Język(i) białek?

przesłanki
przyczynki
perspektywy

Witold Dyrka

Warszawa, 22.02.2016 r.
seminarium Zespołu Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN

Przesłanki

Białko mimo liniowego charakteru jest niewątpliwie konstrukcją przestrzenną. Aminokwasy można sobie wyobrazić jako pewne przestrzenne figury geometryczne, a białko — jako konfigurację tych podstawowych brył, złożoną według pewnych prawideł. Różnym aminokwasom mogą oczywiście odpowiadać różne figury, co jeszcze bardziej komplikuje sprawę.

Dla uproszczenia przyjmijmy, że każdy aminokwas można sobie wyobrazić w postaci płaskiej figury-trójkąta równobocznego. Białku natomiast będzie odpowiadać mozaika otrzymana ze sklejania trójkątów krawędziami. Zamiast więc badać proces syntezy białka z aminokwasów, będziemy rozpatrywać składanie mozaik z trójkątów oraz spróbujemy znaleźć język pozwalający na opisywanie procesu konstruowania mozaik, zakładając, że w zdaniach tego języka mogą występować tylko cztery rodzaje liter.

Z. Pawlak. Gramatyka i matematyka, PWN W-wa 1965

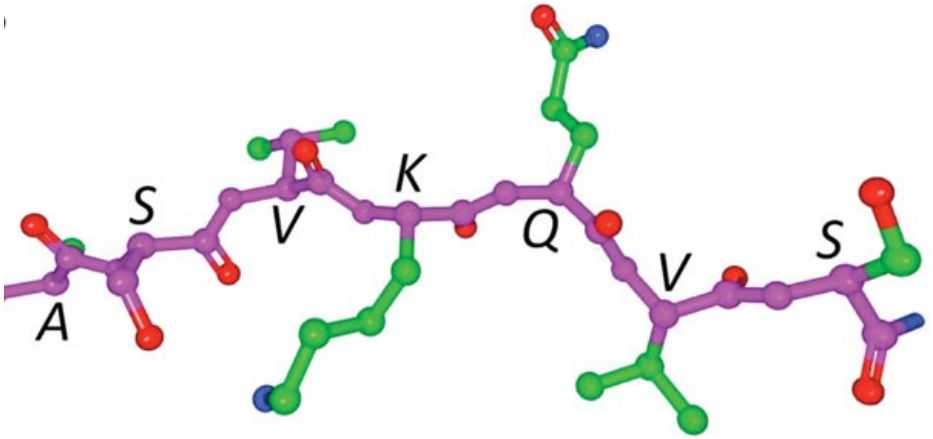
Założenia

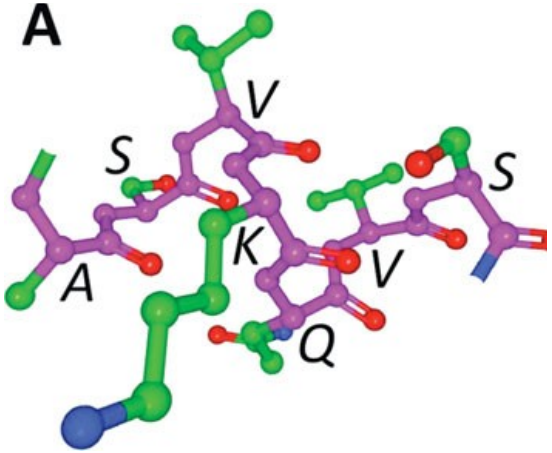
- **Hipoteza termodynamiczna**

  Natywna struktura przestrzenna białka jest zdeterminowana przez sekwencję aminokwasów [Afinsen'73]

- Struktura przestrzenna białka determinuje jego funkcję

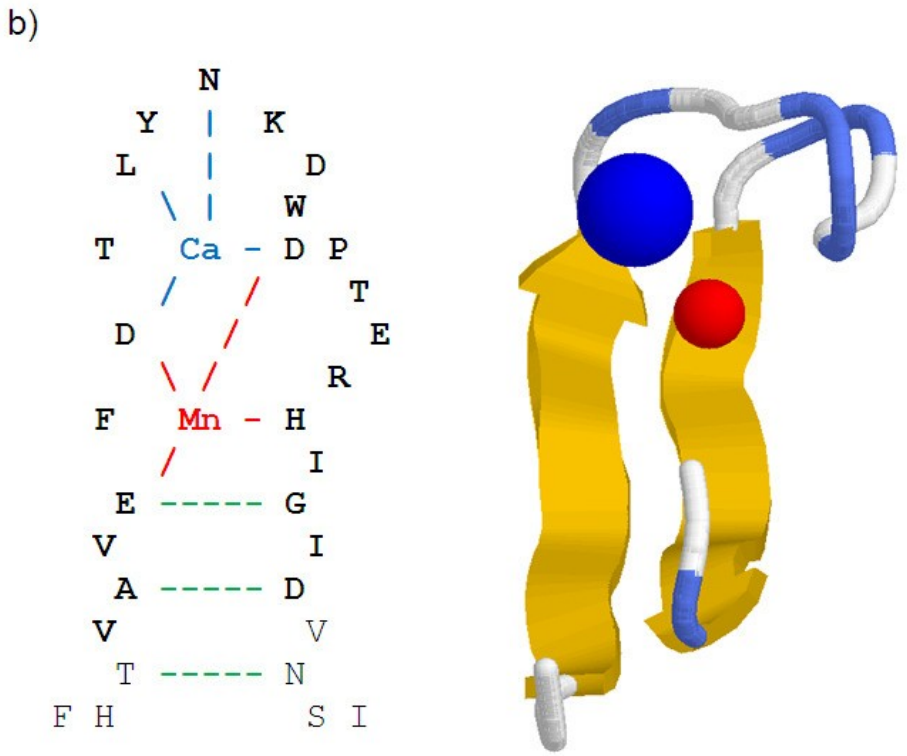
Sekwencja aminokwasowa jako sentencja języka





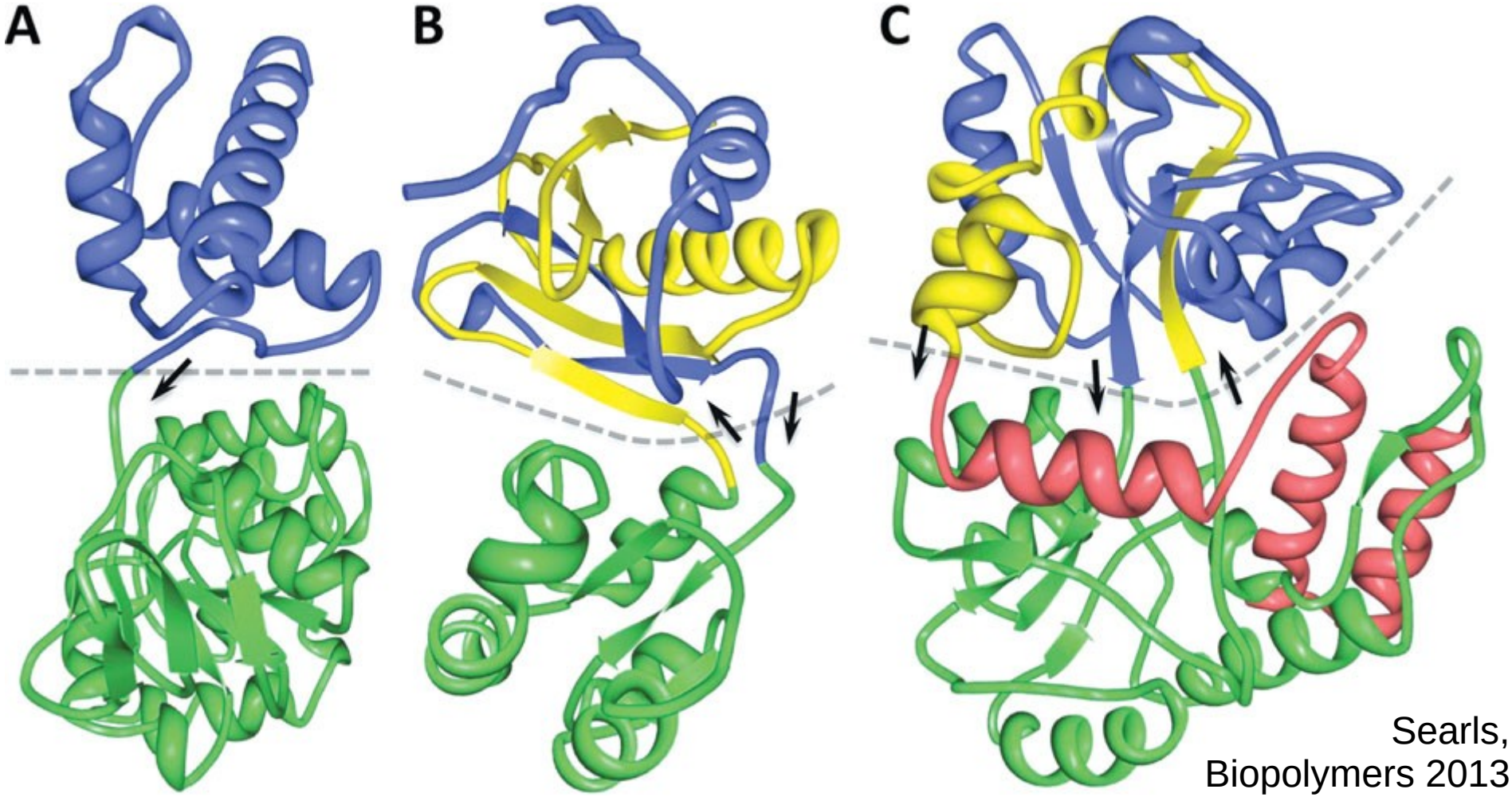
Searls, Biopolymers 2013

Sekwencja → struktura → funkcja



Dyrka & Nebel, BMC Bioinformatics 2009

Struktury syntaktyczne w białkach

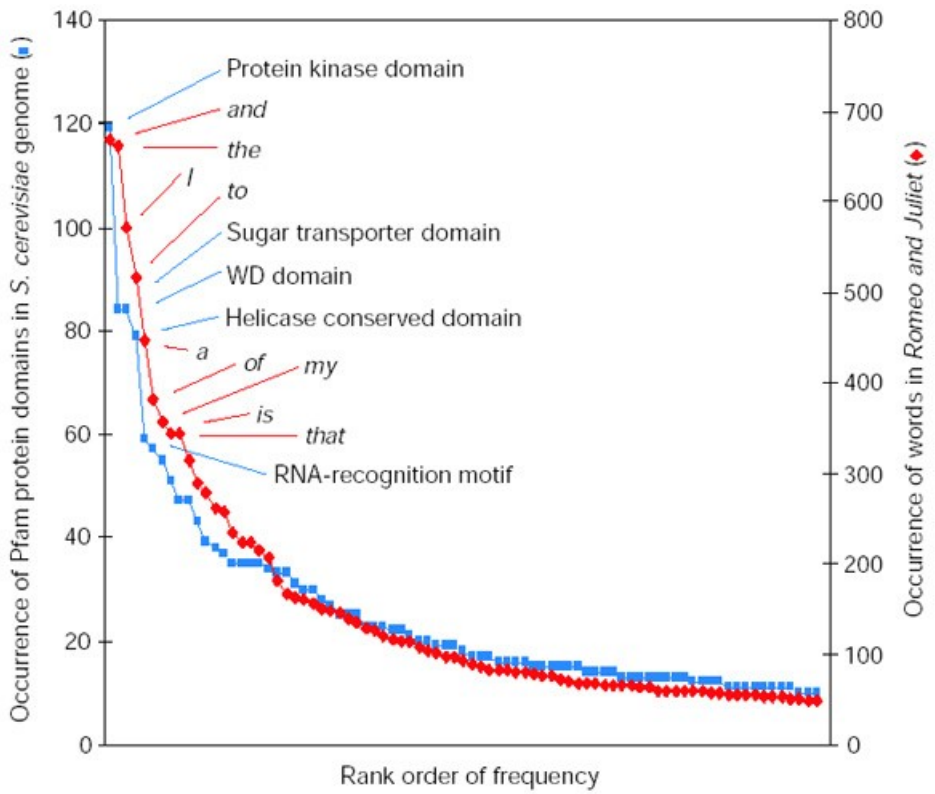


Searls,
Biopolymers 2013

regularne bezkontekstowe kontekstowe

Prawo Zipfa

$$f(k;s,N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$$



Rozkład domen spełnia prawo Zipfa
Searls, Nature 2002

Podobne wyniki:
Qian et al. JMB 2002: rodziny i “foldy” białek
Motomura et al. 2012: najczęstsze motywy

Dlaczego zachowane?

- zasada minimalnego wysiłku
  [Ferrer-i-Cancho & Sole, PNAS 2003]
- proces narodzin-i-śmierci z losowym zabijaniem jako model ewolucji białek
  [Reed & Hughes, PRE 2002]
- immanentna cecha przekształcenia długości wyrazu na jego ranking [Li 1992]
  - krytyka: [Ferrer-i-Cancho & Elvevag, PLOS ONE 2010]

Podwójna artykulacja

- Jednostki niższego poziomu
  (fonemy, które nie mają znaczenia)
  łączą się w jednostki wyższego poziomu
  (morfemy, które mają znaczenie),
  a te łączą się w zdania. [Wikipedia]

- w języku molekularnym?
  - Marcus (1973, 1998, 2004): nukleotydy, kodony, białka
  - Ji (2002):
    - pierwsza artykulacja – interakcje niekowalencyjne:
      - białka łączą się w kompleksy molekularne
    - druga artykulacja – interakcje kowalencyjne:
      - aminokwasy łączą się w białka

Przyczynki

Probabilistyczne gramatyki bezkontekstowe (PCFG)

- Dlaczego
  - co najmniej bezkontekstowe?

    Bezpośrednio reprezentują relacje dalekozasięgowe.

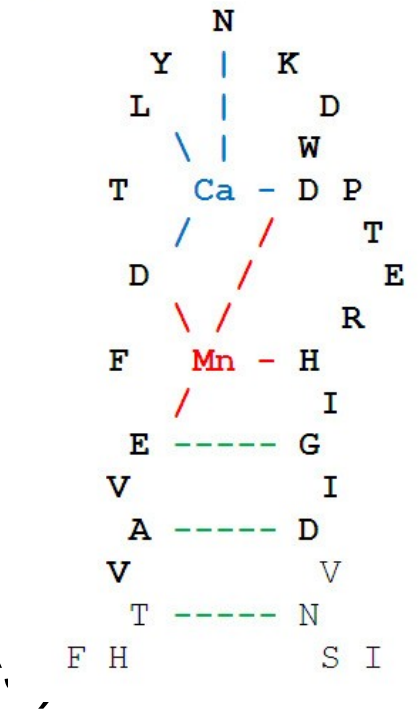
  - co najwyżej bezkontekstowe?

    *Rozsądna złożoność obliczeniowa O(n^*

  - probabilistyczne?

    Uczenie nie wymaga próbki negatywnej

b)



Dyrka & Nebel,
BMC Bioinformatics 2009

PCFG – definicje

Gramatyka probabilistyczna

$G = < \Sigma, N, P, S >$

$\sum$ – alfabet = { A, C, … Y }
N – symbole nieterminalne
P – reguły produkcji i ich p-stwa
S – symbol początkowy $\in$ N

Gramatyka kontekstowa

$P : (N \cup \Sigma)^m N (N \cup \Sigma)^n \rightarrow (N \cup \Sigma)^k$
gdzie m, n, k ≥ 0 oraz m+n+1 ≤ k

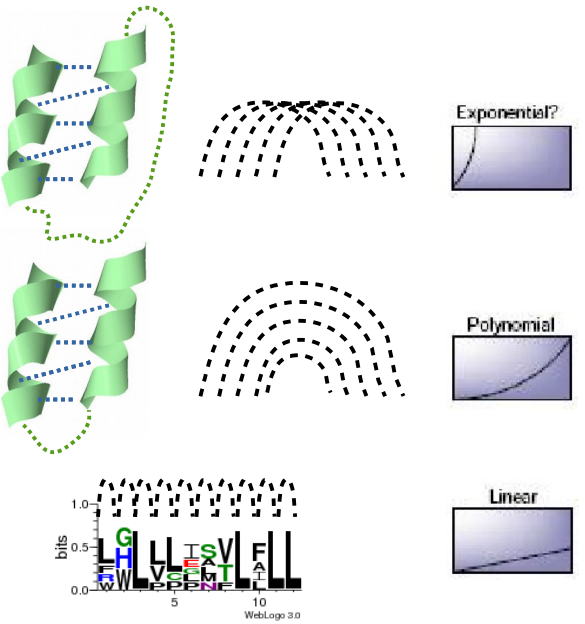
Gramatyka bezkontekstowa

$P : N \rightarrow (N \cup \Sigma)^*$

Gramatyka regularna

$P : N \rightarrow N \Sigma^* \mid \Sigma^*$
(równoważna ukrytemu modelowi Markowa (HMM))

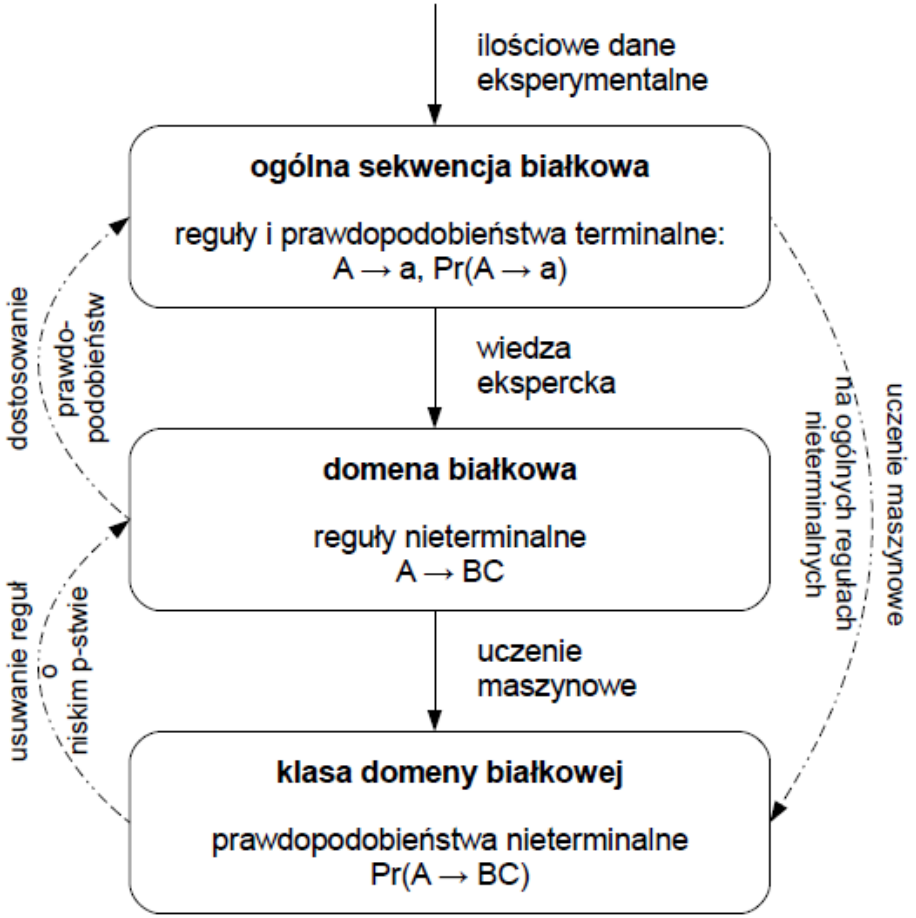
CFG w postaci normalnej CNF
$A \rightarrow BC$
$A \rightarrow a$ *reguły terminalne*
gdzie $A \in N$, $a \in \Sigma$



Gramatyczny model języka białkowego

Trzy warstwy:

1) model ogólnej sekwencji białkowej
   - przypisanie cech ilościowych aminokwasom (ilościowe dane eksperymentalne)
2) syntaktyczny model domeny białka
   - lista reguł opisujących ogólne zasady, np. parowanie helis (wiedza ekspercka)
3) probabilistyczny model klasy domeny
   - przypisanie prawdopodobieństw regułom nieterminalnym (uczenie maszynowe: algorytm genetyczny)



Gramatyczny model języka białkowego

Trzy warstwy:

1) model ogólnej sekwencji białkowej
   - przypisanie cech ilościowych aminokwasom (ilościowe dane eksperymentalne)

2) syntaktyczny model domeny białka
   - lista reguł opisujących ogólne zasady, np. parowanie helis (wiedza ekspercka)

3) probabilistyczny model klasy domeny
   - przypisanie prawdopodobieństw regułom syntaktycznym (uczenie maszynowe: algorytm genetyczny)

Alfabet:

$$a_{i=1..20} \in \Sigma_{AA} = \left\{ \begin{array}{c} A, R, N, D, C, E, Q, G, H, I \\ L, K, M, F, P, S, T, W, Y, V \end{array} \right\}.$$

Reguły terminalne – np. dla dostępności (*acc*):

$$\begin{aligned} acc\_high &\rightarrow \ldots \mid L\,(0.00) \mid K\,(0.18) \mid M\,(0.00) \mid \ldots \\ acc\_med &\rightarrow \ldots \mid L\,(0.03) \mid K\,(0.00) \mid M\,(0.05) \mid \ldots \\ acc\_low &\rightarrow \ldots \mid L\,(0.13) \mid K\,(0.00) \mid M\,(0.11) \mid \ldots \end{aligned}$$

Reguły nieterminalne (tworzenie interfejsów):

$$Inter\_ds \rightarrow \bullet\bullet\ Outer\_dd\ \bullet\bullet \mid \bullet\ Outer\_sd\ \bullet\bullet \mid \varepsilon$$
$$Inter\_ds \rightarrow \bullet\bullet\ Outer\_dd\ \bullet\bullet \mid \varepsilon$$

(Waldispuehl&Steyaert'05)

...

$$\bullet \equiv acc\_high \mid acc\_med \mid acc\_low$$

Prawdopodobieństwa reguł nieterminalnych:

$$\begin{aligned} Inter\_ds \rightarrow\ & acc\_high\ \ Outer\_sd\ \ acc\_med\ \ acc\_med\ \ \mathbf{(0.14)} \\ & acc\_med\ \ Outer\_sd\ \ acc\_med\ \ acc\_med\ \ \mathbf{(0.22)} \\ & acc\_low\ \ Outer\_sd\ \ acc\_med\ \ acc\_low\ \ \mathbf{(0.20)} \\ & acc\_low\ \ Outer\_sd\ \ acc\_med\ \ acc\_med\ \ \mathbf{(0.44)} \end{aligned}$$

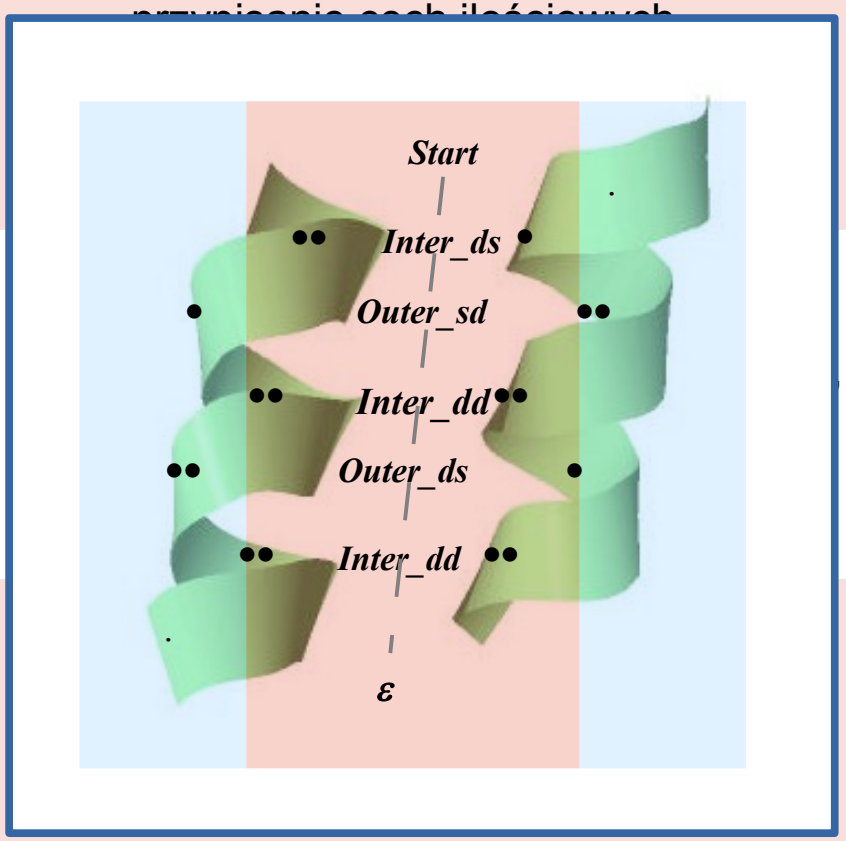
Gramatyczny model języka białkowego

Trzy warstwy:

1) model ogólnej sekwencji białkowej

przypisanie cech ilościowych

2)



3)

Alfabet:

$$a_{i=1..20} \in \Sigma_{AA} = \left\{ \begin{array}{c} A, R, N, D, C, E, Q, G, H, I \\ L, K, M, F, P, S, T, W, Y, V \end{array} \right\}.$$

Reguły terminalne – np. dla dostępności (*acc*):

$acc\_high \rightarrow \ldots \mid L\,(0.00) \mid K\,(0.18) \mid M\,(0.00) \mid \ldots$
$acc\_med \rightarrow \ldots \mid L\,(0.03) \mid K\,(0.00) \mid M\,(0.05) \mid \ldots$
$acc\_low \rightarrow \ldots \mid L\,(0.13) \mid K\,(0.00) \mid M\,(0.11) \mid \ldots$

Reguły nieterminalne (tworzenie interfejsów):

$Inter\_ds \rightarrow \bullet\bullet\ Outer\_dd\ \bullet\bullet \mid \bullet\ Outer\_sd\ \bullet\bullet \mid \varepsilon$
$Inter\_ds \rightarrow \bullet\bullet\ Outer\_dd\ \bullet\bullet \mid \varepsilon$ (Waldispuehl&Steyaert'05)
...
$\bullet \equiv acc\_high \mid acc\_med \mid acc\_low$

Prawdopodobieństwa reguł nieterminalnych:

$Inter\_ds \rightarrow acc\_high\ Outer\_sd\ acc\_med\ acc\_med$ **(0.14)**
$acc\_med\ Outer\_sd\ acc\_med\ acc\_med$ **(0.22)**
$acc\_low\ Outer\_sd\ acc\_med\ acc\_low$ **(0.20)**
$acc\_low\ Outer\_sd\ acc\_med\ acc\_med$ **(0.44)**

Gramatyczny model języka białkowego

Trzy warstwy:

1) model ogólnej sekwencji białkowej
   - przypisanie cech ilościowych aminokwasom (ilościowe dane eksperymentalne)

2) syntaktyczny model domeny białka
   - lista reguł opisujących ogólne zasady, np. parowanie helis (wiedza ekspercka)

3) probabilistyczny model klasy domeny
   - przypisanie prawdopodobieństw regułom syntaktycznym (uczenie maszynowe: algorytm genetyczny)

Alfabet:

$$a_{i=1..20} \in \Sigma_{AA} = \left\{ \begin{array}{c} A, R, N, D, C, E, Q, G, H, I \\ L, K, M, F, P, S, T, W, Y, V \end{array} \right\}.$$

Reguły terminalne – np. dla dostępności (*acc*):

$$\begin{aligned} acc\_high &\rightarrow \ldots \mid L\,(0.00) \mid K\,(0.18) \mid M\,(0.00) \mid \ldots \\ acc\_med &\rightarrow \ldots \mid L\,(0.03) \mid K\,(0.00) \mid M\,(0.05) \mid \ldots \\ acc\_low &\rightarrow \ldots \mid L\,(0.13) \mid K\,(0.00) \mid M\,(0.11) \mid \ldots \end{aligned}$$

Reguły nieterminalne (tworzenie interfejsów):

$$Inter\_ds \rightarrow \bullet\bullet\ Outer\_dd\ \bullet\bullet \mid \bullet\ Outer\_sd\ \bullet\bullet \mid \varepsilon$$
$$Inter\_ds \rightarrow \bullet\bullet\ Outer\_dd\ \bullet\bullet \mid \varepsilon$$

(Waldispuehl&Steyaert'05)

...

$$\bullet \equiv acc\_high \mid acc\_med \mid acc\_low$$

Prawdopodobieństwa reguł nieterminalnych:

$$\begin{aligned} Inter\_ds \rightarrow\ & acc\_high\ \ Outer\_sd\ \ acc\_med\ \ acc\_med\ \ \mathbf{(0.14)} \\ & acc\_med\ \ Outer\_sd\ \ acc\_med\ \ acc\_med\ \ \mathbf{(0.22)} \\ & acc\_low\ \ Outer\_sd\ \ acc\_med\ \ acc\_low\ \ \mathbf{(0.20)} \\ & acc\_low\ \ Outer\_sd\ \ acc\_med\ \ acc\_med\ \ \mathbf{(0.44)} \end{aligned}$$

Dotychczasowe zastosowania modelu (detekcja lub klasyfikacja)

Zadanie	**Wynik**	**Porównanie**
Detekcja miejsc wiążących ligandy *BMC Bioinformatics 2009*	F1 = 0.80-1.00	HMMER2: F1 = 0.79-1.00
Klasyfikacja par helis transmembranowych *Alg Mol Biol 2013*	AUCROC ~ 0.60 (najlepsze do 0.67)	HMMER2,3: AUCROC ~ 0.58 (najlepsze do 0.64)
Rozpoznawanie 6-merów amyloidogennych *F.Thirion, praca mgr PWr, 2013*	AUCROC = 0.87	Waltz (Maurer-Stroh'10): AUCROC ~ 0.90

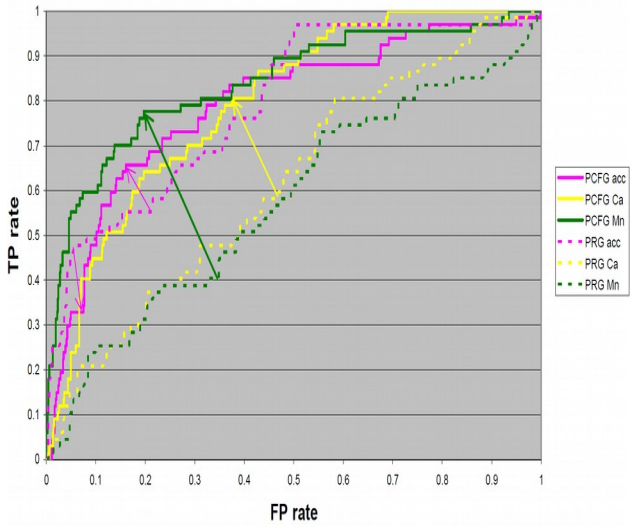
Nieopublikowane:

- detekcja miejsc wiążących RNA (bez powodzenia, problem rozwiązany później przy użyciu SVM przez Spriggs et al. [Bioinformatics 2009])
- detekcja domen tworzących priony (AUCROC=0.77)

Czy poziom bezkontekstowy coś wnosi *w praktyce*?

Detekcja miejsc wiążących
(dł. sekwencji ok. 20aa):

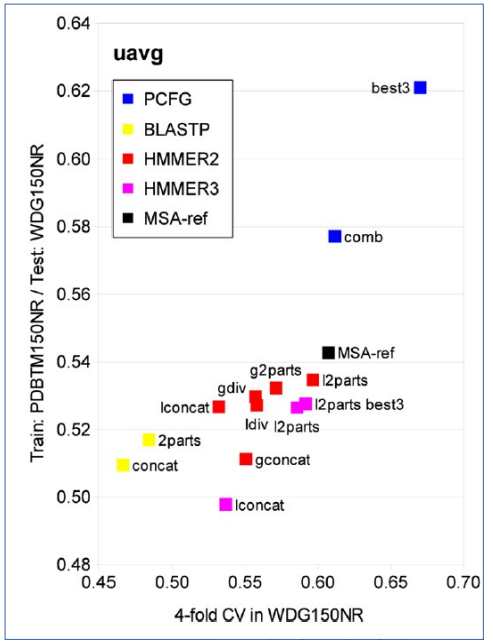
***Tak**, w przypadku niektórych cech*



W.Dyrka, praca MSc Kingston U '07

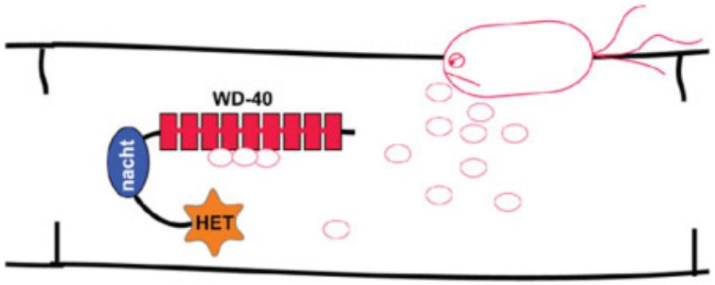
Klasyfikacja par helis
(dł. sekwencji ok. 20aa):

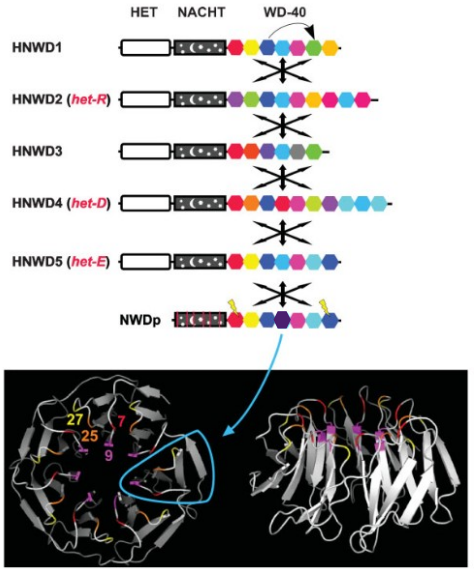
***Tak**, <u>ale</u> wynika to z efektywnej reprezentacji periodyczności helis (cecha regularna), niż z bezpośredniej reprezentacji interakcji dalekozasięgowych*



W.Dyrka et al. Alg Mol Biol 2013

Język receptorów patogenów typu NLR



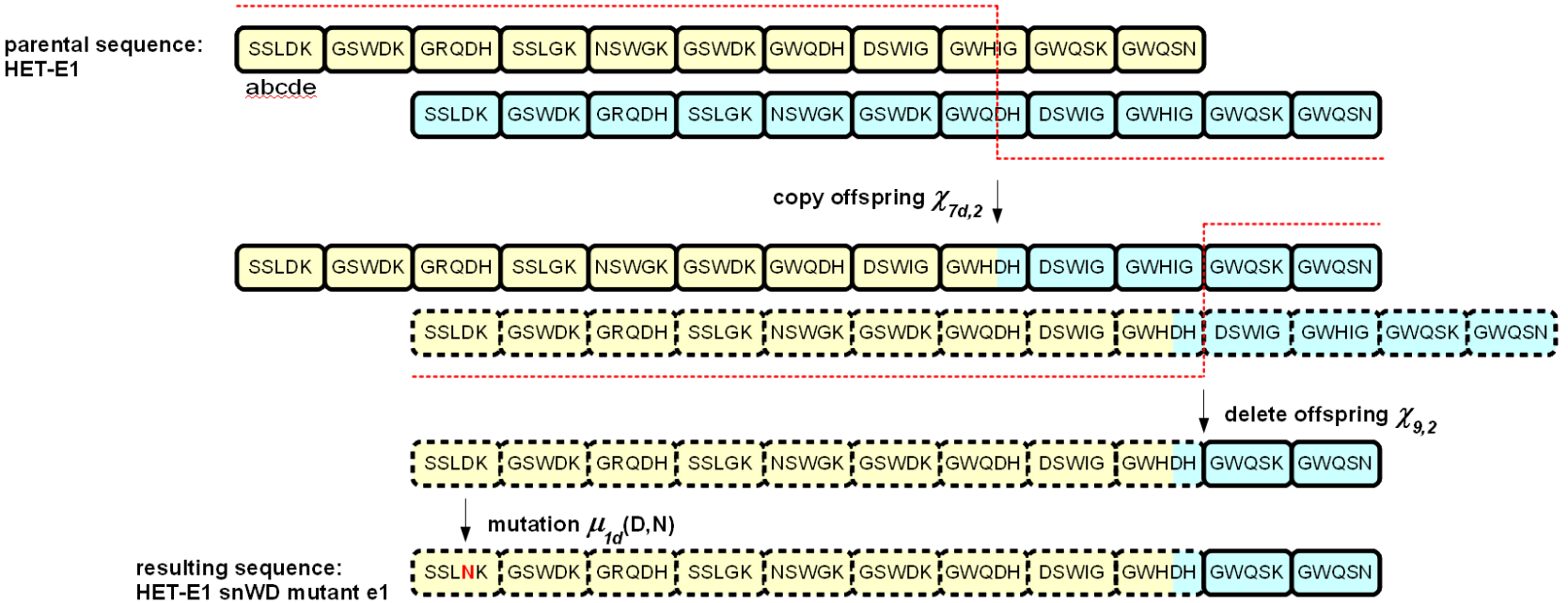


WD repeats	Amino acid positions 7	9	25	27
het-R-1	S	Y	R	V
het-R-2	S	Y	D	V
het-R-4	S	S	R	V
het-R-5	S	S	D	V
het-R-6	S	S	R	I
het-R-7	W	Y	D	V
het-R-9	L	Y	D	V
het-R-10	S	H	D	V
het-R-11	S	S	C	V
het-D$_{\text{het-r hetV}}$-6	W	W	A	S
het-D$_{\text{het-r hetV}}$-5	S	N	T	R
NWD1$_{\text{het-r hetV}}$-2	L	R	M	K
NWD1$_{\text{het-r hetV}}$-6	W	S	G	E
NWD1$_{\text{het-r hetV}}$-7	L	R	G	E
NWDp3$_{\text{het-R het-v}}$-3	L	L	N	G
NWDp3$_{\text{het-R het-v}}$-1	S	R	N	R
NWD2$_{\text{het-r hetV}}$ -1	W	W	G	Y
NWD2$_{\text{het-r hetV}}$ -3	S	L	S	N
NWD2$_{\text{het-r hetV}}$ -5	L	L	H	N
NWDp1$_{\text{het-r hetV}}$ -5	S	W	L	K
NWDp1$_{\text{het-r hetV}}$-7	W	Q	H	K
NWDp1$_{\text{het-r hetV}}$-8	S	Q	H	M
NWDp1$_{\text{het-r hetV}}$-10	X	Q	H	M

Paoletti & Saupe Bioessays 2009, Chevanne et al. BMC Evol Biol 2010

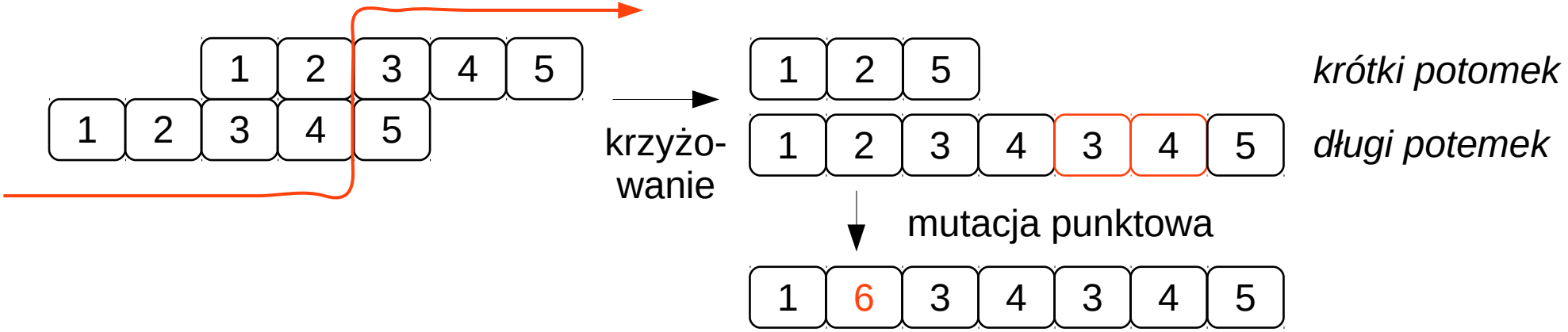
Model rearanżacji powtórzeń

nierówne krzyżowanie + mutacja punktowa



system przepisywania wyrazów

Matematyczny model rearanżacji



Stochastyczny system przepisywania wyrazów z ograniczeniami $\langle \sum, R, P, Q \rangle$

- $\sum$ – alfabet: 20 typów aminokwasów
- R – zbiór reguł przepisywania $u \rightarrow v \in \sum^*$

  modelowanie krzyżowania i mutacji
- P – zbiór prawdopodobieństw reguł

} generacja

- Q – zbiór ograniczeń

→ selekcja

  - np. w formie gramatyki probabilistycznej

Własności modelu

- Model zbiega do jednego stacjonarnego rozkładu sekwencji

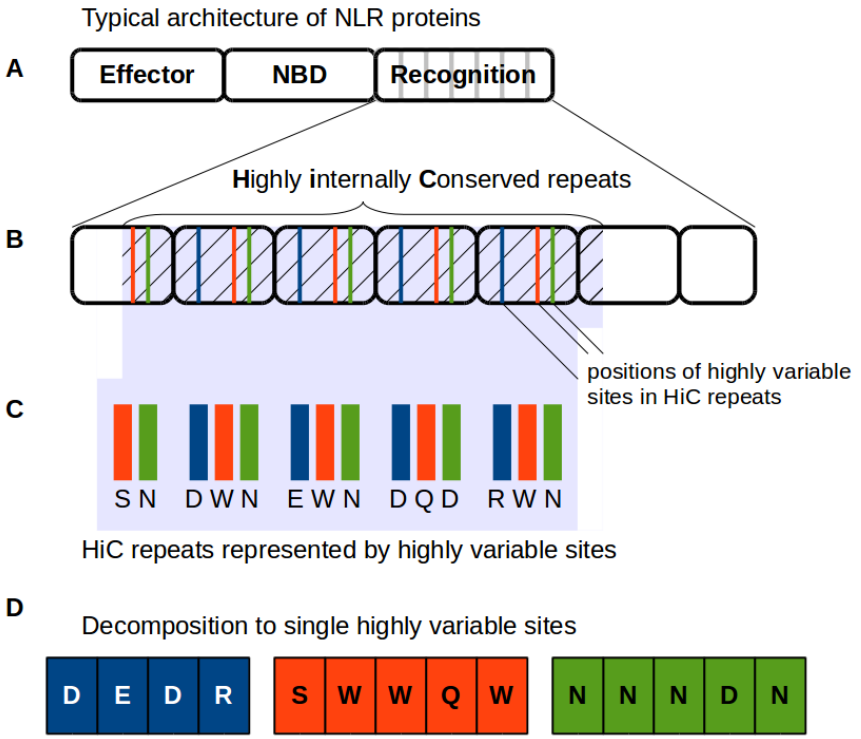
  => także rozkładu liczby powtórzeń i składu aminokwasów

  dla prostych, realistycznych parametrów i ograniczeń

- Wniosek:

  Różnice pomiędzy rozkładem generowanym przez model oraz obserwowanych w rzeczywistości można interpretować jako efekt presji zewnętrznej.

Wstępne badania



Porównanie składu aminokwasów dane biologiczne vs model

- prolina, walina
- kwas asparaginowy i glutaminowy, lizyna, asparagina, tyrozyna
- tryptofan

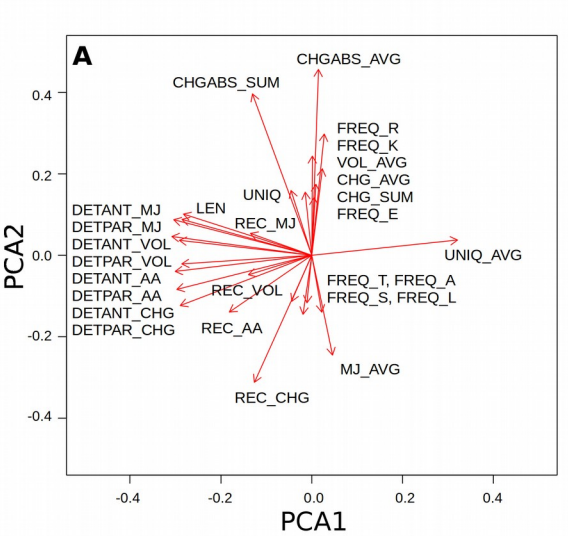
Zgodne z rolą paratopu rozpoznającego patogeny

Analiza presji zewnętrznej

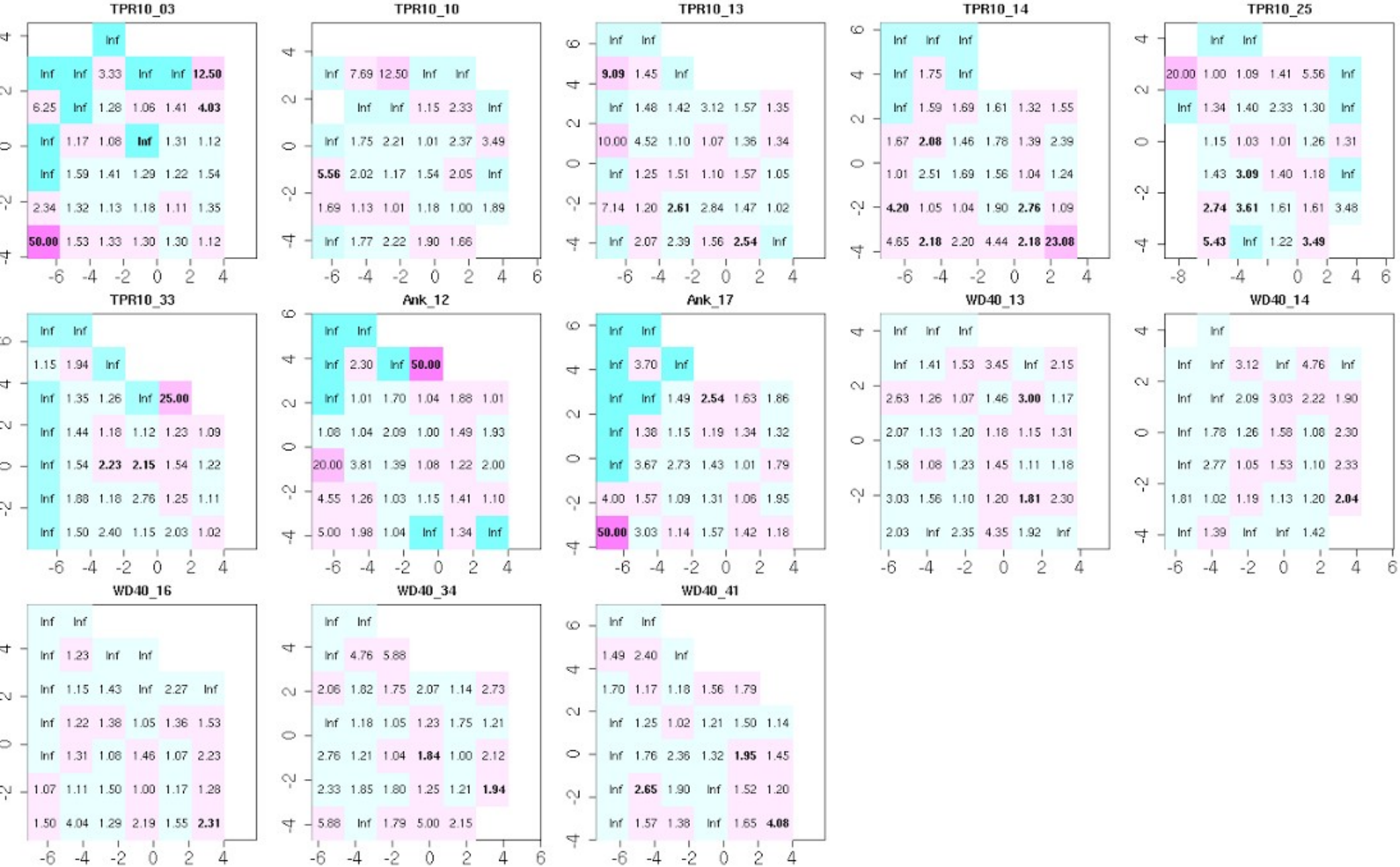
- Sprawdzenie hipotezy biologicznej:
  - hipoteza dot. presji jako ograniczenie na system
  - porównanie przestrzeni rozwiązań modelu z danymi biologicznymi

- Generowanie hipotez biologicznych
  - analiza różnic pomiędzy przestrzenią rozwiązań modelu a danymi biologicznymi – poszukiwanie wzorców

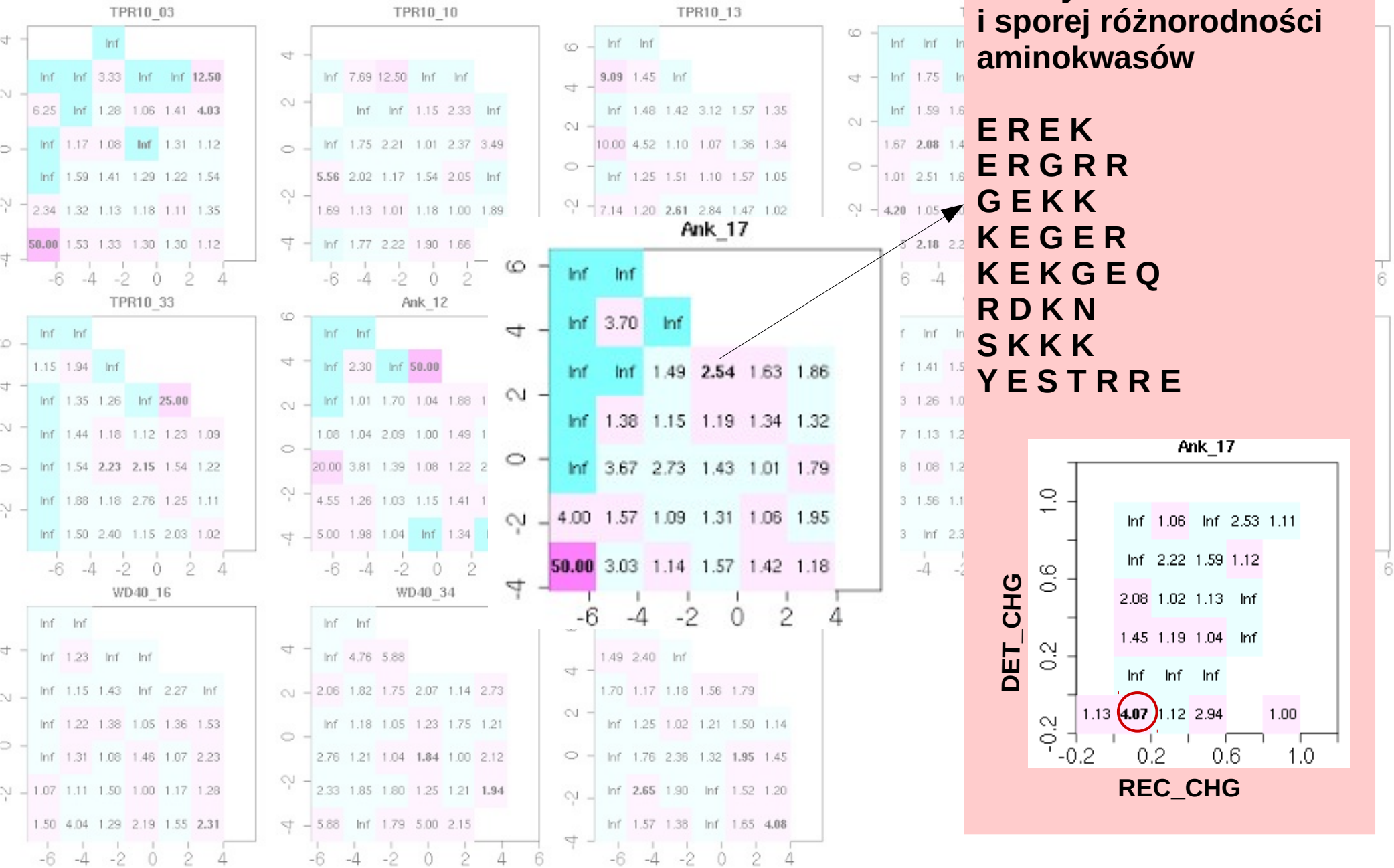
- Definicja przestrzeni rozwiązań
  - wektor cech ilościowych



Przykład: presja na skład aminokwasów



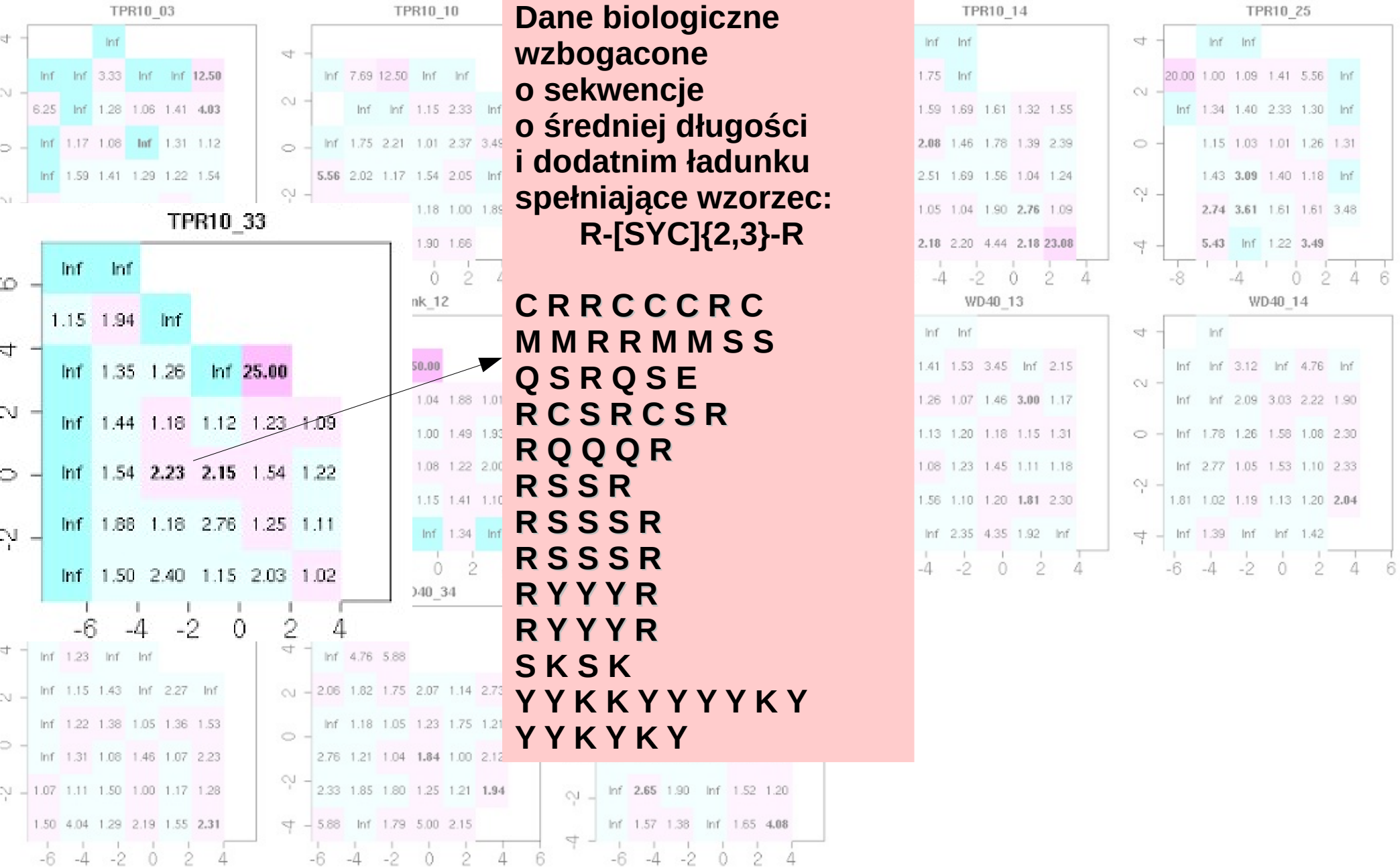
Poszukiwanie wzorców wynikających z presji zewnętrznej



Dane biologiczne istotnie wzbogacone o sekwencje o dużym ładunku i sporej różnorodności aminokwasów

E R E K
E R G R R
G E K K
K E G E R
K E K G E Q
R D K N
S K K K
Y E S T R R E

Poszukiwanie wzorców wynikających z presji zewnętrznej



Dane biologiczne wzbogacone o sekwencje o średniej długości i dodatnim ładunku spełniające wzorzec:

R-[SYC]{2,3}-R

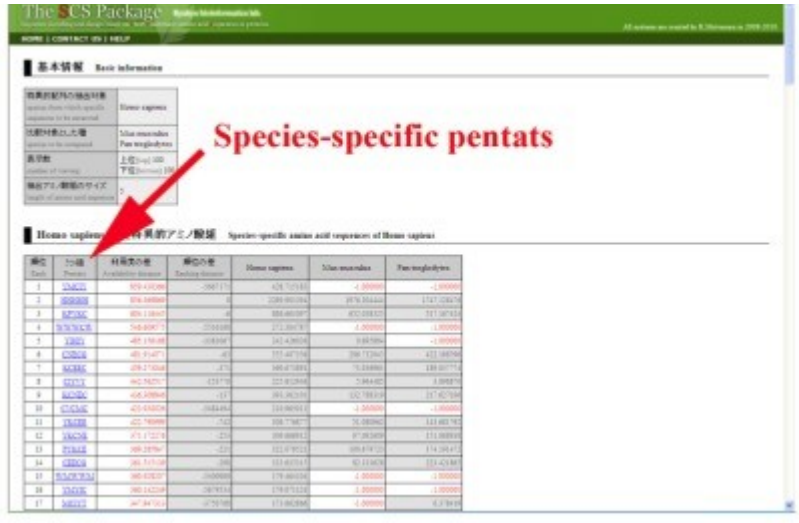
C R R C C C R C
M M R R M M S S
Q S R Q S E
R C S R C S R
R Q Q Q R
R S S R
R S S S R
R S S S R
R Y Y Y R
R Y Y Y R
S K S K
Y Y K K Y Y Y Y K Y
Y Y K Y K Y

Perspektywy

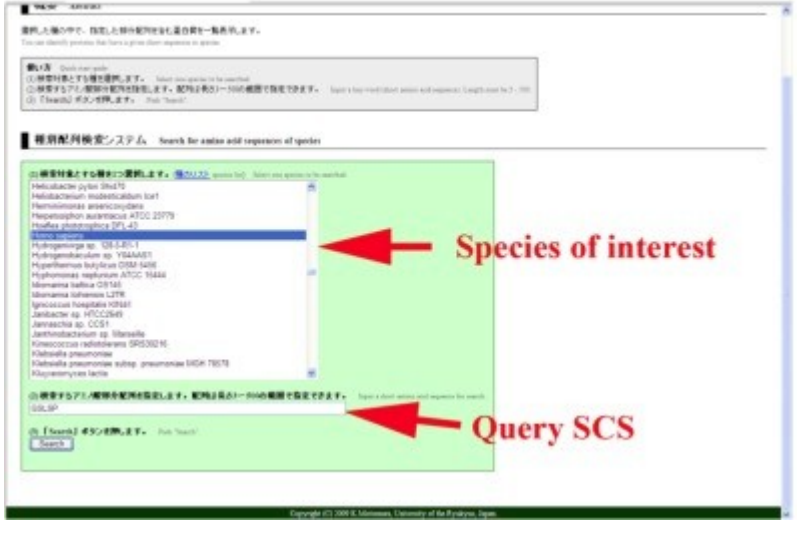
Analiza ilościowa: n-gramy

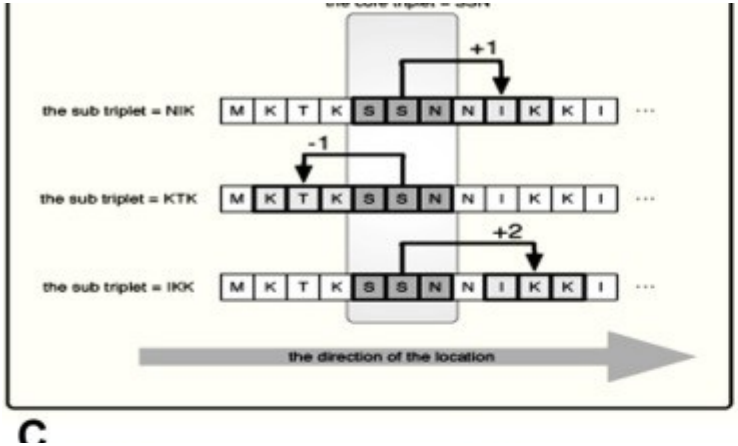
SCS package [Motomura et al. 2013]

C

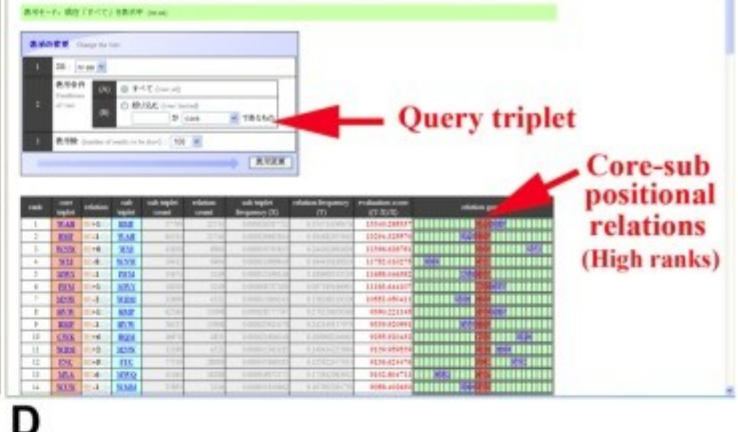


D





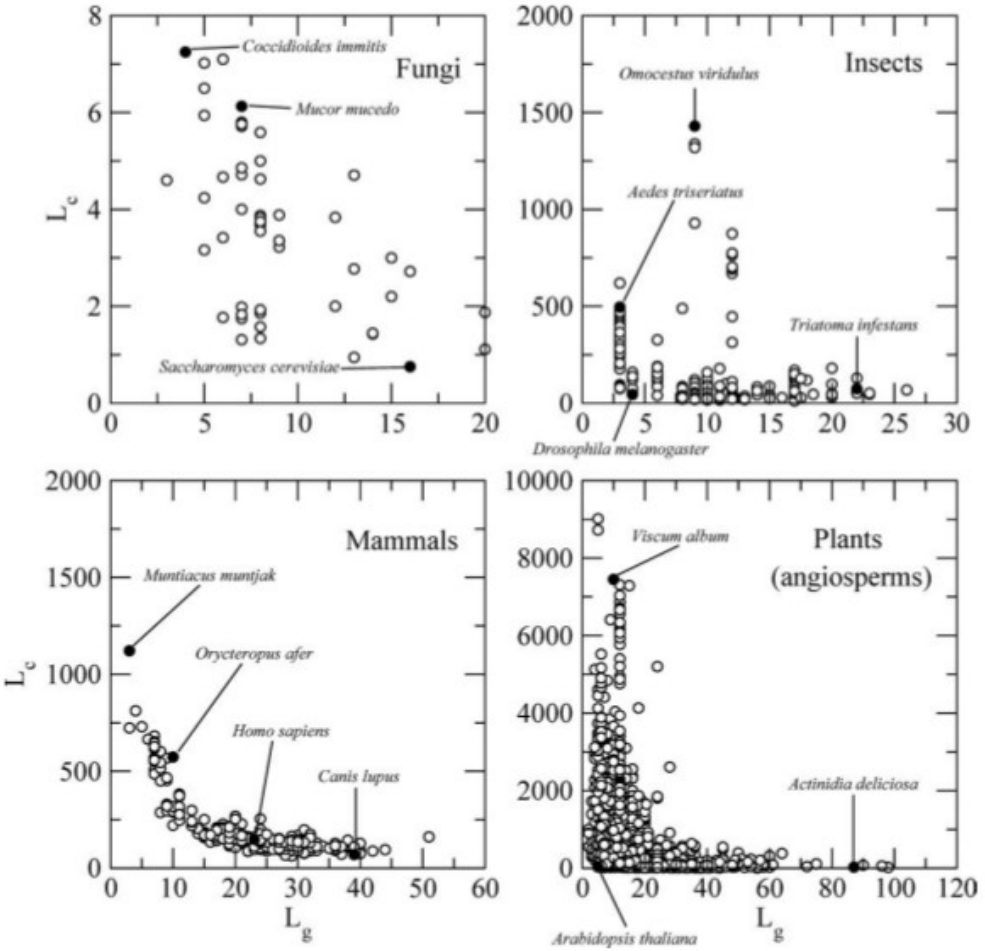
C



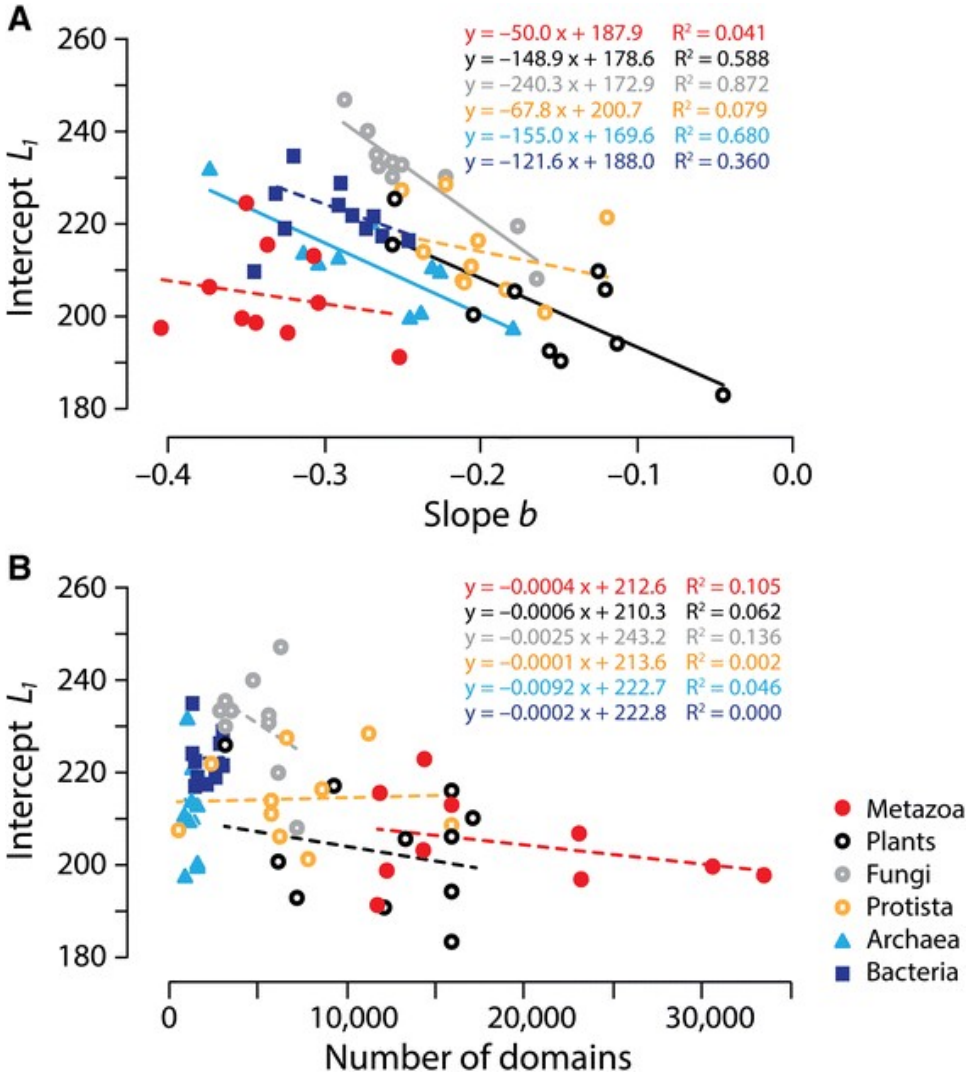
D

Motomura et al.
Comput Struct Biotechnol J. 2013

Analiza ilościowa: prawo Menzeratha



Ferer-i-Cancho & Forns, Complexity 2009;



Shahzad et al. BMC Syst Biol 2015

Analiza ilościowa sekwencji białkowych: ewolucja w czasie i przestrzeni

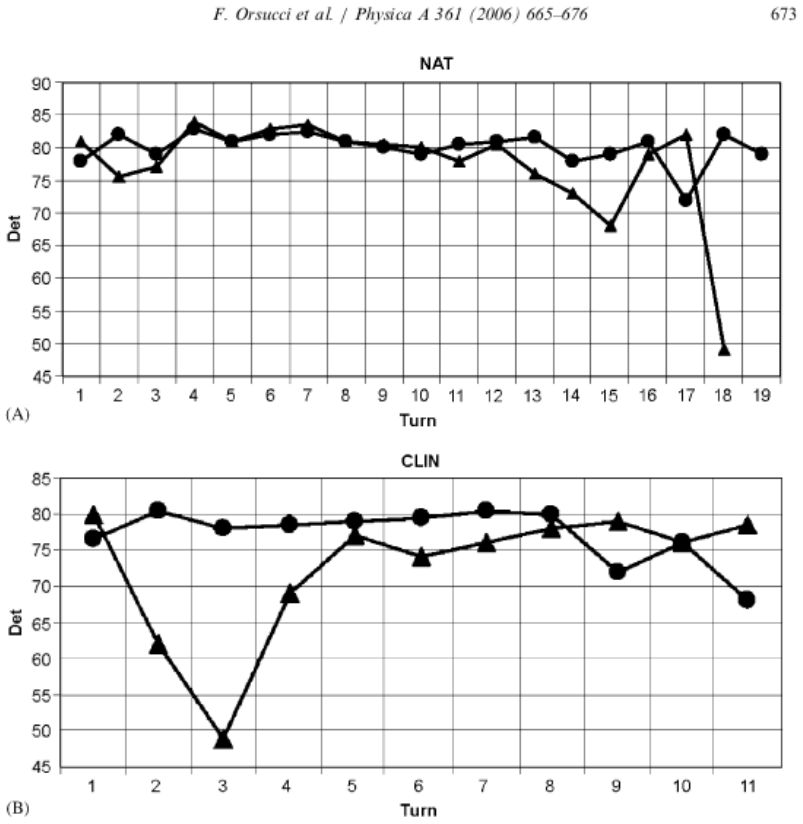
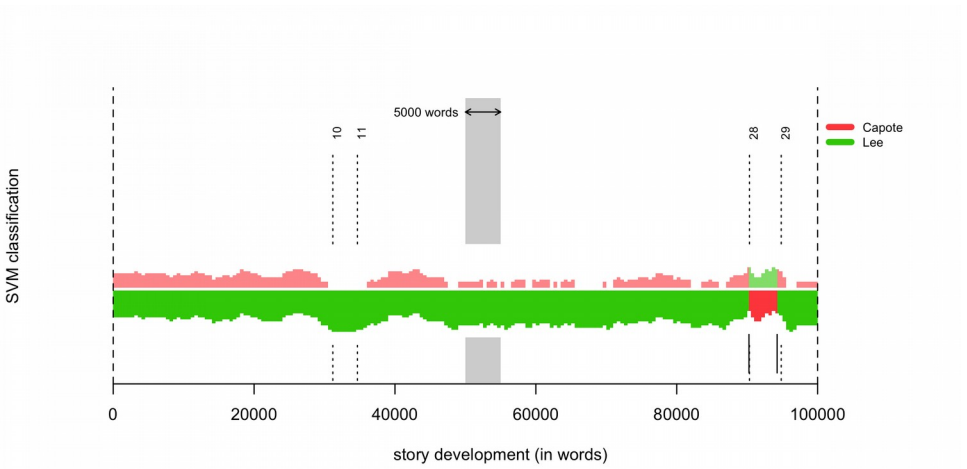


Fig. 3. A and B both represent a semiotic interaction, each dot being a turn in conversation. In A red and blue dots are for friend J or K; in B red dots are for patient P and blue dots for therapist T. The abscissa report the subsequent order of the dialogue (turn), and the ordinate the amount of determinism of the speeches.

Orsucci et al. Physica A 2006



Eder & Rybicki, 2015
https://sites.google.com/site/computationalstylistics/projects/lee_vs_capote

BLAST → HMMER → PCFG? ;-)

The idea that drove me to spend the last four years rewriting HMMER from scratch (again) is pretty simple. BLAST has been the workhorse of computational molecular biology for almost twenty years. The theoretical foundation of biological sequence analysis has been greatly improved since BLAST was written, because of the advent of probabilistic modeling methods such as hidden Markov models. (...) But people still use BLAST; I still use BLAST. Why is that? Theoretical advances in a computational science are all well and good, but someone has to write a practical implementation if these advances are going to be widely used. BLAST is a damn fine implementation; fast, robust, and beautifully supported by NCBI. There are reasonable implementations of HMM methods, including HMMER2 and the UCSC SAM package, but mostly because they're so much slower than BLAST, they have stayed in a different niche — both in how they're used (for profile searches in protein domain databases such as Pfam and SMART) and in terms of mind share. HMMs are still seen as a black box, not as an improved statistical foundation for all of sequence analysis. (…)

So: all this gorgeous theory was languishing, and it was starting to piss me off.

So: HMMER3's ever-so-modest goal is to compete with BLAST.

Sean Eddy's Cryptogenomicon blog: https://cryptogenomicon.org/2008/12/14/hmmer-reloaded/

Grammar-based Classifier System [Unold 2005]

- indukcja gramatyki z pozytywnych i negatywnych przykładów
- algorytm genetyczny pracuje na jednej gramatyce (*podejście Michigan*)

- Dotychczasowe zastosowania bioinformatyczne:
  - rozpoznawanie sekwencji promotorowych *E.coli* (dokładność: 78%) GCS w wersji podstawowej [O.Unold, LNCS 2007]
  - klasyfikacja heksapeptydów amyloidogennych (dokładność: 71%) GCS w wersji rozmytej [O.Unold, Proc ICGI 2012]

Uczenie języków k,l-lokalnie podstawialnych [Coste et al. 2014]

Definition 1 *(Clark and Eyraud, 2007) A language L is* substitutable *iff for any* $x_1, y_1, z_1, x_2, y_2, z_2 \in \Sigma^*, y_1, y_2 \neq \lambda$:

$$x_1 y_1 z_1 \in L \wedge x_1 y_2 z_1 \in L \Rightarrow (x_2 y_1 z_2 \in L \Leftrightarrow x_2 y_2 z_2 \in L).$$

Definition 2 *(Yoshinaka, 2008) A language L is* k, l-substitutable *iff for any* $x_1, y_1, z_1, x_2, y_2, z_2 \in \Sigma^*, u \in \Sigma^k, v \in \Sigma^l, uy_1v, uy_2v \neq \lambda$:

$$x_1 u y_1 v z_1 \in L \wedge x_1 u y_2 v z_1 \in L \Rightarrow (x_2 u y_1 v z_2 \in L \Leftrightarrow x_2 u y_2 v z_2 \in L).$$

Definition 3 *(Coste et al., 2012) A language L is* k, l-local substitutable *iff for any* $x_1, y_1, z_1, x_2, y_2, z_2, x_3, z_3 \in \Sigma^*, u \in \Sigma^k, v \in \Sigma^l, uy_1v, uy_2v \neq \lambda$:

$$x_1 u y_1 v z_1 \in L \wedge x_3 u y_2 v z_3 \in L \Rightarrow (x_2 y_1 z_2 \in L \Leftrightarrow x_2 y_2 z_2 \in L).$$

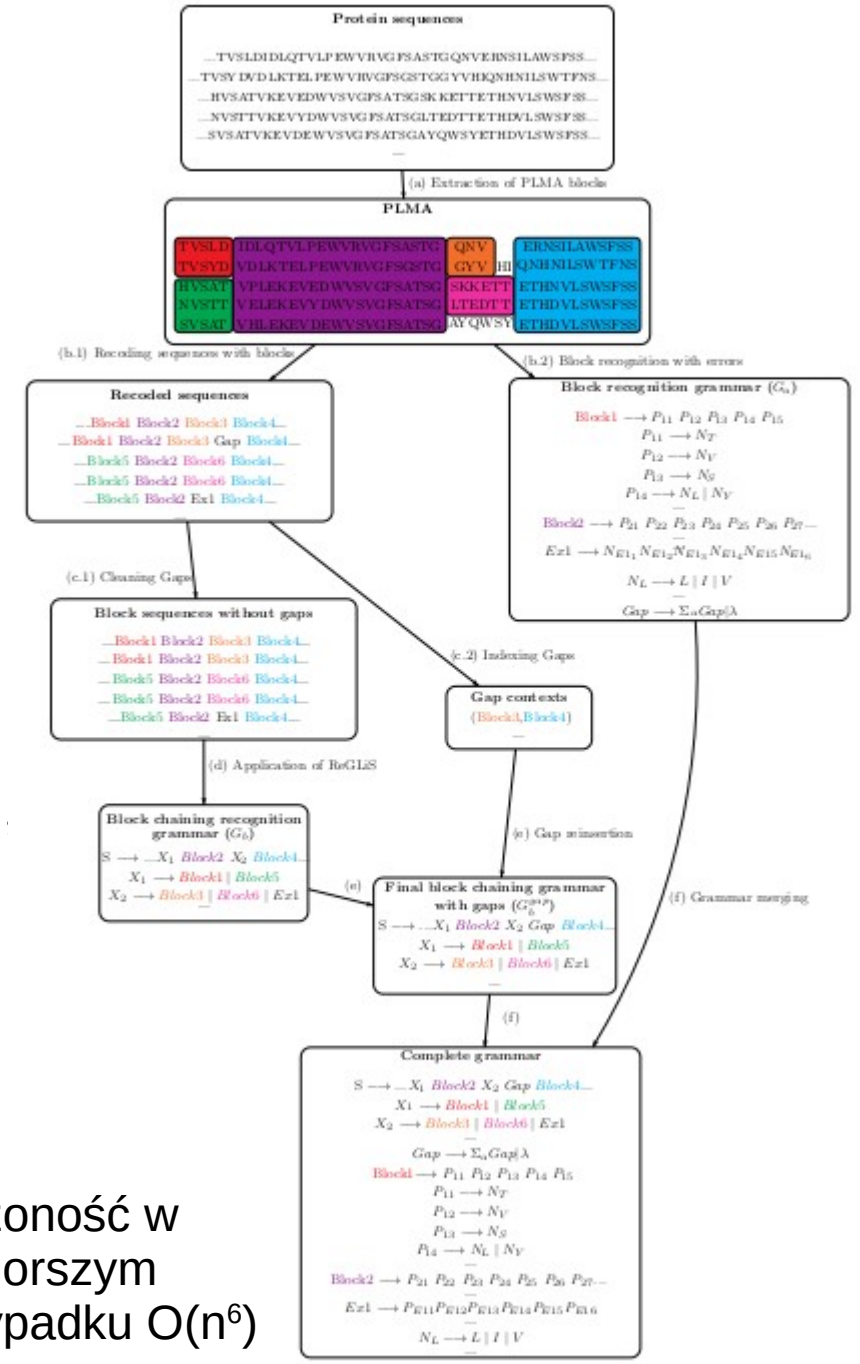
Definition 4 *(Coste et al., 2012) A language L is* k, l-local contextually substitutable *iff for any* $x_1, y_1, z_1, x_2, y_2, z_2, x_3, z_3 \in \Sigma^*, u \in \Sigma^k, v \in \Sigma^l, uy_1v, uy_2v \neq \lambda$:

$$x_1 u y_1 v z_1 \in L \wedge x_3 u y_2 v z_3 \in L \Rightarrow (x_2 u y_1 v z_2 \in L \Leftrightarrow x_2 u y_2 v z_2 \in L).$$

ReGLiS: uczenie k,l-lokalnie podstawialnych języków z pozytywnych przykładów

$\mathbf{K}$ ={"Major General was here yesterday morning.",
"Major General went here yesterday morning.",
"Major General will be there tomorrow morning.",
"He will be gone tomorrow evening."}

$$S \to X_3 X_4 X_2$$
$$X_1 \to was \mid went$$
$$X_2 \to morning \mid evening$$
$$X_3 \to He \mid Major\ General$$
$$X_4 \to will\ be\ X_5\ tomorrow \mid X_1\ here\ yesterday$$
$$X_5 \to there \mid gone$$



Złożoność w najgorszym przypadku O(n^6)

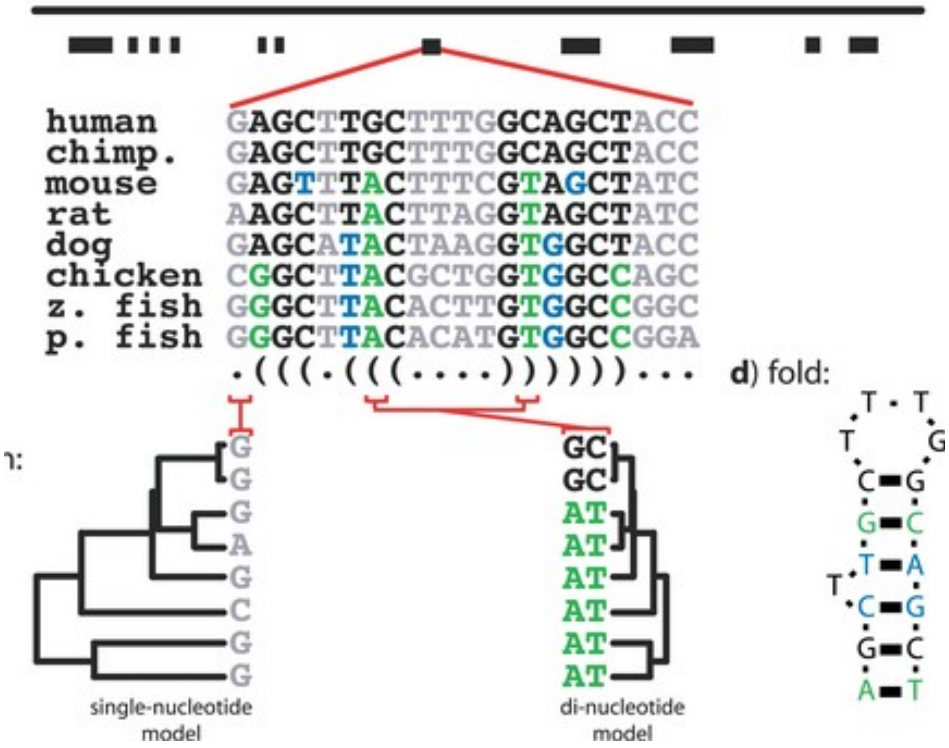
F.Coste et al. Proc ICGI 2014

G.Garet, praca PhD, Uni Rennes 1 (2015)

Filogramatyki [Knudsen & Hein 1999]

$$
\begin{aligned}
\text{begin\_non-structrual} &\rightarrow \text{unpaired} \\
\text{unpaired} &\rightarrow \text{unpaired } x \mid \text{end } x
\end{aligned}
$$

$$
\begin{aligned}
\text{begin\_structrual} &\rightarrow \text{stem\_pair} \\
\text{stem\_pair} &\rightarrow x_l \text{ stem\_pair } x_r \mid x_l \text{ bifurcation } x_r \\
\text{bifurcation} &\rightarrow \text{emit intermediate} \\
\text{emit} &\rightarrow \text{loop\_\&\_bulge} \mid \text{stem\_pair} \\
\text{loop\_\&\_bulge} &\rightarrow \text{end } x \\
\text{intermediate} &\rightarrow \text{bifurcation} \mid \text{emit}
\end{aligned}
$$



Uczenie modelu: EM

Implementacje:

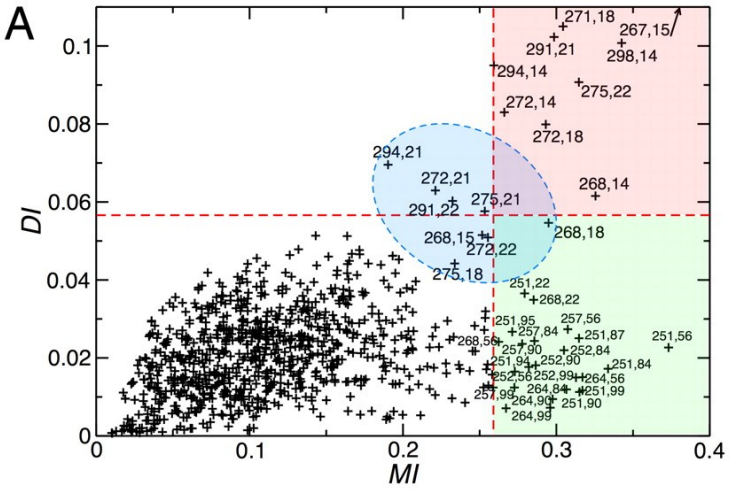
- Pfold (RNA) [Knudsen & Hein NAR'03]
- EvoFold (RNA) [Pedersen et al. PLOS Comp Biol'06]
- XRATE [Klosterman et al. BMC Bioinf'06]

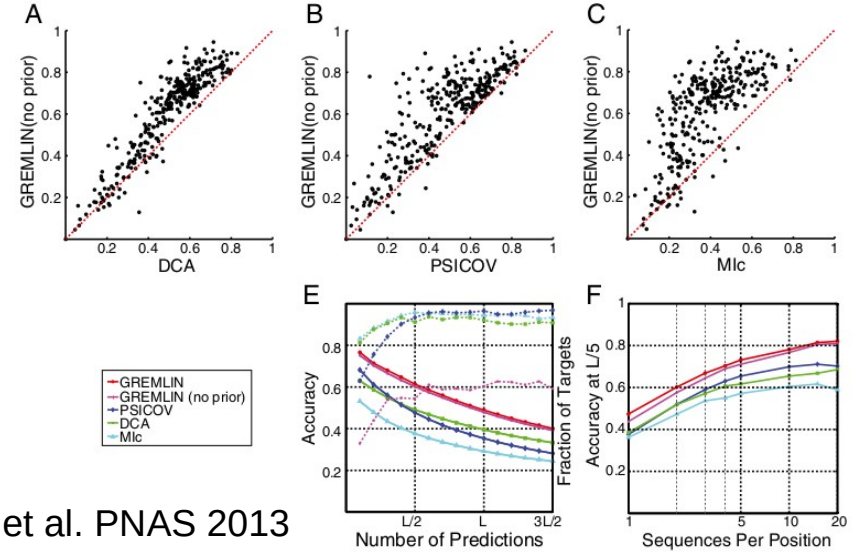
Pedersen et al. PLOS Comp Biol 2006

Direct Coupling Analysis [Weigt et al. 2009]

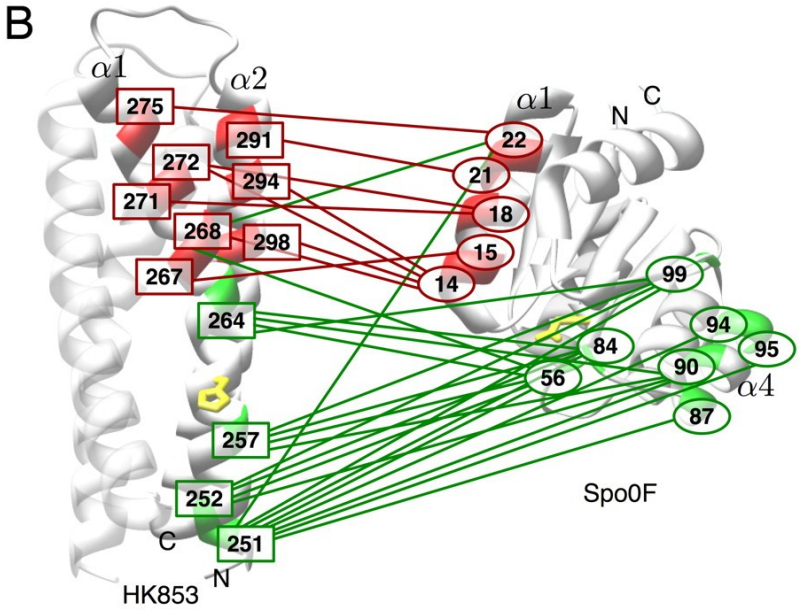
$$MI_{ij}^{(\mathrm{raw})} = \sum_{A_i,A_j=1}^{Q} f_{ij}(A_i,A_j)\ln\frac{f_{ij}(A_i,A_j)}{f_i(A_i)f_j(A_j)}.$$

$$DI_{ij} = \sum_{A_i,A_j} P_{ij}^{(\mathrm{dir})}(A_i,A_j)\ln\frac{P_{ij}^{(\mathrm{dir})}(A_i,A_j)}{f_i(A_i)f_j(A_j)}.$$

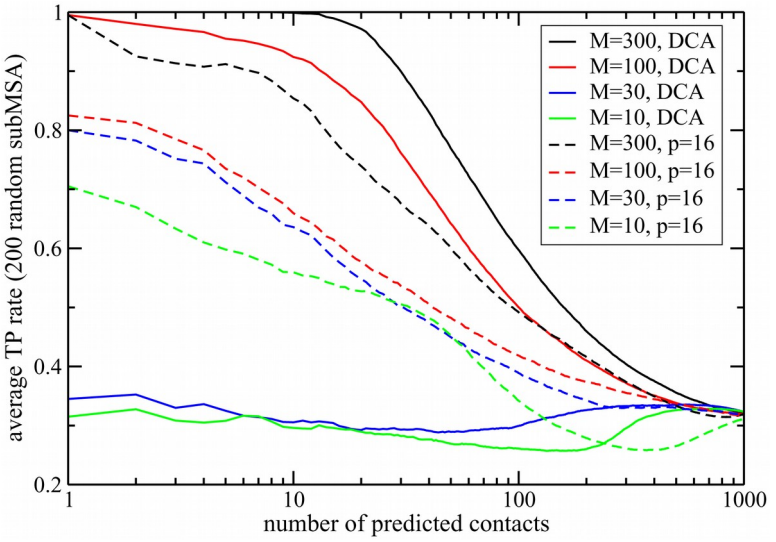




Kamisetty et al. PNAS 2013



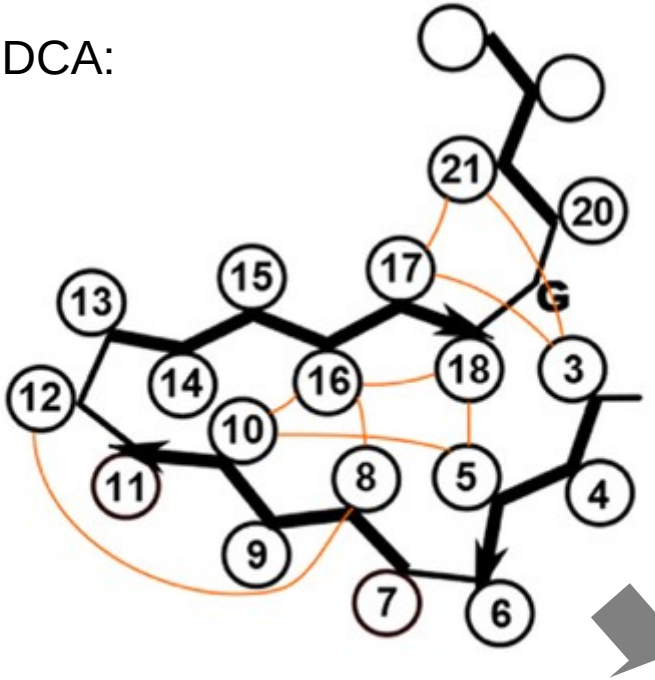
Weigt, White et al. PNAS 2009

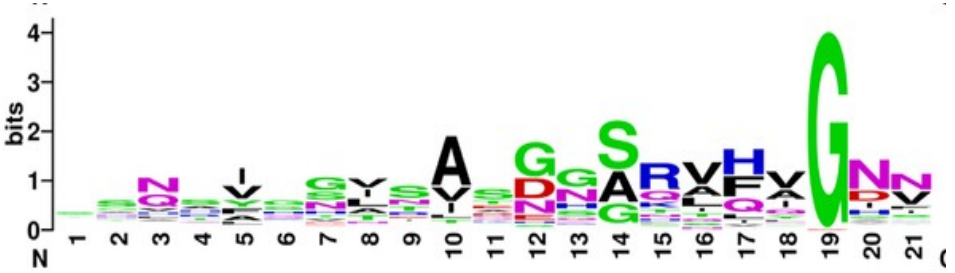


Cocco et al. PLOS Comp Biol 2013

PCFG + DCA

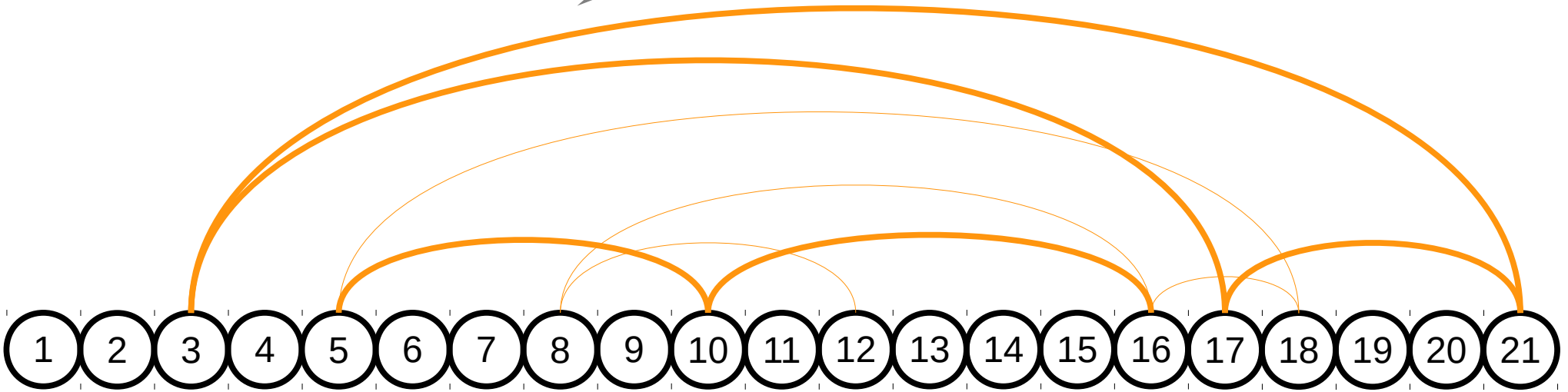
DCA:





- - ((- (- - - -)(- - - - -))(- - -))

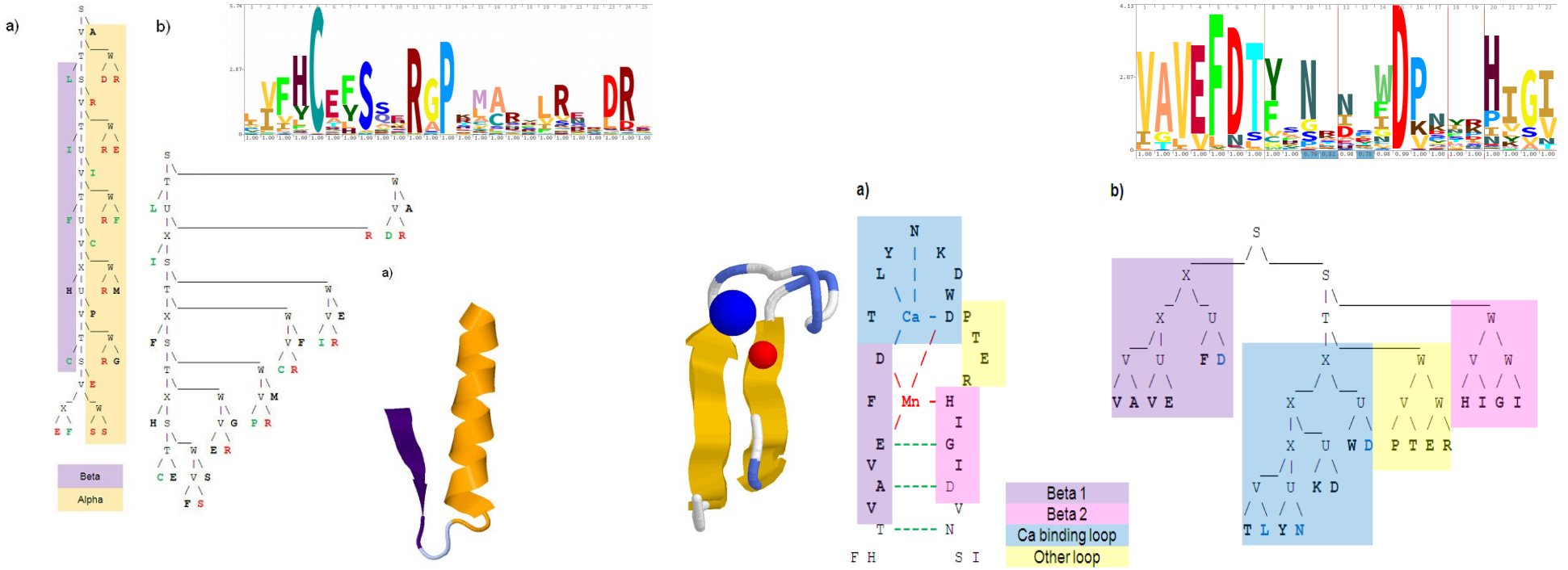
Dodatkowa informacja
przy indukcji gramatyki



Daskalov et al. Sci Rep 2015

PCFG: więcej niż klasyfikator (1)

Drzewo parsowania może oddawać topologię fragmentu białka lepiej niż liniowy profil

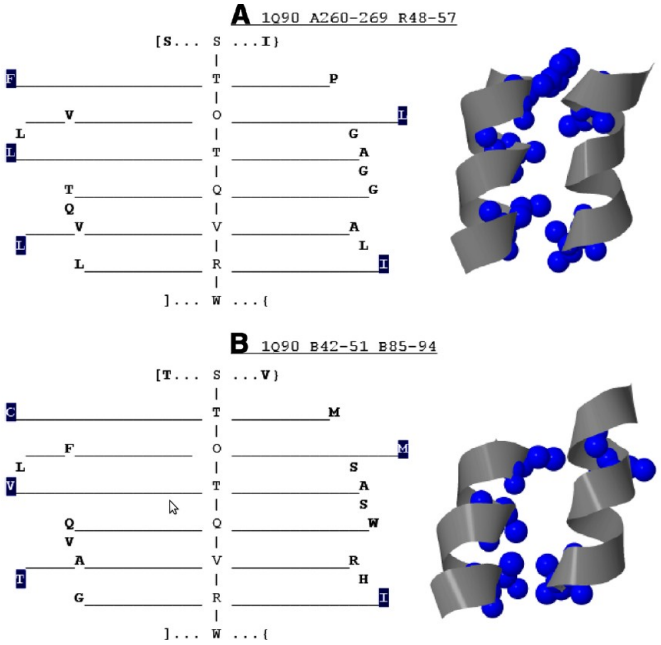


W.Dyrka & J-C.Nebel BMC Bioinformatics 2009

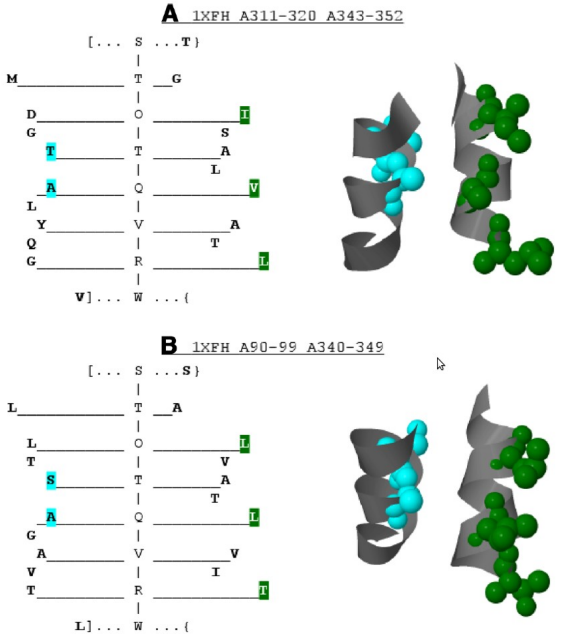
PCFG: więcej niż klasyfikator (2)

Drzewo parsowania *a priori*

- oczekiwany poziom cechy ilościowej na danej pozycji w sekwencji dla wyprowadzeń różniących się tylko co do typu (a nie ilości) generowanych symboli terminalnych



Przewidywany poziom hydrofobowości aminokwasu



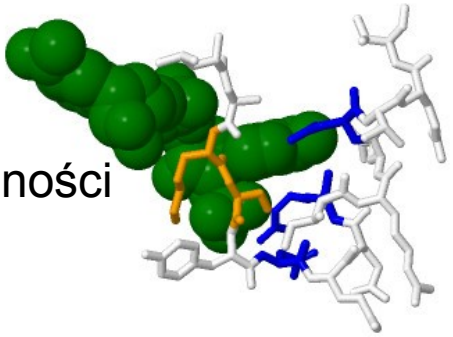
Przewidywany rozmiar aminokwasu

W.Dyrka et al. Alg Mol Biol 2013

PCFG: więcej niż klasyfikator (3)

Analiza prawdopodobnych cyklów w wyprowadzeniach

Wiązanie ligandu NAP:



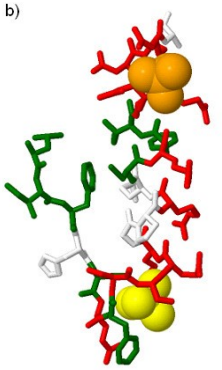
Gramatyka oparta na skłonności aminokwasu do ligandu

S → zU (0.81) | Sz (0.19)
T → pV (0.72) | Tz (0.28)
U → zT (0.72) | Uz (0.28)
V → pU (0.24) | Vz (0.23) | Sp (0.19)
| zp (0.23) | np (0.11)

Najkrótsze wyprowadzenie cyklu:

T => **p**V => **p**S**p** =>
=> **p**z U**p** => **p**zz T**p**

Wiązanie ligandu SO4:



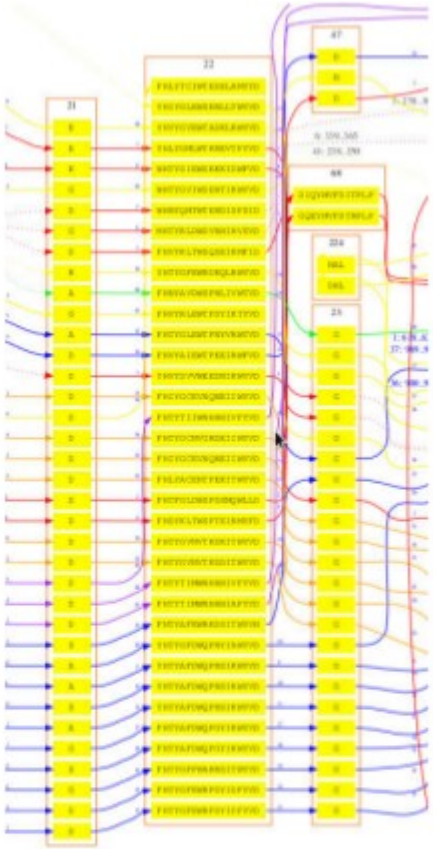
Gramatyka oparta na skłonności do ligandu

S → TW (1.00)
T → XW (0.51) | nU (0.25) | nz (0.25)
U → Xz (0.39) | Xn (0.34) | Xp (0.27)
V → np (0.57) | zp (0.36) | nz (0.07)
W → Vz (0.94) | Vp (0.06)
X → zS (0.54) | nS (0.46)

Najkrótsze wyprowadzenie cyklu:

S =>TW =>TVz =>Tu**p**z
=>Xwu**p**z
=>Xu**p**zu**p**z =>uSu**p**zu**p**z
u = n|z

Protomata [Coste & Kerbellec 2005]



(a) bloc commun à plusieurs familles

(b) projection du concept associé sur la structure 3D

- *to get new insights into the family, when classical MSA are insufficient*
- *to search for new family members in the sequence data banks, with the advantage of a finer level of expressivity than PSSM, Profile HMM or Prosite Patterns*

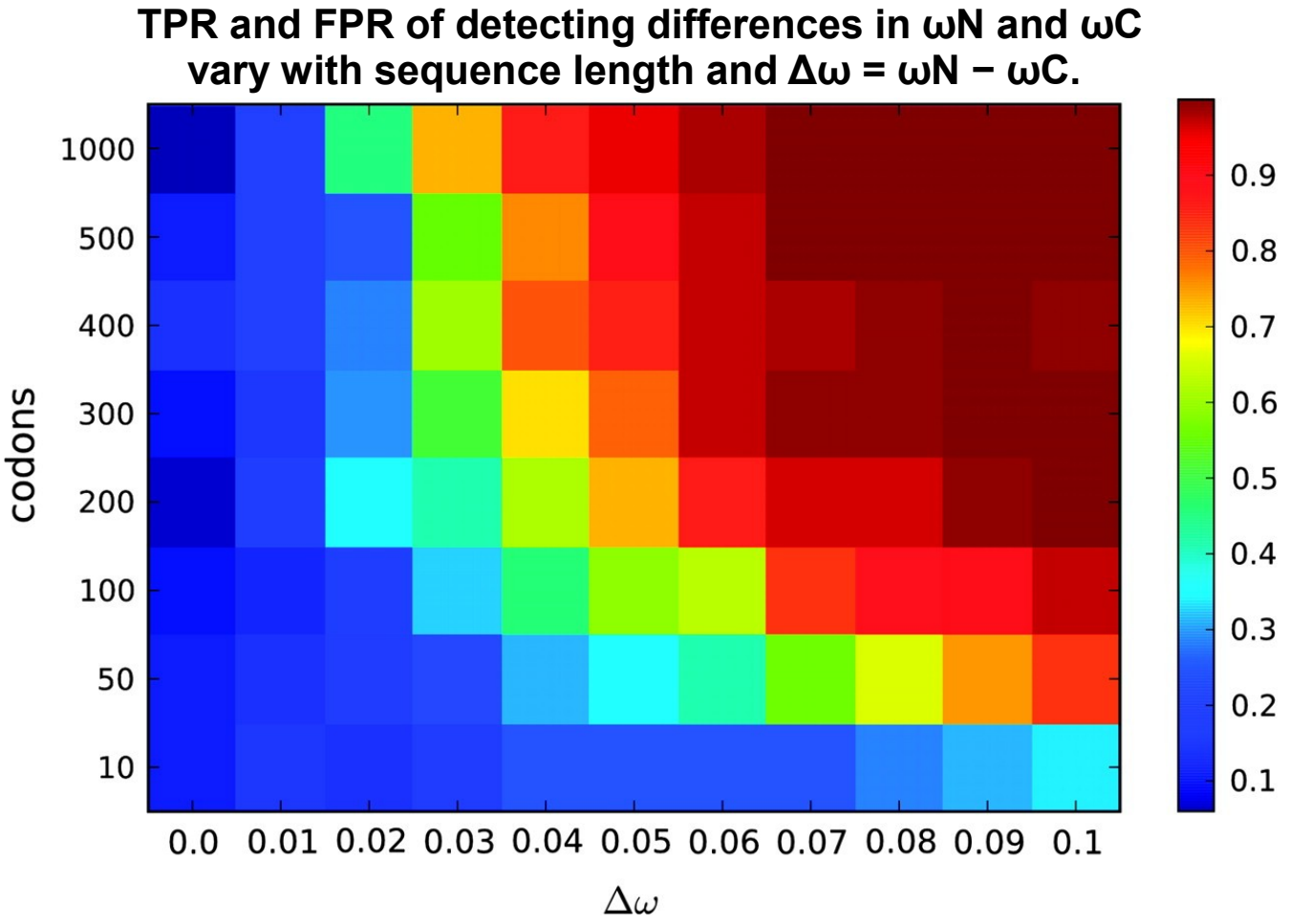
(a) blocs caractéristique de la famille des kappa carraghenases (chaque couleur de séquence représente une famille, ici toutes les séquences passant par le bloc 178 appartiennent donc à la même famille)

G.Garet, praca PhD, Uni Rennes 1 (2015)

Dobór modelu szybkości ewolucji

Model 1:
- taka sama szybkość w C- i N-końcu białka

Model 2:
- różna szybkość w C- i N-końcu białka



Andreas Heger et al. Mol Biol Evol 2009

Dziękuję

Wydział Podstawowych Problemów Techniki
Katedra Inżynierii Biomedycznej
Zespoł Bioinformatyki i Biofizyki Nanoporów

- prof. Małgorzata Kotulska, dr Bogumił Konopka, Florence Thirion, Paweł Woźniak, Jankat Var

Wydział Elektroniki
Katedra Informatyki Technicznej

- prof. Olgierd Unold

Non-Self Recognition in Fungi

- dr Sven J Saupe, dr Mathieu Paoletti, dr Asen Daskalov

Faculty of Computing, Information Systems and Mathematics

- dr Jean-Christophe Nebel

Team MAGNOME / PLEIADE (Bordeaux)

- prof. David J Sherman, dr Pascal Durrens (LaBRI CNRS)

Team DYLISS (Rennes)

- dr François Coste

Usługi obliczeniowe:

- Wrocławskie Centrum Sieciowo-Superkomput.
- Plateforme Fédérative pour la Recherche en Informatique et Mathématiques

Finansowanie:

- British Council (2008-10)
- Ministerstwo Nauki i Szk. Wyższego (2009-11)
- Unia Europejska – Kapitał Ludzki (2009-11)
- Narodowe Centrum Nauki (2011-13, 2016-19)
- Fundacja na rzecz Nauki Polskiej (2012)
- Agence Nationale de la Recherche (2012-15)