

Round-tripping in sequence and structure space

- Searching across abstractions

Jan Holst Jensen
CEO, Biochemfusion



in collaboration with

AKos
Consulting & Solutions GmbH

Bridging the gap

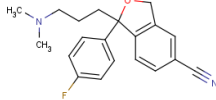
Cheminformatics

Bioinformatics

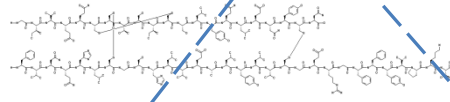
neither-nor

Molecule graphs

Sequences



```
1  GIVEQCCTSICSLYQLENYC  
21 NFNQHLCGSHLVEALYLVC  
41  GERGFFYTPKT
```



```
1  MVSQALRLLCLLLCLGQCCLAAGGVAKASGCETRDMPWRPG  
41  PRRVFTQEEAHCVLHRRRANAFLEELRPGSLERICKRE  
81  QCSFEAREIFKDAERTKLFMISYSDGDCASSPCQNGGS  
121 CRDQLQSYICFCLPAFEGRCNCETHRDDQLICVNEGCGCQ  
161 YCSDHTGTRRSCRCHEGYSLLADGVSCTPVEYPCGRIP  
201 LEKRNASKPQGRIVGCKVCFKGCECPVQVLLLVNCAQLCGG  
241 LLINTIVVVSAAHCFDKIKNWRNLIAVLGEHDLSEHDCDE  
281 QSRRVAQVIIPSTYVPGTINHDIALLRLHQPVVLTDHVV  
321 LCLPERTFSERTLAFVFSLVSGWCQLLDRCATALELMV  
361 NVPRLHTQDCLQQSRKWCDSPNITETHFCACYSDGSKDSC  
401 KGDSGGPHATHRGTWYLTGIVSWGQCCATVGHFGVYTRV  
441 SQYIEWLQKLMRSEPRFCVLLRAPFF
```

100

10k

1M

MW
Da

CWM Global Search – now with a “Proteax” button!

The screenshot shows the CWM Global Search application interface. The top navigation bar includes tabs for Home, Simple search, Advanced search, and Results. Below this is a toolbar with various icons: Modify query, New query, Override query, Add/Remove custom profile, Clear structure box, Import, Open, Save, Reaxys LinkIn, Proteax (circled in red), Help, and Start Global Search. Below the toolbar are two red buttons: "Go to Simple Search" and "Go to Results".

The main content area is titled "You are on the Advanced Search page" and shows "Page 1 of 1". It contains several panels:

- Structure Search:** A yellow box is highlighted with the text "Double click to edit structure". Below it are checkboxes for "Include isomers", "Include Substructures", and "Include similar compounds", along with a "Similarity coefficient (1-100)" set to 90.00.
- CAS number:** A search field with a dropdown menu and a "DOUBLE click here to add new" button.
- Free text:** A search field with a dropdown menu and a "DOUBLE click here to add new" button.
- Filter operations:** A section with a "Filter" dropdown and a "Remove all filter" button.
- Table:** A table with columns "Name" and "Category". The table lists various databases and their categories.

Name	Category
<input type="checkbox"/> ACS Publications	Literature
<input type="checkbox"/> AKOSSAMPLES	Suppliers
<input type="checkbox"/> Bielefeld Academic Search Engine	Open Access
<input type="checkbox"/> BindingDB	Proteins
<input type="checkbox"/> BioMed Central	Open Access
<input type="checkbox"/> BMRB	Spectra
<input type="checkbox"/> Buyersguide	Suppliers
<input type="checkbox"/> CCRIS	Toxicity
<input type="checkbox"/> ChEBI	Bioactivity
<input type="checkbox"/> ChemAxon Chemicalize Search	Search engine
<input type="checkbox"/> ChemBank	Bioactivity
<input type="checkbox"/> ChEMBL	Bioactivity
<input type="checkbox"/> Chemo	Chemical pro
<input type="checkbox"/> ChemE	Suppliers

Marvin

Let's check for Oxytocin look-a-likes

Pos 3 (Ile) and 8 (Leu) are substituted with Glycines and left without explicit hydrogens so any sidechain will be allowed in a substructure mapping.

Clean 3Letter Configure Turn On/Off Format 1/3 Global Search

PLN Hydrogens Editor

H Cys(1) Tyr Gly Gln Asn Cys(1) Pro Gly Gly [NH2]

Protein text - PLN format
H-Cys (1) -Tyr-Gly-Gln-Asn-Cys (1) -Pro-Gly-Gly- [NH2]

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

Sequence Molecule Sum formula: C35 H50 N12 O12 S2 Avg. MW: 894,9747

1 CYGQNCPPG

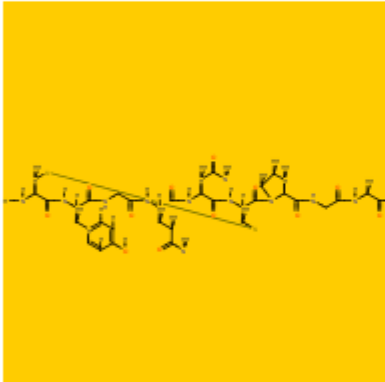
Query structure is generated for you by Proteax; let's start the search

You are on the Advanced Search page

Page 1 of 1

Globalsearch_1 2012-04-29 22:05:31

Double click to edit structure



Structure Search

- Include isomers
- Include Substructures
- Include similar compounds

Similarity coefficient (1-100)

Use the rightmost 'Pin' button to show/hide this pane

Filter Filter operations

Drag a column header and drop it here to group by that column

<input type="checkbox"/>	Name	Category	Structure	SSS/Sim	Text
<input type="checkbox"/>	Open J-Gate	Open Access			<input checked="" type="checkbox"/>
<input type="checkbox"/>	PharmGKB	Drugs			<input checked="" type="checkbox"/>
<input type="checkbox"/>	PLOSONe	Open Access			<input checked="" type="checkbox"/>
<input type="checkbox"/>	PNAS	Literature			<input checked="" type="checkbox"/>
<input type="checkbox"/>	Protein Data Bank	Proteins	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	PubChem	Bioactivity	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	PubMed	Literature	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
<input type="checkbox"/>	PubMedCentral	Literature	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
<input type="checkbox"/>	Quertle	Literature			<input checked="" type="checkbox"/>
<input type="checkbox"/>	SCOPUS	Literature			<input checked="" type="checkbox"/>

We find 101 hits in PubChem

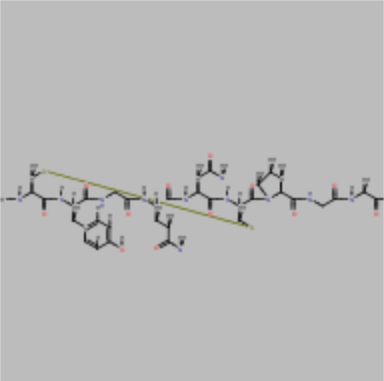
You are on the Results page

Page 1 of 1

2012-04-29 22:05:31

Structure

Double click to view structure



Quicklinks **Globalsearch links**

Remove all filter

2012-04-29 22:05:31

Extreg	Number of hits	Status
RECORD_1	1	OK

Drag a column header and drop it here to group by that column

<input type="checkbox"/>	<input type="text"/>	Datasource	Link	Categr	Keywords	Searchtype	Que
<input checked="" type="checkbox"/>		PUBCHEM	101 PubChem CID(s) found.	Bioactivity		SUBSTRUCTURE	


PubChem hit list


http://www.ncbi.nlm.nih.gov/sites/entrez?db=pccompound&term=5771,577 CWM Global Search v_6.0.2.0 5771 5772 8230 18830 24774... x My NCBI Sign In

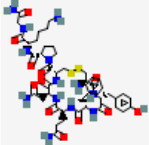
NCBI Resources How To PubChem Compound PubChem Compound 5771 5772 8230 18830 24774 68649 123799 164712 165100 165207 167997 191984 194531 19... Search Help

Display Settings: Summary, 20 per page, Sorted by Default order Send to: Filters: Manage Filters





Results: 1 to 20 of 101 << First < Prev Page 1 of 6 Next > Last >>

1.  [ARGIPRESSIN; Arginine vasopressin; \[Arg8\]-Vasopressin ...](#)
MW: 1084.231600 g/mol MF: C₄₆H₆₅N₁₅O₁₂S₂
IUPAC: (2S)-1-[(4R,7S,10S,13S,16S,19R)-19-amino-7-(2-amino-2-oxoethyl)-10-(3-amino-3-oxopropyl)-13-benzyl-1-...
[Active in 1 BioAssay](#) [Tested in 35 BioAssays](#)
CID: 644077
[Similar Compounds](#) [Same Parent, Connectivity](#) [Mixture/Component Compounds](#) [PubMed \(MeSH Keyword\)](#)

2.  [Lysipressin; Lysopressin; Syntopressin ...](#)
MW: 1056.218200 g/mol MF: C₄₆H₆₅N₁₃O₁₂S₂
IUPAC: (2S)-N-[(2S)-6-amino-1-[(2-amino-2-oxoethyl)amino]-1-oxohexan-2-yl]-1-[(4R,7S,10S,13S,16S,19R)-19-am...
CID: 644076
[Similar Compounds](#) [Same Parent, Connectivity](#) [Mixture/Component Compounds](#) [PubMed \(MeSH Keyword\)](#)



3.  [AC1LCW40; \(2S\)-N-\[\(2S\)-6-amino-1-\[\(2-amino-2-oxoethyl\)amino\]-1-oxohexan-2-yl\]-1-\[\(4R,7S,10S,13S,16R,19R\)-19-amino-7-\(2-amino-2-oxoethyl\)-10-\(3-amino-3-oxopropyl\)-13-benzyl-16-\[\(4-hydroxyphenyl\)methyl\]-6,9,12,15,18-pentaoxo-1,2-dithia-5,8,11,14,17-pentazacycloicosane-4-carbonyl\]pyrrolidine-2-carboxamide; 1-\[\(19-amino-7-\(2-amino-2-oxoethyl\)-10-\(3-amino-3-oxopropyl\)-13-benzyl-16-\(4-hydroxybenzyl\)-6,9,12,15,18-pentaoxo-1,2-dithia-5,8,11,14,17-pentazacycloicosan-4-yl\]carbonyl\]prolyl-N-\(2-amino-2-oxoethyl\) ...](#)
MW: 1056.218200 g/mol MF: C₄₆H₆₅N₁₃O₁₂S₂
IUPAC: (2S)-N-[(2S)-6-amino-1-[(2-amino-2-oxoethyl)amino]-1-oxohexan-2-yl]-1-[(4R,7S,10S,13S,16R,19R)-19-am...
CID: 638316
[Similar Compounds](#) [Same Parent, Connectivity](#)

Actions on your results

-  **BioActivity Analysis**
Analyze the BioActivities of the compounds
-  **Structure Clustering**
Cluster structures based on structural similarity
-  **Structure Download**
Download the structures in various formats
-  **Pathways**
Analyze pathways containing the compounds

Refine your results • What's this?

BioActivity Experiments

- BioAssays, Active (11) 
- BioAssays, Tested (23) 
- Protein 3D Structures (1)
 - Crystal Structure Of Trypsin-Vasopressin Complex (1)

BioMedical Annotation

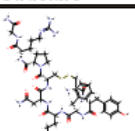
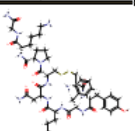
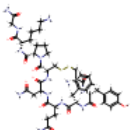
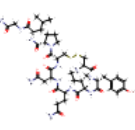
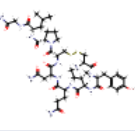
- Pharmacological Actions (19)
 - Vasoconstrictor Agents (15)
- BioSystems (2)

Depositor Category

Now, analyze this...

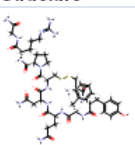
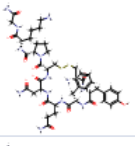
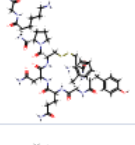
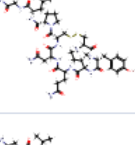
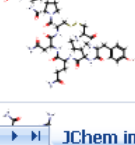
Import PubChem results from SD file with JChem for Excel

The screenshot shows the Microsoft Excel interface with the JChem add-in. The active cell A2 contains the formula `=JCSYSstructure("9FF5B0E6B42565BD82C76EFD856")`. The table below displays the imported data:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Structure	PUBCHEM_COMPOUND_CID											
2		644077											
3		644076											
4		638316											
5		638315											
6		439302											

Convert structures to Protein Line Notation (PLN) with Proteax

The screenshot shows a Microsoft Excel spreadsheet with the following data:

1	Structure	PUBCHEM_COMPOUND_CID	Structure => PLN via Proteax
2		644077	H-Cys(1)-Tyr-Phe-Gln-Asn-Cys(1)-Pro-Arg-Gly-[NH2] name=644077
3		644076	H-Cys(1)-Tyr-Phe-Gln-Asn-Cys(1)-Pro-Lys-Gly-[NH2] name=644076
4		638316	H-Cys(1)-dTyr-Phe-Gln-Asn-Cys(1)-Pro-Lys-Gly-[NH2] name=638316
5		638315	H-Cys(1)-dTyr-Ile-Gln-Asn-Cys(1)-Pro-Leu-Gly-[NH2] name=638315
6		439302	H-Cys(1)-Tyr-Ile-Gln-Asn-Cys(1)-Pro-Leu-Gly-[NH2] name=439302

Show variations as Biochemfusion DerNot expressions (“peptide naming”)

Imported_Gly3_Gly8_open.xlsx - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Developer JChem Proteax

Live Preview Preview options Live Preview 3rd party New protein Edit protein View molecule From JChem structure(s) To JChem structure(s) Copy to JChem sheet Import from file Import from SD file Import from server Export to file Export to MDL@ SD file Compare proteins Formula error help Register/Update license Manage modification database

D2 =PROTEAX_DERNOT_DIFF(C2;E2)

	A	B	C	D	E
1	Structure	PUBCHEM_COMPOUND_CID	Structure => PLN via Proteax	Compared to Oxytocin	Oxytocin reference sequence
2			H-Cys(1)-Tyr-Phe-Gln-Asn-Cys(1)-Pro-Arg-Gly-[NH2] name=644077	F(3)R(8) Oxytocin	H-Cys(1)-Tyr-Ile-Gln-Asn-Cys(1)-Pro-Leu-Gly-[NH2] name=Oxytocin
3			H-Cys(1)-Tyr-Phe-Gln-Asn-Cys(1)-Pro-Lys-Gly-[NH2] name=644076	F(3)K(8) Oxytocin	
4			H-Cys(1)-dTyr-Phe-Gln-Asn-Cys(1)-Pro-Lys-Gly-[NH2] name=638316	{d}YF(2-3)K(8) Oxytocin	
5			H-Cys(1)-dTyr-Ile-Gln-Asn-Cys(1)-Pro-Leu-Gly-[NH2] name=638315	{d}Y(2) Oxytocin	
6			H-Cys(1)-Tyr-Ile-Gln-Asn-Cys(1)-Pro-Leu-Gly-[NH2] name=439302	Oxytocin	

JChem imported structures Sheet2 Sheet3

Ready 100%

Links...

- Proteax for Spreadsheets, by Biochemfusion ApS
 - <http://www.biochemfusion.com/downloads/>
 - Direct data link to JChem for Excel, by ChemAxon
<http://www.chemaxon.com/download/jchem-for-excel/>
 - Structure => PLN conversion will become available with release 2.0 later this year
- CWM Global Search, by AKos GmbH
 - <http://cwmglobalsearch.com/gsweb/>