

WHITEPAPER

MEETING THE CHALLENGES OF REPRESENTING LARGE, MODIFIED BIOPOLYMERS

By Keith T. Taylor, Ph.D., M.R.C.S.

ABSTRACT

This white paper discusses the challenge of developing a chemically-aware method of representation for large, modified biopolymers that is meaningful for both biologists and chemists, reduces redundant information and enables the electronic searching of structural features. The paper compares the methods used by chemists to represent chemical structures with those used by biologists to depict biological structures, employing the synthetic erythropoiesis protein (SEP) to illustrate the extreme complexity encountered at the interface between chemical and biological research where biological entities are often modified to change their biological activity. The paper concludes by explaining the four methods for recording chemical structures supported by the Accelrys Direct (previously Symyx Direct) data cartridge and describes recommended approaches.

Accelrys recognized the growing importance of bio-molecules in the pharmaceutical industry over a decade ago. To meet customer needs, the atom and bond limits in the early molfile format (V2000) were removed with the introduction of the V3000 molfile format. In addition, biopolymer drawing capabilities were included in the new and improved Accelrys Draw (previously Symyx Draw) structure editor, enabling scientists to draw, register, search and report on chemically modified peptide and nucleotide sequences. The Accelrys Direct data cartridge imposes no arbitrary limits on the size or composition of the structures. The combined capabilities of Accelrys Draw and Accelrys Direct meet the challenges of registering large, modified biopolymers today.

The interface between chemistry and biology

Chemical structure diagrams are the universal language of chemistry. Chemists exchange information using diagrams that show how atoms and bonds are connected, and chemists throughout the world understand the meaning of these diagrams regardless of the natural languages spoken by originators

and receivers. Overall this approach works very well even in cases where a structure or substructure in question cannot properly be represented by a tidy arrangement of atoms and bonds. In such cases chemists have adopted conventions that convey the required information. There are, in fact, a number of variants for some small functional groups—the classic example is the nitro group. Chemists argue over which representation is better, but much as British English uses “centre” and American English uses “center,” all chemists understand each other.

Biologists have similar conventions for the representation of biological structures, but they are normally less interested in the fine detail of how all the atoms and bonds are connected within a substance. The properties and behavior of biological molecules are controlled by their overall three-dimensional shape—much as the shape of a house follows the arrangement of the bricks from which it is made and not the chemical composition of the fired clay that constitutes each brick. In most cases, the biologist or biochemist only needs to know which amino acids, in what order, constitute the protein or peptide or the bases that make up RNA and DNA. Adding a small collection of sugars to this mix defines the biological bricks.

Biological systems also contain small molecules that behave as modifiers, for example, neurotransmitters (dopamine, serotonin, acetylcholine, hormones), the plant auxins (indole-3- acetic acid, gibberellic acid), testosterone and adrenaline/epinephrine. Chemists are interested in isolating and describing the structures of these substances as well as synthesizing the substances and potentially more active analogs. Biologists are interested in determining how the substances produce their effects and what this tells us about disease states.

The interface between chemical and biological structures

The two systems work well in isolation but difficulties arise at the interface where biological entities are chemically modified to change their biological activity, or to enhance their bioavailability. An example is the synthetic erythropoiesis protein (SEP) shown in Figure 1.

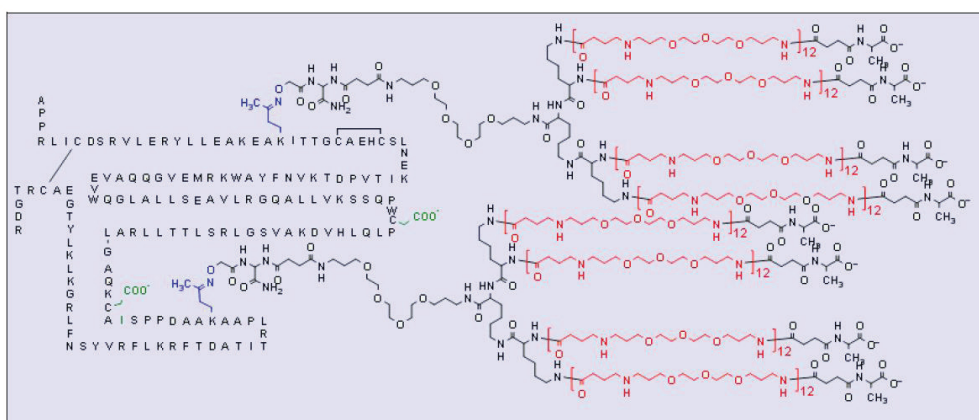


Figure 1: Synthetic erythropoiesis protein - SEP (Gerd G. Kochendoerfer et al: Science, 299, 884, 2003)

SEP has a molecular weight of approximately 50K Daltons and contains more than 7,500 atoms. This amount of information puts a heavy, and largely unnecessary, load on the chemical structure drawing program, the database searching engines and the scientist who has to draw the structure. The mass of information also consumes significant bandwidth when transferred around networks.

Of the 166 residues in this substance, only four are modified. In addition, the two polymer side chains are identical and contribute approximately 5,000 atoms to the structure.

Creating meaningful structures for chemists and biologists

The structure in Figure 1 uses the chemist's style of representation, displaying the amino acids in abbreviated form using the one-letter code convention. Although each amino acid is displayed as a single letter, underneath the abbreviation is the full atom and bond connectivity of the residue (see side chains attached to letter-K).

The synthetic polymer side chains in this representation already exploit the Accelrys shorthand representation for repeating groups. Although only one substructure is shown within each of the square brackets, Accelrys software interprets them as repeating 12 times each.

The challenge—developing a meaningful, efficient, electronically searchable method of representation

The biochemistry industry needs a method of representation for large structures that is meaningful to biologists and chemists, reduces redundant information and enables structural features to be searched using a computer system.

The Accelrys solution. The Accelrys Direct chemistry data cartridge manages both molecular structures and reactions and features industry-leading chemical and stereochemical representation. The representation system uses standard codes for those parts of a structure that are unaltered. It employs full structural representation for non-standard substructures. In addition, the system enables scientists to define and use self-defined or company-specific residues.

Practical considerations. Accelrys Direct has no arbitrary limits on the number or size of structures that it can manage. However, the larger the structure, the more work is involved in matching it against other entries in the database, and matching it with queries. In addition, moving large amounts of atom and bond information around consumes significant system resources. Finally, rendering a large structure puts heavier demands on chemical drawing applications.

Four representation Options

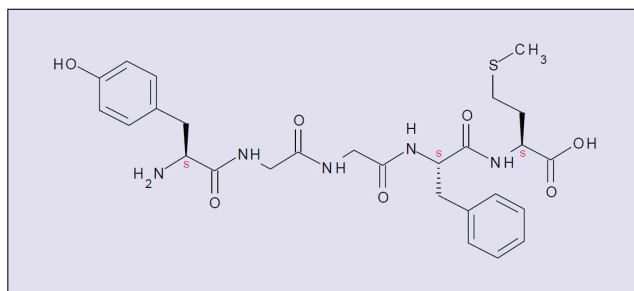
Accelrys Direct offers four options for recording large chemical structures, all of which work for all structure sizes:

- Full atom-bond connectivity
- Full atom-bond connectivity with abbreviations
- Pseudoatom representation
- *Atom (star atom) representation with Data Sgroups

The four options can be mixed, and, as we shall see, the combination of full atom-bond connectivity for novel residues coupled with *atoms for standard residues provides the best balance of flexibility, easy-to-interpret structures, compactness and efficient manipulation.

A simple pentapeptide illustrates these four methods.

Full atom-bond connectivity. This familiar representation method for chemists has proved very useful for describing small molecules. A chemist working with the structure below will understand it to be met-enkephalin, but others less familiar with the chemical structure may not find the interpretation obvious.



Met-enkephalin is a pentapeptide that can be described efficiently in terms of the amino acid residues in order. In this case, the structure is made up from tyrosine, two units of glycine, one unit of phenylalanine, and it is terminated by methionine. Scientists working with substances like this have adopted three-letter and one-letter codes for the amino acid residues. In this terminology, met-enkephalin is represented as either Tyr-Gly-Gly-Phe-Met or YGGFM. These representations are very compact, and the relationship between met-enkephalin and leu-enkephalin (Tyr-Gly-Gly-Phe-Leu, or YGGFL) is clear and obvious. Also, the relationship between met-enkephalin and the neuropeptide Lipotropin C Fragment (Tyr-Gly-Gly-Phe-Met-Thr-Ser-Glu-Lys-Ser-Gln-Thr-Pro-Leu-Val-Thr-Leu-Phe-Lys-Asn-Ala-Ile-Ile-Lys-Asn-Ala-Tyr-Lys-Lys-Gly-Glu) is also easier to see.

The classical chemical structure representation of met-enkephalin contains 71 atoms. The residue representations reduce the number of units to 5, and are more understandable.

Full atom-bond connectivity with abbreviations. The Accelrys Draw structure editor is configured to generate peptides using the standard one-letter or three-letter codes, and Accelrys Direct recognizes the full connectivity of such structures. This abbreviated display affects only the presentation. The number of atoms within the structure remains unchanged at 71.

Y G G F M Tyr—Gly—Gly—Phe—Met

Pseudoatom representation. The Accelrys periodic table (PTable) has provision for 200 entries. The table contains entries for 111 elements and a number of special atom types, for example, Rgroups and the 20 common, naturally occurring amino acids where the pseudoatom symbol corresponds to the amino acid three-letter abbreviation. The entries that do not correspond with elements are called pseudoatoms. There are currently 40 undefined pseudoatoms that can be customized to represent other commonly used residues, in addition to the 20 predefined amino acids.

The pseudoatom representation appears identical to the abbreviated three-letter code form, but it now contains the equivalent of only 5 atoms.

Tyr—Gly—Gly—Phe—Met

This approach is limited to approximately 60 biological pseudoatoms. If you are already using pseudoatoms to represent biological molecules, and the restricted number of additional entries is not a problem, then Accelrys recommends that you stay with this approach.

***Atom (star atom) representation with Data Sgroups.** This option combines two features of the Accelrys Direct data cartridge—the *atom and Data Sgroups—with the object of accommodating a larger number of collapsed representations. The *atom may be considered a NULL-atom, i.e., it occupies a position in a structure but has no mass or defined valency. The Data Sgroup technology enables a name and a mass to be attached to a particular *atom.

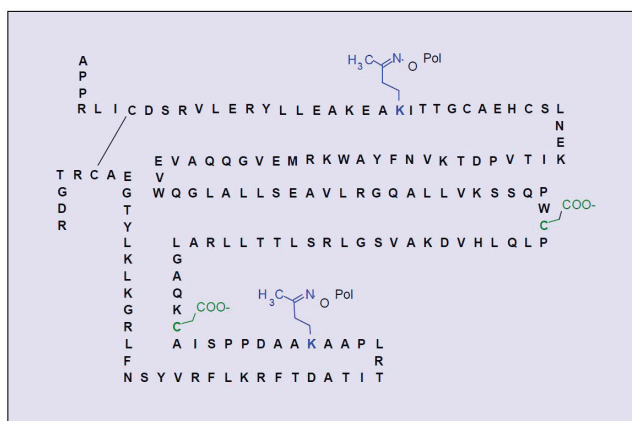
Again, the display appears identical to the abbreviated form, but, unlike the pseudoatom style, there is no arbitrary limit on the length of the codes that can be used to identify the residues— natural residues, unnatural residues and chemically modified residues. Using the *atom representation, the structure contains the equivalent of 5 atoms.

Y G G F M Tyr—Gly—Gly—Phe—Met

There is no limit to the number of *atoms that can be defined and used in a structure. Similarly, the Data Sgroup does not apply arbitrary limits to the names that may be given to a *atom. Sgroup information is structurally significant in the Accelrys Direct database. Hence, *atoms may be used in queries and a sequence can be searched by sub-sequence (or by substructure in chemist's terminology).

Organizations utilizing alternative representation approaches or considering a new system should consider upgrading to the *atom option, as it offers the greatest flexibility.

Custom residues. In the SEP example (Figure 1), the polymer side chain could be defined as a *atom with a custom name and formula mass. PEG side chains are excellent candidates for definition as *atoms or pseudoatoms.



The procedure for creating or modifying residue templates that use the *atom approach (and the pseudoatom approach) is fully described in the document entitled: Chemical Representation that is included in the documentation shipped with Accelrys Direct, Accelrys Draw and Accelrys Cheshire (previously Symyx Cheshire) software.

Mix-&-match. It is possible for a structure to contain any combination of the above four approaches, but it is not recommended to mix the pseudoatom approach with the *atom approach. Choose one of the approaches and standardize on it. Abbreviations may be mixed with any combination because they are only a display feature, which is always backed by full atom-bond connectivity.

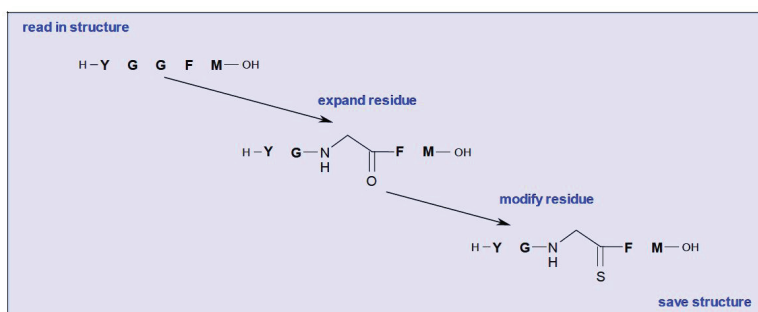
A mix-&-match approach for SEP gives the following representation:

This representation reduces over 7,500 atoms to the equivalent of only 206 atoms. This is a significant reduction that greatly enhances structure manipulation without any loss of information. The nature of the peptide can be readily understood, and the chemically modified regions and the nature of the modifications are immediately obvious. The structure may be searched by sequence and by substructure, enabling comparisons with related substances and the timely development of structure-property/structure-activity tables.

Accelrys Draw supports residue formats

To simplify the structure-drawing workflow, Accelrys Draw has been adapted to work with residue representations. It is delivered with templates for the standard structure-abbreviated forms, and the *atom and pseudoatom representations. It also includes template residues for common PEG residues, and frequently used protecting groups. In addition, the templates can be augmented with company-specific residues.

Accelrys Draw simplifies the drawing and editing process by always working on full structures in the structure-abbreviation style. The full structure form is translated into the *atom (or pseudoatom) format when it is saved, and the reverse conversion takes place when the structure is read in. This enables the scientist to take a *atom-based structure and edit it, perhaps by chemically modifying a residue. When the modified structure is saved, all the residues in the residue definitions are condensed to *atoms, and non-standard regions remain as expanded chemical structures.

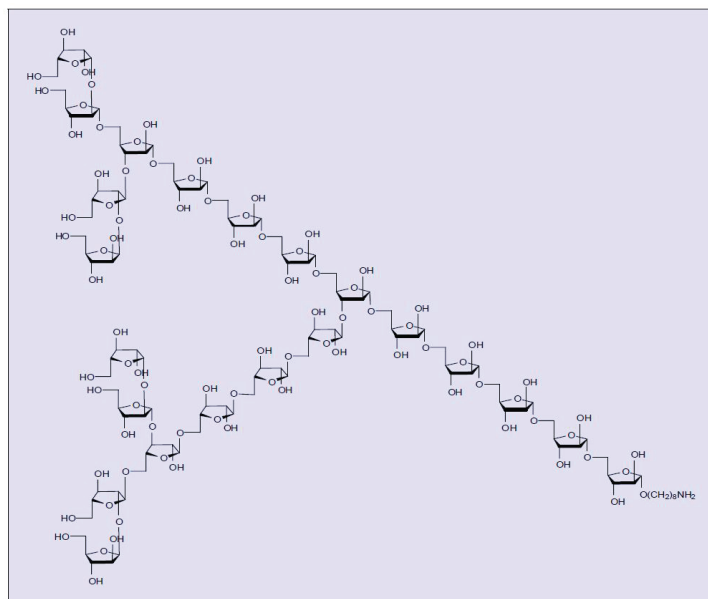


Recommendations

- The *atom or pseudoatom representation is preferred. It is extensible and will be enhanced in future releases.
- If your organization is already using the pseudoatom representation and the limited extensibility is not a problem, then you can continue with its use. Accelrys does, however, strongly recommend that you migrate to the *atom approach because of its greater flexibility.
- Configure Accelrys Draw and Accelrys Direct to use the preferred representation. Consider updating existing structures to the new default. The Accelrys Cheshire chemical scripting language offers capabilities that can automate the upgrade.

DNA and RNA sequences. DNA and RNA sequences may be managed using a similar approach, and templates for the appropriate nucleoside monophosphates are included with Accelrys Draw.

Carbohydrates. Accelrys is researching the extension of the *atom approach to oligosaccharides. Oligosaccharides are the focus of much synthetic effort, for example, the recently published synthetic arabinofuranose shown below. Many synthetic challenges were overcome, including the stereoselective incorporation of four β -arabinofuranoside units.



The large number of possible natural and modified carbohydrates coupled with the need to accommodate multiple linking sites complicates the design of an easy-to-use interface for Accelrys Draw.

Annotation of structures

With Accelrys Direct, you can annotate all or part of a structure. Annotating chains (Chain A, Chain B) to identify an active site is an obvious use of this feature. By using these structuredifferentiating labels, you can restrict queries to a particular chain, or the active site region.

Connection with bioinformatics tools

A sequence obtained as a text string can be imported (pasted) into Accelrys Draw and automatically converted to a Accelrys structure. Similarly, the pure sequence can be extracted from the Accelrys record and registered in a sequence database. Accelrys Direct is a pure Oracle® data cartridge. As such, a trigger can be placed on every column containing a sequence—automatically producing a structure representation from the sequence and registering it into Accelrys Direct.

Property calculations

Currently Accelrys Direct does not incorporate property calculators, but its architecture allows information to be exchanged with industry-standard calculators and the results to be registered in the appropriate table or tables.

Summary

Accelrys Direct, Accelrys Cheshire and Accelrys Draw are a coordinated set of applications supporting the efficient, flexible representation of biopolymers. The recommended *atom representation does not conflict with other applications and representations used by an organization—provided that all names and fields are unique. It is possible to modify structures at the residue level (one or many, contiguous or separated).

Scientists can attach annotations to all or parts of a structure

Scientists can restrict searches to annotated regions of a structure.

Accelrys Cheshire handles the conversion between the full atom-bond format and the *atom format.

Accelrys Cheshire understands both styles. It also understands the pseudoatom style.

Using Accelrys Cheshire scripts, scientists can manipulate structures outside of Accelrys Direct and Accelrys Draw.

To learn more about Accelrys Informatics, go to

<http://accelrys.com/products/informatics>