

IV107 Bioinformatika I

Přednáška 6

Katedra informačních technologií
Masarykova Univerzita Brno

Jaro 2023

/mnt/lexa/Pictures/logo

Předchozí týden

- ▶ GenBank
- ▶ UniProt
- ▶ PDB
- ▶ Gene Ontology
- ▶ KEGG Pathways
- ▶ genomické a proteomické databáze

Vizualizace proteinů

- ▶ QuickPDB (Java) & Co. - $\dot{\zeta}$ Molsoft web app (<https://www.molsoft.com/pdbv.html>)
- ▶ PyMol (Python)
- ▶ VMD+Povray (<https://www.youtube.com/watch?v=wJo3zjRAQoo>)

PovRay raytracing – používá CSG constructive solid geometry

```
sphere{  
  < 0,0,0 >, 180  
  pigment{colorYellow}  
}  
cylinder{  
  < 0,0,0 >, < 150,200,300 >, 60  
  pigment{colorWhite}  
}  
camera{  
  location < 0.0,0.0,800.0 >  
  direction < 0.0,0.0,-1.0 >  
}  
light_source{< 0,0,1000 > colorWhite}
```

Analýza proteinové sekvence

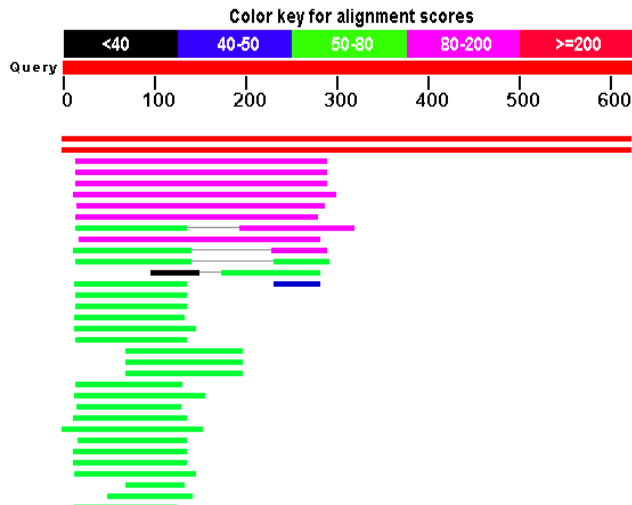
- ▶ strukturní
 - ▶ predikce domén
 - ▶ predikce sekundární struktury
 - ▶ predikce a modelování 3D
 - ▶ homologní
 - ▶ "threading"/"fold recognition"(navlékání)
 - ▶ z fragmentů
 - ▶ ab initio
- ▶ funkční (anotace)
 - ▶ přenos funkce sekvencí podobností (BLAST + GO)
 - ▶ podle příslušnosti k rodině proteinů
 - ▶ podle obsahu motivů (PRINTS—BLOCKS + GO)

001 masaqsfynqssvlkinvmvdddhvfldimsrmlqhskyrdpsvmeiaviav
061 stlkiqrnidliitdyympgmnglqlkkqitqefgnlpvlvmsstdtnkeees
121 fipkpihptdltkiyqfalsnkrngkstlsteqnhkdadvsvpqqitlvpeqa
181 kncsfkdsrvtvnstngscvstdgsrknrkrkpnnggpsddgesmsqpakkkki
241 dlflqairhigldkavpkkilafmsvpyltrenvashlqkyriflrrvaeqgl
301 gidsmfrqthikepyfnyytpstswydt rlnnrsfyskpvhgfgqskllsttr
361 mpynymnrsstyephriqsgsnltlpiqsnlsfpnqpsqneerrsfepvma
421 qvlqfgqlgpsaisghnfnnnmtsrygslipsqpgpshfsygmqsflnnevnt
481 nattqpnldelpqlenlnlyndfgntselpynisnfqfddnkhqqgeadptkf
541 stelnhedgdwtfvninqggsngetsntiaspetntpilninhnqngqgdvp
601 ldpqelvddd fmns lfnndmn

Metody predikce domén

- ▶ vyskytují se ve mnoha proteinech (BLAST)
- ▶ kostra mezi doménami je flexibilní
- ▶ vlastnosti aminokyselin se liší podle pozici vůči prostředí
- ▶ motivy v rámci jedné domény spolu souvisí

Identifikace domén na základě podobnosti (BLAST)



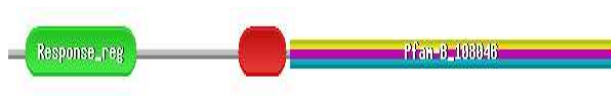
Identifikace domén na základě podobnosti (BLAST + CDD)



<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>

/mnt/lexa/Pictures/logo

Identifikace domén na základě podobnosti (BLAST + PFAM)



Source	Domain	Start	End
PfamA	Response_reg	16	128
PfamA	Myb_DNA-binding	224	274
PfamB	Pfam-B_108046	276	592

Frekvence aminokyselin na rozhraní domén

Table I. Linker propensities

	All	1-linker	2-linker	3-linker	Small	Medium	Long	Helical	Non-helical
Pro	1.299	1.362	1.266	1.332	1.241	1.314	1.309	0.8	1.816
Arg	1.143	1.129	1.137	1.069	1.131	1.132	1.154	1.239	1.038
Phe	1.119	1.122	1.11	0.981	1.368	1.121	1.058	1.09	1.151
Leu	1.085	1.11	1	1.193	1.192	1.106	0.994	1.276	0.885
Glu	1.051	1.054	1.139	0.992	0.736	1.053	1.115	1.199	0.9
Gln	1.047	1.092	0.916	1.111	0.861	0.999	1.2	1.124	0.968
Met	1.032	0.923	1.077	0.998	1.369	1.093	0.782	1.171	0.878
Thr	1.017	1.023	1.018	0.992	0.822	0.988	1.11	0.832	1.189
His	1.014	0.949	1.109	1.034	0.973	1.054	0.992	1.012	1.05
Tyr	1	0.902	1.157	1.12	0.836	1.09	0.866	1.075	0.945
Ala	0.964	0.974	0.938	1.042	1.065	0.99	0.892	1.092	0.843
Val	0.955	0.923	0.959	1.001	1.14	0.957	0.9	0.908	0.999
Ser	0.947	0.932	0.956	0.984	1.097	0.911	0.986	0.886	1.003
Asn	0.944	0.988	0.902	0.828	0.762	0.873	1.144	0.927	0.956
Lys	0.944	0.946	0.952	0.979	0.478	1.003	0.944	1.008	0.893
Ile	0.922	0.928	0.986	0.852	1.189	0.95	0.817	0.912	0.946
Asp	0.916	0.892	0.857	0.97	0.836	0.915	0.925	0.919	0.906
Trp	0.895	0.879	0.971	0.96	1.017	0.939	0.841	0.981	0.852
Gly	0.835	0.845	0.892	0.743	1.022	0.785	0.917	0.698	0.978
Cys	0.778	0.972	0.6856	0.5	1.015	0.644	1.035	0.662	0.896

Převzato z George and Heringa (2002)

DSSP je standardem přiřazení sekundární struktury proteinům v PDB

- ▶ helix

- H alpha helix
- G 3-helix (3/10 helix)
- I 5 helix (pi helix)

- ▶ strand

- B residue in isolated beta-bridge
- E extended strand, participates in beta ladder

- ▶ loop

- T turn (hydrogen bonded)
- S bend (curvature only)

- ▶ coil

- C coil

Přirazení sekundární struktury rodině proteinů z PDB

```
HQKVILVGD GAVGSSYAFAMVLQGI AQEIGIVDI
GARVVVIGA GFVGASYVFALMNQGI ADEIVLIDA
RCKITVVGV GDVGMACAISILLKGL ADELALVDA
YNKITVVGV GAVGMACAISILMKDL ADEVALVDV
DNKITVVGV GQVGMACAISILGKSL TDELALVDV
PIRVLVTGAAGQIAYSLLYSIGNGSVFGKDQPIILVLLDI
```

multiple alignment

```
CCCBBBCCC CHHHHHHHHHHHHHHCC CCCBBBCCC
CCBBBBBCC CHHHHHHHHHHHHCCCC CCBBBBBCC
CCBBBBBCC CHHHHHHHHHHHHCCCC CCBBBBBCC
CCBBBBBCC CHHHHHHHHHHHHCCC CCBBBBBCC
CCBBBBBCC CHHHHHHHHHHHHCCCC CCBBBBBCC
CCCBBBCCC CHHHHHHHHHHHHHHCC CCCBBBCCC
CCBBBBBCCCC CHHHHHHHHHHHHCCCCCCCCC BBBBCC
```

DSSP assignment

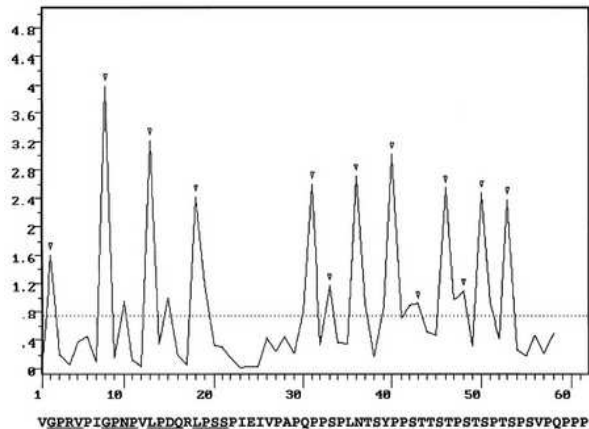
```
CCCBBBCCCC CHHHHHHHHHHHHCCCCCCCCCCC BBBBCCC
```

minimum consensus

```
CCBBBBBCCCC CHHHHHHHHHHHHHHCCCCCCCCC BBBBCC
```

maximum consensus

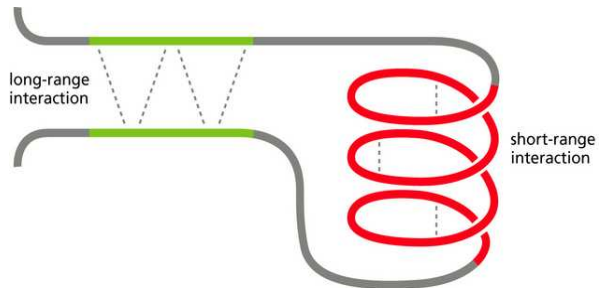
Použití metody Chou-Fasman, 1978



Metoda založena na zastoupení aminokyselin v jednotlivých typech sekundární struktury

/mnt/lexa/Pictures/logo

Blízke a vzdálené interakce



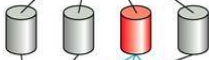
Predikce sekundární struktury neuronovými sítěmi

N-terminal...THIS IS A HIDDEN MESSAGE...C-terminal

input layer



hidden layer



output layer



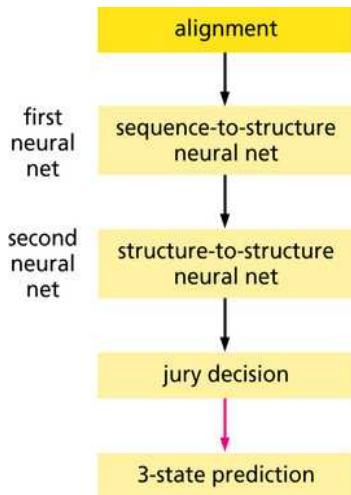
firing result

0.7 0.1 0.2

prediction result

α

Pokročilá predikce sekundární struktury



Predikce závisí od existenci homologů

- homologní** Je k dispozici struktura s podobností $> 20 - 30\%$ identity
- threading** Protein je členem rodiny se známými strukturami
- fragmentová** Protein nese lokální strukturní podobnosti k mnoha proteinem se známou strukturou
- ab initio** Realistické pro krátké sekvence

Princip modelování podle homologů

(A)



(B)

```
HEWL:  -KVFGRL ELAAAMKRHGLDNYRGYSLGNVVA AKFESNFNTQATNRRNTD-GSTDYGILOINSRWV NDGRTF  
LactB:  AEQLTKL EVFRELK- DLKGYGVSLPEVW TTFHTSGYDTQAIIVQND-STEYGLFOINNKIK KDDQNE
```

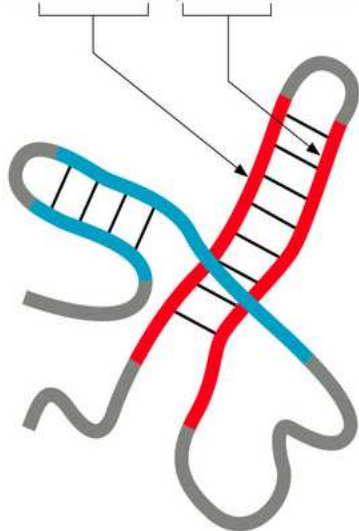


```
HEWL:  GSRNL NIP SALLSSDITASVN AKKIVSDGNGMNAVAVWRNR KGTDVQANIRG R  
LactB:  HSSNI NIS DKFLDDDLTDDIM VKKIL-DKVGINYLAHKAL SE-KLDQNL-- E
```

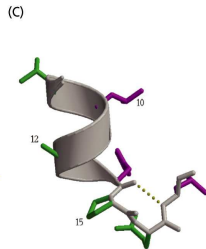
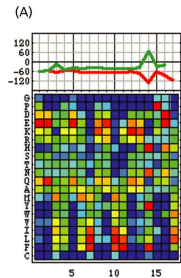
/mnt/lexa/Pictures/logo

Princip "threadingu"

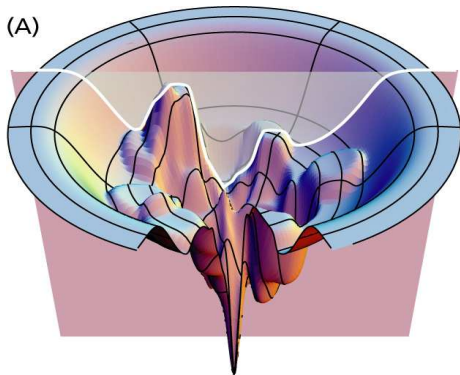
MYTARGETSEQINTHREADING



Určité posloupnosti aminokyselin mají vždy stejnou strukturu

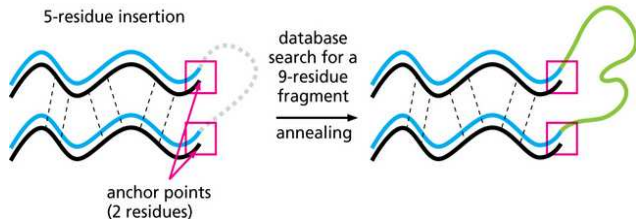


Ab initio modelování - hledání globálního minima

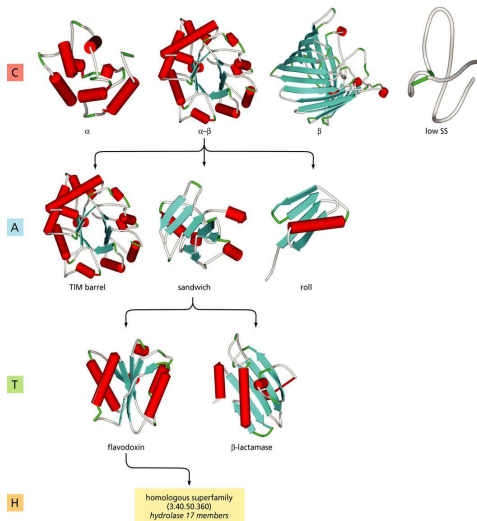


Modelování smyček

Target: VLVATY HDFVLI ...
Template: VLIISYFGNSGREFVIL ...



CATH - Class, Architecture, Topology, Homology



AlphaFold

<https://en.wikipedia.org/wiki/AlphaFold>

/mnt/lexa/Pictures/logo

Outline

Příloha

/mnt/lexa/Pictures/logo

For Further Reading

X

/mnt/lexa/Pictures/logo