PyWaw #65
13.02.2017

Biologia pythona

dr Jacek Śmietański
jacek.smietanski@ii.uj.edu.pl
http://jaceksmietanski.net

O autorze

Instytut Informatyki UJ
Bioinformatyka struktur RNA

Epam
Algorithm Developer, Python

Hobby:
- Bilety z całego świata
- Geocaching

O czym powiem

- Czym jest bioinformatyka i dlaczego jest użyteczna
- Dlaczego bioinformatycy tak lubią pythona
- O bibliotece biopython
- Pokażę kilka innych popularnych narzędzi bioinformatycznych
- O platformie Rosalind
- O projektach realizowanych w moim laboratorium

Bioinformatyka w praktyce

Medycyna
np. medycyna personalizowana

Farmaceutyka
np. projektowanie leków

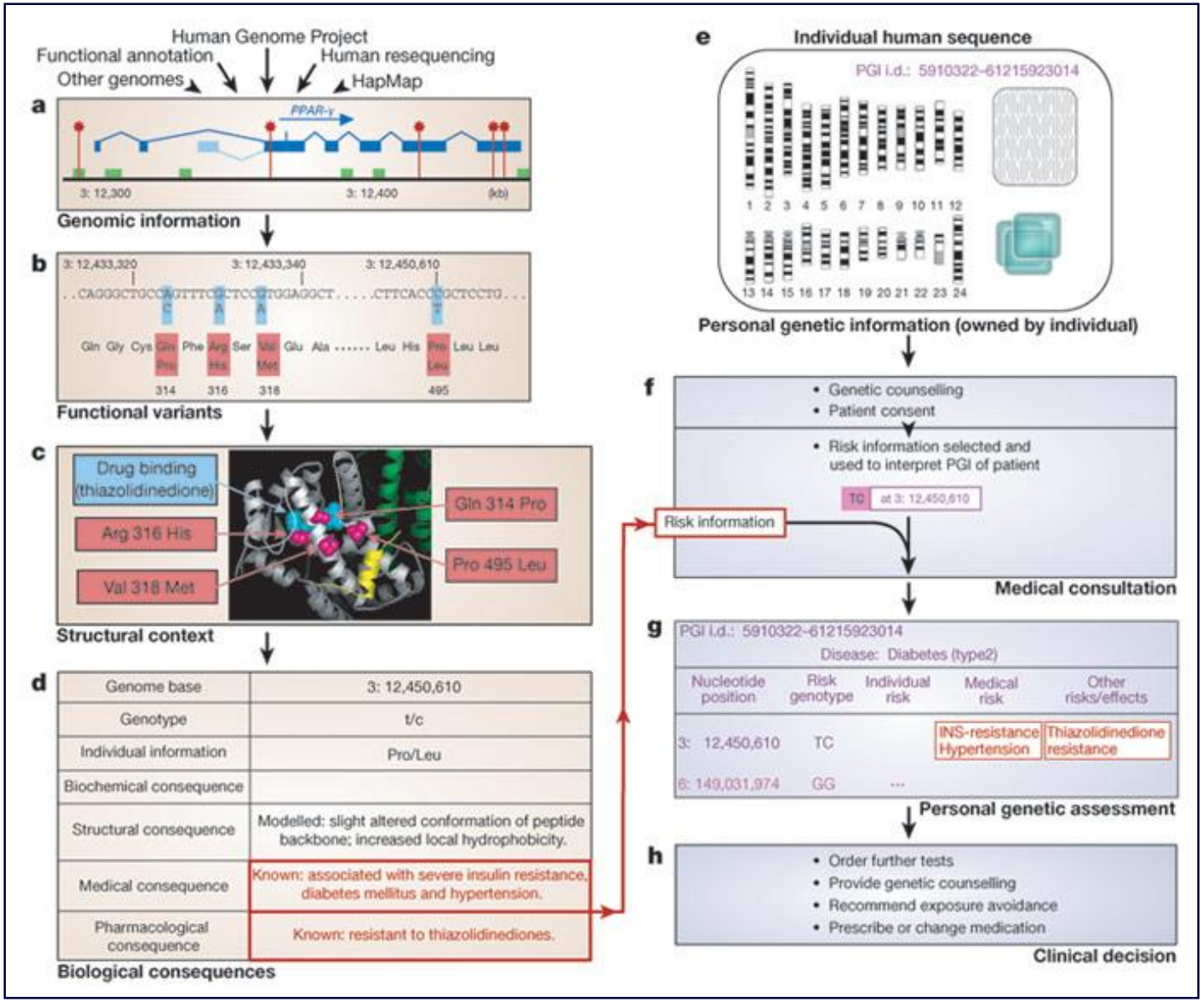
Kryminalistyka
np. identyfikacja sprawców

Sądownictwo
np. ustalanie ojcostwa

Rolnictwo
np. tworzenie nowych odmian

Archeologia
np. badania paleontologiczne

Medycyna personalizowana



W ujęciu algorytmicznym

Bazy danych
- tworzenie i przechowywanie zbiorów danych
- opisywanie danych, identyfikacja powiązań, aktualizacja

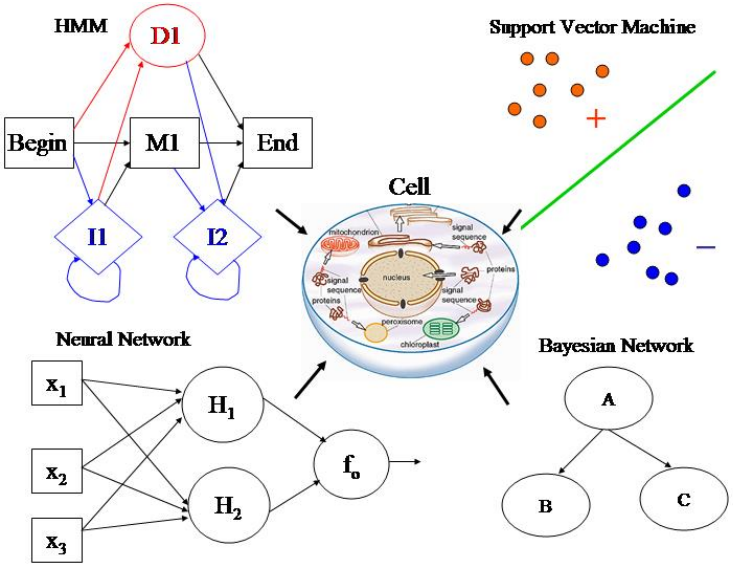
Big data
- analizy porównawcze na ogromnych zbiorach danych

Uczenie maszynowe
- przewidywanie i klasyfikacja (genów, struktur, oddziaływań itp.)

Metody optymalizacji
- modelowanie molekularne, filogenetyka

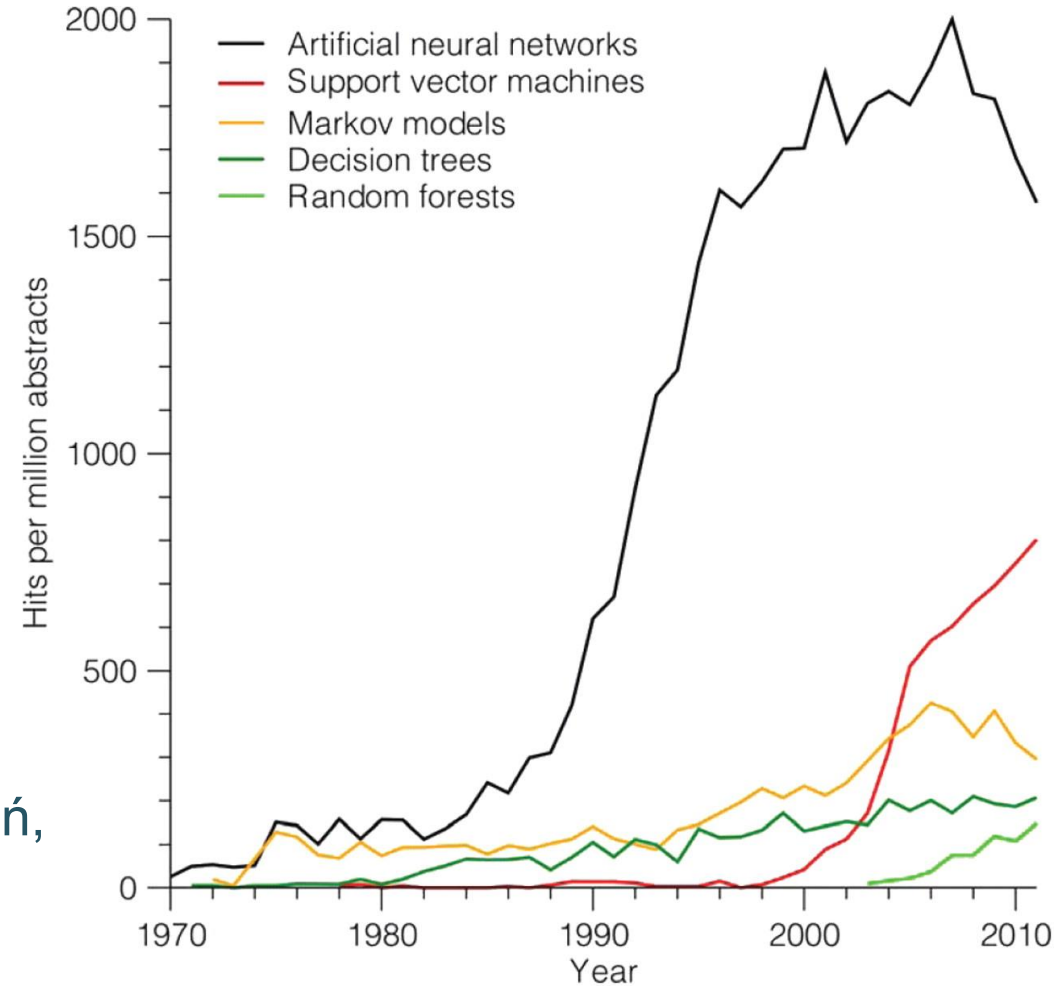
Teoria grafów
- konstrukcja i porównywanie sieci metabolicznych

Uczenie maszynowe w bioinformatyce



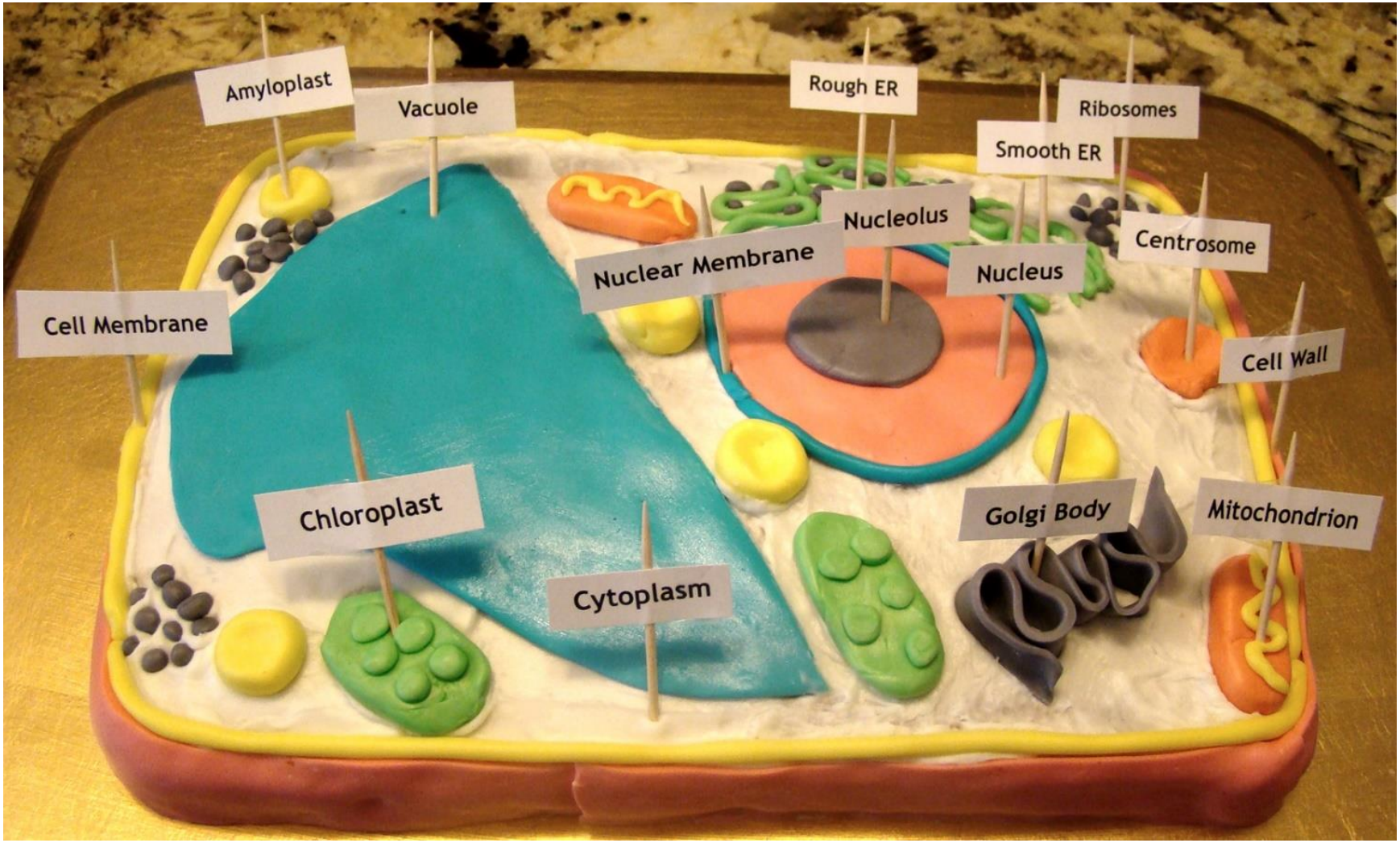
przewidywanie genów,
struktur białkowych, oddziaływań,
klasyfikacja gatunków, …

Zastosowanie technik ML w bioinformatyce:



Źródło: Jensen L J , Bateman A Bioinformatics 2011;27:3331-3332

Obszar rozważań – komórka



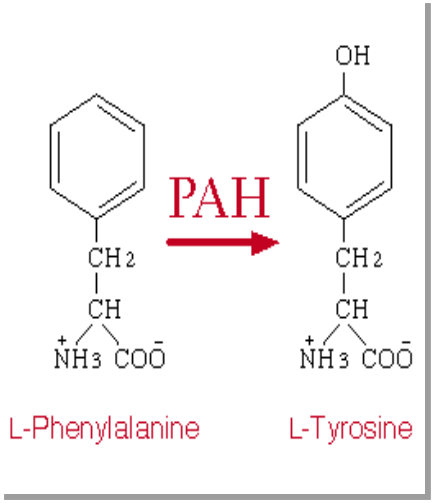
Centralny dogmat bioinformatyki

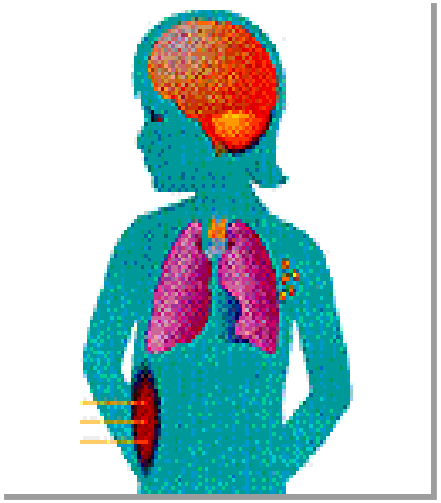
DNA → RNA → Białko

Sekwencja → Struktura → Funkcja → Fenotyp

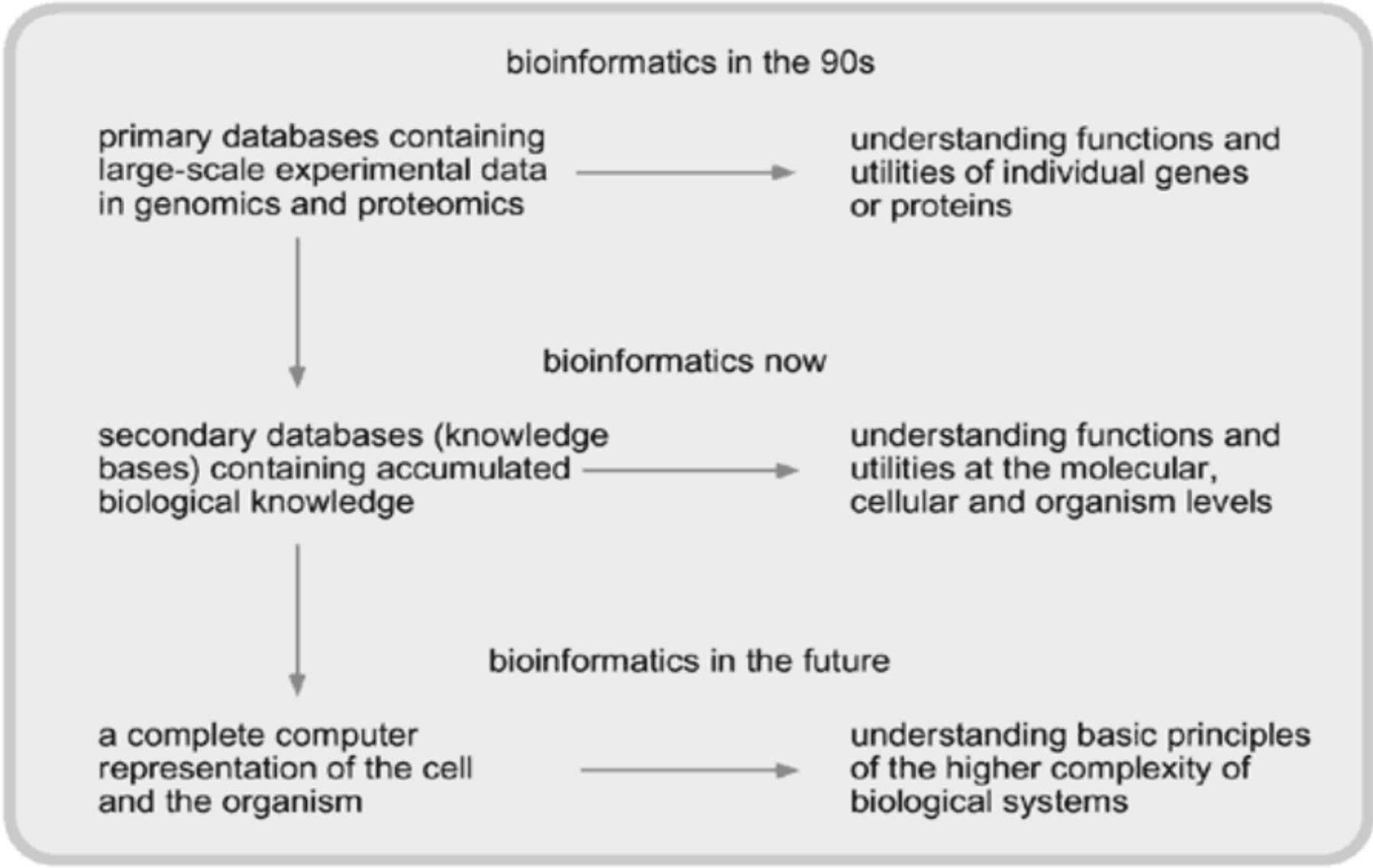
Informacja genetyczna → Struktura molekularna → Funkcja biochemiczna → Fenotyp (objawy)

MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGNV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH

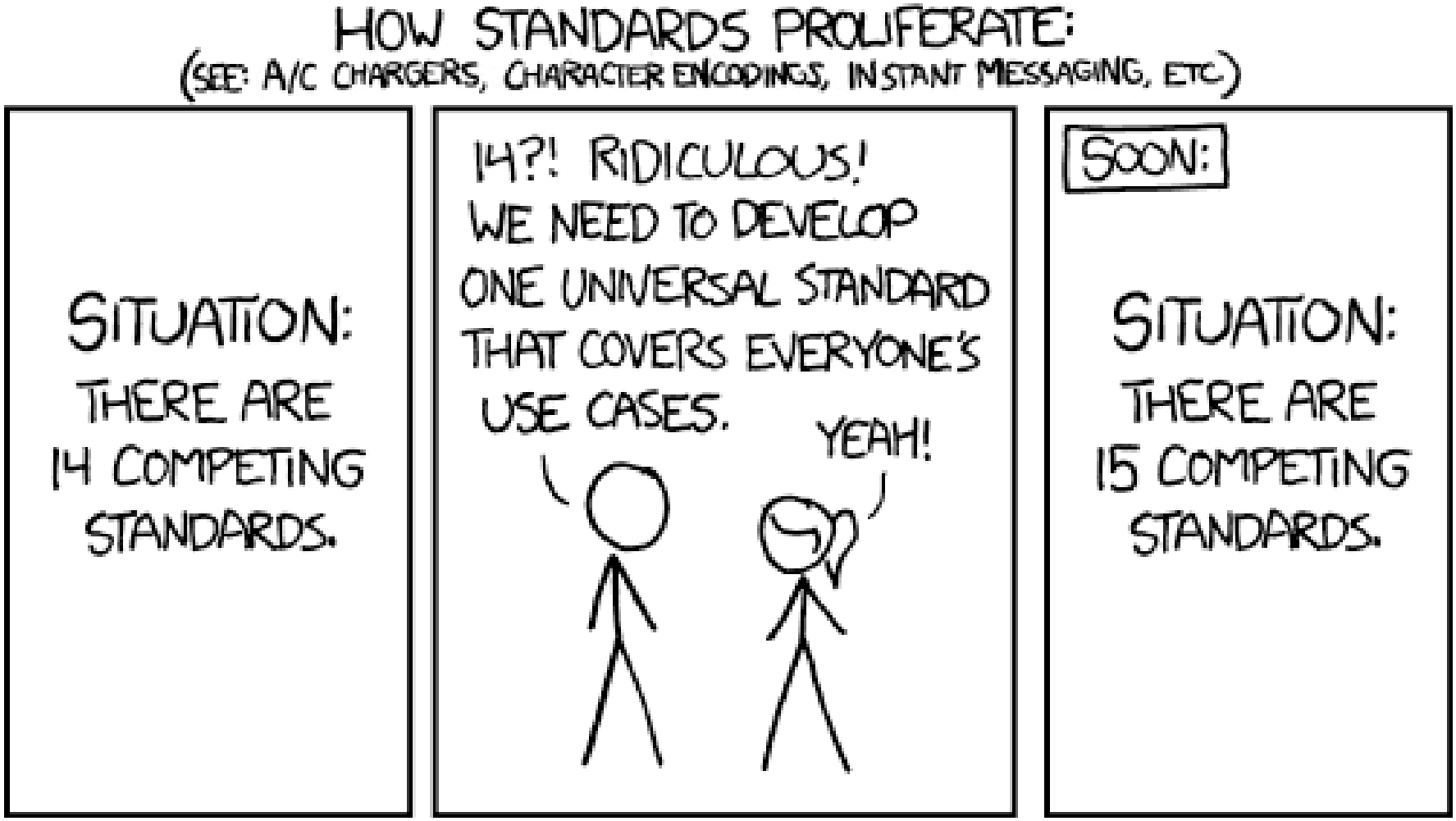




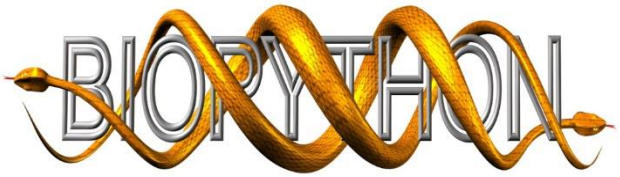
Rozwój bioinformatyki



Standardy


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Biblioteki Bio



Biopython
http://biopython.org

BioJava
http://biojava.org

BioPerl
http://www.bioperl.org

BioRuby
http://bioruby.open-bio.org/

Edit this page on GitHub

Biopython

See also our News feed and Twitter.

Introduction

Biopython is a set of freely available tools for biological computation written in Python by an international team of developers.

It is a distributed collaborative effort to develop Python libraries and applications which address the needs of current and future work in bioinformatics. The source code is made available under the Biopython License, which is extremely liberal and compatible with almost every license in the world.

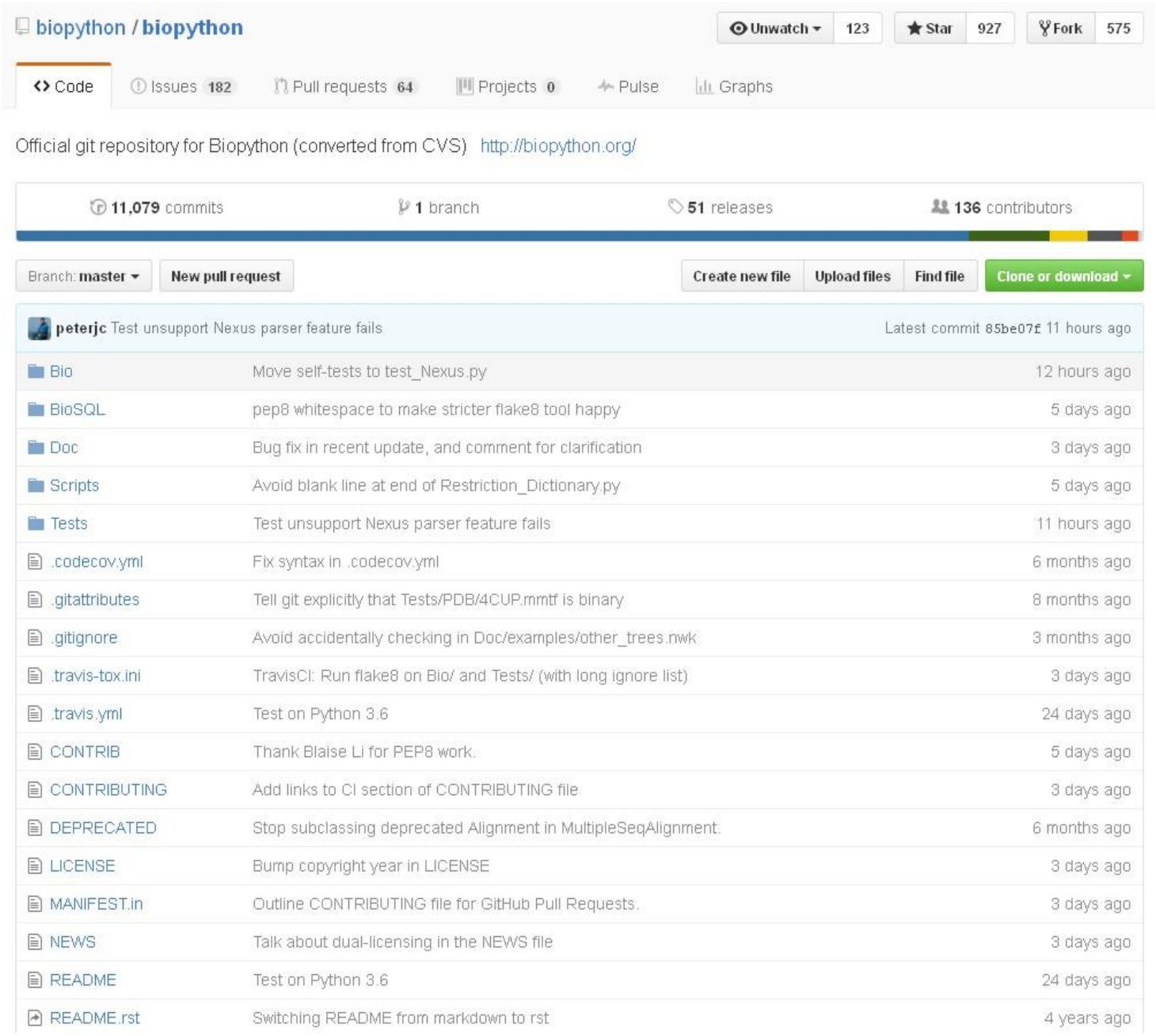
We are a member project of the Open Bioinformatics Foundation (OBF), who take care of our domain name and hosting for our mailing list etc. The OBF used to host our developement repository, issue tracker and website but these are now on GitHub.

This wiki will help you download and install Biopython, and start using the libraries and tools.

Get Started	Get help	Contribute
Download Biopython	Tutorial (PDF)	What's being worked on
Installation help (PDF)	Documentation on this wiki	Developing on Github
	Cookbook (working examples)	Google Summer of Code
	Discuss and ask questions	Report bugs (older issues)

The latest release is Biopython 1.68, released on 25 August 2016.

Biopython - github



biopython / biopython

Unwatch 123 | Star 927 | Fork 575

Code | Issues 182 | Pull requests 64 | Projects 0 | Pulse | Graphs

Official git repository for Biopython (converted from CVS) http://biopython.org/

11,079 commits | 1 branch | 51 releases | 136 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

peterjc Test unsupport Nexus parser feature fails — Latest commit 85be07f 11 hours ago

Bio	Move self-tests to test_Nexus.py	12 hours ago
BioSQL	pep8 whitespace to make stricter flake8 tool happy	5 days ago
Doc	Bug fix in recent update, and comment for clarification	3 days ago
Scripts	Avoid blank line at end of Restriction_Dictionary.py	5 days ago
Tests	Test unsupport Nexus parser feature fails	11 hours ago
.codecov.yml	Fix syntax in .codecov.yml	6 months ago
.gitattributes	Tell git explicitly that Tests/PDB/4CUP.mmtf is binary	8 months ago
.gitignore	Avoid accidentally checking in Doc/examples/other_trees.nwk	3 months ago
.travis-tox.ini	TravisCI: Run flake8 on Bio/ and Tests/ (with long ignore list)	3 days ago
.travis.yml	Test on Python 3.6	24 days ago
CONTRIB	Thank Blaise Li for PEP8 work.	5 days ago
CONTRIBUTING	Add links to CI section of CONTRIBUTING file	3 days ago
DEPRECATED	Stop subclassing deprecated Alignment in MultipleSeqAlignment.	6 months ago
LICENSE	Bump copyright year in LICENSE	3 days ago
MANIFEST.in	Outline CONTRIBUTING file for GitHub Pull Requests.	3 days ago
NEWS	Talk about dual-licensing in the NEWS file	3 days ago
README	Test on Python 3.6	24 days ago
README.rst	Switching README from markdown to rst	4 years ago

Operacje na sekwencjach

```
>>> from Bio.Seq import Seq
>>> my_seq = Seq("AGTACACTGGT")
>>> my_seq
Seq('AGTACACTGGT', Alphabet())
>>> print(my_seq)
AGTACACTGGT
```

Parsowanie plików FASTA

```
>>>from Bio import SeqIO
>>>for seq_record in SeqIO.parse("ls_orchid.fasta", "fasta"):
...         print(seq_record.id)
...         print(repr(seq_record.seq))
...         print(len(seq_record))
```

```
gi|2765658|emb|Z78533.1|CIZ78533
Seq('CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCGTGG...CGC',
SingleLetterAlphabet())
740
...
gi|2765564|emb|Z78439.1|PBZ78439
Seq('CATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCTGTTTACT...GCC',
SingleLetterAlphabet())
592
```

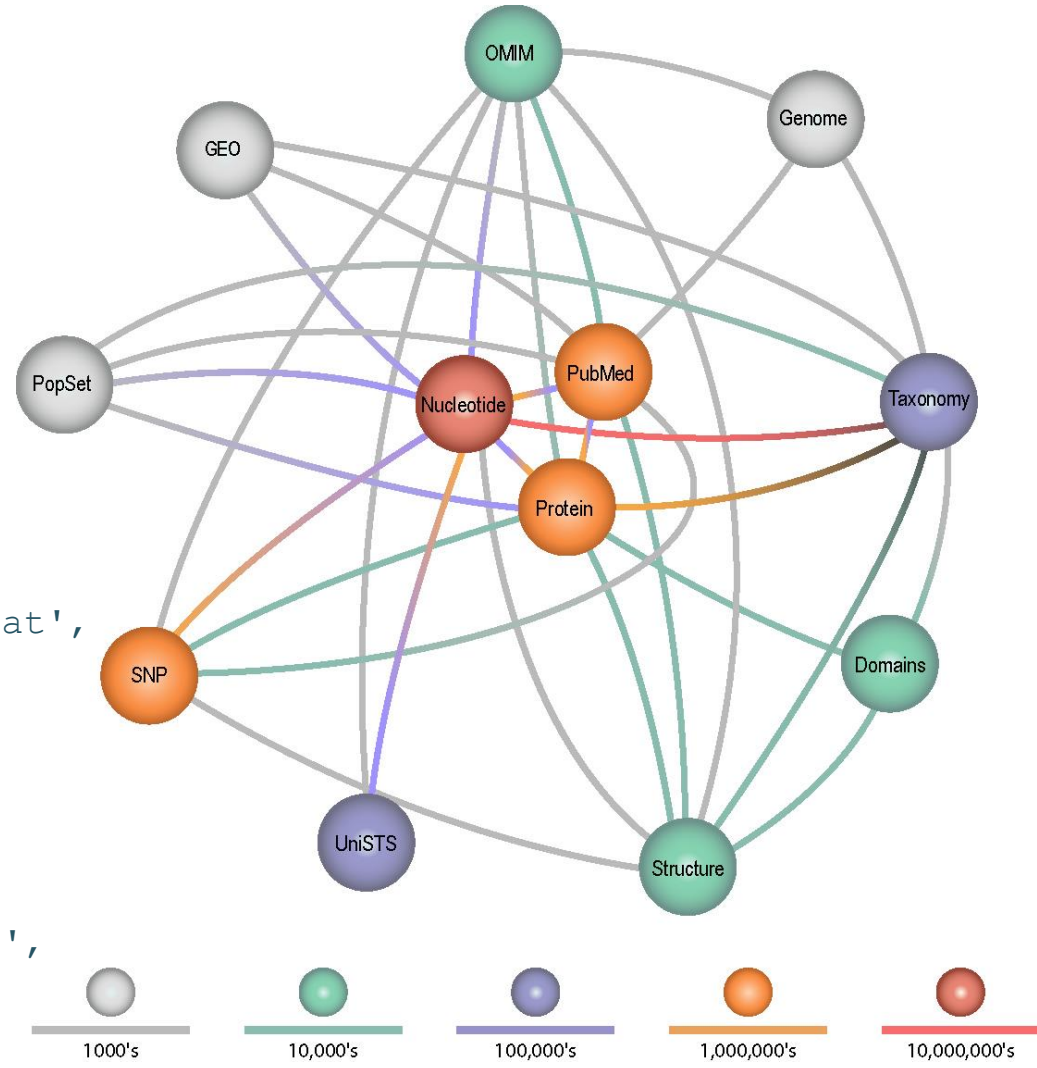
Porównywanie sekwencji

```
>>> from Bio import pairwise2
>>> from Bio import SeqIO

>>> seq1 = SeqIO.read("alpha.faa", "fasta")
>>> seq2 = SeqIO.read("beta.faa", "fasta")

>>> alignments = pairwise2.align.globalxx(seq1.seq, seq2.seq)

>>> print(pairwise2.format_alignment(*alignment[0]))
```

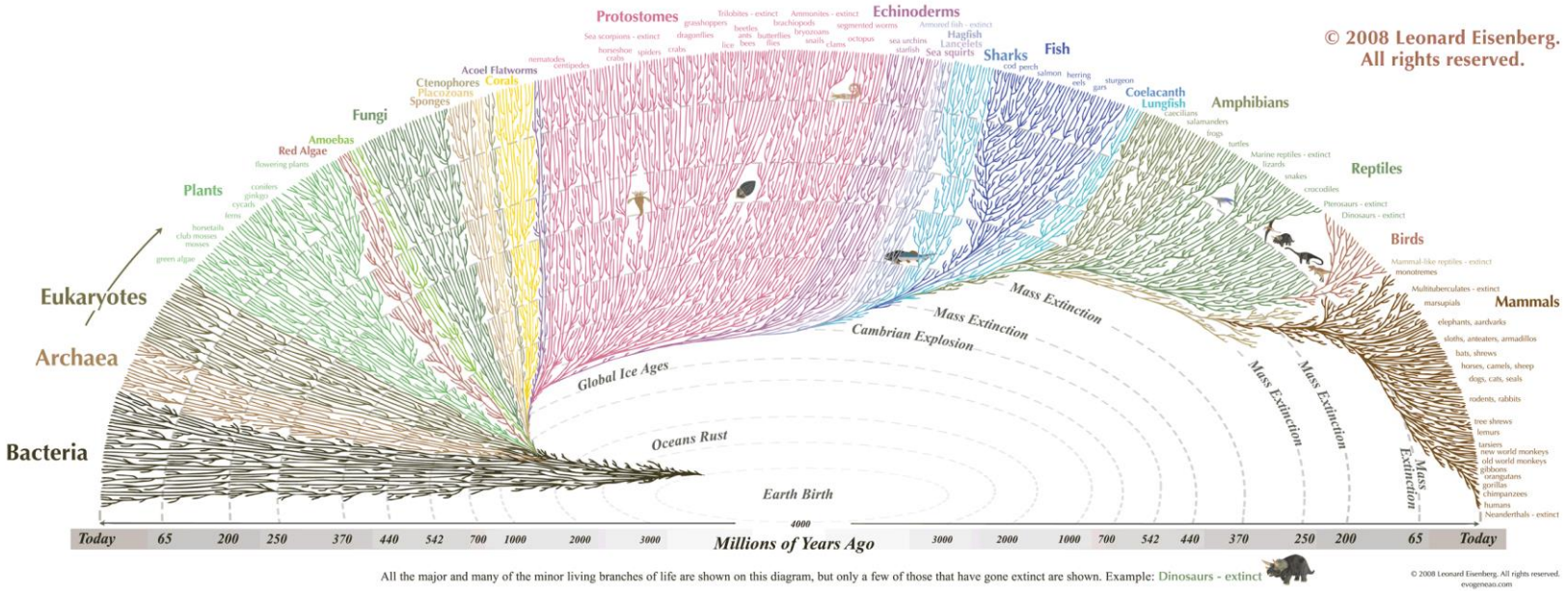
```
MV-LSPADKTNV---K-A--A-WGKVGAHAG---EY-GA-EALE-RMFLSF----PTTK-TY--F...YR-
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...|||
MVHL-----T--PEEKSAVTALWGKV-----NVDE-VG-GEAL-GR--L--LVVYP---WT-QRF...Y-H
 Score=72
```

API do NCBI Entrez

```
>>> from Bio import Entrez
>>> handle = Entrez.einfo()
>>> record = Entrez.read(handle)
>>> record["DbList"]
['pubmed', 'protein', 'nucleotide',
 'nuccore', 'nucgss', 'nucest',
 'structure', 'genome', 'books',
 'cancerchromosomes', 'cdd', 'gap',
 'domains', 'gene', 'genomeprj', 'gensat',
 'geo', 'gds', 'homologene',
 'journals', 'mesh', 'ncbisearch',
 'nlmcatalog', 'omia', 'omim', 'pmc',
 'popset', 'probe', 'proteinclusters',
 'pcassay', 'pccompound', 'pcsubstance',
 'snp', 'taxonomy', 'toolkit',
 'unigene', 'unists']
```



Filogenetyka

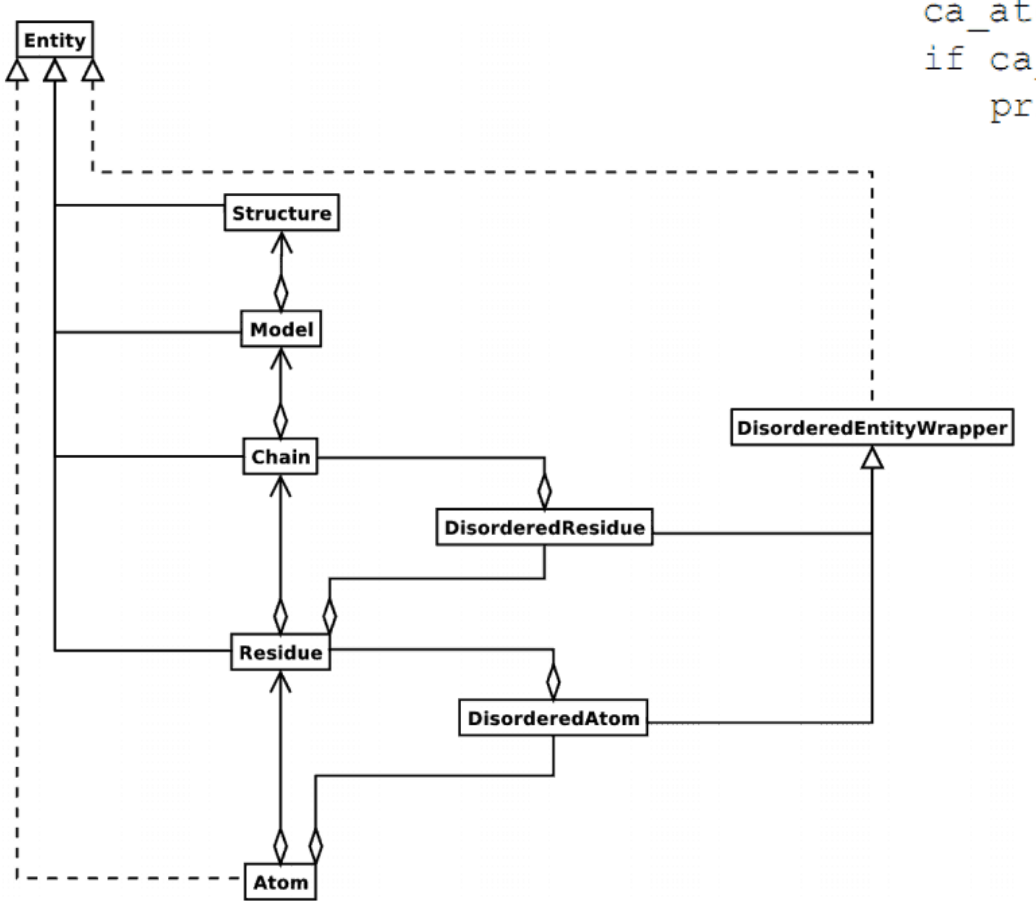



```
>>> from Bio import Phylo
>>> tree = Phylo.read("simple.dnd", "newick")
>>> Phylo.draw_ascii(tree)
```

```
                                  ____________________________ A
                 ________________|
                |                |____________________________ B
                |
 _______________|                 ____________________________ C
|               |________________|
|                                |____________________________ D
|
|                 ____________________________ E
|                |
|________________|____________________________ F
                 |
                 |____________________________ G
```

Struktury

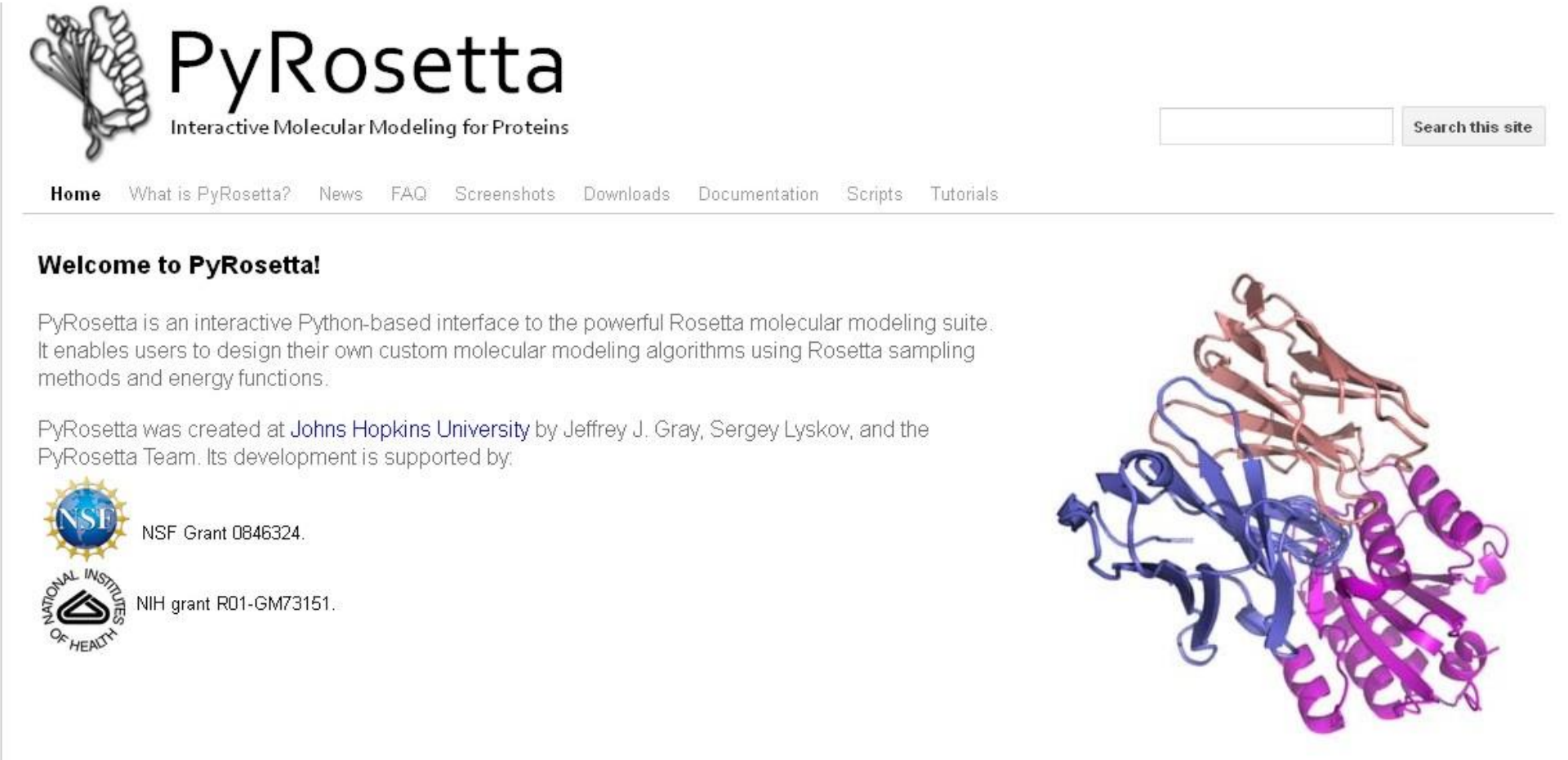
```
from PDBParser import PDBParser
parser=PDBParser(PERMISSIVE=1)
structure=parser.get_structure("1fat", "1fat.pdb")
for model in structure.get_list():
   for chain in model.get_list():
     for residue in chain.get_list():
        if residue.has_id("CA"):
            ca_atom=residue["CA"]
            if ca_atom.is_disordered():
               print residue
```



PyMol

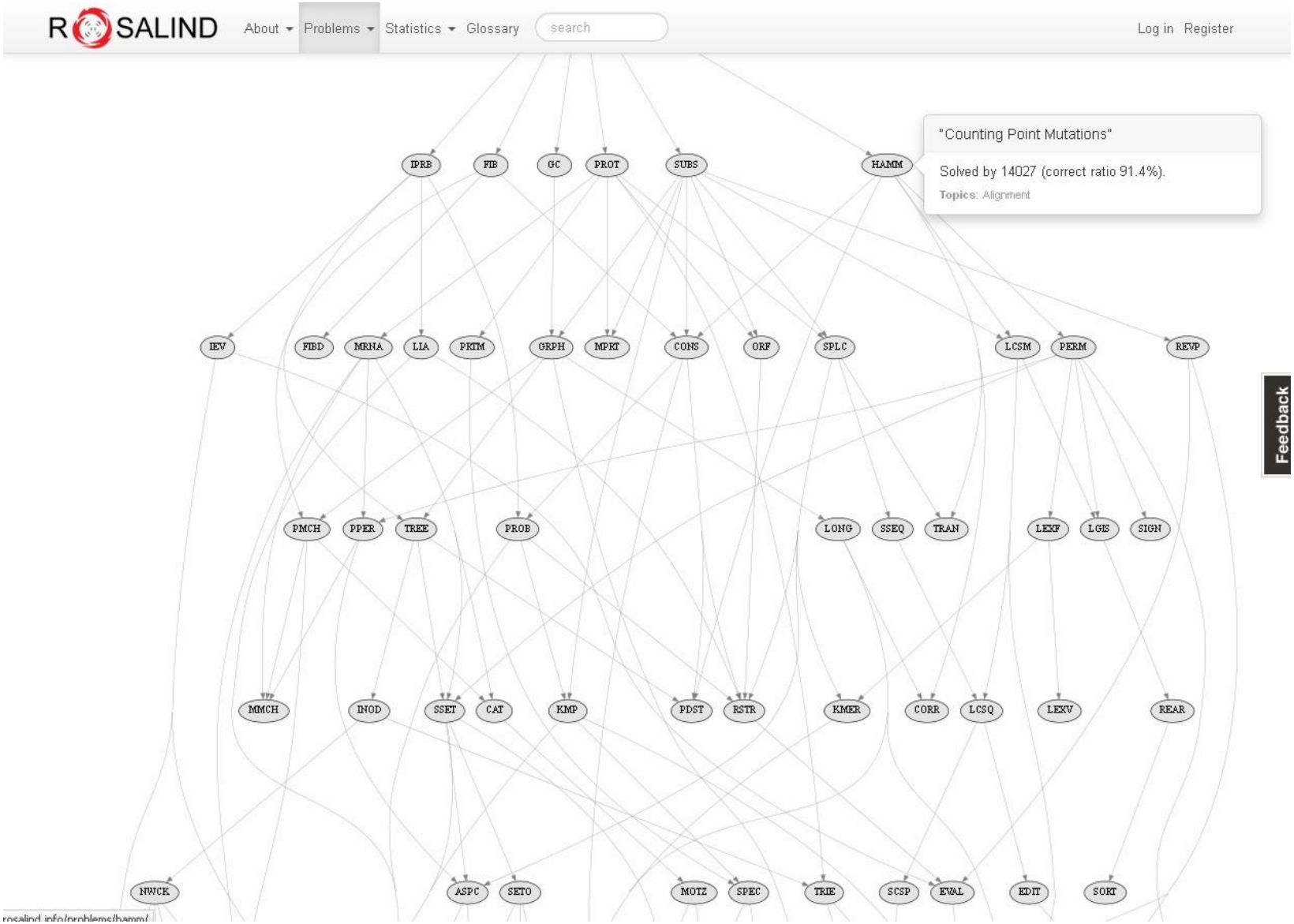
https://www.pymol.org/

http://www.pyrosetta.org/



Rosalind

http://rosalind.info/



Nasze projekty

http://github.com/BipUJ

- Bio.PDBList
- PDB Search
- PyPDBe
- NDB Adapter
- RNA Central tool
- RNA 2D converter
- RNA Secondary Structures Comparison
- …

Dziękuję za uwagę

Kontakt:
jacek.smietanski@ii.uj

I RNA Structures

Źródło: http://bytesizebio.net