

IV108 - Bioinformatika II

Ing. Matej Lexa, PhD.

<http://www.fi.muni.cz/~lexa/>

Út	8:00	B204	Předn·ška
	9:00	B116	Cvičení
Čt	13:00 – 15:00	C506	Konzultace

Navazuje na IV107 (Bioinformatika)

Probíh· paralelně s IV105-6 (Semin·ř Út 17:00 B411)

IV110,IV114 (Projekt Čt 8:00 B411)

Bloky předn·šek:

- 1) Informační obsah a struktura biologick"ch sekvencí
- 2) Vyhled·vací n·stroje
- 3) Předpovíd·ní a manipulace se strukturami biomolekul

Zkouška: Písemn· 50 bodů. Nejlépe hodnocené pr·ce ve cvičení můžou přispívat k celkovému bodovému hodnocení u zkoušky do v"šky 10 bodů. Studijní materi·ly budou specifikov·ny průběžně.

Klasifikace

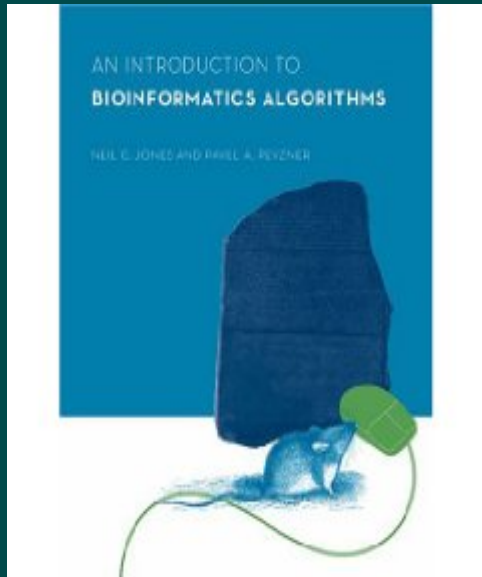
Zkouška

Vyřešen problém z cvičení = 1b (max 10b = 20%)

Písemná zkouška max. 50b (100%)

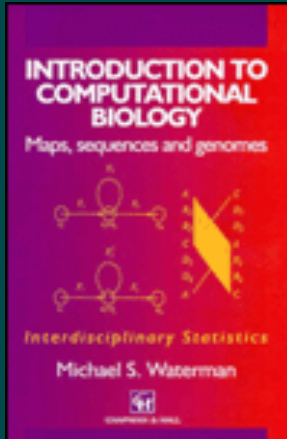
- **A – 91-120 %**
- **B – 81 - 90 %**
- **C – 71 - 80 %**
- **D – 61 - 70 %**
- **E – 51 - 60 %**
- **F – 0 - 50 %**

Studijní literatura

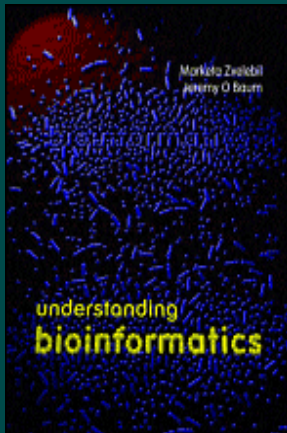


Neil C Jones and Pavel A Pevzner (2004).
An introduction to Bioinformatics
Algorithms.
The MIT Press, 454 s.
ISBN 0-2621-0106-8

Doplňkov· literatura



M.S.Waterman (1995).
Introduction to Computational Biology.
Chapman & Hall/CRC, 448 s.
ISBN 0-4129-9391-0

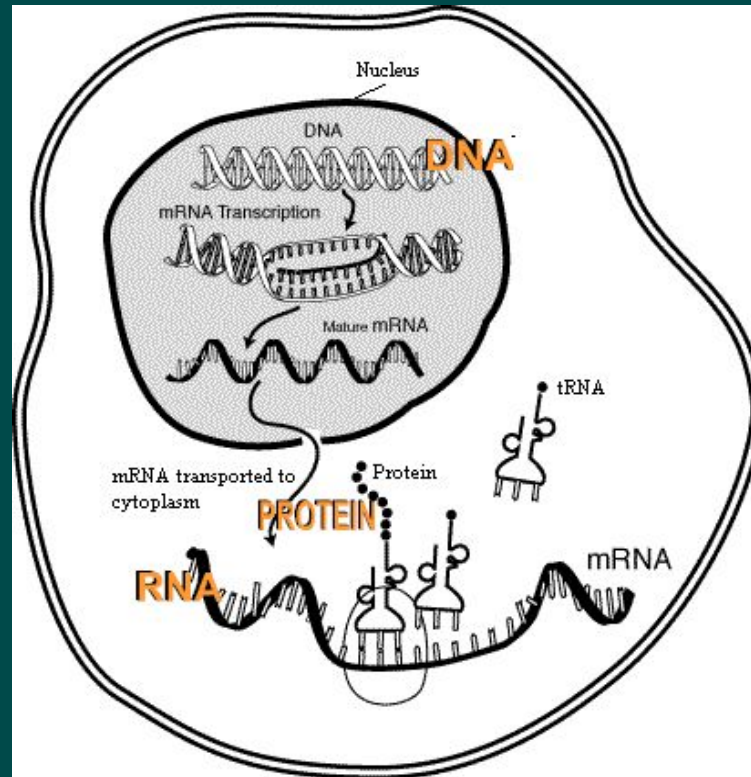


M.J.Zvelebil. And J.O.Baum (2007).
Understanding Bioinformatics.
Taylor & Francis, 750 s.
ISBN: 0-8153-4024-9

Co děl· bioinformatik

- Umí pracovat s velkými datovými soubory
 - Moudrými triky ovláda výkonné počítače
 - V datech hledá zajímavé subsekvence
 - Srovnává podobné sekvence
 - Předpovídá strukturu a funkci genů a proteinů
 - Studuje vývoj sekvencí a organismů
 - Data a výsledky analýz zobrazuje graficky
-
-

DNA - RNA - PROTEIN



SEKVENCE - STRUKTURA - FUNKCE

GENOTYP

```
>chs1  
atgacagaat  
acaggatgac  
tatgacgtga  
cggcttatat  
gatgacc...
```

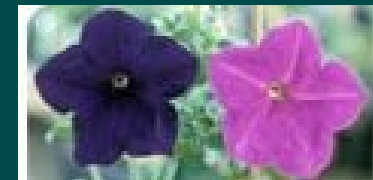
```
>chs1  
MFVDDHLA  
VNQNFYLR  
SHRQL...
```

GEN.KÓD



STRUKTURA

FENOTYP



FUNKCE



Biologick· sekvence (BS)

ACAGTGCGAGCATGACGATGACGCAGCAGATTGACAGAGACGATAGCAGCAT

MASAQSFYLLHLAVDDFMNGAGVLSHERELLYDENKIHDIVISMNDENMNQ

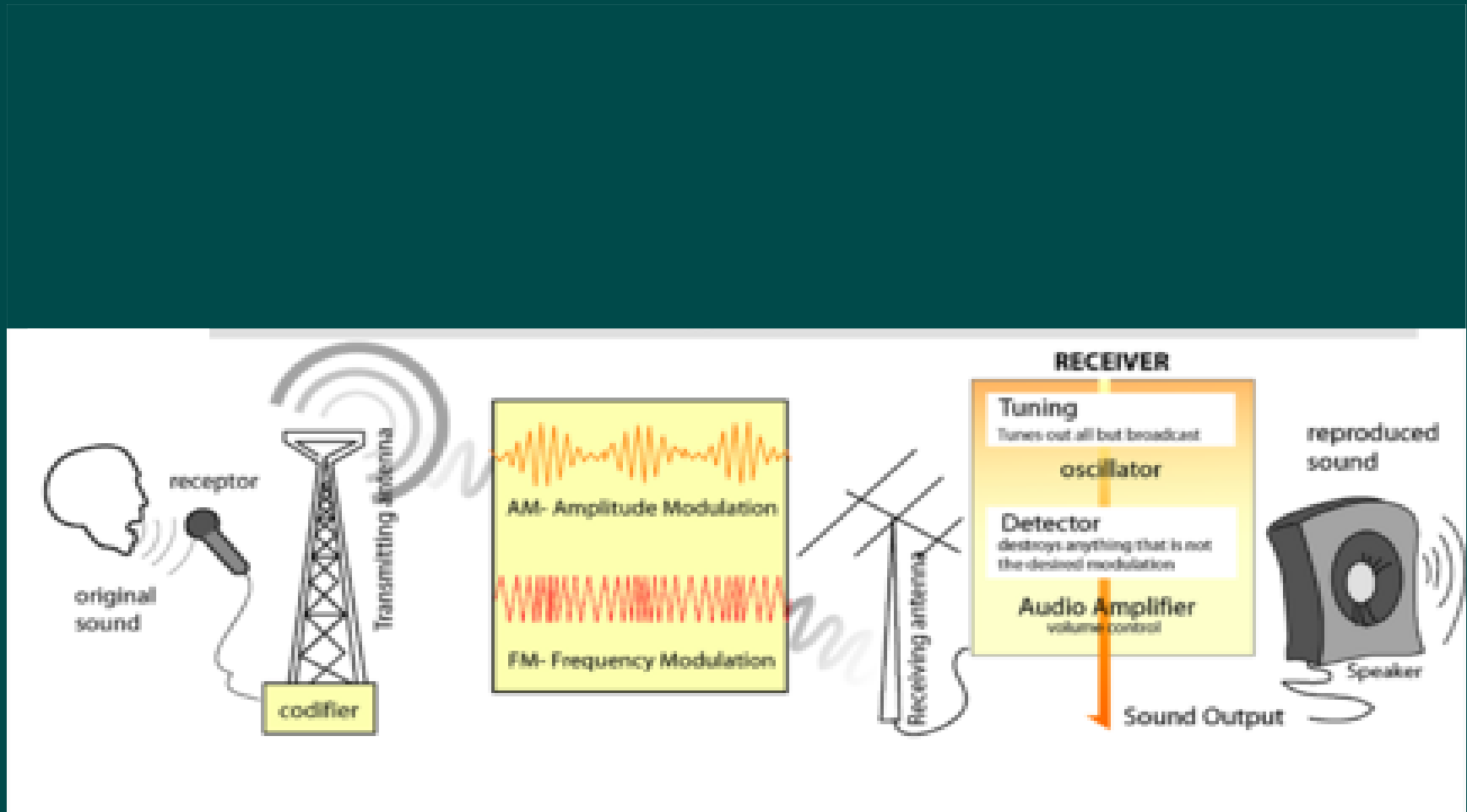
Jazyk

THISISJUSTASIMPLESENTENCEINENGLISHFORYOURINSPIRATION

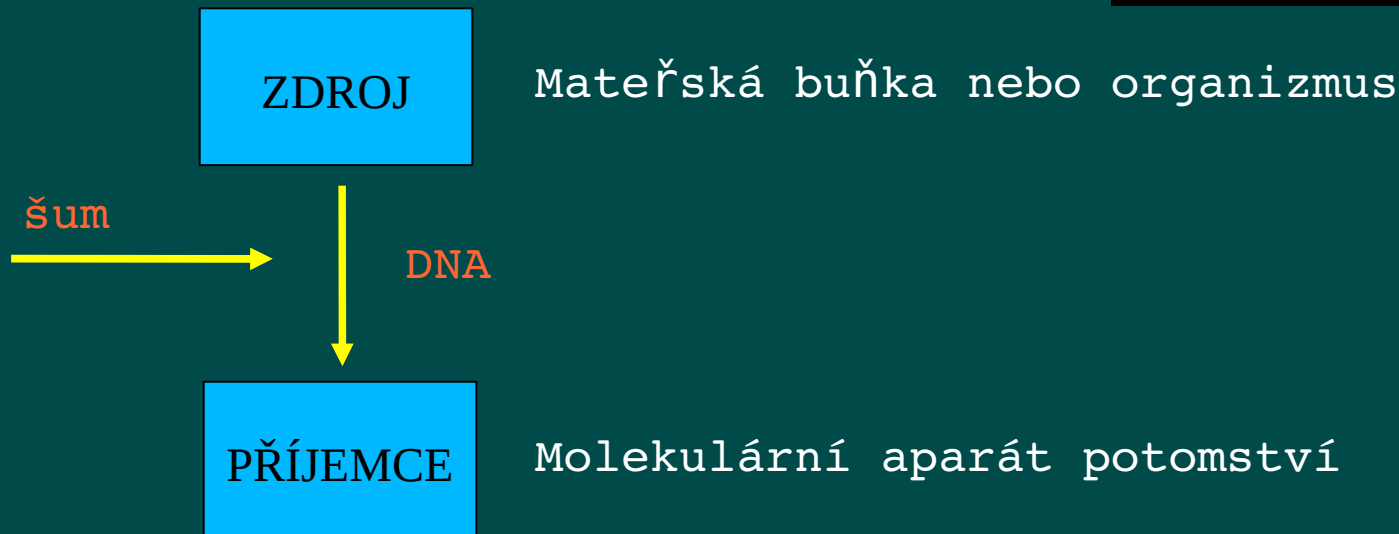
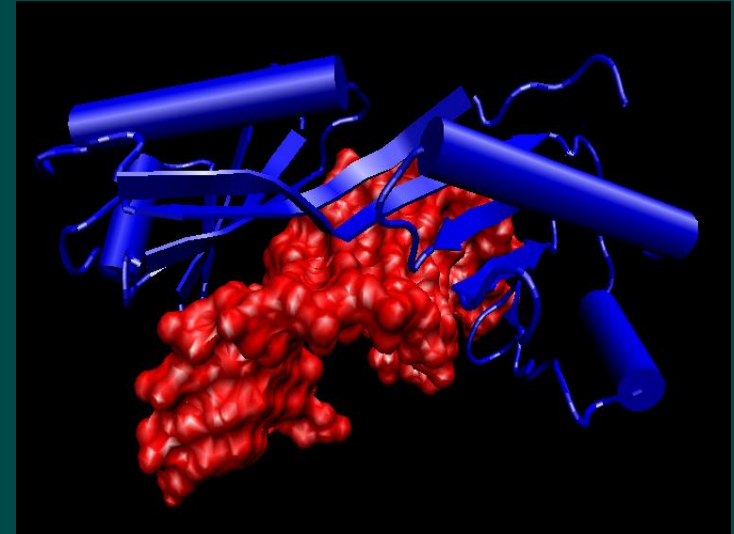


Informace

<http://en.wikipedia.org/wiki/Information>



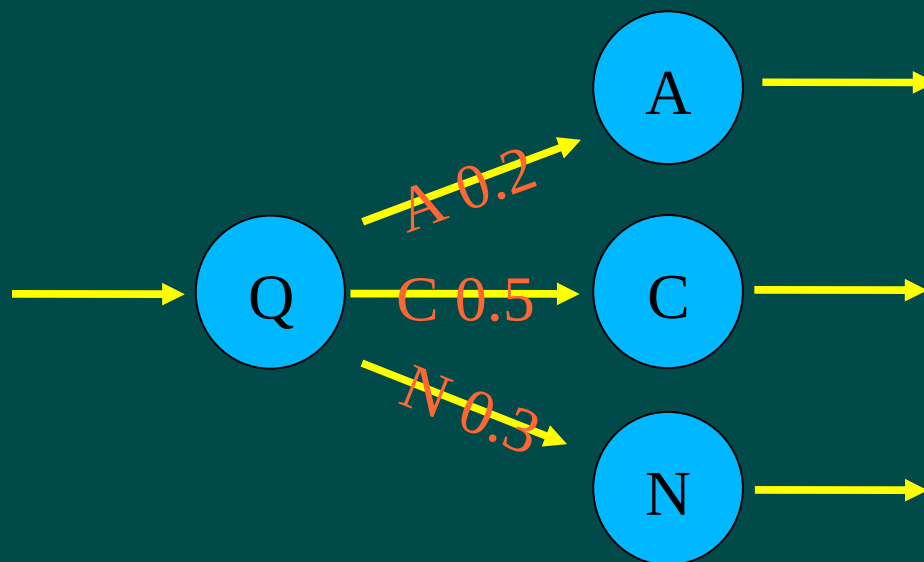
Biologick· sekvence jako informace, život jako komunikace mezi buňkami, DNA jako komunikační kan·l



Sekvence jako Markovův řetězec

NCMKLFQCDSHL

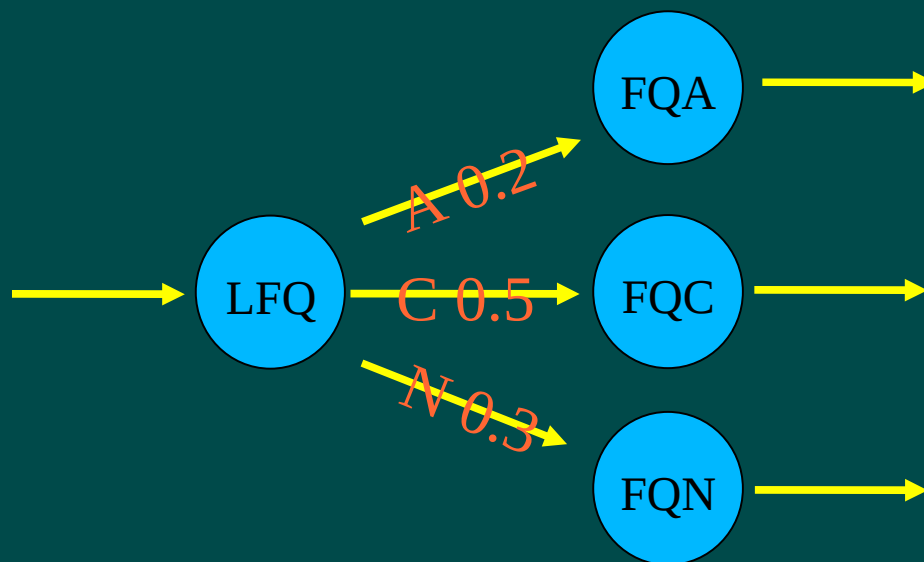
$$P(X_{i+1}|X_i) = P(X_{i+1}|X_0, \dots, X_i)$$



Sekvence jako Markovův řetězec

NCMKLFQCDSHL

NCM, CMK, MKL, KLF, LFQ, FQC, QCD, CDS, DSH, SHL



Frekvence

$$F(x) = P(x) N$$

Je vyšší u řetězců, které jsou součástí často používaných struktur

Vzjemn. informace

$$MI(x,y) = P(x,y) \log (P(x,y) / (P(x)*P(y)))$$

Je vyšší uvnitř struktur než na jejich rozhraní, vyjadřuje korelaci

Entropie

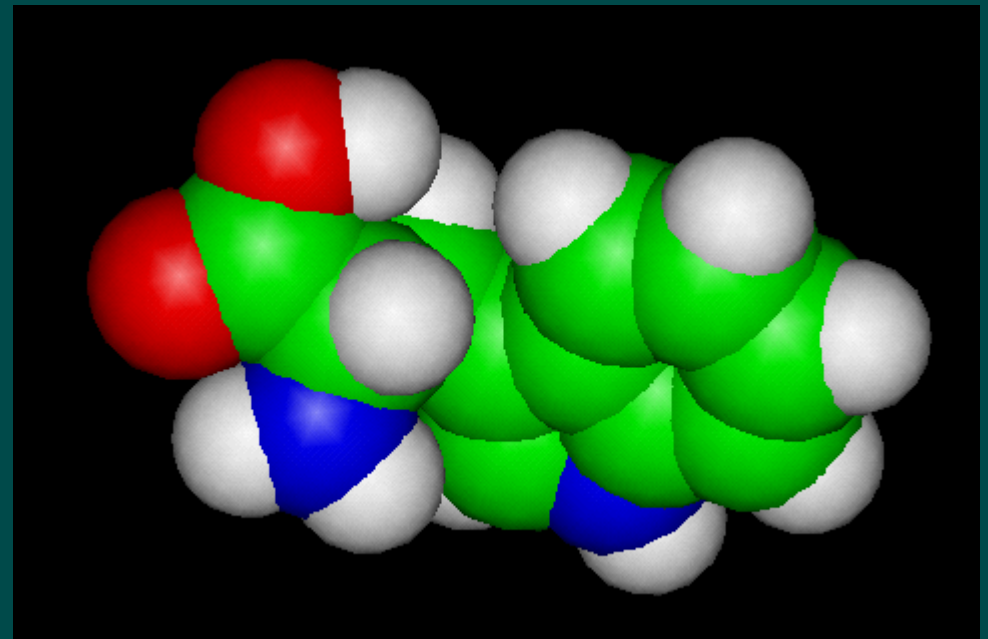
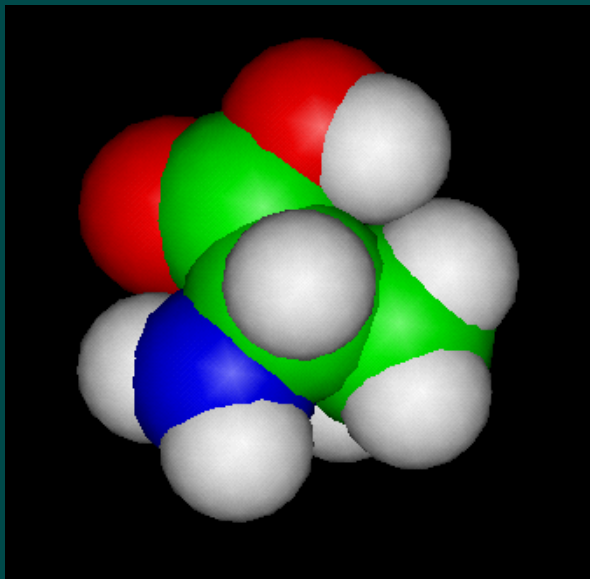
$$H(x) = -\sum p(x)*\log(p(x))$$

Určuje míru neuspořádanosti, nebo taky potřebu informace pro definování určitého stavu

Co vyj·druje frekvence v biologick"ch sekvencích

```
[lexa@peleus prot]$ egrep -c "SSS" ATH1.fa  
15927
```

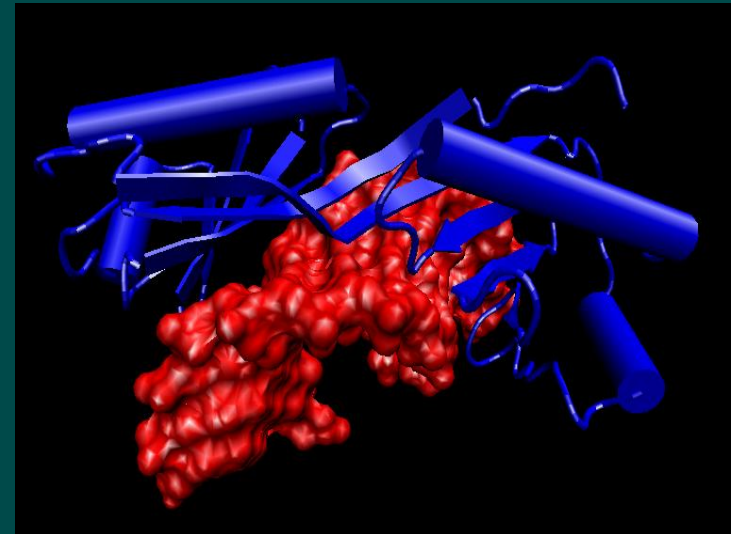
```
[lexa@peleus prot]$ egrep -c "WWW" ATH1.fa  
75
```



Co vyj·druje entropie v biologick"ch sekvencích

TATATAA
TATAAAA
TATATAT
TATAAAT

TATA.A. konsensus
TATAWAW RE, W=[TA]
0000101 entropie



$$H(x) = -\sum p(x) * \log(p(x))$$

Jin" pohled na entropii (podmíněn· entropie)

```
[lexa@peleus prot]$ egrep -c MASAL. ATH1.fa  
19  
[lexa@peleus prot]$ egrep -c MASALL ATH1.fa  
0  
[lexa@peleus prot]$ egrep -c MASALE ATH1.fa  
7
```

$$H(x) = -\sum p(x) * \log(p(x))$$

Co vyj·druje MI v biologick"ch sekvencích

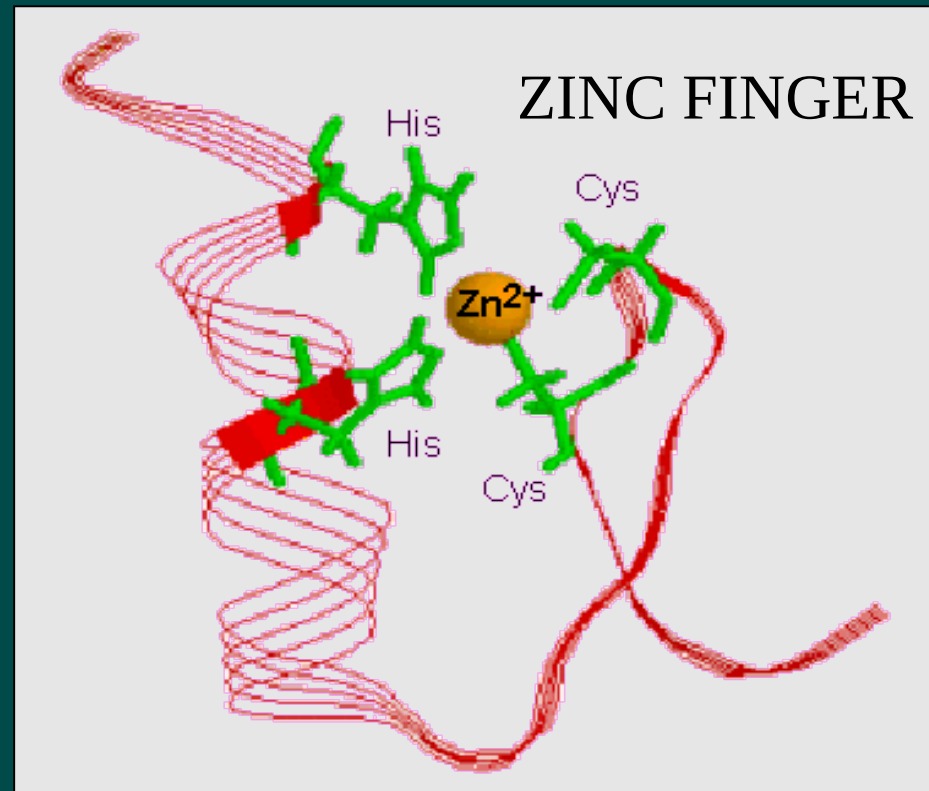
```
$ egrep -c "." ATH1.fa
233194
$ egrep -c "C..C" ATH1.fa
8196
$ egrep -c "H..H" ATH1.fa
7398
$ egrep -c "C..C.+H..H\|H..H.+C..C" ATH1.fa
1005
$ bc
bc 1.06
8196*1000/233194
35
7398*1000/233194
31
0.035*0.031*233194
233.194
```

$$MI(x,y) = P(x,y) \log(P(x,y) / (P(x)*P(y)))$$

Co vyj·druje MI v biologick"ch sekvencích

```
$ egrep -c "." ATH1.fa
233194
$ egrep -c "C..C" ATH1.fa
8196
$ egrep -c "H..H" ATH1.fa
7398
$ egrep -c "C..C.+H..H\|H..H.+C..C" ATH1.fa
1005
$ bc
bc 1.06
8196*1000/233194
35
7398*1000/233194
31
0.035*0.031*233194
233.194
```

$$MI(x,y) = P(x,y) \log(P(x,y) / (P(x)*P(y)))$$



Shannon 1948. A mathematical theory of communication.

1. Zero-order approximation (symbols independent and equiprobable).

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNESBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, ..., n -gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

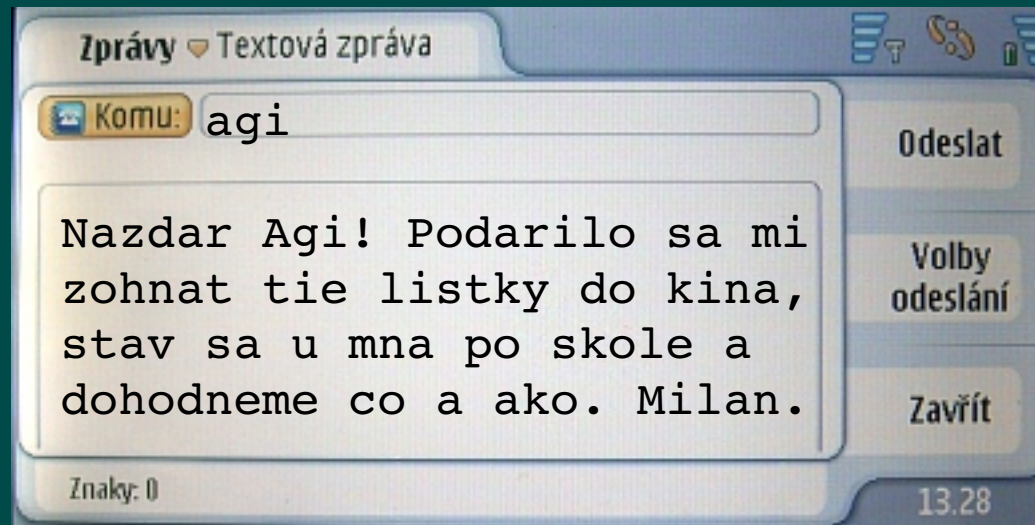
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that

Běžn. sekvence proteinu m. informační obsah několika SMS zpr.v.



Běžn· sekvence proteinu m· informační obsah několika SMS zpr·v.

