Hammock: nástroj pro shlukování velkého množství krátkých peptidových sekvencí

Adam Krejčí

RECAMO, Brno, Czech Republic

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

OP Research and
Development for Innovation

Obsah

1 Motivace: peptidové knihovny
2 Shlukování sekvencí
3 Hammock - software pro shlukování sekvencí
4 Experiment - Doména SH3
5 Experiment - Monoklonální protilátky
6 Experiment - Pouze selekce
7 Závěr

Peptidové knihovny

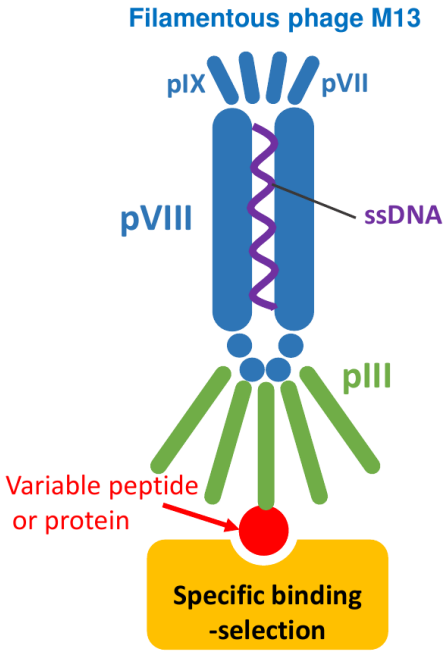
- Laboratorní metody
- Využívají rozsáhlné knihovny obsahující až 10^9 různých náhodných peptidů
- Cíl: selektovat peptidy interagující s *cílem* (protein, DNA, povrch buněk...)
- Phage display, ribosome display, mRNA display...
- Typické využití: mapování epitopů

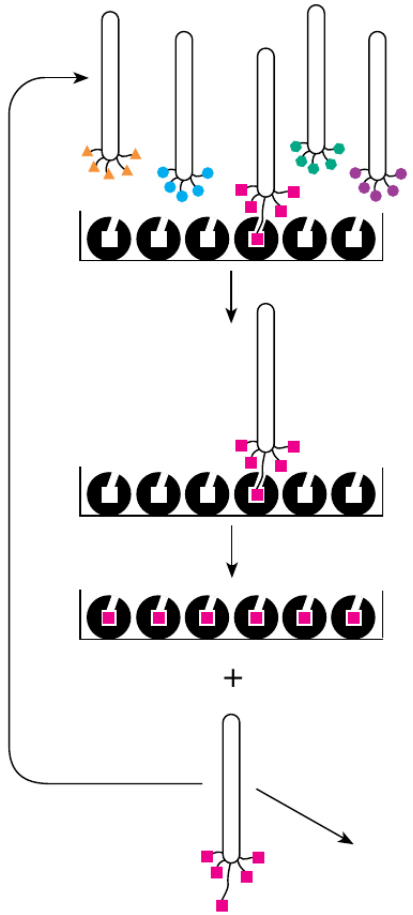
Phage display

- Modifikované fágy nesou krátkou (12 AA) sekvenci na povrchu (= *mimotop*)
- Fág spojuje selektovatelný peptid s čitelnou (DNA) informací
- Protein upevněný na povrchu - *cíl*

- Postup:
  1. Vystavení proteinu fágové knihovně.
  2. Vymyti fágů, které neinteragovaly (nejsou přichyceny na povrchu proteinu).
  3. Uvolnění interagujících fágů a jejich amplifikace v bakteriích.
- Opakování všech kroků. Nakonec sekvenace interagujících fágů.

Phage display



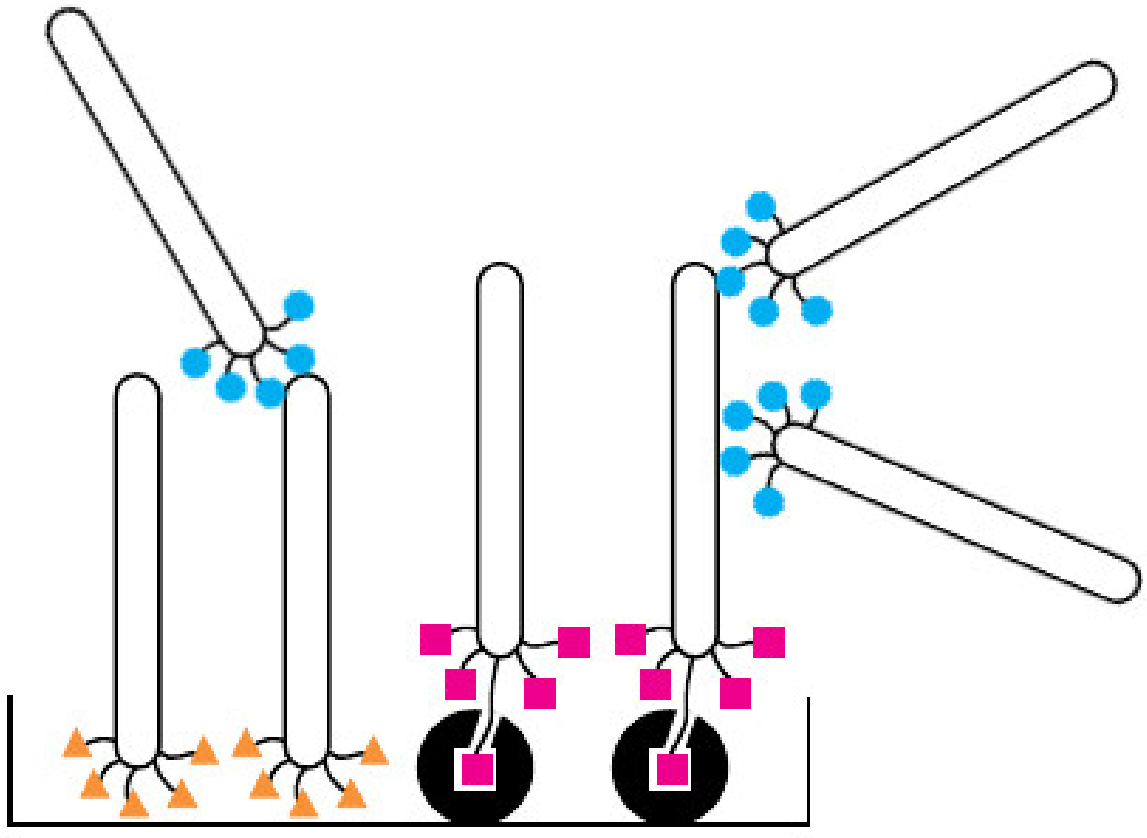


Problémy

Problémy

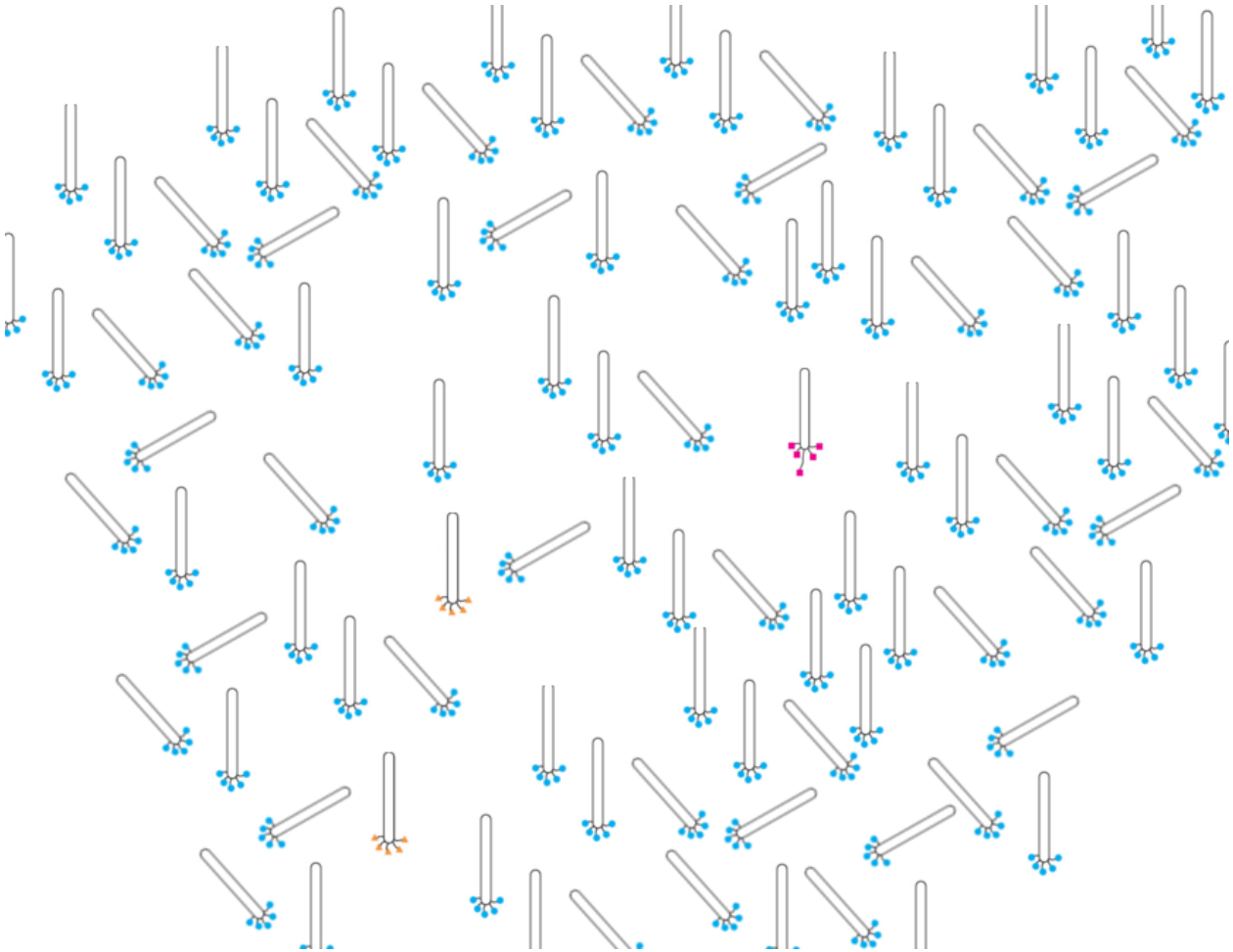
Problémy

- Fágy interagují nespecificky (s povrchem, s jinými fágy...)



Problémy

- Rozdíly v úspěšnosti amplifikace fágů (*"parazitní fágy"*)



Využití NGS

- Výrazně větší množství dat

  $\rightarrow$ Méně kol selekce/amplifikace

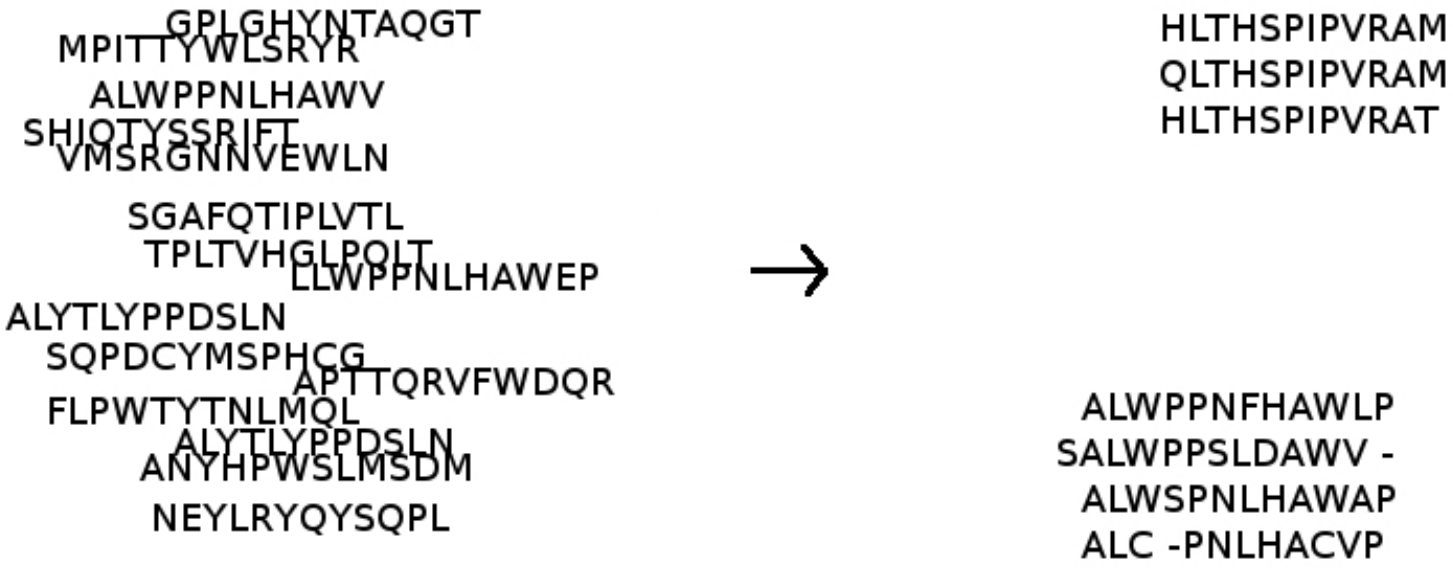
  $\rightarrow$ Sekvenování po každém kole selekce/amplifikace
- Více různých interagujících sekvencí $\rightarrow$ určení vazebného motivu
- Možnost zkoumat více interakcí naráz

  $\rightarrow$ vícenásobně specifické proteiny, proteinové komplexy, celé buňky, krevní sérum...

  $\rightarrow$ Možnost získat řádově miliony unikátních sekvencí

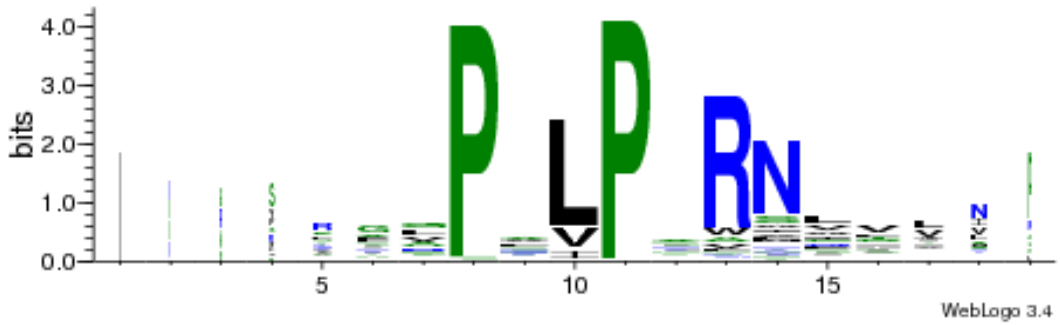
  Jak identifikovat významné sekvenční motivy reprezentující jednotlivé interakce v tak rozsáhlých datech?

Shlukování sekvencí

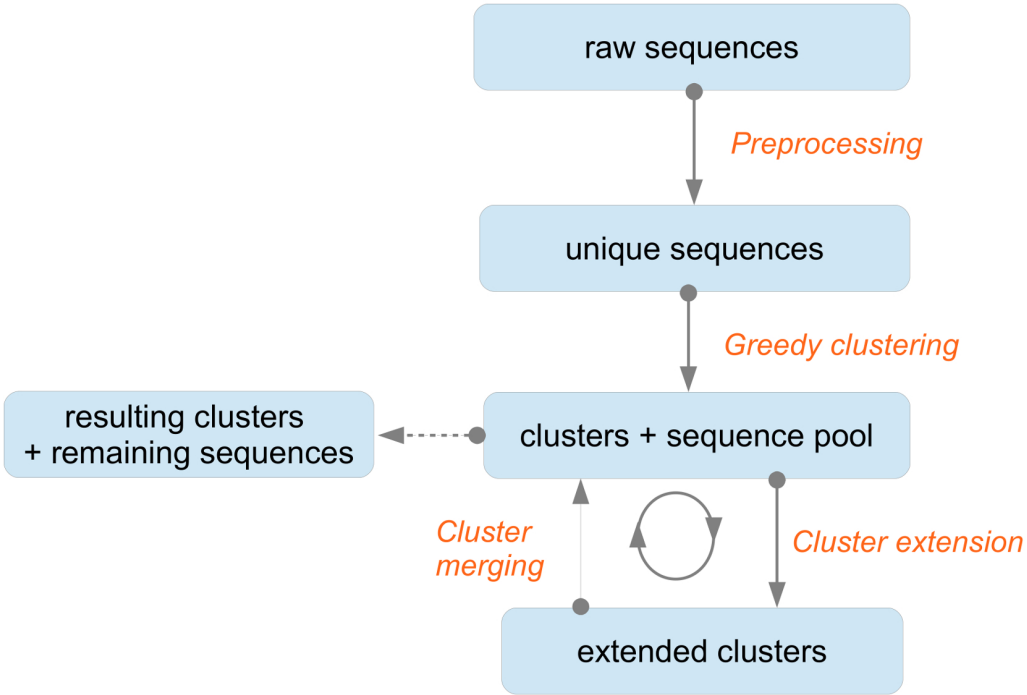


Shluky sekvencí

```
---AIMPEVPQRNWV--
-----APALPLRNDGRY
-----APVLPARIWYDV
---DLGPLLPSRNVS--
----ERPLLPVRNLGI-
----FGPALPARWGGV-
----GIPEIPAGNWAV-
----GIPEIPARNWAV-
-----GPALPARTVGIL
QRMWFLPAVPGL-----
---RAGPMLPDKNLY--
---RAGPMLPDRSLY--
----RGPALPLRVLLH-
----RGPALPLRVLLR-
----RGPVLPLRVLLH-
```



Hammock - software pro shlukování sekvencí



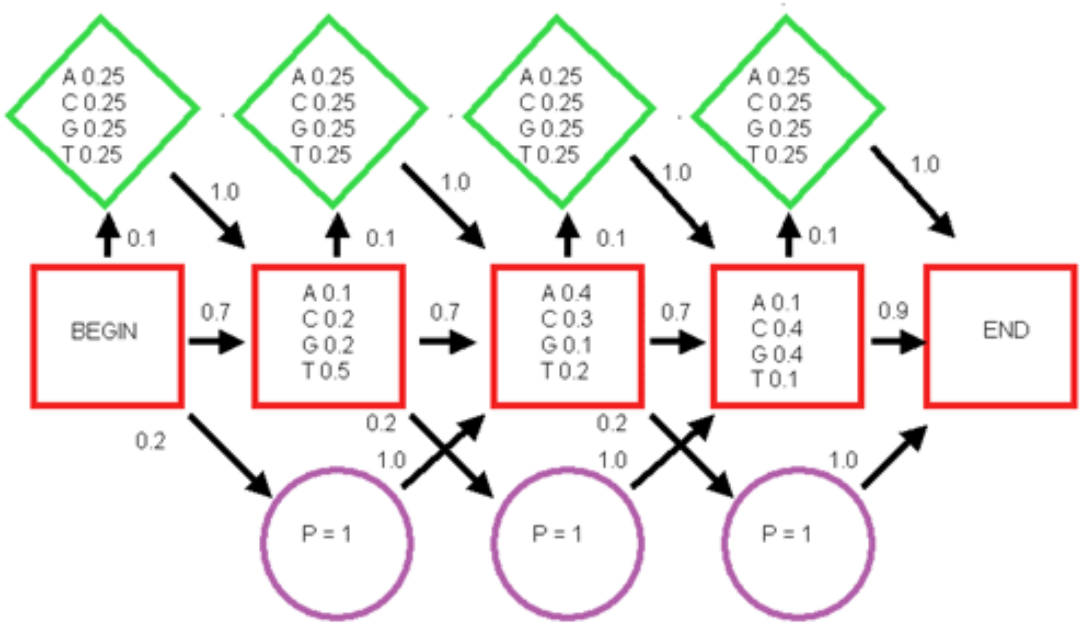
KREJCI, A. et al. Hammock: a hidden Markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets.
Bioinformatics. sep 2015, s. btv522, doi: 10.1093/bioinformatics/btv522

Postup algoritmu

1. Vytvoření iniciálních shluků (”jader”) hladovým algoritmem
2. Vybrána největší jádra, zbytek sekvencí chápán jako neshlukované
3. Shluky (jádra) reprezentovány pomocí HMM
4. Několik iterací:
   1. Rozšiřování shluků - prohledání zbylých sekvencí pomocí HMM
   2. Slučování shluků - s využítím porovnání HMM-HMM
5. Výsledek: Finální shluky + zbylé (neshlukované) sekvence

HMM - Skrytý Markovův Model

- Účel: zarovnání + skóre MSA vs. sekvence
- Stochastický konečný automat s výstupem
- Původní využití: rozpoznání signálu
- varianta pHMM (profile HMM) určena pro reprezentaci MSA
- 3 druhy stavů - match, insertion, deletion
- Rozdíl oproti PSSM - lze reprezentovat vnitřní mezery



Hladový (greedy) shlukovací algoritmus

1. Seřazení sekvencí
2. 1. sekvence je reprezentantem 1. shluku
3. 2. až n. seknvence: porovnání se všemi reprezentanty shluků, následně:
   - Dostatečné skóre $\rightarrow$ vložena do nejpodobnějšího shluku
   - Nedostatečné skóre $\rightarrow$ označena jako reprezentant nového shluku

- $O(n^2)$
- Využití substituční matice
- Omezené zarovnání bez vnitřních mezer

Rlozšiřování shluků

- Hmmer
- Iterativní přidávání nalezených sekvencí do MSA
- Omezení délky zarovnání, počtu *match* stavů HMM, entropie pozic MSA, počtu vnitřních mezer

FINN, R. D. – CLEMENTS, J. – EDDY, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*. 2011, 39, 2, s. 29–37.
ISSN 0305-1048

Slučování shluků

- Založeno na HMM-HMM porovnání (HH-suite)
- Kompletní aglomerativní hierarchické shlukování (shluků)
- Výpočetně náročné ($O(n^3)$)
  - $\rightarrow$ Heuristika
  - $\rightarrow$ Využití informací z minulého kroku
  - $\rightarrow$ Nalezené sekvence vypovídají o podobnosti shluků
- Omezení délky zarovnání, počtu *match* stavů HMM, entropie pozic MSA, počtu vnitřních mezer

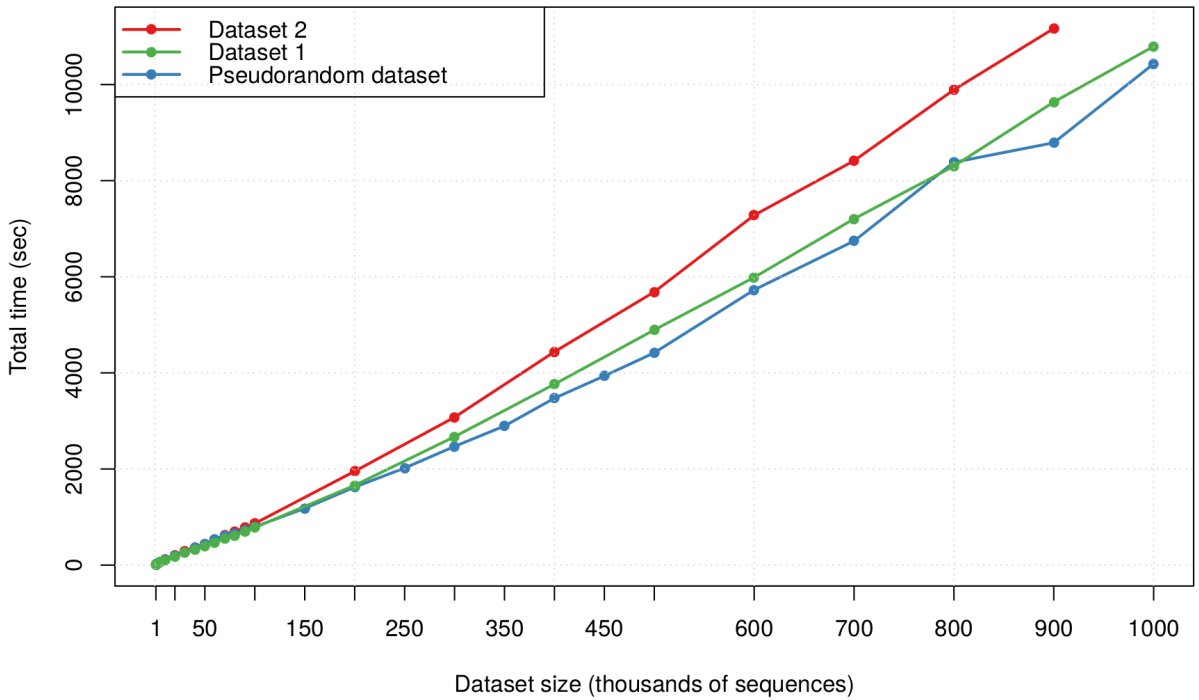
SÖDING, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005, 21, 9, s. 2144.
ISSN 1367-4803

Implementace

- Java + externí procesy
- Všechny kroky paralelizovány
- Modul pro Galaxy

Rychlost

- 10^6 sekvencí za $\sim$ 3h (Desktop, 8 jader)

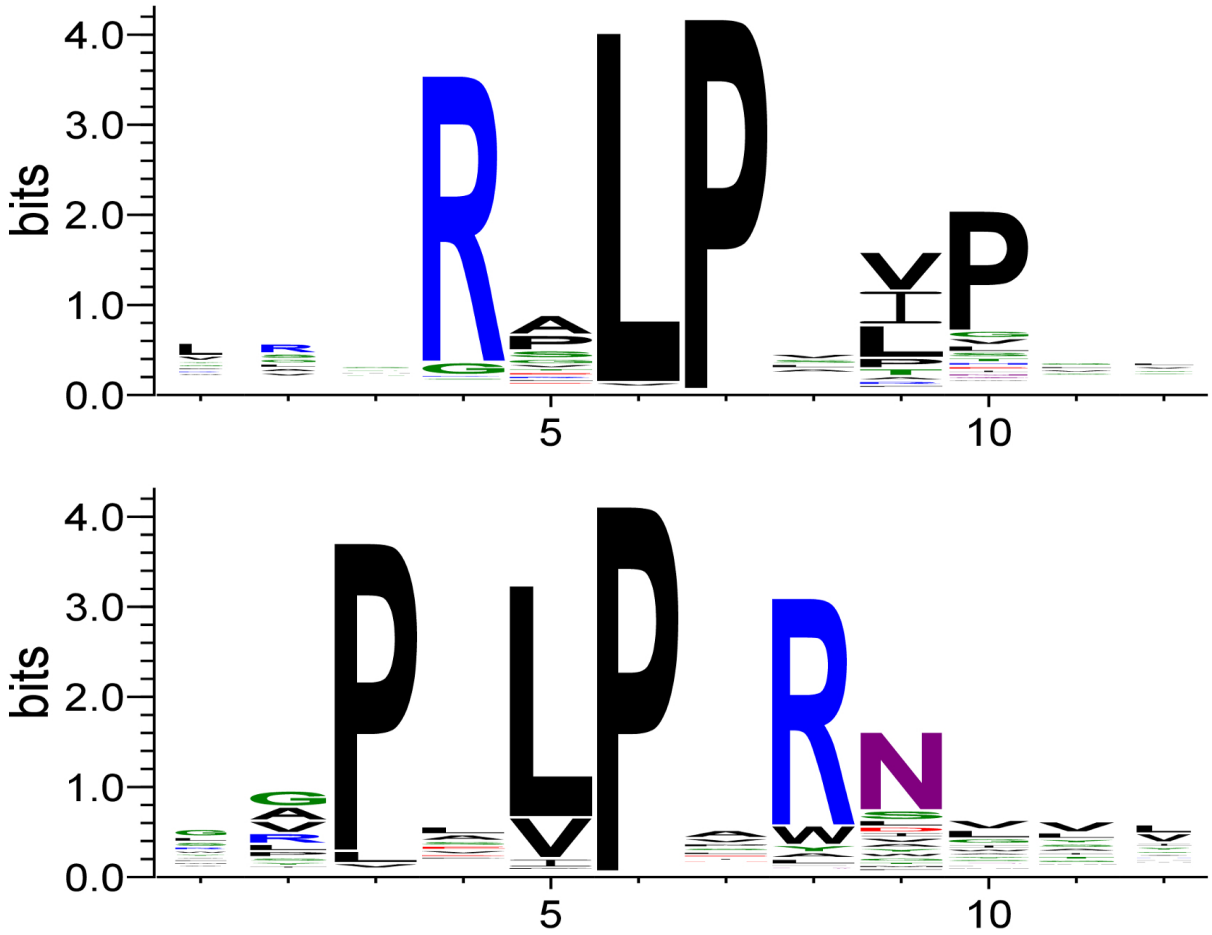


Obsah

1 Motivace: peptidové knihovny
2 Shlukování sekvencí
3 Hammock - software pro shlukování sekvencí
4 Experiment - Doména SH3
5 Experiment - Monoklonální protilátky
6 Experiment - Pouze selekce
7 Závěr

Experiment - Doména SH3

- Malý soubor dat (2457 sekvencí)
- Dříve použito pro testování obdobných nástrojů
- Phage Display, Src SH3 doména
- Doména váže 2 motivy


bits
4.0
3.0
2.0
1.0
0.0
5
10
bits
4.0
3.0
2.0
1.0
0.0
5
10

Porovnání výsledků

Nástroj	Čas	Počet shluků	Počet sekvencí	KLD match	KLD all
Hammock	17 s	2	2153	28.165	24.858
Gibbs	39 min	2	2450	25.151	22.369
MUSI	17 s	2	2456	20.96	19.89

KIM, T. et al. MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets.
Nucleic Acids Research. 2012, 40, 6

ANDREATTA, M. – LUND, O. – NIELSEN, M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach.
Bioinformatics. 2013, s. 8 – 14.
ISSN 1367-4803

Obsah

1 Motivace: peptidové knihovny
2 Shlukování sekvencí
3 Hammock - software pro shlukování sekvencí
4 Experiment - Doména SH3
5 Experiment - Monoklonální protilátky
6 Experiment - Pouze selekce
7 Závěr

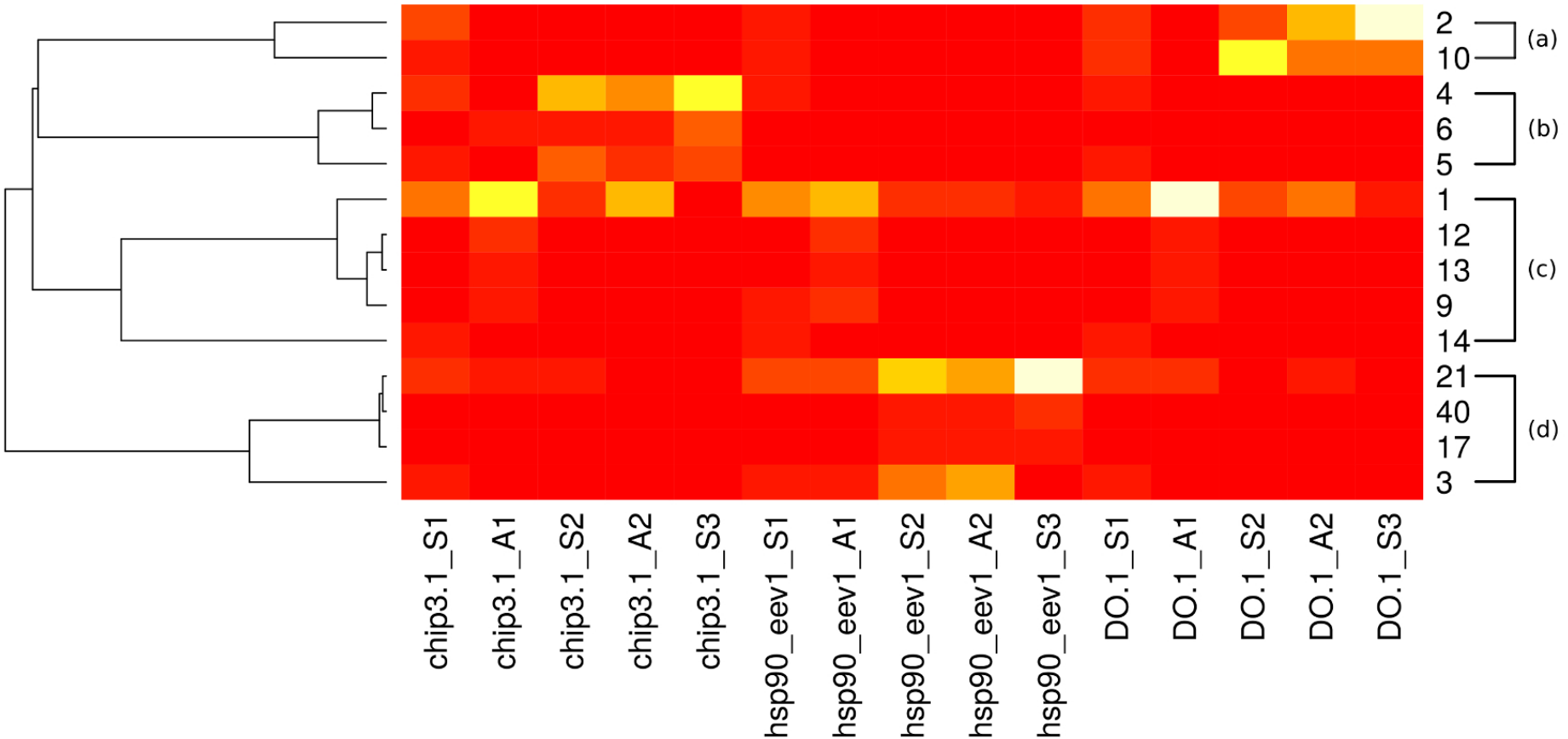
Experiment - Monoklonální protilátky

- Phage display, selekce pomocí 3 monoklonálních protilátek
- 3 kola selekce, 2 kola amplifikace
- Sekvenováno po každém S/A kole
- 74 041 unikátních sekvencí, 316 119 sekvencí celkem

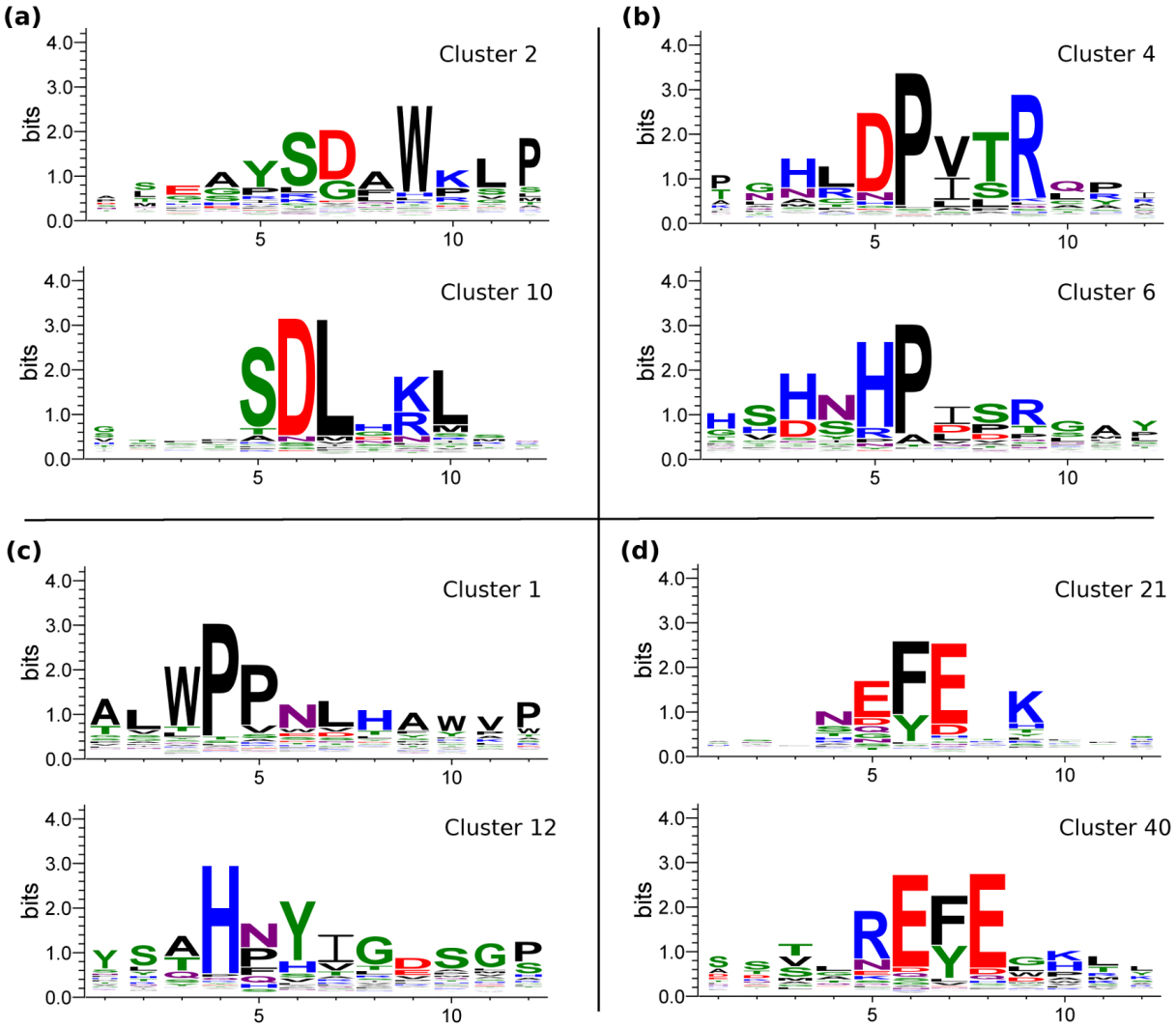
Teoreticky...

- Shluky sekvencí specificky rozpoznaných protilátkou
  $\rightarrow$ Obohaceny pouze v selekčních krocích této protilátky
- Shluky sekvencí s nespecifickou vazbou:
  $\rightarrow$ Obohaceny ve všech selekčních krocích
- Shluky fágů, které se lépe amplifikují
  $\rightarrow$ Obohaceny ve všech amplifikačních krocích

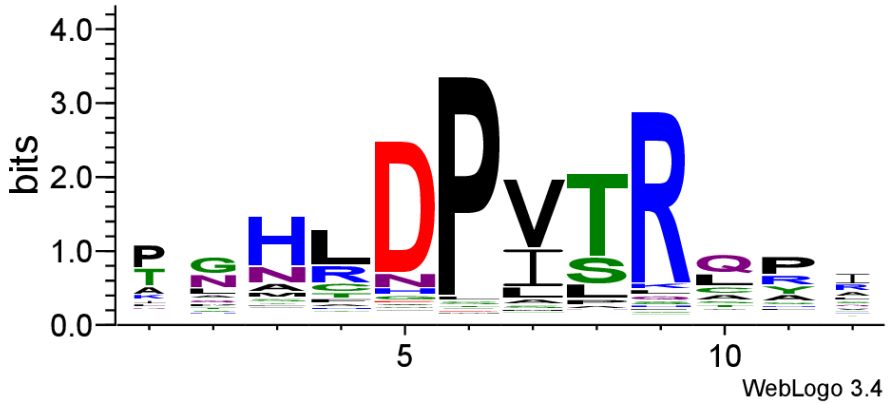
Největší shluky



Největší shluky

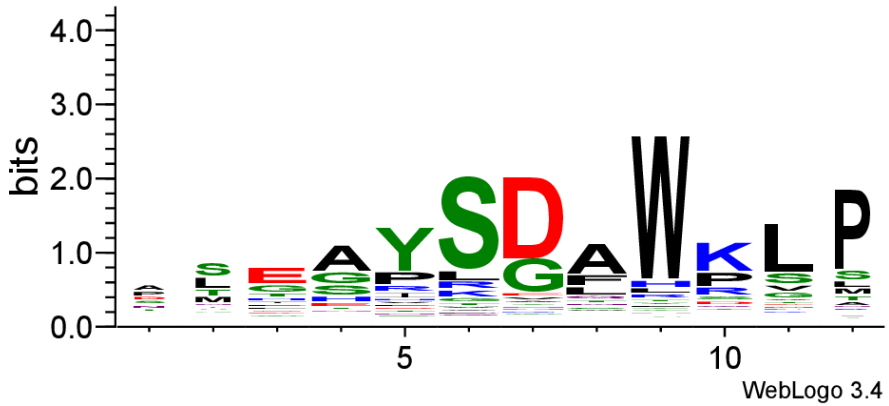


Největší shluky



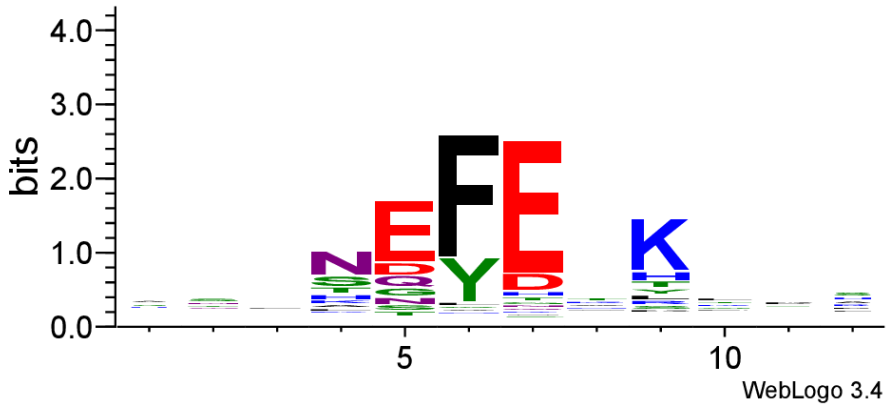
CHIP - FDPVTR

Největší shluky



p53 - FSDLWKLL

Největší shluky



HSP90-alpha - LKEFEGKT

Porovnání výsledků

Nástroj	Čas	Počet shluků	Počet sekvencí	KLD match	KLD all
Hammock	2 m 35 s	74	14421	17.897	17.635
Gibbs	>72 h	-	-	-	-
Gibbs -g 100	14 h 13 m	100	74040	12.622	11.335
MUSI	8 h 20 m	1	74041	0.0	3.22

- Shluky obsahují:

  14 421 (19.5 %) unikátních skvencí

  316 119 (81.1 %) sekvencí celkem

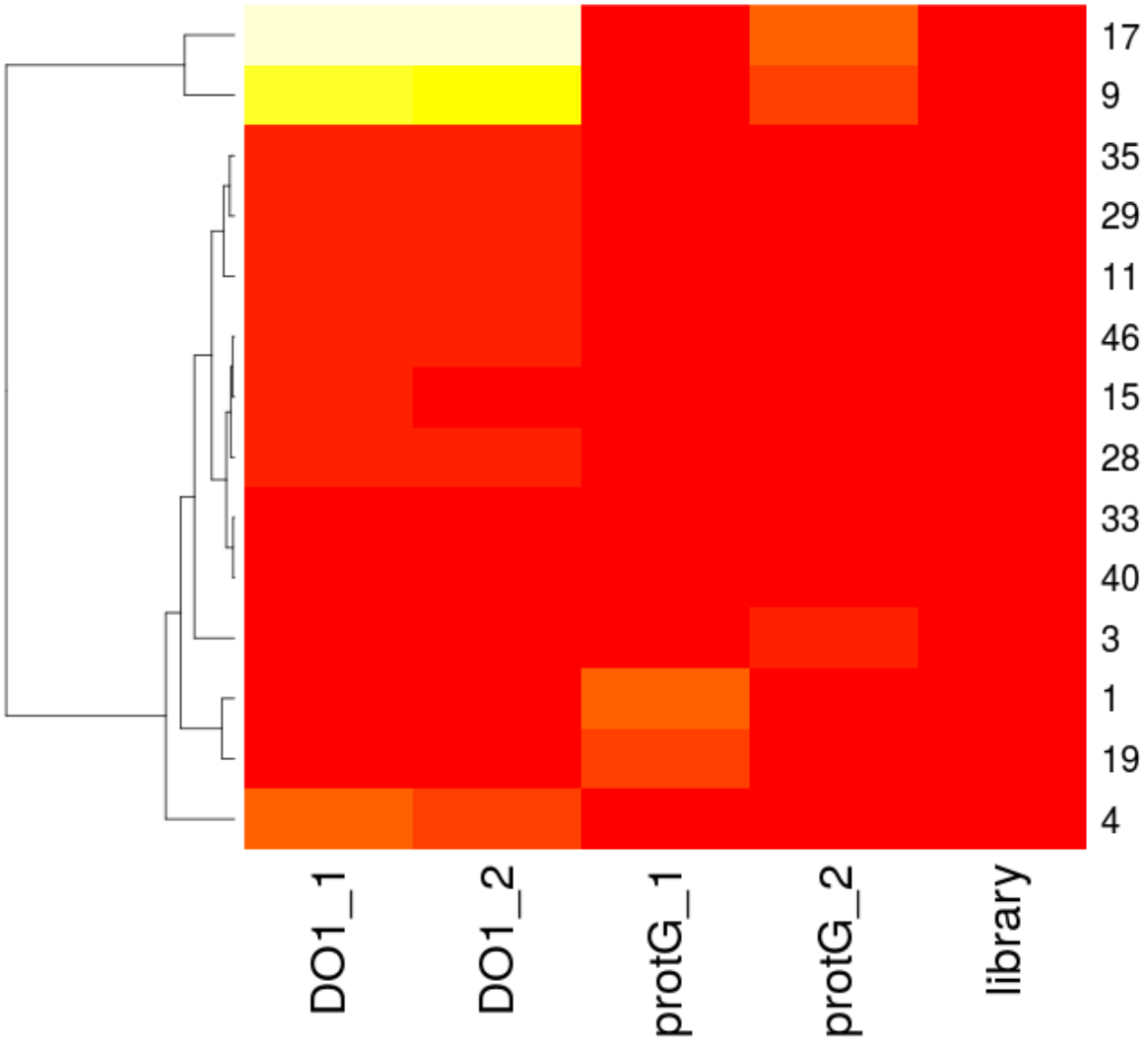
Obsah

1 Motivace: peptidové knihovny
2 Shlukování sekvencí
3 Hammock - software pro shlukování sekvencí
4 Experiment - Doména SH3
5 Experiment - Monoklonální protilátky
6 Experiment - Pouze selekce
7 Závěr

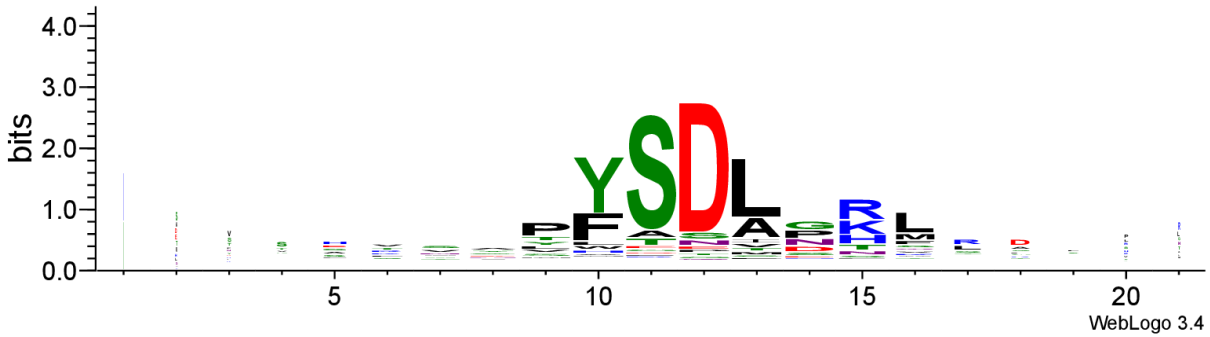
Experiment - Pouze selekce

- Phage display, selekce pomocí protilátky DO-1
- pouze 1 kolo selekce, žádná amplifikace
- Duplikát, sekvenování knihovny + čistého G-proteinu
- 690 491 unikátních sekvencí

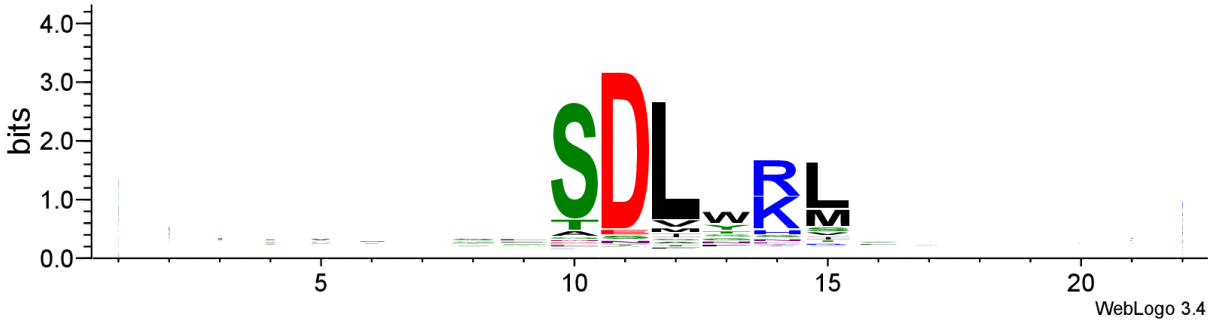
Největší shluky



Největší shluky



Cluster 4



Cluster 17

Obsah

1 Motivace: peptidové knihovny

2 Shlukování sekvencí

3 Hammock - software pro shlukování sekvencí

4 Experiment - Doména SH3

5 Experiment - Monoklonální protilátky

6 Experiment - Pouze selekce

7 Závěr

Shrnutí

- Hammock shlukuje krátké peptidové sekvence
- Až 10^6 sekvencí na desktopu
- Umožňuje provádět Phage display experimenty bez amplifikace
- **Aplikace**
  - Phage display - charakterizace protilátek
  - Phage display - polyklonální sérum po vakcinaci
  - Shlukování CDR3 sekvencí
  - ...

Poděkování

RECAMO
Regional Centre for Applied Molecular Oncology

- Petr Müller
- Borivoj Vojtesek

Edinburgh Cancer Research Centre

- Ted R. Hupp
- Anne-Sophie Huart

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA

- Matej Lexa