

# Structure Searching for Small Sequences in the CAS Registry File<sup>SM</sup>

**Lora Burgess**

**Senior Applications Specialist, CAS**

**PIUG 2008 Boston Biotechnology Meeting**

**February 14, 2008**

[www.cas.org](http://www.cas.org)



*A division of the American Chemical Society*

# CAS REGISTRY<sup>SM</sup> sequence content at a glance

## Sequences from

- Patents
- Journals
- GenBank<sup>®</sup> (all until 3/2005, then only those GenBank in documents indexed by CAS)

Many types of nucleotide and protein sequences, including

- Chemically-modified sequences
- Genetically engineered sequences
- Fusion protein sequences
- Protein Nucleic Acid (PNA) sequences
- Nucleic acid primers and probes
- Cyclic sequences

**CAS currently has referenced more than 540,000 chemically-modified sequences, and additionally > 800,000 referenced sequences from journals and > 11.5 million patent sequences that are not deposited in GenBank**

# What are the patent sources for sequence registrations?

- **Sequences from patents from 51 active patent authorities**

See <http://www.cas.org> for

- A complete list of active authorities
- Details on coverage by patent classification and document types by kind codes

Indexed sequences from 9 major patent offices within 27 days of issuance (CA, DE, FR, EP, GB, JP, RU, US, WO)

- **Sequences from the 'CAS Basic' – the first patent that CAS acquires in a patent family**
- **Sequences from continuations-in-part**

# What are the indexing criteria for patent sequences?

- **1907-1999: Novel sequences in patents**
- **Novel and all other sequences in patents beginning in October 1999**
- **GenBank sequences for nucleic acids and proteins**

## Selected sequences include:

- **Partial sequences**  
Examples:  
Enzyme functional site  
Gapped sequence for an isolated protein
- **Sequences with ambiguities, regardless of the number of unknown or uncertain residues**  
Example:  
**MAQQXXXXVPKFGXXXSDGNVPYTLFGNVPRKGKGGGAAG**
- **Sequences described in the text but not illustrated**

**In 2005, CAS stopped adding to CAS Registry any sequences from patents that contained more than approximately 4,000 sequences**

# What are the indexing criteria for journal sequences?

## *Emphasis is on scientific novelty!*

### Selected sequences include:

- **Newly reported GenBank nucleotide numbers, and**  
All GenBank numbers for translated proteins in the GenBank record Feature Table
- **Nucleotide sequences that are presented-in-document**
- **Primers and probes when considered important to point of paper**
- **Genetic elements (e.g., promoters) if functional information is provided**
- **Protein sequences that are found-in-document**
- **Partial sequences, including gapped sequences**

**In 2007, CAS stopped adding to CAS Registry GenBank sequences in journal articles that referenced more than 1,000 GenBank accession numbers.**

# Factors affecting naming and structuring of sequences

In general, the following holds true...

Sequence Type (length/residues)	Structure Representation	Sequence Codes & Annotation to Describe Atypical Structures	Names <sup>3</sup>
<b>Small PID<sup>1</sup></b> Nucleotides (1-8) Proteins (<~30-32)	Connection Table	No – for nucleotides Yes – for proteins >3	Structure-based
<b>Medium PID</b> Nucleotides (9-≤50) Proteins (~33≤50)	Sequence	Yes <sup>2</sup>	Structure-based
<b>Large PID</b> Nucleotides (>50) Proteins (>50)	Sequence	Yes <sup>2</sup>	Biological <sup>2</sup>
<b>Reported GenBanks</b> (Any Length)	Sequence	No	Biological GenBank number

<sup>1</sup> PID – Presented-In-Document

February 14, 2008 © American Chemical Society, 2008. May not be copied or reprinted without permission from CAS.

<sup>2</sup> Not the case for unclaimed patent sequences

<sup>3</sup> Patent name assigned if sequence is from a patent *A division of the American Chemical Society*



A division of the American Chemical Society

## What kinds of questions can be answered with a sequence structure search?

Has a sequence of interest been modified at a certain position?

Is this peptide known to bind metals?

Short sequences identified with X's in the sequence listing

What kinds of linkers have been used to build drug conjugates?

These are just a few examples of the kinds of questions that a structure query can help answer.

# Sequences with structures

**A searchable connection table structure in CAS REGISTRY has less than 253 non-hydrogen atoms**

**Sequences which fall below this threshold can be structure searched**

- Modified sequences
- Ambiguous sequences, e.g., XAXYX, where X is an uncommon or modified amino acid
- Metal binding peptides
- Organic molecules containing sequence moieties

Note: Not every small sequence has a structure.

# CAS REGISTRY<sup>SM</sup> sequence with modifications

RN 781640-44-6 REGISTRY  
CN Poly(oxy-1,2-ethanediyl), .alpha.-hydro-.omega.-methoxy-, diester with  
N2,N6-dicarboxy-L-lysylglycyl-L-leucyl-L-phenylalanylglycine  
5-[(4S)-4,11-diethyl-3,4,12,14-tetrahydro-3,14-dioxo-4-(1-oxopropoxy)-1H-  
pyrano[3',4':6,7]indolizino[1,2-b]quinolin-9-yl] ester (9CI) (CA INDEX  
NAME)  
FS PROTEIN SEQUENCE  
SQL 5  
NTE modified (modifications unspecified)

type	location	description
modification	Lys-1	undetermined modification

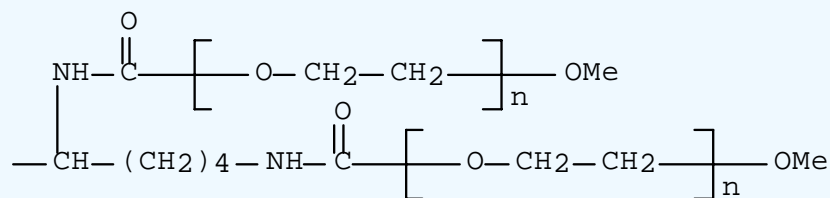
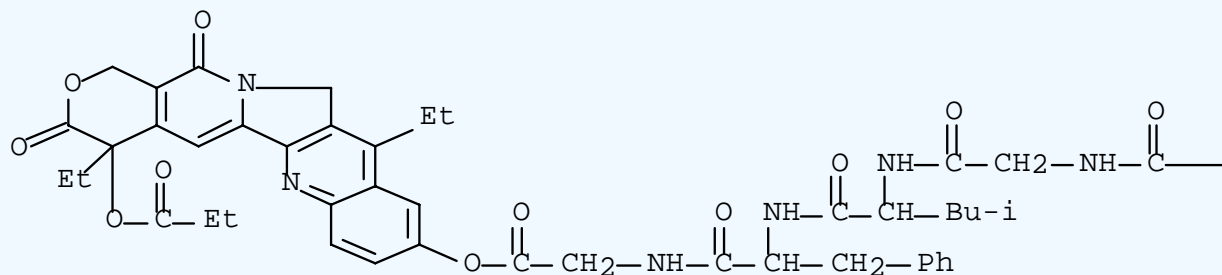
SEQ 1 KGLEFG

The name for this sequence is structure based. Modifications are noted at Lys-1, but not the type.

*(continued on next slide)*

# What is useful about a structure for a sequence?

PAGE 1-A



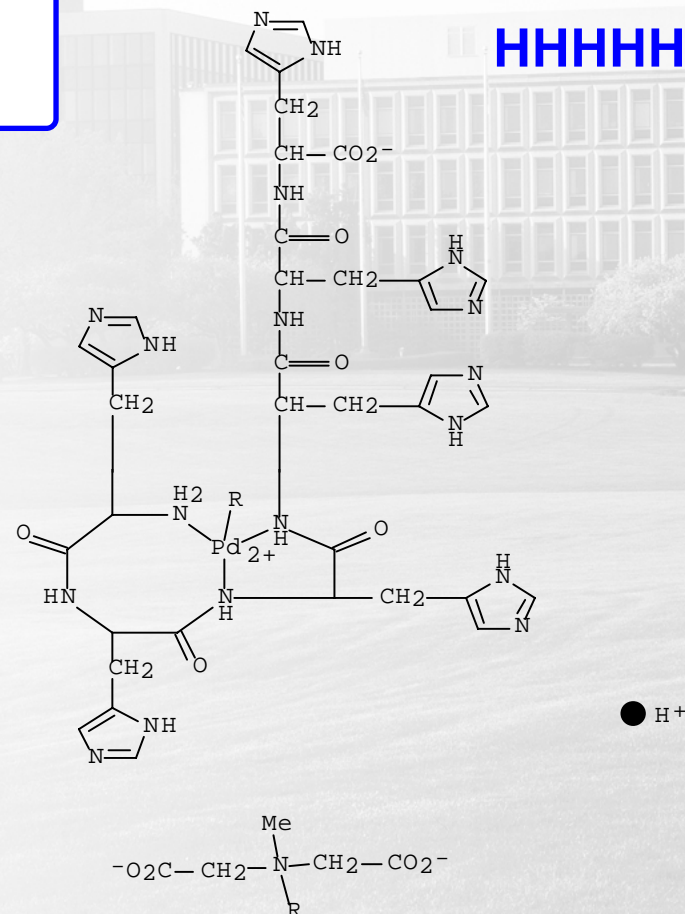
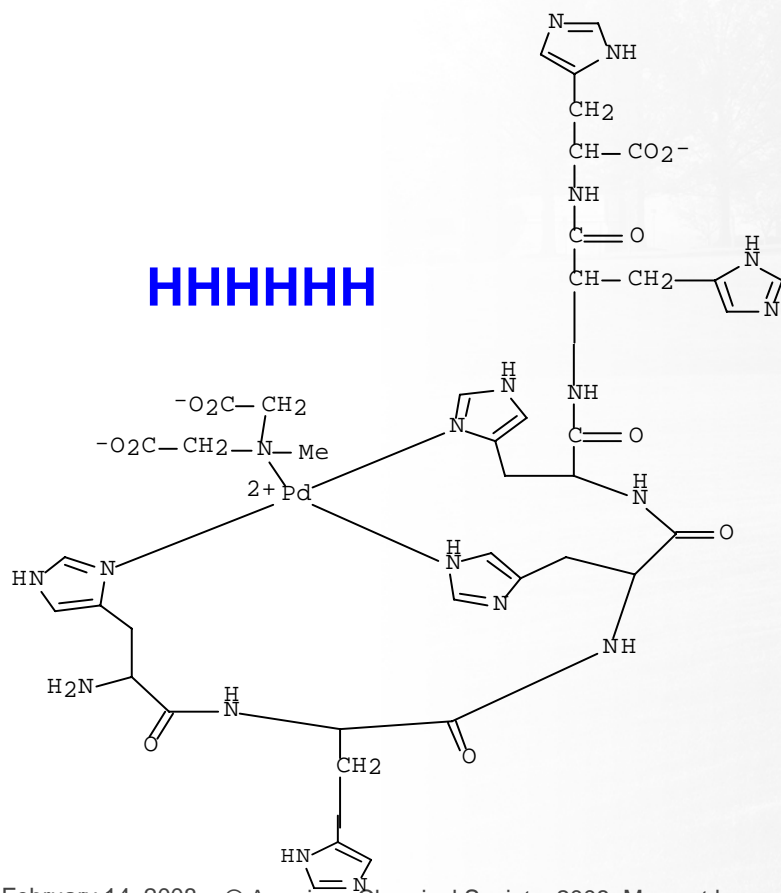
Modifications are clearly shown in the chemical structure representation. Other CAS RNs with the exact same sequence backbone (modified and non-modified) may be retrieved with the SEQLINK command.

**\*\*RELATED SEQUENCES AVAILABLE WITH SEQLINK\*\***

MF (C2 H4 O)<sub>n</sub> (C2 H4 O)<sub>n</sub> C54 H66 N8 O15

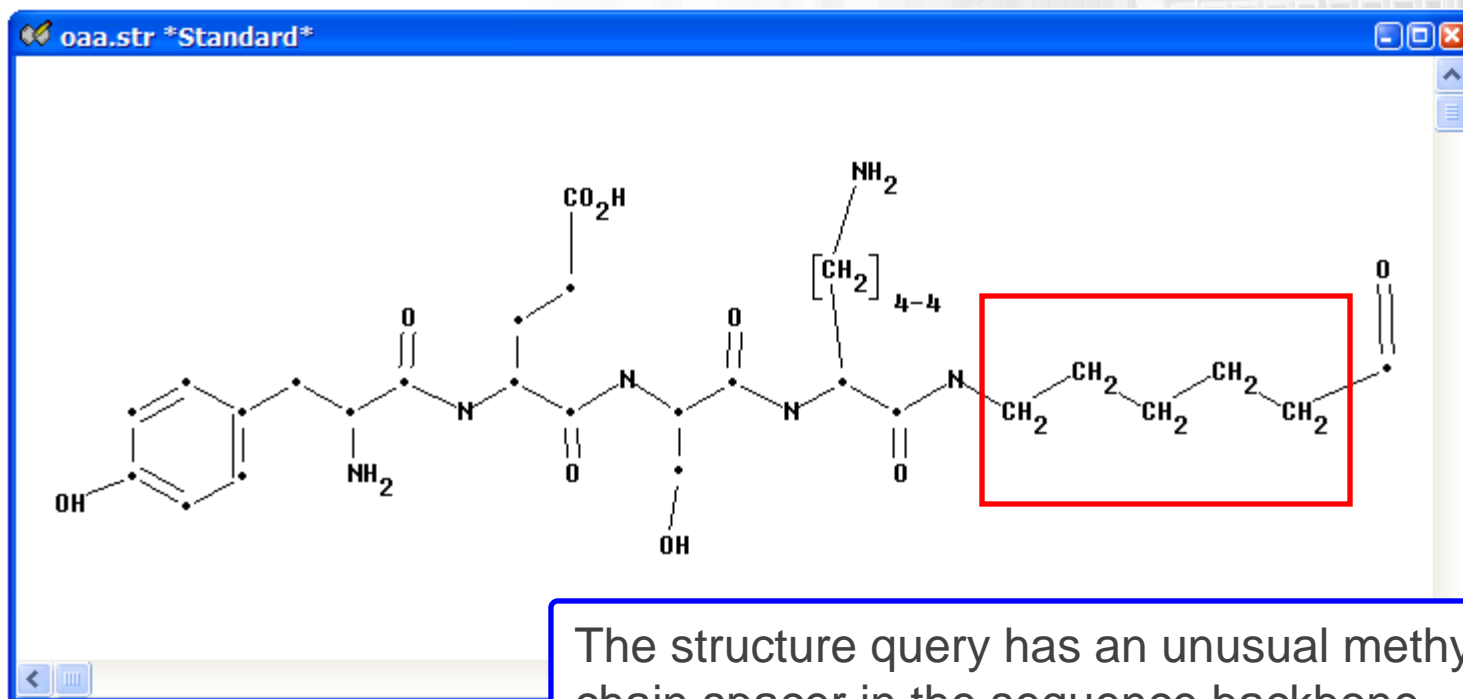
# Is this peptide known to bind metals?

A structure search can distinguish the manner of metal binding- via backbone or side chain functional groups.



# Ambiguous sequences

May be identified with an “X” in the sequence listing, or not listed as a sequence at all



The structure query has an unusual methylene chain spacer in the sequence backbone.

## What kinds of linkers have been used to build drug conjugates?

Conjugates (adducts, carriers, vesicles) are used to deliver a molecule to a target

A conjugate might be made up of a number of different parts:

AB A review. Different anticancer **drugs**, farmorubicin, doxorubicin, paclitaxel and cis-platin were conjugated through a Gly-Phe-Leu-Gly **tetrapeptide side chain** to a water-sol. synthetic **polymeric carrier** based on N-(2-hydroxypropyl)methacryalide (HPMA) ....

CAS indexes the chemistry of a conjugate based on what is revealed about it in a publication

## Possible ways that conjugates may be described in a publication

### 1. The conjugate is completely described

- CAS assigns RNs for the completely characterized substance in REGISTRY
- Uses descriptive text in CAplus<sup>SM</sup> to describe the role of such substances

### 2. The conjugate is partially described

- CAS assigns RNs for the part that is known
- Uses descriptive text in CAplus to describe what the rest of the conjugate is like

### 3. The conjugate has little or no characterization

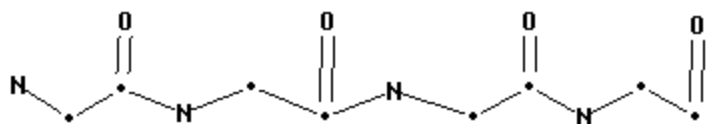
- Uses descriptive text in CAplus to describe what conjugate is like

### 4. The conjugate is described by a Markush structure in a patent

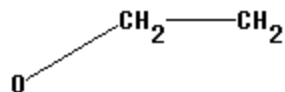
- Makes a Markush structure entry in MARPAT<sup>®</sup>  
If the focus is not the polymer chemistry

# Search question

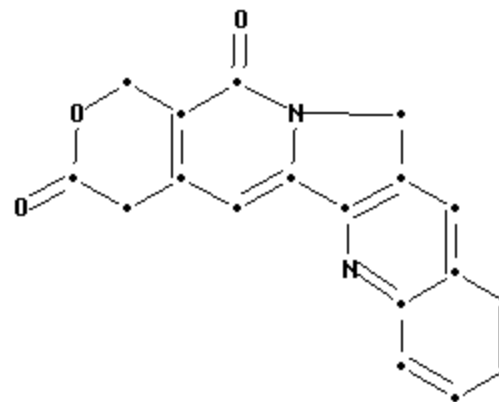
Locate information on pegylated conjugates of camptothecin, linked via a peptide linker at least 4 peptides long



Peptide backbone



PEG - polyethylene glycol



Camptothecin

# A comprehensive search has to take into account

## *CAS bases indexing on author emphasis*

- How substances might appear in documents
- What would be indexed based on the various scenarios described on the previous slide
- Which substances would be assigned RNs
- What would be indexed with subject or concept indexing in CPlus



# Run a SAMPLE substructure search

```
=> S L1 SSS SAM
```

```
SAMPLE SEARCH INITIATED 20:43:11
```

```
SAMPLE SCREEN SEARCH COMPLETED -          83 TO ITERATE
```

```
100.0% PROCESSED          83 ITERATIONS          7 ANSWERS
```

```
SEARCH TIME: 00.00.01
```

```
FULL FILE PROJECTIONS:  ONLINE  **COMPLETE**
```

```
                        BATCH  **COMPLETE**
```

```
PROJECTED ITERATIONS:          1114 TO          2206
```

```
PROJECTED ANSWERS:              7 TO          298
```

```
L2              7 SEA SSS SAM L1
```

A sample search projects the structure will run to completion.

# Evaluate answers with D SCAN

=> **D SCAN**

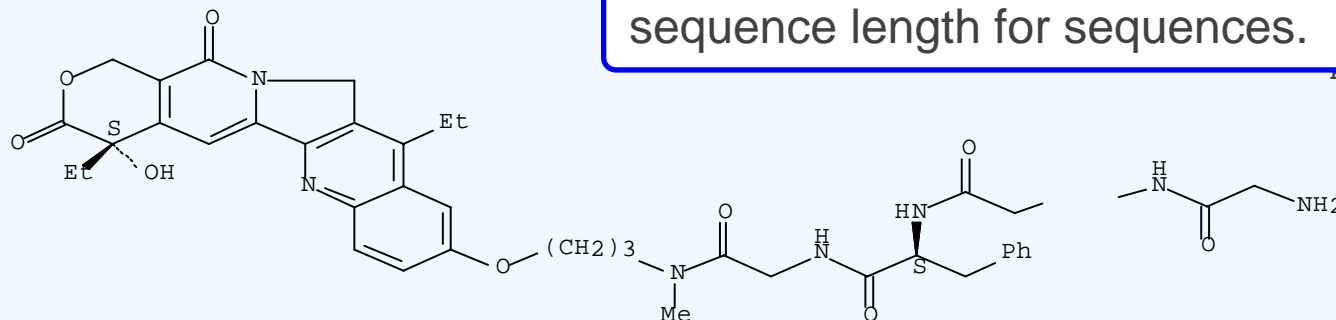
L2 7 ANSWERS REGISTRY COPYRIGHT 2008 ACS on STN

IN Glycinamide, glycyglycyl-L-phenylalanyl-N-[3-[[ (4S)-4,11-diethyl-3,4,12,14-tetrahydro-4-hydroxy-3,14-dioxo-1H-pyrano[3',4':6,7]indolizino[1,2-b]quinolin-9-yl]oxy]propyl]-N-methyl-, monohydrochloride (9CI)

SQL 4

MF C41 H47 N7 O9 . Cl H

D SCAN randomizes answers and shows CA Index Name, molecular formula, chemical structure if available, and sequence length for sequences.



**\*\*RELATED SEQUENCES AVAILABLE WITH SEQLINK\*\***

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1):0

# Run the FULL substructure search

```
=> S L1 SSS FULL
```

```
FULL SEARCH INITIATED 20:43:15
```

```
FULL SCREEN SEARCH COMPLETED -      1536 TO ITERATE
```

```
100.0% PROCESSED      1536 ITERATIONS
```

```
203 ANSWERS
```

```
SEARCH TIME: 00.00.01
```

```
L3      203 SEA SSS FUL L1
```

The 203 substances found in REGISTRY are for substances which are full characterized- all three pieces of the conjugate are identified in a single component of a REGISTRY record.

# Crossover to CAplus

=> **FIL CAPLUS**

=> **S L3**

L4 23 L3

=> **S L4 AND P/DT**

L5 15 L3 AND P/DT

=> **D TI 1-**

L5 ANSWER 1 OF 15 CAPLUS COPYRIGHT 2008 ACS on STN

TI Preparation of camptothecin-peptide conjugates and pharmaceutical compositions containing them

L5 ANSWER 2 OF 15 CAPLUS COPYRIGHT 2008 ACS on STN

TI Preparation of bivalent linkers for drug-peptide and other conjugates

L5 ANSWER 3 OF 15 CAPLUS COPYRIGHT 2008 ACS on STN

TI Complex drug delivery compositions for treating cancer

L5 ANSWER 4 OF 15 CAPLUS COPYRIGHT 2008 ACS on STN

TI Preparation of peptide conjugates in design of somatostatin vectors

• • •

The 203 substances are indexed to 23 references, 15 of which are patent publications. The titles indicate a highly relevant answer set.

## Display answers in more detail

=> **D IBIB ABS HITSTR 1-**

The IBIB ABS HITSTR format is recommended after a substructure search.

ACCESSION NUMBER: 2007:1146

DOCUMENT NUMBER: 147:44909

TITLE: Preparation of camptothecin-peptide conjugates and pharmaceutical compositions containing them

INVENTOR(S): Michel, Matthieu; Ravel, Denis; Ribes, Fabien; Tranchant, Isabelle

PATENT ASSIGNEE(S): Diatos S.A., Fr.

• • •

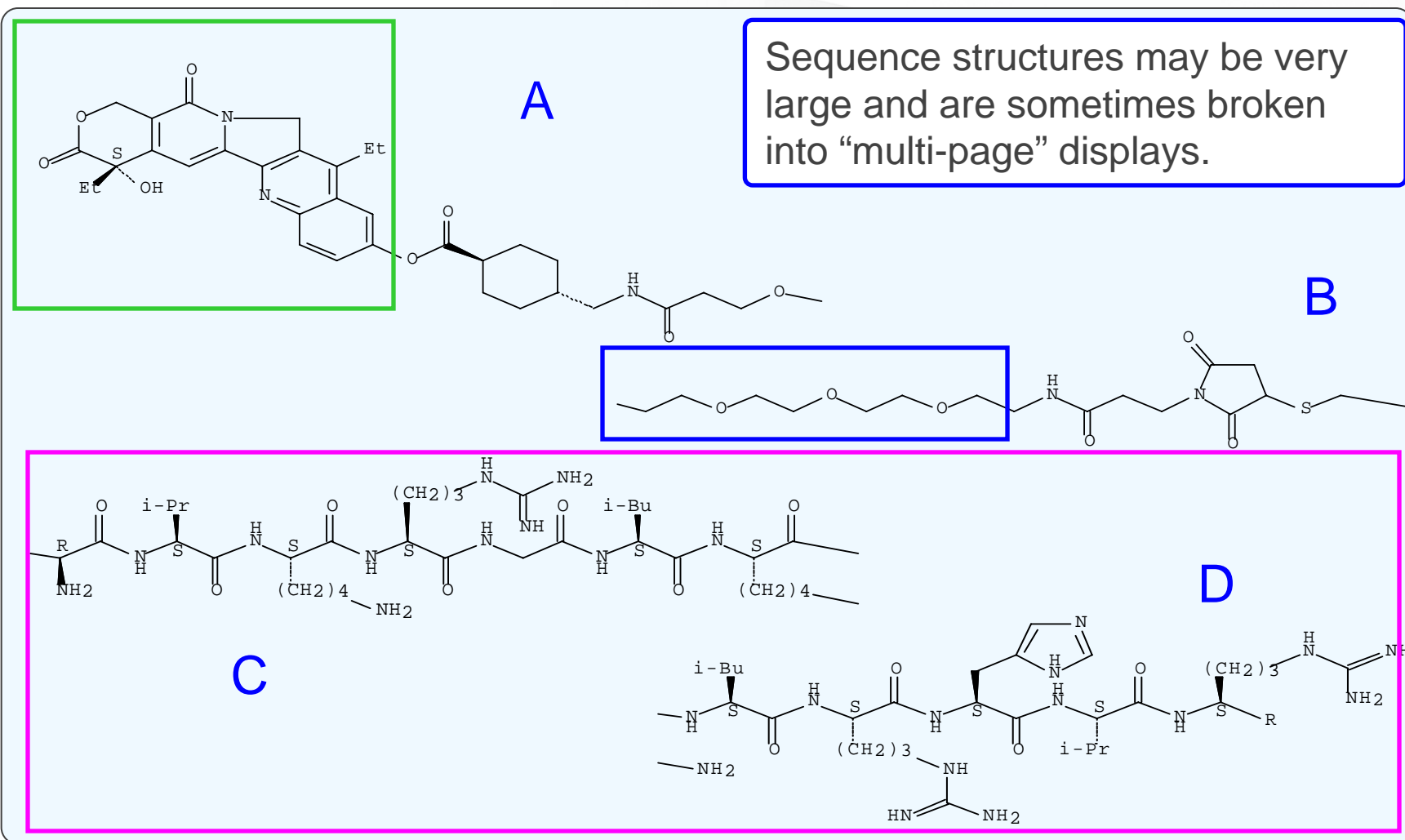
PATENT INFORMATION:

PATENT NO.	KIND	DATE	APPLICATION NO.	DATE
-----	----	-----	-----	-----
WO 2007113687	A2	20071011	WO 2007-IB1697	20070330
PRIORITY APPLN. INFO.:			EP 2006-290500	A 20060330
			US 2006-792312P	P 20060417

OTHER SOURCE(S): MARPAT 147:449092

AB The invention relates to novel camptothecin-peptide conjugates  
• • • for use in the improved delivery of therapeutic drug agents into target cells or tissues. The camptothecin-peptide conjugates provide numerous benefits, including enhancement in terms of aqueous solubility, pharmacokinetics and tissue distribution, enlargement of the therapeutic index, and limitation of the inter-patient metabolic variability. • • •

# HITSTR portion of display (condensed for slide)

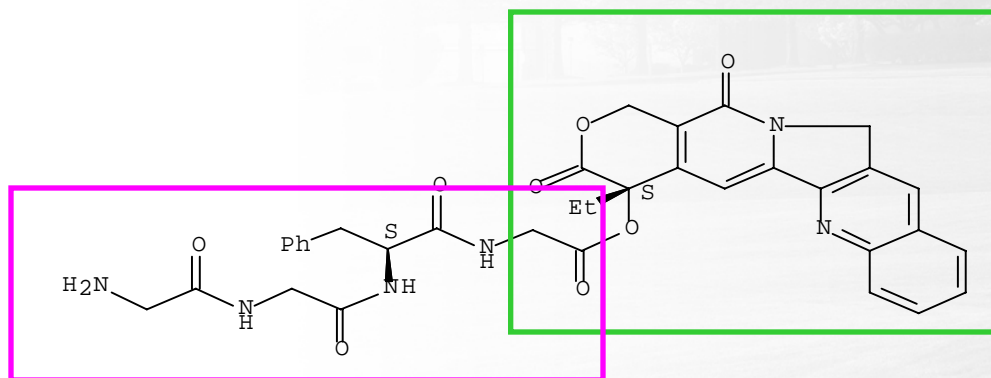


# Comprehensive searching takes into account the many ways a substance or subject might be registered or indexed

Other structure queries should be built so that all combinations are covered

## 1. Drug-linker structure query

- then use text in CAPLUS to add the **polymer**

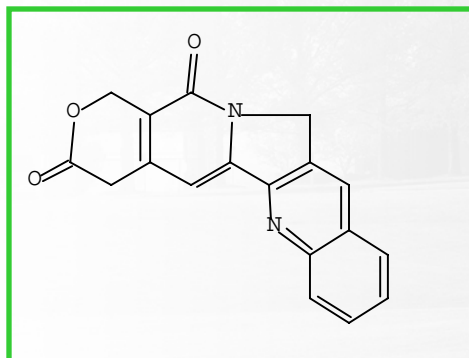
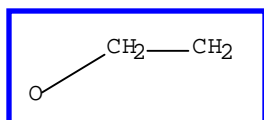


(preparation of modified **PEG**-drug complexes with improved drug targeting and water solubility)

# Comprehensive searching takes into account the many ways a substance or subject might be indexed or registered

## 2. Drug - polymer structure query

- then use text in CAPLUS to add the **linker**



(preparation and coupling of, with **peptide PEG** derivative; hydroxy-substituted 20-acyloxycamptothecin polymer derivs. and use thereof for the manufacture of an antiproliferative medicament)

## 3. Polymer - linker structure query

- then use text in CAPLUS to add the **drug**

## If you really need to be exhaustive

4. Another query will have only the **drug** query
  - use text in CAPLUS to add the **linker** and **polymer**
5. Another query will have only the **polymer** query
  - use text in CAPLUS to add the **drug** and **linker**
6. Another query (perhaps a pure sequence query or substructure) will have only the **linker** possibilities
  - use text in CAPLUS to add the **drug** and **polymer**

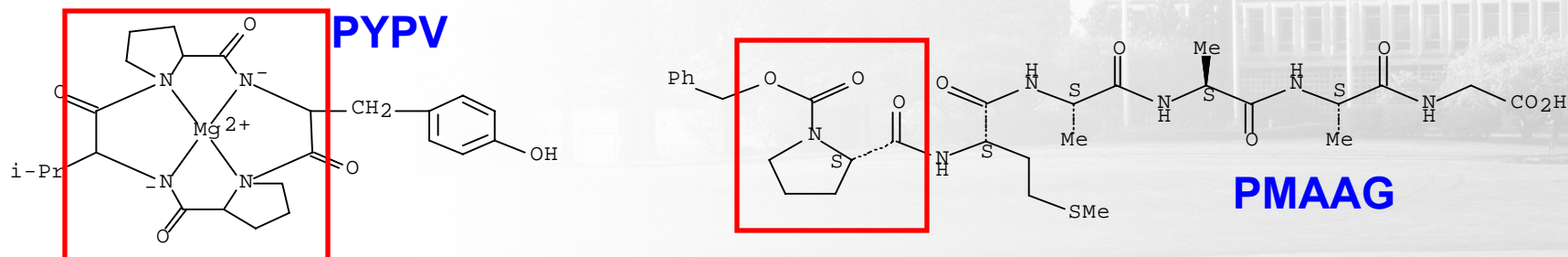
The need behind the question to be searched always drives the balance between comprehensive searching and amount and quality of retrieval.

# Tips for structure searching sequences

## Search Broadly!

- Make liberal use of ring/chain bonds

Especially if Proline may be in the sequence backbone



- Multiple structure queries if MARPAT is to be searched
- Use subsets or filters/screens if structures are projected incomplete
- If searching for metalated peptides be cautious about connecting bonds

Does the metal bind to the peptide backbone, or side chain functional group?

# Support and Training

**CAS Customer Care 1-800-753-4227**

**Instructor led workshops for sequence and structure search**

**STN Virtual workshops**

**Archived e-seminars**

Visit **[www.cas.org](http://www.cas.org)** support pages for more information.

# Acknowledgements

**Alice Humel-Denton, CAS Customer Care**

**Barb Vieira, Senior Product Development Manager, CAS**