Protein Registration

- issues in the structural registration of protein entries

Jan Holst Jensen
CEO and Founder, Biochemfusion

- Enabling biochemformatics

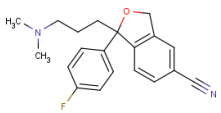
Bridging the gap

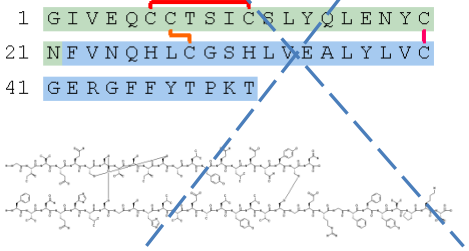
Cheminformatics

Bioinformatics

neither-nor

Molecule graphs





Sequences

100 10k 1M MW Da

Protein representation
Levels of abstraction

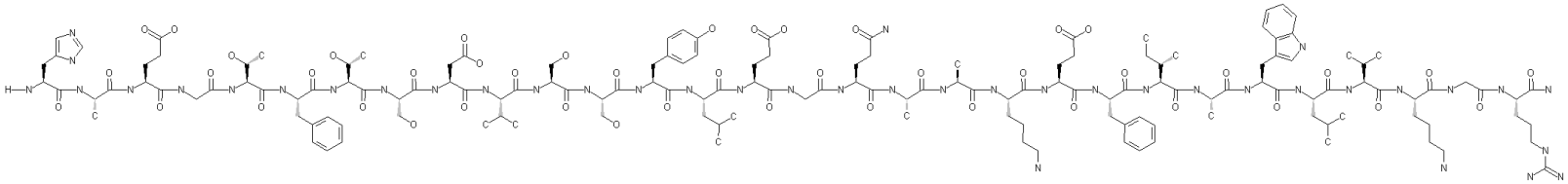
- Name

GLP-1

- Sequence

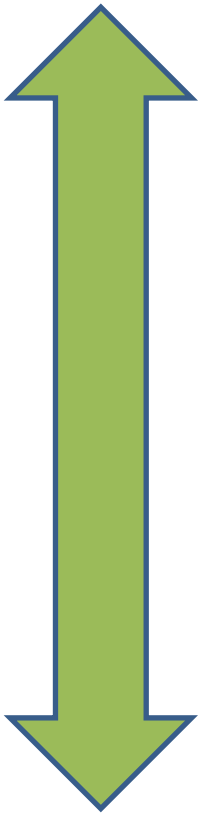
H-HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR-[NH2]

- Atoms and bonds



In the ideal world

- Name
- Sequence
- Atoms and bonds



Free movement up and down the abstraction tree without losing information

CHUCKLES (Daylight monomer toolkit)

CHUCKLES: A method for representing and searching peptide and peptoid sequences on both monomer and atomic levels
Michael A. Siani, David Weininger, Jeffrey M. Blaney
J. Chem. Inf. Comput. Sci., **1994**, 34 (3)

http://pubs.acs.org/doi/abs/10.1021/ci00019a017

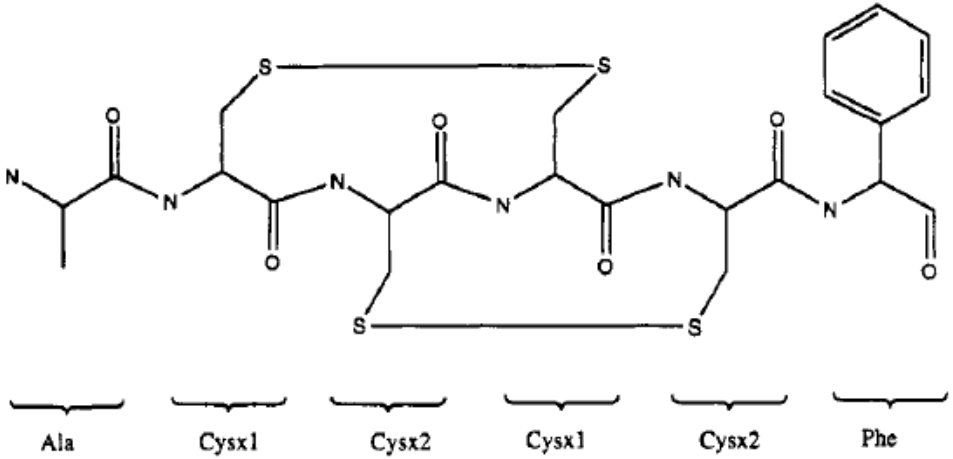
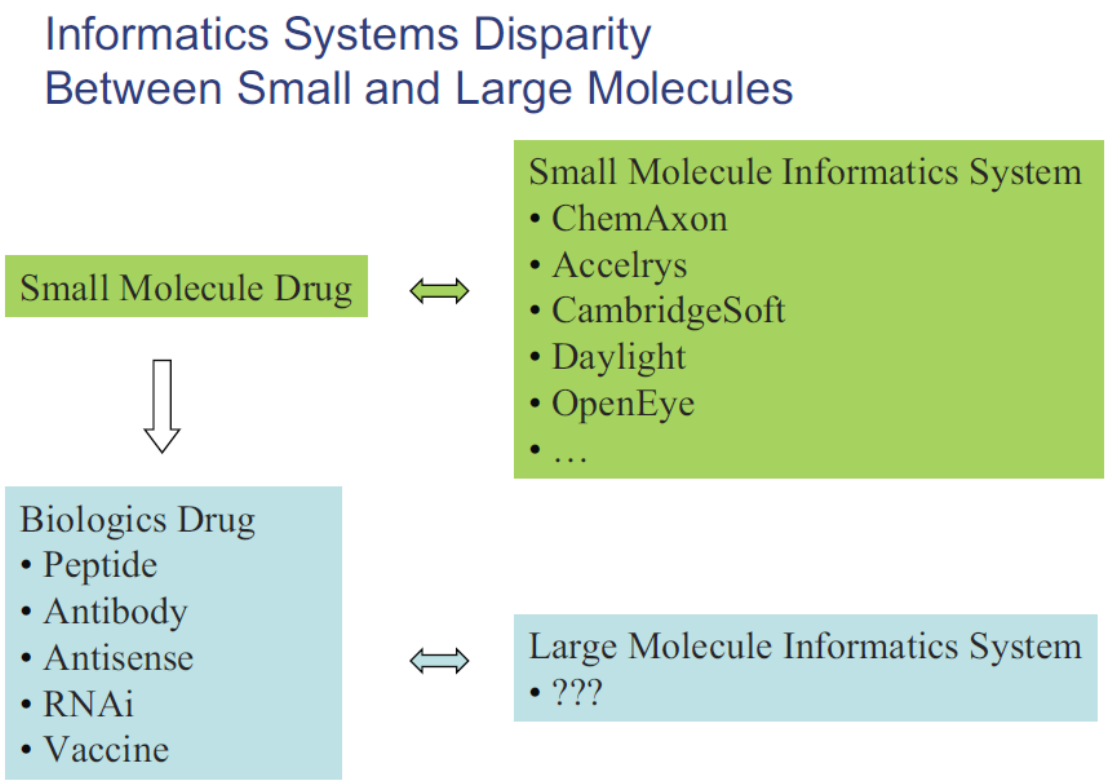


Figure 8. Molecular graph representation of user-input peptide, containing two disulfides. Monomer matches indicated by names are shown below relevant atoms and bonds in graph.

Pfizer Inc.

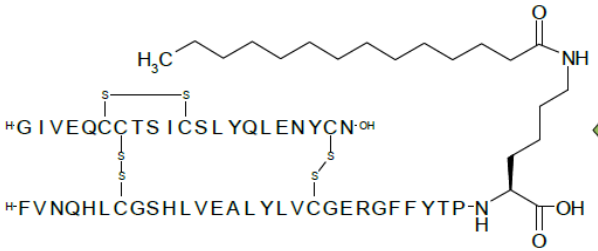
Macromolecular Structure Notation, Editing and Registration

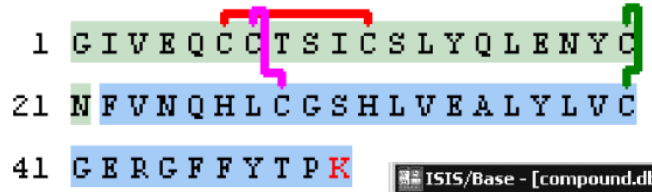
Tianhong Zhang *ChemAxon US UGM 2009*

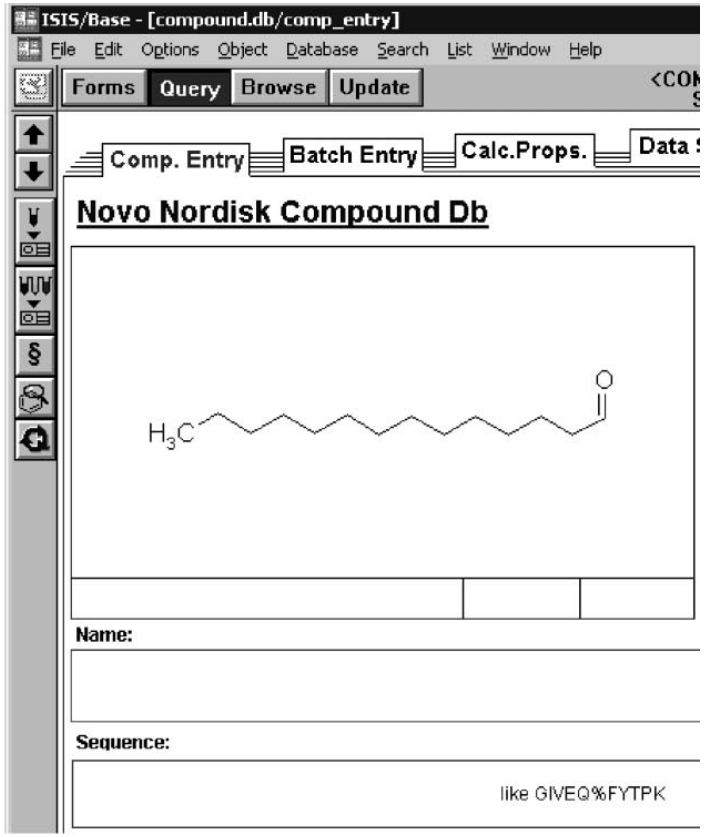


http://www.chemaxon.com/library/ugm-presentations/2009-us/#wednesday-16th

Novo Nordisk A/S







Building a BioChemformatics Database
Jan H. Jensen, Thomas Hoeg-Jensen, and Søren B. Padkjær
J. Chem. Inf. Model., **2008**, 48 (12)

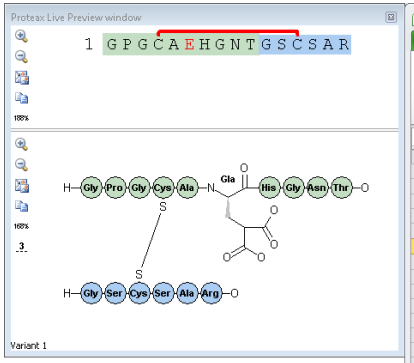
http://pubs.acs.org/doi/abs/10.1021/ci800128b

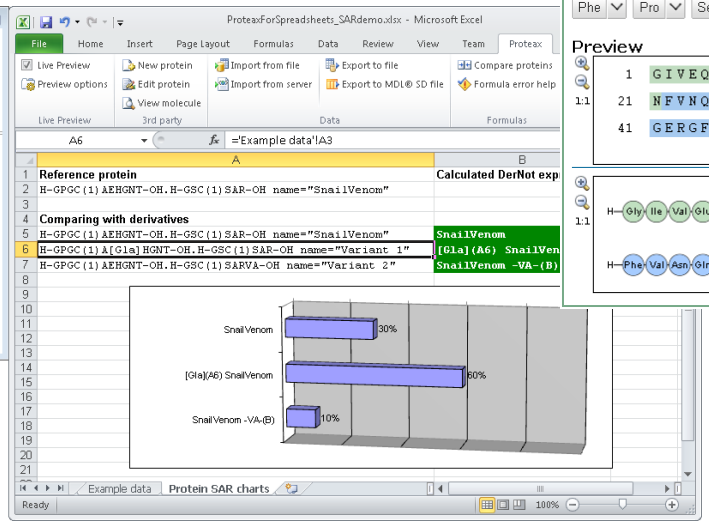
Biochemfusion ApS

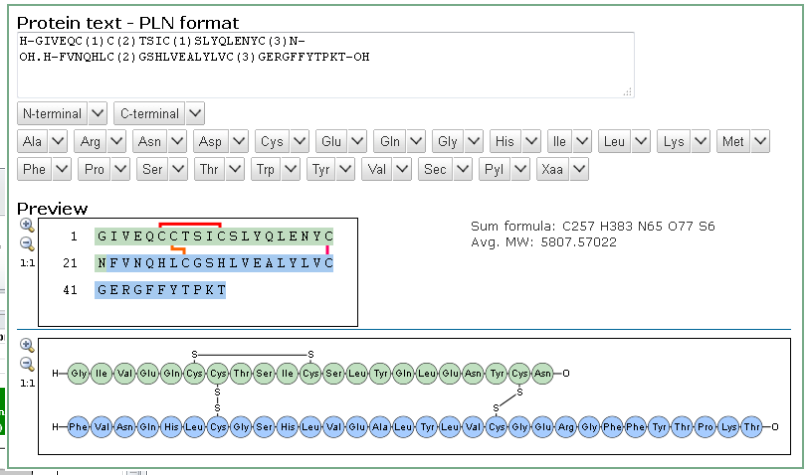
Proteax software suite

- Built for parallel registration
- Handles post-translational and chemical modifications
- IUPAC-like protein naming

http://www.biochemfusion.com/







Parallel registration works

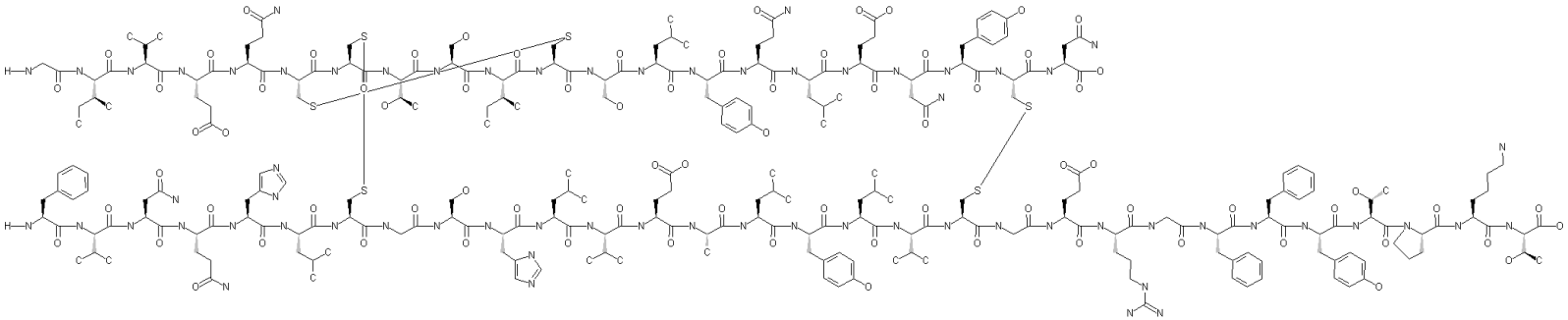
- You don’t have to choose molecule-only or sequence-only
- “Just” do both, and you will have covered the lower two abstraction levels

General issues

- Performance
  - Scaling chemical graphs to the mega-Dalton range
- Uniqueness
  - Normalizing input
- Naming
  - Normalizing and aligning – SAR table production

Performance

CTAB explosion



Human insulin
51 residues => 407 atoms and 418 bonds

Average sequence length of protein in UniProt is greater than 200 residues

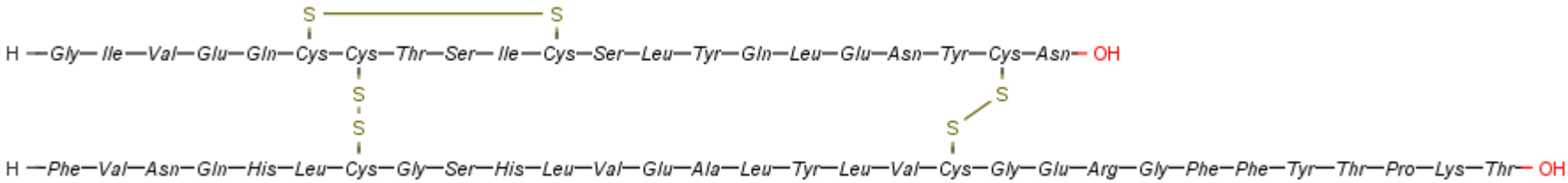
FVIIa (blood clotting factor)	466 residues
Human antibodies	~1200-1500 residues in total
FVIII (blood clotting factor)	2351 residues

The case for condensed structures

- Full structures OK up to 50-100 residues
  - Registration and search performance in normal chemistry databases then declines sharply
- Condensed structures
  - Unmodified residues represented as single atoms – smaller and less complex connection table
  - Suitable for large proteins with a small percentage of modified residues
    - Infrequent modifications seems to be the normal case, although we may become wiser in future; see e.g.

      Extent of Modifications in Human Proteome Samples and Their Effect on Dynamic Range of Analysis in Shotgun Proteomics.
      Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A.
      Mol. Cell. Proteomics **2006**, 5, 2384–2391.

Condensed version

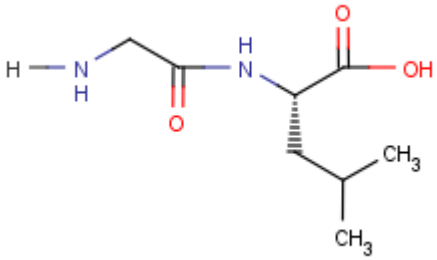


51 residues => 61 atoms and 62 bonds

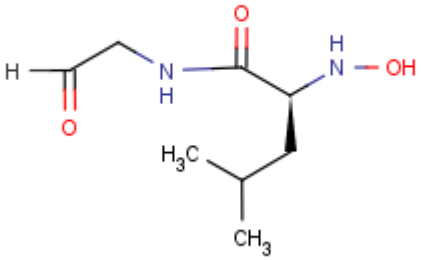
Using the 20 pseudo atoms *Ala*, *Arg*, *Asn* ... *Tyr*, *Val*

Much smaller – but... which orientation ?

H —*Gly*—*Leu*— OH =



or



?

or ... [2 more]

Orientation solutions

- Alpha atoms at pseudo-atom chain terminals define orientation (used at Novo Nordisk)
  - H or C = N-terminal/left peptide bond
  - O or N = C-terminal/right peptide bond
  - Plain connection table and simple rules
- S-group annotations define left and right peptide bonds (used by Accelrys)
  - Requires correct interpretation and handling to survive 3rd party tools

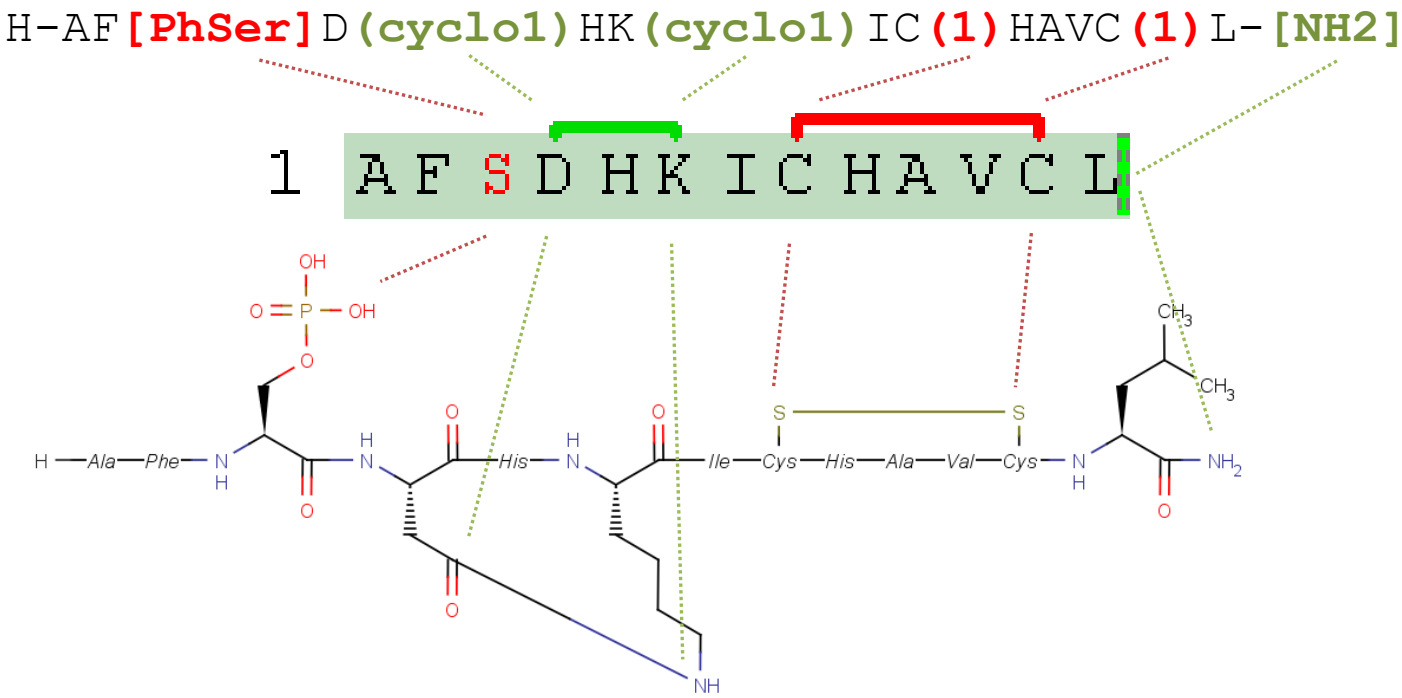
Mixed representation challenges

- Modified residues are stored as full structure
  - Including terminally modified residues
- Chemistry searches that involve backbone features are challenging
  - Are neighboring residues represented as full structure or as pseudo atoms ?
- Are you really trying to search for a subsequence instead ?
  - Combined substructure + sequence searches are easy to do when parallel registration is used

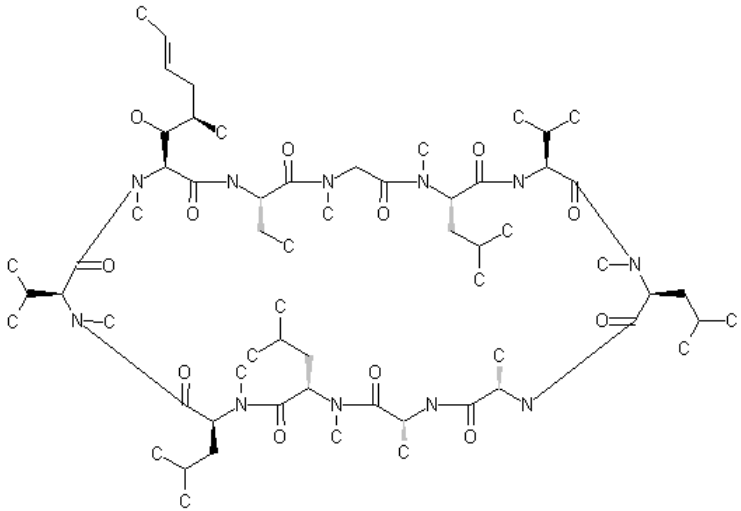
Uniqueness

Protein Line Notation (PLN)

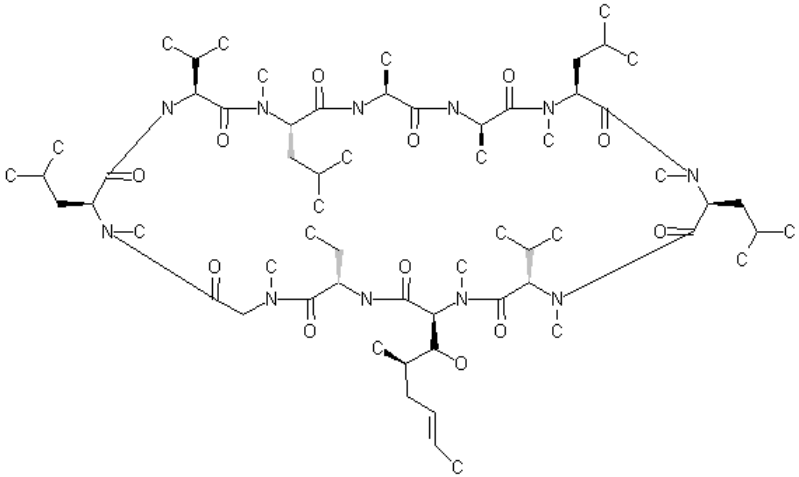
- Protein “SMILES” – sequence with chemical annotations



Pop-quiz – spot the difference



= ?



(cyclo)-[MeBmt]-[Abu]-[MeGly]-[MeLeu]-Val-[MeLeu]-Ala-dAla-[MeLeu]-[MeLeu]-[MeVal]-(cyclo)

(cyclo)-Val-[MeLeu]-Ala-dAla-[MeLeu]-[MeLeu]-[MeVal]-[MeBmt]-[Abu]-[MeGly]-[MeLeu]-(cyclo)

Yes, in fact identical

Normalizing linear chains

- Order chains in lexical sort order

H-GIVEQC(1)C(2)TSIC(1)SLY-OH.
H-QLENYC(3)N-OH.
H-FVNQHLC(2)GSHLVEALYLVC(3)GERGFFYTPKT-OH

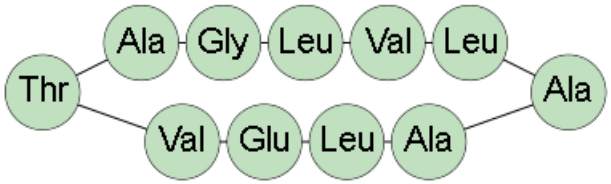
H-FVNQHLC(1)GSHLVEALYLVC(2)GERGFFYTPKT-OH.
H-GIVEQC(3)C(1)TSIC(3)SLY-OH.
H-QLENYC(2)N-OH

- Modified residues and terminals map to unique names

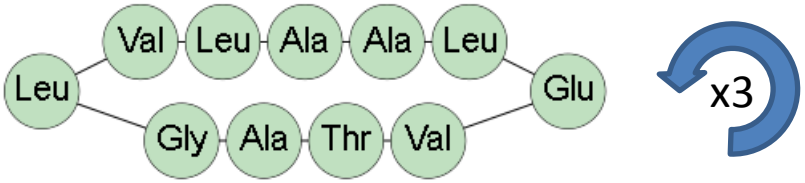
Normalizing cyclic chains

- Rotate until lowest lexical sort order

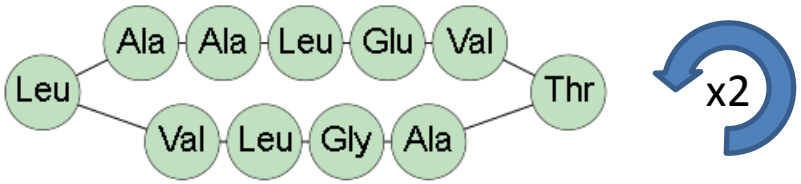
(cyclo)-AGLVLAALEVT-(cyclo)



(cyclo)-VLAALEVTAGL-(cyclo)



(cyclo)-AALEVTAGLVL-(cyclo)



Efficient indexing / lookup

- MD5 checksum of normalized PLN
- 32-character hexadecimal string
  - readily indexed by e.g. Oracle

```
proteax_norm_protein(
  H-GIVEQC(1)C(2)TSIC(1)SLYQLENYC(3)N-OH.
  H-FVNQHLC(2)GSHLVEALYLVC(3)GERGFFYTPKT-OH name="Human insulin"
) = H-FVNQHLC(1)GSHLVEALYLVC(2)GERGFFYTPKT-OH.H-GIVEQC(3)C(1)TSIC(3)SLYQLENYC(2)N-OH
```

```
proteax_norm_protein_chksum(
  H-GIVEQC(1)C(2)TSIC(1)SLYQLENYC(3)N-OH.
  H-FVNQHLC(2)GSHLVEALYLVC(3)GERGFFYTPKT-OH name="Human insulin"
) = d507f1997d3716ae974b17d0b4f08b04
```

Naming

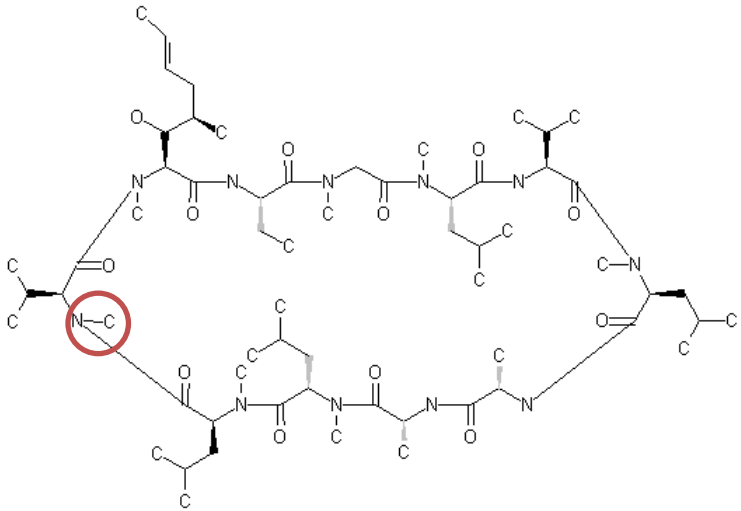
DerNot expressions

- **Der**ivatives **Not**ation
  - IUPAC-like protein naming

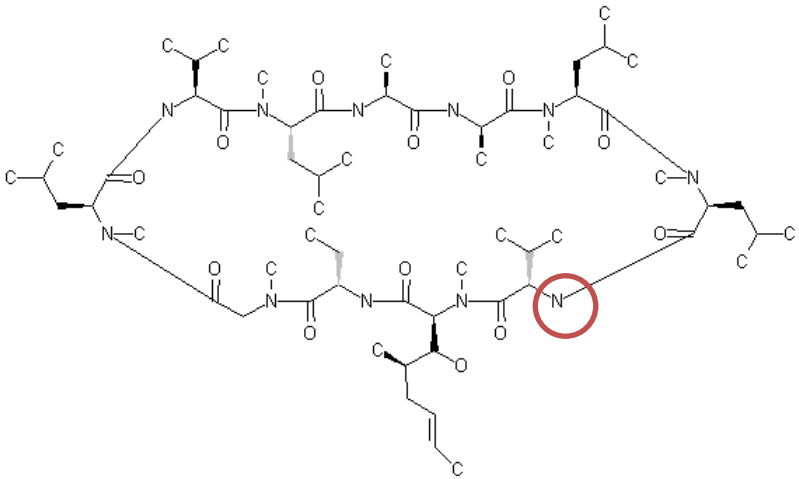
```
des-(B30) endo-R(B28) K(A4)A(B2) "Human Insulin"
```

deletions insertions substitutions

Pop-quiz #2 – spot the difference



= ?



(cyclo)-[MeBmt]-[Abu]-[MeGly]-[MeLeu]-Val-[MeLeu]-Ala-dAla-[MeLeu]-[MeLeu]-[MeVal]-(cyclo) name="Cyclosporin CsA"

(cyclo)-Val-[MeLeu]-Ala-dAla-[MeLeu]-[MeLeu]-Val-[MeBmt]-[Abu]-[MeGly]-[MeLeu]- (cyclo)

Proteax_DerNot_Diff(<right>, <left>) => V(11) "Cyclosporin CsA"

Protein SAR table

- Human insulin analouges

Lispro	H-GIVEQC(1)C(2)TSIC(1)SLYQLENYC(3)N-OH.H-FVNQHLC(2)GSHLVEALYLVC(3)GERGFFYTKPT-OH
Aspart	H-GIVEQC(1)C(2)TSIC(1)SLYQLENYC(3)N-OH.H-FVNQHLC(2)GSHLVEALYLVC(3)GERGFFYTDKT-OH
Glulisine	H-GIVEQC(1)C(2)TSIC(1)SLYQLENYC(3)N-OH.H-FVKQHLC(2)GSHLVEALYLVC(3)GERGFFYTPET-OH
Glargine	H-GIVEQC(1)C(2)TSIC(1)SLYQLENYC(3)G-OH.H-FVNQHLC(2)GSHLVEALYLVC(3)GERGFFYTPKTRR-OH

- Calculate names and add T_{max} data

Lispro	K(B28)P(B29) "Human insulin"	0.5 - 1.5
Aspart	D(B28) "Human insulin"	1 - 1.5
Glulisine	K(B3)E(B29) "Human insulin"	1 - 1.6
Glargine	G(A21) "Human insulin" -RR-(B)	~24

Summary

- Parallel registration works
  - Multiple representations enable traversal of the abstraction tree
- Chemical graphs *can* scale to the mega-Da range through condensed representations
- Using a consistent protein notation enables
  - Fast uniqueness checks
  - Protein naming and SAR table generation

Acknowledgements

- Thanks to former colleagues at Novo Nordisk A/S
  - Thomas Høeg-Jensen
  - Søren B. Padkjær
  - Christian Vind
  - Steen Aagaard Sørensen

- Also thanks to
  - Gerd Blanke, StructurePendium GmbH
  - Thomas Dörner, Independent consultant
  - Tianhong Zhang, Pfizer Inc.
  - Numerous people at MDL/Symyx/Accelrys
  - Numerous people at ChemAxon

Warning: Commercials follow

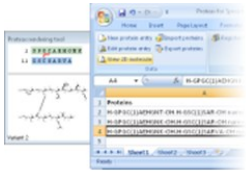
Proteax® – easily integrated

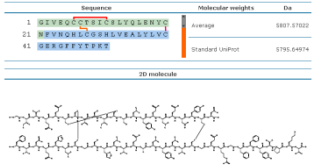
Proteax Desktop
.NET, Java®, Python, C/C++, Pascal

Scripting, programming

Proteax for Spreadsheets
Spreadsheet functions, macro programming

MS Excel®
OpenOffice.org®





Java® applet, Flash®-based, JavaScript™, .NET, and Pascal visualizers

Structure and sequence visualization

Proteax for KNIME
KNIME nodes for graphical processing of protein data

KNIME™ graphical workflows

Proteax toolkit

Proteax Cartridge
Database integration
registration and querying

Oracle®

```
select * from compounds
 where proteax.mw_avg(protein_text) > 5000;
```

PostgreSQL®

Try it

- Proteax for MS Excel and Oracle
  - http://www.biochemfusion.com/downloads/
- Works in
  - Microsoft Excel 2003, 2007, 2010
    - Both 32-bit and 64-bit Office versions
  - Oracle 10 and 11 on both Windows and Linux
- Proteax on the web
  - http://www.proteax.dk/proteax_live/