3RD EDITION | ADVANCED TRAINING COURSE IN **MOLECULAR BIOENGINEERING**

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Peptides: Where Chemistry Meets Code

Dr. William J. Zamora R

UCR
UNIVERSIDAD DE COSTA RICA

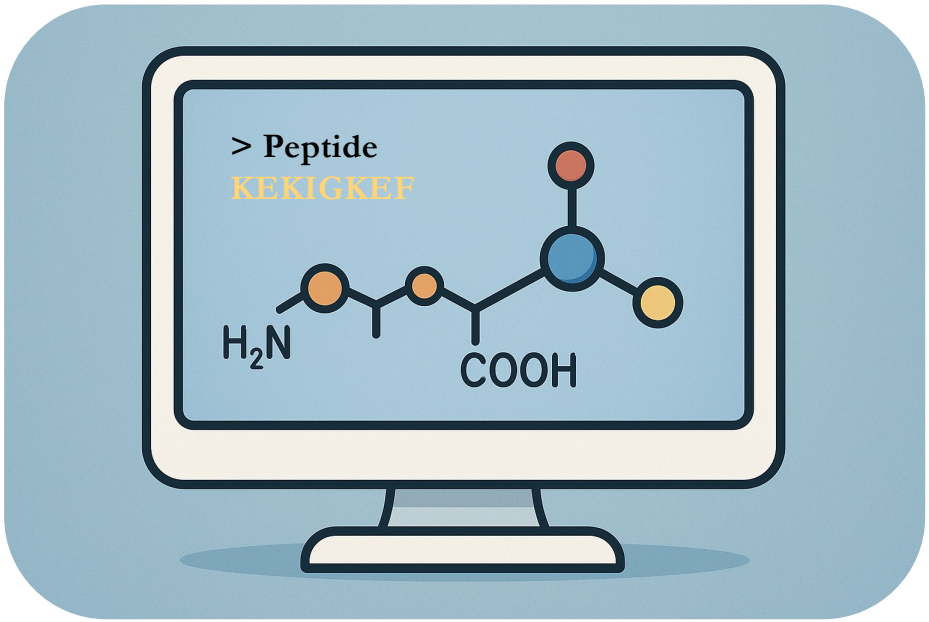
Bio^3
COMPUTACIONAL BIOPHYSICS, BIOLOGY AND BIOINFORMATICS GROUP

aToxCIA
Laboratorio de Toxicología Computacional e Inteligencia Artificial
LEBI Laboratorio de Ensayos Biológicos

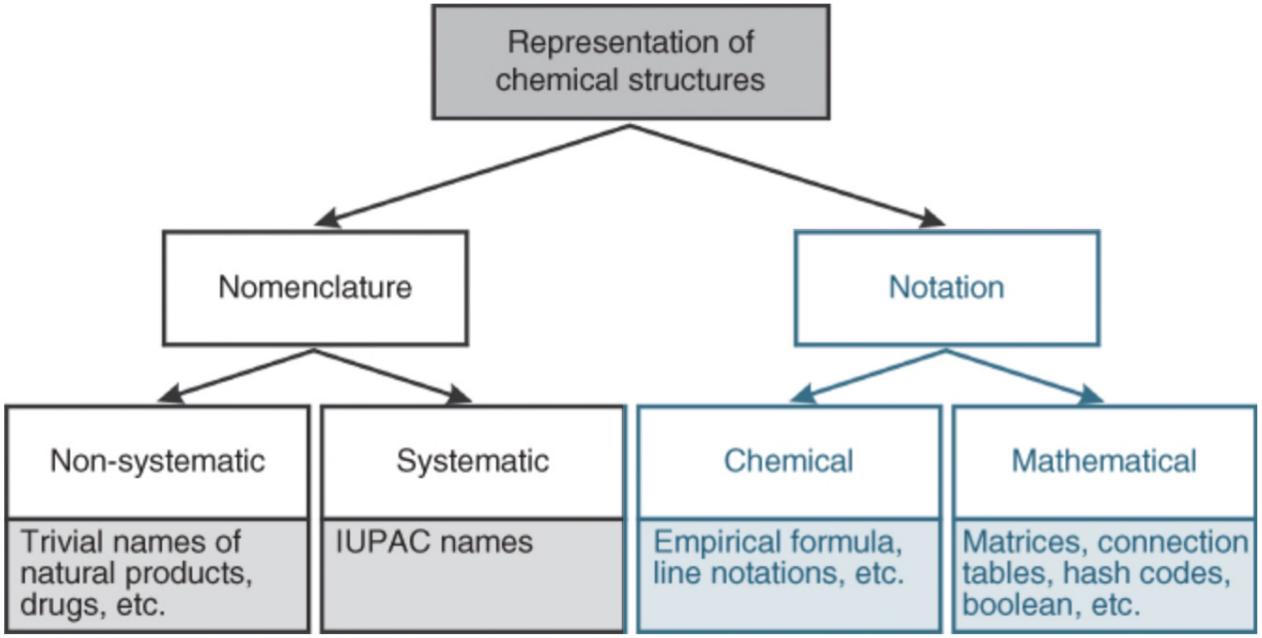
UNA
UNIVERSIDAD NACIONAL
COSTA RICA

Introduction

STRUCTURES ON COMPUTERS



Molecules on Computers

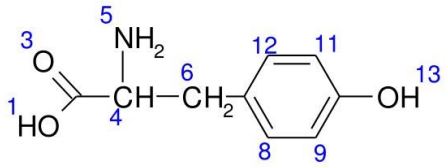


Molecules on Computers

Notations

Chemical Notations

2.1.2 Line Notations



(1950s to the 1970s)

- Wiswesser Line Notation (WLN) (obsolete)

  QVYZ1R DQ

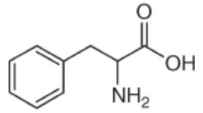
  — **Q** = OH, **V** = -CO-, **Z** = -NH2, **R** = benzene

- ROSDAL (Beilstein)

  Representation Of Structure Diagram Arranged Linearly

  10-2=30,2-4-5N,4-6-7=-12-7,10-130

(1980s – 2000s)

Type	Nomenclature/notation
Trivial name	Phenylalanine
Systematic name (IUPAC)	2-Amino-3-phenylpropanoic acid
Structure diagram	
Empirical formula	$C_9H_{11}NO_2$
SMILES	NC(Cc1ccccc1)C(O) = O
SLN	C[1]H:CH:CH:CH:CH:C(:@1)CH2CH(NH2)C(=O)OH
InChI	InChI = 1S/C9H11NO2/c10-8(9(11)12)6-7-4-2-1-3-5-7/h1-5,8H,6,10H2,(H,11,12)
InChIKey	COLNVLDHVKWLRT-UHFFFAOYSA-N

Wiswesser, William J. (**1951**). Simplified chemical coding for automatic sorting and printing machinery. Willson Products Inc., Reading, PA.

Barnard, J.M., Jochum, C.J., and Welford, S.M. (**1989**) ACS Symp. Ser., 400, 76–81.

Molecules on Computers

SMILES (Simplified Molecular Input Line Entry System)

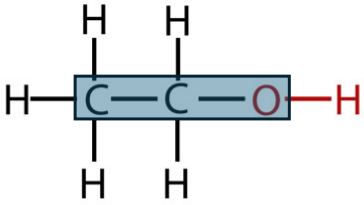
Organic Subset

Element	Valence
B	3
C	4
N	3 or 5
O	2
P	3 or 5
S	2, 4 or 6
halogens	1
*	unspecified

SMILES	Name
CC	ethane
CCO	ethanol
NCCCC	n-butylamine
CCCCN	n-butylamine

SMILES	
C-C	same as: CC
C-C-O	same as: CCO
C-C=C-C	same as: CC=CC

Ethanol



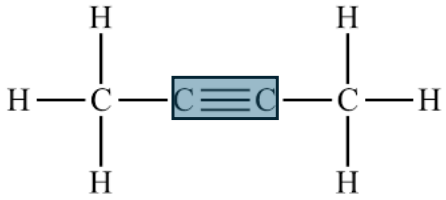
Structural formula

Molecules on Computers

SMILES (Simplified Molecular Input Line Entry System)

Bonds

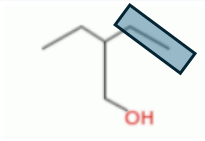
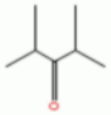
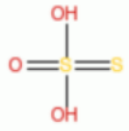
SMILES	Name
C=C	ethene
C#N	hydrogen cyanide
CC#CC	2-butyne
CCC=O	propanol
[Rh-](Cl)(Cl)(Cl)(Cl)$[Rh-](Cl)(Cl)(Cl)Cl	octachlorodirhenate (III)



Molecules on Computers

SMILES (Simplified Molecular Input Line Entry System)

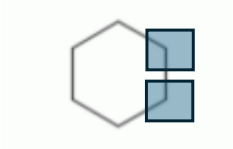
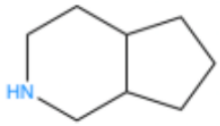
Branches

Depiction	SMILES	Name
	CCC(CC)CO	2-ethyl-1-butanol
	CC(C)C(=O)C(C)C	2,4-dimethyl-3-pentanone
	OS(=O)(=S)O	thiosulfate

Molecules on Computers

SMILES (Simplified Molecular Input Line Entry System)

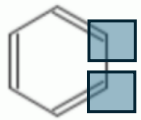
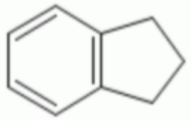
Rings

Depiction	SMILES	Name
	C1CCCCC1	cyclohexane
	N1CC2CCCC2CC1	perhydroisoquinoline

Molecules on Computers

SMILES (Simplified Molecular Input Line Entry System)

Aromaticity

Depiction	SMILES	Name
	c1ccccc1 C1=CC=CC=C1	benzene
	c1ccc2CCCc2c1 C1=CC=CC(CCC2)=C12	indane

Molecules on Computers

Structural Data Files (sdf)

HEADER BLOCK (1. FORMULA, 2.PROGRAM, 3. COMMENT)

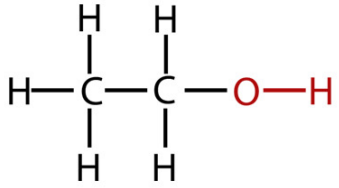
```
C2H6O
APtclcactv0601210243D 0   0.00000     0.00000
  9  8  0  0  0  0  0  0  0  0999 V2000
    0.0072   -0.5687   -0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2854    0.2499   -0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9857   -0.1365   -0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.1304    0.3147    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.0392   -1.1972    0.8900 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0392   -1.1972   -0.8900 H   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3175    0.8784    0.8900 H   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3175    0.8784   -0.8900 H   0  0  0  0  0  0  0  0  0  0  0  0
   -2.1422   -0.4239   -0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  4  1  0  0  0  0
  1  5  1  0  0  0  0
  1  6  1  0  0  0  0
  2  7  1  0  0  0  0
  2  8  1  0  0  0  0
  2  9  1  0  0  0  0
  3  4  1  0  0  0  0
M  END
$$$$
```

COUNT LINE BLOCK

ATOM BLOCK

BOND BLOCK

```
OpenBabel0601210100D
  9  8  0  0  0  0  0  0  0  0999 V2000
    1.0195    0.0236   -0.0982 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5341    0.0228   -0.0910 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0144   -0.9180    0.8583 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.6335    0.7322   -0.8363 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.6279    0.2939    0.8880 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.6349   -0.9748   -0.3308 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.9260    1.0118    0.1654 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.9215   -0.2573   -1.0749 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.6732   -0.6549    1.7301 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  4  1  0  0  0  0
  1  5  1  0  0  0  0
  1  6  1  0  0  0  0
  2  3  1  0  0  0  0
  2  7  1  0  0  0  0
  2  8  1  0  0  0  0
  3  9  1  0  0  0  0
M  END
$$$$
```



Molecules on Computers

Molfile2 (.mol2)

@<TRIPOS>MOLECULE

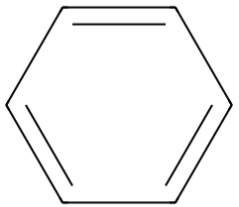
1.The first data line is the **name of the molecule**

2. The second data line contains the number of **atoms, bonds, substructures, features, and sets associated with the molecule**

3. The third data line is the molecule type. The supported types are **SMALL, BIOPOLYMER, PROTEIN, NUCLEIC_ACID, and SACCHARIDE**

4. The fourth data line tells the type of **charges** associated with the molecule. The supported types are NO_CHARGES, DEL_RE, GASTEIGER, GAST_HUCK, HUCKEL, PULLMAN, GAUSS80_CHARGES, AMPAC_CHARGES, MULLIKEN_CHARGES, DICT_CHARGES, MMFF94_CHARGES, and USER_CHARGES

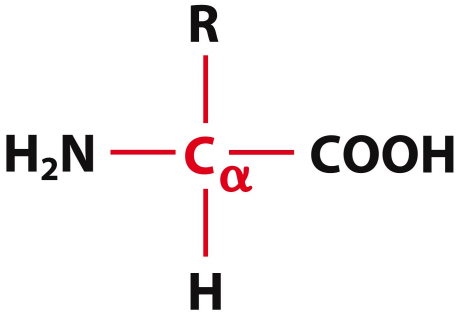
```
@<TRIPOS>MOLECULE
benzene

 12 12 0 0 0
SMALL
NO_CHARGES
****
Any comment about the molecule goes here. No need for a pound sign here

@<TRIPOS>ATOM
      1 C          -0.7600    1.1691   -0.0005 C.ar    1  BENZENE       0.000
      2 C           0.6329    1.2447   -0.0012 C.ar    1  BENZENE       0.000
      3 C           1.3947    0.0765    0.0004 C.ar    1  BENZENE       0.000
      4 C           0.7641   -1.1677    0.0027 C.ar    1  BENZENE       0.000
      5 C          -0.6288   -1.2432    0.0001 C.ar    1  BENZENE       0.000
      6 C          -1.3907   -0.0751   -0.0015 C.ar    1  BENZENE       0.000
      7 H          -1.3536    2.0792    0.0005 H       1  BENZENE       0.000
      8 H           1.1243    2.2140   -0.0028 H       1  BENZENE       0.000
      9 H           2.4799    0.1355   -0.0000 H       1  BENZENE       0.000
     10 H           1.3576   -2.0778    0.0063 H       1  BENZENE       0.000
     11 H          -1.1202   -2.2126   -0.0005 H       1  BENZENE       0.000
     12 H          -2.4759   -0.1340   -0.0035 H       1  BENZENE       0.000
@<TRIPOS>BOND
     1     1     2   ar
     2     2     3   ar
     3     3     4   ar
     4     4     5   ar
     5     5     6   ar
     6     1     6   ar
     7     1     7    1
     8     2     8    1
     9     3     9    1
    10     4    10    1
    11     5    11    1
    12     6    12    1
```

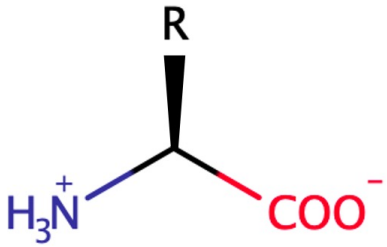


Peptides on Computers

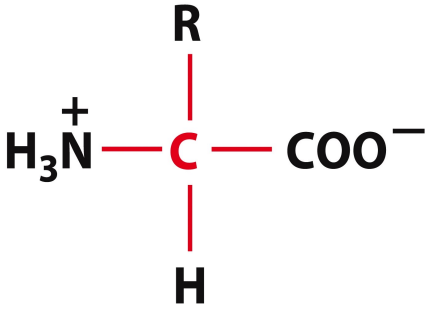
L - AMINO ACIDS

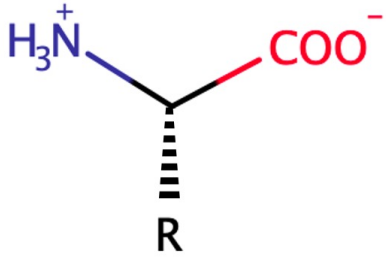
Neutral: non-ionic





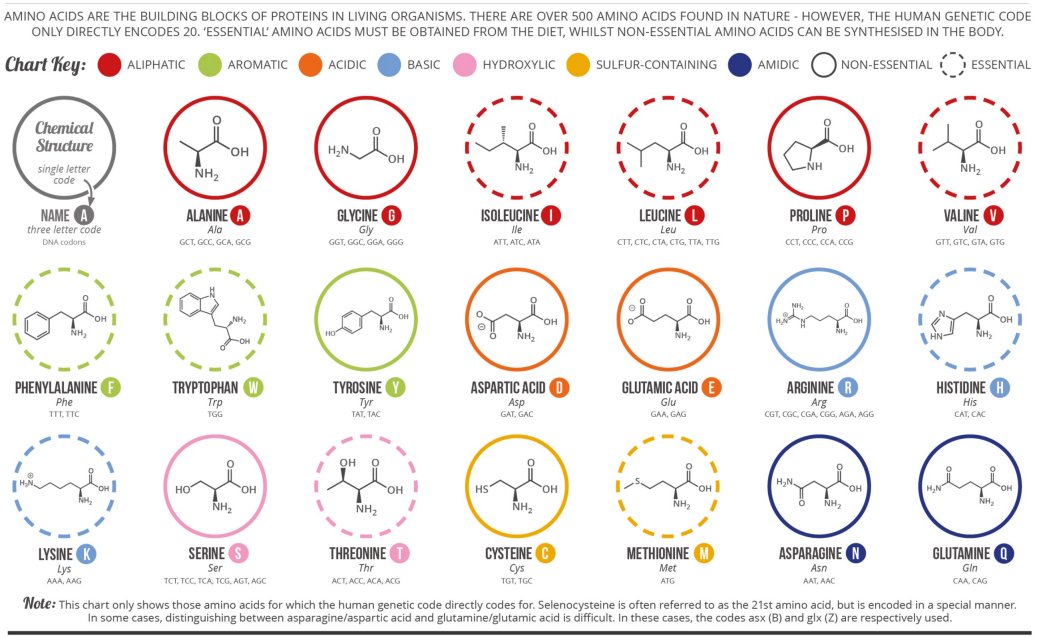
Neutral: zwitterion





Peptides on Computers

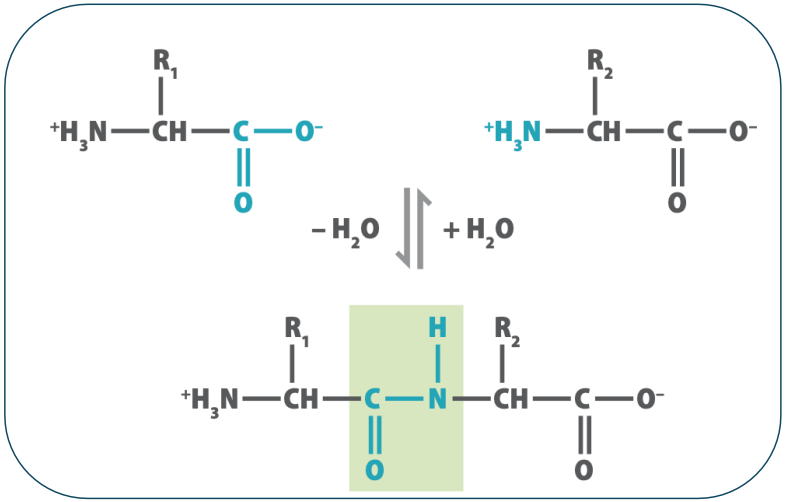
L - AMINO ACIDS

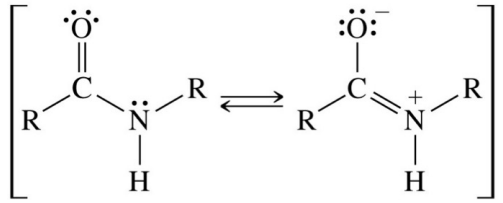


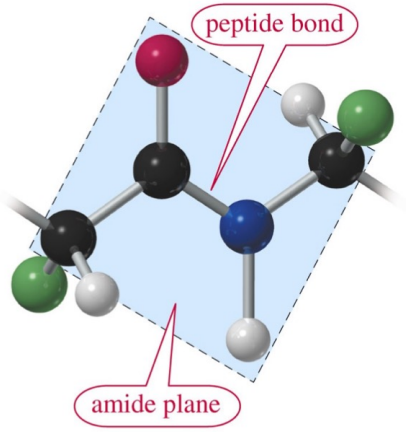
IUPAC amino acid code	Three letter code	Amino acid
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Peptides on Computers

PEPTIDE BOND

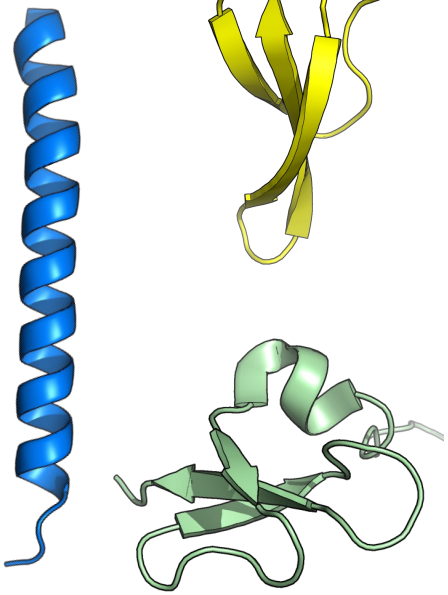


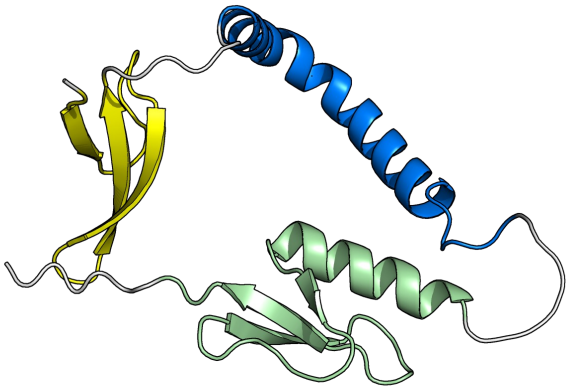




Peptides on Computers

PEPTIDES





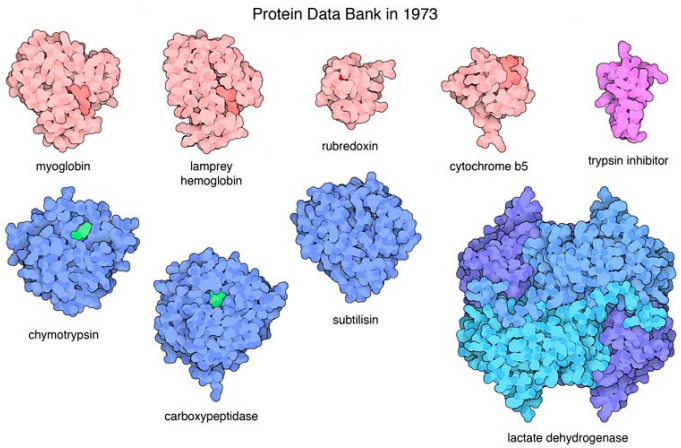
"Until we have the structural details that we can really appreciate exactly what the proteins are doing"

M. Williamson; U. Sheffield

Peptides on Computers

PROTEIN DATABASE (PDB)

RCSB PDB
PROTEIN DATA BANK



Established in **1971** at **Brookhaven National Laboratory** (leadership of Walter Hamilton) and originally contained **7 structures**.

1973 lead by Tom Koetzle and Joel Sussman in **1994**.

In **1998** Led by Helen M. Berman at the Research Collaboratory for Structural Bioinformatics (RCSB) in response to an RFP and a lengthy review process (together with Rutgers + UCSD/SDSC + CARB/NIST)

In **2003**, the Worldwide Protein Data Bank (wwPDB) was formed to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community.

In **2005** CARB/NIST left RCSB

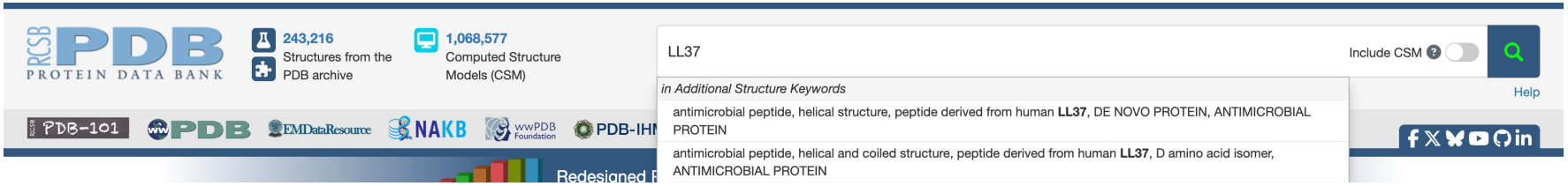
Stephen K. Burley became RCSB PDB Director in **2014**. Helen M. Berman currently serves as Director Emerita.

In **2019** UCSF joined RCSB PDB

In **2021** *50th Anniversary* of the PDB with symposia, materials, and more.

Peptides on Computers

PROTEIN DATABASE (PDB)



1 to 1 of 1 **Structure** | Page 1 of 1 | 25 | Sort by ↓ Score

Explore in 3D

5XNG | **pdb_00005xng**

Download File | View File

EFK17A structure in Microgel MAA60

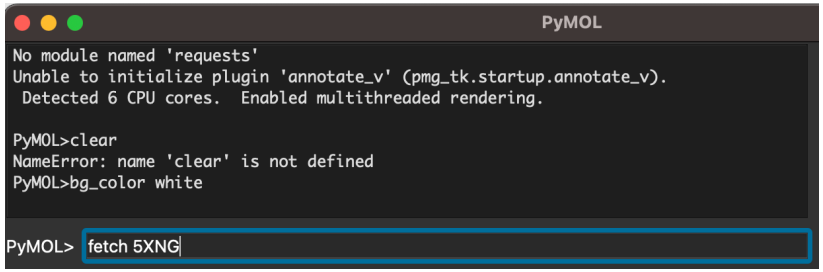
Datta, A., Bhunia, A.

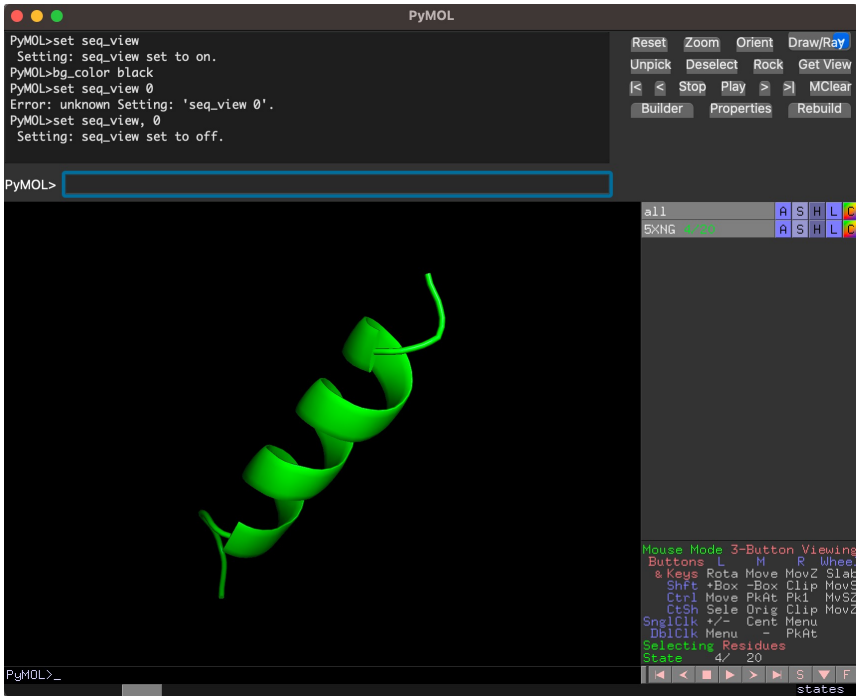
(2017) ACS Appl Mater Interfaces **9**: 40094-40106

Released	2018-04-18
Method	SOLUTION NMR
Organisms	Homo sapiens
Macromolecule	Cathelicidin antimicrobial peptide (protein)

Peptides on Computers

PROTEIN DATABASE (PDB)

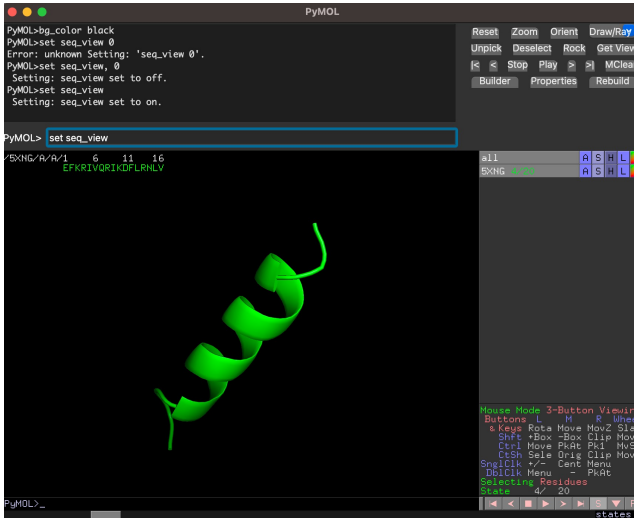




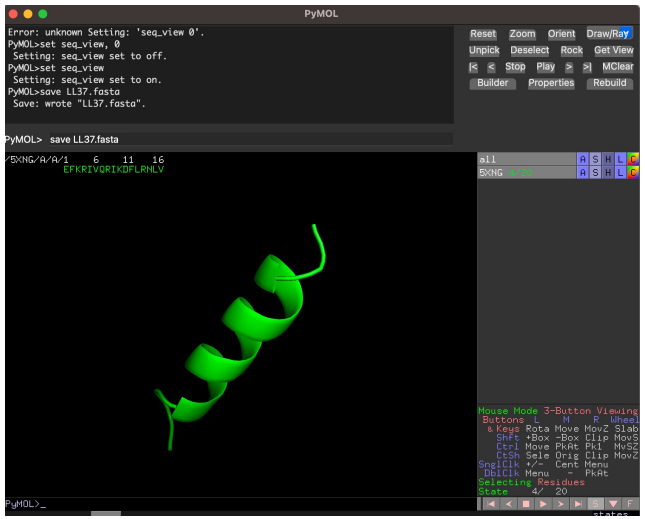
Peptides on Computers

PROTEIN DATABASE (PDB)

set seq_view



save file_name.fasta



LL37.fasta

Peptides on Computers

FASTA Format

>Peptide 1; LL37 derived
EFKRIVQRIKDFLRNLV

- Created in **1985** by David J. Lipman and William R. Pearson
- Text-based, bioinformatic data format used to store nucleotide or amino acid sequences
- Is a shortening of "FAST-All"
- An evolution of previous tools "FAST-P" (protein) and "FAST-N" (nucleotide)
- Sequences in FASTA are represented by single alphabetical codes outlined in IUPAC

IUPAC amino acid code	Three letter code	Amino acid
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Peptides on Computers

FASTA Format

```
> library('seqinr')
> ll37=read.fasta(file = "LL37.fasta",seqtype = "AA")
> ll37
$`5XNG_A`
 [1] "E" "F" "K" "R" "I" "V" "Q" "R" "I" "K" "D" "F" "L" "R" "N" "L" "V"
attr(,"name")
[1] "5XNG_A"
attr(,"Annot")
[1] ">5XNG_A"
attr(,"class")
[1] "SeqFastaAA"

> length(ll37[[1]])
[1] 17
> ll37[[1]][1:17]
 [1] "E" "F" "K" "R" "I" "V" "Q" "R" "I" "K" "D" "F" "L" "R" "N" "L" "V"
```

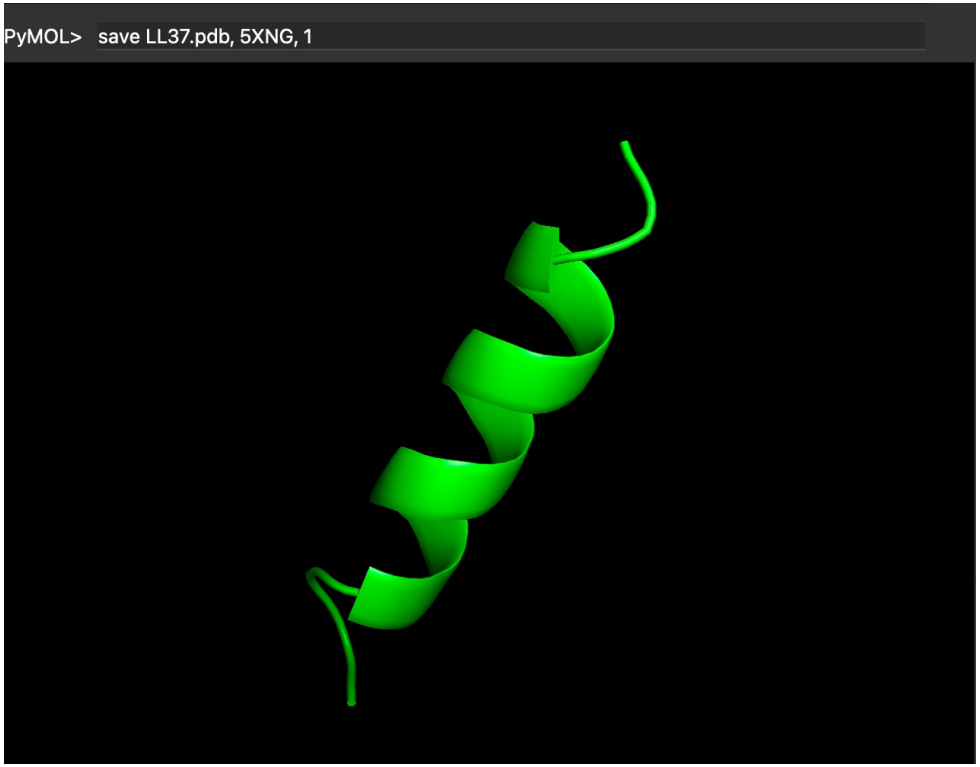
Peptides on Computers

FASTA Format

```
> ll37[[1]][1:17]
 [1] "E" "F" "K" "R" "I" "V" "Q" "R" "I" "K" "D" "F" "L" "R" "N" "L" "V"
> seq=ll37[[1]][1:17]
> is.element(seq,"E")
 [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE FALSE FALSE
> which(is.element("E",seq))
[1] 1
>
> is.element(seq,"R")
 [1] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[13] FALSE  TRUE FALSE FALSE FALSE
> which(is.element(seq,"R"))
[1]  4  8 14
```

Peptides on Computers

PDB files (.pdb)



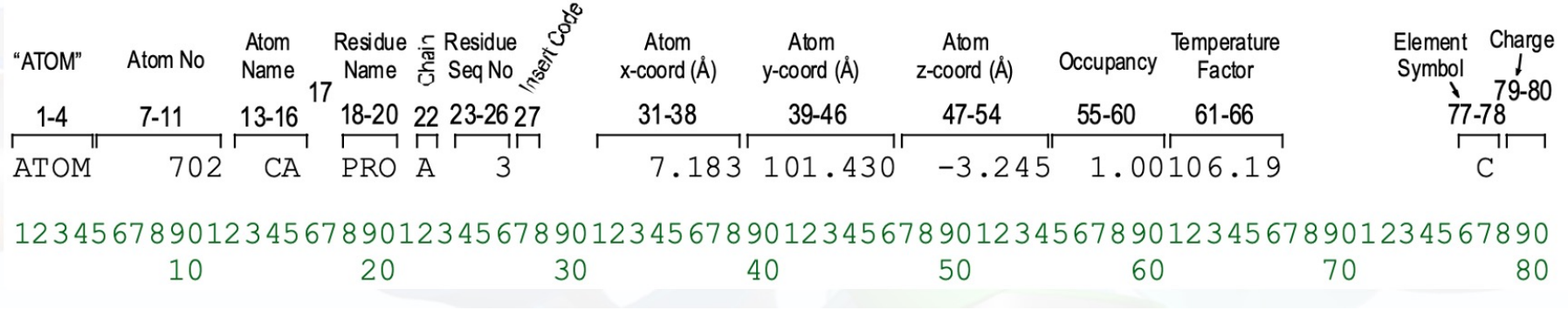
Peptides on Computers

PDB files (.pdb)

```
ATOM      1  N   GLU A   1       1.329   0.000   0.000  1.00  4.00      A    N
ATOM      2  CA  GLU A   1       2.093  -0.001  -1.242  1.00 74.25      A    C
ATOM      3  C   GLU A   1       2.494  -1.421  -1.633  1.00 55.24      A    C
ATOM      4  O   GLU A   1       2.954  -1.663  -2.749  1.00 42.32      A    O
ATOM      5  CB  GLU A   1       3.341   0.872  -1.101  1.00 52.41      A    C
ATOM      6  CG  GLU A   1       3.135   2.089  -0.214  1.00 61.12      A    C
ATOM      7  CD  GLU A   1       4.383   2.942  -0.096  1.00 52.52      A    C
ATOM      8  OE1 GLU A   1       4.482   3.716   0.880  1.00 74.33      A    O
ATOM      9  OE2 GLU A   1       5.260   2.838  -0.979  1.00 32.22      A    O
ATOM     10  HA  GLU A   1       1.464   0.409  -2.018  1.00 72.33      A    H
```

Peptides on Computers

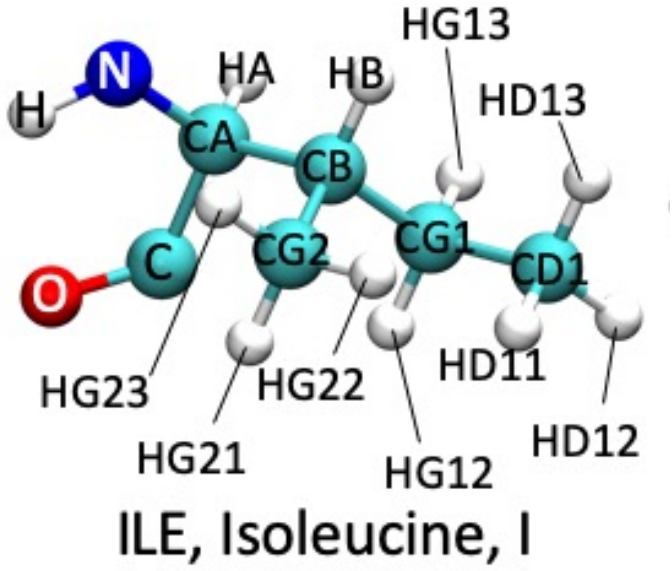
PDB files (.pdb)



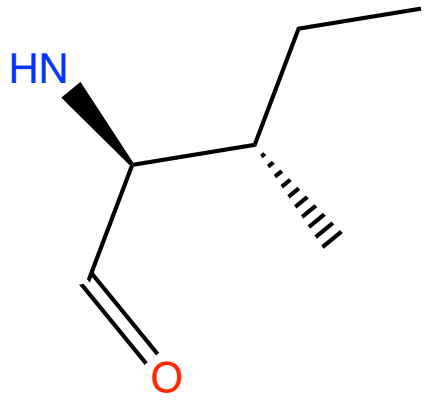
Peptides on Computers

PDB files (.pdb)

_L_ - AMINO ACIDS

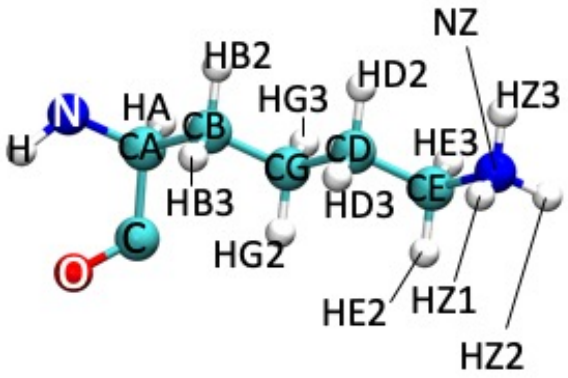


ILE, Isoleucine, I

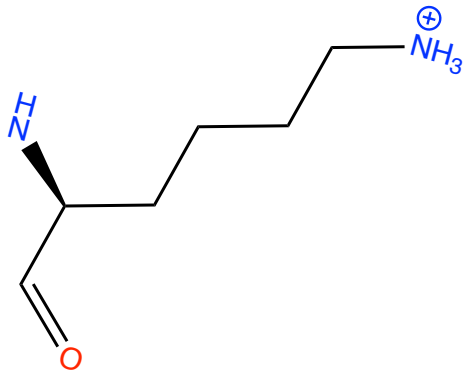


Peptides on Computers

PDB files (.pdb) **_L_ - AMINO ACIDS**



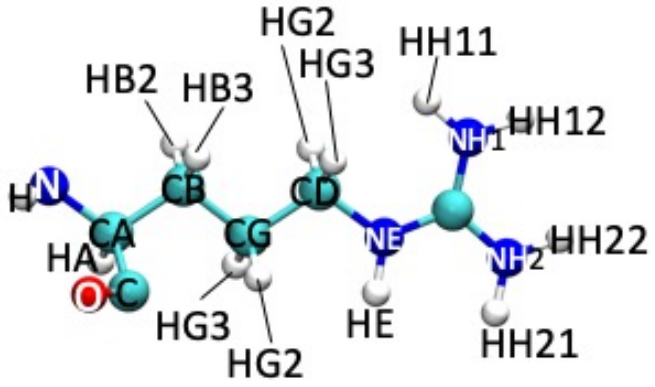
LYS, Lysine, K



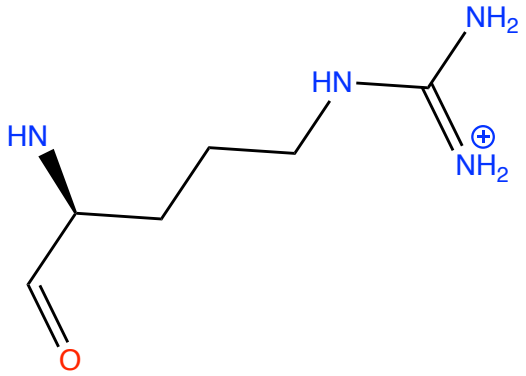
Peptides on Computers

PDB files (.pdb)

L - AMINO ACIDS



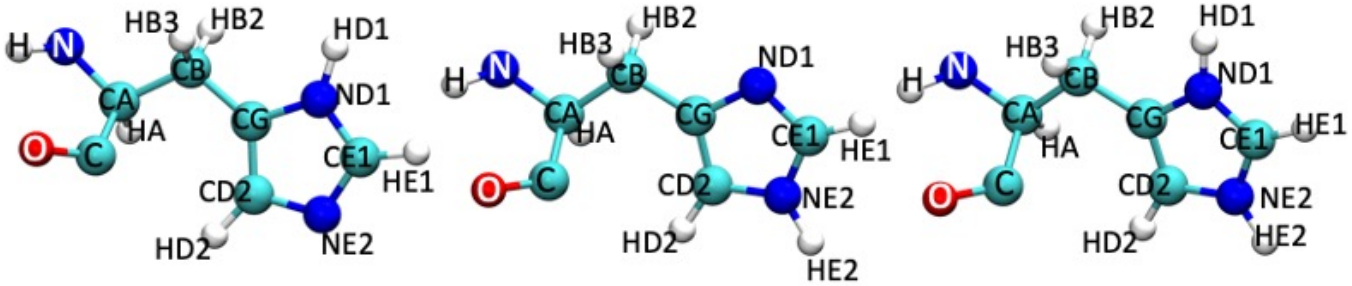
ARG, Arginine, R



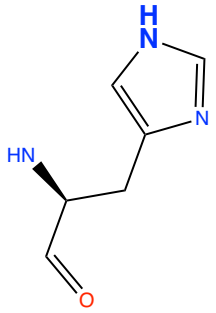
Peptides on Computers

PDB files (.pdb)

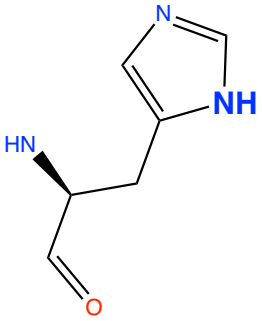
L - AMINO ACIDS



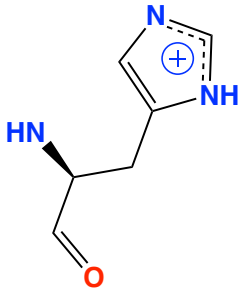
HID, Histidine*, H



HIE, Histidine, H

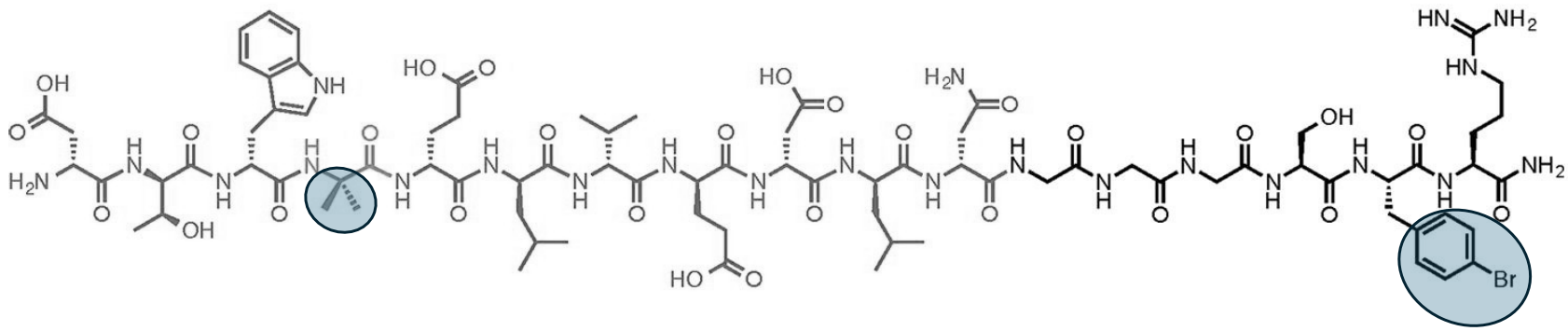


HIP, Histidine*, H



Introduction

HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation



CC(C)C[C@@H](C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](CC(=O)N)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@](C)(C(=O)N[C@@H](CC1=CNC2=C1C=CC=C2)C(=O)N[C@@H]([C@@H](C)O)C(=O)N[C@@H](CC(=O)O)C(=O)N)[X11])NC(=O)[C@H](CC(=O)N)NC(=O)CNC(=O)CNC(=O)CNC(=O)[C@H](CS)NC(=O)[C@H](CC3=CC=C(C=C3)O)NC(=O)[C@H](CCC

Peptides on Computers

HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation

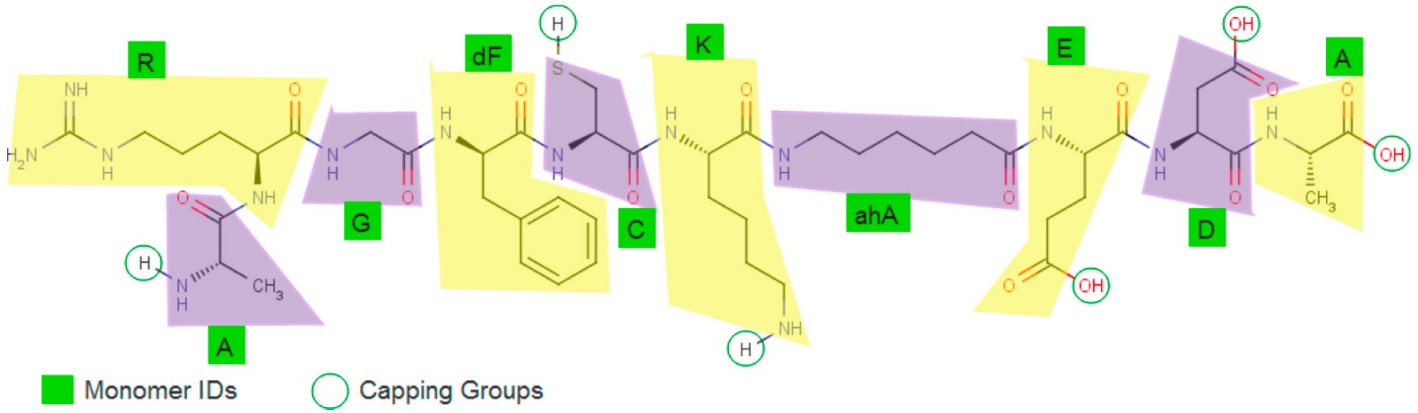
"SMILES like"

Structure hierarchy consists of four levels

1. Complex Polymer
2. Simple Polymer
3. Monomer
4. Atom

Peptides on Computers

HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation

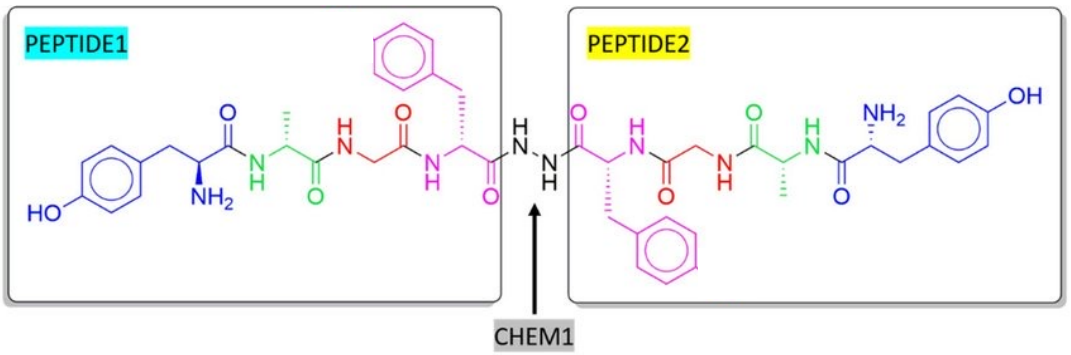


PEPTIDE polymer

A.R.G.[dF].C.K.[ahA].E.D.A.

Peptides on Computers

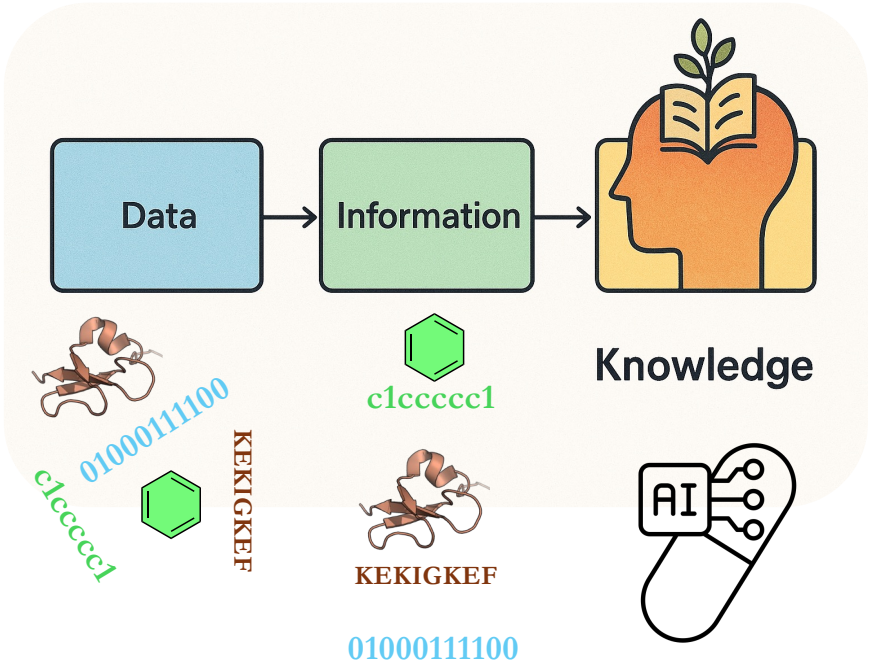
HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation



- Dollar sign "$" to separate the major sections of the complex polymer notation
- Curly brackets "{}" to enclose simple polymer notation in the simple polymer list and to enclose polymer attribute in polymer attribute list
- Vertical pipe "|" to separate simple polymers, connections, hydrogen bonds, and polymer attributes
- Comma "," to separate the three components in a connection or hydrogen bond string
- Dash "-" to connect the source and target in a connection or hydrogen bond
- Colon ":" to separate monomer position from attachment point or hydrogen bond code, and attribute name from attribute value

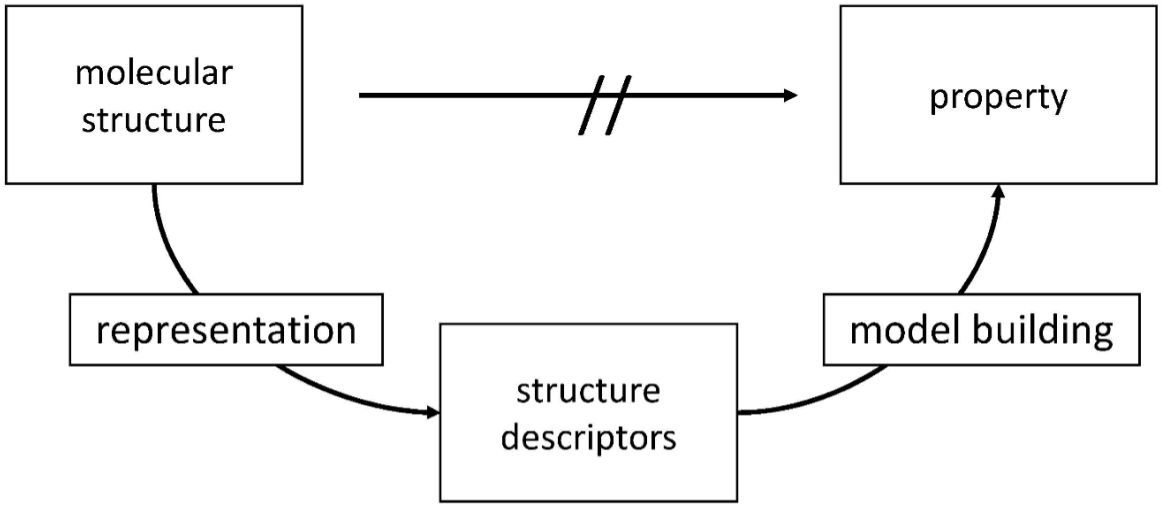
Quantitative Structure-Activity/Property Relationship
QSAR & QSPR

INDUCTIVE LEARNING



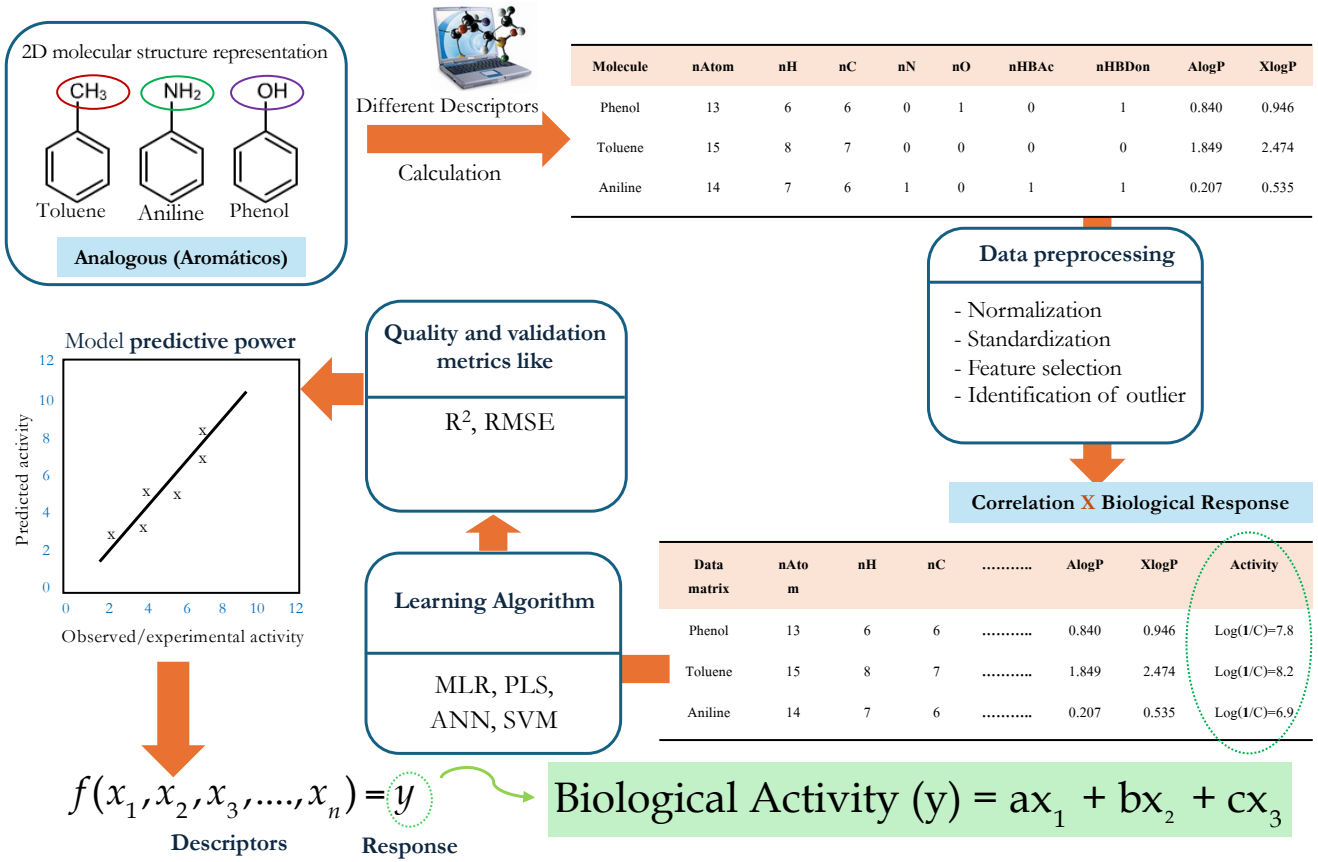
Quantitative Structure-Activity/Property Relationship
QSAR & QSPR

The Basic Approach to QSAR/QSPR



*Relationships between many chemical and, particularly, **activity/property** data of compounds and their structure are too complex to be directly predicted on first principles*

Quantitative Structure-Activity/Property Relationship
QSAR & QSPR



Quantitative Structure-Activity/Property Relationship
QSAR & QSPR

The QSAR Assessment Framework

Principle 1 – Defined endpoint

- A model can be used for **regulatory applications** if the **endpoint** predicted is clearly **defined** and considered to be of **regulatory relevance**

- Experimental data of compounds must be adequately and transparently reported
(e.g. CAS number and/or SMILES/HELM)

- Include all available **experimental details**

- A description of the **data curation** procedure

- **Commercial** models may not disclose all information for business reasons
(It is up to regulators to determine the minimum level of transparency that is acceptable)

Quantitative Structure-Activity/Property Relationship
QSAR & QSPR

The QSAR Assessment Framework

Principle 2 – Unambiguous algorithm

- **Unambiguous** description of the algorithm leading to the prediction (model equation)
- Description of **all inputs** and options should be provided (assessors to reproduce calculations)
- Model **accessibility** is evaluated (ensure the correct and reproducible use of the model)
- Preference for models that are **commercially** or **freely available**
- Complex or commercial models, not all information may be available (It is up to regulators)

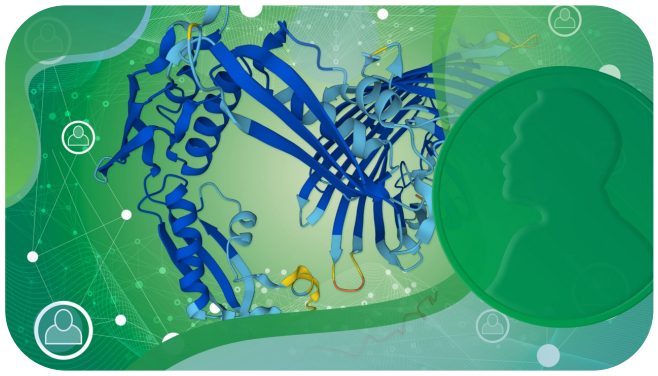
Quantitative Structure-Activity/Property Relationship
QSAR & QSPR

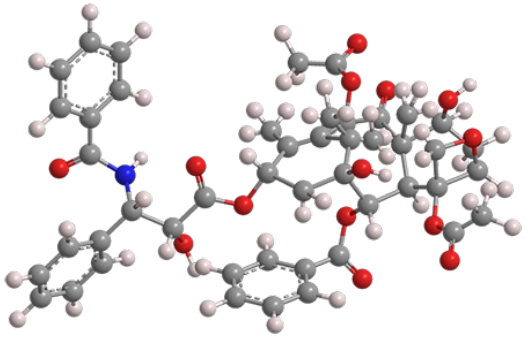
The QSAR Assessment Framework

Principle 3 – Defined domain of applicability (AD)

- **AD** is the response and **chemical structure space** for which a model can make predictions with a certain **degree of reliability**.
- QAF focuses on verifying that the AD is **clearly defined** and described by the model developers

2024 Nobel Prize for Chemistry





Placlitaxel

Quantitative Structure-Activity/Property Relationship
QSAR & QSPR

The QSAR Assessment Framework

Principle 4 – Goodness-of-fit, robustness and predictivity

Goodness-of-fit and robustness: ***internal validation***
(the performance of the model when predicting structures used for its development)

Predictivity of a model: ***external validation***
(performance of the model when predicting structures not used for its development)

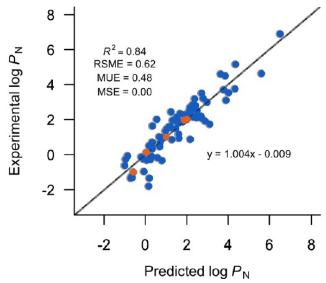


Fig. 4 Comparison between experimental and predicted *n*-octanol/water log P_N using the MLR-3 model for the training (blue) and test (orange) set

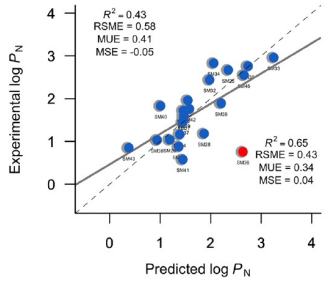


Fig. 5 Comparison between experimental and the multiple linear regression method for determining the *n*-octanol/water log P_N for the SAMPL7 dataset. Red point illustrates the outlier founded in our method. Top left, statistical analyses are shown for all compounds and bottom right, after exclusion of SM36

The QSAR Assessment Framework

Principle 5 – Mechanistic interpretation

- **Recommends** that a QSAR model should be associated with a mechanistic interpretation
- Knowledge of the **relationship** between the predictive property and toxicology
- **Structural fragments and descriptors** included in the model algorithm should be explained in relation to the property being predicted.

Thank you for your attention

Graciès!

Merci!

Gracias!

Thanks!

Obrigado!

william.zamoraramirez@ucr.ac.cr