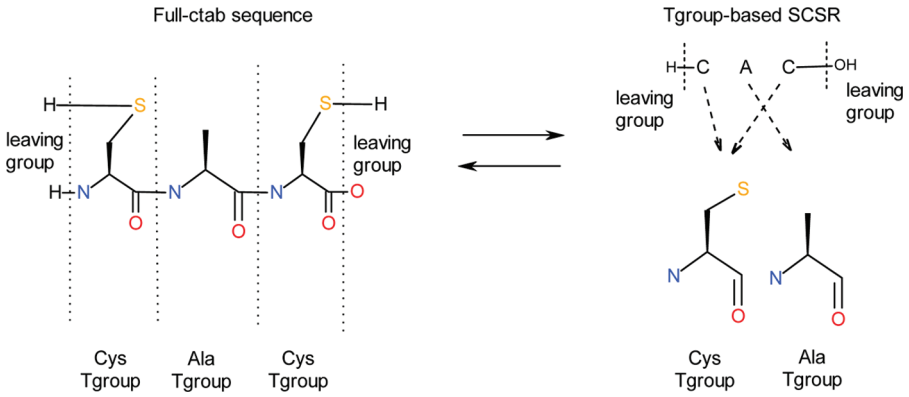


Figure 10. Conversion between the full-ctab sequence format and the Tgroup-based SCSR.

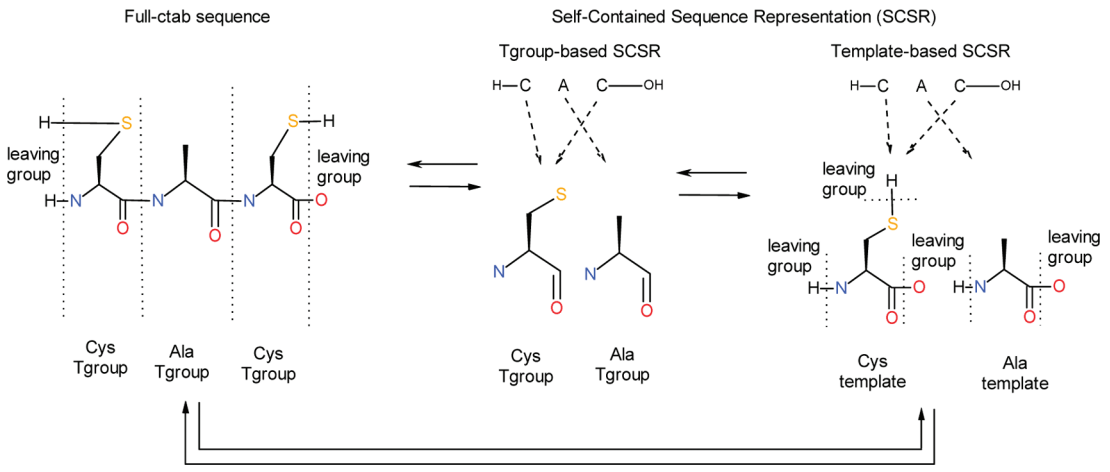


Figure 11. Conversion between the full-ctab sequence and the template-based SCSR.

Because of these features, the Tgroup-based SCSR is particularly suitable for internal manipulation of sequences and applications, such as the generation of the sequence NEMA key (see the NEMA Section).

Finally, consider the conversion between the full-ctab sequence and the template-based SCSR. This conversion is a two-step process (Figure 11): (a) the conversion between the full-ctab sequence and the Tgroup-based SCSR; and (b) the conversion between the Tgroup-based SCSR and the template-based SCSR.

The first step has been discussed above, and the second step has been partially discussed in the Alternate Form of SCSR Section. That is, the task of conversion between the Tgroup-based SCSR and the template-based SCSR is the conversion between the Tgroups and the corresponding templates. To convert each Tgroup in the full-ctab sequence into the corresponding template, we need to add the leaving groups to the attachment point atoms of each Tgroup. However, the full-ctab sequence may not contain all the necessary leaving groups for each Tgroup. Which leaving group must be used to connect to each attachment point atom of a Tgroup? Where does this leaving group come from?

One solution is to attach a set of unique templates that were used to create the full-ctab sequence to the sequence itself. For example, if a template-based SCSR is expanded to the full-ctab format, then the existing templates must not be deleted. This will lead to the shortcoming of carrying duplicate data in the full-ctab format. Another solution is to use a list of external templates. If the Tgroup of an external template is identical to a Tgroup in the full-ctab sequence, then the leaving groups of that template are used to convert the corresponding Tgroup into a template. The comparison of the Tgroups of the full-ctab sequence and the external template can be conveniently done using the Tgroup NEMA key (see the NEMA section). Because a set of global templates is required to enforce the business rules, those global templates can be used as external templates for converting a full-ctab sequence into the template-based SCSR.

The above processes have three characteristics:

(a) Conversion between the full-ctab sequence format and the template-based SCSR is a two-step process.

(b) The contraction of a full-ctab sequence into the template-based SCSR must rely on the external templates to determine the leaving groups of the attachment point atoms of each Tgroup of the full-ctab sequence.

(c) The chemistry contents of the two formats might not be exactly identical because the template-based SCSR contains additional leaving groups that might not exist in the full-ctab sequence format.

Because the template-based SCSR contains the templates used to create the sequences, this representation is used to extend the V3000 molfile format to support SCSR.

As mentioned previously, SCSR is general and can represent different types of biopolymers. Figure 12 shows an example of the contraction and expansion of a simple DNA sequence using Accelrys Draw.

dx.doi.org/10.1021/ci2001988 |J. Chem. Inf. Model. 2011, 51, 2186–2208

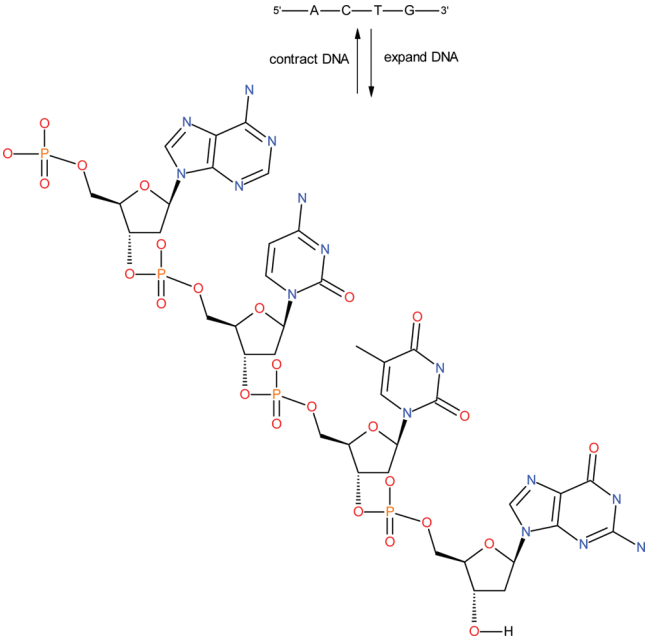


Figure 12. Contraction and expansion of a simple DNA sequence using Accelrys Draw.

Representation of Modified Residues in SCSR. Chemical modifications of residues play an important role in controlling the bioactivities and the functions of biopolymers. To allow SSS of the modified, cross-linked, and unnatural residues, those residues are kept in explicit chemistry.

It should be noticed that such explicit chemistry is encoded in such a way that this explicit chemistry can be replaced with the residue name for rendering purpose only, so that the whole sequence can be shown in the standard bioinformatics format—the sequence view. This encoding method is called abbreviation. An abbreviated structure (also called an abbreviation or abbreviation Sgroup) is a type of chemical Sgroup that displays a text label to represent all or part of a molecular structure. The abbreviated structure can be expanded to display the underlying structure (see Figure 13a, for example). Abbreviated structures are equivalent to the same structures without the abbreviations. The abbreviation approach also allows the cross-linked residues to be displayed in the structure view, where all residues are shown using their residue names, but the bonds between residues, including cross-links, are explicitly shown (Figure 13b).

Bond Directions and Cyclic Biopolymers. Unlike a small molecular representation that is based on an undirected graphical model, SCSR is based on a partially directed graphical model. SCSR contains bond direction information on residue connections. This information is offered via the priorities and the directions of the Tgroup and templates (Figure 4b) and the template atoms in SCSR (Figure 14). And thus cyclic peptides can be represented unambiguously without needing to expand any template atom. It also removes ambiguity in cross-linked structures. For example, the two cyclic peptides, Figure 15a and b, look identical, but they are not exactly the same, as shown in Figure 15a′ and b′, respectively, where the bond directions are explicitly marked. These two cyclic sequences can be distinguished from each other using the sequence NEMA key technology (Figure 15).

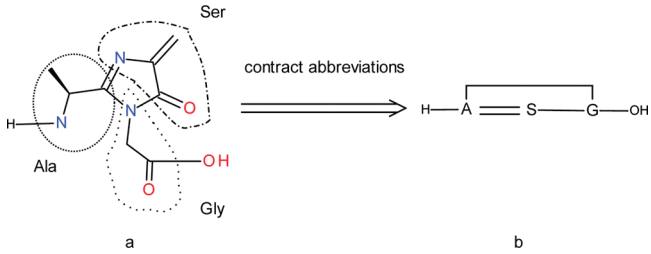


Figure 13. Modifications across the residue boundaries. (a) The residues alanine, serine, and glycine are cross-linked and modified to form a five-membered ring. (b) Contracting the abbreviated structures of alanine, serine, and glycine in the structure view of Accelrys Draw shows that there is a double bond between alanine and serine residues and a cross-link between alanine and glycine residues. This three-residue formed substructure is common to a set of lyases (see Table 1).

NEMA

What is NEMA? NEMA[11] stands for the newly enhanced Morgan algorithm. It includes a set of technologies for accurate perception of stereochemistry, canonicalization of molecular structures, conversion of the graph-based chemical structure into the canonical linear notation (the full NEMA name), and generation of NEMA keys from the full NEMA names for small and large molecules alike. NEMA offers a new method of stereochemical recognition that meets the industry's need for improved chemical representation of tetrahedral and geometric stereogenic centers. The NEMA method extends stereochemistry recognition to axial chirality, for example, allenes and atropisomers, such as hindered biaryls. It supports both two- and three-dimensional (2D and 3D) stereochemistry perception. The NEMA algorithm for stereochemistry perception has recently been being extended to deal with more complex stereochemical geometries, such as trigonal bipyramidals and square pyramidals. The stereo perception method is the heart of the NEMA technologies, especially for dealing with biomolecules that have many stereogenic centers.

With the dramatic increase of data, large databases that contain tens of millions of molecular structures are now widely available, for example, the PubChem database. Conversion of graph-based molecular structures into canonical NEMA names turns the CPU-intensive graph-matching based structure search problem into a simple string comparison task, leading to a significant performance gain. Another advantage of the full NEMA name is that it does not lose any chemical structural information. However, the NEMA name has two short comings. One is that different molecular structures may have different lengths of the full NEMA names. The other is that the full NEMA names of large molecules can be very long. Therefore, comparison of the full NEMA names is still not very efficient and convenient.

NEMA Key. The above problems can be tackled by using the SHA-256 secure hash algorithm[18] to compress the full NEMA name into a short, fixed length (30 characters) string called the NEMA key.[19] To reduce the possibility of the hash collision, we replaced the 10-number set (0 1 2 3 4 5 6 7 8 9)[19] with the 32-character set (1 2 3 4 5 6 7 8 9 A B C D E F G H J K M N P Q R S T U V W X Y Z) for encoding the NEMA key. In this new NEMA key representation, each character can represent 5 bits. The 30-character long NEMA key represents 150 bits, which is 60 bits more than the original NEMA key representation based on the 10-number set.

NEMA key-based duplicate checking of a large SCSR-based sequence database is more than an order of magnitude faster than

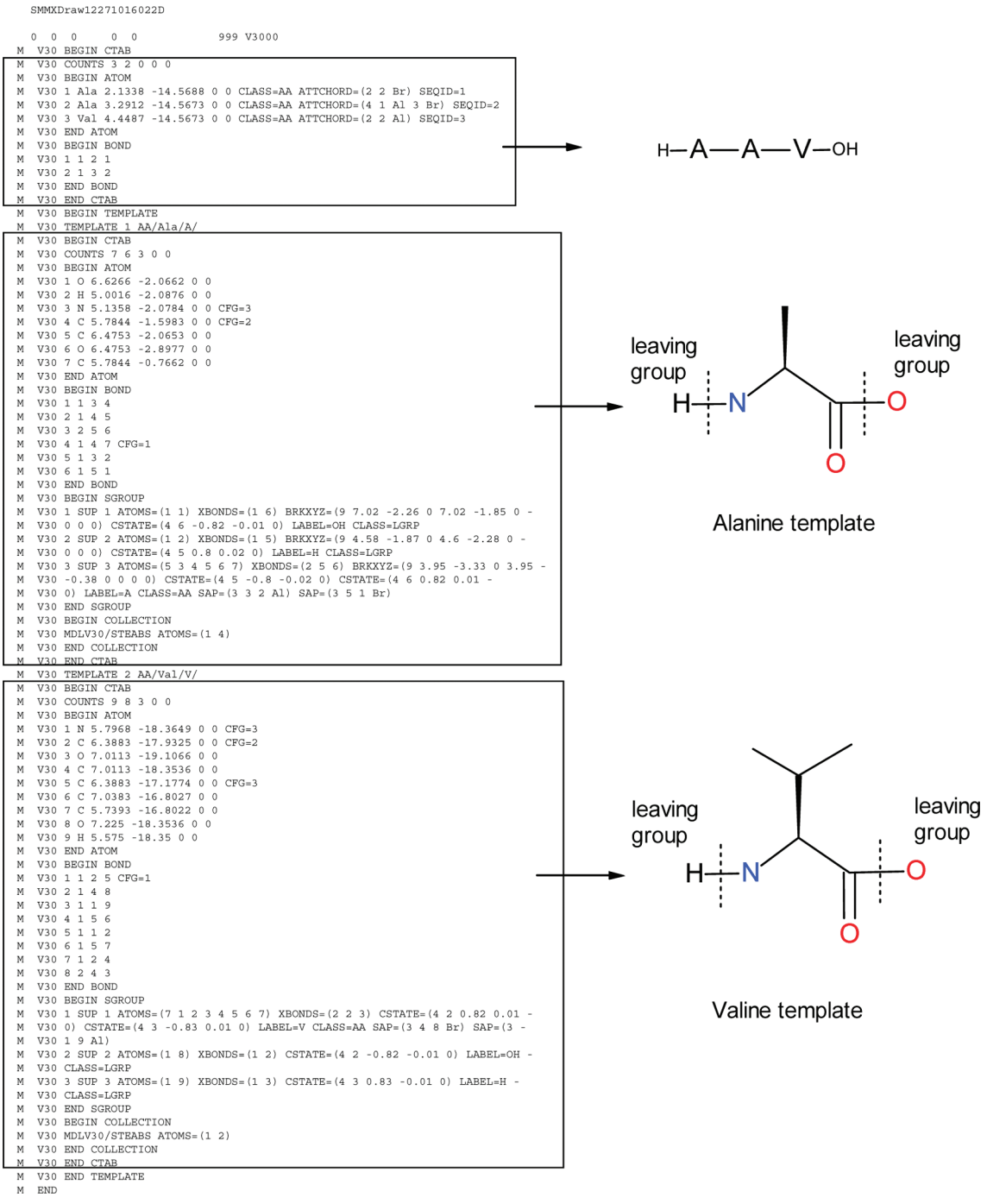


Figure 14. Example of the SCSR-enhanced V3000 molfile format. Note that Al and Br associated with each template atom determine the backbone direction of a sequence.

the graph-matching based approach. The only problem of the NEMA key method is that because of the nature of the hash algorithm, there may exist the possibility of hash collisions. For example, the InChIKey is a fixed length (25 characters) condensed representation of the IUPAC International Chemical Identifier (InChI),[20] and has gained popularity in the cheminformatics community in recent years. In 2010 Goodman[21] at University of Cambridge discovered InChIKey collisions using the stereoisomers of spongistatin I, which contains only 24 tetrahedral stereogenic centers and 2 double bonds. Although this problem has been discovered and widely discussed, no fix was included in the latest InChI v. 1.03 (2010) software release,[22] indicating the complexity of the problem. Our tests confirmed Goodman's discovery. A total of 16 565 InChIkey collisions were found from all 67 108 864 ($= 2^{26}$) stereoisomers of spongistatin I. The upper bound on collision resistance of the InChIKey for stereoisomers is 2^{16}, which equals 6.6×10^4.

In contrast, NEMA generated unique NEMA keys for all 67 108 864 stereoisomers of spongistatin I without any collision. This is because the NEMA key has much higher collision resistance than the InChIKey. As mentioned previously, each stereo NEMA key consists of 30 characters, and each character

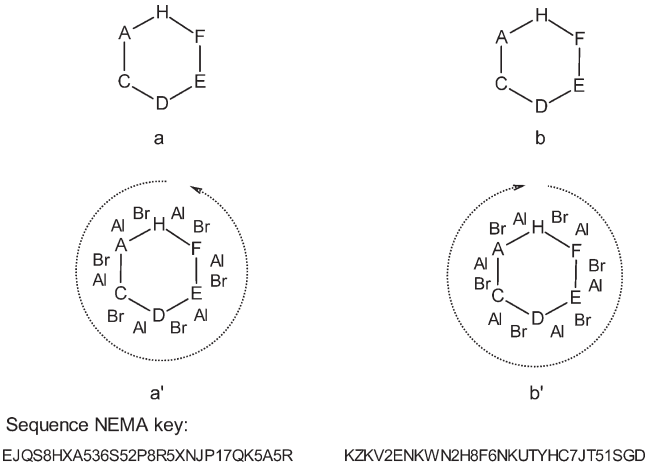


Figure 15. Bond directions in cyclic sequences. The drawings a′ and b′ of this figure show that SCSR contains the backbone direction information. The direction goes from Br to Al (see Figure 4b for the meaning of Al and Br). This information can be used to distinguish two stereoisomers. One application of this direction information is the sequence NEMA key. The sequence NEMA keys of these two cyclic peptides are different, indicating that they represent two different stereoisomers.

represents 5 bits. And thus, each NEMA key internally represents 150 bits. The upper bound on collision resistance of the NEMA key is 2^{75}, which equals 1.8×10^{21}. The total number of sequences in all databases is certainly much, much smaller than 1.8×10^{21} [for example, all databases of the UniProt release 2011_02 contains only 68 782 492 (6.9×10^7)] entries,[23] the hash collision of the sequence NEMA key appears unlikely at present.

Below we discuss some new types of NEMA keys that are specifically designed to handle biopolymers.

Tgroup NEMA Key. The Tgroup NEMA key is generated for a Tgroup using the same algorithm that is used for generating the stereo molecular NEMA key, except that the attachment point information of the Tgroup is also taken into account. Similarly, a leaving group (Lgroup) NEMA key can also be generated for each Lgroup of a template or a sequence. The combination of the Tgroup NEMA key, Lgroup NEMA keys, and the attachment point information of a template constitutes the template key. These NEMA keys allow fast comparison of structures of Tgroups and templates during sequence contraction and expansion and validation of templates and sequences during sequence registration and searching. The Tgroup NEMA key also makes it possible to efficiently generate sequence NEMA keys for large sequences.

Sequence NEMA Key. Theoretically, the existing NEMA key algorithm could also be used to generate stereo NEMA keys for biopolymers. This could be achieved by first expanding all the compressed residues into the full chemistry representation and then generating the NEMA key. However, two factors make this approach impractical. First, the full chemistry representation of a large biopolymer can be huge and thus generation of the NEMA key for such a molecule would be very slow. Second, biopolymers are large molecules with many stereogenic centers. Therefore, the number of possible stereoisomers of a large biopolymer can be enormous, and thus, the possibility of hash collisions would be unavoidable. The consequence of a hash collision would be that the key would not be unique to a single biopolymer.

To tackle these challenges, we have developed a new type of NEMA key for biopolymers called the sequence NEMA key. The sequence NEMA key is generated from the biopolymer structure itself by taking into account all chemistry detail, including stereochemistry, and thus is independent of any template atom symbol.

The algorithm of generating the sequence NEMA key is as follows:

(a) Convert a fully or partially expanded sequence into a maximally compressed Tgroup-based SCSR.
(b) Generate the Tgroup NEMA key for each unique Tgroup of the sequence.
(c) Replace each template atom symbol with the corresponding Tgroup NEMA key.
(d) Generate the sequence NEMA key from the Tgroup NEMA key contained sequence structure.

If a sequence consists of two or more disconnected subsequences, then the sequence NEMA keys for each subsequence are first generated, and then sorted NEMA keys are compressed using the same hash algorithm to produce the final overall sequence NEMA key.

Since a Tgroup contains many atoms and some Tgroups contain several stereogenic centers, the above algorithm can efficiently generate NEMA keys for large biopolymers and also significantly reduces the possibility of hash collisions. The sequence NEMA key also takes into account the information on the connection directions of the residues in SCSR and thus can distinguish two cyclic sequences that differ only in residues connection directions (Figure 15). The sequence NEMA key offers an efficient method for the exact sequence search and duplicate checking in sequence databases.

Substructure Search of Biopolymers. In principle, substructure searching (SSS) on any part of a biomolecular sequence encoded in SCSR is possible. This can be achieved by first expanding all the template atoms of both the target sequences and the query subsequence (if not a query substructure). However, for large sequences, SSS is slow. Considering that biochemists are mainly interested in the chemical structures of modified, cross-linked, and unnatural residues in a sequence, we can limit our SSS operations to only noncompressed structural portions of biopolymer sequences. In our implementation, to facilitate the substructure search on modified structural fragments, all modified residues are indexed using the FastSearch indexing approach.[24] More specifically, FastSearch indexes real chemical atoms plus about 10 global template atoms in any direction from the real atoms. This allows one to create a query which has a template atom connected to a real atom and takes advantage of the "context" provided by the template atom.

After the FastSearch index file is built for the biopolymer database, SSS on any nonstandard residues can be performed the same way we search small molecule databases. SSS finds chemically identical regions, regardless of what they are named. FastSearch indexing of areas with explicit chemistry makes searching perform and scale well. It complements existing sequence searching tools, such as BLAST.

Extension of V3000 Molfile Format to Support SCSR. The Accelrys chemical table file (CTfile) formats[13] are the de facto standard for representation and communication of chemical structure information in cheminformatics. The V3000 molfile format has been extended to support SCSR by introducing the TEMPLATE block.

TEMPLATE Block. A Template block defines one or more template definitions:

```
M V30 BEGIN TEMPLATE
[template-definition]*
M V30 END TEMPLATE
```

where * means 1 or many template definitions.

A template definition begins with a line that defines template properties, followed immediately with a single ctab block to provide the ctab definition:

```
M V30 TEMPLATE index -
M V30 [ name | class/name [/alternate_name1[...] ] -
M V30 [ COMMENT=template_comment ]
M V30 BEGIN CTAB
...
M V30 END CTAB
```

Template definitions can be present only at the root ctab of a CTfile, not within Rgroup blocks or within any other multictab blocks.

ATOM Block. The Atom block has been enhanced to support collapsed template definitions. An atom block specifies all node information for the connection table. It must precede the bond block. It has the following format:

```
M V30 BEGIN ATOM
M V30 index type x y z aamap –
M V30 [CHG=val] [RAD=val] [CFG=val] [MASS=val] –
M V30 [VAL=val] –
MV30 [HCOUNT=val] [STBOX=val] [INVRET=val] [EXACHG=val] –
M V30 [SUBST=val] [UNSAT=val] [RBCNT=val] –
M V30 [ATTCHPT=val] –
M V30 [RGROUPS=(nvals val [val ...])] –
M V30 [ATTCHORD=(nvals nbr1 val1 [nbr2 val2 ...])] –
M V30 [CLASS=template_class] –
M V30 [SEQID=sequence_id] –
...
M V30 END ATOM
```

The "..." indicates other atom/sgroup line(s) might be present before the "END" of that object block.

(a) ATTACHORD
This property has been augmented to allow textual attachment ids. Only Rgroup atoms and collapsed template atoms can provide ATTACHORD information. Rgroup atoms support only integer ATTACHORD values, while collapsed template atoms support text values (such as Al and Br).

(b) SEQID
Currently this property supports a positive integer value to capture residue sequence id information.

(c) CLASS
This property provides the class information for a collapsed template atom. Thus for a collapsed alanine template atom (AA/Ala), its CLASS = AA and its name = Ala.

SGROUP Block. The Sgroup block has new properties to support collapsed template definitions. An Sgroup block defines all Sgroups in the molecule, including superatoms. The format is as follows:

```
M V30 BEGIN SGROUP
[M V30 DEFAULT [CLASS=class] –]
M V30 index type extindex –
M V30 [ATOMS=(natoms atom [atom ...])] –
M V30 [XBONDS=(nxbonds xbond [xbond ...])] –
M V30 [CBONDS=(ncbonds cbond [cbond ...])] –
M V30 [PATOMS=(npatoms patom [patom ...])] –
M V30 [SUBTYPE=subtype] [MULT=mult] –
MV30 [CONNECT=connect] [PARENT=parent] [COMPNO=compno] –
M V30 [XBHEAD=(nxbonds xbond [xbond ...])] –
M V30 [XBCORR=(nxbpairs xb1 xb2 [xb1 xb2 ...])] –
M V30 [LABEL=label] –
M V30 [BRKXYZ=(9 bx1 by1 bz1 bx2 by2 bz2 bx3 by3 bz3])* –
MV30 [ESTATE=estate] [CSTATE=(4 xbond cbvx cbvy cbvz)]* –
M V30 [FIELDNAME=fieldname] [FIELDINFO=fieldinfo] –
M V30 [FIELDDISP=fielddisp] –
M V30 [QUERYTYPE=querytype] [QUERYOP=queryop] –
M V30 [FIELDDATA=fielddata] ... –
M V30 [CLASS=class] –
M V30 [SAP=(3 aidx lvidx id)]* –
M V30 [BRKTYP=bracketType] –
M V30 [SEQID=sequence_id] –
...
M V30 END SGROUP
```

(a) SEQID
Currently this property supports a positive integer value to capture residue sequence id information for an expanded template definition. Only abbreviation Sgroups (type = SUP) support SEQID property information. An example of the SCSR-enhanced V3000 molfile is shown in Figure 14, and the corresponding molfile can be downloaded in the Supporting Information section.

RESULTS

SCSR Significantly Reduces the Size of Large Biopolymers. The new representation can store large biopolymers in a compact manner. For example, a sequence that consists of 3000 natural amino acids contains about 22 050 heavy atoms in the explicit chemistry format. The average number of heavy atoms for the 20 natural amino acids is 7.35, which does not include the C-terminal leaving group. Thus, the same sequence encoded in the SCSR format contains only 3000 template atoms plus 147 heavy atoms of the 20 unique amino acid Tgroups. That is only 14% (= 3147/22 050) of the size of the full structure format. It should be noticed that there are only very limited number of unique natural residues (for example, 20 unique natural amino acids). Therefore, the larger a sequence is, the higher the compression.

Even for sequences that contain modified residues, which are kept in explicit chemistry, SCSR can still provide good compression. Consider the longest sequence (Titin) in UniProtKB. This sequence consists of 35 213 amino acids with 21 modified residues, 45 disulfide bonds, and 5 cross-links. The SCSR format of this sequence contains 35 929 atoms. Among those atoms, 35 097 atoms are template atoms, the remaining 832 atoms are chemical element atoms, N, C, O, H, P, and S, of the modified residues and residues involving disulfide bonds and cross-links. The SCSR contains 20 unique natural amino acid Tgroups. If the sequence is fully expanded, then it will contain about 256 200 heavy atoms. And thus, the SCSR-based sequence is, coincidently, 14% (= 35 360/256 200) of the size of the full structure format.

Building Sequence Database from Converting UniProtKB to the SCSR Format. To validate the efficiency of SCSR, we developed a conversion program called UniProtConverter. It takes a UniProt format file as input, extracts the sequence information and derives the modifications, and outputs the SCSR record as a molfile. This program has been integrated with Accelrys Direct and Accelrys Draw and allows the converted sequences and selected data (such as UniProt's sequence, entry name) to be directly registered into an Oracle table.

We downloaded the XML format of the UniProtKB database and converted the records into the SCSR-based molfile format. Additional detail about converting UniProtKB into the SCSR format can be found in the Supporting Information. A structure view of the SCSR-based sequence of the lantibiotic Pep 5 protein is given in Figure 16. Over 500 000 SCSR-based sequences have been stored in Accelrys Direct sequence database. Other UniProt data, such as entry name, accession number, and sequence, are also stored in the same Oracle

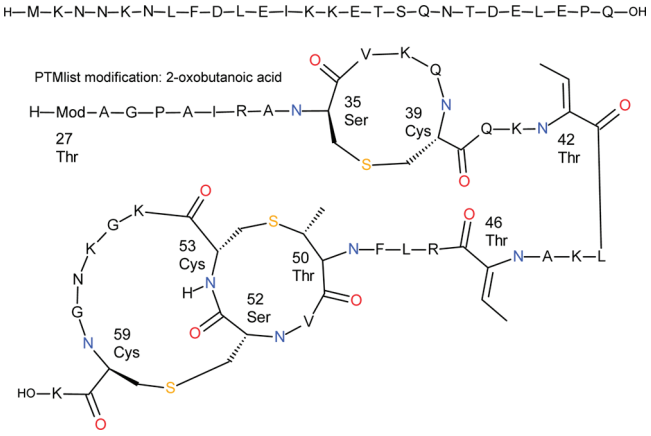


Figure 16. The SCSR format of the lantibiotic Pep5 protein sequence shown in Figure 1. Note that the templates associated with the sequence are not shown. The modified residues and residues with cross-links are marked with their three-letter symbol and SEQID in this figure for helping explanation only. The notation "PTMlist modif..." is associated with the Mod atom, indicating that the structure of the modified residue cannot be converted by the current version of the UniProtConverter program and thus is described as a "Mod" atom, and its modification is described as "2-oxobutanoic acid" in the UniProt's PTMlist.[25]

table. This "biocheminformatics" database is used for other investigation shown below.

SCSR-Based Sequence NEMA Key Allows Better Data Integration. Data integration involves combining data that reside in different sources. For example, new sequence data need to be registered into a central bioinformatics repository, or two merged companies need to merge their databases. Data integration is complicated, and many problems remain unsolved.[26] For example, Novo Nordisk's protein structural identity method can cause false positives.[9]

Besides, although the fragment-based screening method[7] can eliminate most of the database items during the exact structure search[7] of small molecule databases, this approach is not useful for structure searching of biomolecule databases because biopolymers consist of only a limited set of residues, and thus the screening keys are not selective. As such, the graph-based exact structure search method is too slow to be useful for duplicate checking and searching of biopolymer databases.

Here we show that the SCSR-based sequence NEMA key can be used as an efficient tool to check whether two given sequences have the same chemical structure.

Biopolymers are large chiral molecules. The SCSR-based NEMA key algorithm enables efficient generation of sequence NEMA keys for biopolymers by taking into account all the chemical structure information, including stereochemistry, disulfide bridges, and cross-links. The sequence NEMA key for the longest sequence (Titin) in UniProtKB is quite compact: TP3N8Y4G969QMKH9WDUQTKDQPWVURZ. This method also takes into account the bond direction information. This allows the generation of different sequence NEMA keys for the two cyclic sequences that are identical except the backbone directions (Figure 15). This method has been integrated with Accelrys Direct and allows precalculation of sequence NEMA keys for all sequences in the Direct sequence database for fast exact sequence structure searching and duplicate checking, an important step for data integration and integrity.

UniProtKB uses some annotation methods to minimize redundancy. If a given amino acid can have or not have a post-translational modification (PTM), the PTM is described as a partial PTM. If different PTMs can occur on the same amino acid, then the PTM is annotated as an alternate one. We ran a redundancy check of our sequence database to find out the number of sequence pairs that have the same UniProt sequence but differ in SCSR. First, each UniProt sequence is compared with all other sequences using the Oracle text comparison method. If two identical sequences are found, then their sequence NEMA keys are compared. The whole process took 69.5 min on a Dell Precision M6500 laptop, and the sequence NEMA key comparison took about 10% of the total time. Surprisingly, a total of 16 048 pairs of sequences that have the same UniProt sequences, but different sequence NEMA keys were found. This test led to some interesting findings. We provide two examples here: First, the UniProt "rule" of "partial modified residues" seems not to be followed strictly. For example, the sequences and modified residues of M2_I73A5[27] and M2_I72A3[28] are identical except that M2_I73A5 has additional three modified residues at positions 20, 82, and 93 (Figure 17). It seems that these three modified residues should be treated as partial modified residues and thus the two entries could be merged into one. The reason why they were treated as separate entries is because the two sequences come from different virus strains of the influenza A virus.

Second, PAPA1_MYCTA and PAPA1_MYCTU have the same sequences, but the former does not have mutagenesis, while the latter has two mutagenesis (H → A and D → A at positions 171 and 175, respectively). To reduce the redundancy, it seems that the concepts of partial and alternate modified residues should also be applied to mutagenesis.

The above examples show that the sequence NEMA key offers a fast method for sequence structure searching and redundancy checking. It is complementary to the existing bioinformatics approaches and also has potential applications in the data integration of traditional bioinformatics databases, such as UniProtKB.

SCSR Supports SSS of Biopolymers. SSS is one of the most important search methods in cheminformatics. Chemists use it to retrieve a set of chemical structures that all contain a given query substructure from a molecular structure database. The substructure matching is NP-complete[29] and CPU extensive. Here we show that the SCSR-based sequence database can be searched efficiently using the Accelrys FastSearch method.[24]

As an example, a biochemist might want to know: "How many sequences in the UniProtKB contain the substructure shown in Figure 18a, and what features do those sequences have in common?"

Using the structure Figure 18a as a query to perform SSS in our sequence database led to a total of 256 hits. The 256 sequences retrieved can be grouped into eight classes based on the modified structure that match the query (Table 1). The three residues that form the modified structures that contain the query Figure 18a are: ASG, CSG, SSG, AYG, EYG, MYG, NYG, and QYG, respectively. The majority of the sequences (237 in total) contain the ASG structure. The eight classes of sequences can be further grouped into two clusters. The first cluster includes the first three classes: ASG, CSG, and SSG, which have the same second and third residues of SG. The sequences of this cluster are all lyases: 47 phenylalanine ammonia-lyases, such as PAL3_PETCR and PALY_AMAMU, and 190 histidine ammonia-lyases, such as HUTH_AGRVS and HUTH_HUMAN.

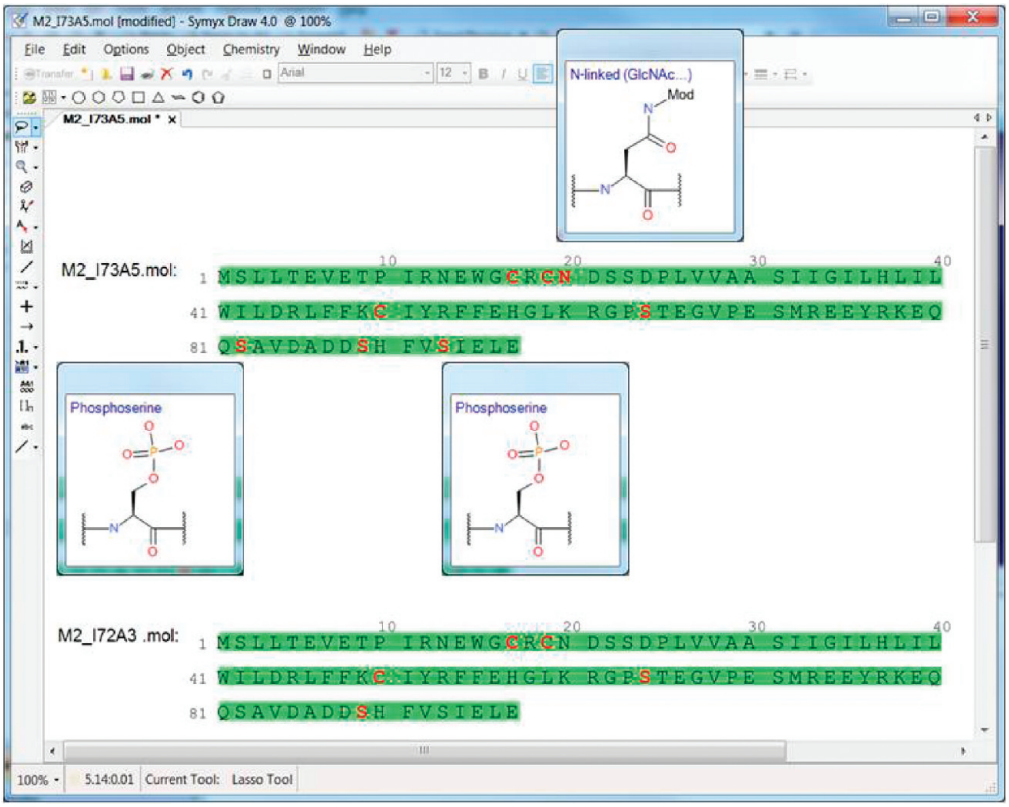


Figure 17. The sequences and modified residues (marked in red) of M2_I73A5 and M2_I72A3 are identical except that M2_I73A5 has additional three modified residues at positions 20, 82, and 93. The chemical structures of those three residues are shown in this figure.

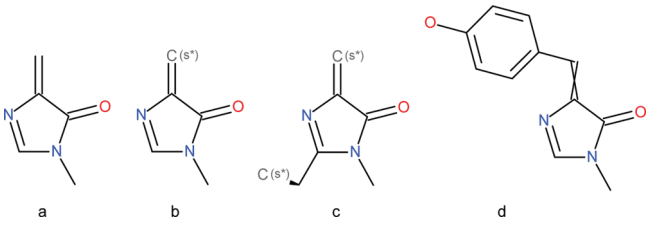


Figure 18. The SSS queries. Queries a and b are similar except that in b, the query option (s*) specifies that the carbon atom marked with (s*) can only have the number of nonhydrogen substituents as drawn.

The second cluster includes the remaining five classes: AYG, EYG, MYG, NYG, and QYG, which have the same second and third residues of YG. The 11 sequences of this cluster are all fluorescent or nonfluorescent pigment proteins that show certain colors. For instance, NFCP_CONPS and NFCP_CONGI are mauve nonfluorescent pigment proteins, while GFPL_DISST and GFPL_CLASP are GFP-like green fluorescent pigment proteins.

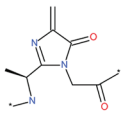
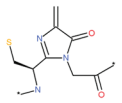
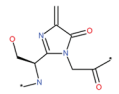
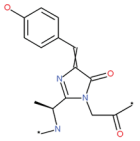
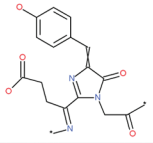
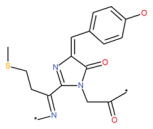
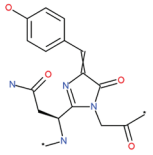
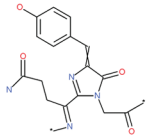
Another common feature of the 256 sequences is that all of them contain only the modified residues that matched the query structure Figure 18a but no other modifications and cross-links. The only exception is three histidine ammonia-lyases, HUTH_BOVIN, HUTH_HUMAN, and HUTH_MOUSE, that belong to the class ASG. These three lyases have the same length of 657 residues, and all contain three additional modified residues (phosphoserine) at the positions 631, 635, and 648.

More specific sequences can be retrieved by using more specific SSS queries. For example, using the structures Figure 18b and d as SSS queries led to the retrieval of only 245 sequences that belong to the first cluster and 11 sequences that belong to the second cluster, respectively. Using the structure Figure 18c as a query led to retrieval of only 237 sequences that belong to class ASG. Two hits are shown in Figure 19. The significant reduction of the structural complexity of biopolymers in the SCSR format makes SSS an efficient tool. For example, the longest sequence (Titin) in UniProtKB contains about 256 200 heavy atoms in the full chemistry format. In the SCSR format, it contains only 832 real chemical atoms from the modified and cross-linked residues that are the main focus of building the FastSearch index and performing SSS. Using the Accelrys Direct sample web application[30] from the Mozilla Firefox browser, SSS took an average of 0.086 s per query to search the above database for four SSS queries shown in Figure 18 on the Dell laptop.

The above examples show that PTM plays an important role in the functions of proteins and that the SCSR-based SSS offers a useful tool for searching sequences that contain a common modified substructure. SSS complements existing sequence searching tools, such as BLAST.

SCSR-Based Sequence View versus Bioinformatics Sequence View. A picture is worth a thousand words. The "picture" of molecule—the chemical structure—is the most important language common to all chemists. Although the bioinformatics sequence is simple (Figure 1), it lacks the ability to show the chemical structures of the residues. In contrast, SCSR allows the sequence view to be implemented in such a way that it not only displays the sequence in the format similar to that of bioinformatics sequence view but also has the ability to display disulfide bridge and cross-links as well as chemical structures of any residue in a convenient, intuitive manner, if needed.

Table 1. Sequences Retrieved Using the SSS Query Figure 18

UniProt Sequence Name	Three residues	Modified structure	Number of Sequences
HUTH_AGRVS HUTH_HUMAN ... PAL3_PETCR	ASG		237
HUTH_STRGR HUTH_STRAW HUTH_STRCO HUTH_CAEEL	CSG		4
HUTH1_FUSNN HUTH2_FUSNN HUTH_ALKMQ PALY_AMAMU	SSG		4
NFCP_CONPS NFCP_CONGI	AYG		2
NFCP_HETCR	EYG		1
RFP_PARAC NFCP_ANESU	MYG		2
GFPL1_ZOASP	NYG		1
GFPL_DISST GFPL_CLASP NFCP_MONEF NFCP_GONTE RFP_DISSP	QYG		5

The following example is used to compare the sequence views of a typical bioinformatics system, UniProt, and Accelrys Draw. The focus is on what benefits the SCSR-based sequence view can bring to the scientists. Ribonuclease S-2 (RNS2_NICAL) is a stylar glycoprotein that is associated with expression of self-incompatibility in the potato. From the UniProt Web site,[31] it can be seen that this protein consists of 214 amino acids that form two subsequences, a signal peptide (positions 1−22) and a ribonuclease S-2 chain (positions 23−214). The protein has three active sites at the positions 53, 109, and 113. It contains two disulfide bonds at 67 ↔ 116 and 175 ↔ 204, respectively. This protein also has three residues with the N-linked glycosylation at positions 49, 50, and 160 (Figure 20a).

The ribonuclease S-2 sequence was converted from the UniProt format to the SCSR-extended molfile format. We loaded this molfile into Accelrys Draw, and the sequence view of this protein is shown in Figure 20b. The two subsequences are differentiated using different background colors. The modified residues are highlighted in red. The disulfide bridges are shown as red lines. The active sites are not highlighted because this data was not captured by the current version of the UniProtConverter program.

We now compare UniProt and Accelrys Draw with regard to viewing chemical detail about the active site and residue modification. Clicking 53 or the line at the active site 53 in the graphical view of the UniProt Web site leads to a new webpage that simply highlights the histidine symbol, H, in yellow (Figure 21a), but no chemical structure information about that active site is available. In contrast, in the sequence view of Accelrys Draw, when the user right-clicks the histidine residue no. 53 and selects "View Structure", the structure of histidine that carries active sites (two nitrogen atoms in the ring) is shown (Figure 21b). The chemical structure of any standard amino acid residue can be shown in the same way.

For amino acid modification, in the UniProt Web site, clicking CAR_000106 in the glycosylation 49 line leads to the webpage shown in Figure 22a. For comparison, in Accelrys Draw, pointing the mouse cursor to the modified asparagine residue no. 49 brings up a display that contains its structure (Figure 22b).

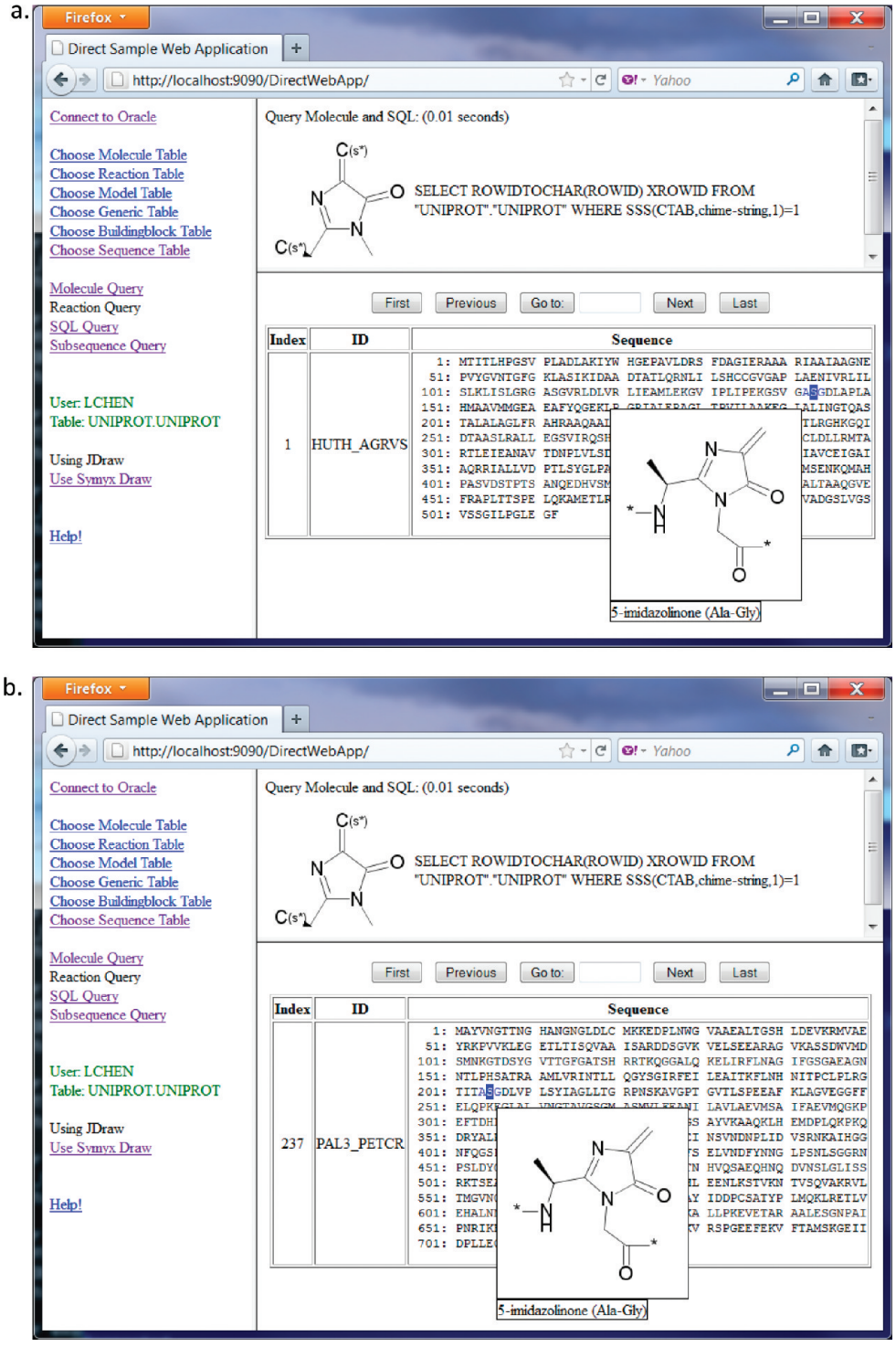


Figure 19. A total of 237 sequences that belong to class ASG were retrieved from the test database using query Figure 18c. Two of them are shown here. (a) The first hit is HUTH_AGRVS. (b) The last hit is PAL3_PETCR. Both hits contain only three modified residues, ASG. The structure formed by those modified residues can be shown via a mouse hovering to any of three residues: A, S, or G. In this figure, the mouse was on serine (S).

The mod atom attached to the nitrogen of the asparagine side chain and the N-linked (GlcNAc...) data indicate that the modification is an N-linked glycan.

For the disulfide bond, in the UniProt Web site, clicking 175 ↔ 204 or the corresponding line in the graphical view leads to the webpage shown in Figure 23a, where the two cysteine residues that are linked by a disulfide bond are highlighted in yellow, but no additional information is available. For comparison, in Accelrys Draw, the two cysteine residues that are linked with a disulfide bridge are highlighted in red and linked with a red line, and pointing the mouse cursor to either the cysteine at position 175 or the cysteine at position 204 will lead to automatically displaying the structures of the two cysteine residues and their disulfide bridge (Figure 23b).

dx.doi.org/10.1021/ci2001988 |J. Chem. Inf. Model. 2011, 51, 2186–2208

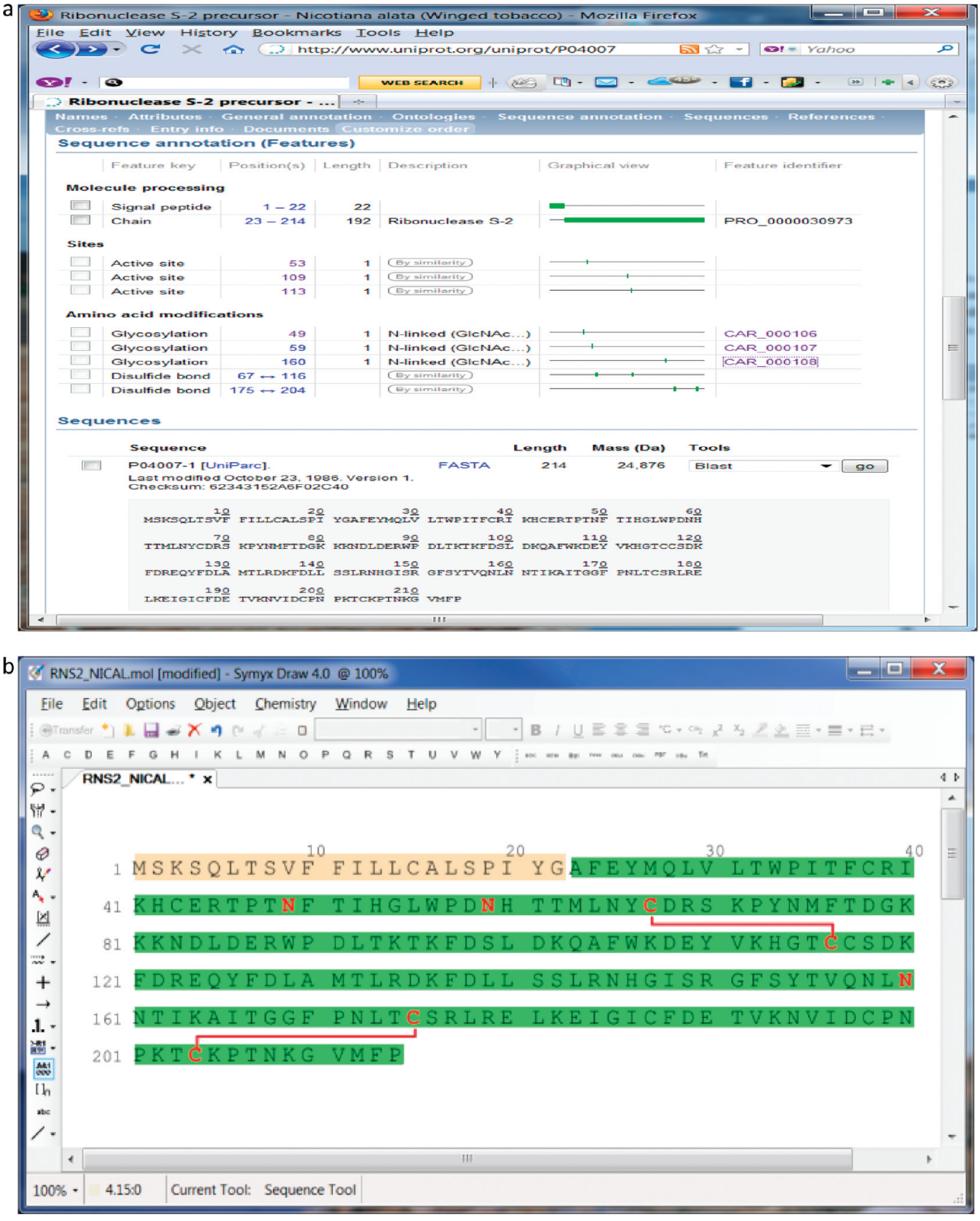


Figure 20. SCSR-based sequence view versus bioinformatics sequence view. (a) The sequence and its annotation of ribonuclease S-2 in the UniProt system. (b) The sequence view of ribonuclease S-2 in Accelrys Draw.

It should be noted that the chemical structure of any modified and cross-linked residue can be displayed by hovering the mouse over the residue name, without any mouse click.

The above example demonstrates the strength of the SCSR-based sequence view over the traditional bioinformatics sequence view.

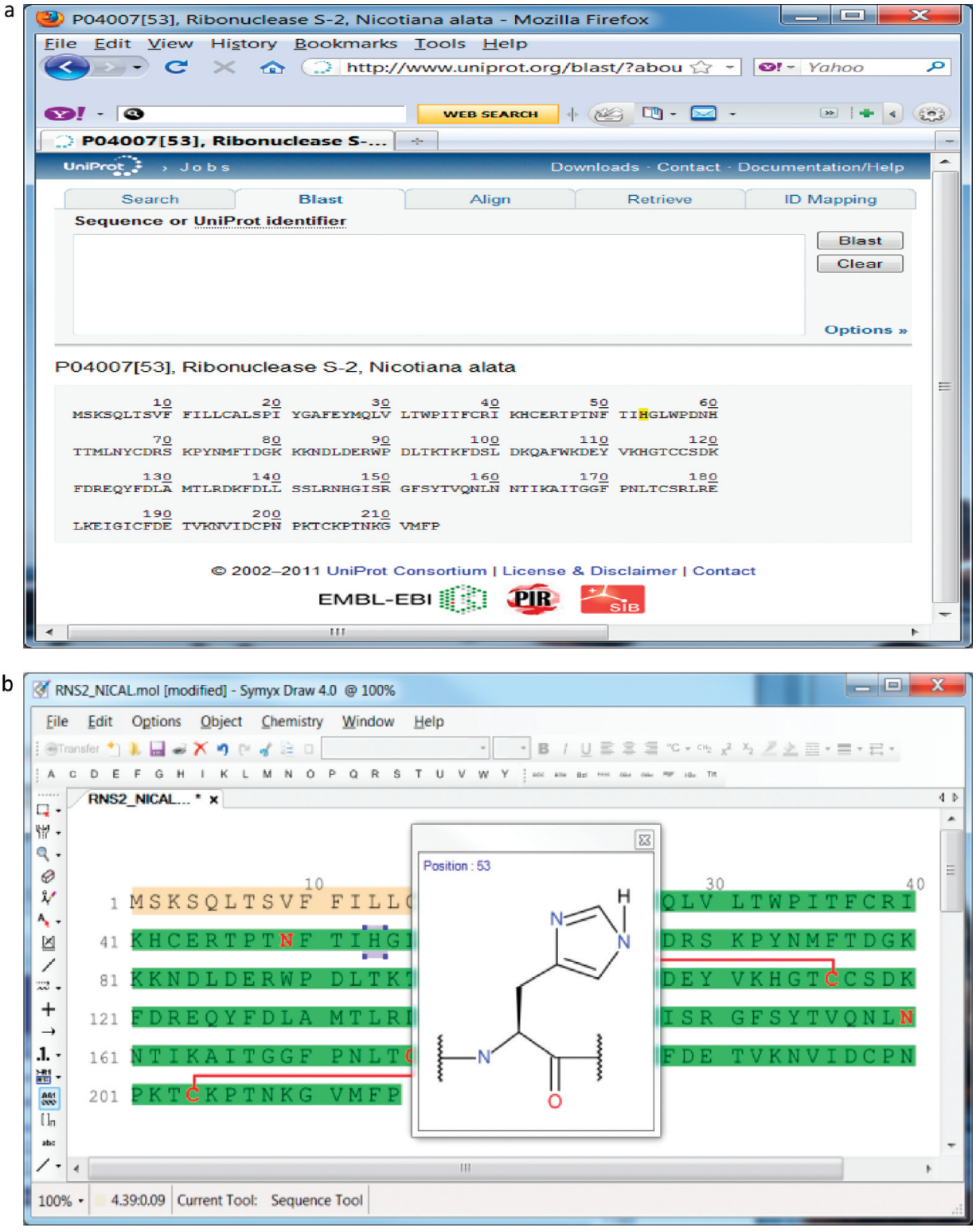


Figure 21. SCSR-based sequence view versus bioinformatics sequence view. Compare active sites: (a) UniProt only highlights the active site residue name, H, in yellow and (b) structure of the active site residue, His, is shown in Accelrys Draw.

SCSR-Extended Molfile Stores Sequences with Full Chemistry Detail. The Accelrys CTfile formats[13] are the most widely used structure file format in cheminformatics. For example, the default format for downloading chemical structures from the PubChem Download Service[32] is the Accelrys' SDfile format,[13] which is a collection of molfiles plus additional data. We have enhanced the V3000 molfile format to support SCSR. We demonstrated previously that the SCSR-based molfile can be used to build a large sequence database, and this database can be searched in a way similar to that used to search small molecule databases.

dx.doi.org/10.1021/ci2001988 |J. Chem. Inf. Model. 2011, 51, 2186–2208

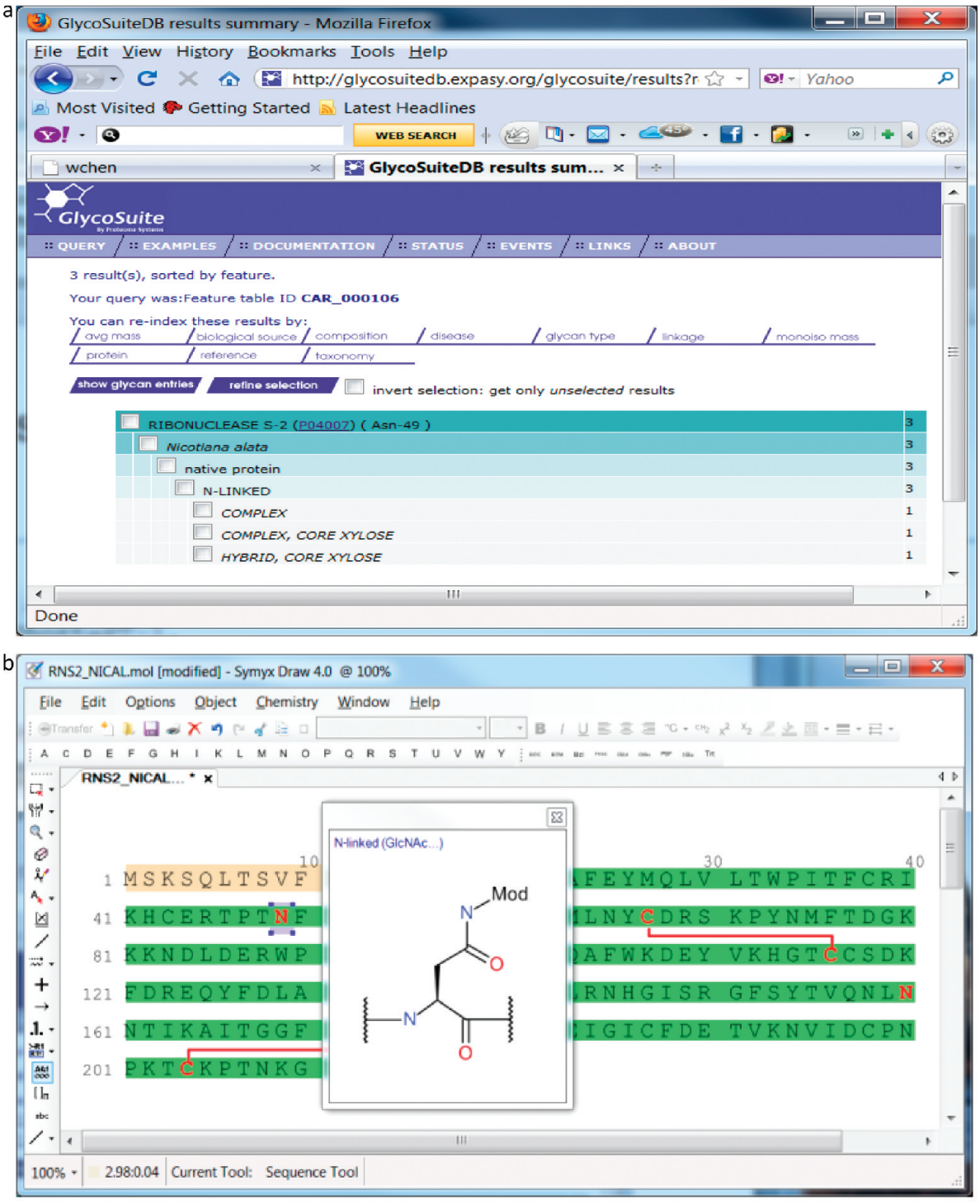


Figure 22. SCSR-based sequence view versus bioinformatics sequence view. Compare modified residues: (a) information on the glycosylation of Asn-49 in the UniProt system and (b) structure of the Asn residue that carries a N-linked glycosylation is shown in Accelrys Draw.

The sequence data in UniProtKB is compiled from the literature. The description of residue modifications using short texts can cause a loss of accuracy in chemical structures that exist in the original literature. This not only causes difficulties in converting sequences from the UniProt format to a chemical structure format like the SCSR format but, more importantly, brings up the question about the value to perform such conversion. For example, the UniProt Web site cites Weil et al's article[33] for the lantibiotic Pep5 protein sequence. In their paper, the chemical structures of all the modified and cross-linked residues are explicitly drawn. In fact, the structure view of the isolated pre-Peps sequence shown in that paper (reproduced in Figure 24) is similar to that of Figure 25. As such, it would make more sense to record such sequences directly in the SCSR format, if this format were available. This would ensure that all chemical structure information on a sequence that is available in the original literature is recorded accurately.

dx.doi.org/10.1021/ci2001988 |J. Chem. Inf. Model. 2011, 51, 2186–2208

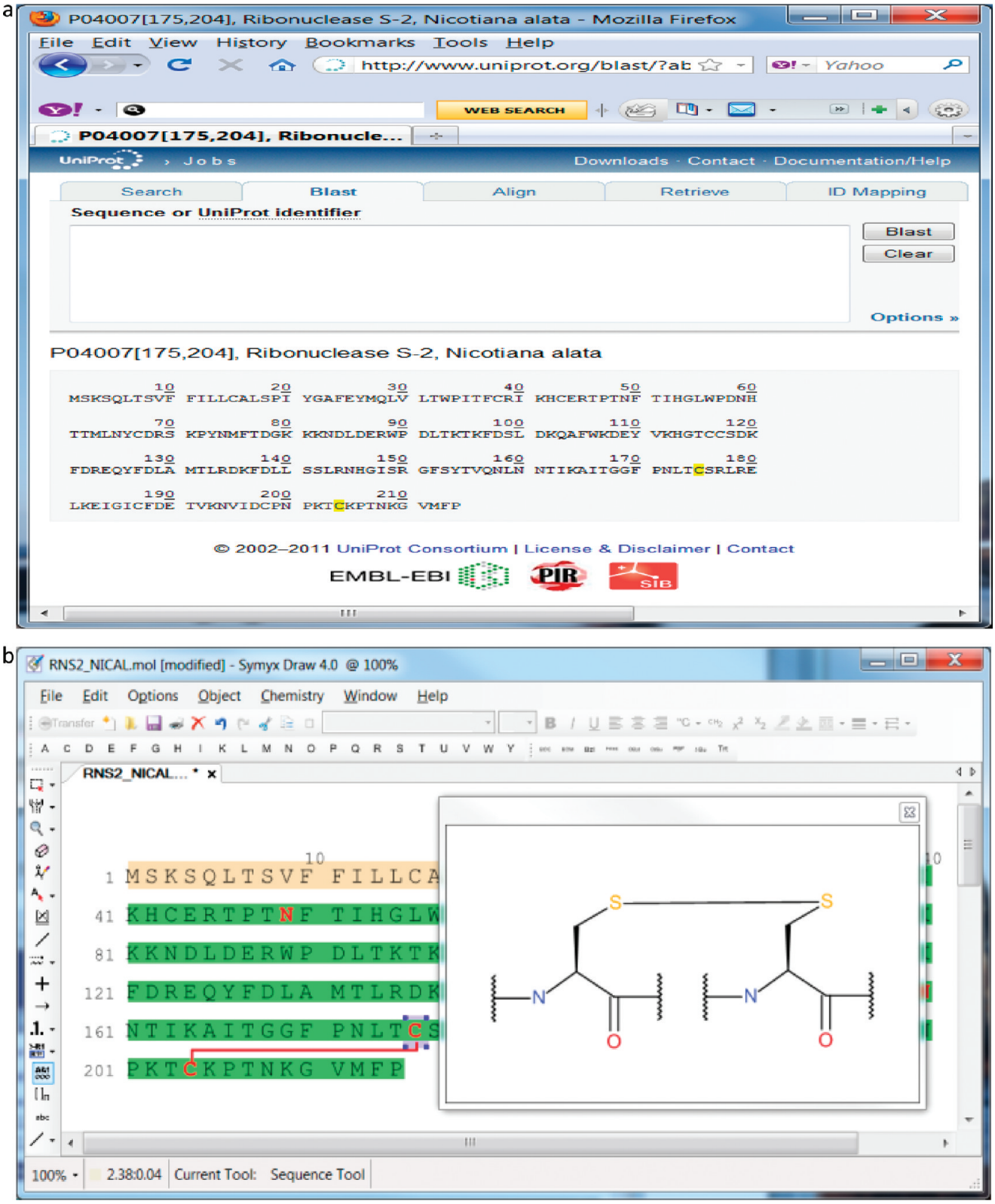


Figure 23. SCSR-based sequence view versus bioinformatics sequence view. Compare cross-links: (a) two cysteine residues that are linked via a disulfide bridge are highlighted in yellow in UniProt and (b) structures of the two cysteine residues and their disulfide bridge are shown in Accelrys Draw.

Therefore, we recommend that in the future, all protein sequences be recorded in both the existing bioinformatics formats and the SCSR-extended molfile format. Other types of sequences that contain unnatural residues should also be recorded in the SCSR format. This will allow all chemical structure information that was obtained through experiments

dx.doi.org/10.1021/ci2001988 |J. Chem. Inf. Model. 2011, 51, 2186–2208

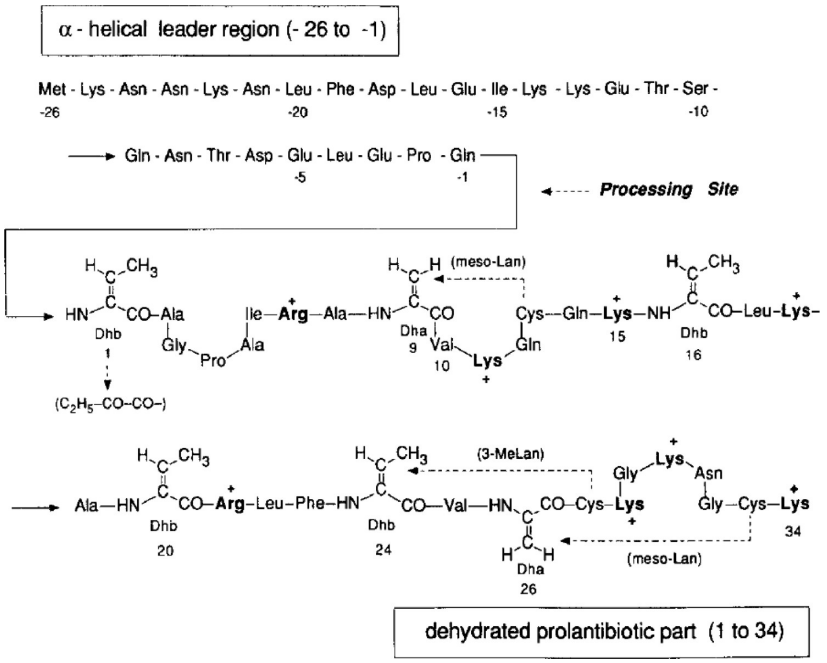


Figure 24. Structure of the isolated pre-peps sequence. (Courtesy of John Wiley and Sons, Inc.).

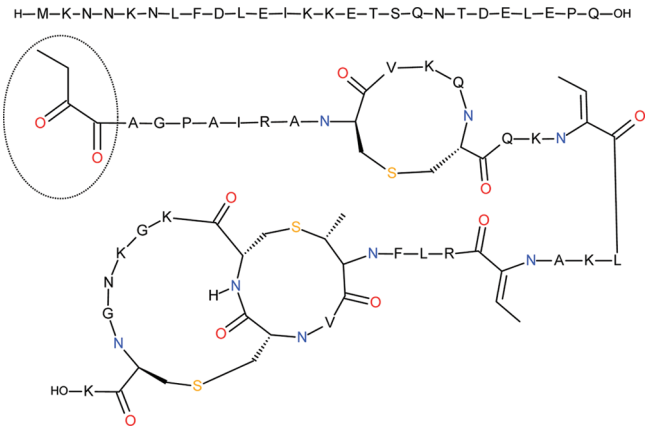


Figure 25. Correct structure of the lantibiotic pep5 protein sequence. The modified threonine structure is marked with a dotted cycle.

and/or sequence analysis to be recorded in full chemistry detail.

CONCLUSION

The power of the self-contained sequence representation (SCSR) is the ability to represent sequences in a compressed format without losing any chemistry detail. SCSR is the first general, comprehensive, and condensed representation of biopolymers at the chemistry level. SCSR addresses the performance challenge, the chemistry-loss hurdle, and the problem of representing pure cyclic sequences. SCSR allows keeping modified, cross-linked, and unnatural residues in explicit chemistry to emphasize their importance. SCSR can represent different types of biopolymers and any combination of them. In fact, SCSR can represent any macromolecular structure that consists of a limited number of structural units. SCSR may be extended to create a more compact sequence representation. For example, a protein domain may be defined as a "domain template", and thus the basic SCSR of protein may be further compressed into the second level of SCSR.

The SCSR offers flexibility of what, how, and when the full detail of chemistry should be used: (a) The chemistry details of the whole sequence can be used (such as deriving the structure activity relationship (SAR) from the full structures of a set of small sequences like peptides),[34] or only the structures of modified residues are needed (such as deriving SAR for a set of any-size biopolymers but the physicochemical properties are calculated from the structures of only modified residues (more detail on the calculation of physicochemical properties from the SCSR-based biopolymers can be found in the Supporting Information). (b) The chemistry detail of a sequence can be accessed by expanding template atoms and can also be used without expanding template atoms (such as the sequence NEMA key generation). (c) The chemistry detail can be used "statically" (such as generating NEMA keys once for a database) or dynamically (such as displaying chemical structure of a residue in a SCSR-based sequence view).

The SCSR-enhanced V3000 molfile format is the first general and comprehensive chemical structure file format that can record biopolymer structures in a compressed format without loss of chemical detail. It addresses the need for recording and transferring biopolymers with complete chemistry detail and yet provides good performance in cheminformatics, thus offering a powerful solution to help build biopolymer structure databases.

SCSR provides a solid framework for the future development of new cheminformatics tools and systems that can efficiently manage and handle large collections of biopolymers at the chemistry level. SCSR lays the foundation for the integration of the bioinformatics and cheminformatics systems to form a more comprehensive biocheminformatics system, such as Accelrys Direct. In this next-generation system, the bioinformatics and the SCSR-based sequences (and other biology data[35] and chemistry data) are stored in the same database side-by-side and

can be synchronized smoothly. This system offers unified search and handling across large sequences and small-molecule structures. It enables biologists and chemists to use respectively the bioinformatics and cheminformatics tools to search and handle biomolecular sequences and, more importantly, enables scientists to perform more sophisticated analysis and searches (such as the combination of SSS with ClustalW[36]) that are impossible using the standalone bioinformatics and cheminformatics systems, respectively.

■ ASSOCIATED CONTENT

Supporting Information. A sample global template molfile, additional details of converting UniProtKB to the SCSR format, and the calculation of physicochemical properties from the SCSR-based biopolymers. This information is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: williamlingran.chen@accelrys.com. Telephone: (925) 543 7541.

Present Addresses

‡PerkinElmer, 100 Cambridge Park Drive, Cambridge, MA 02140, United States.

■ ACKNOWLEDGMENT

This work has benefited from previous work by others of MDL Information Systems, the precursor organization to Symyx Technologies and Accelrys, Inc., and especially Dr. John Laufer who developed a compressed file format for use in earlier desktop programs. The authors thank Drs. F. Brown, P. Flook, and M. Hahn for valuable comments on the manuscript. In particular, we thank Dr. T. Albert for helpful suggestions for editing the manuscript.

■ REFERENCES

(1) Gathering clouds and a sequencing storm. *Nat. Biotechnol.* **2010**, *28* (1), 1.

(2) Sboner, A.; Habegger, L.; Pflueger, D.; Terry, S.; Chen, D. Z.; Rozowsky, J. S.; Tewari, A. K.; Kitabayashi, N.; Moss, B. J.; Chee, M. S.; Demichelis, F.; Rubin, M. A.; Gerstein, M. B. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.* **2010**, *11* (10), R104.

(3) Neumann, H.; Wang, K.; Davis, L.; Garcia-Alai, M.; Chin, J. W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **2010**, *464*, 441–444.

(4) Gibson, D. G.; Glass, J. I.; Lartigue, C.; Noskov, V. N.; Chuang, R.-Y.; Algire, M. A.; Benders, G. A.; Montague, M. G.; Ma, L.; Moodie, M. M.; Merryman, C.; Vashee, S.; Krishnakumar, R.; Assad-Garcia, N.; Andrews-Pfannkoch, C.; Denisova, E. A.; Young, L.; Qi, Z. Q.; Segall-Shapiro, T. H.; Calvey, C. H.; Parmar, P. P.; Hutchison, C. A.; Smith, H. O.; Venter, J. C. Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **2010**, *329*, 52–56.

(5) Brown, F. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annu. Rep. Med. Chem.* **1998**, *33*, 375–384.

(6) Chen, W. L. Chemoinformatics: Past, Present, and Future. *J. Chem. Inf. Model.* **2006**, *46*, 2230–2255.

(7) Chen, L. Substructure and Maximal Common Substructure Searching. In *Computational Medicinal Chemistry and Drug Discovery*; Bultinck, P., Winter, H. D., Langenaeker, W., Tollenaere, J. P., Eds.; Marcel Dekker: New York, 2004; pp 483−513.

(8) Taylor, K. T. Meeting the challenges of representing large, modified biopolymers. White paper; Symyx, San Ramon, CA; http://www.symyx.com/products/pdfs/biopolymer_whitepaper.pdf (accessed July 11, 2011).

(9) Jensen, J. H.; Hoeg-Jensen, T.; Padkjær, S. B. Building a BioChemformatics Database. *J. Chem. Inf. Model.* **2008**, *48*, 2404–2413.

(10) Chen, W. L. Self-Contained Sequence Representation: A Proposal; Unpublished work; Elsevier MDL, San Ramon, CA, 2005.

(11) Chen, W. L. The NEMA Algorithm for Stereochemistry Perception; Unpublished work; Elsevier MDL, San Ramon, CA, 2006.

(12) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(13) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.

(14) Symyx Direct bridges the gap between bioinformatics and cheminformatics, 2010. http://www.biovalley.com/content.cfm?nav=6&content=19&command=details&id=14130 (accessed June 30, 2011). BioValley: The Life Sciences Network: Illkirch, France.

(15) *New Drawing Software From Accelrys Bridges Chemistry and Biology*; Accelrys: San Diego, CA, 2010; http://ir.accelrys.com/releasedetail.cfm?releaseid=537233 (accessed June 30, 2011).

(16) *Accelrys Draw* − no fee; Accelrys: San Diego, CA; http://accelrys.com/products/informatics/cheminformatics/draw/no-fee.php (accessed July 11, 2011).

(17) *CTfile Formats*; Accelrys: San Diego, CA; http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php (accessed July 11, 2011).

(18) *NIST Secure Hashing*; NIST: Gaithersburg, MD; http://csrc.nist.gov/groups/ST/toolkit/secure_hashing.html (accessed July 11, 2011).

(19) Taylor, T. T.; Chen, W. L. NEMA key based exact match searching. White paper; Symyx Technologies Inc.: San Ramon, CA, 2008; http://www.symyx.com/products/pdfs/nema_whitepaper.pdf (accessed July 11, 2011).

(20) *The IUPAC International Chemical Identifier (InChITM)*; IUPAC: Research Triangle Park, NC; http://www.iupac.org/inchi (accessed July 11, 2011).

(21) *InChIKey Collision: Two isomers of spongistatin: One InChIKey*; The Goodman Group, University of Cambridge: Cambridge, U.K.; http://www-jmg.ch.cam.ac.uk/data/inchi (accessed July 12, 2011).

(22) Release Notes of IUPAC International Chemical Identifier (InChI): InChI version 1, software version 1.03, 2010.

(23) *UniProt*, release 2011_07; European Bioinformatics Institute, Swiss Institute of Bioinformatics, and Protein Information Resource (Georgetown University Medical Center: Hinxton, U.K., Lausanne, Switzerland, and Washington, D.C.; ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/relnotes.txt (accessed July 12, 2011).

(24) Christie, B. D.; Leland, B. A.; Nourse, J. G. Structure Searching in Chemical Databases by Direct Lookup Methods. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 545–547.

(25) *Ptmlist.txt*. http://www.uniprot.org/docs/ptmlist (accessed July 11, 2011).

(26) Ziegler, P.; Dittrich, K. R. Three Decades of Data Integration - All Problems Solved? In 18th IFIP World Computer Congress (WCC 2004), Toulouse, France, August 22−27, 2004; WCC: Toulouse, France, 2004; Building the Information Society, vol *12* , pp 3−12.

(27) P63232 (M2_I73A5) reviewed, UniProtKB/Swiss-Prot; http://www.uniprot.org/uniprot/P63232 (accessed June 30, 2011).

(28) Q463X4 (M2_I72A3) reviewed, UniProtKB/Swiss-Prot; http://www.uniprot.org/uniprot/Q463X4 (accessed June 30, 2011).

(29) Cook, S. A. The complexity of theorem proving procedures. *Proceedings, Third Annual ACM Symposium on the Theory of Computing*; ACM: New York, 1971, pp 151−158.

(30) The Accelrys Direct sample web application is a Java-based Web search tool. It was developed to demonstrate how to retrieve and display SCSR-based sequences stored in the Accelrys Direct database.
(31) P04007 (RNS2_NICAL) reviewed, UniProtKB/Swiss-Prot; http://www.uniprot.org/uniprot/P04007 (accessed July 12, 2011).
(32) PubChem Download Service; http://pubchem.ncbi.nlm.nih.gov//pc_fetch/pc_fetch.cgi (accessed July 11, 2011)
(33) Weil, H.-P.; Beck-Sickinger, A. G.; Metzger, J.; Stevanovic, S.; Jung, G.; Josten, M. Biosynthesis of the lantibiotic Pep5. Isolation and characterization of a prepeptide containing dehydroamino acids. *Eur. J. Biochem.* **1990**, *194*, 217–223.
(34) Mangoni, M. L.; Carotenuto, A.; Auriemma, L.; Saviello, M. R.; Campiglia, P.; Gomez-Monterrey, I.; Malfi, S.; Marcellini, L.; Barra, D.; Novellino, E.; Grieco, P. Structure-Activity Relationship, Conformational and Biological Studies of Temporin L Analogues. *J. Med. Chem.* **2011**, *54*, 1298–1307.
(35) Pihl, T. D.; Ribaudo, R. K. The need for a biological registration system. *IDrugs* **2010**, *13*, 388–93.
(36) Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T. J.; Higgins, D. G.; Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **2003**, *31*, 3497–3500.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on August 22, 2011, with minor text errors in the Template block of the NEMA section. The corrected version was published ASAP on August 26, 2011.