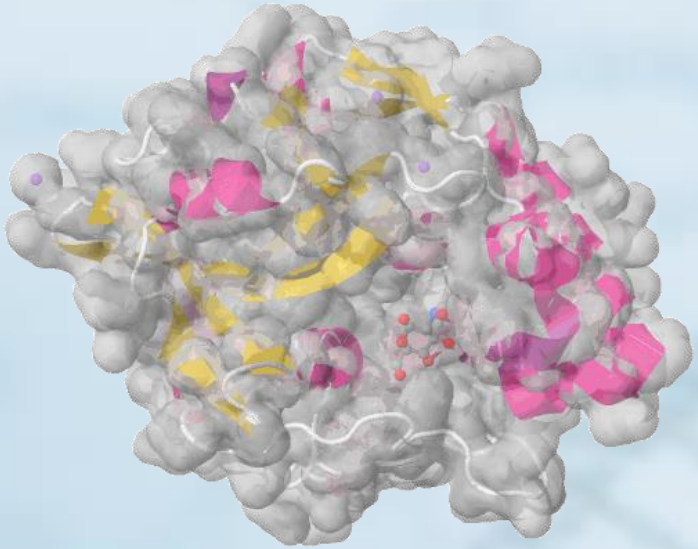
Využití metod strojového učení v bioinformatice



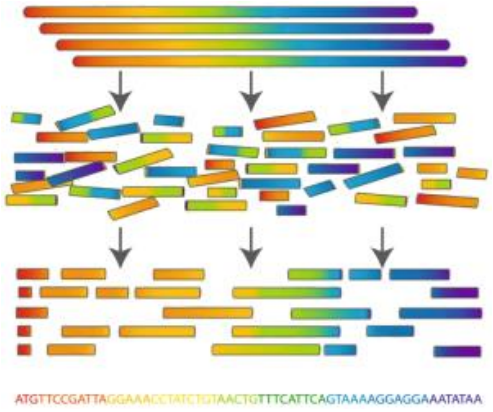
David Hoksza

SIRET Research Group
Katedra softwarového inženýrství,
Matematicko-fyzikální fakulta
Karlova Univerzita v Praze

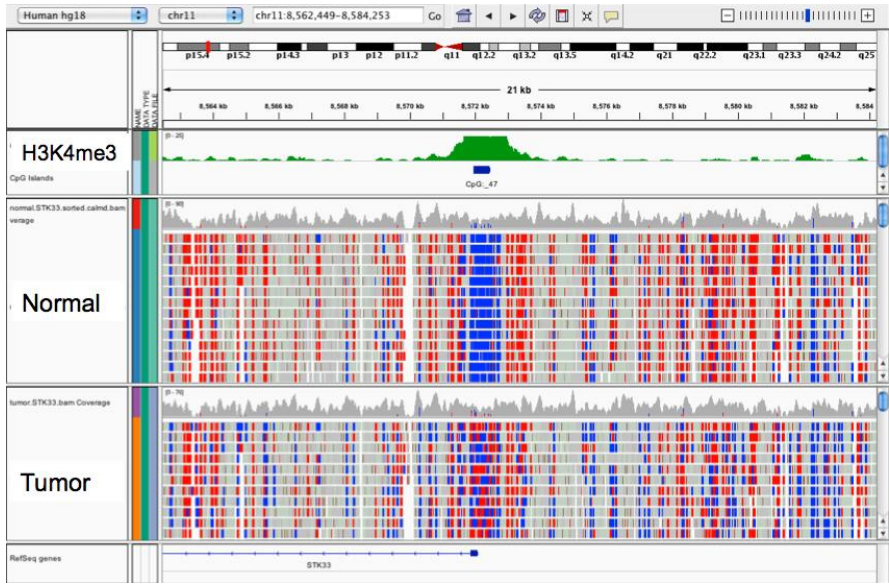
Bioinformatika

Bioinformatika

- Obor zabývající se vývojem a aplikací počítačových metod pro **získávání**, **ukládání**, **analýzu** a **interpretaci** biologických dat
  - Zaměření na molekuly DNA, RNA a proteinů



- cAMP- and cGMP-dependent protein kinase phosphorylation site
  - `[RK](2)-x-[ST]`
- Tyrosine kinase phosphorylation site
  - `[RK]-x(2)-[DE]-x(3)-Y or [RK]-x(3)-[DE]-x(2)-Y`



Typické úlohy v bioinformatice

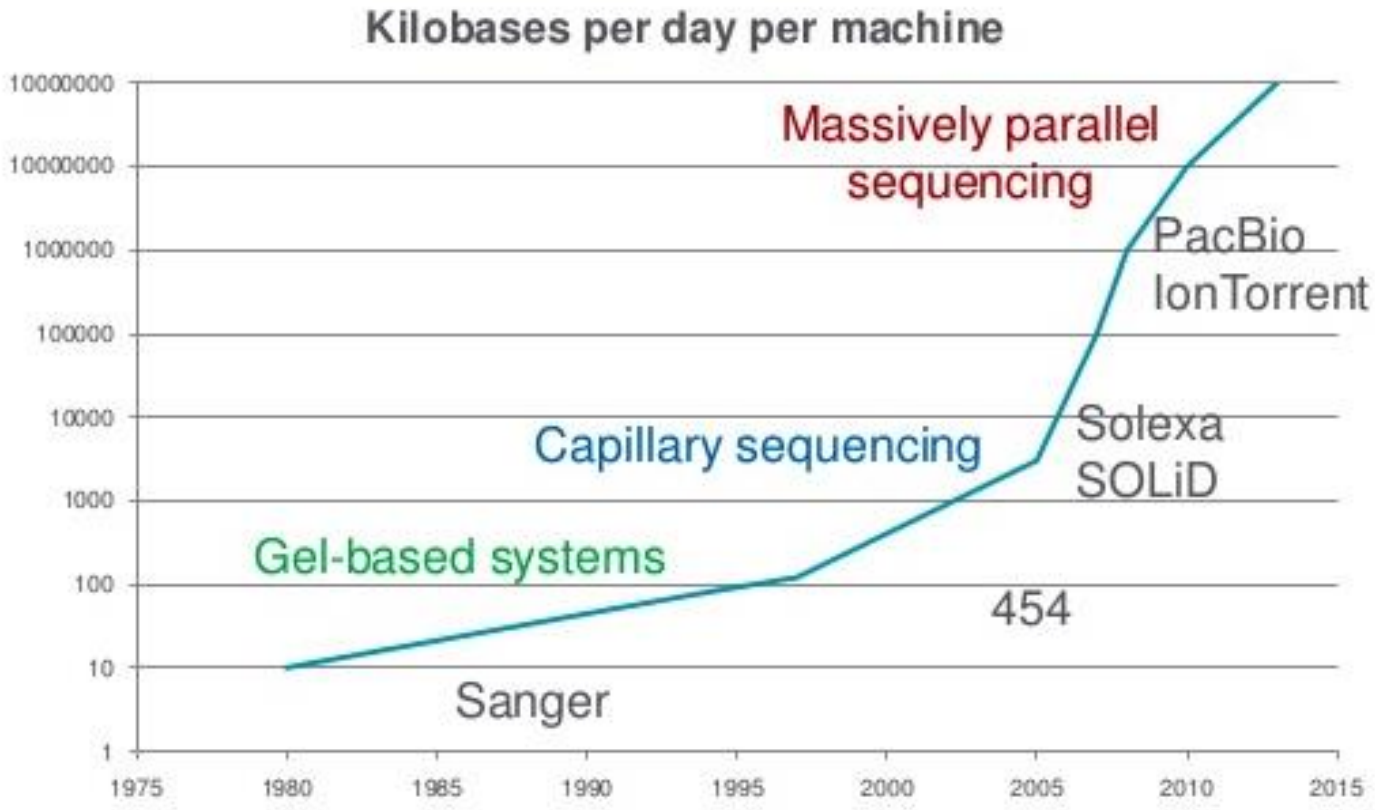
- Skládání a mapování genomů
- Podobnostní modelování sekvencí DNA, RNA, proteinů
- Fylogenetická analýza
- Analýza DNA/RNA
  - Hledání sekvenčních motivů
  - Hledání mutací ve vybraných populacích
- …

- Predikce struktury proteinů (RNA)
- Identifikace interakcí
- Molekulární modelování (docking)
- …

- Data genové exprese
- Data hmotnostní spektrometrie
- …

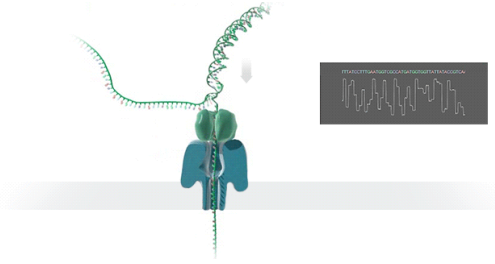
Aplikační domény bioinformatiky

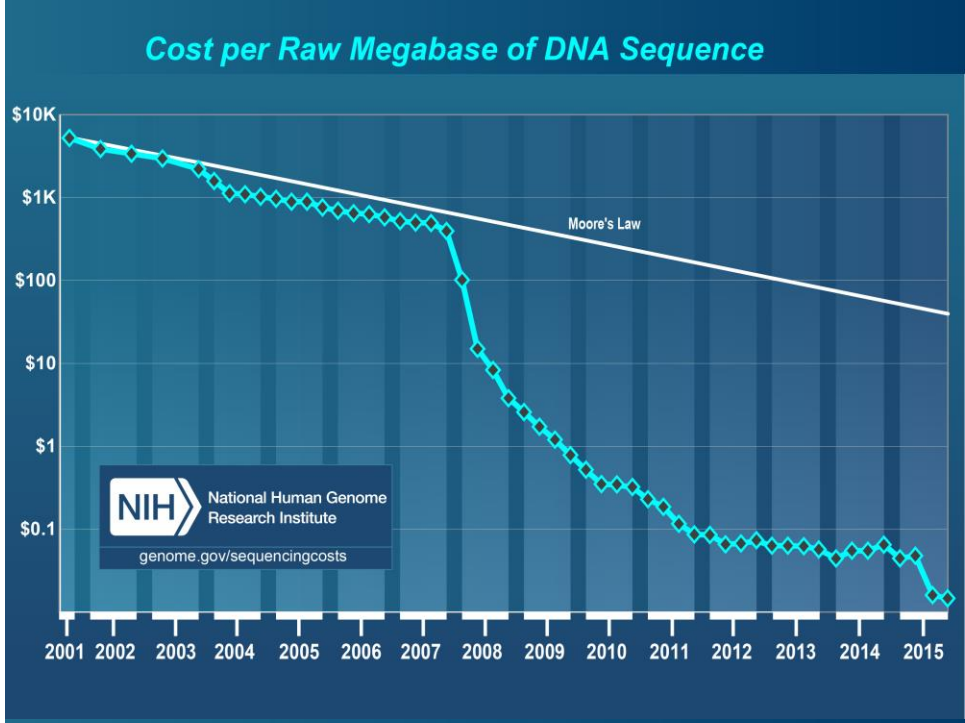
- **Základní výzkum**
  - Funkční genomika, evoluční genomika
- **Medicína**
  - Genetické komponenty nemocí, preventivní medicína, personalizovaná medicína, genová terapie, počítačový vývoj léčiv
- **Mikrobiologie**
  - Biotechnologie, epidemiologie, boj se znečištěním
- **Zemědělství**
  - Odolnější plodiny, resistence proti škůdcům, zvýšení nutriční hodnoty plodin

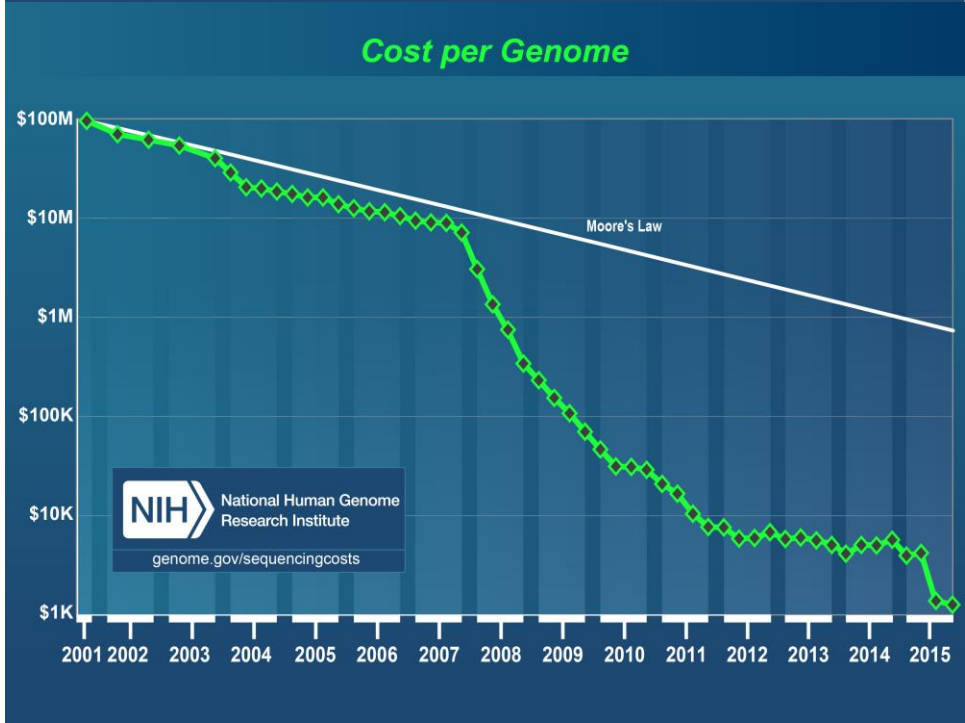

Kilobases per day per machine
10000000
10000000
1000000
100000
10000
1000
100
10
1
1975
1980
1985
1990
1995
2000
2005
2010
2015
Massively parallel sequencing
PacBio
IonTorrent
Solexa
SOLiD
Capillary sequencing
Gel-based systems
454
Sanger


Flowcell
MinION Mkl device
USB port

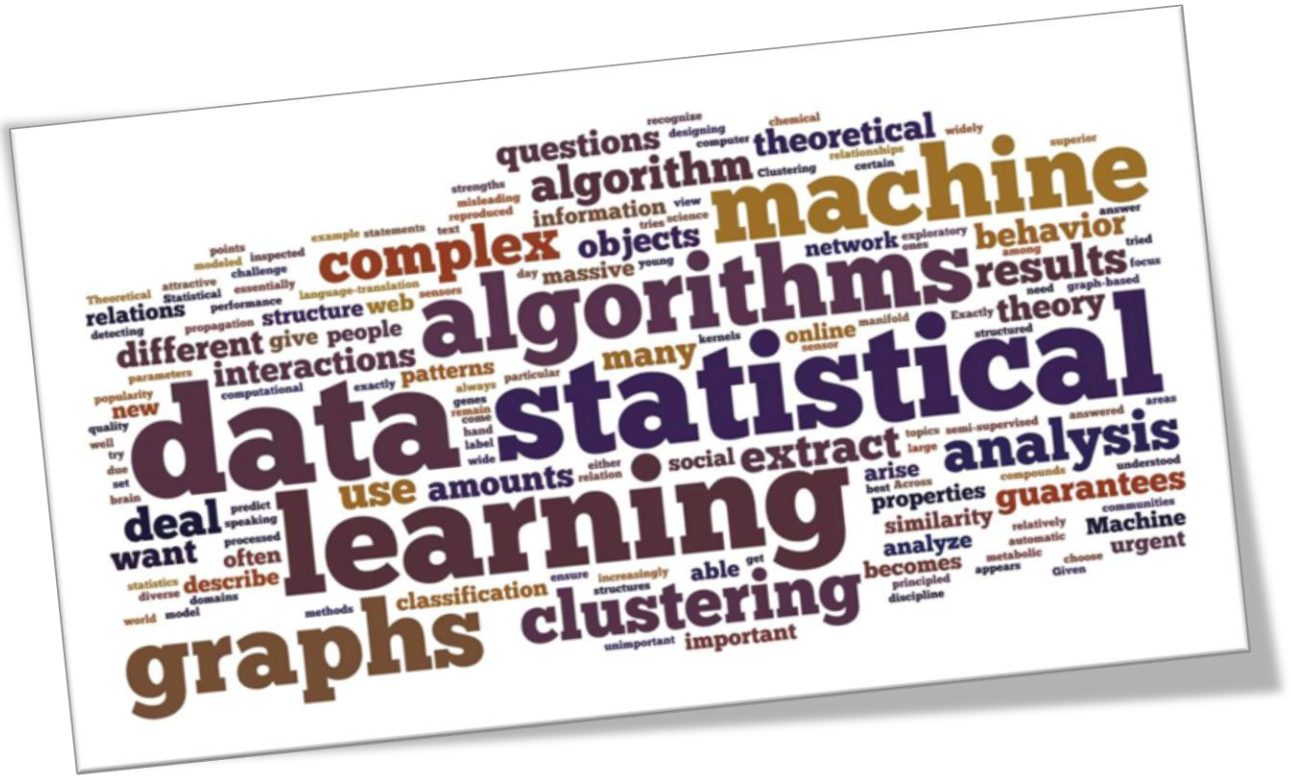

MinION




Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

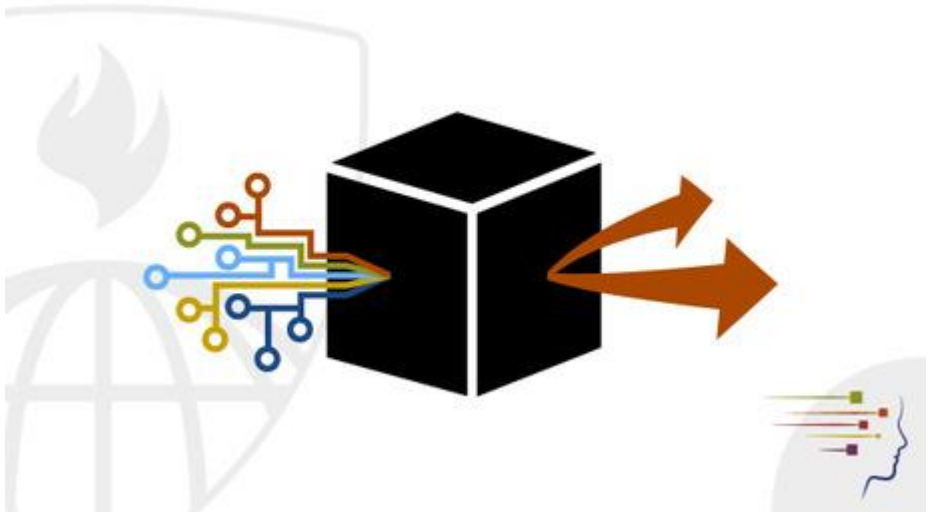
Strojové učení



Strojové učení

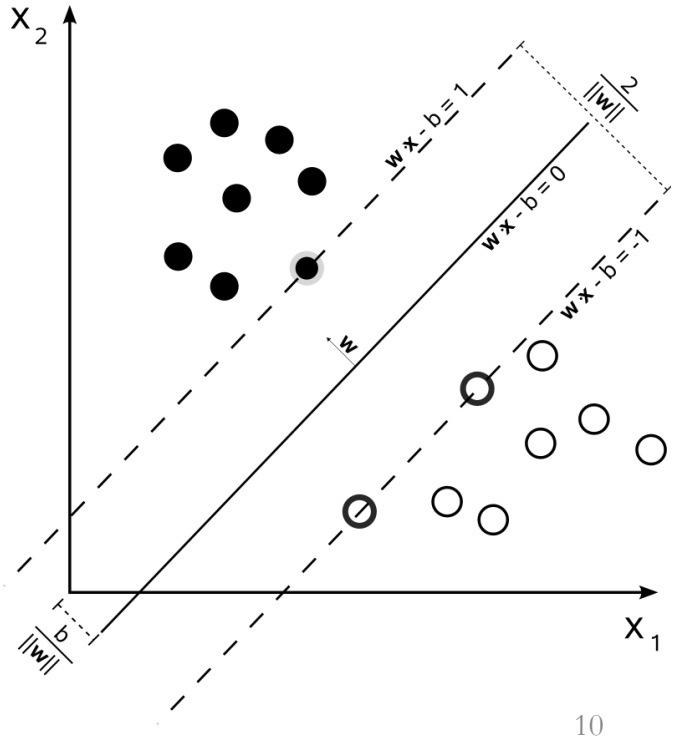
- Množství dat
- Počítačový výkon je dostupný a levný
- Na SU lze nahlížet jako na jiný způsob programování

Rozhodovací procesy, predikce



Strojové učení

- Skupina algoritmů schopných **identifikovat (naučit se) vzory v datech** a tuto znalost **aplikovat na dosud neviděných příkladech**
  - Regrese
  - **Klasifikace**
  - Shlukování
- Typy
  - **Strojové učení s učitelem (supervised learning)**
  - Strojové učení bez učitele (unsupervised learning)
- Typické přístupy
  - Rozhodovací stromy
  - Support vector machines
  - Neuronové sítě
  - Markovovské modely
  - Pravděpodobnostní grafické modely



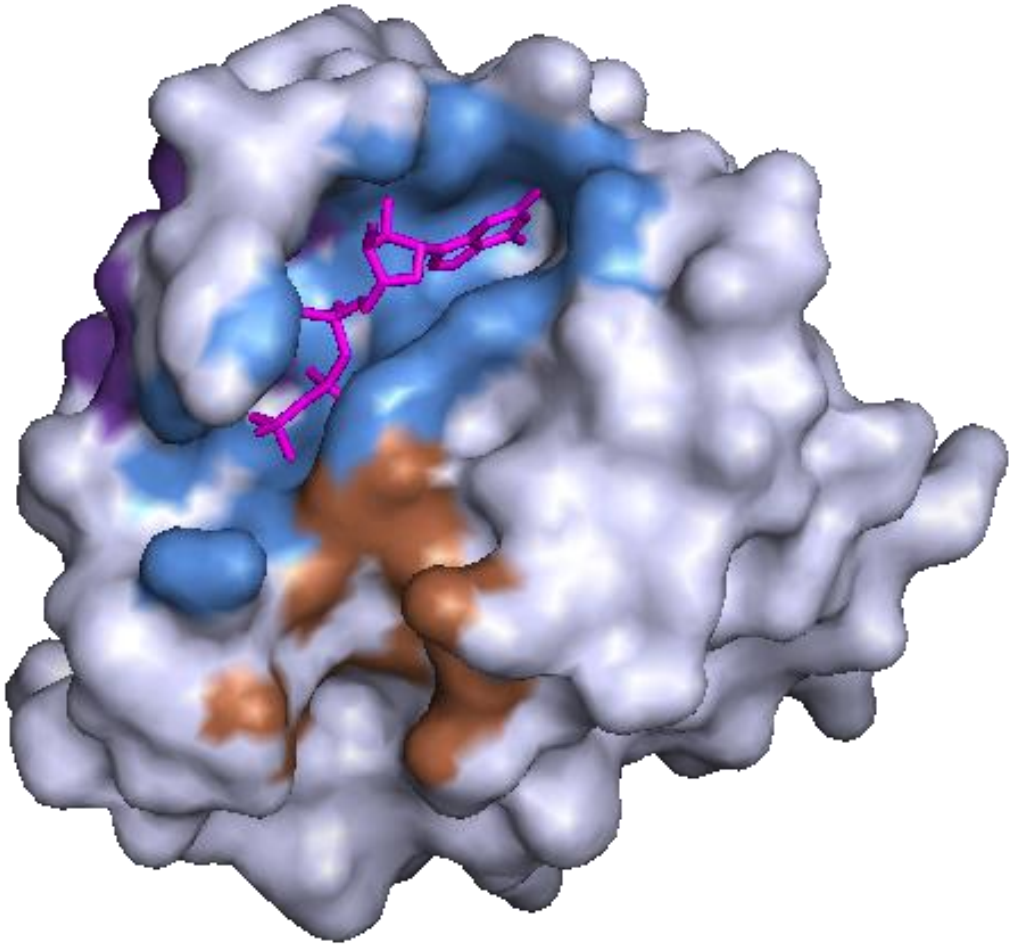
Bioinformatické úlohy řešené strojovým učením

- Identifikace sekvenčních motivů
  - Rozpoznávání a funkční anotace genů
  - Hledání protein-vázajících míst v DNA
  - Rozpoznání sekundární struktury
  - Opravy sekvenačních chyb
- Identifikace signálů
- Predikce vazebných míst na úrovni struktury
- Predikce bioaktivních malých molekul (léčiv)
- …

Identifikace (protein-ligand) aktivních míst na povrchu proteinu

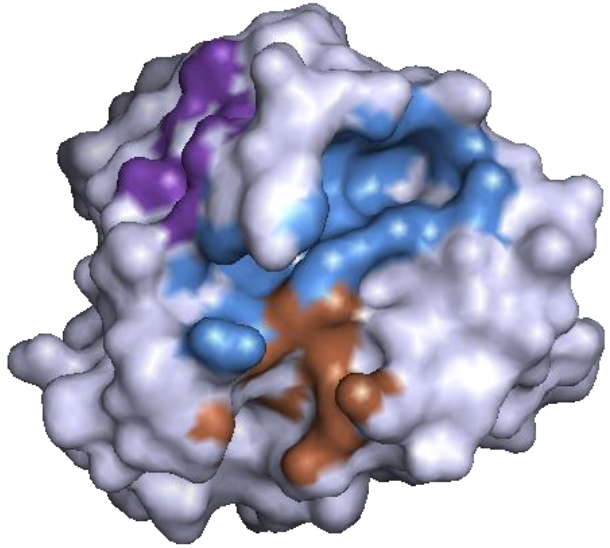
Protein-ligand interakce

- **Aktivní místa (kapsy)**
- **Motivace**
  - Identifikace potenciálních **cílů pro léčiva**
  - Predikce **vedlejších účinků** léčiv
  - **Predikce funkce** neznámého proteinu



P2RANK

- Počítačová metoda (algoritmus) schopný **identifikovat místa na povrchu proteinu**, na které se může s vysokou pravděpodobností **vázat nespecifikovaný ligand**



- **Vstup**: počítačová reprezentace proteinové struktury
- **Výstup**: seznam míst na povrchu proteinu pravděpodobně schopných vázat ligand

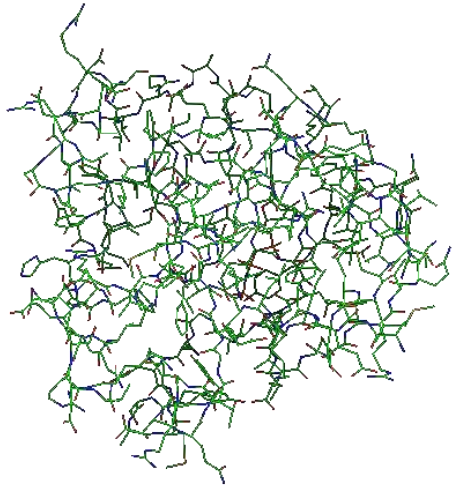
Informatický pohled na protein

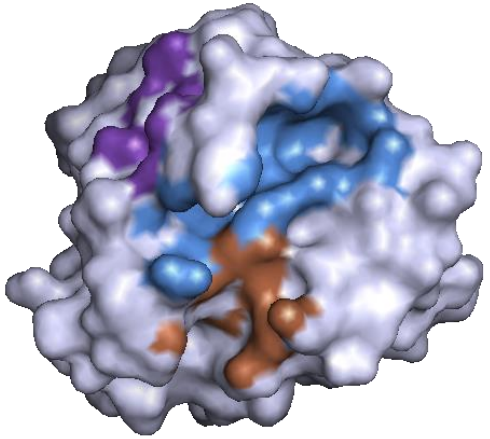
Sekvence

- **Řetěz** aminokyselin → **lineární sekvence** písmen (slovo)
  - Písmena reprezentují aminokyseliny (ARNDCEQGHILKMFPSTWYV)

Struktura

- **Pozice** jednotlivých atomů v 3D prostoru





```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM     32  N   AARG A  -3      11.281  86.699  94.383  0.50 35.88           N
ATOM     33  CA  AARG A  -3      12.353  85.696  94.456  0.50 36.67           C
ATOM     34  C   AARG A  -3      13.559  86.257  95.222  0.50 37.37           C
ATOM     35  O   AARG A  -3      13.753  87.471  95.270  0.50 37.74           O
ATOM     36  CB  AARG A  -3      12.774  85.306  93.039  0.50 37.25           C
ATOM     37  CG  AARG A  -3      11.754  84.432  92.321  0.50 38.44           C
ATOM     38  CD  AARG A  -3      11.698  84.678  90.815  0.50 38.51           C
ATOM     39  NE  AARG A  -3      12.984  84.447  90.163  0.50 39.94           N
ATOM     40  CZ  AARG A  -3      13.202  84.534  88.850  0.50 40.03           C
ATOM     41  NH1AARG A  -3      12.218  84.840  88.007  0.50 40.76           N
ATOM     42  NH2AARG A  -3      14.421  84.308  88.373  0.50 40.45           N
ATOM     43  N   BARG A  -3      11.296  86.721  94.521  0.50 35.60           N
ATOM     44  CA  BARG A  -3      12.333  85.862  95.041  0.50 36.42           C
ATOM     45  C   BARG A  -3      12.759  86.530  96.365  0.50 36.39           C
ATOM     46  O   BARG A  -3      12.924  87.757  96.420  0.50 37.26           O
ATOM     47  CB  BARG A  -3      13.428  85.746  93.980  0.50 36.60           C
ATOM     48  CG  BARG A  -3      12.866  85.172  92.651  0.50 37.31           C
ATOM     49  CD  BARG A  -3      13.374  85.886  91.406  0.50 37.66           C
ATOM     50  NE  BARG A  -3      12.644  85.487  90.195  0.50 38.24           N
ATOM     51  CZ  BARG A  -3      13.114  85.582  88.947  0.50 39.55           C
ATOM     52  NH1BARG A  -3      14.338  86.056  88.706  0.50 40.23           N
```

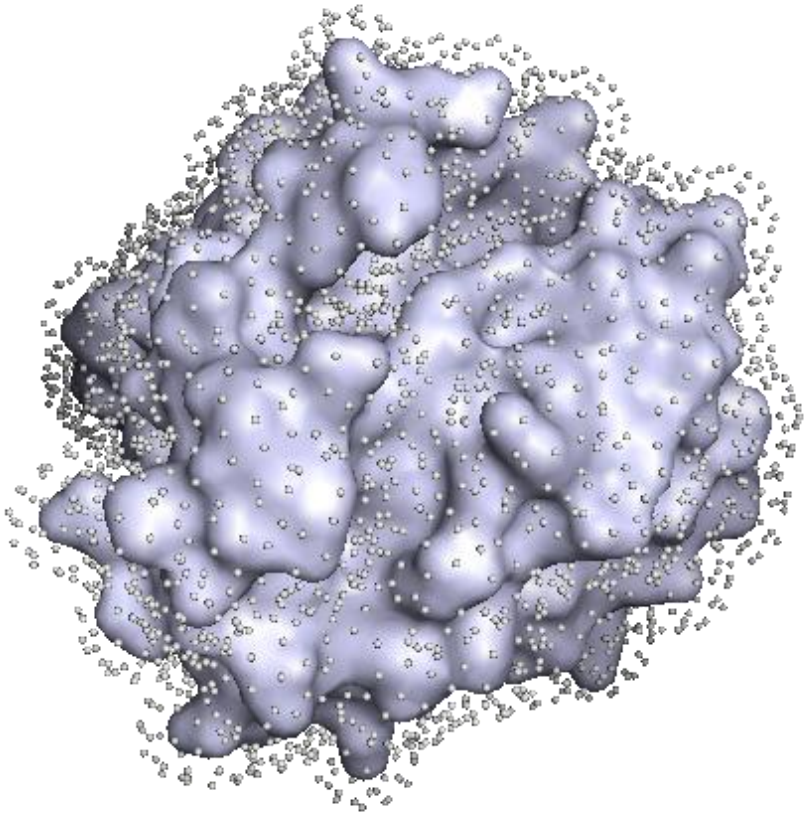
P2RANK princip

- **Využití informací o existujících** aktivních místech a jejich rysech pro **rozpoznání typově podobných** míst na neznámém proteinu

  - **Jak určit kapsu na dosud neviděném proteinu?**
    - Strojové učení (s učitelem)
      - Fáze učení: naučení modelu pro rozpoznání rysů bodů aktivních míst
      - Fáze rozpoznávání: aplikace modelu na povrch neznámého proteinu

  - **Jak popsat rysy povrchu proteinu?**
    - Projekce fyzikálně chemických vlastností aminokyselin na povrch proteinu

Algoritmus – učící fáze



1. Získání známých protein-ligand komplexů

2. Potažení povrchů proteinů **sítí bodů**

3. **Extrakce vektoru fyzikálně-chemických vlastností** pro každý z bodů každého proteinu

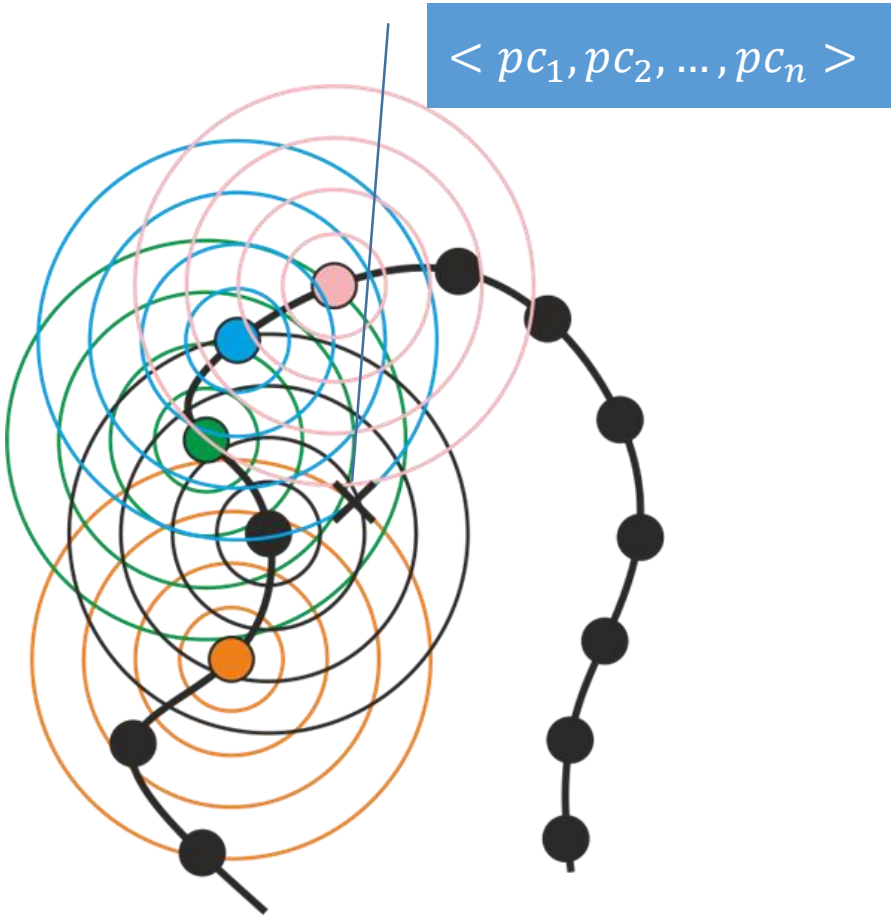
4. **Vybudování modelu**, který pro daný bod (vektor) bude schopný určit, s jakou pravděpodobnostní je tento součástí kapsy = **strojové učení**

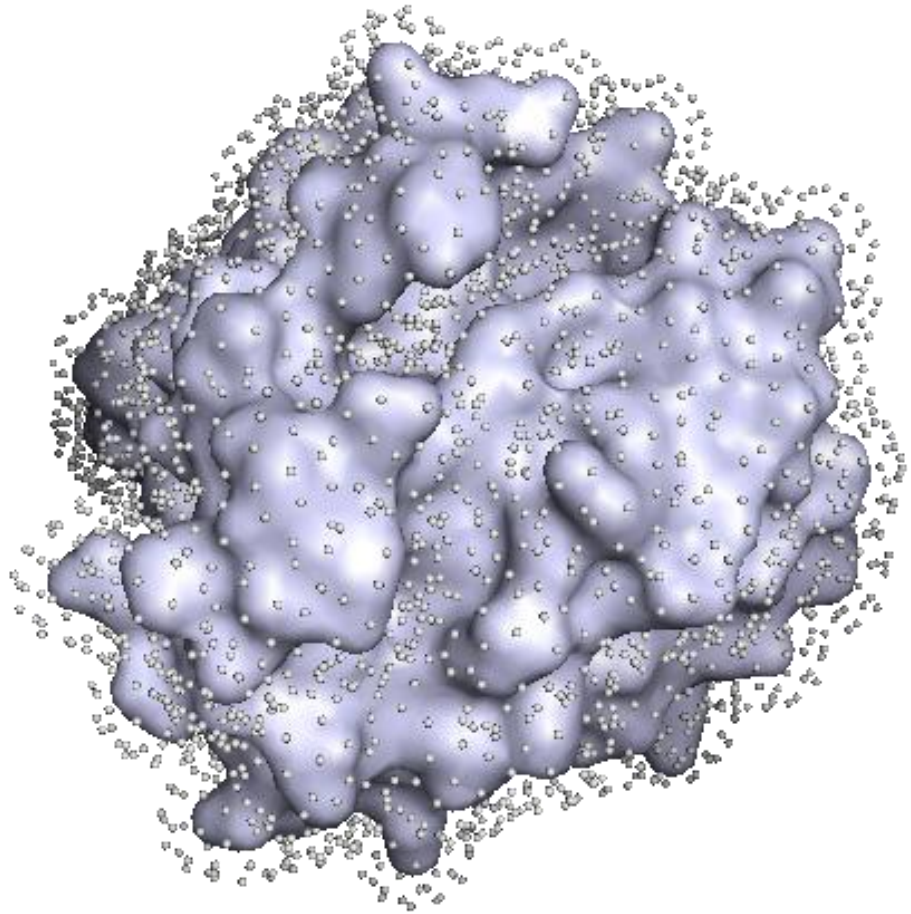
P2RANK - extrakce vlastností

- Okolo **30 atributů** popisující fyzikálně-chemické vlastnosti aminokyselin a lokálního okolí daného bodu

$$\mathrm{PC}(V) = \frac{1}{m} \sum_{\mathrm{A}_i \in \mathrm{A}(V)}^{m} \mathrm{PC}(\mathrm{A}_i) \cdot w(dist(V, \mathrm{A}_i))$$

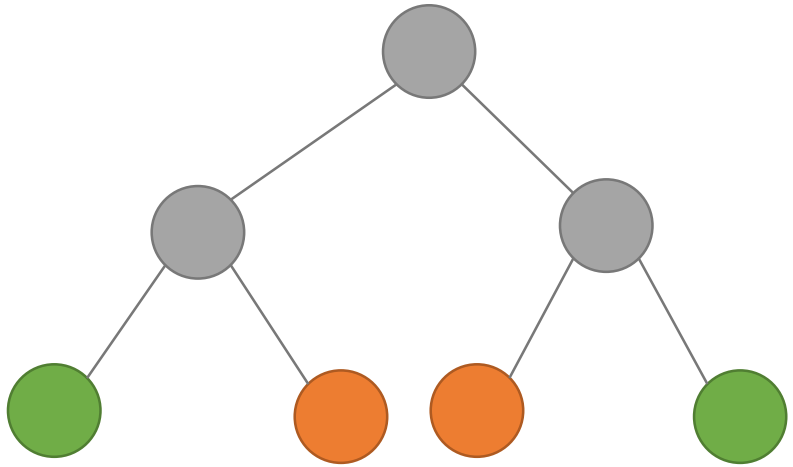
$$w(d) = \begin{cases} 1, & d \leq 4\ \text{Å} \\ (4/d)^2 & d > 4\ \text{Å} \end{cases}$$

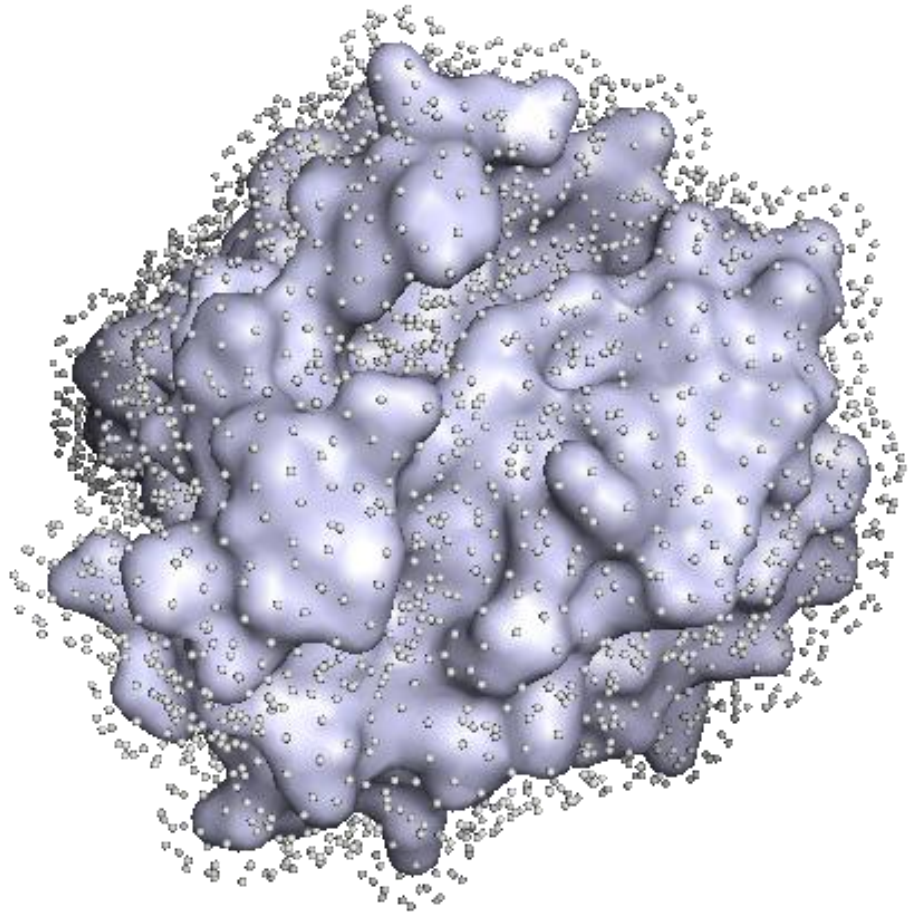




Rozhodovací stromy

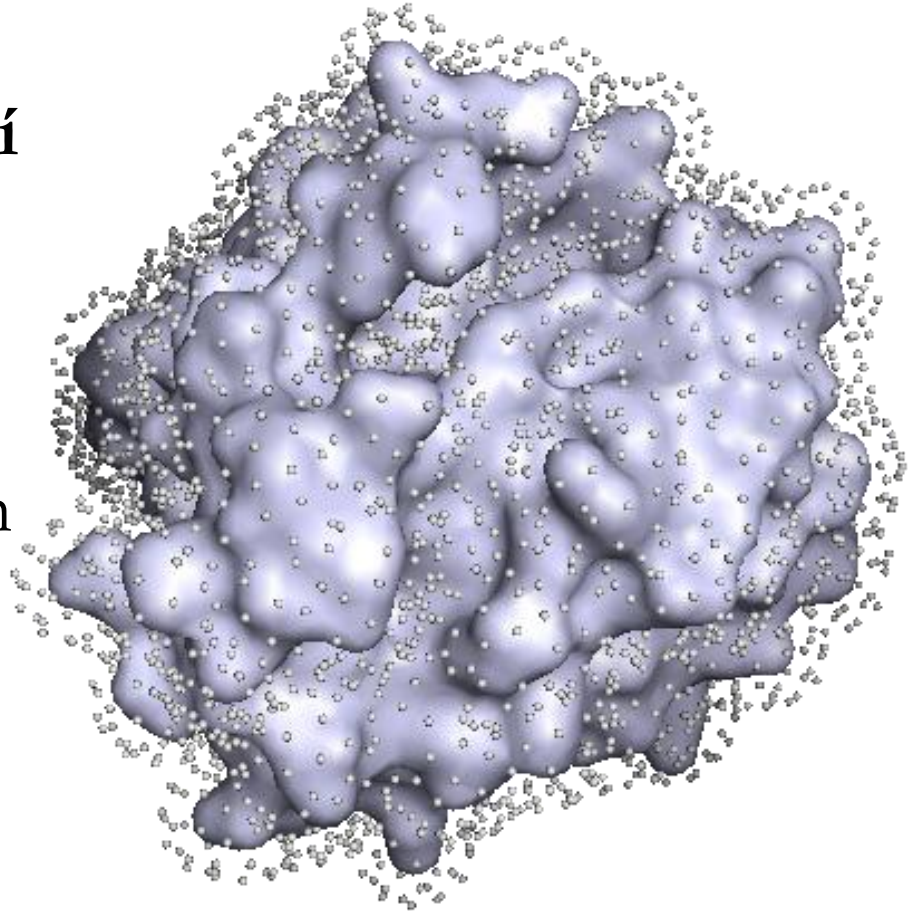
- Model reprezentován **stromem**
  - Vnitřní uzly — pravidla/otázky
  - Hrany — odpovědi
  - Listy — rozhodnutí

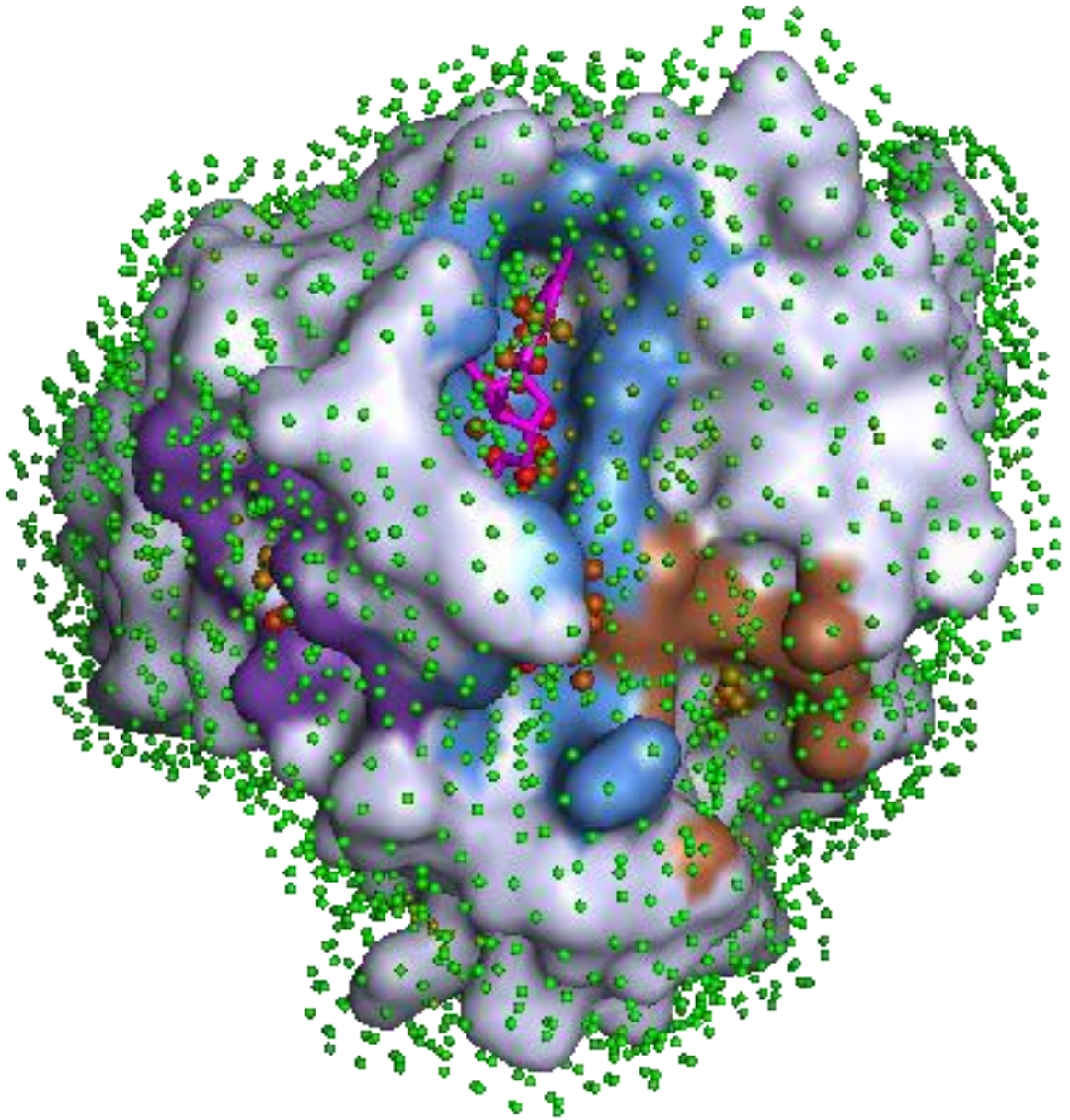


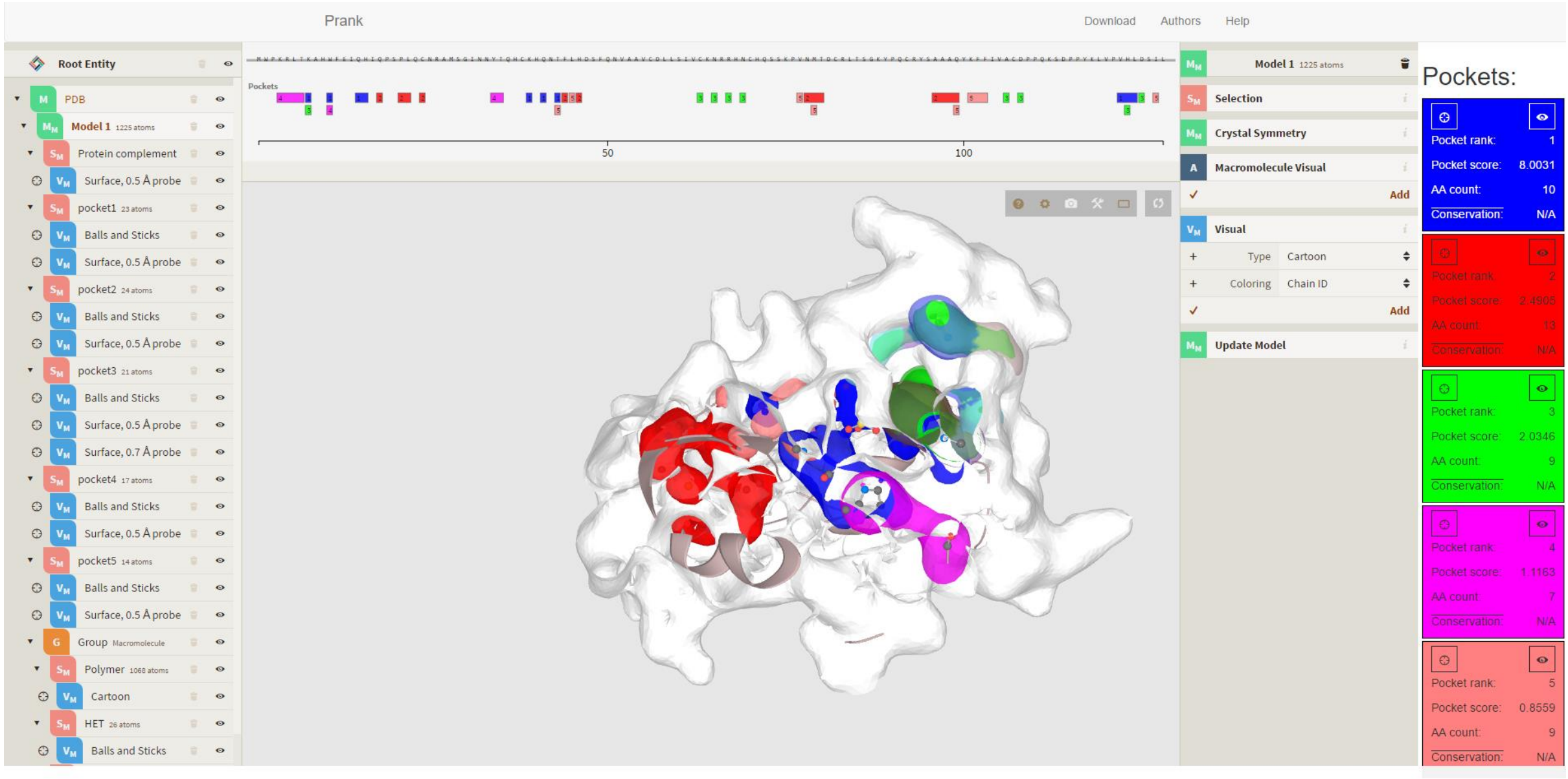


Algoritmus – fáze rozpoznávání

1. **Potažení povrchu** neznámého proteinu **sítí bodů**
2. **Aplikace modelu** pro každý bod sítě → vazebné skóre bodu
3. Vypuštění bodů s nízkým vazebným skórem
4. **Identifikace shluků** vysoce skórujících bodů → **kapsa**
5. **Ohodnocení** kapes součtem skór jejich bodů






Prank
Download
Authors
Help
Root Entity
PDB
Model 1 1225 atoms
Protein complement
Surface, 0.5 Å probe
pocket1 23 atoms
Balls and Sticks
Surface, 0.5 Å probe
pocket2 24 atoms
Balls and Sticks
Surface, 0.5 Å probe
pocket3 21 atoms
Balls and Sticks
Surface, 0.5 Å probe
Surface, 0.7 Å probe
pocket4 17 atoms
Balls and Sticks
Surface, 0.5 Å probe
pocket5 14 atoms
Balls and Sticks
Surface, 0.5 Å probe
Group Macromolecule
Polymer 1068 atoms
Cartoon
HET 26 atoms
Balls and Sticks
Pockets
50
100
Model 1 1225 atoms
Selection
Crystal Symmetry
Macromolecule Visual
Add
Visual
Type Cartoon
Coloring Chain ID
Add
Update Model
Pockets:
Pocket rank: 1
Pocket score: 8.0031
AA count: 10
Conservation: N/A
Pocket rank: 2
Pocket score: 2.4905
AA count: 13
Conservation: N/A
Pocket rank: 3
Pocket score: 2.0346
AA count: 9
Conservation: N/A
Pocket rank: 4
Pocket score: 1.1163
AA count: 7
Conservation: N/A
Pocket rank: 5
Pocket score: 0.8559
AA count: 9
Conservation: N/A

Jaké znalosti jsou třeba pro vývoj P2RANKu?

- Základy biologie, proteomiky
- Znalost zdrojů biologických dat (PDB)
- Programování
- Pokročilá algoritmizace
- Pravděpodobnost a statistika

http://bioinformatika.mff.cuni.cz/

Studijní program | Svět bioinformatiky | Bioinformatický seminář

Bioinformatika je obor, ve kterém se potkává biologie a informatika při studiu biologických dat jako jsou DNA, RNA a proteiny.

Biologie těží ve 21. století z neobyčejného rozvoje metodických přístupů (sekvenování, proteomika), který vedlo a dále vede k bezprecedentnímu nárůstu cenných biologických dat. Je velmi pravděpobné, že se v dohledné době stane sekvenování kompletního genomu standardním diagnostickým vyšetřením a Evropan absolvuje sekvenování svého genomu několikrát za život. I

Dotazy