Chem/bioinformatics visualization projects at Charles University

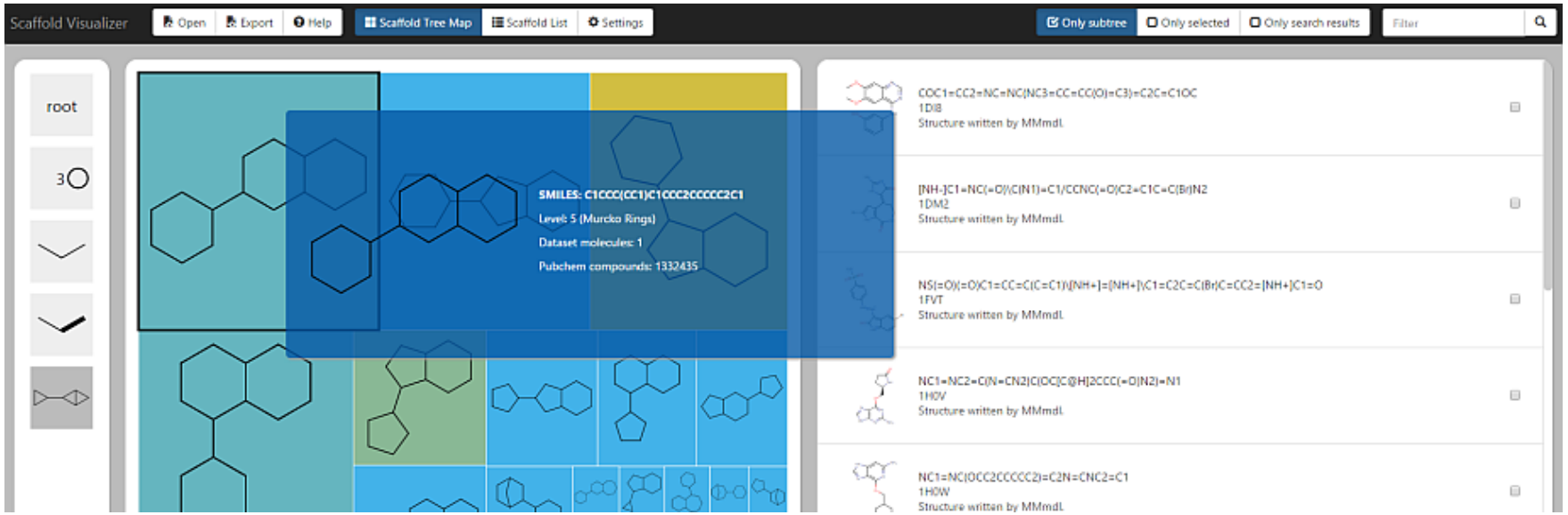
David Hoksza

SIRET Research Group

Department of Software Engineering

Faculty of Mathematics and Physics

Charles University in Prague



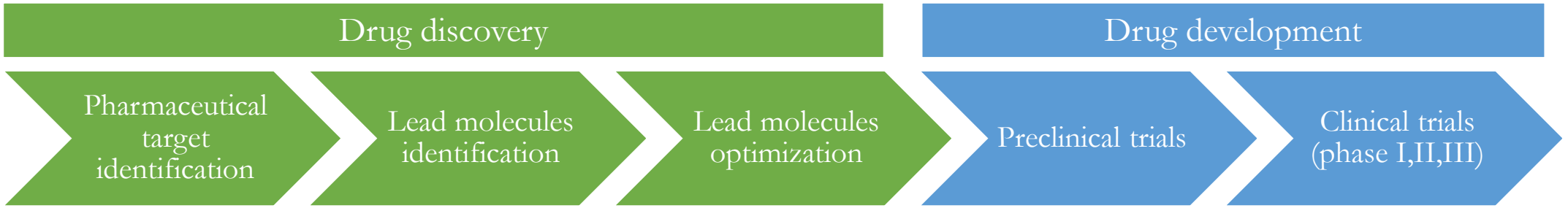
Scaffvis

Scaffold-based hierarchical visualization of large datasets on the background of chemical space

Cheminformatics

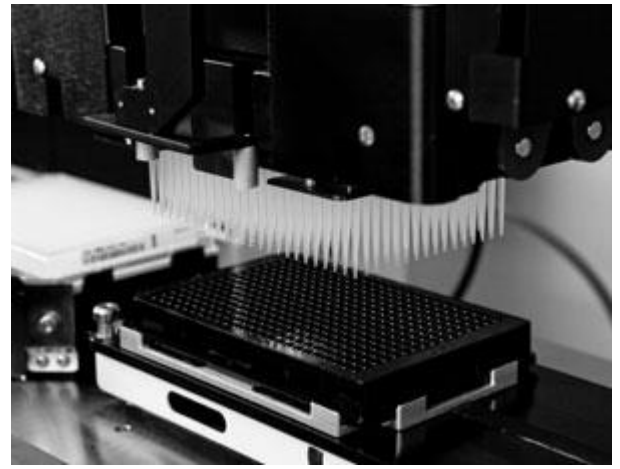
- Chemical informatics was defined by F. Brown as

"mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization"



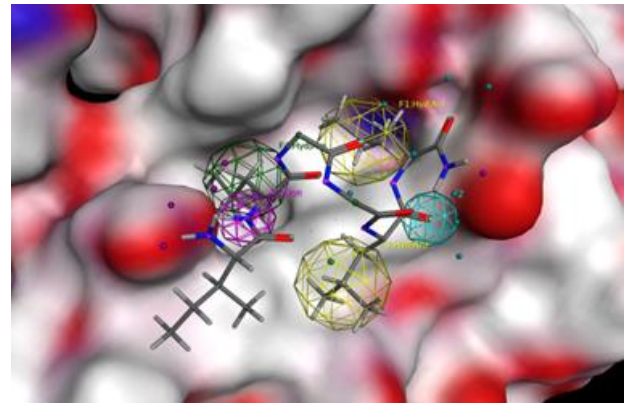
Computer-based identification of bioactive compounds

- High-throughput **screening** (HTS)
  - Laboratory method capable to test thousand of compounds in parallel when searching for **bioactive candidates** (leads)



source: Hybrigenics Services

- **High-throughput virtual screening (HTVS)**
  - Computational method for analyzing virtual libraries of chemical compounds
    - Capability of quickly testing millions of compounds
    - No need to physically own the compounds
    - Ability to test virtual, not yet synthetized, compounds
    - Less reliable than classical biological screening



Chemical datasets visualization

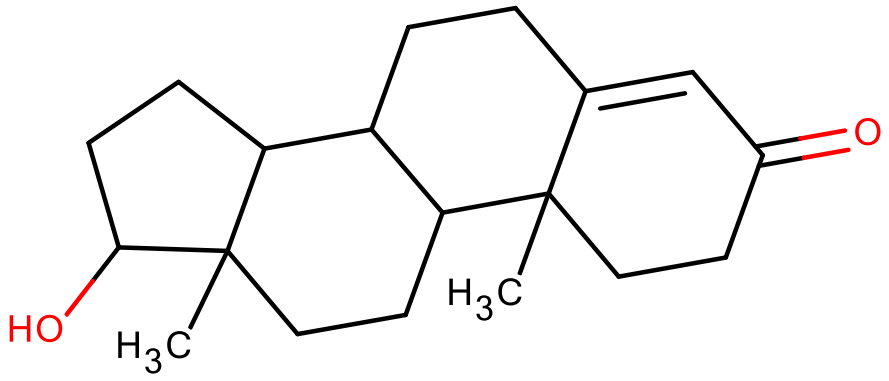
- Direct visualization
  - Structure-based approach
  - Property-based approach

  } 2D embedding - mapping/dimensionality reduction

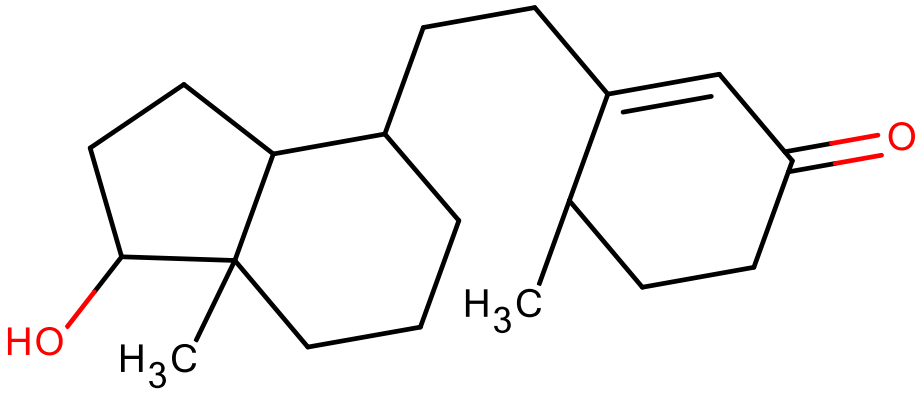
- (Hierarchical) clustering
  - Individual molecules
  - Hierarchy on common (structural) features

Similarity of molecules

Computer scientist's perception



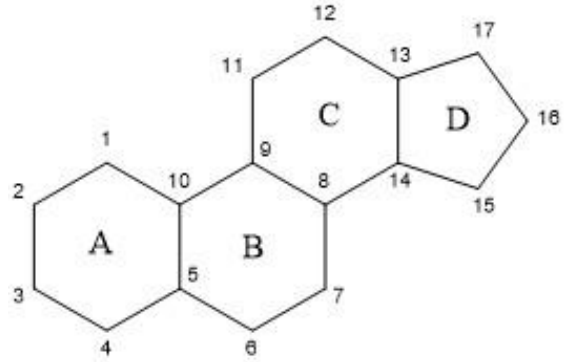
Chemist's perception

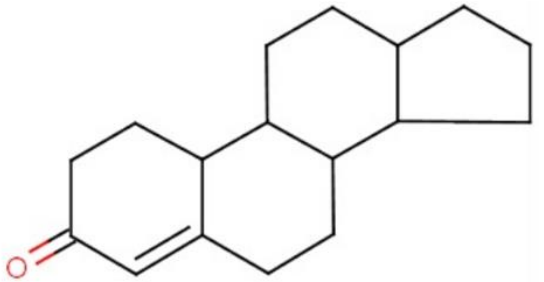


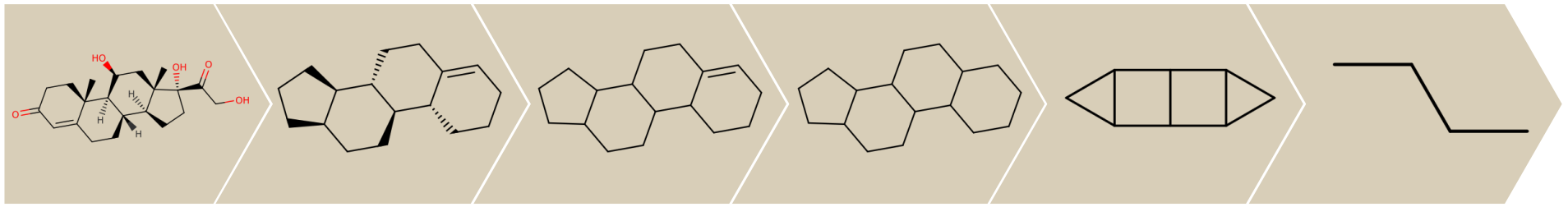
Molecular scaffolds → hierarchy

- Scaffold = ring-system | core | framework
  - Core functional or structural elements
  - Molecular backbone connected to biological activity

- Different types of scaffolds



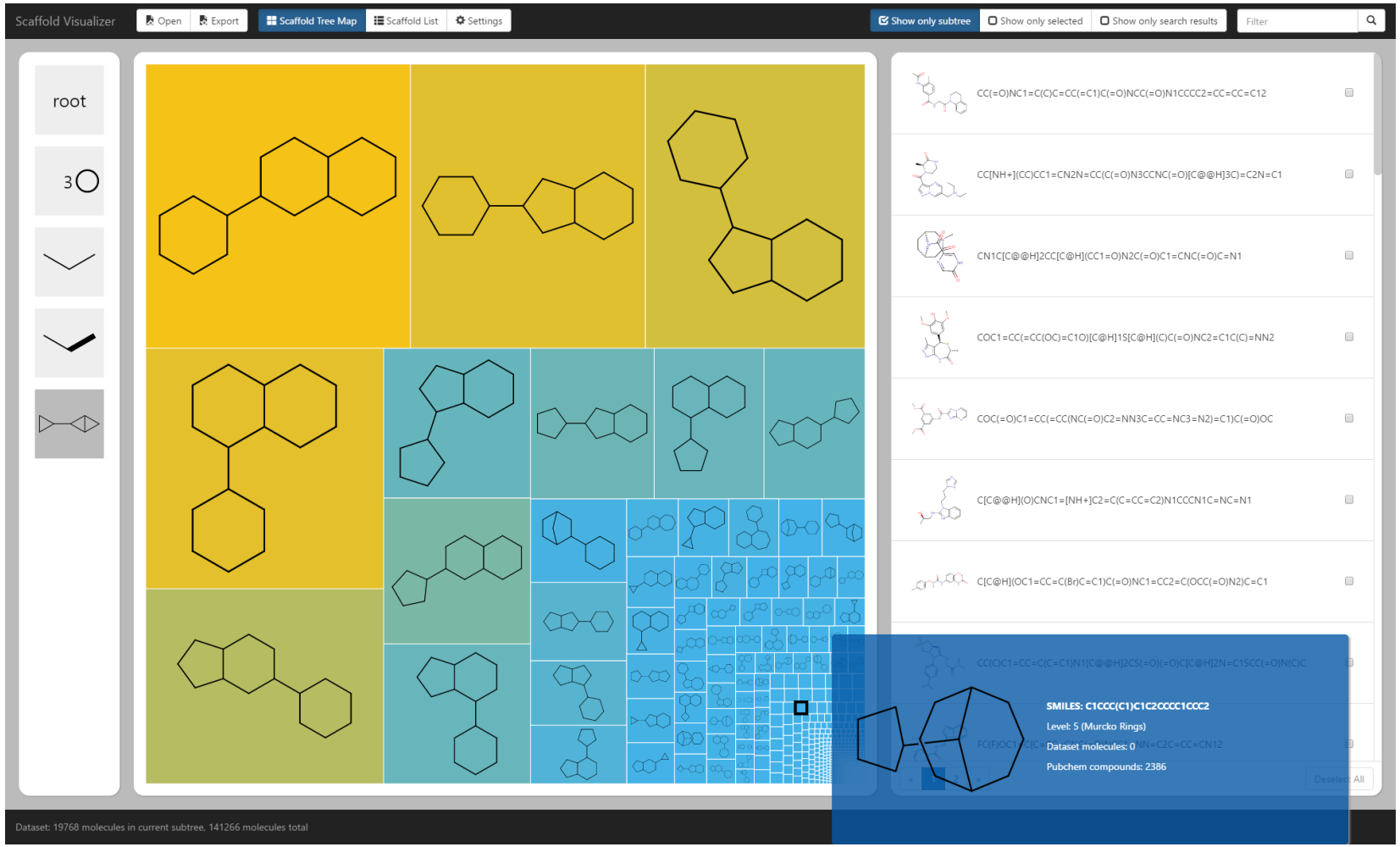




Hydrocortison scaffold hierarchy

Scaffvis

- **Zoomable tree map**
  - Subset selection, export, listview, …
- Arbitrary chemical **background**
  - **Color coding**
  - **Size-based coding**
  - Chemical space



http://scaffvis.projekty.ms.mff.cuni.cz https://github.com/velkoborsky/scaffvis

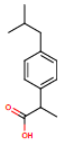
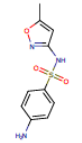
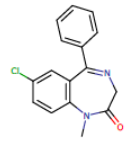
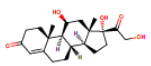
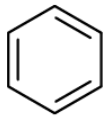
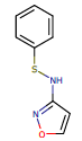
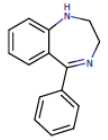
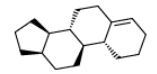
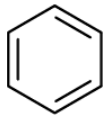
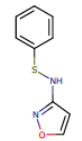
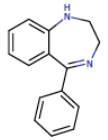
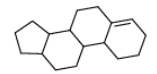
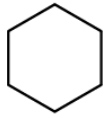
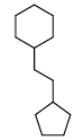
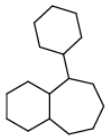
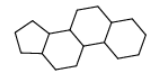
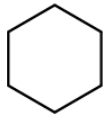
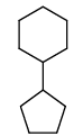
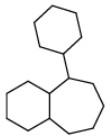
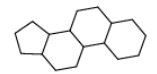
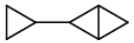
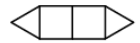
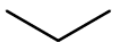
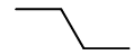
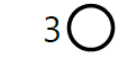
Scaffvis project

- Definition of scaffold hierarchy
- Build the hierarchy from the known chemical space
- Build zoomable tree map to visualize a set of compounds

Hierarchy construction

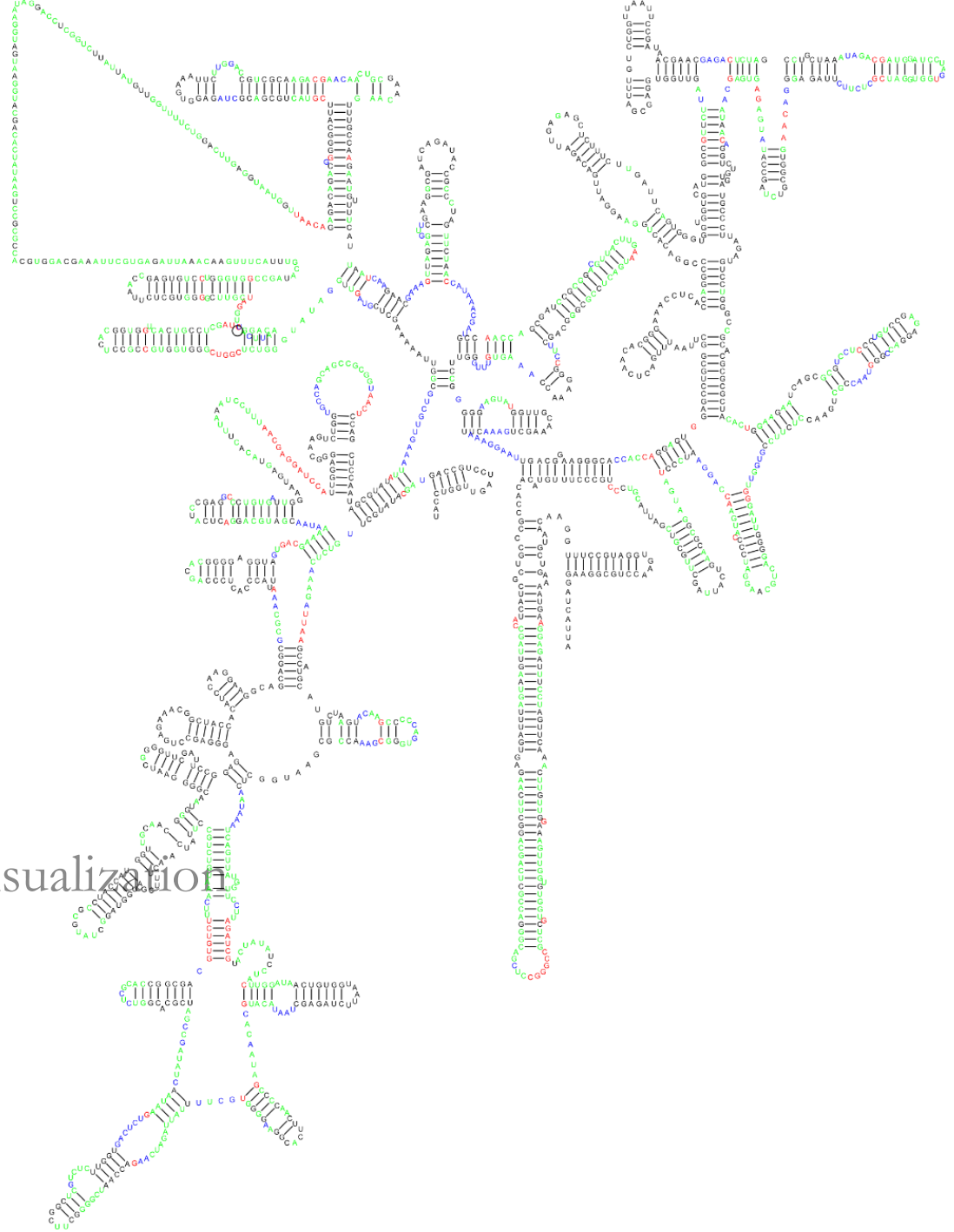
- Known chemical space → Pubchem
  - 91 millions compounds → 20 millions scaffolds
- Representability on screen
  - Reasonable branching factor

Level	Number of scaffolds by branching factor							
	0-100		101-400		401-1600		>1600	
0	0	0.00 %	1	100.00 %	0	0.00 %	0	0.00 %
1	69	67.65 %	15	14.71 %	7	6.86 %	11	10.78 %
2	49 781	99.94 %	28	0.06 %	0	0.00 %	0	0.00 %
3	118 902	100.00 %	0	0.00 %	0	0.00 %	0	0.00 %
4	137 032	99.56 %	455	0.33 %	125	0.09 %	21	0.02 %
5	595 555	99.92 %	472	0.08 %	30	0.01 %	1	0.00 %
6	1 274 080	99.40 %	5 756	0.45 %	1 602	0.13 %	350	0.03 %
7	7 476 752	100.000 %	35	0.00 %	0	0.00 %	0	0.00 %

Level	Ibuprofen	Sulfameth-oxazole	Diazepam	Hydrocor-tisone
Native				
8				
7				
6				
5				
4				
3				
2				
1				

TRAVeLer

Template-based RNA secondary structure visualization

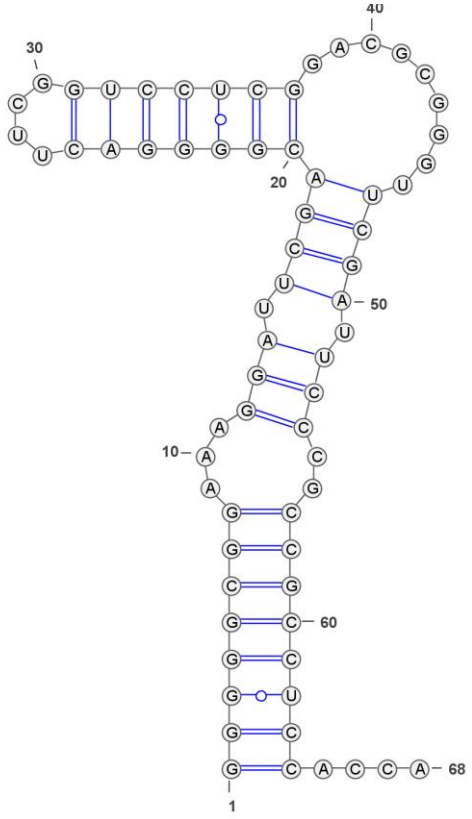


RNA structure

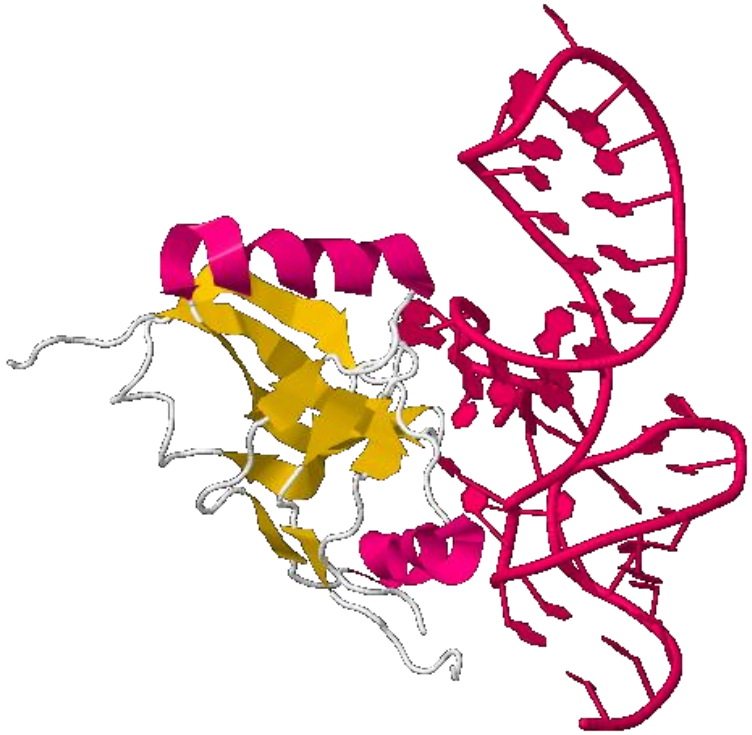
Primary structure

GGGGGCGGAAAGGAUUC
GACGGGGACUUCGGUCCU
CGGACGCGGGUUCGAUUC
CCGCCGCCUCCACCA

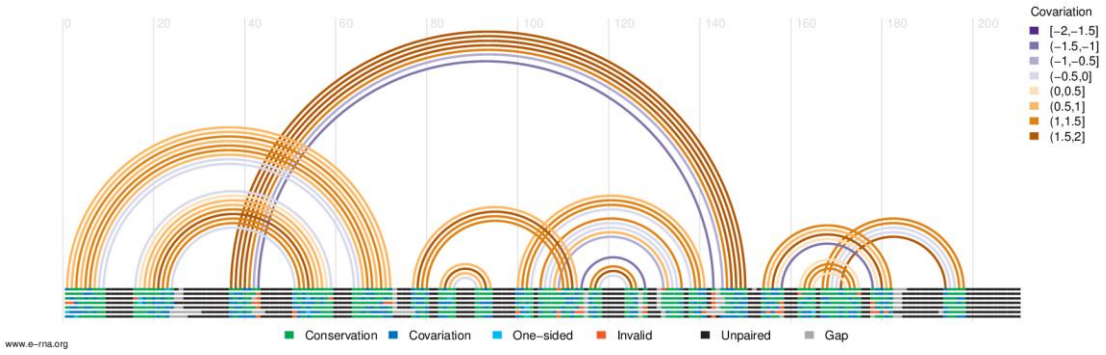
Secondary structure



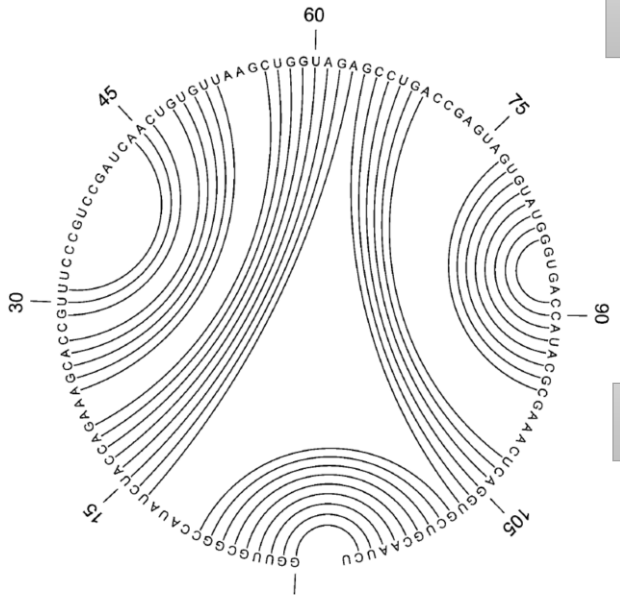
Tertiary structure



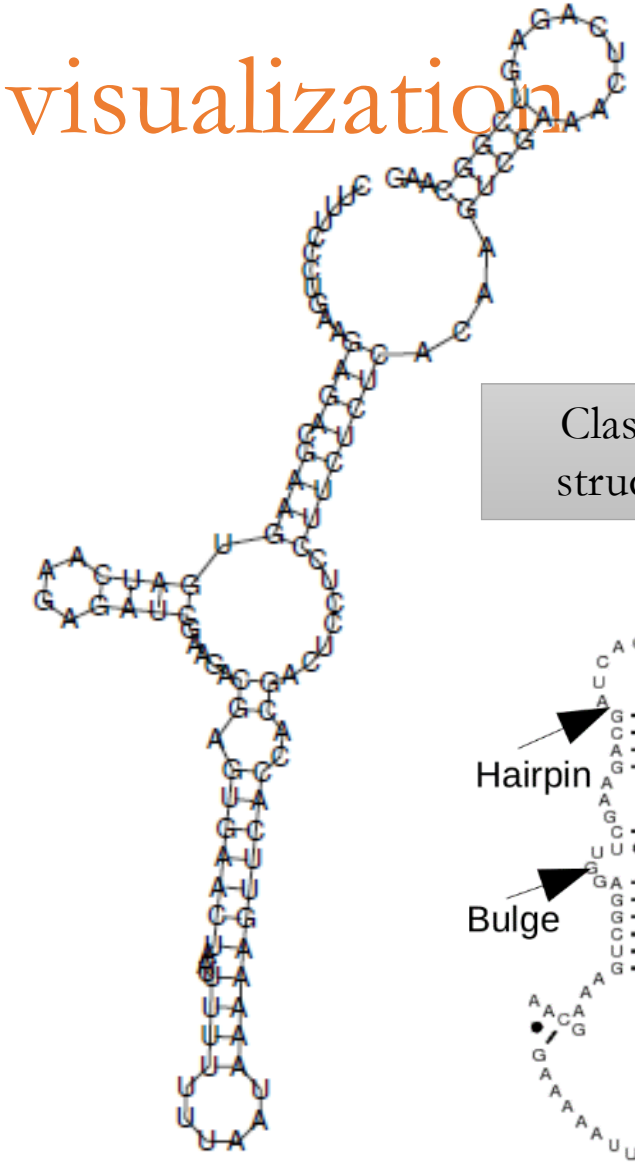
RNA secondary structure visualization



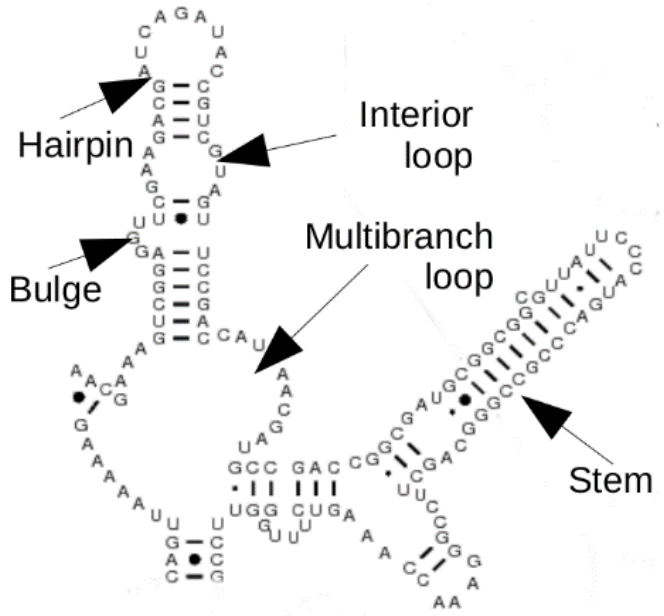
Arc graph

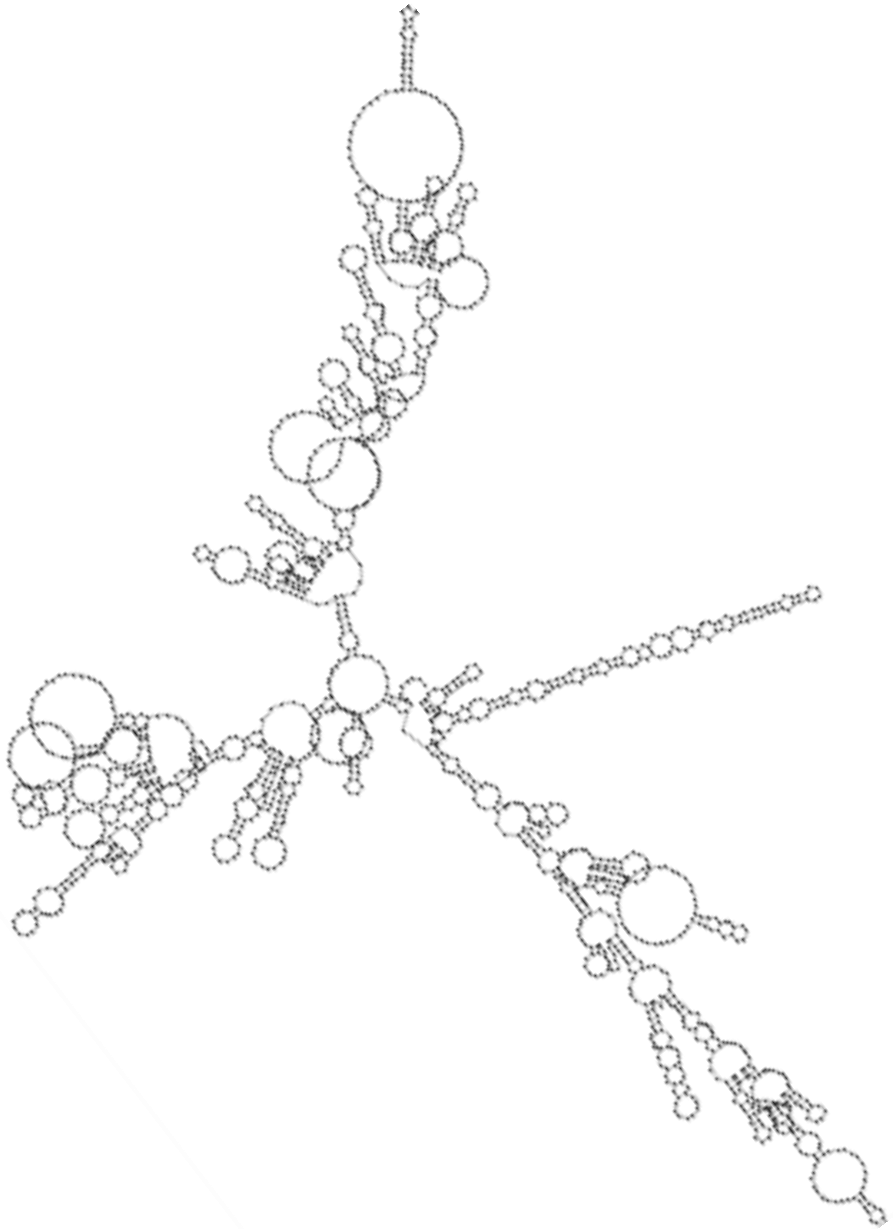


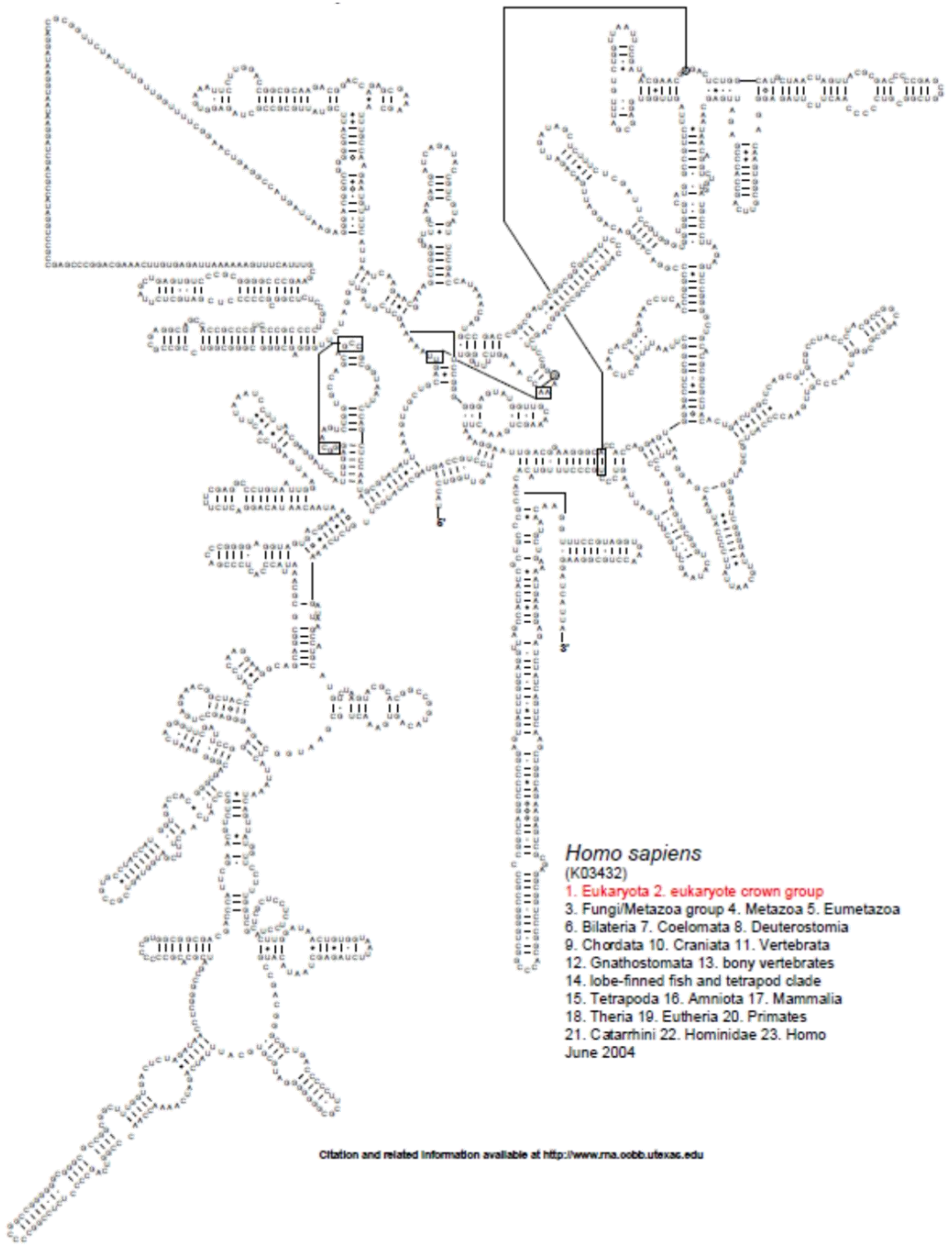
Circular graph



Classical structure






Homo sapiens
(K03432)
1. Eukaryota 2. eukaryote crown group
3. Fungi/Metazoa group 4. Metazoa 5. Eumetazoa
6. Bilateria 7. Coelomata 8. Deuterostomia
9. Chordata 10. Craniata 11. Vertebrata
12. Gnathostomata 13. bony vertebrates
14. lobe-finned fish and tetrapod clade
15. Tetrapoda 16. Amniota 17. Mammalia
18. Theria 19. Eutheria 20. Primates
21. Catarrhini 22. Hominidae 23. Homo
June 2004
Citation and related information available at http://www.rna.ccbb.utexas.edu

Algorithm outline

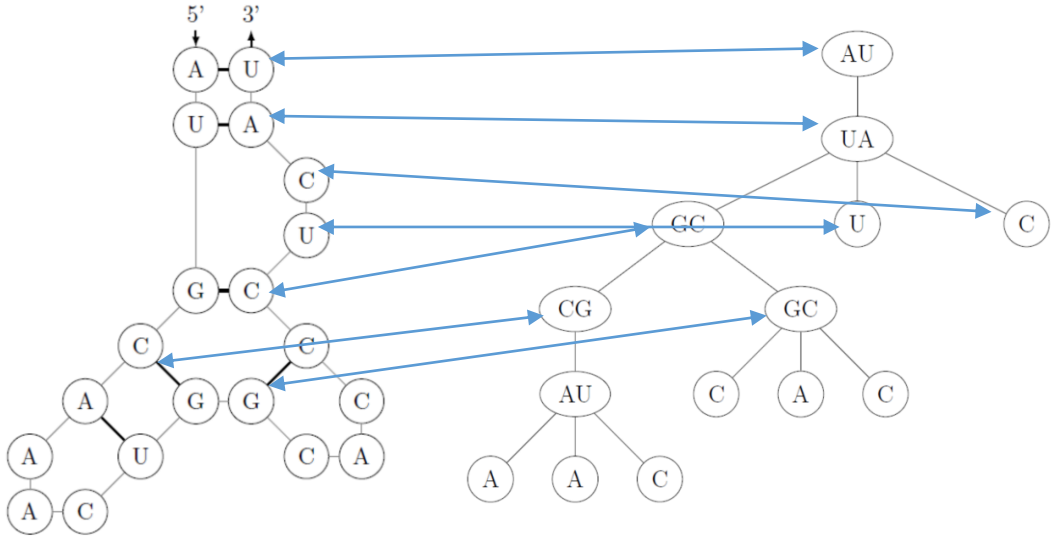
- Template-based visualization → preservation of **common motifs**
  - Template = homologous structure with known optimal layout

- **Input**: RNA secondary **structure without layout** (**target**) and secondary **structure with** existing **layout** (**template**)

1. **Convert** target and template structure into **tree representation**
2. Compute **tree edit distance between template and target** → sequence of tree edit operations
3. **Map** the tree edit operations **to visual operations** to convert template layout to target layout

RNA tree edit distance

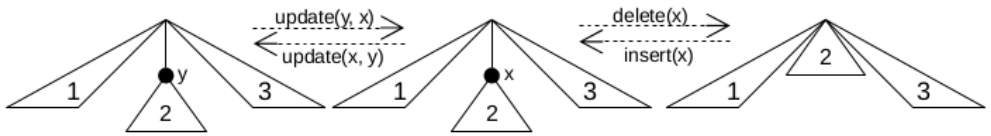
- **Structure → tree**
  - Base pairs → inner nodes
  - Unpaired nucleotides → leafs



AUGCAAACUGGCACCCUCAU
(((((...))(...))..))

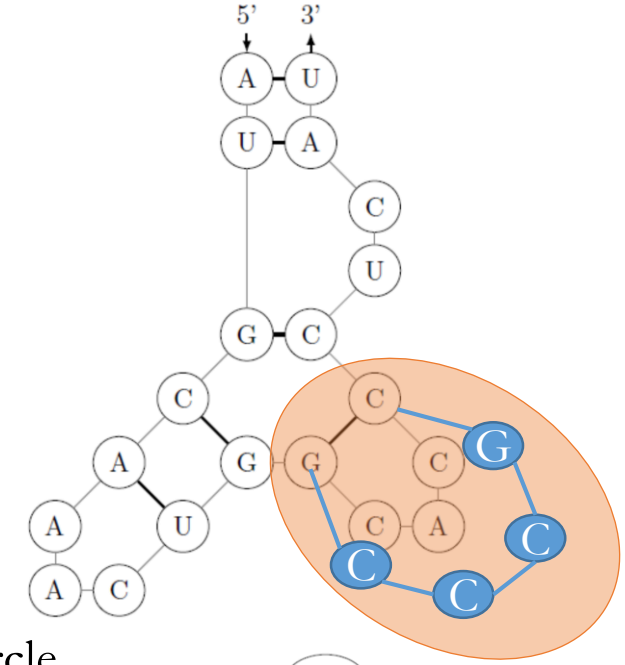
Tree edit distance

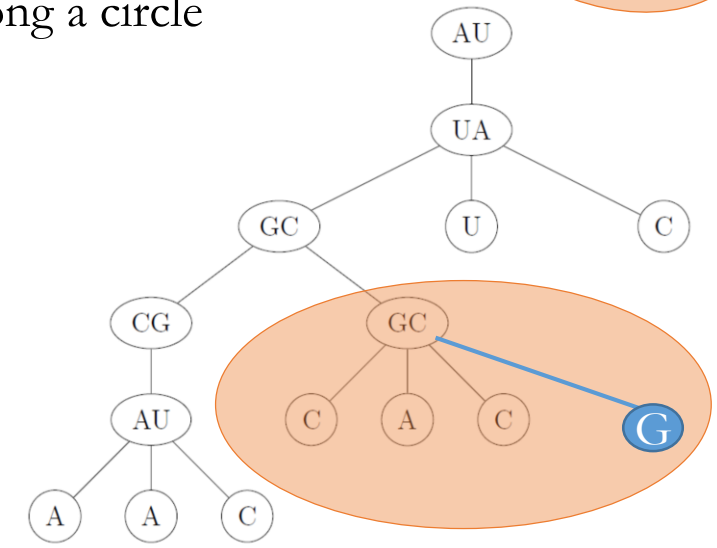
- Modification of string edit distance
- Operations
  - **Update** – relabeling
  - **Delete** – deletion of a node and reconnection of children to the parent
  - **Insert** – insertion of a node between two connected nodes and reconnection of children



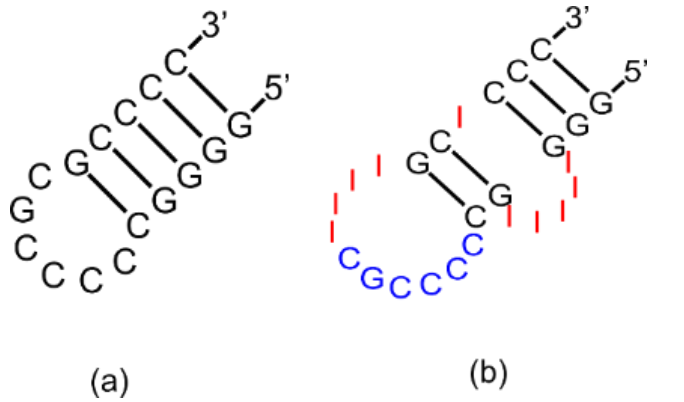
Visual operations

- **Update**
  - Relabeling
- **Insert**
  - **Leaf node**
    - **No siblings** → formation of a **new loop**
    - **Existing siblings** → **loop extension** → uniform distribution along a circle
  - **Inner node**
    - **Insert base pair** at given position
    - **Shift** all its **descendants**
- **Delete**
  - Analogy to insert
- Multi-branch loops treated individually

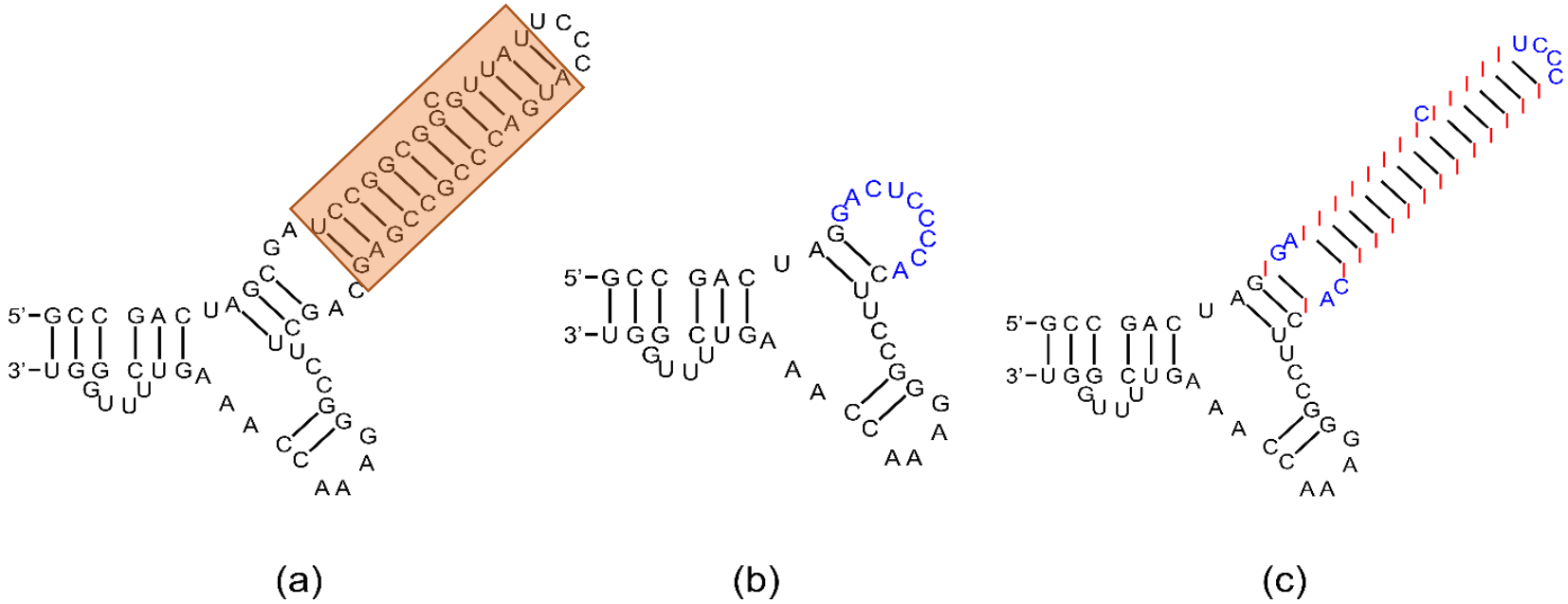




Examples



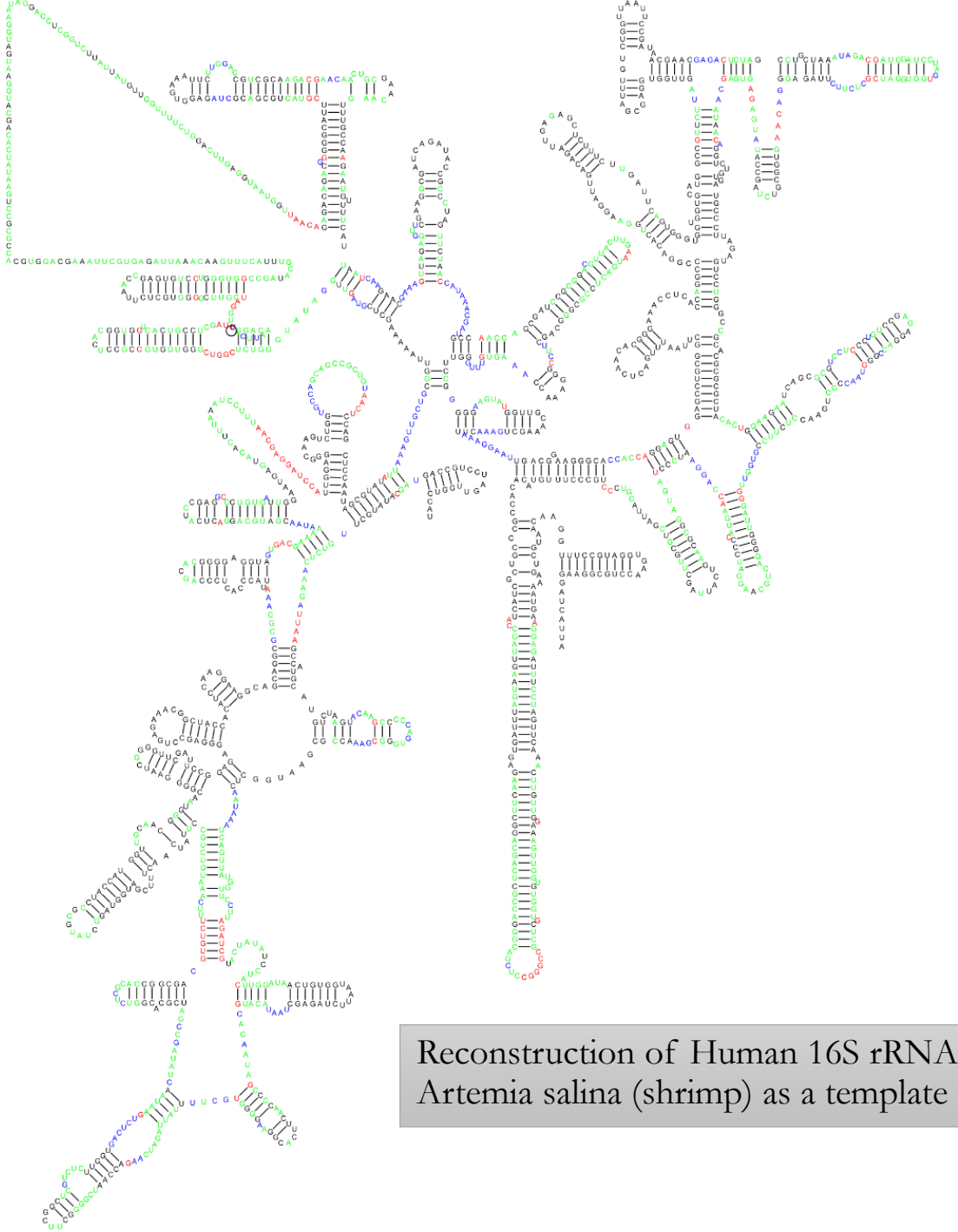
Insertion into both stem and loop parts of a hairpin



Substantial deletion and reinsert of one branch of a multibranch loop

Ribosomal RNA test

- **Reconstruction** of visualizations of known **16S ribosomal subunits** from the **Metazoa** kingdom
  - Layout taken from the CRW database
- 16 organisms
- Every pair of organisms tested → **272 layouts**
- **3 crossings per layout on average**



Reconstruction of Human 16S rRNA using Artemia salina (shrimp) as a template

https://github.com/rikiel/TRAVeLer

Questions?