Biologická sekvence (BS)

ACAGTGCGAGCATGACGATGACGCAGCAGATTGACAGAGACGATAGCAGCAT

MASAQSFYLLHLAVDDFMNGAGVLSHERELLFYDENKIHDIVISMNDENMNQ

Jazyk

THISISJUSTASIMPLESENTENCEINENGLISHFORYOURINSPIRATION

Frekvence

F(x) = P(x) N
Je vyšší u řetězců, které jsou součástí často používaných struktur

Vzájemná informace

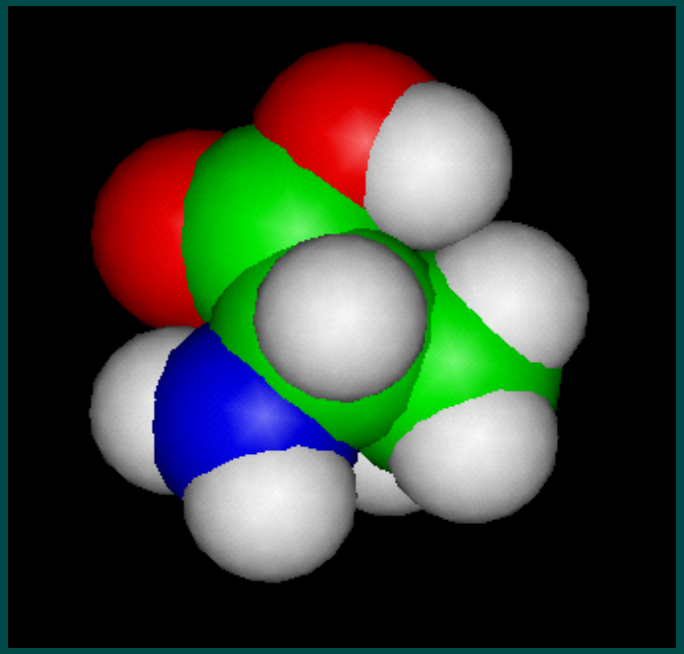
*MI(x,y) = P(x,y) log (P(x,y) / (P(x)*P(y)))*
Je vyšší uvnitř struktur než na jejich rozhraní, vyjádřuje korelaci

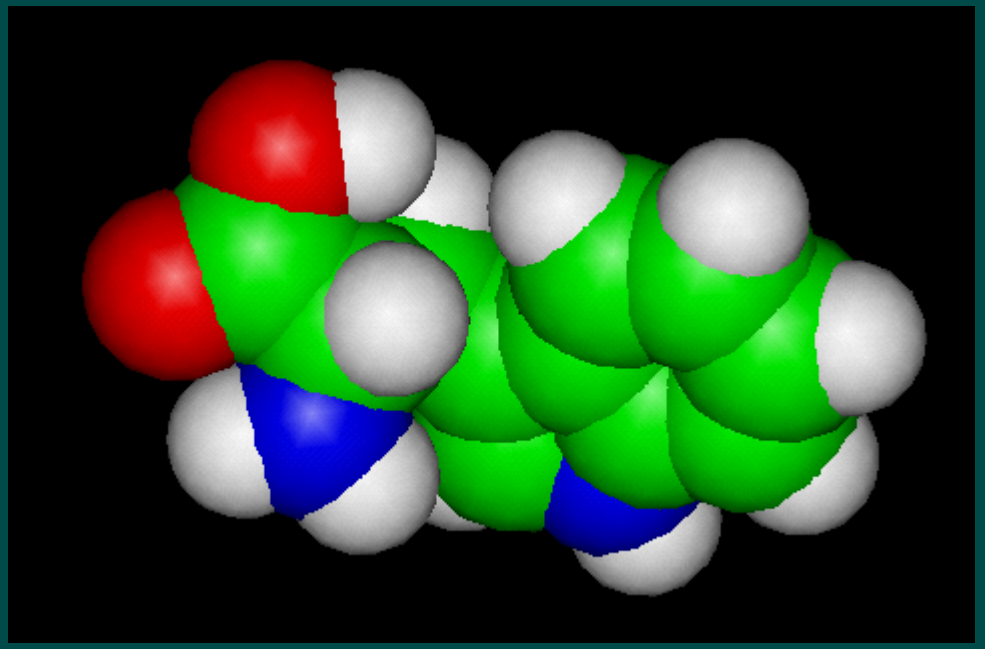
Entropie

*H(x) = - p(x)*log(p(x))*
Určuje míru neuspořádanosti, nebo taky potřebu informace pro definování určitého stavu

Strukturní interpretace frekvence

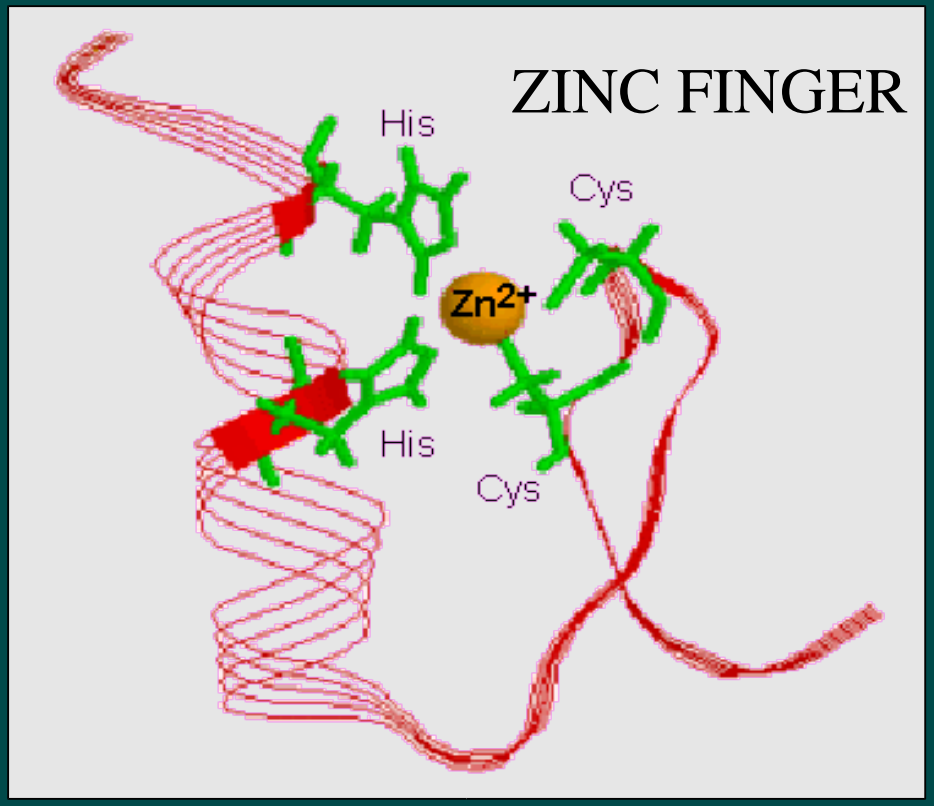
fyzikálně-chemické a prostorové vlastnosti





Strukturní interpretace vzájemné informace

vazební místo



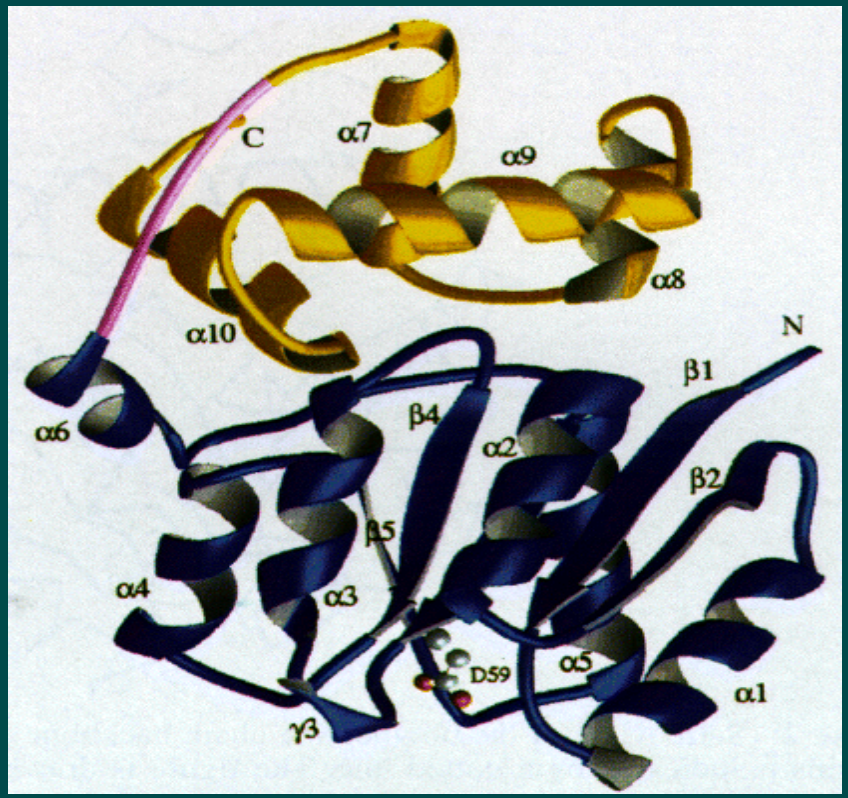
Další možnosti strukturní interpretace statistických veličin a výpočtů

doména

vysoká f vysoká f

SHLQFNMIDIVISK

nízká f



Why bother?

BUCKDIDNOTREADTHENEWSPAPERSORHEWOULDHAVEKNOWNTHATTROUBLEWASBREWING

BUCKDIDNOTREADTHENEWSPAPERSORHEWOULDHAVEKNOWNTHATTROUBLEWASBREWING

BUCKDIDNOTREADTHENEWSPAPERSORHEWOULDHAVEKNOWNTHATTROUBLEWASBREWING

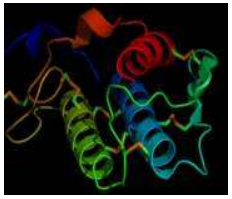
Mental image:

MASAQSFYLLHLAVDDFMNGAGVLSHERELLFYIMASKRDLDENCVIGARAKIHDIVISMNDENMN

MASAQSFYLLHLAVDDFMNGAGVLSHERELLFYIMASKRDLDENCVIGARAKIHDIVISMNDENMN

MASAQSFYLLHLAVDDFMNGAGVLSHERELLFYIMASKRDLDENCVIGARAKIHDIVISMNDENMN

Protein:

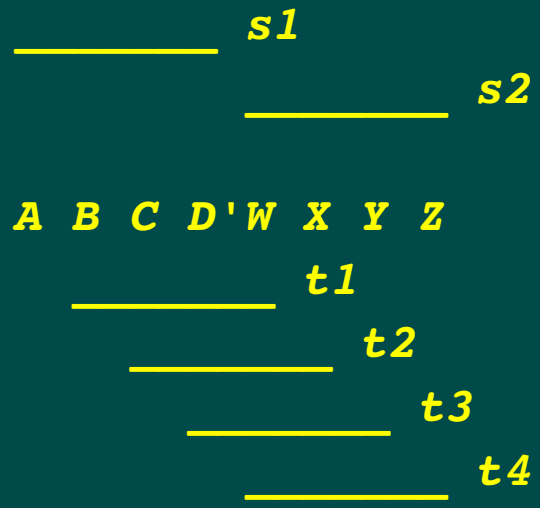


Kubota, Lee, 1999. Mostly-unsupervised statistical segmentation of Japanese: application to kanji.

kanji, hiragana, katakana – znaky různé úrovně

kanji jsou na úrovni našich slabik a tvoří polovinu slov

sekvence kanji se často dají segmentovat různými způsoby



漢英字典刊行会

Kubota, Lee, 1999. Mostly-unsupervised statistical segmentation of Japanese: application to kanji.

Pro každou mezeru se vypočítá hodnota (s1+s2)/(t1+...+tn)

漢英字典刊行会

Vstupní data pro analýzu textu ve formátě FASTA

```
>SENTENCE
THECALLOFTHEWILD

>SENTENCE
BYJACKLONDON

>SENTENCE
CHAPTERONE

>SENTENCE
BUCKDIDNOTREADTHENEWSPAPERSORHEWOULDHAVEKNOWNTHATTROUBLEWASBREWING
```

Vyhodnocení frekvence 4-gramů v textu

```
__<BUC KDIDNO 0 0 68 396 22 20 14 691 594 0 0.28
_<BUCK DIDNOT 0 68 396 22 20 14 691 594 1407 1 29.116
<BUCKD IDNOTR 68 396 22 20 14 691 594 1407 101 0 1.274
BUCKDI DNOTRE 396 22 20 14 691 594 1407 101 86 0 1.647
UCKDID NOTREA 22 20 14 691 594 1407 101 86 282 1 0.064
CKDIDN OTREAD 20 14 691 594 1407 101 86 282 799 0 0.554
KDIDNO TREADT 14 691 594 1407 101 86 282 799 149 0 0.824
DIDNOT READTH 691 594 1407 101 86 282 799 149 270 1 7.055
IDNOTR EADTHE 594 1407 101 86 282 799 149 270 5248 0 0.321
DNOTRE ADTHEN 1407 101 86 282 799 149 270 5248 1830 0 0.434
NOTREA DTHENE 101 86 282 799 149 270 5248 1830 471 0 6.81
OTREAD THENEW 86 282 799 149 270 5248 1830 471 145 1 0.695
TREADT HENEWS 282 799 149 270 5248 1830 471 145 139 0 0.126
READTH ENEWSP 799 149 270 5248 1830 471 145 139 74 0 0.082
EADTHE NEWSPA 149 270 5248 1830 471 145 139 74 74 1 3.303
ADTHEN EWSPAP 270 5248 1830 471 145 139 74 74 82 0 3.782
DTHENE WSPAPE 5248 1830 471 145 139 74 74 82 169 0 2.283
THENEW SPAPER 1830 471 145 139 74 74 82 169 232 0 1.186
HENEWS PAPERS 471 145 139 74 74 82 169 232 364 0 2.008
ENEWSP APERSO 145 139 74 74 82 169 232 364 442 0 1.412
NEWSPA PERSOR 139 74 74 82 169 232 364 442 33 0 1.36
EWSPAP ERSORH 74 74 82 169 232 364 442 33 13 0 1.027
WSPAPE RSORHE 74 82 169 232 364 442 33 13 183 0 0.291
SPAPER SORHEW 82 169 232 364 442 33 13 183 65 0 0.438
PAPERS ORHEWO 169 232 364 442 33 13 183 65 1221 1 1.681
APERSO RHEWOU 232 364 442 33 13 183 65 1221 477 0 3.32
PERSOR HEWOUL 364 442 33 13 183 65 1221 477 1184 1 7.206
ERSORH EWOULD 442 33 13 183 65 1221 477 1184 2985 0 0.5
```

Segmentace textu v angličtině

~50%

2-gram

<	**7.046**
THE	**3.965**
CALL	**1.771**
OF	**5.683**
THE	**3.37**
WILD	**0.843**
<	**0.628**
BY	**15.17**
JACK	**10.951**
LOND	**3.267**
ON	**4.759**
<	**8.495**
CHAP	**5.136**
TER	**2.424**
ONE	**1.69**
<	**4.565**
INTO	**3.996**
THE	**6.199**
PRIM	**2.914**
ITI	**4.348**
VE	**1.674**

~20%

4-gram

<	1.024
THEC	2.834
ALL	10.841
OFTHEW	4.86
ILD	19.2
<	2.062
BY	4.632
JACK	2.758
LONDON	14.962
<	2.025
CHAPTERONE	0.922
<	10.137
IN	1.555
TOTHEP	4.058
RIMI	3.24
TIVE	6.681

~20%

2-gram

<	1.396
BU	1.717
CK	57.205
DID	3.357
NOT	3.116
READ	3.744
THE	1.733
NEW	8.714
SPAP	3.266
ER	2.745
SOR	18.096
HE	3.303
WOU	2.25
LD	2.73
HA	4.572
VE	5.867
KNOW	6.71
NTH	2.046
ATT	1.74
ROU	11.28
BLEWASB	5.806
REW	3.149
ING	22.372

~35%

4-gram

<	8.735
BUCK	29.116
DI	1.647
DNOT	7.055
REA	6.81
DTHEN	3.782
EWS	2.008
PAPERSOR	7.206
HEW	1.587
OULD	6.122
HAVE	25.589
KNOWN	7.595
THAT	8.573
TROUBL	12.29
EWAS	8.537
BREWING	3.078

Segmentace sekvence z PDB

CGVGFIANLRGKPDH	2.908
TLVE	2.489
QALKALGC	6.761
MEH	2.355
RGG	3.459
CSAD	1.952
NDSGD	1.636
GAGV	3.156
MTAIP	3.338
RELLAQ	6.626
WFNT	1.612
RNLPM	3.229
PDGDRLGVGM	2.648
VFLPQ	1.967
EPSAREVARAY	1.781
VEEVV	1.553
RLEKLTVLG	3.571
WREVPVNS	1.521
DVLGI	1.919
QAKN	1.57
NQ	1.514
PHIEQILVT	3.613
CPEG	2.37
CAGDELDRRL	1.989
YIARSIIGKKLAEDF	1.593

Obr. - Stereo pohled na identifikované segmenty

Weisser D, Klein-Seetharaman J (2004). Identification of fundamental building blocks in protein sequences using statistical association measures. ACM SAC 2004

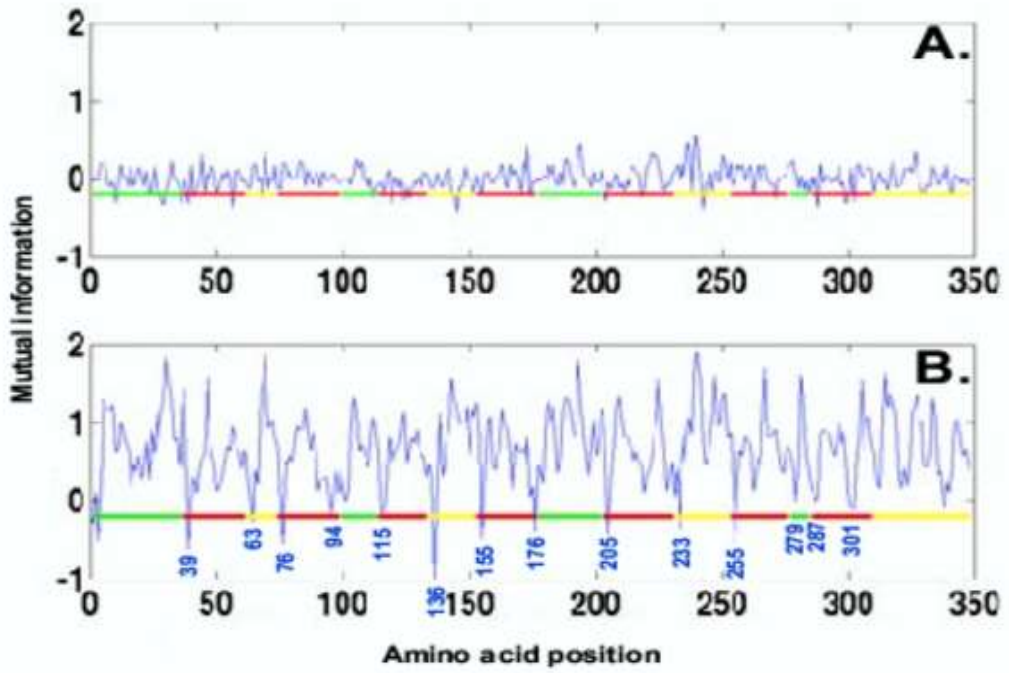


Figure 2. Mutual information values along the rhodopsin sequence using different datasets (A. human, B. GPCR) to generate mutual information values. Horizontal lines use the same color code as in Figure 1 indicating the positions of the segments belonging to each of ec, cp and helices domains based on expert knowledge. The positions of breakpoints indicated by mutual information minima are shown as blue labels in B.

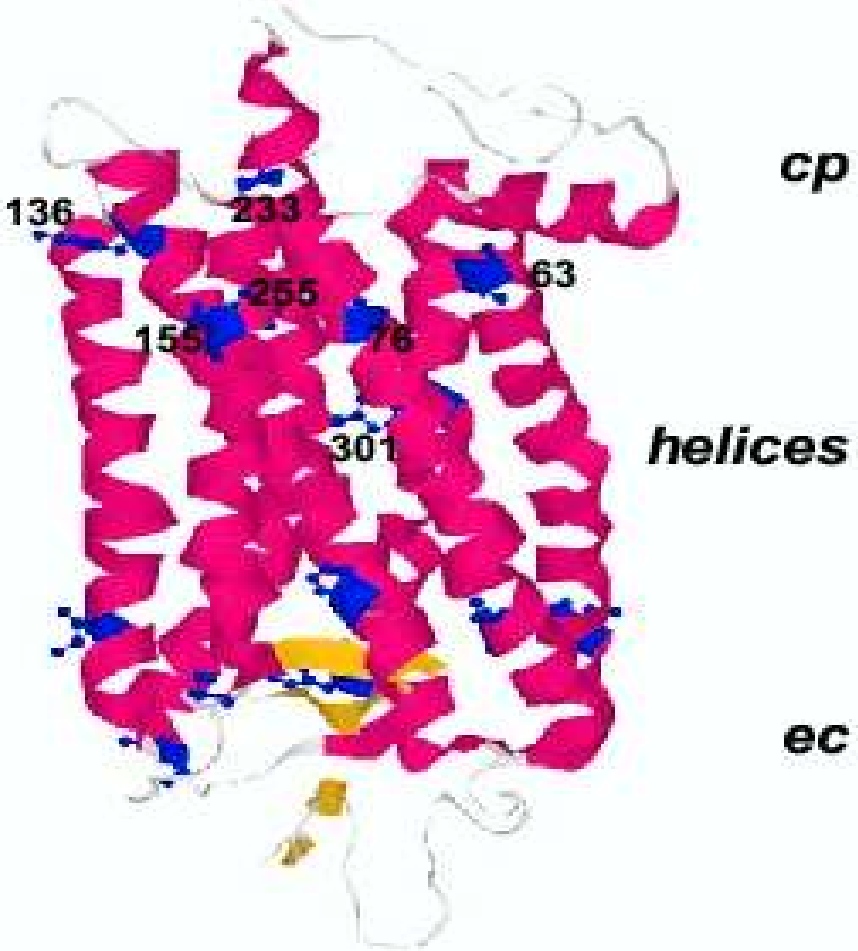
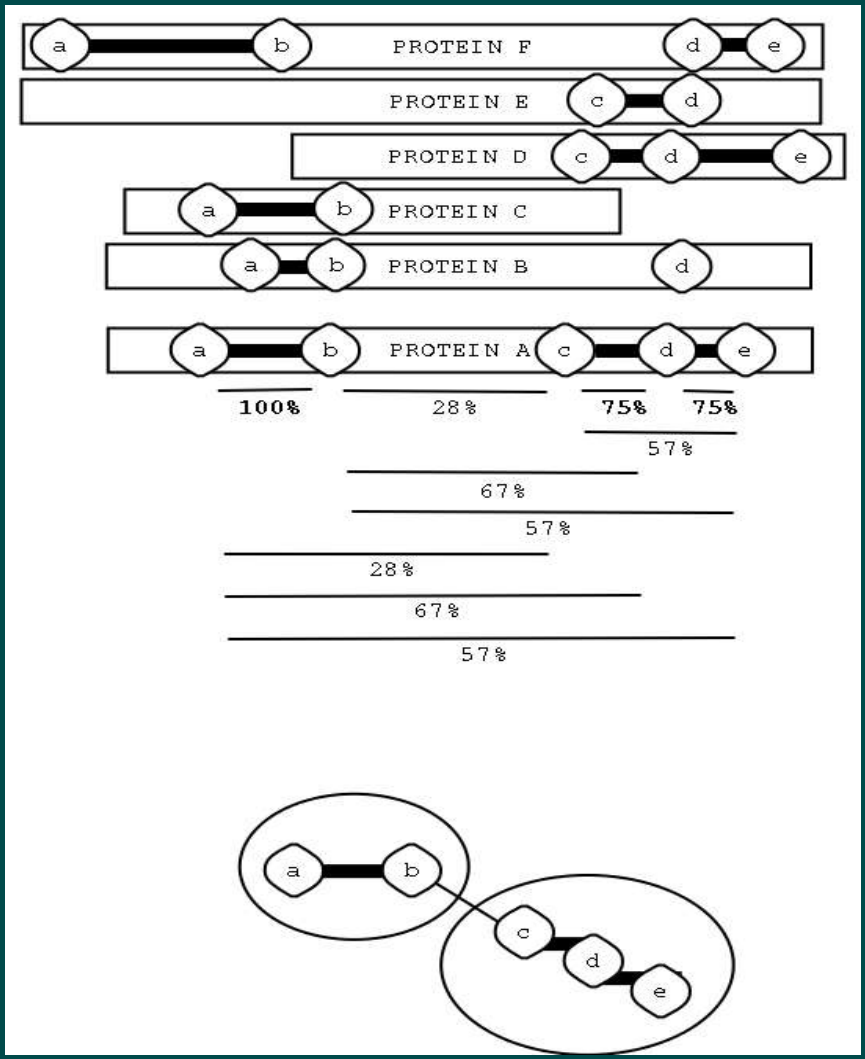


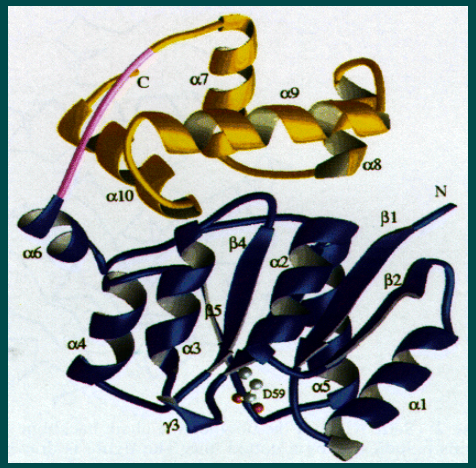
Figure 7. Tertiary structure of rhodopsin based on X-ray crystallography [3, 4]. All secondary structure elements (helices and beta-sheet) are shown in red, all loops are shown in gray. The position of breakpoints identified by mutual information (Figure 2B) is shown in blue. The corresponding amino acid residues are shown as ball-and-stick representation. The molecule is oriented as shown in the secondary structure model in Figure 1A.

Spoluvýskyt krátkých sekvencí v proteinech

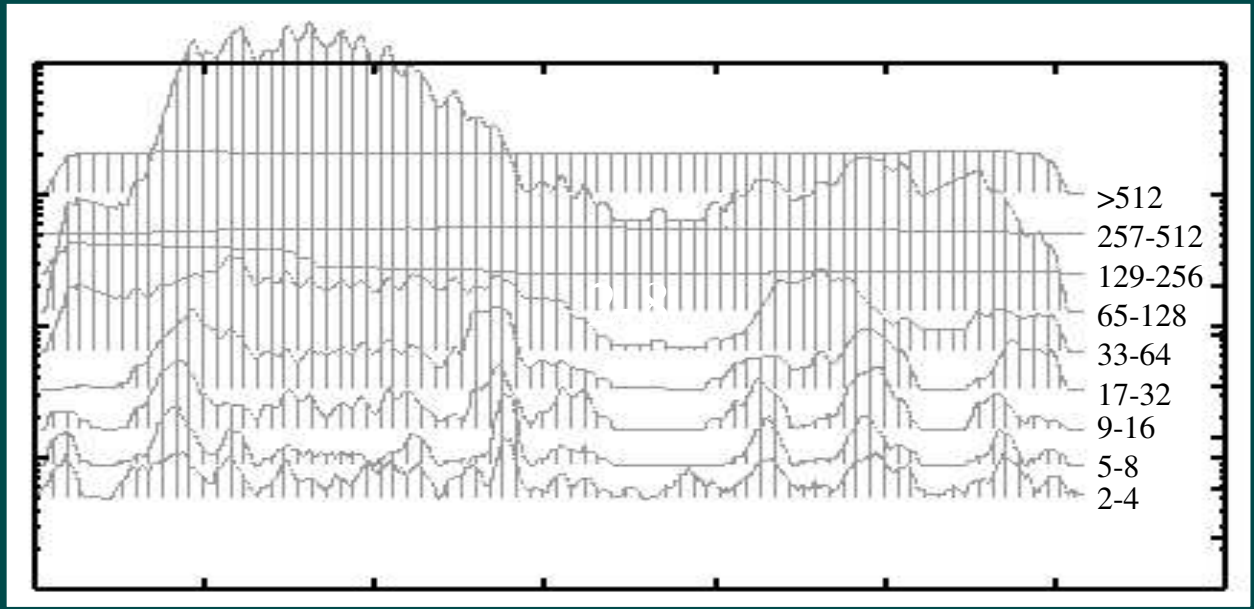
```
a = SHLQFMV
b = DHLDDRK
c = ...
```

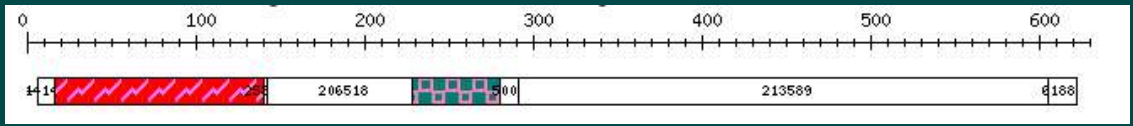


Jedním z důvodů spoluvýskytu krátkých sekvencí je, že spolu vytvářejí samostatní doménu, která se vyskytuje ve větším počtu proteinů



Vyhodnocení hledání domén





Počet korelací procházejících daným místem proteinu Atg07210 porovnaný se záznamem v databáze PRODOM

Celková struktura sekvence jak se jeví při srovnání s ostatními sekvencemi v databázích pomocí BLASTu