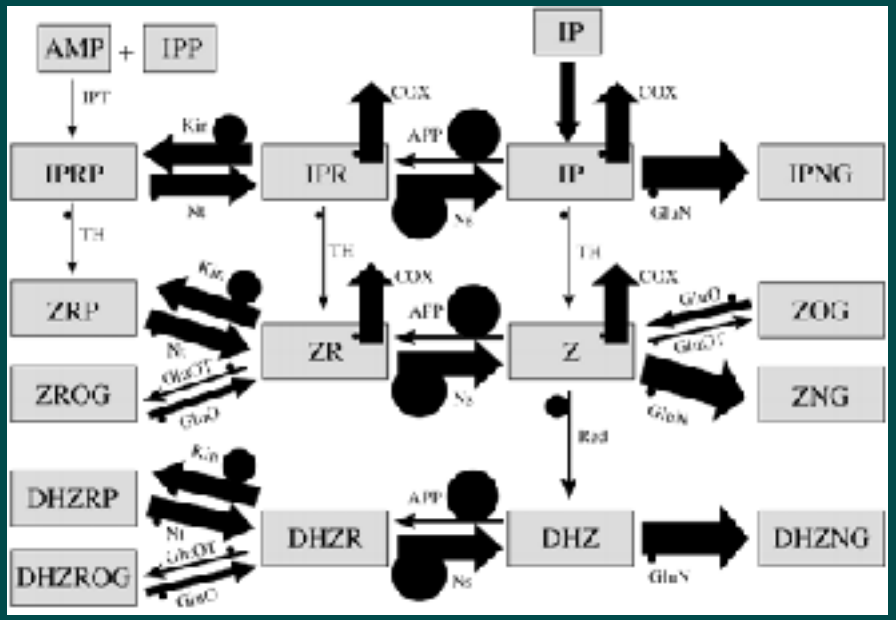
# Štatistická segmentácia biologických sekvencií: Čierna mágia alebo krok správnym smerom?

**Matej Lexa**
**Fakulta Informatiky MU Brno**

# Numerické modely biologických procesov

- *Príjem, metabolizmus a transport dusíka v rastlinách*
- *Pohyb vody a živín pri kvapkovej závlahe rastlín*
- *Rast a rozdelenie biomasy a dusíka medzi korene a nadzemnú časť*
- *Premeny cytokinínov v rastlinách*

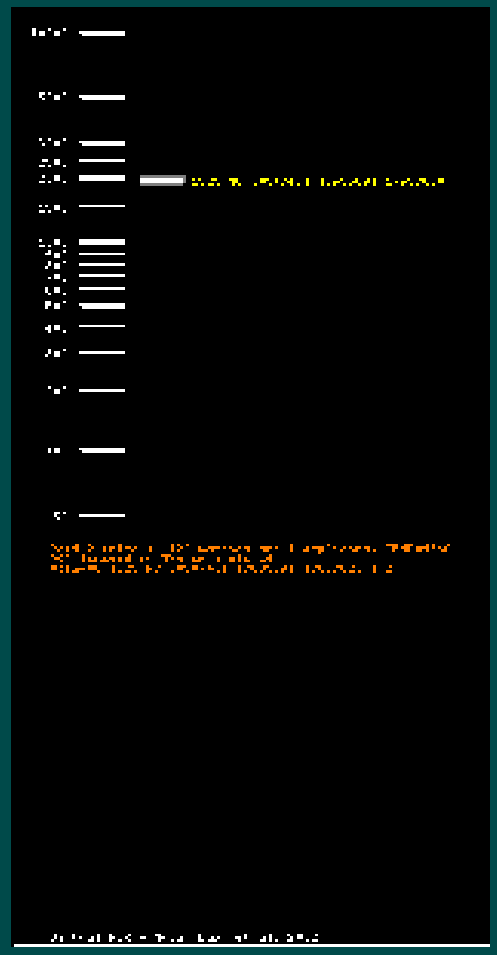


# Numerické modely biologických procesov

- *Príjem, metabolizmus a transport dusíka v rastlinách*
- *Pohyb vody a živín pri kvapkovej závlahe rastlín*
- *Rast a rozdelenie biomasy a dusíka medzi korene a nadzemnú časť*
- *Premeny cytokinínov v rastlinách*

## Simulácia PCR

*VPCR 2.0* *http://www.sci.muni.cz/lmfr/vpcr/*



## Numerické modely biologických procesov

- *Príjem, metabolizmus a transport dusíka v rastlinách*
- *Pohyb vody a živín pri kvapkovej závlahe rastlín*
- *Rast a rozdelenie biomasy a dusíka medzi korene a nadzemnú časť*
- *Premeny cytokinínov v rastlinách*

## Simulácia PCR

*VPCR 2.0* *http://www.sci.muni.cz/lmfr/vpcr/*

## Hľadanie približného výskytu krátkych reťazcov

*PRIMEX 1.0* *http://bioinformatics.cribi.unipd.it/primex/*

## Numerické modely biologických procesov

- *Príjem, metabolizmus a transport dusíka v rastlinách*
- *Pohyb vody a živín pri kvapkovej závlahe rastlín*
- *Rast a rozdelenie biomasy a dusíka medzi korene a nadzemnú časť*
- *Premeny cytokinínov v rastlinách*

## Simulácia PCR

*VPCR 2.0* *http://www.sci.muni.cz/lmfr/vpcr/*

## Hľadanie približného výskytu krátkych reťazcov

*PRIMEX 1.0* *http://bioinformatics.cribi.unipd.it/primex/*

## Analýza biologických sekvencií (BS)

*Analýza výskytu krátkych reťazcov a ich kolokácií*

# Biologická sekvencia (BS)

*MASAQSFYLLHLAVDDFMNGAGVLSHERELLFYDENKIHDIVISMNDENMN*

# Biologická sekvencia (BS)

*MASAQSFYLLHLAVDDFMNGAGVLSHERELLFYDENKIHDIVISMNDENMN*

# BS majú podobné vlastnosti ako jazyk

- *predstavuje kombinatoriálny systém na určitej abecede*
- *sekvencie majú rozdielne zastúpenie n-gramov*
- *vyvíja sa pod vplyvom prostredia a podlieha selekčným tlakom*
- *kóduje zložitejšiu štruktúru*
- *tvoria ju elementy, ktoré majú analogické kategórie v jazyku*

# Biologická sekvencia (BS)

*MASAQSFYLLHLAVDDFMNGAGVLSHERELLFYDENKIHDIVISMNDENMN*

# BS majú podobné vlastnosti ako jazyk

- *predstavuje kombinatoriálny systém na určitej abecede*
- *sekvencie majú rozdielne zastúpenie n-gramov*
- *vyvíja sa pod vplyvom prostredia a podlieha selekčným tlakom*
- *kóduje zložitejšiu štruktúru*
- *tvoria ju elementy, ktoré majú analogické kategórie v jazyku*

# Toto poznanie formuje nové smery v bioinformatike

- *Biosemiotika (Hoffmeyer 1991, 1997)*
- *Analógia biológia/jazyk (Pattee 1980, Sereno 1991)*
- *Center for Biological Language Modelling (J.K.Seetharaman, Pittsburg, PA, USA)*
- *Biological Language Conferences (2003, 2004)*

# Pracovná verzia analógie biológia/jazyk pre proteíny

| | |
|---|---|
| *proteín* | *veta* |
| *doména , motív* | *fráza* |
| *segment (?)* | *slovo* |
| *perióda sekundárnej štruktúry* | *slabika* |
| *aminokyselina* | *písmeno, hláska* |
| | |
| *funkcia domény, proteínu* | *význam frázy, vety* |
| *funkcia segmentu, vazba* | *význam slova* |
| | |
| *malé molekuly* | *mentálne reprezentácie objektov* |
| *chemické reakcie* | *myslenie, učenie* |
| *metabolická dráha* | *rozhovor* |
| | |
| *štruktúra proteínu* | *neurologický obraz, myšlienka* |
| *pravidlá skladania proteínov* | *syntax* |

# Čo je to slovo?

- *reťazec ohraničený medzerami*
- *slovo je najmenšia časť jazyka, ktorá má svoj definovaný význam*

## Čo je to slovo?

- *reťazec ohraničený medzerami*
- *slovo je najmenšia časť jazyka, ktorá má svoj definovaný význam*

## Ak abeceda nemá medzery a nepoznáme význam?

- *reťazec nachádzajúci sa v slovníku*

## Čo je to slovo?

- *reťazec ohraničený medzerami*
- *slovo je najmenšia časť jazyka, ktorá má svoj definovaný význam*

## Ak abeceda nemá znak pre medzery?

- *reťazec nachádzajúci sa v slovníku*

## Ak slovník neobsahuje všetky tvary (skloňovanie)?

- *reťazec morfologicky podobný reťazcu nachádzajúcemu sa v slovníku*

## Čo je to slovo?

- *reťazec ohraničený medzerami*
- *slovo je najmenšia časť jazyka, ktorá má svoj definovaný význam*

## Ak abeceda nemá znak pre medzery?

- *reťazec nachádzajúci sa v slovníku*

## Ak slovník neobsahuje všetky tvary (skloňovanie)?

- *reťazec morfologicky podobný reťazcu nachádzajúcemu sa v slovníku*

## Ak jazyk nemá slovník, alebo slovník nestačí?

- *??? - pozrieme sa na japončinu, skúsime to s BS*

# Prečo je dobré vedieť určiť slová?

*V jazyku sú to slová, ktoré podliehajú syntaktickým pravidlám a vytvárajú tak konečný význam vety.*

# Prečo je dobré vediet určiť slová?

*V jazyku sú to slová, ktoré podliehajú syntaktickým pravidlám a vytvárajú tak konečný význam vety.*

*Je predpoklad, že pochopenie štruktúry a funkcie biologických sekvencií bude jednoduchšie, ak by sme poznali lexikon a gramatiku biologického jazyka a tie by mohli byť založené práve na slovách.*

# Prečo je dobré vediet určiť slová?

*V jazyku sú to slová, ktoré podliehajú syntaktickým pravidlám a vytvárajú tak konečný význam vety.*

*Je predpoklad, že pochopenie štruktúry a funkcie biologických sekvencií bude jednoduchšie, ak by sme poznali lexikon a gramatiku biologického jazyka a tie by mohli byť založené práve na slovách.*

*Napríklad pri počítačovej analýze japonského textu boli úspešnejšie postupy, ktoré najprv našli čo najvačší počet kandidátov na slová a potom použili pravidlá syntaxe aby určili, ktoré z nich sú skutočne použité*

*než*

*programy, ktoré sa snažili určiť identitu slova len z jeho blízkeho okolia.*

# Prečo je dobré vediet určiť slová?

*V jazyku sú to slová, ktoré podliehajú syntaktickým pravidlám a vytvárajú tak konečný význam vety.*

*Je predpoklad, že pochopenie štruktúry a funkcie biologických sekvencií bude jednoduchšie, ak by sme poznali lexikon a gramatiku biologického jazyka a tie by mohli byť založené práve na slovách.*

*Napríklad pri počítačovej analýze japonského textu boli úspešnejšie postupy, ktoré najprv našli čo najvačší počet kandidátov na slová a potom použili pravidlá syntaxe aby určili, ktoré z nich sú skutočne použité*

*než*

*programy, ktoré sa snažili určiť identitu slova len z jeho blízkeho okolia.*

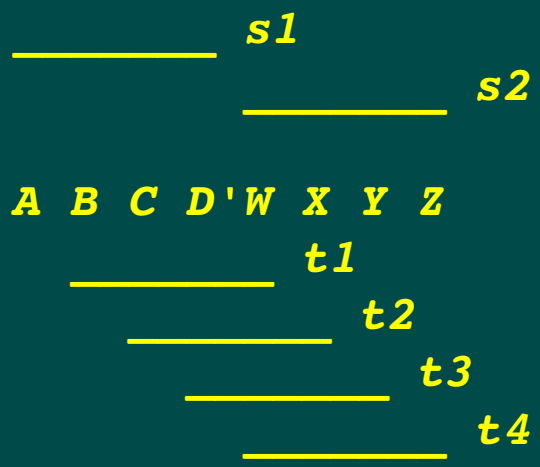
*To pripomína povestný strop 75-80% pri určovaní sekundárnej štruktúry neurálnymi sieťami či SVM v okne asi 20 aminokyselín.*

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

*kanji, hiragana, katakana – znaky roznej úrovne*

*kanji sú na úrovni našich slabík a tvoria polovicu slov*

*sekvencie kanji sa často dajú segmentovať viacerými sposobmi*




漢英字典刊行会

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

*Pre kazdu medzeru sa vypocita hodnota (s1+s2)/(t1+...+tn)*


漢英字典刊行会

# Vstupné dáta pre analýzu textu vo formáte FASTA

```
>SENTENCE
THECALLOFTHEWILD

>SENTENCE
BYJACKLONDON

>SENTENCE
CHAPTERONE

>SENTENCE
BUCKDIDNOTREADTHENEWSPAPERSORHEWOULDHAVEKNOWNTHATTROUBLEWASBREWING
```

# Vyhodnotenie frekvencie 4-gramov v texte

```
__<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
```

# Segmentácia textu v angličtine

*~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 |

# Vyhodnotenie frekvencie 4-gramov v PDB

```
_<CGVGFIAN 0 7 143 860 568 439 437 432 0.407
<CGVGFIANL 7 143 860 568 439 437 432 643 1.287
CGVGFIANLR 143 860 568 439 437 432 643 505 1.141
GVGFIANLRG 860 568 439 437 432 643 505 977 1.245
VGFIANLRGK 568 439 437 432 643 505 977 488 0.876
GFIANLRGKP 439 437 432 643 505 977 488 620 1.227
FIANLRGKPD 437 432 643 505 977 488 620 309 0.763
IANLRGKPDH 432 643 505 977 488 620 309 127 0.767
ANLRGKPDHT 643 505 977 488 620 309 127 145 1.16
NLRGKPDHTL 505 977 488 620 309 127 145 292 0.662
LRGKPDHTLV 977 488 620 309 127 145 292 1144 1.754
RGKPDHTLVE 488 620 309 127 145 292 1144 1377 2.209
GKPDHTLVEQ 620 309 127 145 292 1144 1377 434 2.908
KPDHTLVEQA 309 127 145 292 1144 1377 434 418 1.059
PDHTLVEQAL 127 145 292 1144 1377 434 418 744 0.287
DHTLVEQALK 145 292 1144 1377 434 418 744 938 0.862
HTLVEQALKA 292 1144 1377 434 418 744 938 1168 2.489
TLVEQALKAL 1144 1377 434 418 744 938 1168 1418 1.18
LVEQALKALG 1377 434 418 744 938 1168 1418 1162 0.942
VEQALKALGC 434 418 744 938 1168 1418 1162 363 1.026
EQALKALGCM 418 744 938 1168 1418 1162 363 26 0.812
QALKALGCME 744 938 1168 1418 1162 363 26 37 0.593
ALKALGCMEH 938 1168 1418 1162 363 26 37 63 0.946
LKALGCMEHR 1168 1418 1162 363 26 37 63 164 3.082
KALGCMEHRG 1418 1162 363 26 37 63 164 176 6.761
ALGCMEHRGG 1162 363 26 37 63 164 176 738 1.9
```

# Segmentácia textu v PDB

```
<              20.428
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
```

# Segmentácia textu v PDB

```
<                20.428
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 pohľad na segmenty s hranicou > 3*

# 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 NBNESEBYA 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 DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRL.

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

# Chang et al., 1995. Automatic construction of a Chinese electronic dictionary.

*Na rozdiel od práce ktorá na japonských textoch používala frekvenciu výskytu n-gramov, v tejto práci kombinujú autori tri ukazovatele*

- *Frekvenciu*
- *Vzájomnú informáciu*
- *Entropiu*

# Chang et al., 1995. Automatic construction of a Chinese electronic dictionary.

*Na rozdiel od práce ktorá na japonských textoch používala frekvenciu výskytu n-gramov, v tejto práci kombinujú autori tri ukazovatele*

- *Frekvenciu*
- *Vzájomnú informáciu*
- *Entropiu*

# Frekvencia

*Hranice slov sú zvyčajne v oblastiach s nízkou frekvenciou n-gramov*

# Chang et al., 1995. Automatic construction of a Chinese electronic dictionary.

*Na rozdiel od práce ktorá na japonských textoch používala frekvenciu výskytu n-gramov, v tejto práci kombinujú autori tri ukazovatele*

- *Frekvenciu*
- *Vzájomnú informáciu*
- *Entropiu*

## Frekvencia

*Hranice slov sú zvyčajne v oblastiach s nízkou frekvenciou n-gramov*

## *Vzájomná informácia*

*MI(x,y) = log ( P(x,y) / (P(x)*P(y)) )*
*n-gramy x a y s vysokými hodnotami sa vyskytujú spolu častejšie než možno očakávať z frekvencie ich výskytu, k tomu dochádza často vnútri slov a frází*

# Chang et al., 1995. Automatic construction of a Chinese electronic dictionary.

*Na rozdiel od práce ktorá na japonských textoch používala frekvenciu výskytu n-gramov, v tejto práci kombinujú autori tri ukazovatele*

- *Frekvenciu*
- *Vzájomnú informáciu*
- *Entropiu*

## Frekvencia

*Hranice slov sú zvyčajne v oblastiach s nízkou frekvenciou n-gramov*

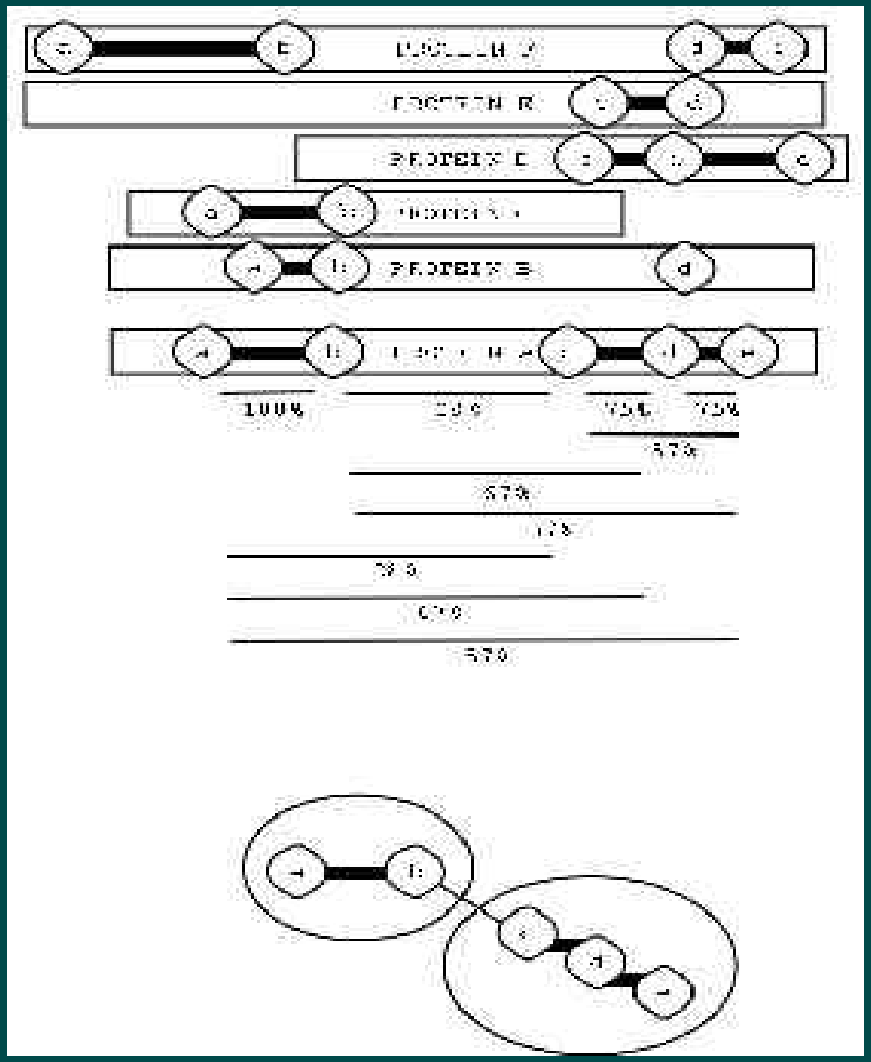
## *Vzájomná informácia*

*MI(x,y) = log ( P(x,y) / (P(x)*P(y)) )*
*n-gramy x a y s vysokými hodnotami sa vyskytujú spolu častejšie než možno očakávať z frekvencie ich výskytu, k tomu dochádza často vnútri slov a frází*

## *Entropia*

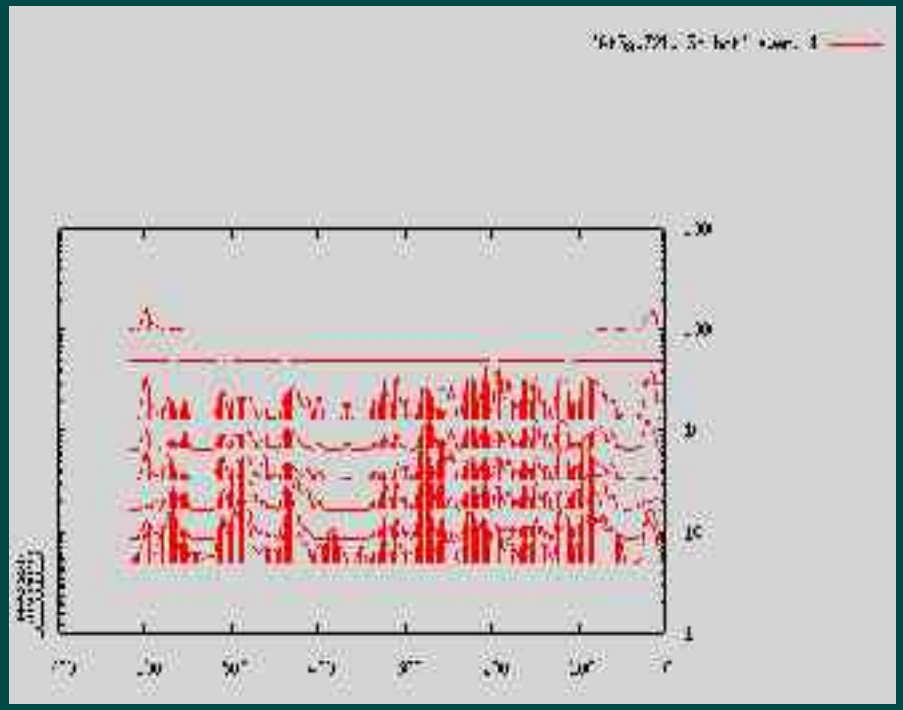
*H(x) = - p(x,c)*log(p(x,c))*
*Určuje neusporiadanost v ľavom a pravom okolí n-gramu x, tá je vyššia na hranici slov a frází*

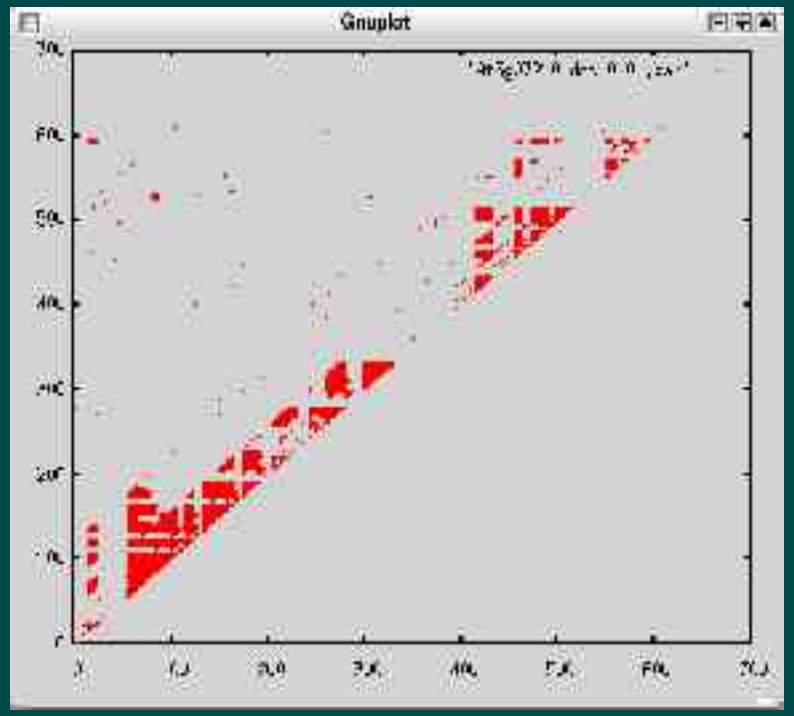
# Spoluvýskyt krátkych sekvencií v proteínoch



*Jedným z dovodov spoluvýskytu krátkych sekvencií je, že spolu vytvárajú samostatnú doménu, ktorá sa vyskytuje vo viacerých proteínoch*

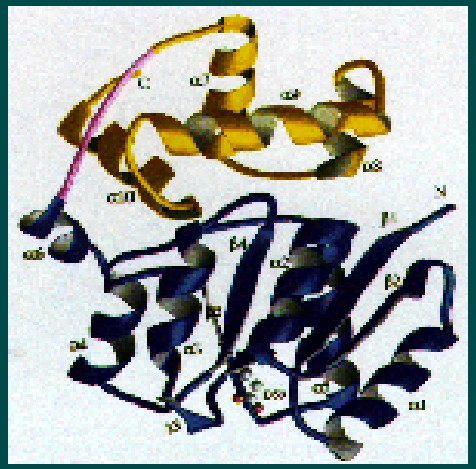
# *Hľadanie domén pomocou kolokácií krátkych segmentov*



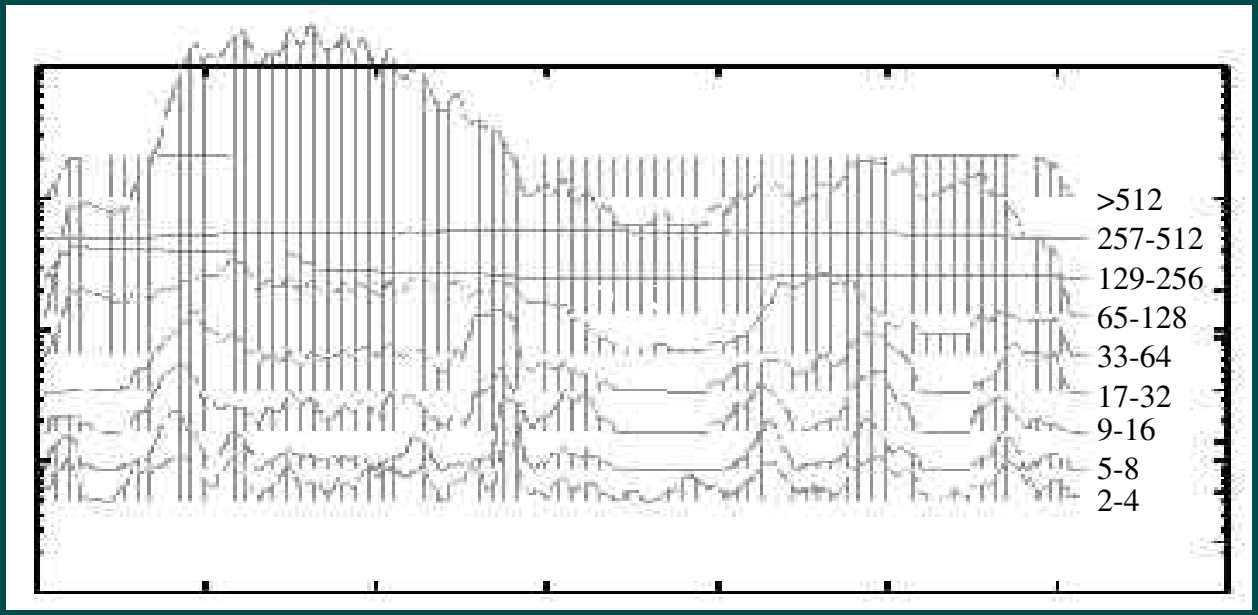


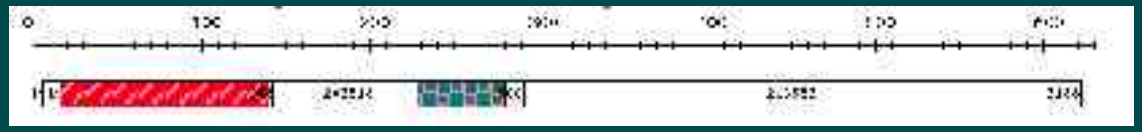

*A — Počet kolokácií vychádzajúcich z jednotlivých pozícií sekvencie*

*B — "Dotplot" korelovaných párov ukazuje ktoré oblasti sekvencie tvoria vyššie jednotky*



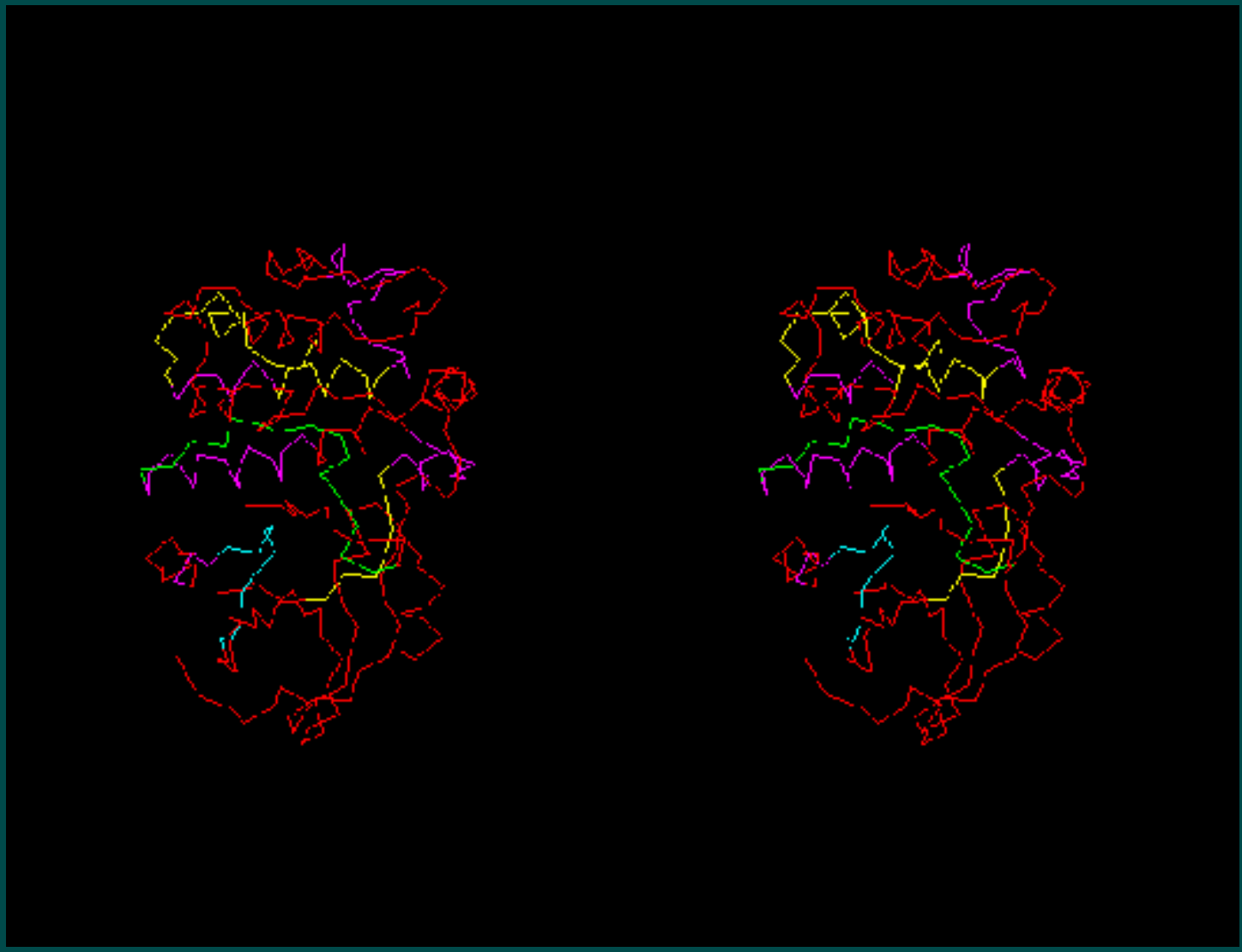
# Vyhodnotenie hľadania domén



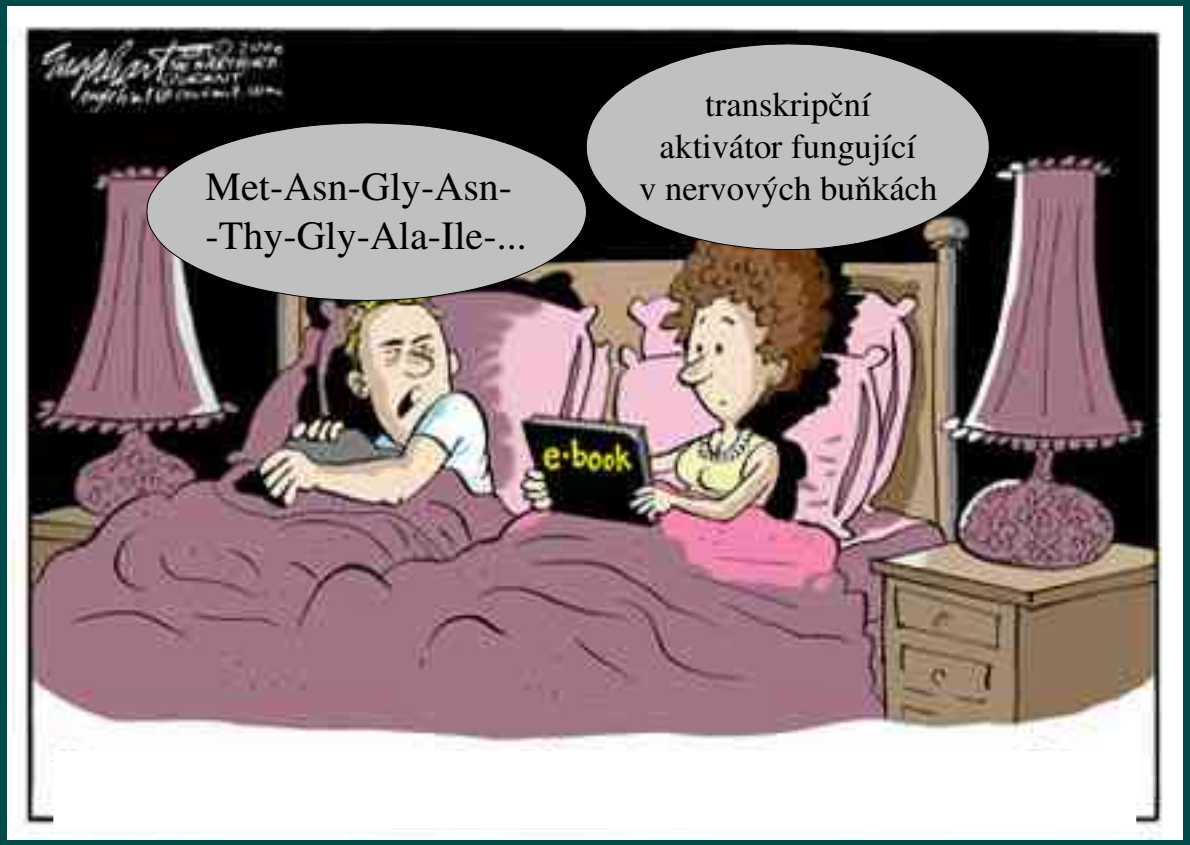



*Počet korelácií prechádzajúcich daným miestom proteínu Atg07210 porovnaný so záznamom v databáze PRODOM*

# *Priestorové usporiadanie segmentov s vysokou mierou asociácie v sekvencii*



# *Cieľom je dopracovať sa k slovníku často používaných slov v biologických sekvenciách*



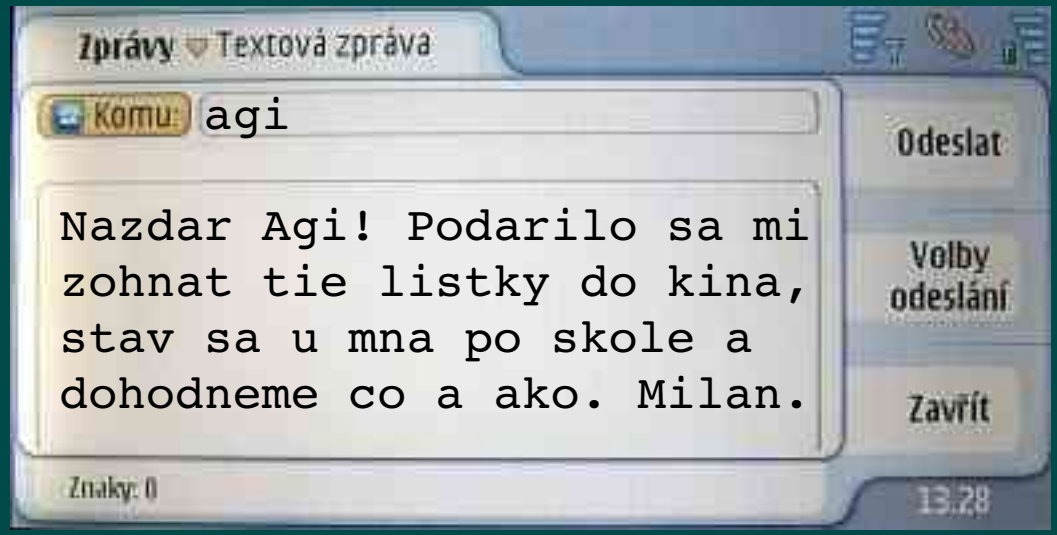
# Bussemaker et al., 2000. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.

*Postupne budujú slovník dlhších a dlhších slov s priradenými pravdepodobnosťami, ktoré sa nachádzajú v promótoroch s frekvenciou vyššou, než je možné očakávať v náhodnej sekvencii.*

| *motif* | *factor* | *-log(P)* |
|---|---|---|
| *TTTCCNNNNNNGGAAA* | *MCM1* | *6* |
| *ATACANNNNTACAT* | *?* | *10* |

# *Bioinformatik sa musí pozerať na biologické dáta inými očami*

***Bežná sekvencia proteínu má informačný obsah niekoľkých SMS správ.***



***Bežná sekvencia proteínu má informačný obsah niekoľkých SMS správ.***