

## Building a BioChemformatics Database

Jan H. Jensen,<sup>\*,†</sup> Thomas Hoeg-Jensen,<sup>‡</sup> and Søren B. Padkjær<sup>§</sup>

Scientific Computing, Novo Nordisk Park A2P, Novo Nordisk A/S, DK-2760 Maaloev, Denmark,  
Diabetes Protein and Peptide Chemistry, Novo Nordisk Park D6.1.142, Novo Nordisk A/S,  
DK-2760 Maaloev, Denmark, and Protein Structure and Biophysics, Novo Nordisk Park G8.2.78,  
Novo Nordisk A/S, DK-2760 Maaloev, Denmark

Received April 12, 2008

The structural registration of chemically modified macromolecules is vital for the development of biopharmaceuticals. However, registration and search of such complex molecules has so far posed formidable challenges performance-wise, since today's chemistry-oriented databases do not scale well to macromolecules. As a practical consequence, macromolecules tend to be stored in protein databases with a focus on protein sequence only, and salient chemistry details are therefore lost. This article describes protein format extensions and the use of pseudoatoms for representing natural amino acids in chemical structures to allow high-performance registration and retrieval of large macromolecules. The representations include exact chemical modifications and enable lossless conversion between chemistry and sequence formats. Registration is done in parallel in both sequence and chemistry formats, and users can register and retrieve molecules in either format as they choose, resulting in what we call a BioChemformatics database. Having both sequence and chemistry formats available on-demand allows for the construction of protein SAR tables with mixed sequence and chemistry information. Likewise, searching may combine sequence and chemistry terms and be performed in standard vendor applications like MDL's ISIS/Base or in-house applications using standard SQL queries.

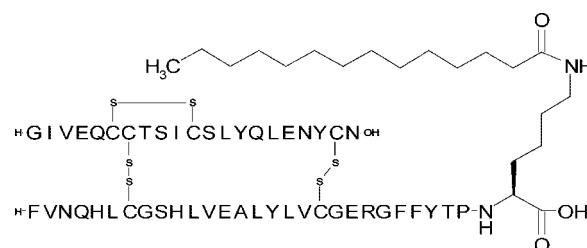
### INTRODUCTION

Controlled preparation and detailed registration of proteins with post-translational and chemical modifications is key to successful engineering of drugs based on naturally occurring proteins.<sup>1</sup> Recent findings indicate that post-translational modifications are much more common than previously believed.<sup>2</sup> Also, the available methods for preparation of semisynthetic or fully synthetic proteins have improved dramatically in recent years.<sup>3,4</sup>

The insulin analogue detemir (Levemir) seen in Figure 1 is a well-known example of an engineered variant of a human hormone.<sup>5,6</sup> Another example is recombinant Factor VIIa, an essential component for hemostasis. Factor VIIa is depicted in Figure 2 along with the set of glycane structures believed to attach to four of its residues. These glycane structures putatively control tissue-specificity and are amenable to protein chemistry engineering to yield analogues of the molecule better suited for drug use.

These two examples are representative of compounds that size-wise are in the upper range of chemoinformatics and the midrange of bioinformatics. From a bioinformatics/protein perspective, insulins are relatively short sequences, with only around 50 amino acids. However, from a chemoinformatics/chemistry perspective insulins are very large molecules.

Bioinformatics applications and tools use a sequence level description of molecules as chains of amino acids. Each amino acid is coded for by an alphabetic letter, and the



**Figure 1.** Insulin detemir. The chemical modification on the lysine prolongs and stabilizes the effect of injected insulin significantly.

resulting string of letters thus yields the sequence of the macromolecule. The formats used for electronic exchange of these macromolecules may combine the sequence with textual annotations that describe various features of the molecule (disulfide bridges, chains, secondary structure, etc.) but rarely any exact chemistry details, e.g. the UNIPROT format.<sup>7</sup>

On the other hand, chemoinformatics applications describe a molecule as a network of atoms and bonds, as in the MDL molfile format.<sup>8</sup> This is a very successful approach for small molecules, but much information is redundant when describing macromolecules, as most atoms and bonds are located in the amino acid chains. More importantly, the large number of atoms and bonds in macromolecules renders a full-structure based system very slow when working with proteins having more than 100–200 amino acids. Examples of such large proteins as full chemical structures in MDL molfile format are available on the Web.<sup>9</sup>

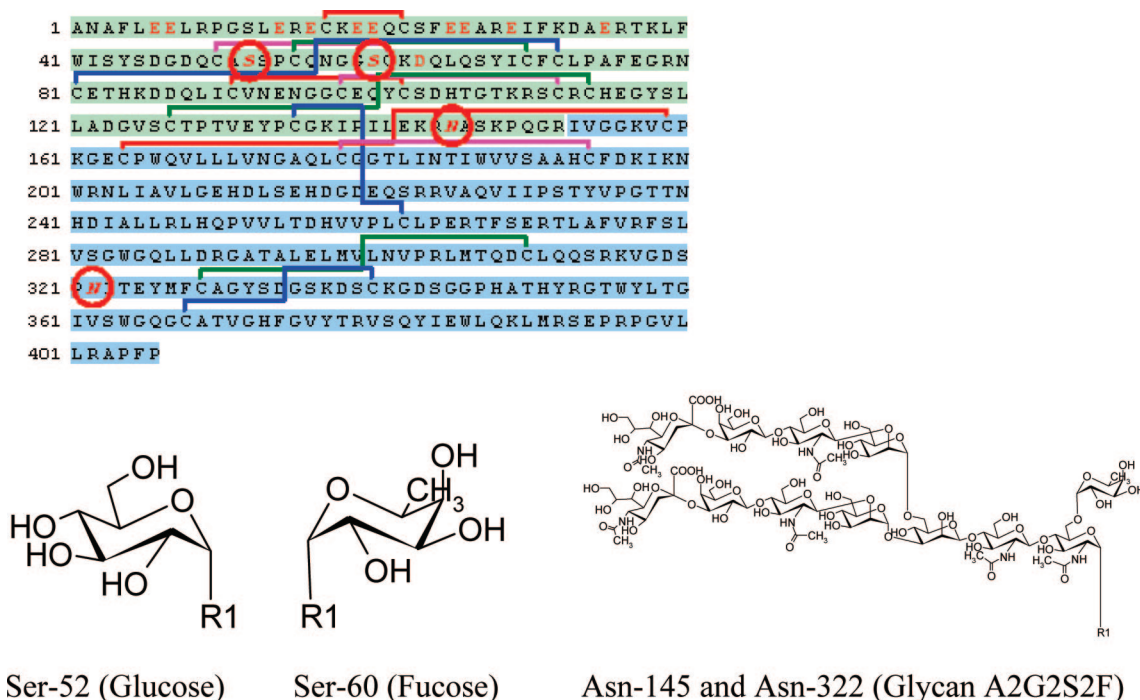
Protein chemists are therefore faced with a dilemma when they consider molecule registration. They like their molecules stored in a chemistry database to capture vital chemical

\* Corresponding author e-mail: jan@jan-holst.dk.

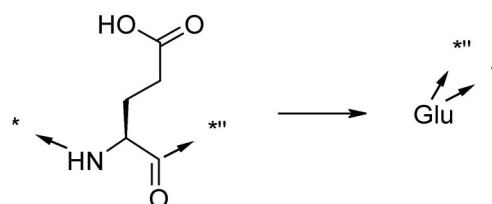
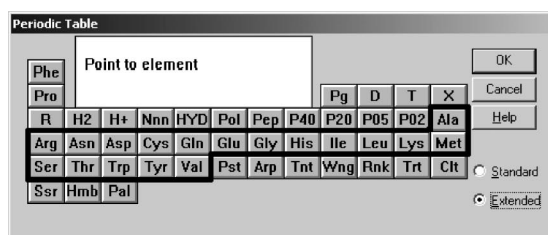
† Now employed at Biochemfusion.

‡ Diabetes Protein and Peptide Chemistry, Novo Nordisk A/S.

§ Protein Structure and Biophysics, Novo Nordisk A/S.



**Figure 2.** FVIIa with what is currently believed to be the normal set of glycans. The sequence display is rendered by the Novo Nordisk Protein Editor. The residues with glycans attached have been emphasized by red circles by the authors.



**Figure 3.** The Extended periodic table of MDL ISIS/Draw which includes atom symbols, also known as “pseudoatoms”, for the 20 natural amino acids (outlined by the authors). To the right an example representing a whole Glu residue with a single pseudoatom. The pseudoatoms that represent natural amino acids are called “residue atoms” in the main text.

details and they would also like the molecules to be accessible by bioinformatics tools to e.g. analyze the effect of point mutations. The dilemma is painfully obvious in the case of molecules in a size-range where both formats are equally appropriate, e.g. insulins that have around 50 amino acids. An ideal solution for registration of biopharmaceutical compounds would therefore consist of the following:

- \* A database that accepts compounds for registration in either protein or chemistry format.

- \* The database will provide compounds in protein or chemistry format on request regardless of the original registration format.

- \* Queries can be performed at the chemistry level, the sequence level, or a combination of both.

A solution that seems to fulfill these requirements was already implemented: The CHUCKLES approach is able to treat compounds as both sequence and chemical structure entities.<sup>10</sup> However, CHUCKLES does not specifically address the performance problems inherited in registering large or even midsize proteins like FVIIa in a chemistry database.

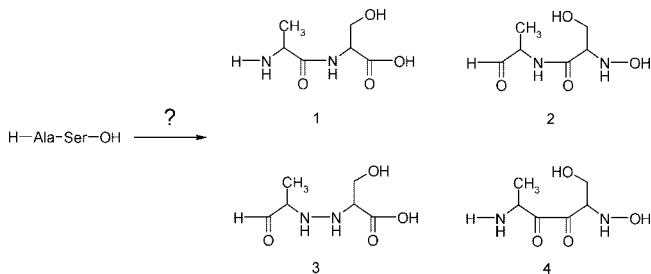
#### TRIMMING THE CHEMISTRY FORMAT

To solve the registration and search database performance issues we needed to find a more efficient way of storing

macromolecules in our MDL-based chemistry database.<sup>11</sup> We focused on representing redundant information in a simpler manner, in particular the recurring residue structures of the protein chains. The concept of “pseudoatoms” to represent residue structures already exists in the “extended periodic table” of MDL’s ISIS/Draw. The basic idea is that each residue structure can be represented by a single “residue atom” with two single bonds, representing the peptide bonds, as shown in Figure 3.

This method yields a very efficient representation of residues, on average reducing both the number of atoms and the number of bonds by approximately a factor 7 (note that hydrogen atoms are not registered in either format, because they are inherent to the structures). Perhaps more importantly, the residue atom method reduces the complexity of the atom-bond network graph considerably, thus speeding up molecule registration immensely.

However, the naive use of residue atoms carries the risk of losing chemical information. Take for instance the two-residue peptide in Figure 4 which may represent any of the four full atomic structures shown (not considering L- and D-isomers). The problem is that the residue atom does not contain any information about which peptide bond is attached to the residue’s N-terminal and which one to its C-terminal.



**Figure 4.** Representing residue structures with single residue atoms introduces structural ambiguity. The given sequence H-Ala-Ser-OH represents by convention only structure 1, but if in fact either of structures 2–4 were the structure on hand, there would be no way to represent it unambiguously by a residue-atom-based structure.

A potential solution to the orientation problem could be to attach annotations to the peptide bonds indicating their N- or C-terminal affiliation. An example of this is the “MDL condensed representation” implemented by MDL in their chemistry drawing tool MDL Draw.<sup>12</sup> However, these kind of annotations (also known as S-group data in MDL terminology) tend to get lost when the molecules are run through various third-party tools, whereas ordinary bonds and atoms are always preserved.

Instead of annotating the bonds we chose to define strict rules for the use of residue atoms which would ensure that the correct chemistry can be inferred by simply examining the immediate chemical surroundings of the residue atoms. Thus any chemistry package able to read the plain atom and bond information will be able to extract, and preserve, the full information represented by the residue atoms.

First we constrained every residue atom to have exactly two single bonds. Our practical implementation treats cysteine slightly different, both for its frequent participation in disulfide bonds and for layout reasons. Cysteine-rules are explained later as they have no direct bearing on the general principle. Next we constrained any chain of residue atoms within a molecule to be terminated by certain atoms only:

- \* The N-terminal peptide bond of a residue atom chain must connect to an “H” or a “C” atom.

- \* The C-terminal peptide bond must connect to an “O” or “N” atom.

This means that terminal residue atoms may bind only to

- (a) standard free peptide terminal atoms or

- (b) another residue represented by a full chemical structure, e.g. part of a protein backbone.

Figure 5 shows examples of how expansion to full atomic detail works on the basis of these rules.

## LIMITATIONS OF RESIDUE ATOM REPRESENTATION

Pure cyclic peptides cannot be drawn using residue atoms alone. Consider a peptide with the sequence GIVE (Gly-Ile-Val-Glu) shown in Figure 6(a). The representation does not provide any clue as to whether you should read the cycle clockwise or counterclockwise to follow peptide bonds. The solution is to represent one or more of the residues using a full structure, e.g. the Gly shown in Figure 6(b). Any arbitrary residue may be chosen.

## THE CYS EXCEPTION

Residue atoms must have exactly two bonds—the peptide bonds. One exception is however the cysteine residue atom

Cys. Cysteine very often participates in disulfide bridges, and to keep the representation compact we would like to use residue atoms for those cases also. In order to be able to infer orientation without ambiguity we found that Cys should represent not a full cysteine structure but rather the cysteine minus the sulfur atom. Both layout and representation issues support this model.

Consider a peptide with the sequence MCASCQ, having the two cysteines bound in a disulfide bridge. In the case where the Cys residue atom includes the sulfur it may be drawn like Figure 7(a). The drawing is neither aesthetically pleasing nor is it chemically precise since the disulfide bond might just as well be a peptide bond and the Ala and Ser residue would then attach to the cysteine side chains. Furthermore, free cysteines would need an explicit hydrogen added in order for molecular weights to be calculated correctly by ordinary tools. This would obscure the layout and introduce ambiguity in the representation in certain cases, as in Figure 7 (b), where you cannot tell which of the explicit hydrogens is bound to the cysteine nitrogen and which to the sulfur.

The solution we chose instead represents cysteine as a combination of a Cys residue atom and an externally added sulfur atom, as shown in Figure 8(a). The peptide from Figure 7(a) is shown in this representation mode in Figure 8(b) which gives an acceptable layout and has the added advantage of behaving correctly chemistry-wise if disulfide bonds are added or deleted directly in a normal molecule drawing tool. The rules for determining bond orientation of a Cys atom are still kept simple: It must have exactly three single bonds, one of which must go to a sulfur atom; the other two bonds represent the peptide bonds as normal.

## IMPLEMENTING RESIDUE ATOM SUPPORT

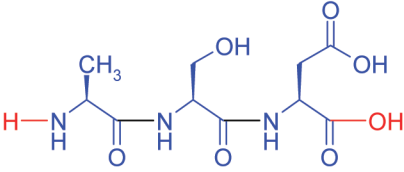
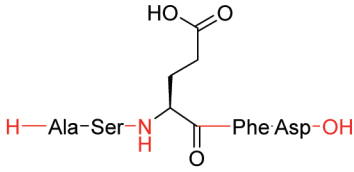
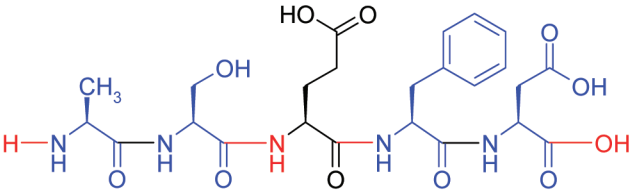
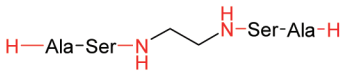
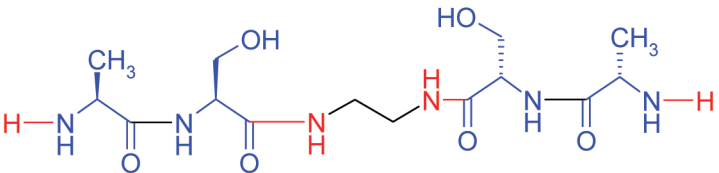
To make residue atoms work in real life, we needed some way to extend our standard chemistry drawing tool, ISIS/Draw, so it would understand the rules for using residue atoms properly. The first practical solution to this was a plug-in for ISIS/Draw. The plug-in was able to do the expansions shown in Figure 5 and also import proteins from files in UNIPROT format.

On the server side we successfully managed to add the residue atom usage rules to the chemistry database. The database can then check the rules upon registration of a molecule ensuring that no ambiguous structures get stored.

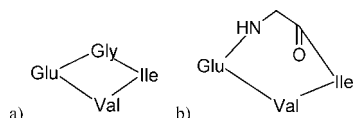
However, we found it difficult to integrate the ISIS/Draw plug-in satisfactorily into the normal workflow of the application, e.g. real-time enforcement of the “two-single-bonds-per-residue-atom” rule and the rules for residue atom orientation. Also, midsize to large proteins were problematic to work with display-wise and we found it hard to incorporate a high-level sequence view of the protein, which would be useful for interfacing to standard bioinformatics tools. A better solution was required, and we started searching for sequence editors with the ability to also handle chemical modifications.

## THE NOVO NORDISK PROTEIN EDITOR

Several commercial sequence editors were investigated, but none had the sequence-structure duality we were looking for. We therefore decided to build a dedicated protein editor.

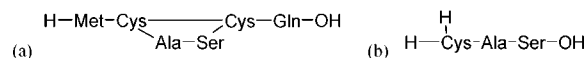
Input peptide	Corresponding full structure
H—Ala-Ser-Asp—OH	
 <p>[In this case, the person drawing this 5-peptide structure has decided to draw the glutamic acid as a full structure. The terminal atoms of the glutamic acid orient its neighboring residue atoms.]</p>	
 <p>[Two peptide chains linked by a simple structure. The nitrogens orient the serines.]</p>	
<p>H—Gly-Ile—H</p> <p>[Gly-Ile aldehyde]</p>	<p>This will produce an error like:</p> <p>Cannot infer direction of residue atom chain "Gly-Ile".</p> <p>In this case one of the residues must be drawn as a full structure.</p>

**Figure 5.** Examples of expanding residue atoms into full residue structures using the described rules for orienting the peptide bond directions. Terminal atoms and peptide bonds at the end of residue atom chains are highlighted in red; structures resulting from residue atom expansion are shown in blue.



**Figure 6.** (a) Attempting to represent cyclic peptide with residue atoms alone will not work. The structure can be read both clockwise and counterclockwise yielding two different structures. (b) Drawing one of the residues as a full structure, in this case the Gly, forces a fixed orientation.

The editor would use the UNIPROT (then SWISSPROT) format as its primary format but with format extensions to allow for the specification of exact chemical modifications wherever needed. The advantages of building gently on a standard protein format were obvious: We could use our registered proteins directly in standard bioinformatics tools—e.g. search them with BioWisdom's SRS system<sup>13</sup> or run them through standard BLAST searches<sup>14</sup> and various



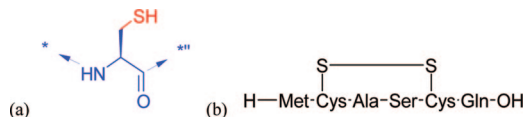
**Figure 7.** Problems arising when using Cys residue atoms that include the sulfur atom of cysteine. (a) A peptide with the sequence MCASCQ having a disulfide bridge between the two cysteines. You cannot tell whether the Cys-Cys bond represents a disulfide bridge or a peptide bond. (b) Another example of ambiguous chemistry. Here you cannot tell which explicit hydrogen is bound to the cysteine nitrogen and which is bound to the sulfur.

analysis and prediction tools, e.g. GPMW,<sup>15</sup> BioPerl,<sup>16</sup> and SciTegic's Pipeline Pilot.<sup>17</sup>

#### SIDE CHAIN AND TERMINAL MODIFICATIONS

Figure 9 shows how a chemist can specify a chemical modification in our protein editor via simple structural drawing.





**Figure 8.** (a) We chose to represent cysteine with an external sulfur atom (red) and the rest of the cysteine residue structure represented by the Cys residue atom (blue). (b) The peptide from Figure 7(a) can now be represented unambiguously and is aesthetically acceptable. Furthermore, disulfide bridges can be added or deleted in a normal drawing tool without producing incorrect chemistry (implicit hydrogens on the sulfur atoms are automatically taken care of).

Besides explicitly drawing a structure the chemist may also choose from a list of standard side chain modifications (e.g., acetyl, carboxyl, etc.). The list of standard modifications allows us to easily convert between the controlled vocabulary of modification names in UNIPROT format and our extended format.

The modification structure attachment point is specified as an atom number referring to fixed atom numbers within a set of in-house reference molecules that we use to represent the 20 natural amino acid residues. If the modification attaches to the N-terminal or the C-terminal of the residue the attachment point is simply specified as “N-terminal” or “C-terminal” instead of an atom number. A more suitable and general numbering would be to follow the locant numbering recommended by IUPAC.<sup>18</sup>

### CROSS-LINKS

Cross-links other than disulfides are often used in synthetic peptides. Such cross-links are modeled as modification structures with two or more R-group atoms as shown in Figure 10.

Further details on the extended UNIPROT format are available in the Supporting Information.

### CONVERTING SEQUENCE FORMAT TO CHEMISTRY FORMAT

Since the exact chemistry of all modifications is known, a full atomic level representation of the protein can be generated directly. When using residue atoms to represent all unmodified residues the chemical representation also becomes very compact and highly efficient. This enables us to register even large proteins, >3000 residues, in both protein format and chemistry format simultaneously without performance issues.

### CONVERTING CHEMISTRY FORMAT TO SEQUENCE FORMAT

To let the Protein Editor work with proteins already registered as full chemical structures and to allow bioinformatics tools access to the sequence information in these structures, we implemented a conversion algorithm dubbed “mol2uniprot”. This algorithm takes an arbitrary MDL molfile as input and builds the corresponding representation in extended UNIPROT format.

A basic outline of the algorithm is as follows:

- (1) Find residue substructures.
- (2) Gather residue chains.
- (3) Perceive modifications.

In step 1 residues are mapped via ordinary substructure searches. It is important to execute the mappings in a fixed

order since many residues are substructures of other residues. E.g. mapping glycines first would be a problem, because all nonglycine residues would then appear in the output as glycines with side chain extensions.

The mapped molecule is then traversed in step 2 to find all adjacent residues bound by peptide bonds and gather those into chains. Side chain and cross-link modification structures may contain residue substructures that are not part of a linear chain. An example of this is the side chain modification of Lys-26 in the GLP-1 analogue liraglutide, Lys26Nε-hexadecanoyl-γ-glutamyl Arg34 GLP-1(7-37), Figure 11.<sup>19,20</sup>

A minimum chain length constraint of 2 is therefore enforced to reduce spurious “false chains” in the output. If these very short “false chains” are detected the contained residue structures are marked up and the algorithm reruns step 1, but now the marked up structures are excluded from the substructure mappings. It is necessary to rerun step 1 since substructures within a “false chain” may overlap/mask parts of a residue in a genuine chain and thereby cause the mapping to change.

Finally, the remaining molecule fragments that are non-residues are isolated in step 3—these must be chemical modifications. Fragments bound to a single residue are side chain modifications, and fragments bound to several residues are cross-links.

Step 1 was implemented as MDL Cheshire<sup>21</sup> scripts and steps 2 and 3 were implemented in Object Pascal (Borland Pascal/Delphi).<sup>22</sup> An example conversion is shown in Figure 12, where the Protein Editor, which includes the mol2uniprot algorithm, converts the detemir insulin into its corresponding extended UNIPROT format. Note how both chemical modifications and disulfide bridges are preserved by the conversion.

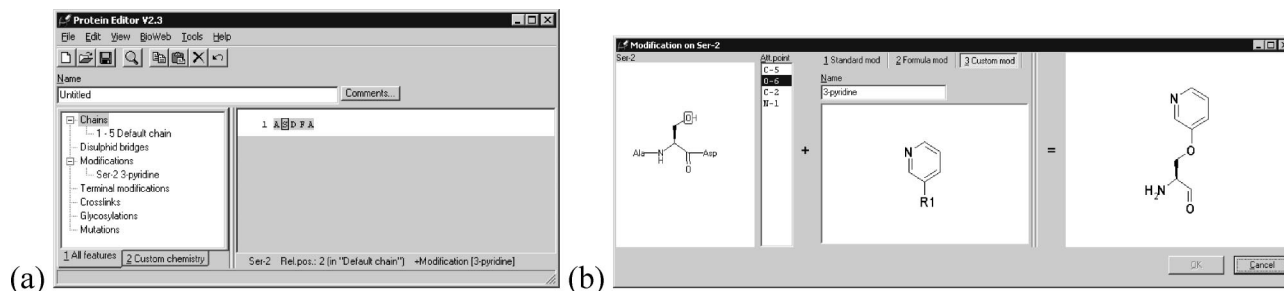
The algorithm has been able to convert all protein-related molfile registrations in our corporate compound database to extended UNIPROT format without chemical information loss (several thousand compounds). The conversion is fully automatic without the need for manual intervention or curation.

### REGISTRATION IN A BIOCHEMFORMATICS DATABASE

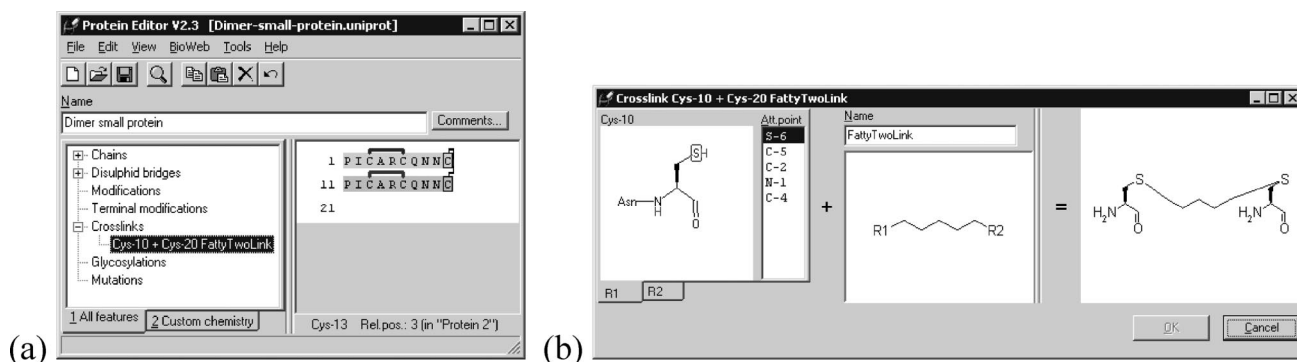
With all these tools in place the database is ready to accept macromolecules in both chemistry and protein format and also provide them in either format as requested. The way compound registration works is illustrated in Figure 13.

Protein chemists working with the protein editor register their molecules in extended UNIPROT format, Figure 13(a). The UNIPROT format version is stored in a suitable protein database schema. Upon registration the protein’s sequence is extracted and stored for future searching, and the protein’s corresponding chemistry representation is calculated on-the-fly and registered in the chemistry database.

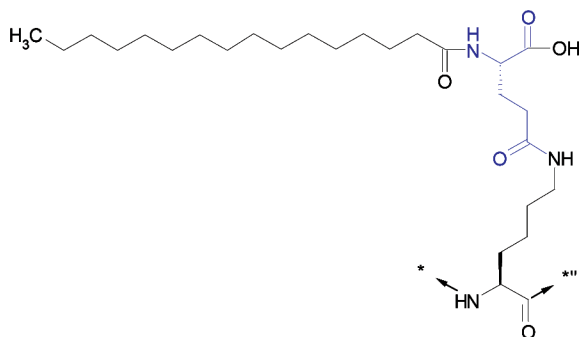
Medicinal chemists preferring the use of small molecule chemistry tools register molecules as usual in the chemistry database, perhaps as a biosequence drawn in MDL ISIS/Draw, Figure 13(b). Molecules above a minimum molecular weight are run through mol2uniprot. If mol2uniprot produces protein output the protein representation of the molecule is registered in the protein database schema, making the sequence searchable.



**Figure 9.** An example of how to modify a residue within a sequence using the Novo Nordisk Protein Editor. (a) Double-click highlighted serine to go to the modification editor shown in (b). (b) Draw modification molecule (in the middle) and attach it to the serine by selecting the reactive atom/the attachment point.



**Figure 10.** How a chemist may construct an arbitrary cross-link in the Novo Nordisk Protein Editor. (a) The residues that should participate in the cross-link are highlighted. The user then invokes the cross-link editor shown in (b). (b) A modification structure is drawn in the middle field, and each residue's reactive site (an atom is highlighted) is linked to one R-group label of the modification structure.



**Figure 11.** Side chain modification occurring in liraglutide. If the mol2uniprot algorithm did not have a minimum chain length constraint, the highlighted Glu residue would be represented in the output protein as a one-residue chain reacting via a cross-link to the main protein chain. Instead, the Glu residue is represented as normal full-structure chemistry.

All molecules with sequence content are thus accessible in both chemistry and sequence format as needed.

## SEARCHING A MIXED-REPRESENTATION DATABASE

The condensed chemical representation improves registration and retrieval performance dramatically but presents a challenge query-wise. Since any modified residue will need to be stored as a full structure, the database will contain molecules where any given residue may be represented by a residue atom in one molecule and by a full residue structure in another. Thus the peptide MVASQ with and without an acetylation on the serine will be stored as the molecular structures shown in Figure 14(a). This influences searching both with regards to molecular identity as well as substructure searching.

## MOLECULAR IDENTITY

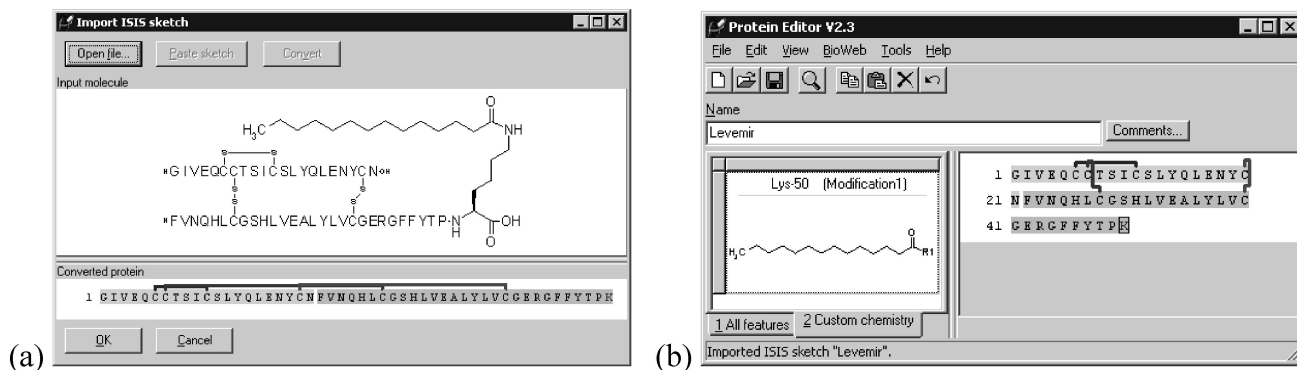
The mixed representation may cause two structurally identical peptides to fail an exact match since you are allowed to freely mix full-structure residues and residue atoms. E.g. the cyclic peptide in Figure 6 could have any one, or any number, of its residues drawn as full structure. Trying to force a full-structure exact match is not practical for large proteins for performance reasons.

The approach we have taken for determining protein structural identity is to first compare protein chain sequences via their protein representations. When we find full sequence identity we iterate through the chemical modifications and compare these (usually very small structures) by a normal chemistry exact match.

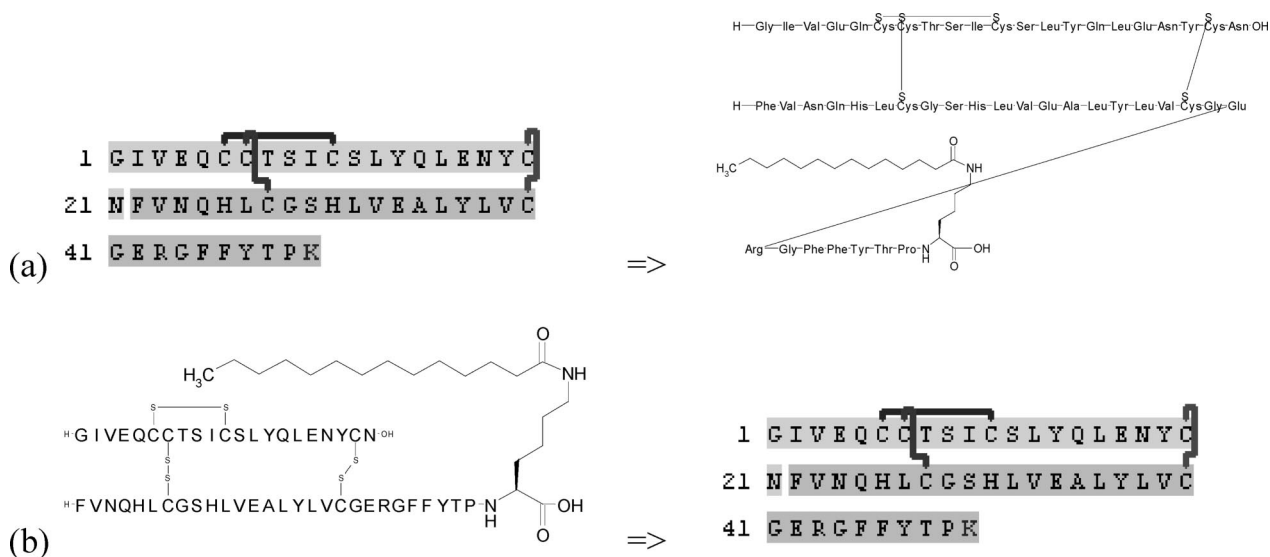
In order to determine sequence identity the protein chain sequences must first be normalized. A fully cyclic chain is normalized by shifting its sequence until it has the lowest lexical sort order possible. E.g. "EFGACD" will be shifted to "ACDEFG". If the cyclic chain is repetitive, e.g. "EFGACDEFGACD", modification structures must be taken into account (see the next paragraph on lexically identical chains) to decide which subsequence part should be listed first. Next, all chains of the peptide are ordered by a simple lexical sort.

Lexically identical chains must be further sorted based on their modifications, if any, to ensure correct later comparison of individual modification structures. Currently we resort to a lexical sort of the sum formulas of the modification structures. Disulfide bridges are currently not a part of the identity check, and this will eventually cause false positives.

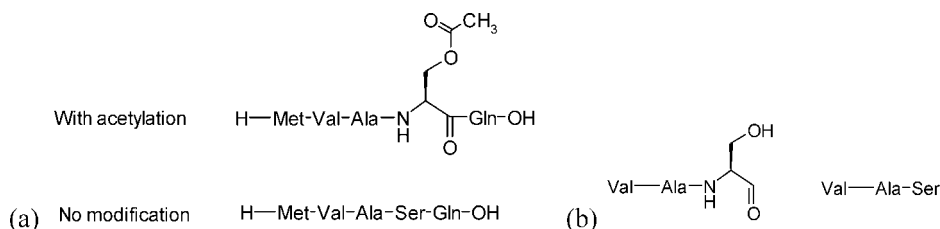
The sequence identity check will fail if two protein entries with differing sequences represent the same chemistry. This can happen if a chemist accidentally registers a side chain



**Figure 12.** Conversion of MDL molfile input (a) to extended UNIPROT format (b). The conversion retains chemical modifications (here on Lys-50) and cross-links (here three disulfide bridges).



**Figure 13.** Parallel registration of small molecule and protein representations of the same compound. (a) A protein chemist registers an extended UNIPROT entry in the protein database. The protein database generates a chemical representation of the full protein and registers that in the chemistry database. (b) A medicinal chemist registers an ISIS/Draw sketch in the chemistry database. A background job examines the new structure and calculates the corresponding UNIPROT entry and registers that in the protein database.



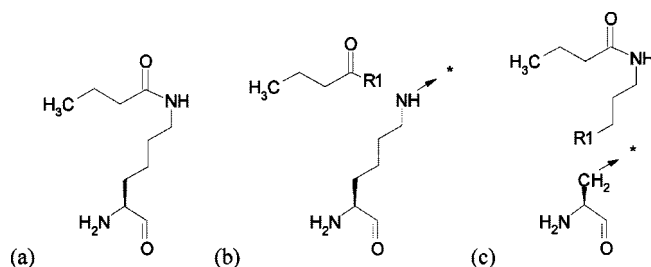
**Figure 14.** (a) Two different valid ways of representing a serine residue in a mixed-representation database. The structure on top has an acetylated serine, and therefore the serine is stored as a full structure. If the serine is unmodified, as in the lower structure, it will be stored as a single residue atom. (b) Neither of the two substructure queries shown will find both the registered structures from (a).

modification which is partly a normal residue side chain as shown in Figure 15.

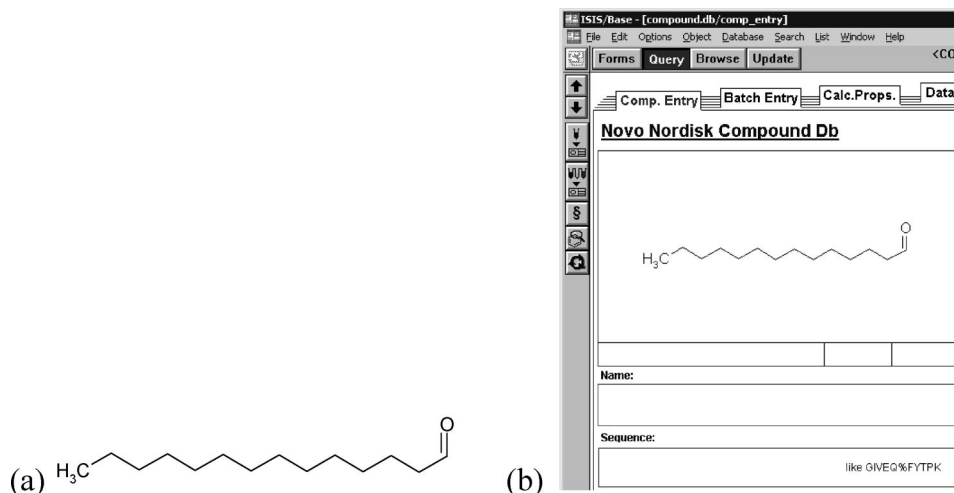
To avoid this problem, the protein representations being compared should always be based on the output of mol2uniprot. This will ensure that structurally identical residues are always resolved to the same residue code.

### SUBSTRUCTURE SEARCHING

Providing a fully generic substructure search capability against a mixed-representation database is a major task. It would be roughly equivalent to implementing a Markush substructure search,<sup>23,24</sup> although it would seem that you could take advantage of the severe connectivity constraints



**Figure 15.** (a) A modified lysine can be represented as (b) a lysine with a small fatty acid attached or (c) an alanine with a lysine side chain plus the fatty acid attached. The protein representation's sequence will then contain differing residue codes: (b) will be modeled as a modified "K" and (c) will be a modified "A".



**Figure 16.** (a) Simply searching for detemir-like insulins using this query finds several thousand hits in the Novo Nordisk compound database. (b) Combining the chemical substructure with a sequence query constrains the result set to a manageable few hundred hits.

Drag a column header here to group by that column				
NNC number	Mod. position	Mod. AA	Modification	Receptor affinity (%)
0100-0000-0304	50	Lys	<chem>CCCCCCCCCCCCCCCC=O</chem>	52
0100-0000-0377	50	Lys	<chem>CCCCCCCCCCCCCCCC(O)CCCCCCCCCCCCCCCC=O</chem>	41
0100-0000-0409	50	Lys	<chem>CCCCCCCCCCCCCCCC(O)CCCCCCCCCCCCCCCC=O</chem>	88.4
0100-0000-0413	50	Lys	<chem>CCCCCCCCCCCCCCCC(O)CCCCCCCCCCCCCCCC=O</chem>	43.7

**Figure 17.** In the Novo Nordisk compound database chemical modifications have been extracted and made available as separate structures. This aids visualization of protein SAR tables since only the chemical differences between compounds are shown instead of the full (large) compounds.

imposed on residue atoms to simplify the implementation. However, in light of the actual query needs we decided to take a more pragmatic approach.

The most common type of protein substructure search involves looking for side chain chemical details in proteins having certain sequence motifs. A real-life example is given in Figure 16 where we would like to find insulins having chemical modifications similar to that of insulin detemir (Figure 1).

Thanks to the parallel registration of compound representations this type of query can be executed by combining a sequence text query with a chemistry substructure query. The text query runs against the protein database's sequences and the substructure query against the chemistry database. The mol2uniprot algorithm could be used to assist end-users in transforming full structure sequence chemistry queries into separate sequence and chemistry queries. With this same approach the search problem in Figure 14 is resolved to searching the protein database for sequences containing the text "VAS".

If the end-user decides to search for multiple chain fragments in one query, the fragments will need to be run separately against the sequence database as you cannot know the order in which the fragments may occur in the target sequence. Found hits will need to be filtered to remove overlapping hits.

Substructure-like sequence queries are possible by utilizing the substructure relationships between the natural amino acids: Gly is a substructure of all other amino acids and will therefore match any residue, Phe is a substructure of Tyr, and so on. A plain text sequence query can therefore be transformed to a substructure sequence search via regular expressions.<sup>25</sup> E.g. to search for proteins that have the substructure "GFY" (Gly-Phe-Tyr) within a chain you could run the regular expression query "[[:upper:]]FY]Y" against the protein chain sequences.

#### SUBSTRUCTURE QUERY LIMITATIONS WITHIN A MIXED-REPRESENTATION DATABASE

Searching for chemistry details within a single modified residue structure or a cross-linked residue works well as all



modified residues are guaranteed to be stored as full structures. However, if the chemistry query spans multiple residues by including peptide bonds the search will most likely fail as you will not know whether neighboring residues are stored as residue atoms or full structures, so some insight into the registration is required. Longer sequence chemistry queries can be readily detected and converted into separate chemistry and sequence queries, but a more troublesome query would be a one- or two-residue sequence chemistry query: The minimal chain length constraint enforced by mol2uniprot makes it unclear whether the target structure will have been mapped into a sequence text or not.

On the other hand, the protein representation enables query types that would be very hard to express in pure chemistry terms, e.g. "find all insulins having a chemical modification in position B30 where the side chain structure contains a C-6 fatty acid". The parallel registration also enables the execution of queries that cannot be performed efficiently in a chemistry-only database, e.g. searching for long sequence motifs.

### PROTEIN CHEMISTRY SAR

Proteins and peptides in SAR tables present a challenge display-wise due to their sheer size. One useful approach is suggested by Vielmetter et al.<sup>26</sup> where point mutations and sequence variations are extracted and emphasized. In addition to this our protein representation allows extraction and display of chemical variations.

In the example SAR table shown in Figure 17 mol2uniprot has been used to extract the side chain modifications of the original full structure registrations of insulins. The presentation of the SAR is thus much clearer than if the whole insulin molecule was shown and differences in modification structures are easier to discern. Compare e.g. with Table 1 in Jonassen et al.<sup>27</sup>

### FUTURE DIRECTIONS

The extended UNIPROT protein format has already demonstrated its utility in several biopharmaceuticals projects in-house as a primary registration format. It has also proven useful for extracting new knowledge from and enabling better searching of existing chemical structures, e.g. our range of insulin and GLP-1 analogues.

Future work includes something as mundane (but not trivial) as residue numbering within and across protein entries. Better facilities for creating whole compounds or modifications from simple text entries are being investigated, e.g. creating glycane tree molecules on-the-fly from a textual glycane notation. Also, novel predictor tools combining the sequence and chemistry data to yield more precise descriptions of biopharmaceutical compounds are being looked into.

The current representation of modification and linker structures only allows for strictly additive chemistry. The only displacement that occurs is deprotonation which is dealt with via implicit hydrogens. Reactive substitution is not possible without first mutating the residue to e.g. glycine and then rebuilding the remaining residue structure on top of that. This could be an area of improvement.

The database also lacks good tools to handle structures where the sites of modification are not well characterized, like structures where a glycan modification may be known,

but the site of modification is uncertain or the attachment is in fact mixed between several sites.

### ACKNOWLEDGMENT

Jan H. Jensen thanks Steen Aagaard Sørensen, Scientific Computing, Novo Nordisk, for discussions on sequence normalization and identity.

**Supporting Information Available:** Novo Nordisk A/S UNIPROT extensions—appendix 2 of the Protein Editor manual. This material is available free of charge via the Internet at <http://pubs.acs.org>.

### REFERENCES AND NOTES

- (1) Veronese, F. M.; Pasut, G. PEGylation, successful approach to drug delivery. *Drug Discovery Today* **2005**, *10*, 1451–1458.
- (2) Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A. Extent of Modifications in Human Proteome Samples and Their Effect on Dynamic Range of Analysis in Shotgun Proteomics. *Mol. Cell. Proteomics* **2006**, *5*, 2384–2391.
- (3) Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B. H. Synthesis of proteins by native chemical ligation. *Science* **1994**, *266*, 776–779.
- (4) Muir, T. W.; Sondhi, D.; Cole, P. A. Expressed protein ligation: A general method for protein engineering. *Proc. Nat. Acad. Sci. U.S.A.* **1998**, *95*, 6705–6710.
- (5) Markussen, J.; Havelund, S.; Kurtzhals, P.; Ersen, A. S.; Halstrom, J.; Hasselager, E.; Larsen, U. D.; Ribøl, U.; Schaffer, L.; Vad, K.; Jonassen, I. Soluble, fatty acid acylated insulins bind to albumin and show protracted action in pigs. *Diabetologia* **1996**, *39*, 281–288.
- (6) Goldman-Levine, J. D.; Lee, K. W. Insulin Detemir-A New Basal Insulin Analog. *Ann. Pharmacother.* **2005**, *39*, 502–507.
- (7) UniProt Knowledgebase user manual. <http://www.expasy.org/sprot/userman.html> (accessed August 24, 2008).
- (8) MDL molfile format specification. <http://www.mdll.com/downloads/public/ctfile/ctfile.jsp> (accessed August 24, 2008).
- (9) Condensed chemical representation. [http://www.biochemfusion.com/doc/condensed\\_representation.html](http://www.biochemfusion.com/doc/condensed_representation.html) (accessed August 27, 2008).
- (10) Siani, M. A.; Weininger, D.; Blaney, J. M. CHUCKLES: A Method for Representing and Searching Peptide and Peptoid Sequences on Both Monomer and Atomic Levels. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 588–593.
- (11) MDL Direct. [http://www.mdll.com/products/framework/rel\\_chemistry\\_server/index.jsp](http://www.mdll.com/products/framework/rel_chemistry_server/index.jsp) (accessed August 24, 2008).
- (12) MDL Draw. [http://www.mdll.com/products/framework/mdl\\_draw/index.jsp](http://www.mdll.com/products/framework/mdl_draw/index.jsp) (accessed August 24, 2008).
- (13) SRS. <http://www.biowisdom.com/navigation/srs/srs> (accessed August 24, 2008).
- (14) Altschul, Stephen, F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
- (15) GPMW. <http://www.gpmw.com/> (accessed August 24, 2008).
- (16) BioPerl. <http://www.bioperl.org/> (accessed August 24, 2008).
- (17) SciTegic data analysis and reporting platform. <http://accelrys.com/products/scitegic/> (accessed August 24, 2008).
- (18) IUPAC Nomenclature and Symbolism for Amino Acids and Peptides, section "3AA-2.2 Designation of locants". <http://www.chem.qmul.ac.uk/iupac/AminoAcid/AA1n2.html#AA22> (accessed August 24, 2008).
- (19) Elbrond, B.; Jakobsen, G.; Larsen, S.; Agerso, H.; Jensen, L. B.; Rolan, P.; Sturis, J.; Hatorp, V.; Zdravkovic, M. Pharmacokinetics, pharmacodynamics, safety, and tolerability of a single-dose of NN2211, a long-acting glucagon-like peptide 1 derivative, in healthy male subjects. *Diabetes Care* **2002**, *25*, 1398–1404.
- (20) Madsen, K.; Knudsen, L. B.; Agerso, H.; Nielsen, P. F.; Thøgersen, H.; Wilken, M.; Johansen, N. L. Structure-activity and protraction relationship of long-acting glucagon-like peptide-1 derivatives: Importance of fatty acid length, polarity, and bulkiness. *J. Med. Chem.* **2007**, *50*, 6126–6132.
- (21) MDL Cheshire. [http://www.mdll.com/products/framework/chemistry\\_rules/](http://www.mdll.com/products/framework/chemistry_rules/) (accessed August 24, 2008).
- (22) Delphi2007 for Win32. <http://www.codegear.com/products/delphi/win32> (accessed August 24, 2008).
- (23) Schoch-Grübler, U. (Sub)structure searches in databases containing generic chemical structure representations. *Online Rev.* **1990**, *14*, 95–108.

- (24) Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 16. The Refined Search: An Algorithm for Matching Components of Generic Chemical Structures at the Atom-Bond Level. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 1–7.
- (25) Regular expressions. [http://www.opengroup.org/onlinepubs/009695399/basedefs/xbd\\_chap09.html](http://www.opengroup.org/onlinepubs/009695399/basedefs/xbd_chap09.html) (accessed August 24, 2008).
- (26) Vielmetter, J.; Tishler, J.; Ary, M. L.; Cheung, P.; Bishop, R. Data management solutions for protein therapeutic research and development. *Drug Discovery Today, BioSilico* **2005**, 10, 1065–1071.
- (27) Jonassen, I.; Havelund, S.; Ribel, U.; Plum, A.; Loftager, M.; Hoeg-Jensen, T.; Volund, A.; Markussen, J. Biochemical and Physiological Properties of a Novel Series of Long-Acting Insulin Analogs Obtained by Acylation with Cholic Acid Derivatives. *Pharm. Res.* **2006**, 23, 49–55.

CI800128B