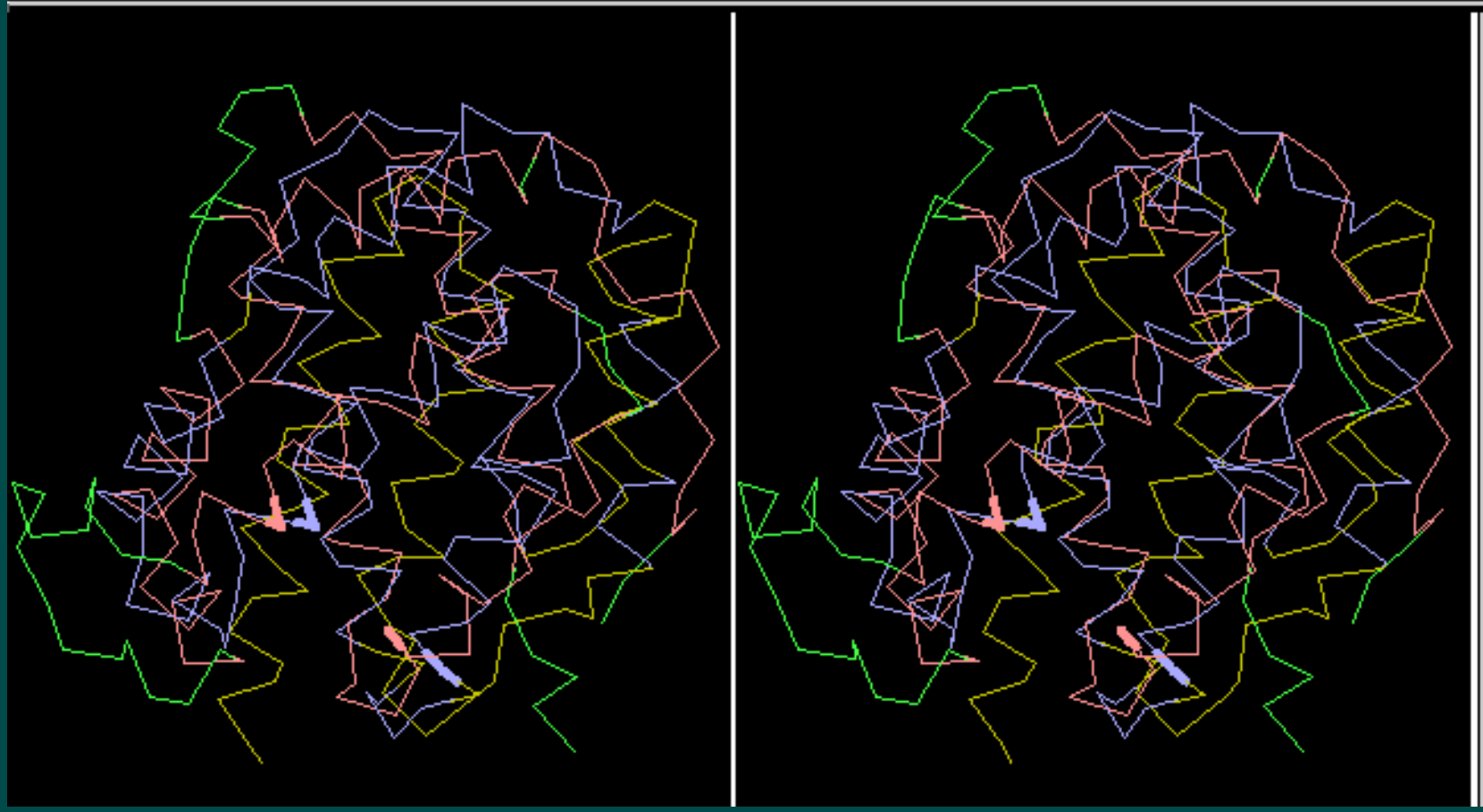
Proč srovnávat struktury?

```
COMPND    COLICIN *A (C-TERMINAL DOMAIN) (PORE-FORMING DOMAIN)
AUTHOR    M.W.PARKER,J.P.M.POSTMA,F.PATTUS,A.D.TUCKER,D.TSERNOGLOU
  NR. STRID1 STRID2  Z   RMSD LALI LSEQ2 %IDE REVERS PERMUT NFRAG TOPO PROTEIN
!   4: 1col-A 1hlb    6.2  3.1  119   157    9      0      0    10 S    HEMOGLOBIN (SEA CUCUMBER)
#
```



```
138 H E == A H   101
139 H S == R H   102
140 H W == T H   103
141 H V == H H   104
142 H L == .
143 T S == D H   105
144 T G == L H   106
145   I == N T   107
146   A == K T   108
147 H S == .
148 H S == .
149 H V == V     109
150 H A == G     110
151 H L == A H   111
152 H G == D H   112
153 H I == H H   113
154 H F == Y H   114
155 H S == N H   115
156 H A == L H   116
157 H T == F H   117
158 H L == A H   118
159 H G == K H   119
```

Proč srovnávat struktury?

seskupování proteinů do skupin s podobnou strukturou

určit významnost individuálních pozic ve struktuře

najít podobnosti, které nejsou viditelné na úrovni sekvence

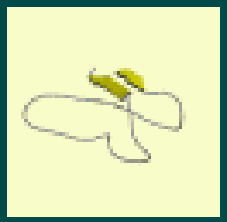
Proč srovnávat struktury?

seskupování proteinů do skupin s podobnou strukturou (CATH, SCOP)

Třída 1 ALFA

Třída 2 BETA

Třída 3 ALFA/BETA



Třída 4 BEZ SEKUNDÁRNÍ STRUKTURY

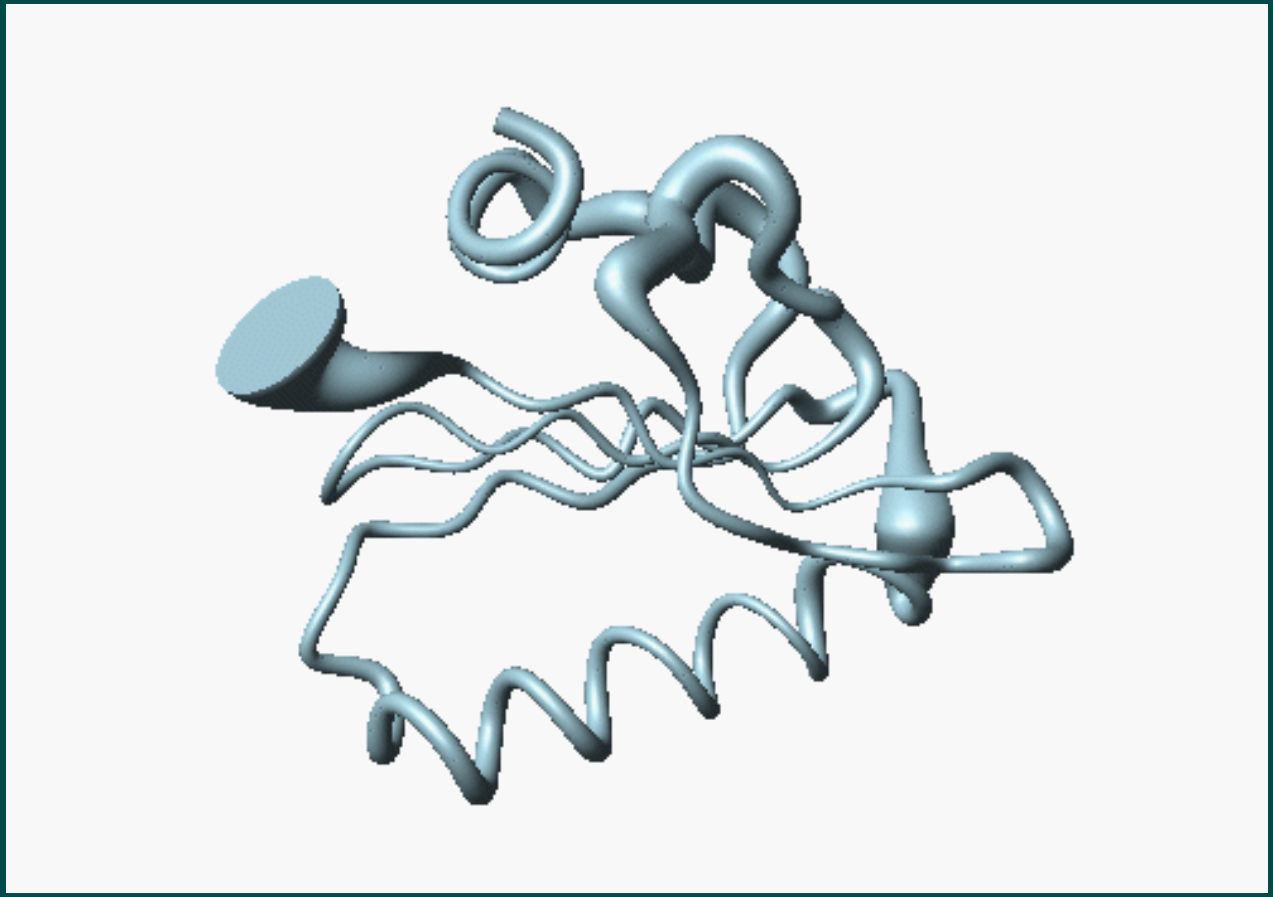
Počty v různých strukturních třídách proteinů evidované v databázi SCOP

SCOP: Structural Classification of Proteins. **1.69** release
25973 PDB Entries (1 Oct 2004). 70859 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	218	376	608
All beta proteins	144	290	560
Alpha and beta proteins (a/b)	136	222	629
Alpha and beta proteins (a+b)	279	409	717
Multi-domain proteins	46	46	61
Membrane and cell surface proteins	47	88	99
Small proteins	75	108	171
Total	945	1539	2845

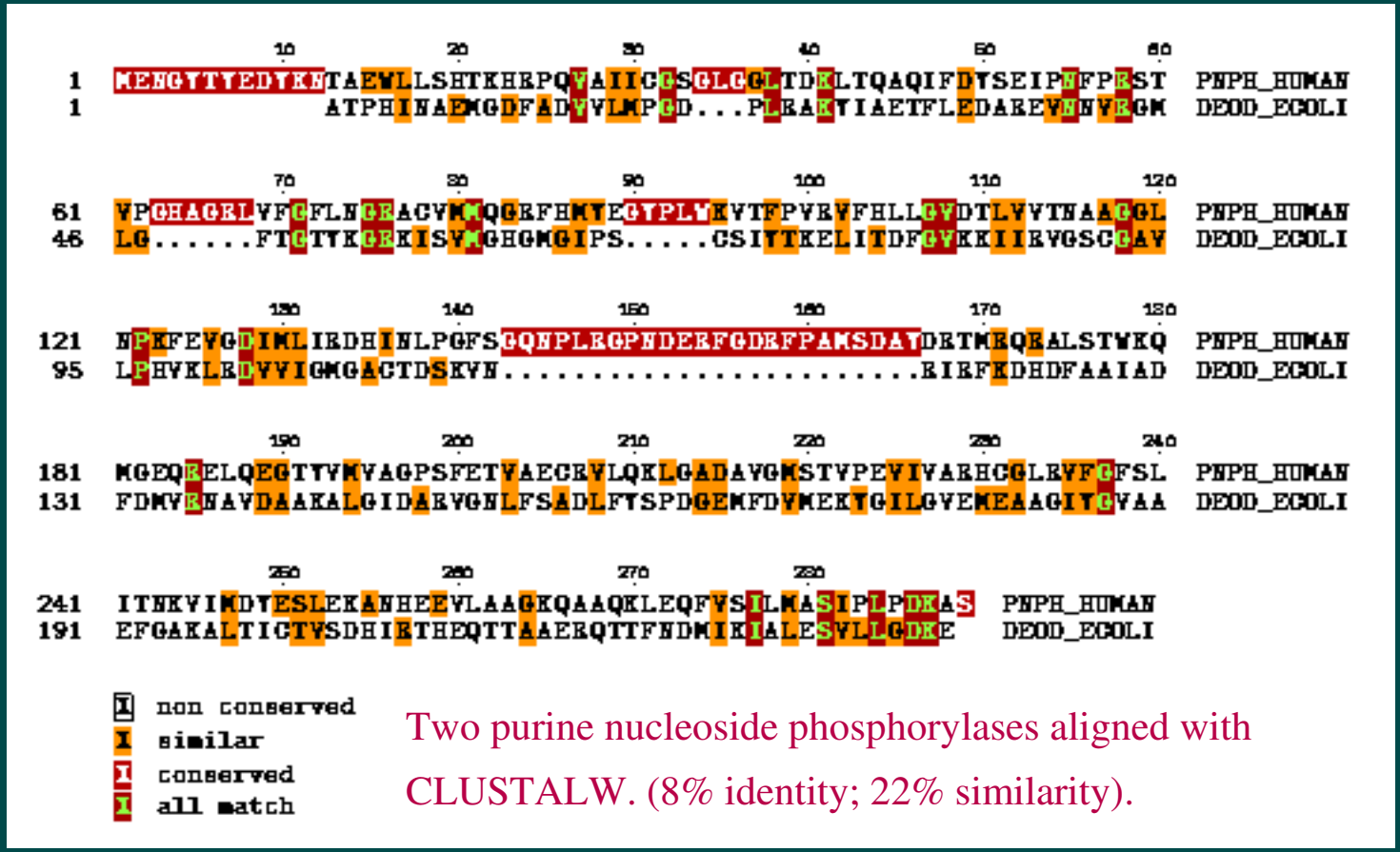
Proč srovnávat struktury?

určit významnost individuálních pozic ve struktuře



Proč srovnávat struktury?

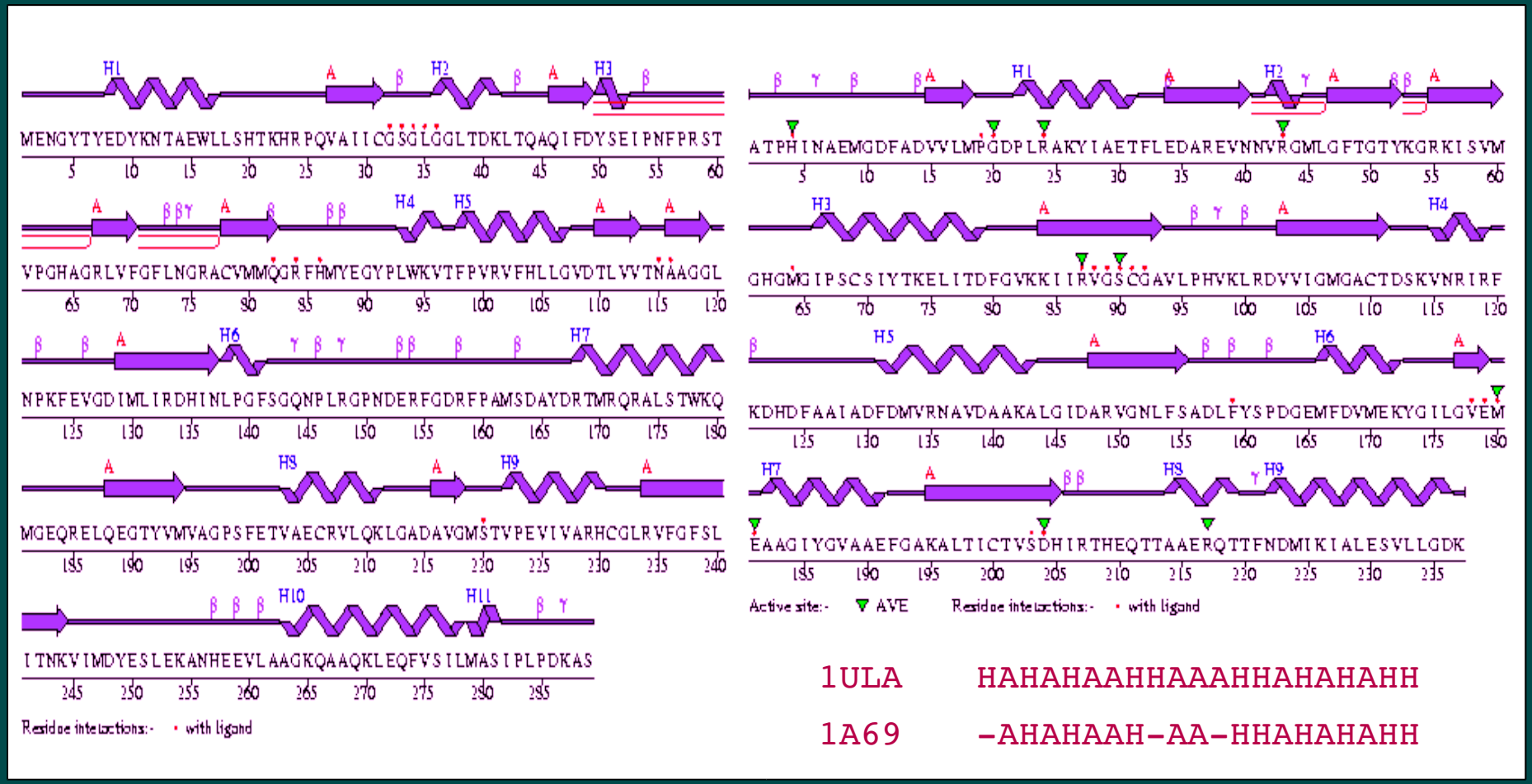
najít podobnosti, které nejsou viditelné na úrovni sekvence



Two purine nucleoside phosphorylases aligned with CLUSTALW. (8% identity; 22% similarity).

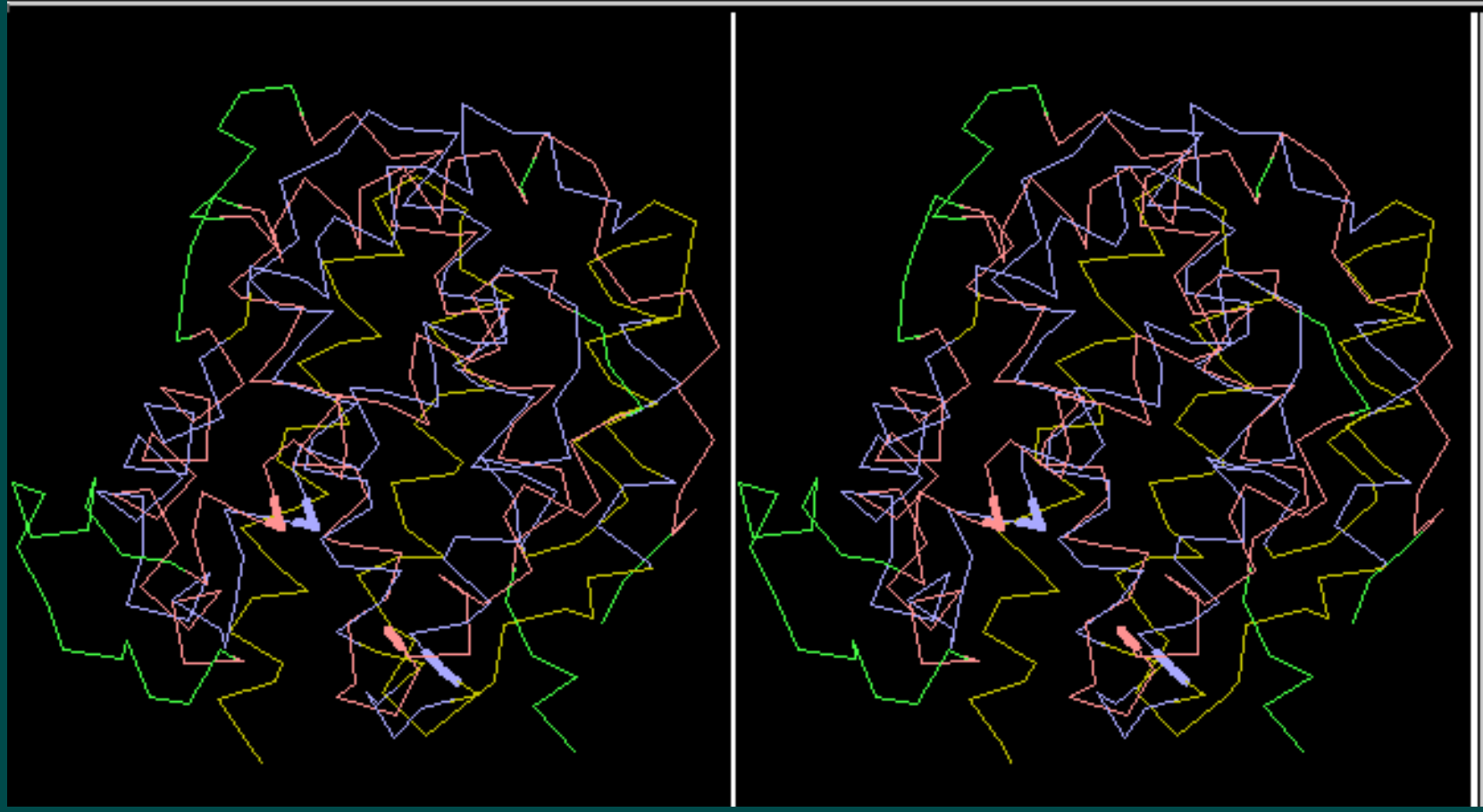
Proč srovnávat struktury?

najít podobnosti, které nejsou viditelné na úrovni sekvence



Proč srovnávat struktury?

```
COMPND    COLICIN *A (C-TERMINAL DOMAIN) (PORE-FORMING DOMAIN)
AUTHOR    M.W.PARKER,J.P.M.POSTMA,F.PATTUS,A.D.TUCKER,D.TSERNOGLOU
  NR. STRID1 STRID2  Z   RMSD LALI LSEQ2 %IDE REVERS PERMUT NFRAG TOPO PROTEIN
!   4: 1col-A 1hlb    6.2  3.1  119   157    9      0      0    10 S    HEMOGLOBIN (SEA CUCUMBER)
#
```



```
138 H E == A H   101
139 H S == R H   102
140 H W == T H   103
141 H V == H H   104
142 H L == .
143 T S == D H   105
144 T G == L H   106
145   I == N T   107
146   A == K T   108
147 H S == .
148 H S == .
149 H V == V     109
150 H A == G     110
151 H L == A H   111
152 H G == D H   112
153 H I == H H   113
154 H F == Y H   114
155 H S == N H   115
156 H A == L H   116
157 H T == F H   117
158 H L == A H   118
159 H G == K H   119
```

RMSD – Root Mean Square Distance

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{n}}$$

$$RMSD(\mathbf{y},\mathbf{x}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{x}_i)\bullet(\mathbf{y}_i - \mathbf{x}_i)}$$

RMSD je určitou obdobou standartní odchylky

Identické struktury mají RMSD=0
Podobné struktury mají RMSD v rozmezí 0-5 angstromů

Nejnižší RMSD není vždy nejlepší z jiných hledisek, např.
pokud si nějaká část struktury vzájemně neodpovídá, bylo by lepší její vliv vůbec nebrat v úvahu.
RMSD je citlivé na velikost porovnávaných proteinů

Procento zarovnaných pozic

Kromě RMSD je užitečné vědět, jaké procento délky polypeptidického řetězce se nám povedlo navzájem přiřadit. To můžeme vypočítat jako

$$N / \min(L(A), L(B))$$

kde L(X) je délka proteinu X a N je počet přiřaděných aminokyselin. Aminokyseliny se považují za přiřazené, pokud jejich vzdálenost v překryvu struktur je menší než určena hranice (obyčejně několik angstromů). Hodnoty se pohybují od 0 do 100%.

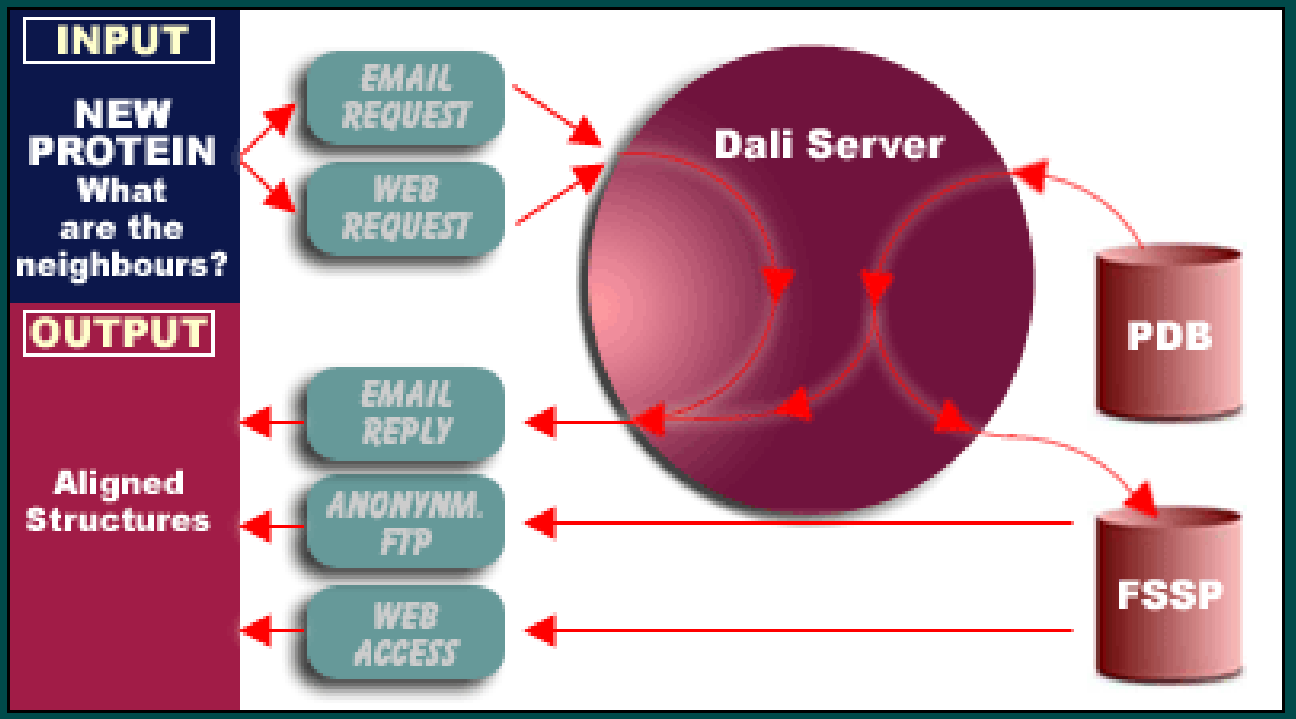
Nástroje pro porovnávání struktur

Výpočet transformace potřebné k optimálnímu překrytí dvou struktur patří do třídy NP-hard algoritmů. Zložitost tkví kromě jiného v tom, že nevíme předem říct, jak namapovat atomy jedné struktury na strukturu druhou. Prakticky použitelné algoritmy v této oblasti používají heuristický přístup, z čeho vyplývá, že mají své silné i slabé stránky, pokud jde o schopnost najít nejlepší řešení.

Geometric hashing, distance matrix alignment, graph theory (shortest path, largest common subgraph), knot theory

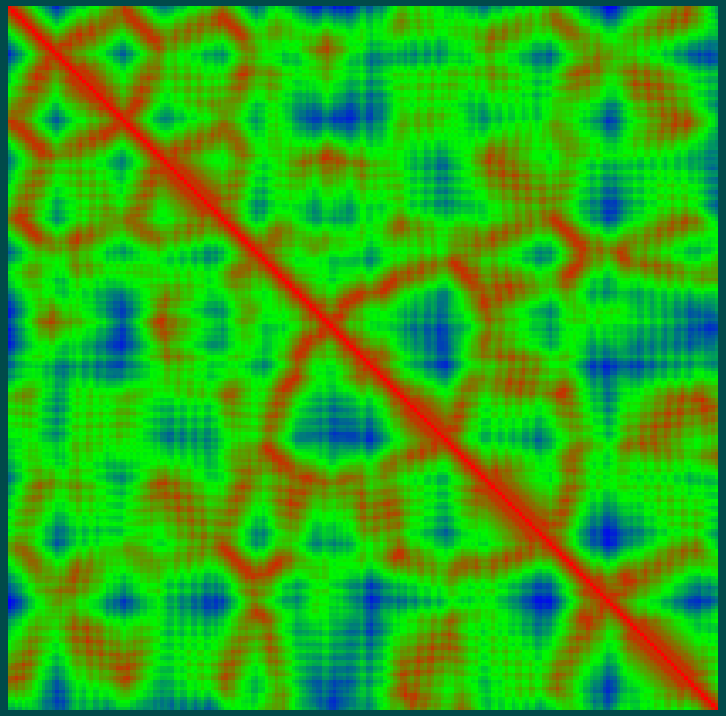
C-ALFA	http://bioinfo3d.cs.tau.ac.il/c_alpha_match/
DALI	http://www.ebi.ac.uk/dali/
VAST	http://www.ncbi.nih.gov/Structure/VAST/
FLEXPROT	http://bioinfo3d.cs.tau.ac.il/FlexProt/

DALI - Distance Matrix Alignment

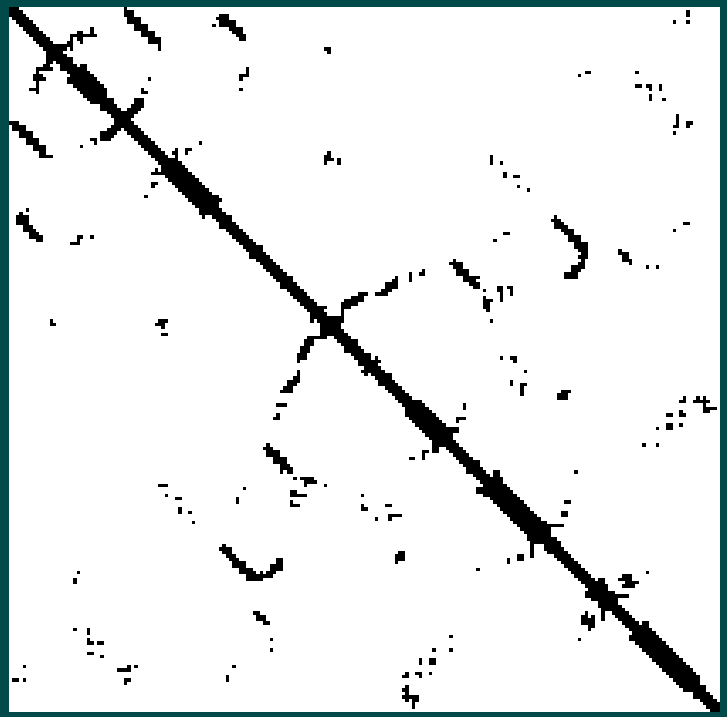


1AUG

DISTANCE MATRIX

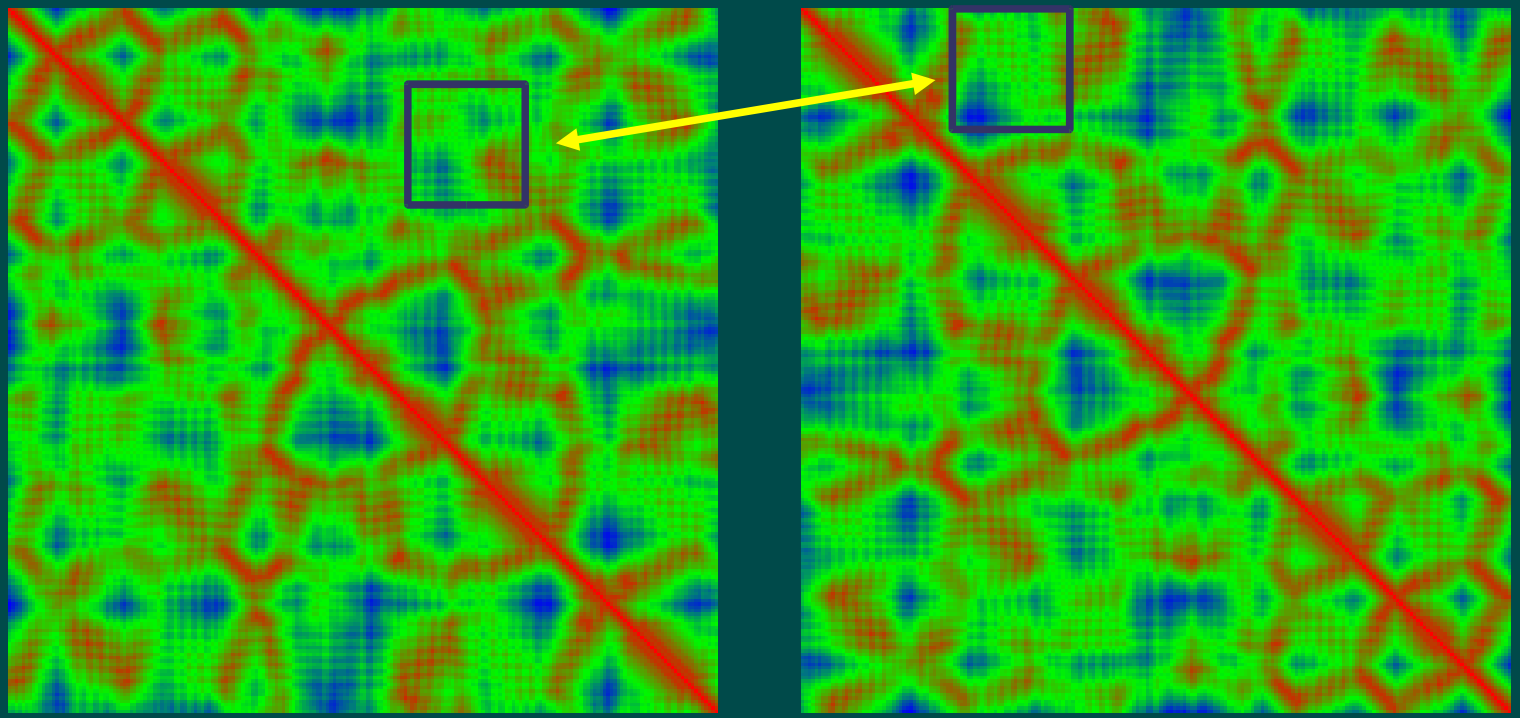


CONTACT MAP



DALI - Distance Matrix Alignment

Vytvoření seznamu elementárních podobností



DALI - Distance Matrix Alignment

$$S=\sum_{i=1}^{L}\sum_{j=1}^{L}\phi(i,j),$$

Skóre porovnávání dvou matic

$$\phi^{R}(i,j)=\theta^{R}-\left|d_{ij}^{A}-d_{ij}^{B}\right|$$

Skóre porovnávání dvou pozic

$$\phi^{E}(i,j)=\begin{cases}(\theta^{E}-\dfrac{\left|d_{ij}^{A}-d_{ij}^{B}\right|}{d_{ij}^{*}})w(d_{ij}^{*}),\ i\neq j\\ \theta^{E},\ i=j\end{cases}$$

$$w(r)=e^{-\frac{r^{2}}{\alpha^{2}}}$$

“Elastické” skóre porovnávání dvou pozic

DALI - Distance Matrix Alignment

Z různých míst seznamu elementárních podobností se program paralelně snaží budovat větší oblasti zarovnání, tak jak je popsáno níže. Výsledkem je pak několik nejlepších globálních či lokálních překrytí dvou molekul.

OPTIMIZACE MONTE CARLO

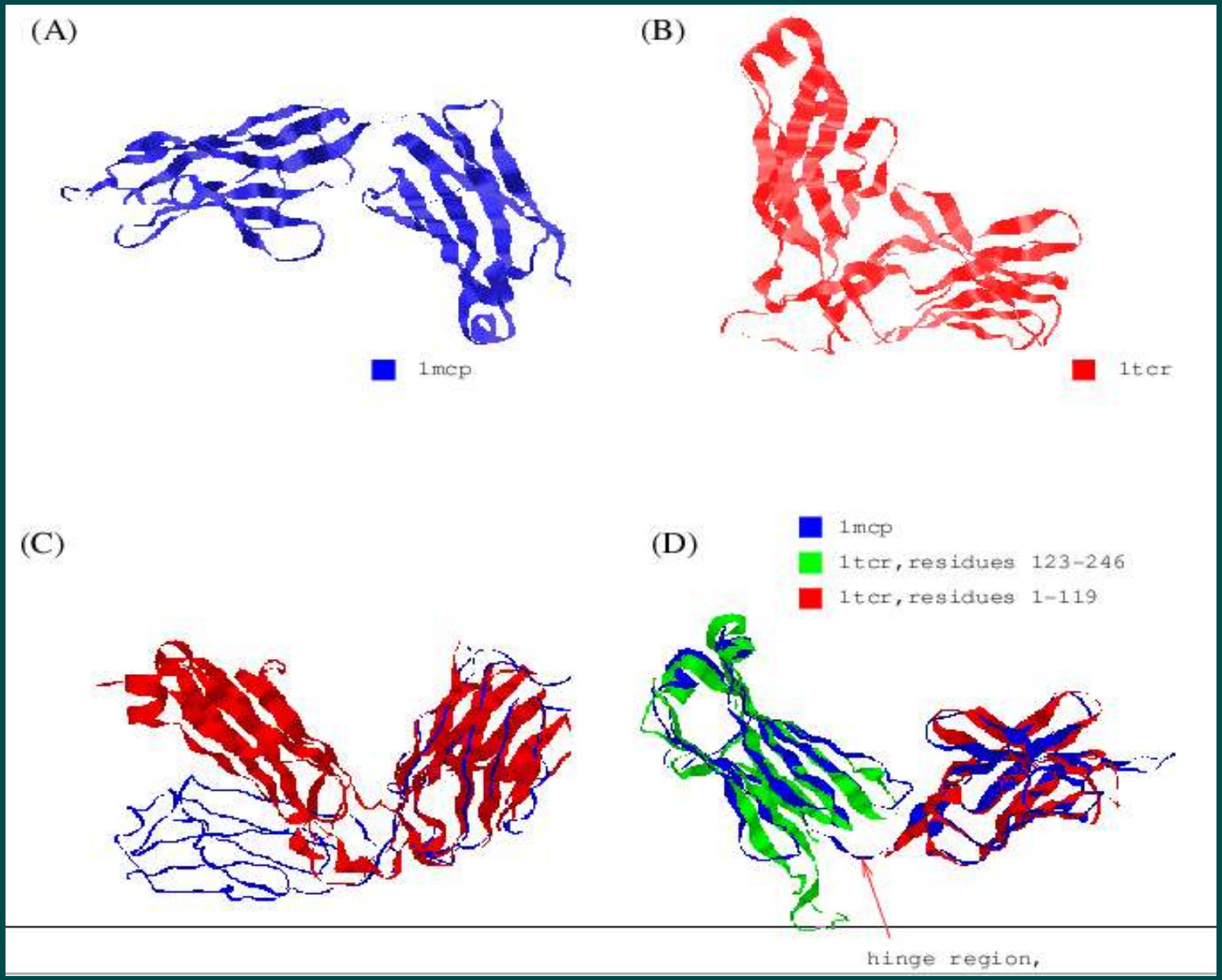
1) EXPANZE

Nabírání nových elementárních podobností se seznamu

2) ZKRACOVÁNÍ

Odstraňování úseků, které po expanzi zhoršují celkové skóre

FlexProt – Flexible Structural Alignment



Rychlé vyhledávání strukturálních podobností

Převedení souřadnic atomů tvořících proteinovou páteř na posloupnost torzních úhlů psi, phi a omega umožní rychle vyhledat oblasti možné strukturální podobnosti a dopočítat RMSD či %překryvu tradičními metodami na malém objemu dat.

```
>1AAF_A 09-APR-92 ELECTRON_TRANSPORT
QRGNFRNQRKIIKCFNCGKEGHIAKNCRAPRKRGCWKCGKEGHQMKDCTERQA
h@rn@D\LjpRR@TXNNlTJv@JNXPRRTTJR@VJNXNvTFN@@XZNTFj@Fh
```