

# Примјена рачунара у биологији



**Владимир Филиповић**

[vladaf@matf.bg.ac.rs](mailto:vladaf@matf.bg.ac.rs)

# Хемијске формуле и реакције у радовима



# Chemical Markup Language CML

# Chemical Markup Language - CML

XML је данас најкоришћеније приступ за обезбеђивање семантике у научним документима, као што су MathML (за математику), SBML/BIO-PAX (за биологију), GML and KML (за геометрију), SVG (за графику) and NLM-DTD, ODT and OOXML (за структурисање докумената).

Chemical Markup Language (ChemML or CML) је приступ за управљање информацијама о молекулима коришћењем алата као што су језик за означавање XML и програмског језика и окружења Java.

Ово је прва доменски специфична имплементација која је стриктно базирана на XML-у.

CML обезбеђује подршку за највећи део хемије, нарочито молекула, једињења, реакција, спектра, кристала и рачунарске хемије (compchem).

# Chemical Markup Language – CML (2)

Традиционално, хемијске информације су биле чуване у различитим форматима (типovima) датотека, што је спречавало њихову поновну искористивост.

CML користи преносивост XML-а да би помогао креирање интероперабилних докумената од стране CML аутора и хемичара. Постоји велики број алата који могу генерисати, обрађивати и прегледати CML документе. Издавачи могу, коришћењем CML-а, дистрибуирати хемију преко XML докумената.

CML може да подржи широк опсег хемијских појмова:

- молекуле
- реакције
- спектре и аналитичке додатке
- рачунарску хемију
- хемијску кристалографију и материјале

# Chemical Markup Language – CML (3)

```
<cml convention="conventions:molecular" xmlns="http://www.xml-cml.org/schema"
conventions="http://www.xml-cml.org/convention/" cmlDict="http://www.xml-
cml.org/dictionary/cml/" nameDict="http://www.xml-cml.org/dictionary/cml/name/">
  <molecule id="sulfuric acid">
    <formula concise="sulfuric acid"/>
  </molecule>
  <molecule id="">
    <formula title="[Cu(NH3)4]2+ SO42-]>
      <formula formalCharge="+2">
        <atomArray elementType="Cu"/>
        <formula count="4">
          <atomArray elementType="N H" count="1 3"/>
        </formula>
      </formula>
      <formula formalCharge="-2">
        <atomArray elementType="S O" count="1 4"/>
      </formula>
    </formula>
  </molecule>
</cml>
```

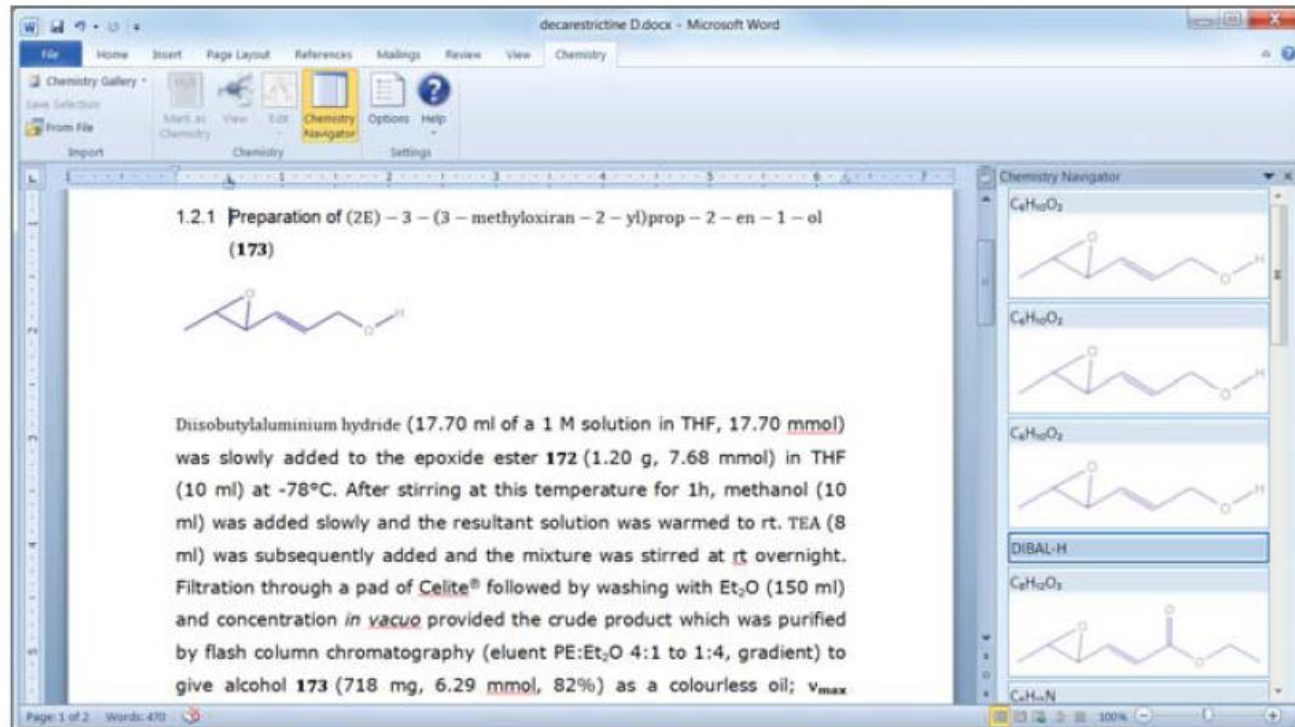


# Chemistry Add-in for Word

# Chemistry Add-inn for Word

Свака наука има свој језик. За успех ма ког научног истраживања је од огромног значаја могућност комуникације и сарадње коришћењем језика те науке.

У хемији и биологији, не само да постоји специфичан језик, већ је то специфичан језик са својим специфичним симболима.



*Name and 2D views of the same chemical shown in the document, along with the Chemistry Navigator; which displays all of the chemistry zones within the current document.*



# Chemistry Add-inn for Word (2)

Програм Chemistry Add-in for Word олакшава студентима, хемичарима и истраживачима да убаце и модификују хемијске информације, као што су ознаке, формуле и 2D прикази, из и

у Microsoft Office Word.

Он упреже моћ језика

Chemical

Markup

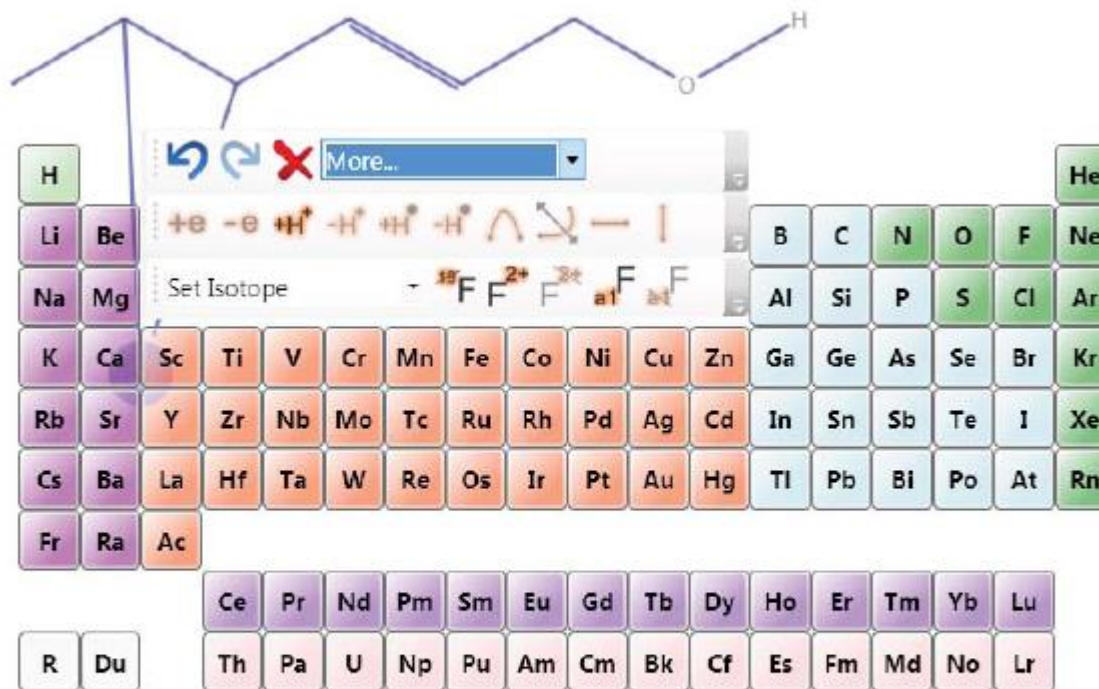
Language

(CML) који

представља

XML за

хемију.



*The 2D Editor showing the Periodic Table of Elements lookup.*

# Chemistry Add-inn for Word (3)

CML допушта истраживачима не само да креирају и мењају хемијске формуле у Word-у, већ и да у документ укључе и податке који се садрже у овим формулама.

Chemistry Add-in и CML обезбеђују да хемиски документни буду отворени, читљиви и доступни људима и другим технологијама.

Поред функционалности ауторства, Chemistry Add-in допушта да корисник маркира тзв. „инлајн“ хемијске зоне, исцртавање високо квалитетних визуалних приказа хемијских структура, као и могућност чувања и излагања хемијских информација са богатом семантиком у глобалној хемијској заједници.

Chemistry Add-in supports подржава сценарије објављивања и истраживања података за ауторе, читаоце, издаваче и све остале у информатичкој хемијској заједници.

# Chemistry Add-inn for Word (4)

Коришћењем Chemistry Add-in, могуће је:

- Креирати „инлајн“ хемијске зоне (контроле које садрже информације о молекулу) ради репрезентације хемијских података.
- Креирати хемијску зону уношењем уобичајеног назива (на енглеском – нпр. “water”), а потом конвертовати ту хемијску зону да буде представљена на жељени начун.
- Пребацити са тривијалног назива молекула на његову концизну формулу или на његову 2D репрезентацију.
- Представити молекуле помоћу високо квалитетног 2D структурног дијаграма и користити уграђени едитор за модификовање структуре.

# Chemistry Add-inn for Word (5)

Коришћењем Chemistry Add-in, могуће је:

- Извршити увоз молекула из других алата за графичко едитовање, мењати њихову структуру и сачувати измењене молекуле у Галерији, тако да ти исти молекули могу бити коришћени и у разним другим документима.
- Сачувати и изложити хемијске информације на семантички богат начин.

# Bioclipse



# Bioclipse

Bioclipse is a free and open source workbench for the life sciences.

Bioclipse is based on the Eclipse Rich Client Platform (RCP) which means that Bioclipse inherits a state-of-the-art plugin architecture, functionality, and visual interfaces from Eclipse, such as help system, software updates, preferences, cross-platform deployment etc.

Bioclipse provides advanced functionality in fields such as cheminformatics, bioinformatics, semantic web, spectrum analysis, drug discovery, safety assessment, and general chemistry education.

# Bioclipse (2)

Bioclipse is developed as a collaboration between the Proteochemometric Group , Dept. of Pharmaceutical Biosciences, Uppsala University, Sweden, and the Cheminformatics and Metabolism Team at the European Bioinformatics Institute (EBI).

Bioclipse is released under Eclipse Public License (EPL) + exception, see the License Statement, putting no constraints on choice of backend and/or license for creating plugins for Bioclipse; it is totally open for both open source plugins as well as commercial.

# Bioclipse (4)

Some examples of functionality is listed below:

## **Cheminformatics**

- Cheminformatics in Bioclipse is mainly based on the Chemistry Development Kit (CDK), and contains a framework for managing and analyzing chemical compounds.
- Bioclipse supports editing in 2D, processing large collections of molecules in tables, calculation of various types of properties, and much more cheminformatics functionality.
- The Jmol application is integrated in Bioclipse as an editor, and provides advanced interactive 3D visualizations.



# Bioclipse (5)

## Illustration: Work with collections of molecular structures

Bioclipse

Bioclipse Navigator

ChEBI\_complete.sdf

ola

ChEBI\_complete.sdf

Sample Data

- 2D structures
- 3D Structures
- Javascripts
- PDB
- SDF

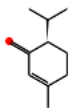
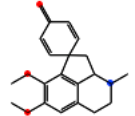
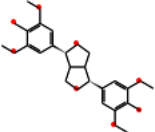
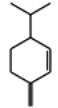
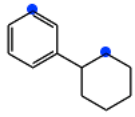
	2D-structure	IUPAC Names	SMILES	Last Modified
9		(6S)-3-methyl-6-(propan-2-yl)cyclo...	<chem>CC(C)[C@@H]1CCC(C)=CC...</chem>	25 Mar 2008
10		(8a'R)-5',6'-dimethoxy-1'-methyl-2',...	<chem>[H][C@]12CC3(C=CC(=O)C...</chem>	10 Mar 2008
11		(7alpha,7'alpha,8alpha,8'alpha)-3,3',...	<chem>[H][C@]12CO[C@H](c3cc(O...</chem>	06 May 2008
12		(4S)-p-mentha-1(7),2-diene(6S)-3-...	<chem>[H][C@]1(CCC(=C)C=C1)C(...</chem>	23 May 2008
13		3-[(2S)-piperidin-2-yl]pyridine	<chem>[H][C@]1(CCCC1)c1ccncc1</chem>	15 Jan 2008

Table | Single Molecule | Headers

Outline | Properties

Property | Value

General

- Has 2D Coords: yes
- Has 3D Coords: no
- Molecular Format: N/A
- Molecular Formula: C10H16H16
- Molecular Mass: 136.2344

Molecular Properties

- Beilstein Registry Number: 19027674229885
- CAS Registry Number: 498-15-7
- Charge: 0
- ChEBI ID: CHEBI:7
- ChEBI Name: (+)-car-3-ene
- Formulae: C10H16
- Gmelin Registry Number: 663435
- InChI: InChI=1/C10H2/c1-7-4-5
- InChI: InChI=1/C10H2/c1-7-4-5
- InChIKey: QTEBNBFBDXVOON-BDAKN
- InChIKey: QTEBNBFBDXVOON-BDAKN
- IUPAC Names: (1S,6R)-3,7,7-trimethylbicy
- KEGG COMPOUND Database ID: C11382
- Last Modified: 10 May 2006
- Mass: 136.23404
- PubChem Database ID: 11533292
- SMILES: CC1=CCC2C(C1)C2(C)(C)
- SMILES: CC1=CCC2C(C1)C2(C)(C)
- Synonyms: (+)-3-Carene(+)-Delta(3)-

# Bioclipse (6)

## Illustration: Editing of chemical structure

The screenshot displays the Bioclipse software interface with the following components:

- Bioclipse Navigator:** A tree view on the left showing a hierarchy of files. Under "Sample Data", there are "2D structures" (0037.cml, 0037.mdl, ATP.mol, polycarpol.mol, reserpine.mol, thiamin.mol) and "3D Structures" (2-methylpropanal, pentan-1-ol.mol, pentanal.cml, propan-2-ol.cml, tetracosane.cml). Other folders include "Javascripts", "PDB", "SDF", and "Fragments2.sdf".
- Central Canvas:** Displays the chemical structure of reserpine, a complex alkaloid with a pentacyclic core and a long side chain ending in a phosphate group.
- Properties Panel:** Located at the bottom, it shows a table of properties for the selected molecule.

Property	Value
<b>General</b>	
Has 2D Coords	yes
Has 3D Coords	no
InChI	InChI=1/C10H19NSO13P3/c11-8-5-9(13-2-12-8)15(3-14-5)10-7(17)6(16)4
InChIKey	VIYUSENGHRDRMH-UHFFFAOYAU
Molecular Format	MDL Molfile (2D)
Molecular Formula	C10H19NSO13P3
Molecular Mass	510.2055
SMILES	O=P(O)(O)OP(=O)(O)(O)OP(=O)(O)OCC3OC(N2CNC1C(N)CNC12)C(O)C3(O)

The right sidebar contains an "Outline" panel listing atoms and bonds, and a "Bonds" panel listing bond types.

# Bioclipse (7)

Illustration: Compound visualized in 3D with Jmol

The screenshot displays the Bioclipse application window. The central 3D viewer shows a ball-and-stick model of a diazepam molecule, with a semi-transparent orange surface representing a molecular docking or interaction field. The molecule features a benzodiazepine core with a diazepam substituent.

On the left, the 'Bioclipse Navigator' panel shows a hierarchical tree of chemical structure reports. The 'drugs' folder is expanded, listing various compounds, with 'diazepam.mol' selected.

On the right, the 'Outline' panel lists atoms C1 through C17, with C4, C5, C7, and N8 highlighted in green. Below it, the 'Properties' panel shows the 'General' tab for the selected molecule, displaying metadata such as 'name: diazepam.mol' and 'size: 2,419 bytes'.

At the bottom, the 'Javascript Console' panel shows the following code and output:

```
org.mozilla.javascript.Undefined@dee89b  
fromCml fromSMILES fromString  
  
> mol=cdk.fromSMILES("C1CCCCC1CCOCCO")  
CDKMolecule:C10H20O2  
  
> cdk.calculateMass(mol)  
152.10616882301883
```

The status bar at the bottom indicates the current file path: 'BOdata/Blue Obelisk Chemical Structure Repository/drugs/diazepam.mol'.

# Bioclipse (8)

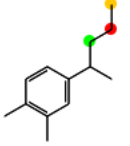
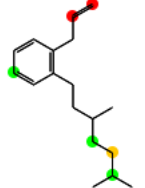
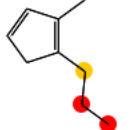
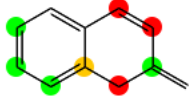
## Pharmacology and Drug Discovery

- Bioclipse is equipped with many features that simplify pharmacological research and drug discovery. Quantitative Structure-Activity Relationships (QSAR) is a methodology to relates the responses between several chemical structures and a target, for example to measure the toxicity of drugs in the human body, described by mathematical descriptors and modelled using statistical methods.
- QSAR models can for example predict if a compound is toxic, and scientists can get decision support for changing the chemical structure before even testing it in the wet lab.
- Bioclipse has many features to work with QSAR and similar fields. A QSAR Project is available with a Builder on a QSAR project file, completely describing the project model. A editor is available, with tabs to select chemical structures, choose mathematical descriptions (descriptors), and all other metadata for performing the analysis.

# Bioclipse (9)

Illustration: Predicting site-of-metabolism on multiple molecules

The screenshot displays the Bioclipse software interface. On the left, the 'Bioclipse Navigator' shows a file tree with '20Mols2d.sdf' selected. The main window, titled '20Mols2d.sdf', contains a table with four rows of molecule data. Each row includes a number, a 2D chemical structure, and the molecule's name. The molecules are adrenaline, alprenolol, clomethiazole, and coumarin. At the bottom left, the 'MetaPrint2D Report' shows the results of the last run, indicating a successful status and a calculation time of 49 ms.

	2D-structure	cdk:Title
1		adrenaline
2		alprenolol
3		clomethiazole
4		coumarin

MetaPrint2D Report

Last MetaPrint2D run

Status OK  
Database ALL  
Operator DEFAULT  
Calculation time 49 ms

# Bioclipse (10)

## Bioinformatics

- Bioinformatics in Bioclipse concerns primarily the management and analysis of biological sequences (DNA, RNA, and protein).
- Bioinformatics in Bioclipse relies heavily on BioJava, which provides core bioinformatics functionality, and a graphical editor for sequence alignments.
- Various clients for Web services are also available to facilitate downloading of e.g. biological sequences and annotations, as well as for bioinformatics analysis.

# Bioclipse (11)

Illustration: Sequence Editor displaying a wrapped alignment

The screenshot displays the Bioclipse Sequence Editor interface. The main window shows a sequence alignment for two sequences, OPSD\_FELCA and OPSD\_SHEEP, with a consensus sequence. The alignment is wrapped across multiple lines. The Bioclipse Navigator on the left shows the project structure, including 'MyProject', 'MyQSARproject', and 'seqs' with files 'zf\_zf2\_aligned.fasta', 'zf.fasta', 'zf2.fasta', and 'zf3.fasta'. The bottom panel shows the 'Sequences' tab with a 'Source' view and a 'JavaScript Console' with a query result for 'OPSD\_SHEEP'.

Sequences | Source

Outline

Properties | Progress | JavaScript Console

```
> biows.queryUniProtKB("OPSD_SHEEP")
[Protein OPSD_SHEEP:
'mngtegnfyvpsnktgvvrspfeapqyilaepwqfsmloaymflilvgfpinftlyvtvghkklrtplnyilln
lavadlfmvfggfttlytshgyfvfgptgcnlegffatlggeialwslvlaieryvvvckpmsnfrfgenhaimgv
aftwmalacaapplvgwsryipagmqscgalftlkepinnesfviymfvhfsipliviffcygqlvftvkeaaaq
qesattqkaekvtrmviimviaflcwlpyagvayifithagsdfgpi fmtipaffaksssvynpviymnkqfrn
cmittlccgknplgddeastvsktetsqvapa']
```

# Bioclipse (12)

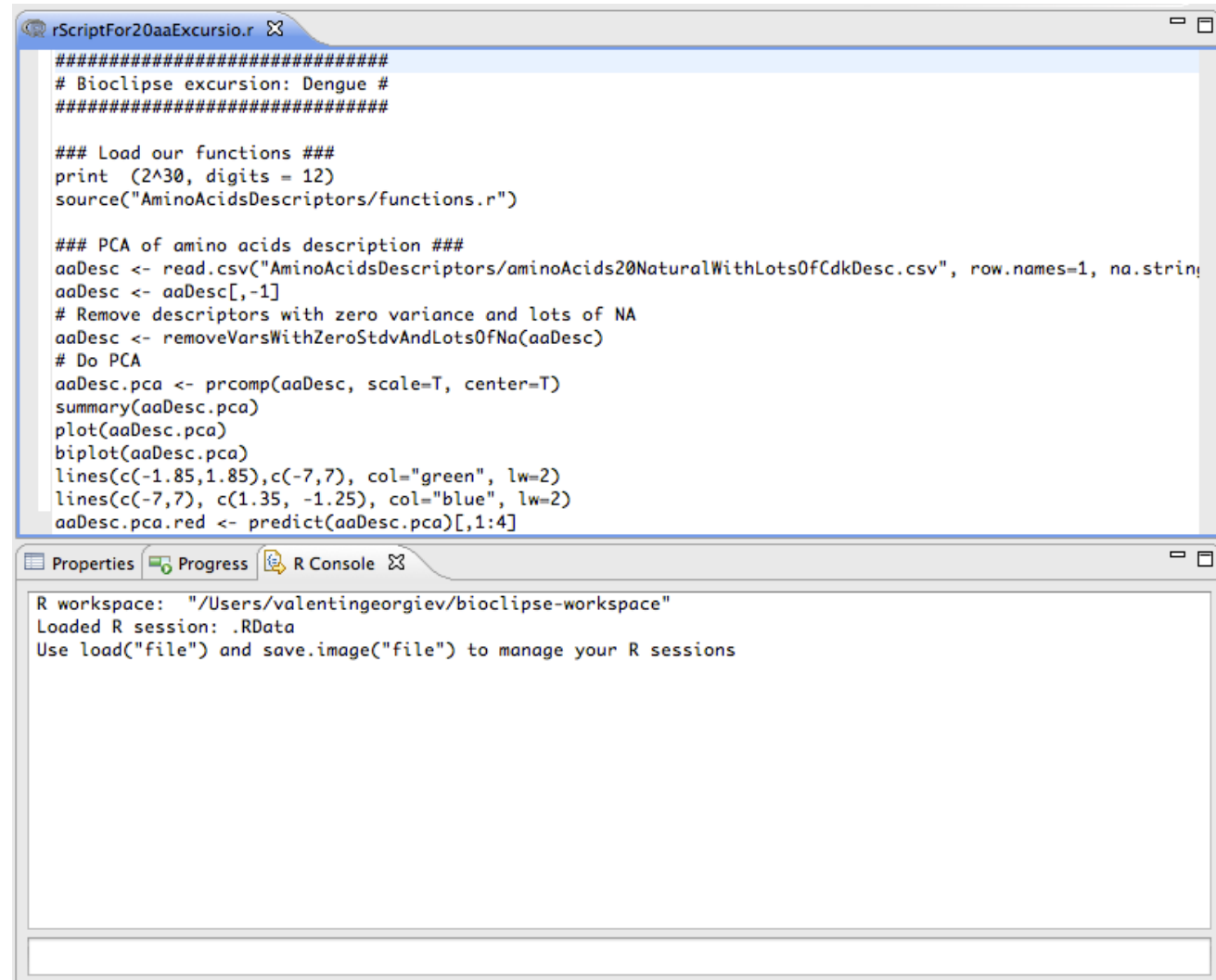
## Bioclipse-R Itegration

- The statistical computing language R is now integrated in Bioclipse.
- R is a free and open source software that runs on Windows, Mac OS, and a wide variety of UNIX platforms and similar systems (including FreeBSD and GNU/Linux).
- The Bioclipse-R feature provides a graphical R editor and a interactive R console for easy running of R scripts, snippets and commands, and the plotting capabilities of R.



# Bioclipse (13)

## Illustration: Integration of the R into Bioclipse workbench



```
rScriptFor20aaExcursio.r
#####
# Bioclipse excursion: Dengue #
#####

### Load our functions ###
print (2^30, digits = 12)
source("AminoAcidsDescriptors/functions.r")

### PCA of amino acids description ###
aaDesc <- read.csv("AminoAcidsDescriptors/aminoAcids20NaturalWithLotsOfCdkDesc.csv", row.names=1, na.string="")
aaDesc <- aaDesc[,-1]
# Remove descriptors with zero variance and lots of NA
aaDesc <- removeVarsWithZeroStdvAndLotsOfNa(aaDesc)
# Do PCA
aaDesc.pca <- prcomp(aaDesc, scale=T, center=T)
summary(aaDesc.pca)
plot(aaDesc.pca)
biplot(aaDesc.pca)
lines(c(-1.85,1.85),c(-7,7), col="green", lw=2)
lines(c(-7,7), c(1.35, -1.25), col="blue", lw=2)
aaDesc.pca.red <- predict(aaDesc.pca)[,1:4]
```

Properties Progress R Console

R workspace: "/Users/valentingeorgiev/bioclipse-workspace"  
Loaded R session: .RData  
Use load("file") and save.image("file") to manage your R sessions

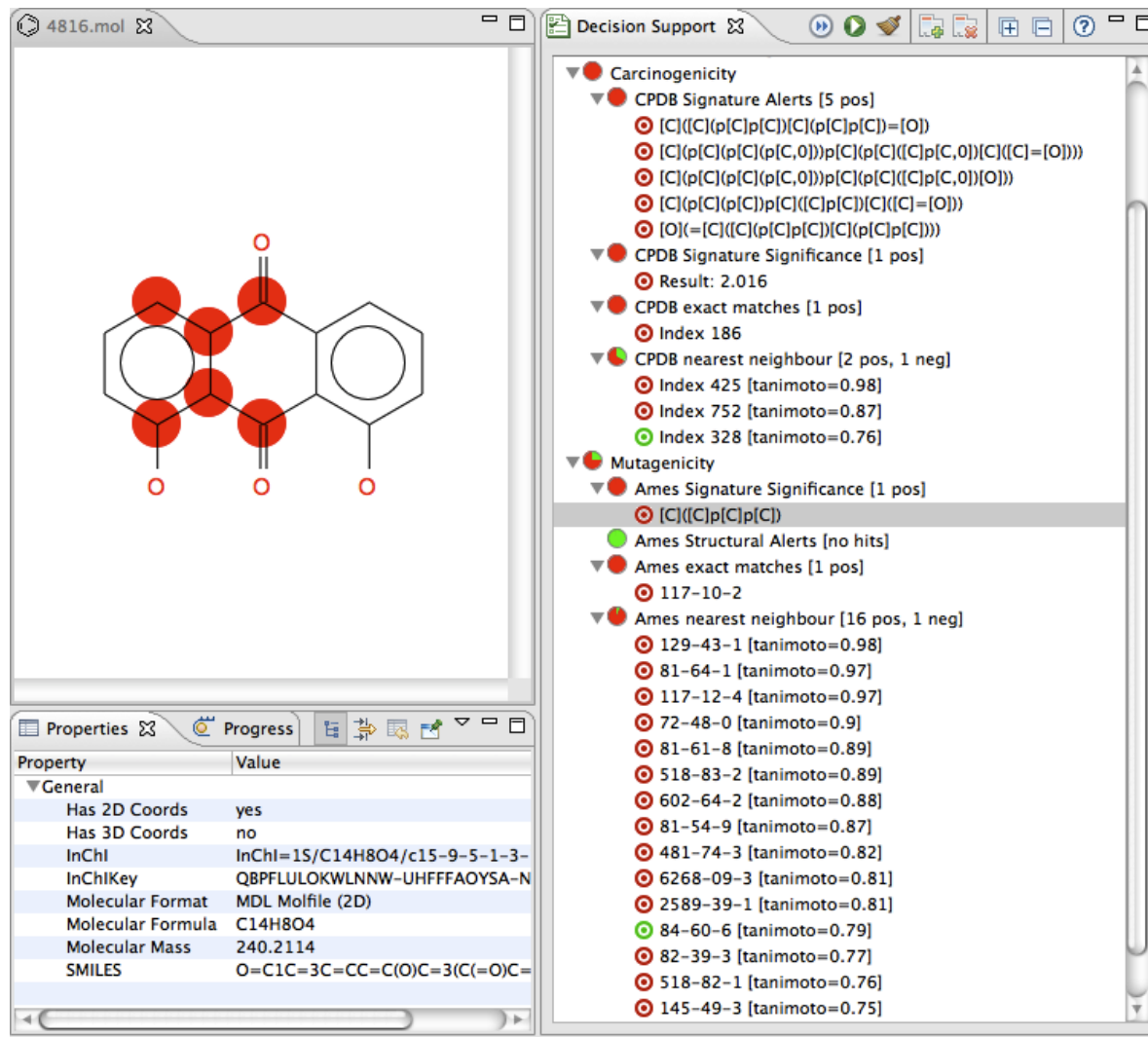
# Bioclipse (14)

## Decision support

- Bioclipse facilitates working with decision support systems, for example when predicting the susceptibility of a drug for certain patients. By sequencing the DNA of the patient (or e.g. a virus in the patient), it is possible to predict what drugs that would best attack the disease. An example of this is HIVPred, which is implemented as an XMPP cloud service and a plugin in Bioclipse to facilitate invocation.
- Bioclipse also has a feature for decision support in safety assessment with graphical views and editors for executing and integrating various computational models to predict the safety of chemical compounds.

# Bioclipse (15)

## Illustration: Chemical liability assessment in the Bioclipse workbench



**4816.mol**

**Decision Support**

- Carcinogenicity**
  - CPDB Signature Alerts [5 pos]
    - [C]([C](p[C]p[C])C)(p[C]p[C])=O
    - [C](p[C](p[C](p[C,0]))p[C](p[C]([C]p[C,0])[C]([C]=O)))
    - [C](p[C](p[C](p[C,0]))p[C](p[C]([C]p[C,0])[O]))
    - [C](p[C](p[C])p[C]([C]p[C])C)([C]=O
    - [O](=[C]([C](p[C]p[C])C)(p[C]p[C]))
  - CPDB Signature Significance [1 pos]
    - Result: 2.016
  - CPDB exact matches [1 pos]
    - Index 186
  - CPDB nearest neighbour [2 pos, 1 neg]
    - Index 425 [tanimoto=0.98]
    - Index 752 [tanimoto=0.87]
    - Index 328 [tanimoto=0.76]
- Mutagenicity**
  - Ames Signature Significance [1 pos]
    - [C]([C]p[C]p[C])
  - Ames Structural Alerts [no hits]
  - Ames exact matches [1 pos]
    - 117-10-2
  - Ames nearest neighbour [16 pos, 1 neg]
    - 129-43-1 [tanimoto=0.98]
    - 81-64-1 [tanimoto=0.97]
    - 117-12-4 [tanimoto=0.97]
    - 72-48-0 [tanimoto=0.9]
    - 81-61-8 [tanimoto=0.89]
    - 518-83-2 [tanimoto=0.89]
    - 602-64-2 [tanimoto=0.88]
    - 81-54-9 [tanimoto=0.87]
    - 481-74-3 [tanimoto=0.82]
    - 6268-09-3 [tanimoto=0.81]
    - 2589-39-1 [tanimoto=0.81]
    - 84-60-6 [tanimoto=0.79]
    - 82-39-3 [tanimoto=0.77]
    - 518-82-1 [tanimoto=0.76]
    - 145-49-3 [tanimoto=0.75]

**Properties**

Property	Value
<b>General</b>	
Has 2D Coords	yes
Has 3D Coords	no
InChI	InChI=1S/C14H8O4/c15-9-5-1-3-
InChIKey	QBPFLLULOKWLNW-UHFFFAOYSA-N
Molecular Format	MDL Molfile (2D)
Molecular Formula	C14H8O4
Molecular Mass	240.2114
SMILES	<chem>O=C1C=3C=CC=C(O)C=3(C(=O)C=</chem>

# Bioclipse (16)

Illustration: Chemical liability assessment for multiple structures with results in spreadsheet

The screenshot displays the Bioclipse software interface. On the left is the 'Bioclipse Navigator' panel showing a file tree with folders like 'ames-test', 'balloontest', 'ChiralSignatures', 'Drugbank', 'ds-data', 'Gists', 'MyExperiment', 'myproject', 'olaQQQ', 'Sample Data', 'test', 'Virtual', 'wee', and 'XMPP Service Bindings'. The 'Drugbank' folder is expanded, showing a list of files including 'possible', '1-100.sdf', '1108.cml', '236.mol', '238.mol', '240.mol', '242.mol', '249.mol', '4816.mol', 'DB00249.mol', 'DB01108.mol', 'DB04816.mol', 'drugbank.sdf', 'metapyrilene.cml', 'oxeladin.cml', 'oxolamine.cml', and 'part1.sdf'. The main window shows a spreadsheet titled 'part1.sdf' with columns: '2D-structure', 'DRUGBANK\_ID', 'Ames Consensus', 'Ames exact mat...', 'Ames nearest n...', 'CPDB nearest n...', 'Ames Signature...', and 'CPDB Signature...'. The spreadsheet contains 12 rows of data, each representing a chemical structure and its associated liability assessment results. The results are color-coded: green for 'NEGATIVE' and red for 'POSITIVE'. The bottom status bar indicates '0 items selected'.

	2D-structure	DRUGBANK_ID	Ames Consensus	Ames exact mat...	Ames nearest n...	CPDB nearest n...	Ames Signature...	CPDB Signature...
5		DB00139	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]	INCONCLUSIVE [...]	NEGATIVE [1 hits]	NEGATIVE [1 hits]
6		DB00140	POSITIVE [1 hits]	INCONCLUSIVE [...]	POSITIVE [1 hits]	INCONCLUSIVE [...]	POSITIVE [1 hits]	POSITIVE [1 hits]
7		DB00141	NEGATIVE [1 hits]	INCONCLUSIVE [...]	INCONCLUSIVE [...]	INCONCLUSIVE [...]	NEGATIVE [1 hits]	NEGATIVE [1 hits]
8		DB00142	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]
9		DB00143	POSITIVE [1 hits]	POSITIVE [1 hits]	POSITIVE [1 hits]	INCONCLUSIVE [...]	POSITIVE [1 hits]	NEGATIVE [1 hits]
10		DB00144	NEGATIVE [1 hits]	INCONCLUSIVE [...]	INCONCLUSIVE [...]	INCONCLUSIVE [...]	NEGATIVE [1 hits]	NEGATIVE [1 hits]
11		DB00145	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]	INCONCLUSIVE [...]	NEGATIVE [1 hits]	NEGATIVE [1 hits]
12		DB00146	NEGATIVE [1 hits]	INCONCLUSIVE [...]	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]	NEGATIVE [1 hits]

# Захвалница

Део материјала презентације је преузет са адресе  
<http://research.microsoft.com/en-us/projects/chem4word/>

Део материјала презентације је преузет са адресе  
<http://www.xml-cml.org/>

Део материјала презентације је преузет са адресе  
<http://www.bioclipse.net/>