

Creating readable protein SAR tables through consistent naming

Jan Holst Jensen, Biochemfusion ApS, Denmark

jan@biochemfusion.com

*This is the commercial white paper version of
Biochemfusion's poster presentation at MipTec 2011.*

biochemfusion

Introduction

During the development of pharmaceutical compounds drug candidates are often compared in a Structure Activity Relationship table. When the drug candidate is a small molecule, display clutter may be minimized by removing recurring fragments through e.g. R-group decomposition.¹

The current industry trend is moving towards drugs based on proteins and peptides and so there is a need for similar methods to produce readable protein SAR tables.

Proteins are typically described via their sequences, or on an abstracted level by abbreviations or trivial names. When comparing closely related protein variants in a SAR table full sequences or full structures present too much detail to be useful.

Using trivial names and abbreviations will most likely result in inconsistent or ambiguous names, and they often won't give a sufficiently precise description of the molecule. If we however can calculate consistent and chemically precise protein names from the protein sequence the problem is in principle solved.

This paper will present Biochemfusion's notation for chemically precise naming of protein variants and examples of how protein SAR tables and compound registration benefit from it. The notation is partly based on the existing IUPAC recommendations² and has been named *DerNot* for **DER**ivatives **NOT**ation.³

Actual SAR table examples using DerNot notation are found from page 5 forward.

DerNot notation

Names that obey the rules of the DerNot notation are called DerNot expressions.

DerNot expressions can be used to express the difference between two similar proteins. Vice versa, you may apply a DerNot expression to a protein in order to generate a new derivative.

The DerNot notation is based on plain text with a consistent and simple syntax so it can be both generated and parsed by an algorithm of manageable complexity. The plain text format also ensures that the notation can be exchanged between researchers and diverse IT systems without corruption.

A DerNot expression lists the deletions, insertions, and substitutions that you need to apply to get from protein A to protein B.

If a protein variant is created from a protein with the trivial name "Ref" by deleting residue 3, inserting an alanine (Ala/A) after residue 5, and substituting residue 8 with arginine (Arg/R), the corresponding DerNot expression will be:

```
des-(3) endo-A(5) R(8) "Ref"
```

Post-translationally or chemically modified residues and terminals are handled through the same notational means as is used in Biochemfusion's Protein Line Notation (PLN).⁴ Changing residue 2 of Cyclosporin CsA to a norvaline residue, producing Cyclosporin CsG, may therefore be done via the following DerNot expression. Norvaline is known with the modification name "Nva".

```
[Nva](2) "Cyclosporin CsA"
```

Chain extension works by prepending or appending subsequences to a given chain. Sanofi®'s Insulin Glargine/Lantus® is created by a single in-chain glycine residue substitution plus extension of the insulin B-chain by two arginines. The DerNot expression for this compound is shown below.

```
G(A21) "Human insulin" -RR-(B)
```

When multiple proteins are lined up in a SAR table, all compared to a single reference protein, the calculated DerNot expressions will repeatedly contain the trivial name of the reference protein. To minimize redundancy and to make the presentation compact DerNot expressions may be abbreviated to anonymous form.

The anonymous form uses an asterisk as a placeholder for the the trivial name of the reference, e.g. the anonymous form of the above DerNot expression will be

```
G(A21) * -RR-(B)
```

A more detailed description of the DerNot notation can be found in the DerNot specification.³

Implementation challenges

Reading a DerNot expression and applying the resulting edit operations to a reference protein is a fairly straightforward operation. The reverse operation, comparing a protein variant against a reference and calculating the corresponding DerNot expression, involves a couple of alignment challenges as discussed below.

To calculate a DerNot expression an algorithm must align the individual chains of the two proteins and finally output the perceived minimal number of edit operations (insertions, deletions, substitutions, and chain extensions) to get from one protein to the other.

Chain swapping

Proteins may consist of multiple chains and the proteins being compared may have their chains listed in different order. All chain pair combinations must therefore be tested and the most similar chain pairs will be used as the basis for generating the final DerNot expression.

```

1  G I V E Q C C T S I C S L Y Q L E N Y C
21 N F V N Q H L C G S H L V E A L Y L V C
41 G E R G F F Y T P K T
  
```

Fig 1: Human Insulin, A chain listed first (green), then B chain (blue).

```

1  F V N Q H L C G S H L V E A L Y L V C G
21 E R G F F Y T P K G I V E Q C C T S I C
41 S L Y Q L E N Y C N
  
```

Fig 2: Human Insulin variant des-(B30) "Human Insulin". The insulin B chain is here listed first, followed by the A chain.

In the example above, comparing Fig 2 against Fig 1, the chain reversal should be ignored. The algorithm should be able to calculate a DerNot expression for the protein in Fig 2 that is des-(B30) "Human Insulin".

Cyclic chain rotation

When comparing cyclic chains you will have no guarantee that the reference and the variant agree on which residue is the "first" residue of the chain. The algorithm should therefore check whether the protein variant needs to be rotated to minimize the resulting DerNot expression.

```

1  T A G L V L A A L L V
  
```

Fig 3: Cyclosporin CsA.

```

1  V L A A L L V T A G L
  
```

Fig 4: A cyclosporin variant rotated 4 residues. This is Cyclosporin CsB, also known as: A(A2) "Cyclosporin CsA".

Comparing Fig 4 against Fig 3 the algorithm should be able to deduce that Fig 4 should be rotated 4 residues, producing A(A2) "Cyclosporin CsA".

SAR table examples

One application of the DerNot notation is to use calculated DerNot expressions to highlight structural differences in protein SAR tables.

In the example below, a number of successful insulin analogues have been registered in a database system together with associated data on their peak activity time Tmax after subcutaneous injection.⁵

The database has Biochemfusion's Proteax cartridge installed which can compare protein pairs and calculate corresponding DerNot expressions.

Cmp.No	Sequence	Name	DerNot diff.	Param.	Value	Unit
C00001		Human insulin *		Tmax	1 - 3	hrs
C00002		Lispro	endo-P(B29) des-P(B28) *	Tmax	0.5 - 1.5	hrs
C00003		Aspart	D(B28) *	Tmax	1 - 1.5	hrs
C00004		Glulisine	K(B3)E(B29) *	Tmax	1 - 1.6	hrs
C00005		Glargine	G(A21) *-RR-(B)	Tmax	20.5 - 26.3	hrs
C00006		Detemir	des-T(B30) [N6-C14fattyacid-lysine](B29) *	Tmax	6 - 8.1	hrs

Human Insulin (compound "C00001") has in this case been chosen as the reference compound that all other variants are compared against. The resulting calculated DerNot expression is listed in the "DerNot diff." column using the anonymous notation for maximum brevity.

The protein sequence has been shown graphically in the "Sequence" column to illustrate the inherent difficulty in using full sequences for visual comparison.

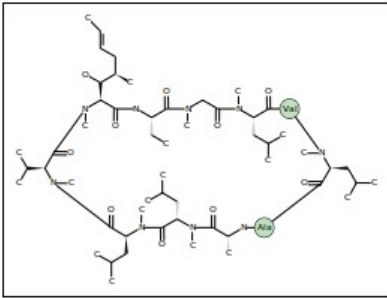
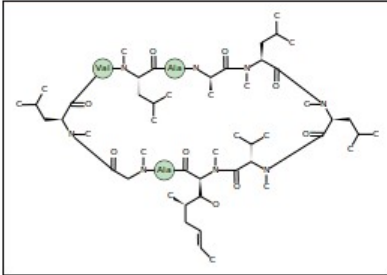
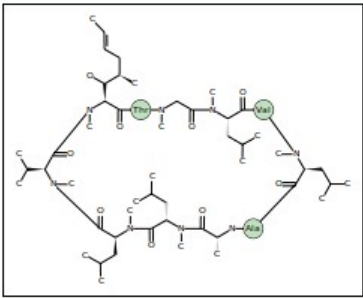
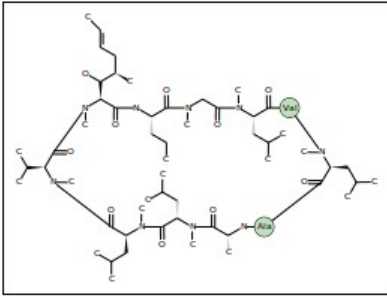
To illustrate the dynamic nature of the database system, you may choose another reference compound and re-run the SAR report. If we choose Insulin Aspart (compound "C00003") as the reference compound instead the SAR report will become the one below.

Cmp.No	Sequence	Name	DerNot diff.	Param.	Value	Unit
C00001	<pre> 1 GIVEQCCTSIICSLYQLENYC 21 NFNQHLGSHLVEALYLVC 41 GERGFFYYTPKT </pre>	Human insulin	P(B28) *	Tmax	1 - 3	hrs
C00002	<pre> 1 GIVEQCCTSIICSLYQLENYC 21 NFNQHLGSHLVEALYLVC 41 GERGFFYYTPKT </pre>	Lispro	endo-P(B29) des-D(B28) *	Tmax	0.5 - 1.5	hrs
C00003	<pre> 1 GIVEQCCTSIICSLYQLENYC 21 NFNQHLGSHLVEALYLVC 41 GERGFFYYTDKT </pre>	Aspart	*	Tmax	1 - 1.5	hrs
C00004	<pre> 1 GIVEQCCTSIICSLYQLENYC 21 NFNQHLGSHLVEALYLVC 41 GERGFFYYTPET </pre>	Glulisine	K(B3)PE(B28-29) *	Tmax	1 - 1.6	hrs
C00005	<pre> 1 GIVEQCCTSIICSLYQLENYC 21 GFVNQHLGSHLVEALYLVC 41 GERGFFYYTPKTRR </pre>	Glargine	G(A21)P(B28) * -RR-(B)	Tmax	20.5 - 26.3	hrs
C00006	<pre> 1 GIVEQCCTSIICSLYQLENYC 21 NFNQHLGSHLVEALYLVC 41 GERGFFYYTPK </pre>	Detemir	des-T(B30) P[N6-C14fattyacid-lysine](B28-29) *	Tmax	6 - 8.1	hrs

Another example is a comparison of cyclosporin toxicity, shown below (hypothetical LD50 values).⁶

In this example the user has chosen to show the molecular structure in the "Molecule" column instead of the sequence, as the cyclosporins contain many post-translational modifications. The molecular structure is calculated on-the-fly by the Proteax Cartridge.

Cyclosporin CsB has been deliberately rotated 4 residues clock-wise to demonstrate that the calculated DerNot expressions are rotation-invariant. The calculated DerNot expressions will always use the numbering as defined by the reference compound, here "Cyclosporin CsA", residue 1 being the residue at the top left of the structure.

Cmp.No	Molecule	Name	DerNot diff. Param.	Value	Unit
C00007		Cyclosporin CsA *	LD50	1400 - 1510	mg/kg
C00008		Cyclosporin CsB A(2) *	LD50	1600 - 1820	mg/kg
C00009		Cyclosporin CsC T(2) *	LD50	800 - 1200	mg/kg
C00012		Cyclosporin CsG [Nva](2) *	LD50	2500 - 3005	mg/kg

Easing compound registration

DerNot expressions can also be used to simplify registration of protein derivatives.

Instead of entering a full protein sequence a researcher can start from a known reference protein and apply a DerNot expression to that. An underlying database system can then register the generated full sequence and molecule.

In the example shown here, a researcher is using Cyclosporin CsA as a reference compound. Once the reference compound is selected, its sequence and chemical structure is automatically loaded in Biochemfusion's protein editor - the component outlined with a green box.

The screenshot displays the Biochemfusion protein editor interface. At the top, a table lists reference compounds, with 'Cyclosporins' (ID: C00007) selected. Below this, the 'Project' field is set to 'Cyclosporins' and the 'Ref. compound' field is set to 'C00007'. A 'DerNot expression' field is empty, and an 'Apply to ref.' button is present. The main editor area is outlined in green and contains the following sections:

- Protein text - PLN format:** A text box containing the DerNot expression: `(cyclo)-[MeBmt][Abu][MeGly][MeLeu]V[MeLeu]A{d}A[MeLeu][MeLeu][MeVal]-(cyclo)` with the name `name="Cyclosporin CsA"`.
- Amino acid selection:** A grid of buttons for selecting amino acids: Ala, Arg, Asn, Asp, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val, Sec, Pyl, Xaa.
- Preview:** A section showing the protein sequence `1 TAGLVLAALLV` and the chemical structure of Cyclosporin CsA. The sum formula is `C62 H111 N11 O12` and the average molecular weight is `1202.61124`.

Red arrows point to the 'Cyclosporins' entry in the table, the 'C00007' field, and the chemical structure, with the text 'Reference compound selected' and 'Reference compound automatically loaded in editor'.

The researcher may now enter a DerNot expression to modify the loaded reference compound. Once the expression has been applied the researcher will immediately see the resulting sequence and structure.

2	insulins	C00001	Human insulin
3	Cyclosporins	C00007	Cyclosporin CsA

Project Ref. compound

DerNot expression ← DerNot expression entered and applied to reference compound

Protein text - PLN format

(cyclo) - [MeBm] [Ala] [MeGly] [MeLeu] V [MeLeu] A {d} A [MeLeu] [MeLeu] [MeVal] - (cyclo)
name="A(A2) *"

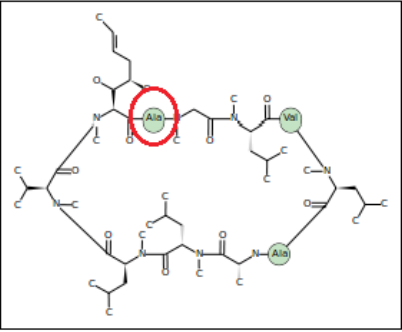
N-terminal C-terminal

Ala Arg Asn Asp Cys Glu Gln Gly His Ile Leu Lys
Met Phe Pro Ser Thr Trp Tyr Val Sec Pyl Xaa

Preview

1 T A E L V L A A L L V Sum formula: C61 H109 N11 O12
Avg. MW: 1188.58466

1:1



1:1 ← Resulting molecule updated on-the-fly

The benefits of using DerNot expressions for registration will naturally increase with the length of the protein sequence.

In short

DerNot notation defines a convenient protein naming scheme that is both human- and machine-readable. Together with Biochemfusion's Protein Line Notation that enables registration of chemically well-defined protein and peptide sequences you will have a new set of protein chemistry tools at your disposal. Tools that let you create readable protein SAR tables with consistent and precise names and a novel way to register protein variants by name.

The demonstration database system described in this paper is available on the web: http://www.proteax.dk/demo_db/ (requires a modern browser with HTML5 support)

References

- [1] R-group decomposition, see e.g.
<http://www.chemaxon.com/jchem/doc/user/RGroupDecomposition.html>
http://accelrys.com/products/pdf/rgroup_decomposition_reprint.pdf
- [2] IUPAC "Nomenclature and Symbolism for Amino Acids and Peptides"
<http://www.chem.qmul.ac.uk/iupac/AminoAcid/AA1n2.html#AA1>
- [3] Biochemfusion DerNot 1.0 specification
http://www.biochemfusion.com/doc/Biochemfusion_DerNot_1.0_spec.pdf
- [4] Biochemfusion PLN 1.1 specification
http://www.biochemfusion.com/doc/Biochemfusion_PLN_1.1_spec.pdf
- [5] Tmax values of insulin analogues approximated from data available at
<http://www.drugs.com/ppa/insulin-analogs.html>
- [6] Cyclosporin CsA LD50 values approximated from a material data safety sheet; remaining LD50 values generated at random.

Insulin Glulisine/Apidra[®] and Insulin Glargine/Lantus[®] was developed by Sanofi S.A. Sanofi[®], Apidra[®], and Lantus[®] are registered trademarks of Sanofi S.A.

<http://www.sanofi.com>

Insulin Aspart/NovoLog[®] and Insulin Detemir/Levemir[®] was developed by Novo Nordisk A/S. Novo Nordisk[®], NovoLog[®], and Levemir[®] are registered trademarks of Novo Nordisk A/S.

<http://www.novonordisk.com>

Insulin Lispro/Humalog[®] was developed by Eli Lilly and Company. Lilly[®] and Humalog[®] are registered trademarks of Eli Lilly and Company.

<http://www.lilly.com>

Biochemfusion[®] and Proteax[®] are registered trademarks of Biochemfusion ApS.

<http://www.biochemfusion.com>