Predikce struktury proteinu ze sekvence

Sekvence <-> Přírodní zákony

Sekvence <-> Databáze sekvencí

Sekvence <-> Databáze struktur

Frekvence

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

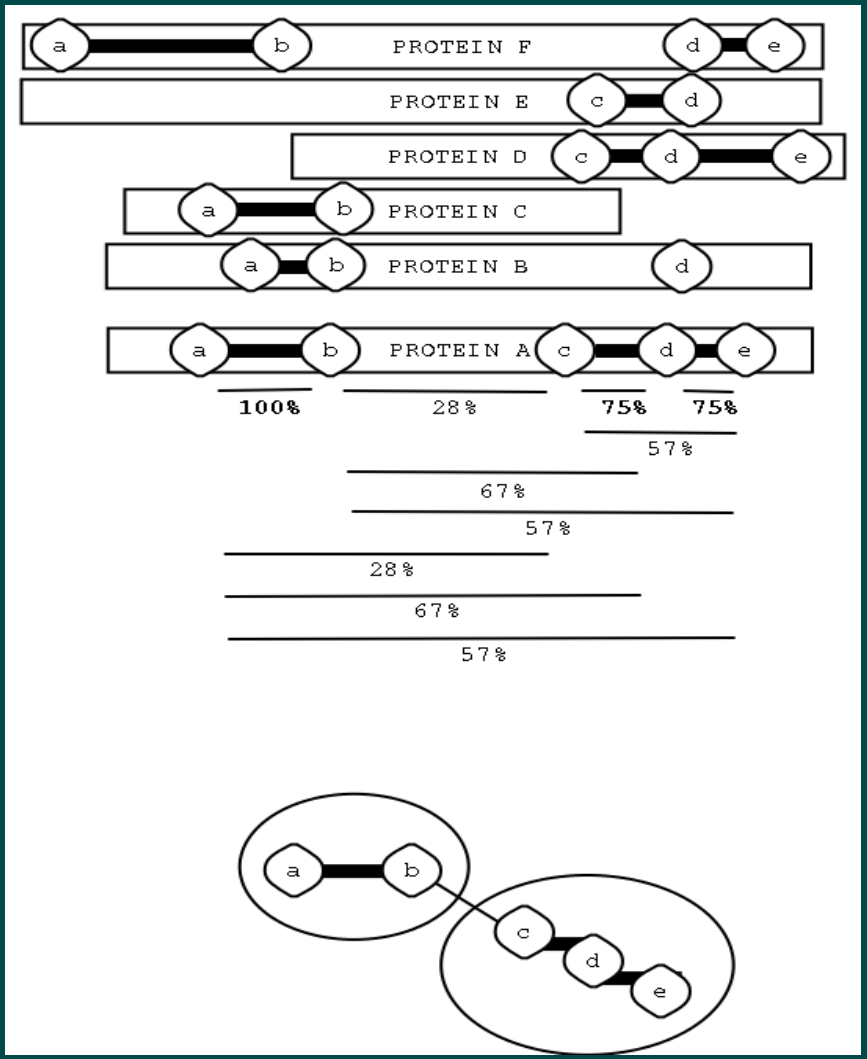
Vzájemná informace

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

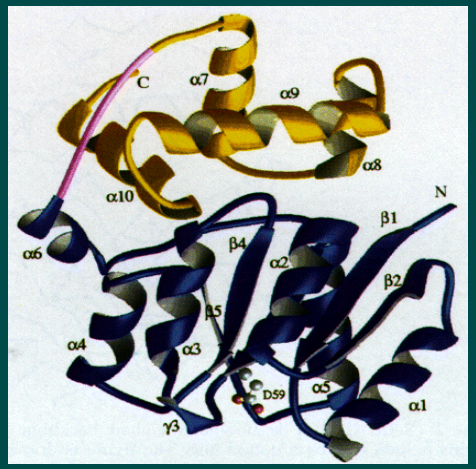
Entropie

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

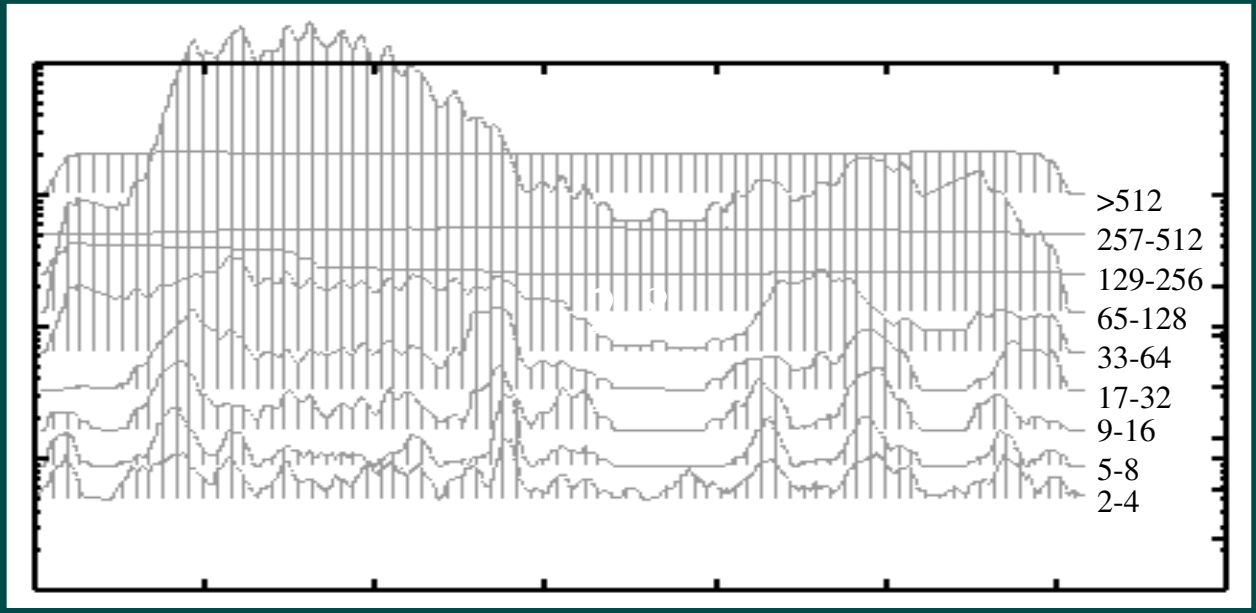
Spoluvýskyt krátkých sekvencí v proteinech

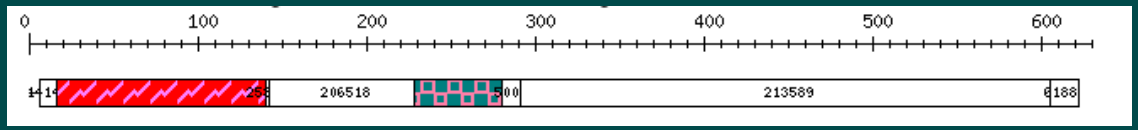


Jedním z důvodů spoluvýskytu krátkých sekvencí je, že spolu vytvářejí samostatní doménu, která se vyskytuje ve větším počtu proteínů



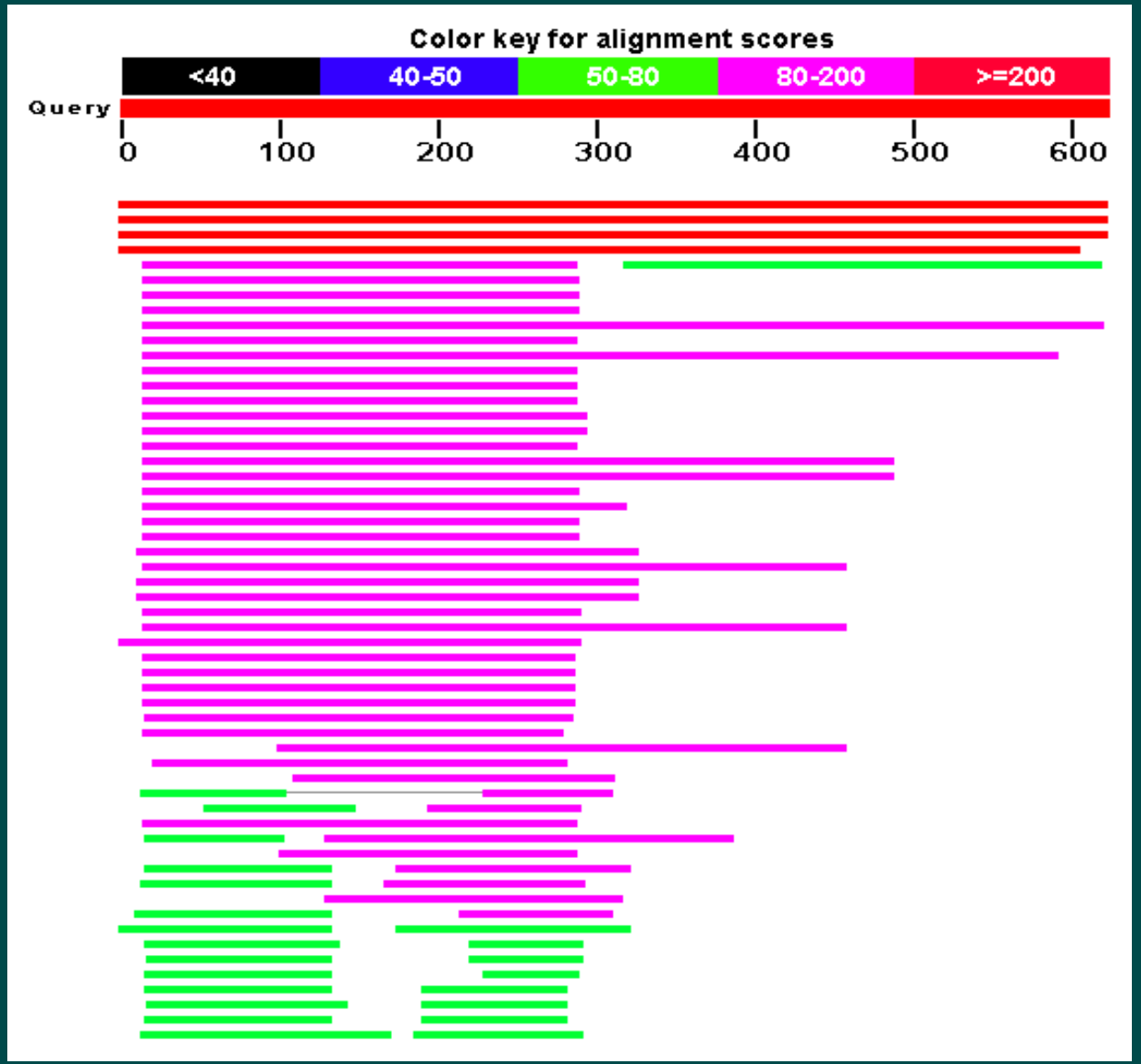
Vyhodnocení hledání domén





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 proteinových databázích

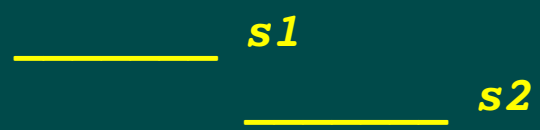


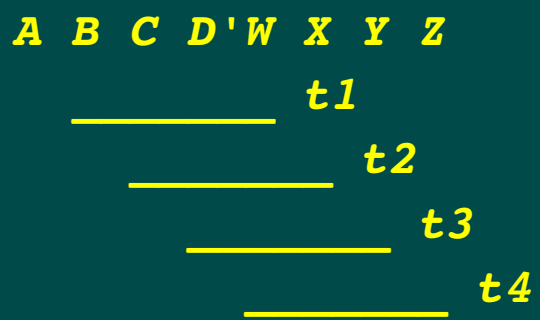
Kubota, Lee, 1999. Mostly-unsupervised statistical segmentation of Japanese: application to kanji.

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
```

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

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

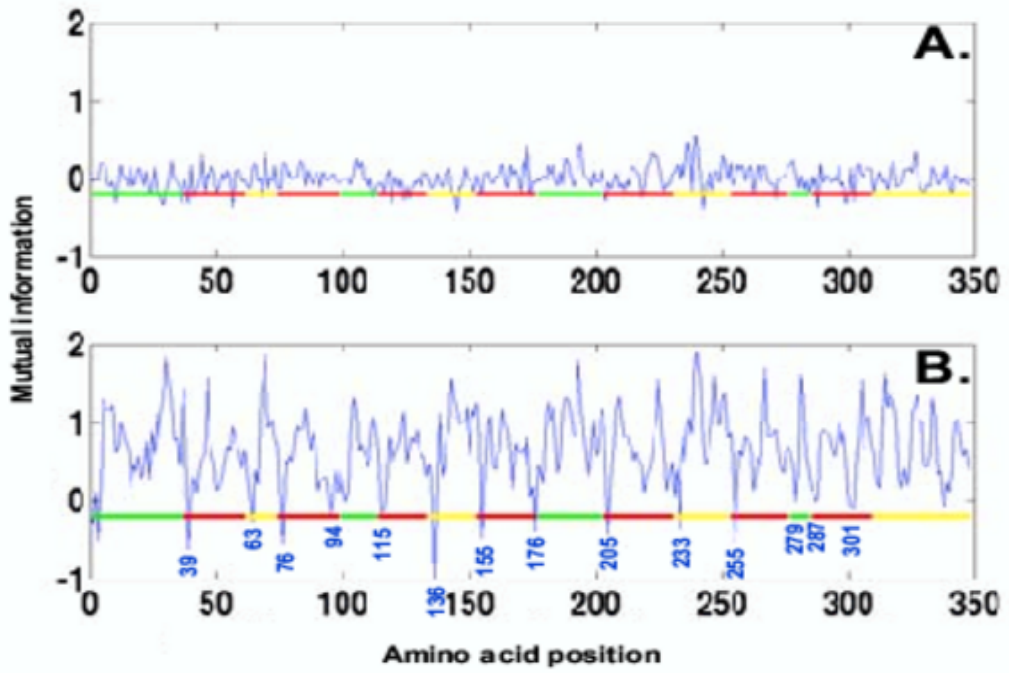


Figure 2. Mutual information values along the rhodopsin sequence using different datasets (A. human, B. GPCR) to generate mutual information values. Horizontal lines use the same color code as in Figure 1 indicating the positions of the segments belonging to each of ec, cp and helices domains based on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

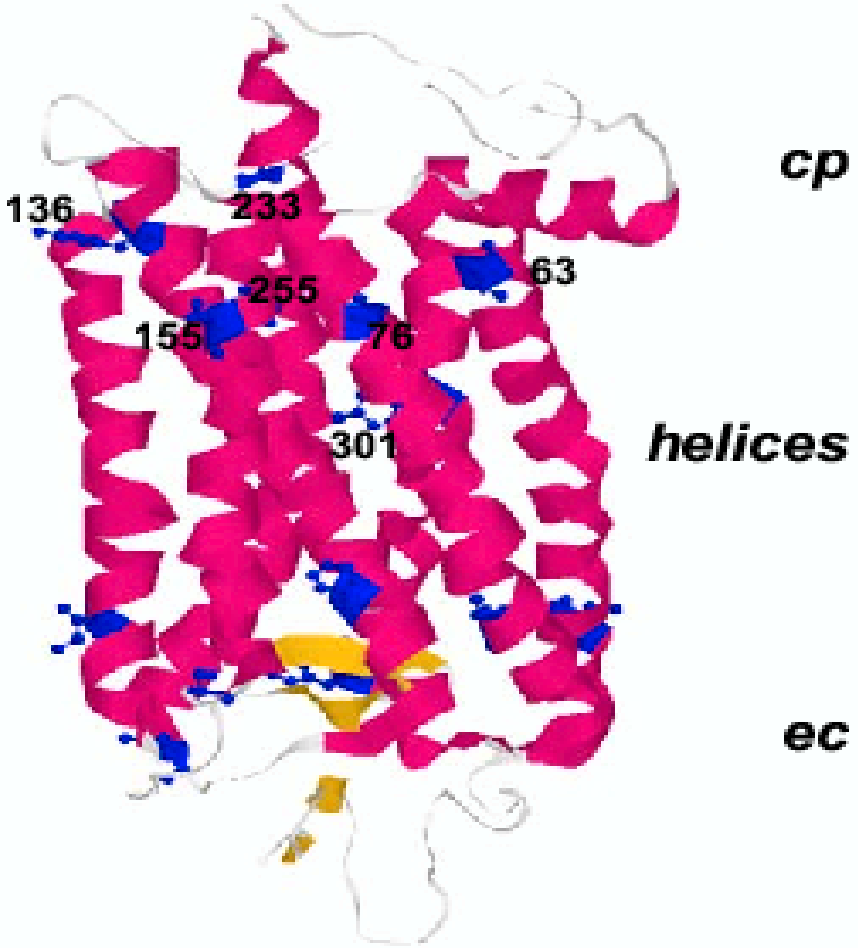


Figure 7. Tertiary structure of rhodopsin based on X-ray crystallography [3, 4]. All secondary structure elements (helices and beta-sheet) are shown in red, all loops are shown in gray. The position of breakpoints identified by mutual information (Figure 2B) is shown in blue. The corresponding amino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

Predikce struktury proteinu ze sekvence

http://hpcio.cit.nih.gov/Folding.htm

1) Modelování struktury na základě homologie
   (BLAST PDB)

2) Threading
   (ENERGY FUNCTION/PDB
   http://compbio.ornl.gov/structure/prospect

3) Skládání ze sekvenčně-strukturních fragmentů
   (I-SITES=CLUSTERING/PDB)

4) Ab initio modelování
   (ENERGY FUNCTIONS)

Threading

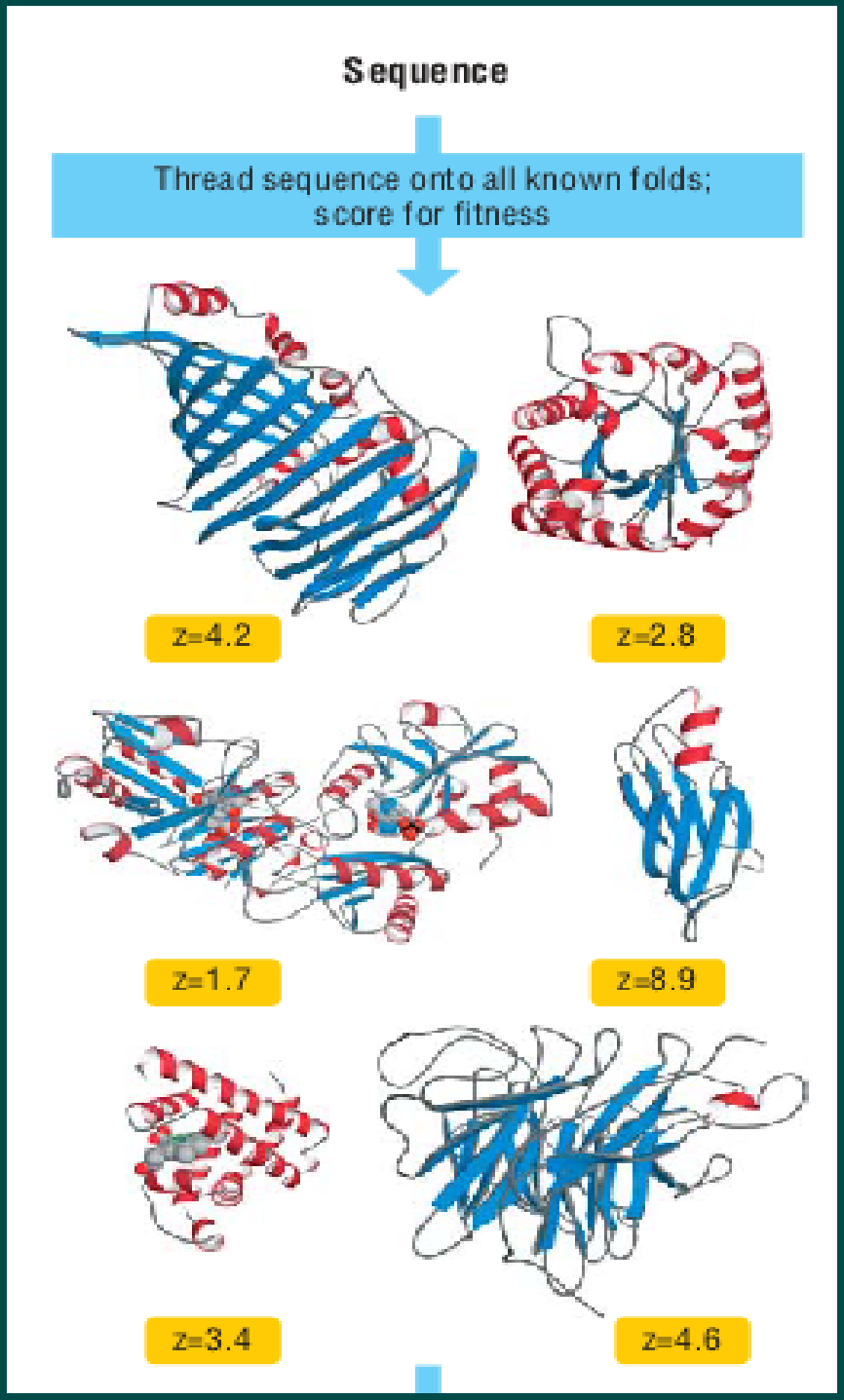
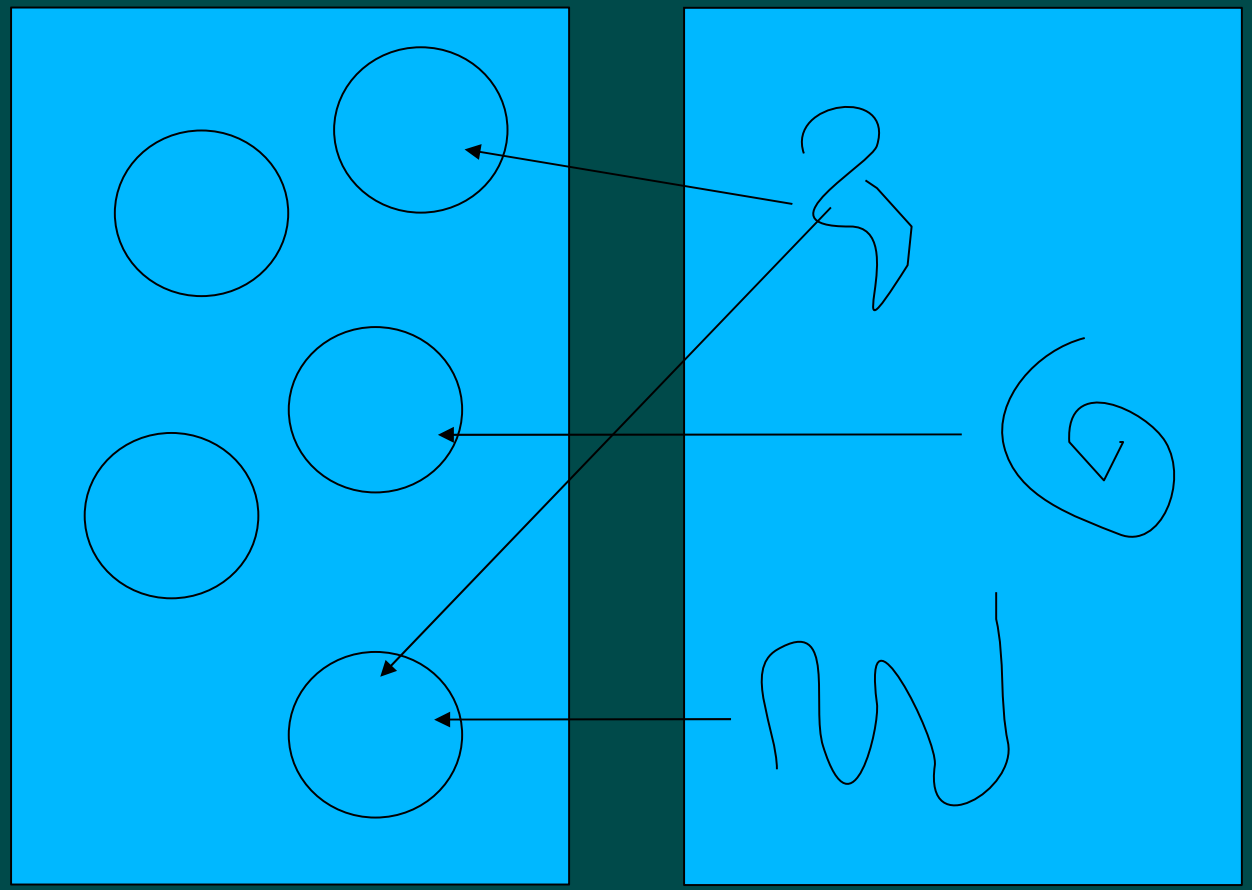


Figure 4-25 The method of profile-based threading A sequence of unknown structure is forced to adopt all known protein domain folds, and scored for its suitability for each fold. The z-value relates the score for the query sequence to the average score for a set of random sequences with the same amino-acid composition and sequence length. A very high z-score indicates that the sequence almost certainly adopts that fold. Sequences can be submitted online for threading by PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/index.html).

I-Sites

Princip

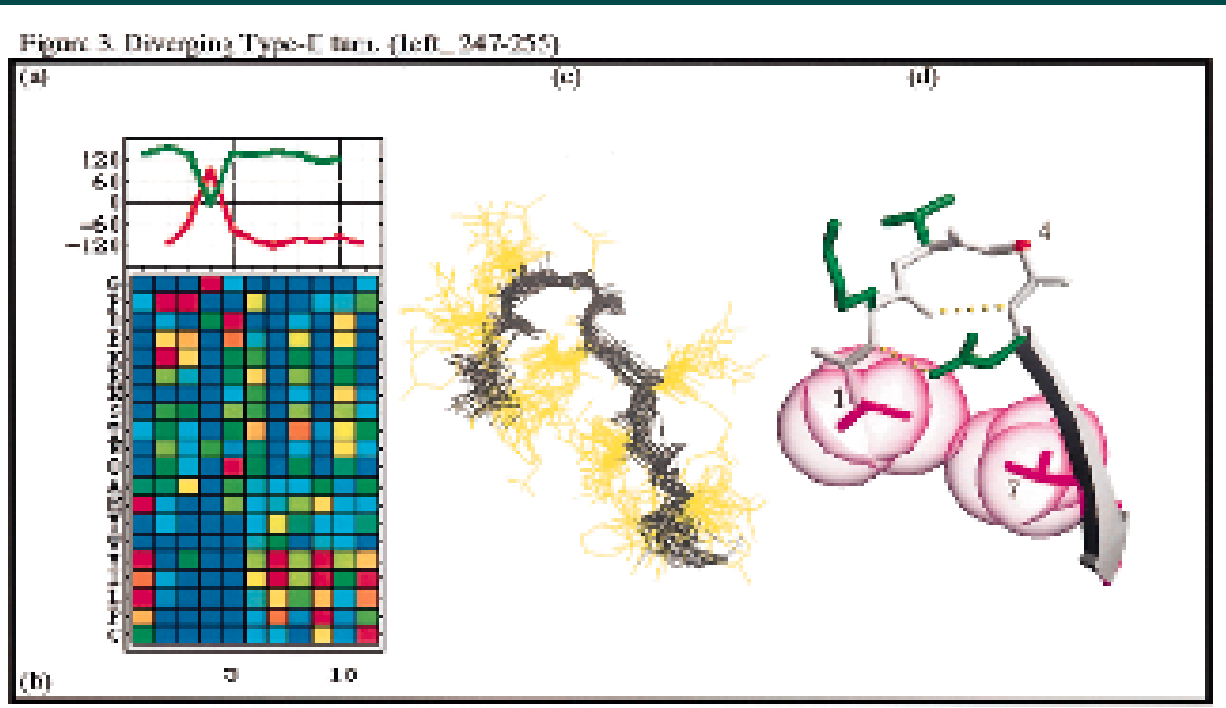


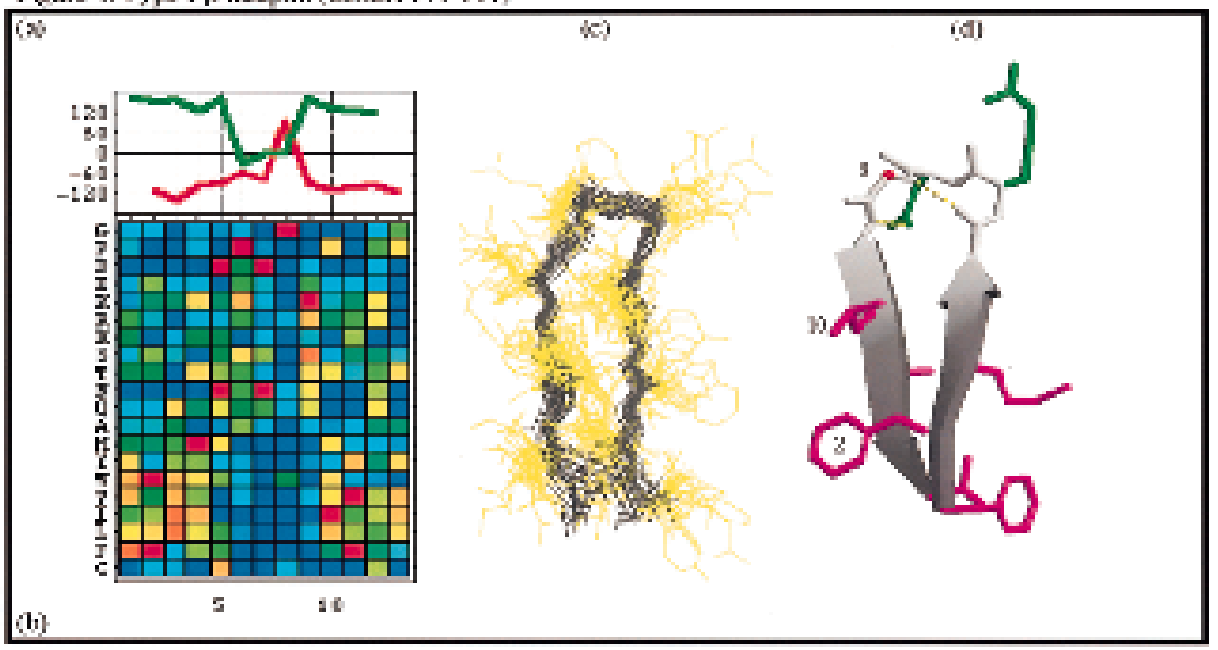
SEKVENCE

STRUKTURY

I-Sites

PSSM
Paradigma

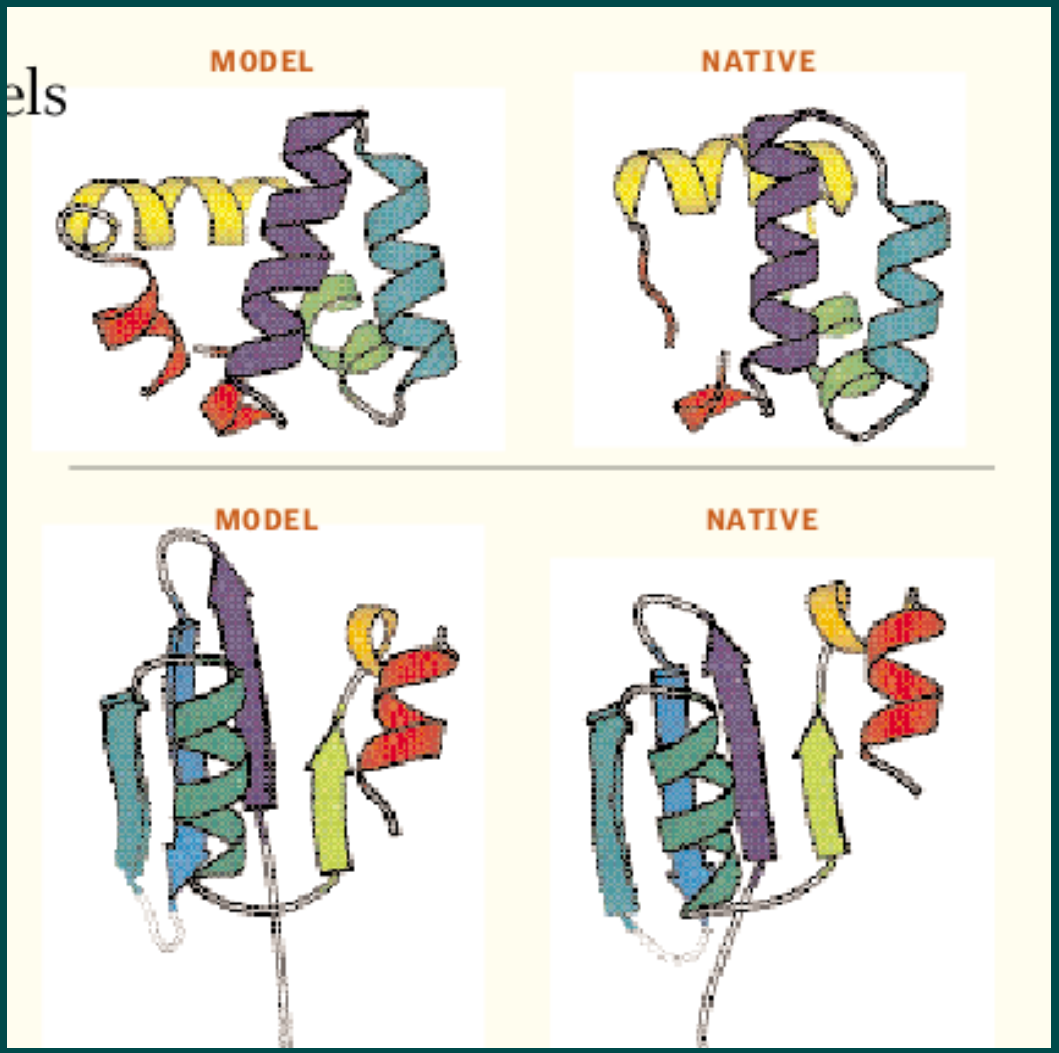




http://www.bioinfo.rpi.edu/applications/i-sites/Isites/

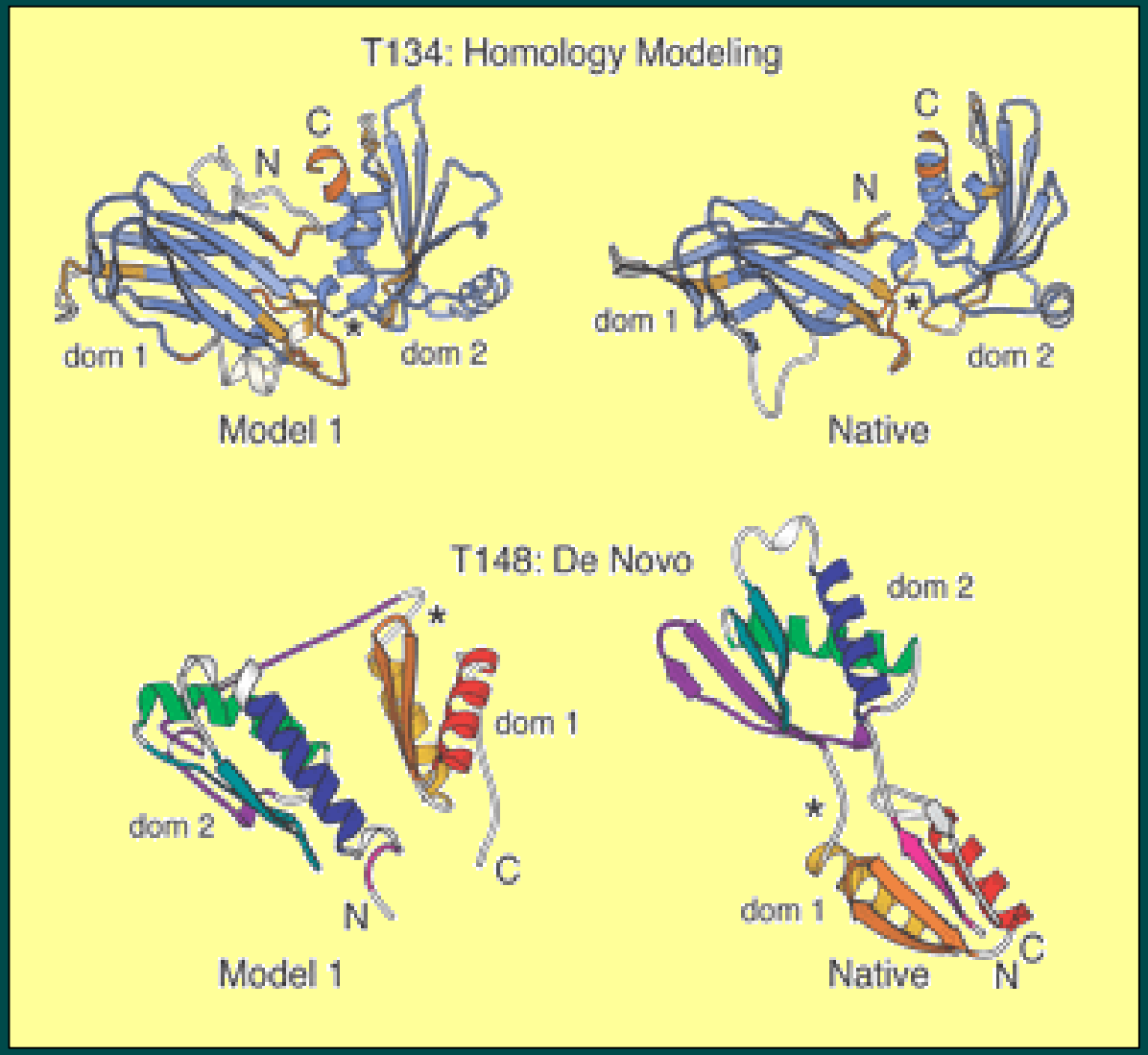
Příklady predikce struktury (Rosetta)

http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php



Příklady predikce struktury (Robetta)

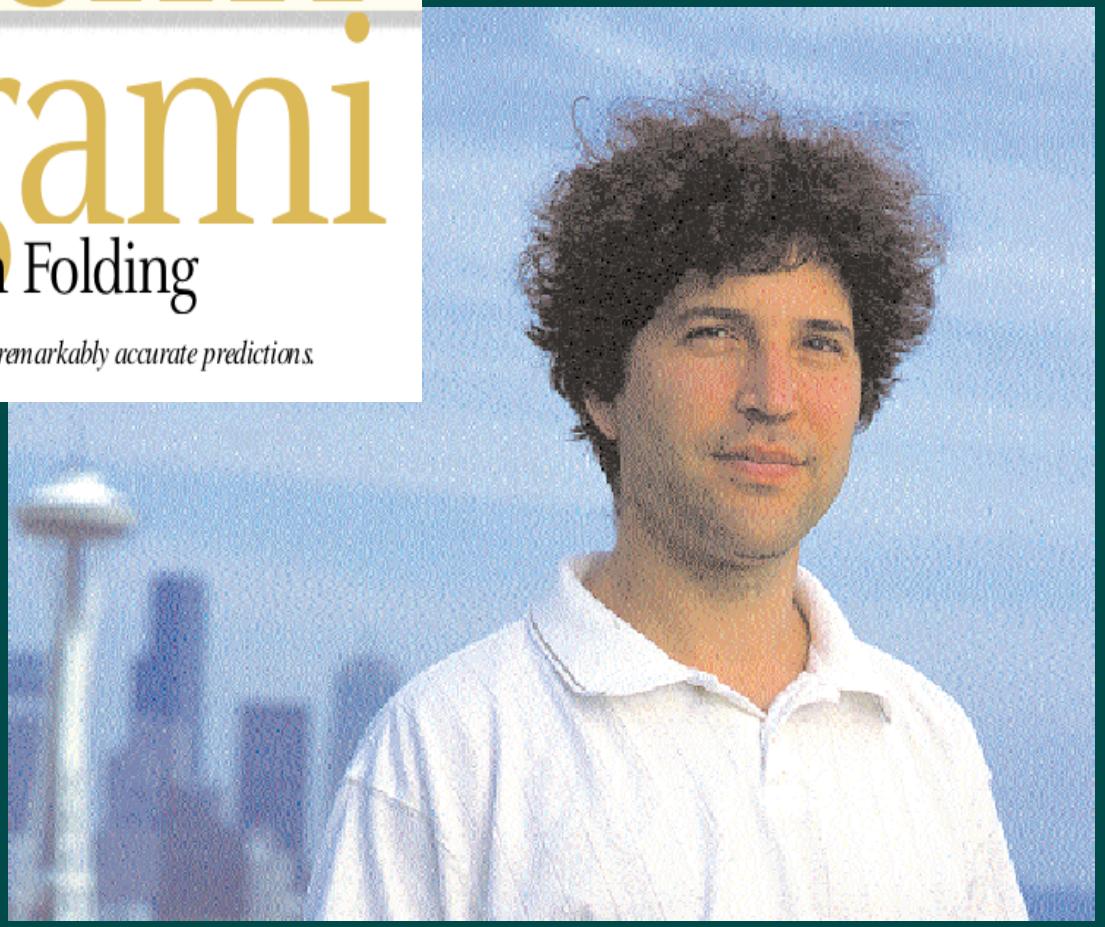
http://robetta.bakerlab.org/



Rosetta Tackles the

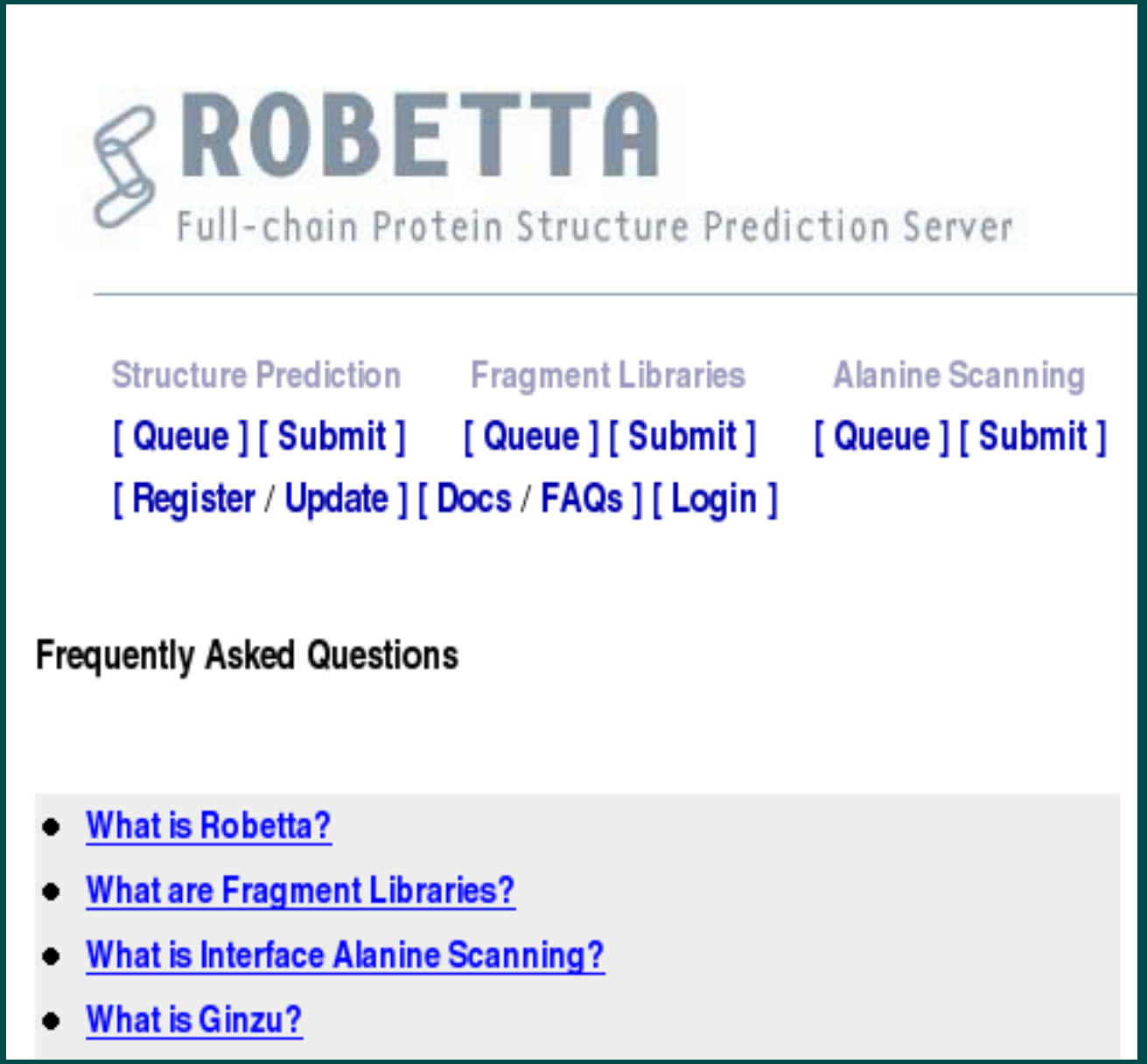
Extreme Origami of Protein Folding

David Baker's model is producing remarkably accurate predictions.



http://www.hhmi.org

http://robetta.bakerlab.org/documents/faqs.jsp



Často používané modely sekvence

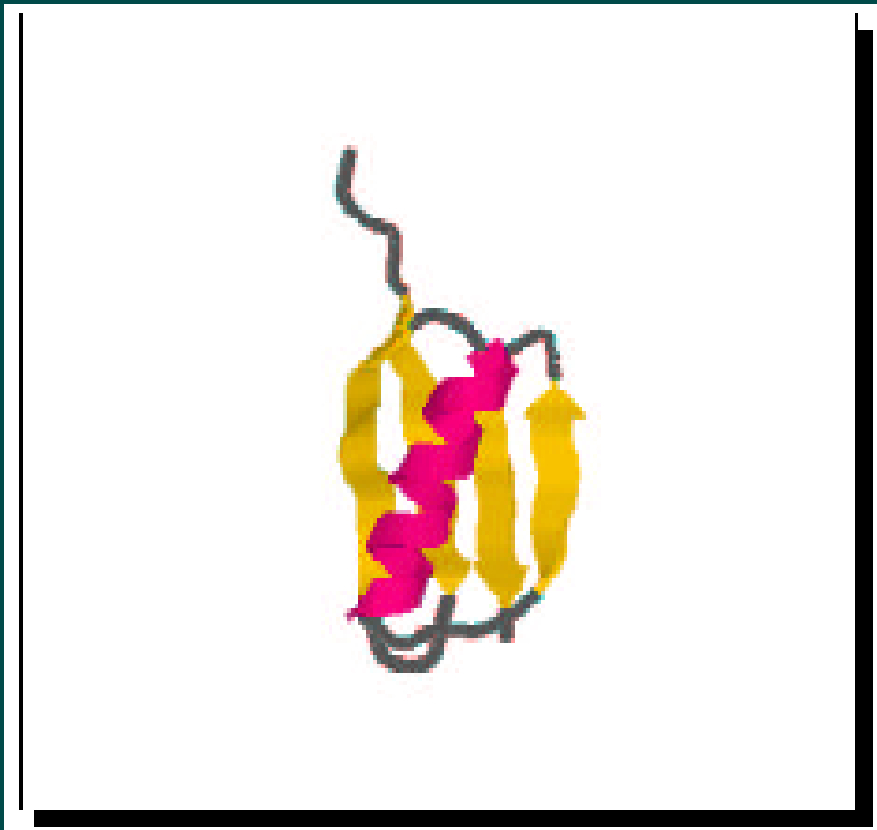
PSSM – Positional specific scoring matrix
(PWM – Positional weight matrix)

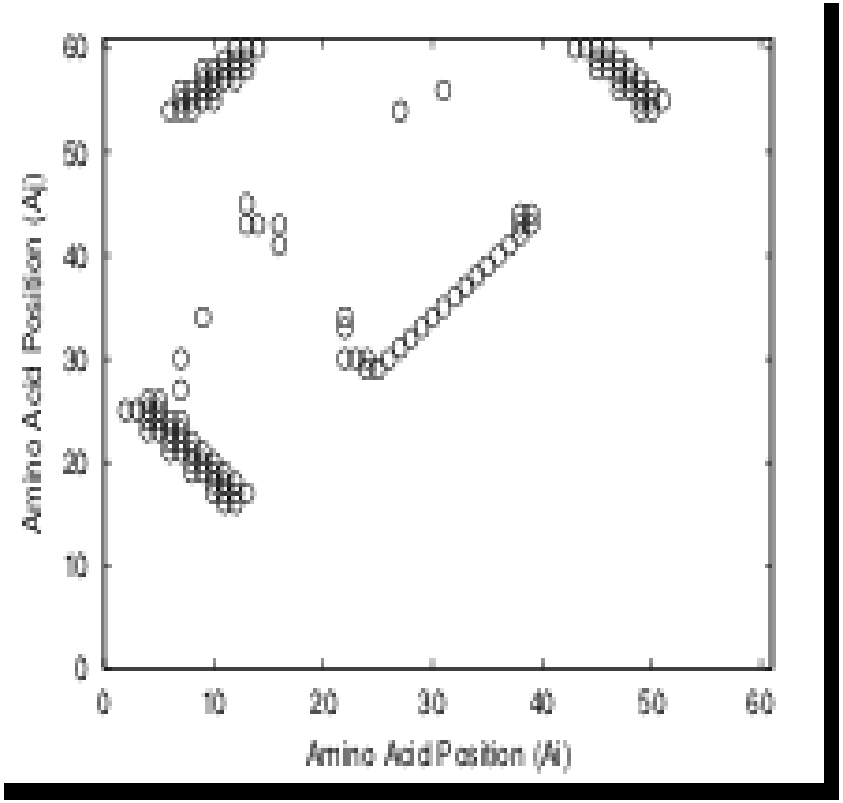
```
    1   2   3   4   5 ...
A  12   0   3   2   1
C   0   1 45 11   2
D   1   3   2   0   1
...
```

HMM – Hidden Markov Model

```
STATE 1 -p1-> STATE 2 -p2-> STATE 3
   |             |             |
OUTPUT        OUTPUT        OUTPUT
```

Kontaktní mapy





Analýzou kontaktních map je možné najít přírodou často používané typy kontaktů