

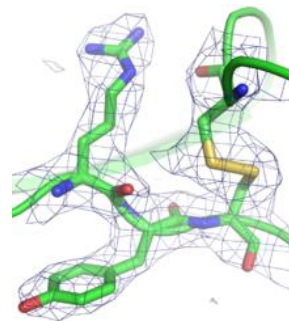
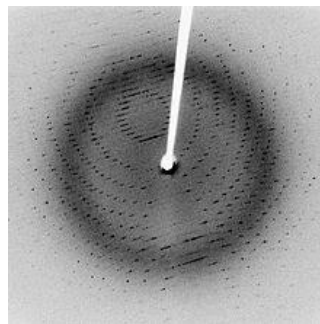
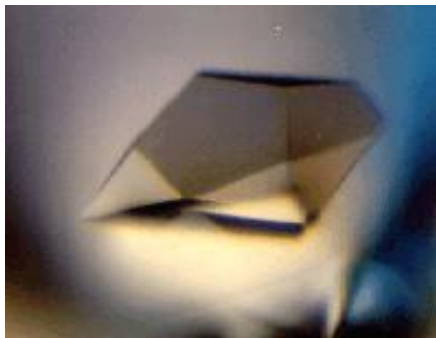
Modelowanie struktury białek i ocena funkcjonalna poprawności modelu

Małgorzata Kotulska

Witold Dyrka, Bogumił Konopka, Monika Rybicka,
Paweł Gąsior, Paweł Woźniak, Tomasz Gołda

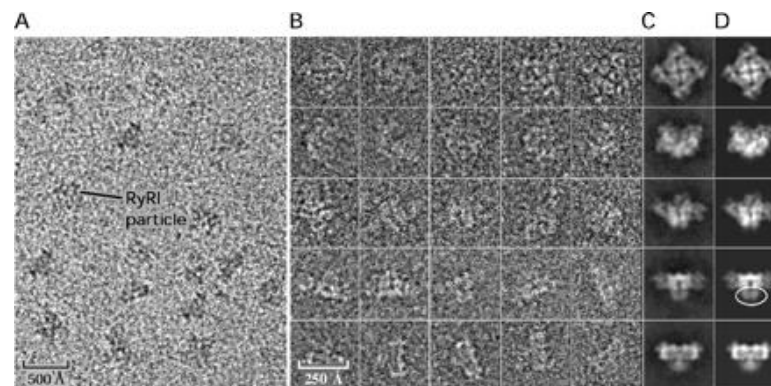
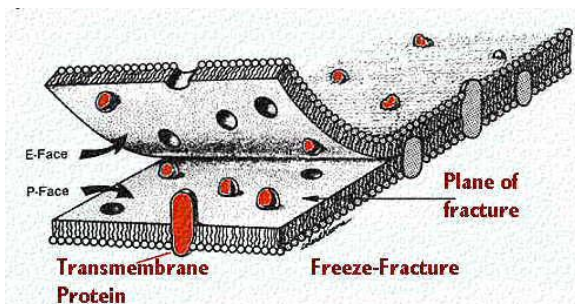
Jakie są metody eksperymentalne?

Obraz dyfrakcji kryształu białka z promieniowania Roentgena

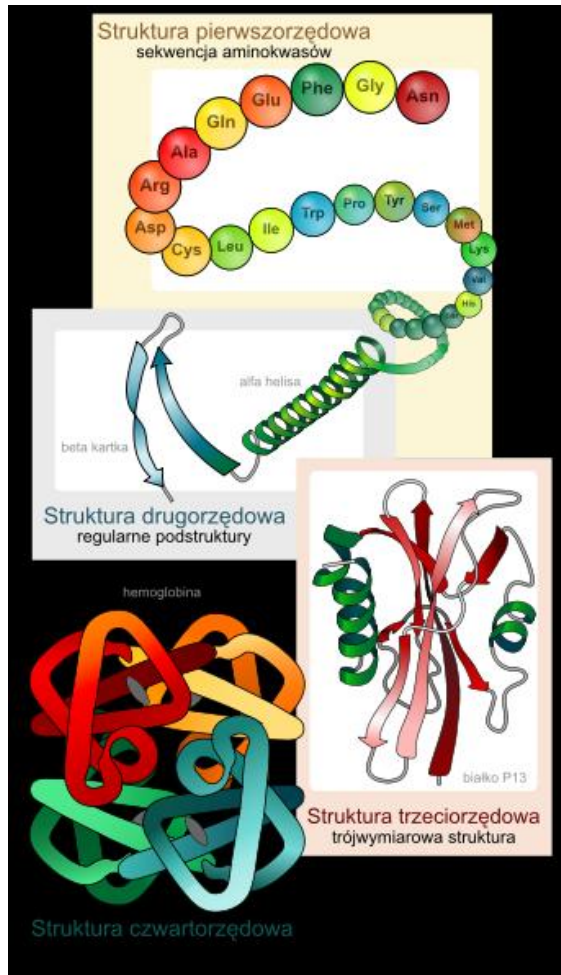


Nuclear magnetic resonance spectroscopy (NMR)

Mikroskopia Elektronowa



Motywacja biologiczna



Przewidywanie struktury przestrzennej białek, dla których nie można łatwo wyznaczyć jej doświadczalnie, wyłącznie z sekwencji aminokwasów.

Dotyczy to białek **trudnych w krystalizacji, bardzo dużych** lub występujących w specyficznym środowisku naturalnym, np. **transmembranowych**.

Jak wielu białek dotyczy ten problem?

UniProtKB/Swiss-Prot - **533 049** sekwencji białkowych (16.11.2011)

UniProtKB/TrEMBLt - **18 215 214** sekwencji białkowych
przetłumaczonych z DNA (16.11.2011)

PDB - **77 243** trójwymiarowych struktur białkowych (16.11.2011)

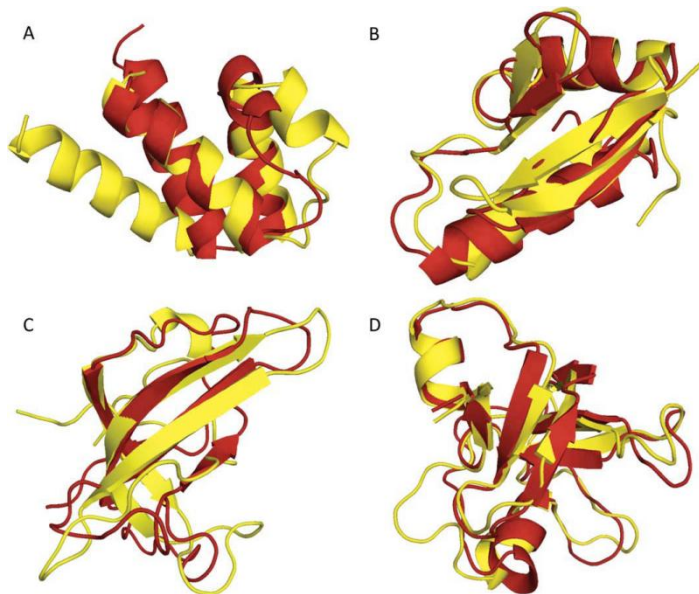
Klasyczne metody modelowania struktury trzeciorzędowej

- Ab-initio / de Novo
- Porównawcze:
 - Homologiczne
 - Nawlekanie

Skuteczność istniejących metod Testowane w konkursie CASP

- Przewidywanie struktury drugorzędowej 50-90%
 - Przewidywanie struktury trzeciorzędowej 10-90%
 - Brak znaczącego postępu skuteczności
- ✓ Wynik zależy od metody oraz trudności białka
- ✓ Brak postępu w wynikach przewidywania w ciągu ostatnich edycji CASP

Predykcja miejsc kontaktowych jako obiecujący kierunek działania



Najlepsze modele odtworzone wyłącznie w oparciu o przewidziane miejsca kontaktowe (CASP 8)

M.L. Tress, A. Valencia, Proteins 2010; 78:1980–1991.

- **CASP 4 (2000)**

„Detecting residue-residue contacts in proteins (RR)” – **6 grup**

- **CASP 9 (2010)**

„Detecting residue-residue contacts in proteins (RR)” – **18 grup**

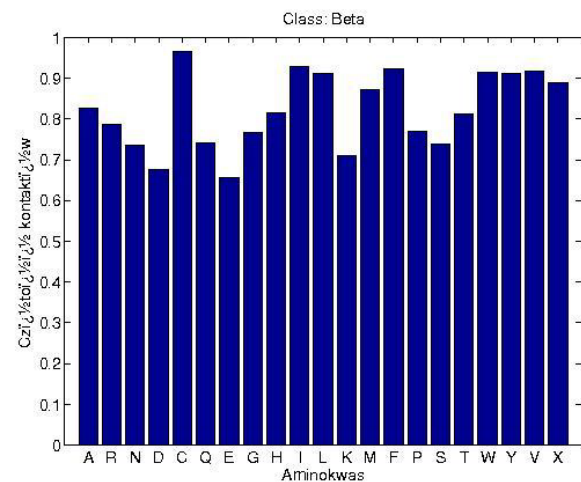
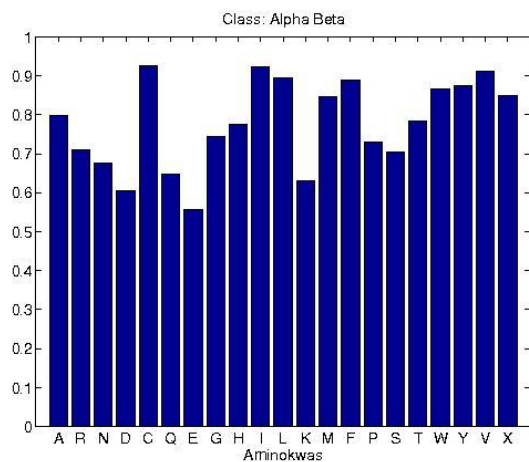
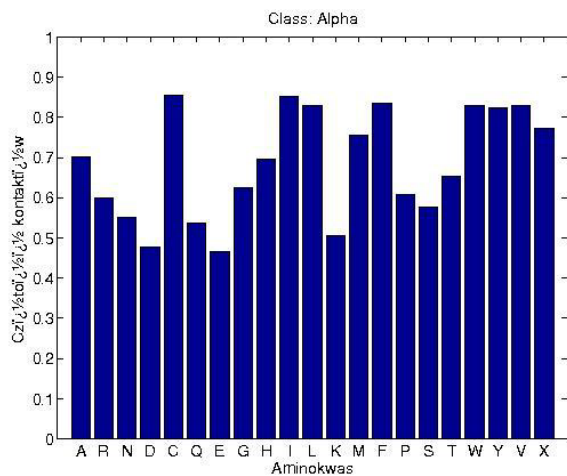
(3 kategorie odległości: $x > 6$, $x > 12$, $x > 24$ aminokwasów)

Co to jest miejsce kontaktowe?



Miejsca kontaktowe

Charakterystyka kontaktów zależy od klasy białek



Kontakty helisa-helisa

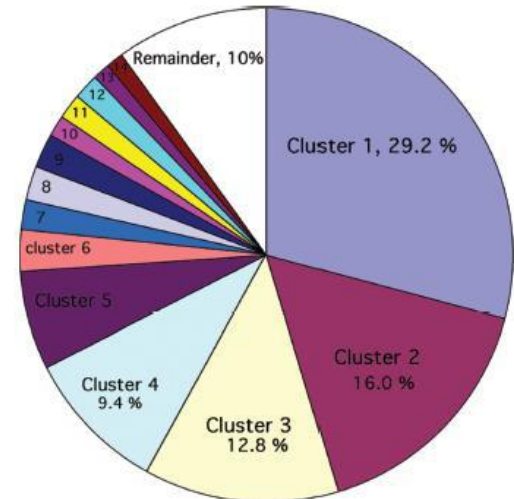
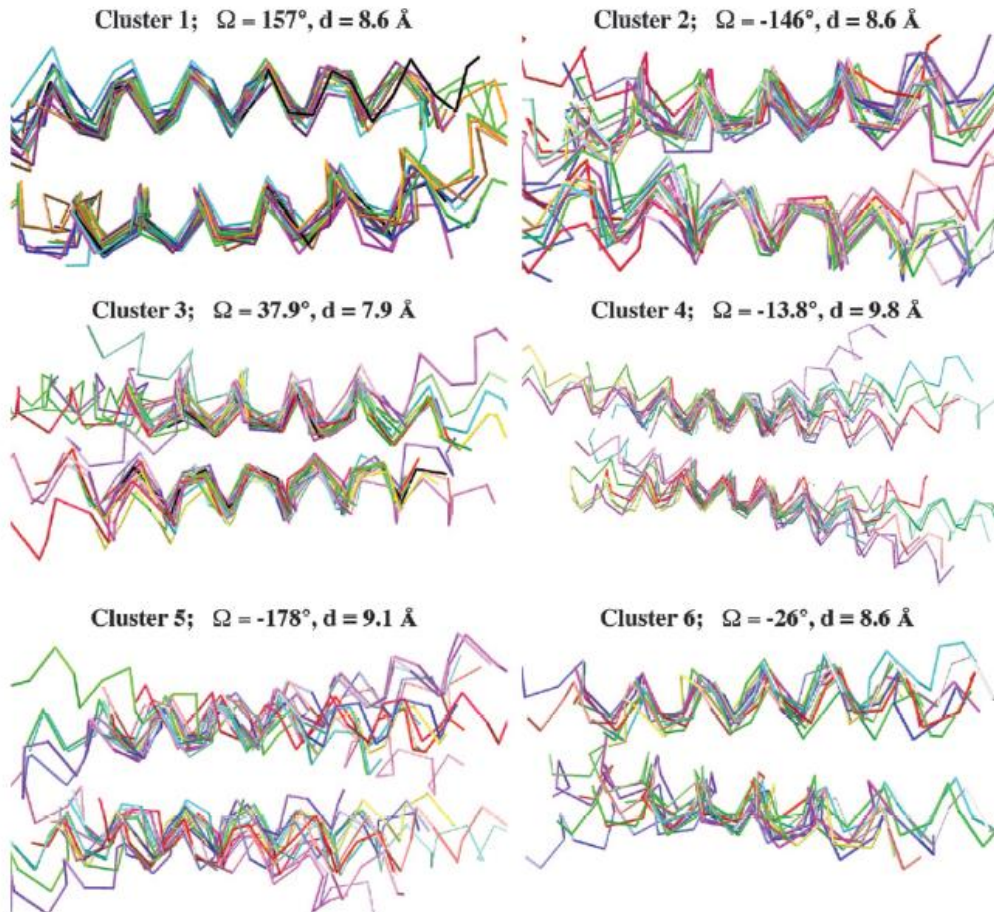


Fig. 1. Pie chart showing the fraction of the total number of pairs that fall within a given cluster.

Probabilistyczny model gramatyczny języka białkowego

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

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

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

acc_high → ... | L (0.00) | K (0.18) | M (0.00) | ...

acc_med → ... | L (0.03) | K (0.00) | M (0.05) | ...

acc_low → ... | L (0.13) | K (0.00) | M (0.11) | ...

Reguły nieterminalne (tworzenie interfejsów):

Inter_ds → •• *Outer_dd* •• | • *Outer_sd* •• | ε

Inter_ds → •• *Outer_dd* •• | ε (Waldispuehl&Steyaert'05)

...

• ≡ *acc_high* | *acc_med* | *acc_low*

Prawdopodobieństwa reguł nieterminalnych:

Inter_ds → *acc_high* *Outer_sd* *acc_med* *acc_med* (0.14)

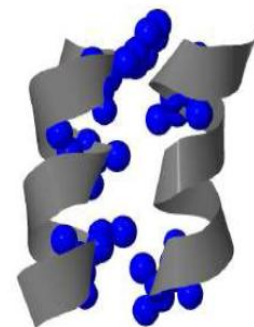
acc_med *Outer_sd* *acc_med* *acc_med* (0.22)

acc_low *Outer_sd* *acc_med* *acc_low* (0.20)

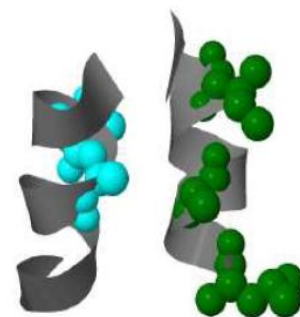
acc_low *Outer_sd* *acc_med* *acc_med* (0.44)

Wyniki klasyfikacji

- Ocena klasyfikatora: 4-krotna walidacja krzyżowa (ROC)
- Najlepsze pojedyncze klasyfikatory – w oparciu o cechy:
 - dostępność (hydrofilowość) dla klasy 1,4
 - rozmiar van der Waalsa dla klasy 2,3



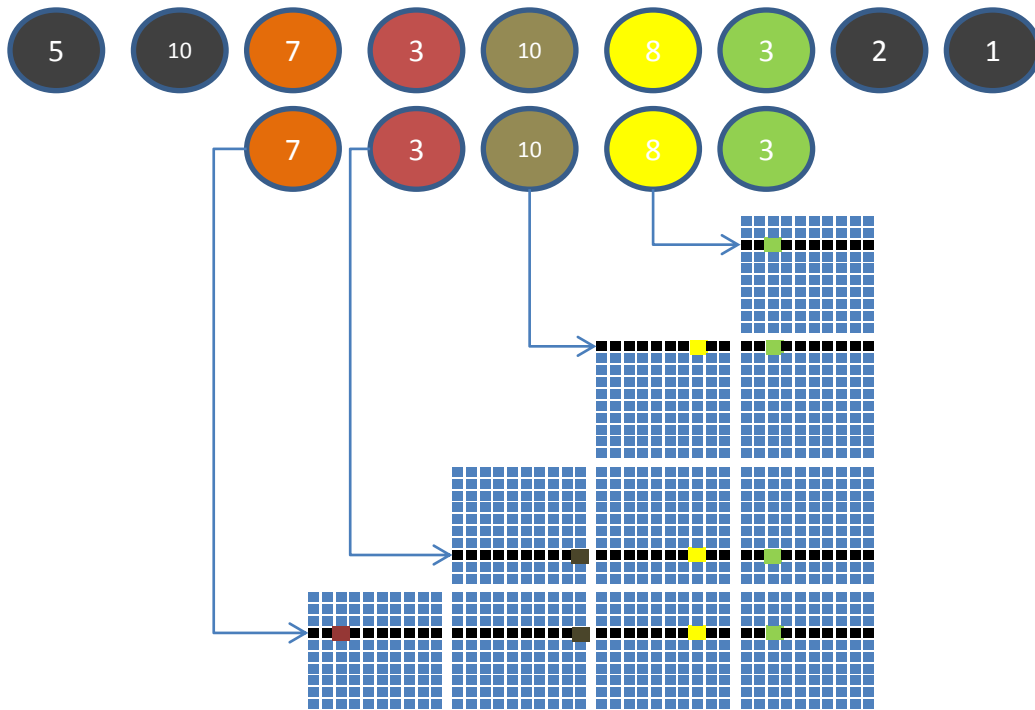
klasa 1



klasa 2

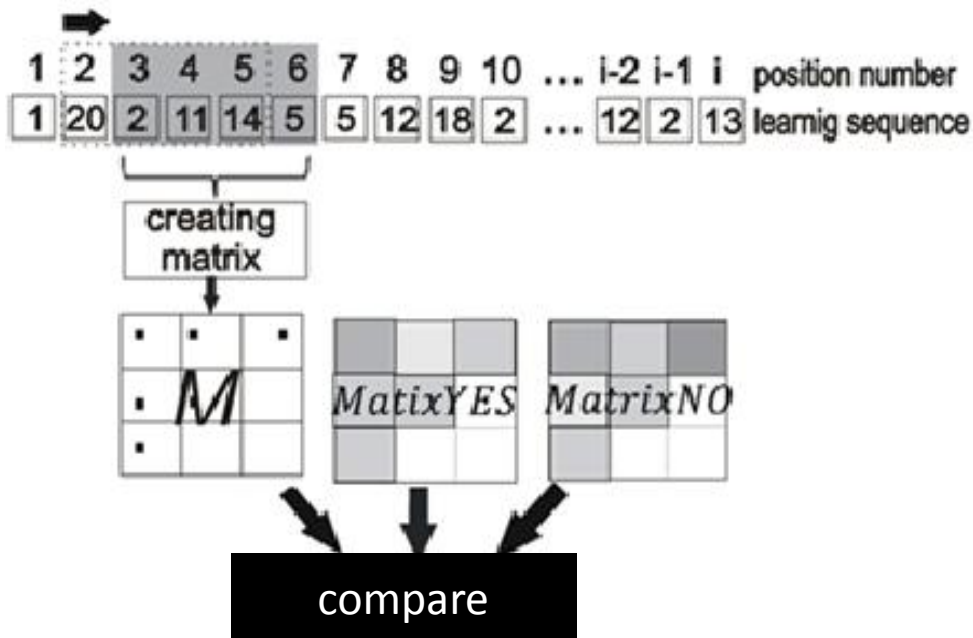
Klasa	Liczność	AUC ROC		$R_c(Pr=1.0)$
		pojedynczy	multi-klas.	multi-klas.
c1	67	0.59	0.72	0.20
c2	35	0.68	0.81	0.14
c3	23	0.71	0.78	0.26
c4	15	0.70	0.84	0.41
średnia	$\Sigma=140$	0.64	0.77	0.22*

* oznacza możliwość przypisania 22% par struktury trzeciorzędowej z dokładnością $<1.5\text{\AA}$ ($C\alpha$)

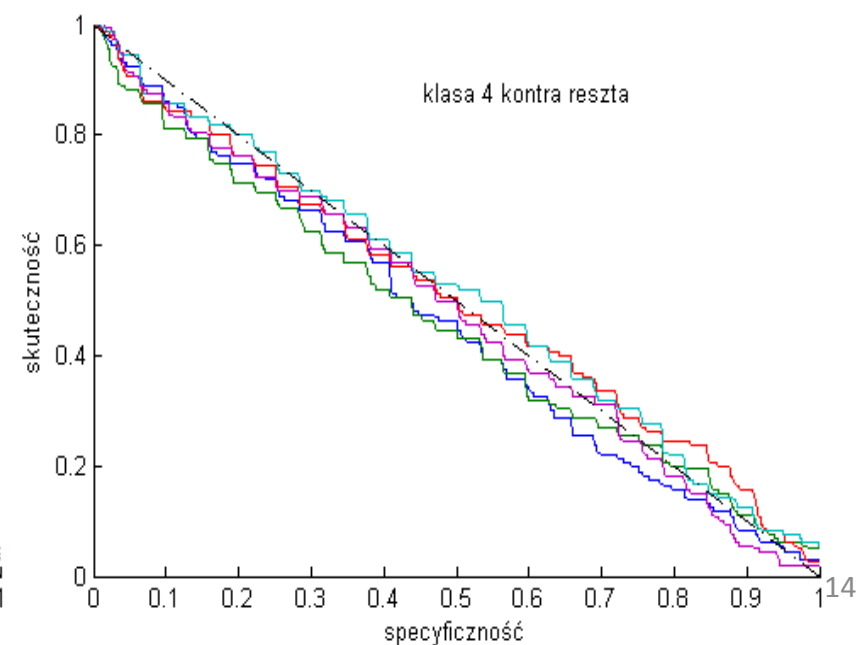
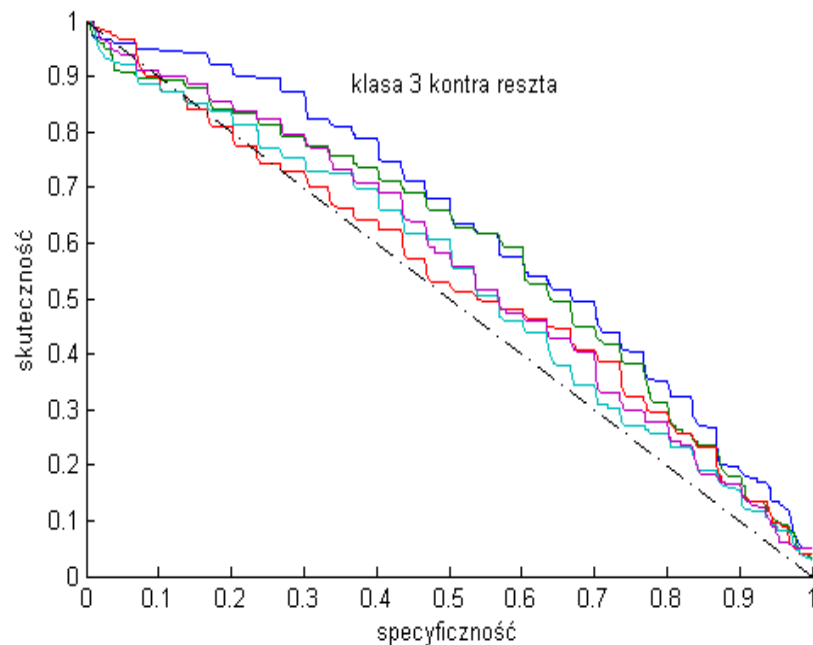
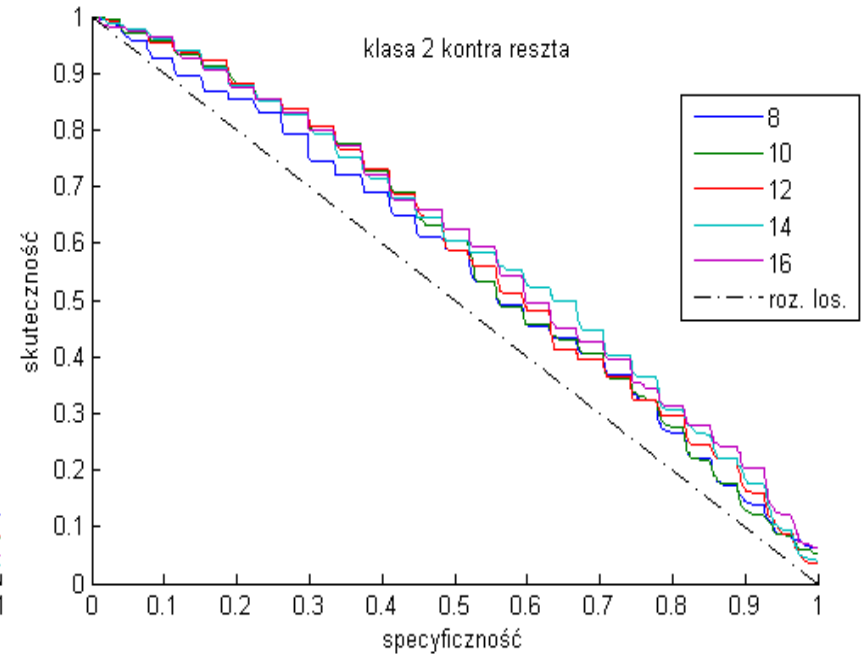
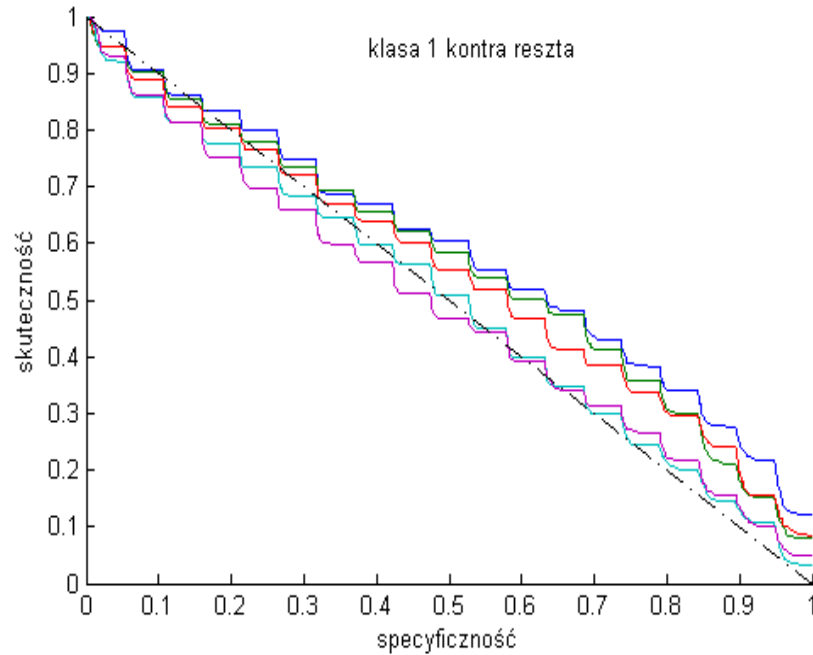


Uczenie kowariancyjne z wykorzystaniem aminokwasów

- ❑ Ograniczenie informacji - wybór małego fragmentu
- ❑ Na podstawie danych uczących stworzenie MatrixYES oraz MatrixNO
- ❑ Wybór najbardziej podobnej macierzy do MatrixYES i najbardziej różniącej się od MatrixNO

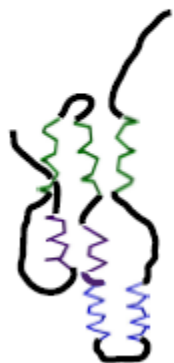


Krzywa ROC dla różnych okien

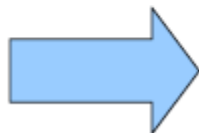


Struktura a funkcja

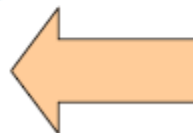
Predict helix-helix contact conformation



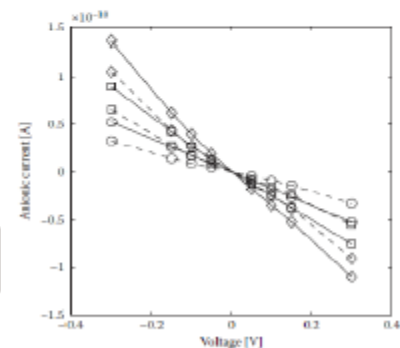
Constrain
ab initio
prediction



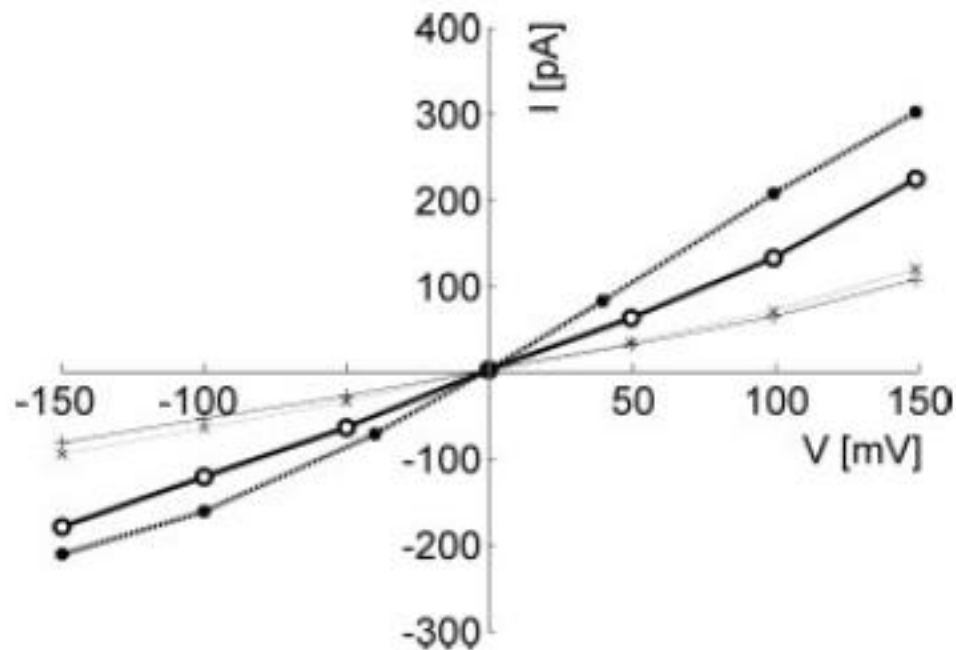
Validate
ab initio
prediction



Calculate transport
characteristics for
channel model



Modelowanie funkcjonalne przepływów



Modelowanie funkcjonalne przepływów

- dyskretne: *dynamika molekularna (MD)*
- ciąгло-dyskretne: *dynamika Browna (BD)*
- ciągłe: model *Poissona-Nernsta-Plancka*

$$J = -D \left(\nabla n + n \cdot \nabla \Phi \cdot \frac{ze}{kT} \right)$$

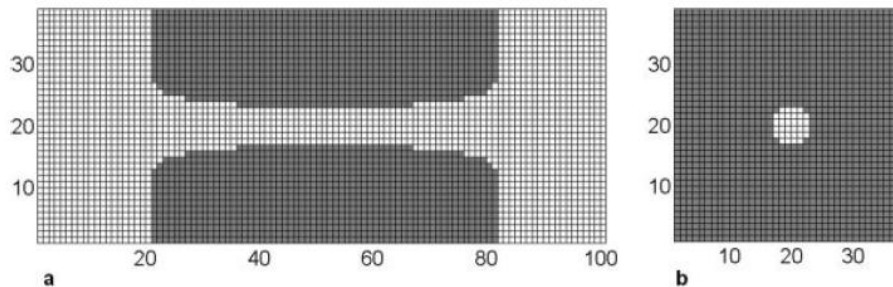
Nernst-Planck

$$\varepsilon_0 \nabla \cdot [\varepsilon(r) \nabla \Phi(r)] = -e \sum_v z_v n_v(r) - \rho$$

Poisson

Modelowanie dyskretne przepływu

Siatka o wymiarach $[h_x h_y h_z]$, warunki stacjonarne przepływu: $\nabla \cdot \mathbf{J} = 0$



$$n_i = \frac{\sum_{j=1}^6 n_j \frac{D_i + D_j}{h_j} \left[1 + \frac{ez}{2kT} (\Phi_j - \Phi_i) \right]}{\sum_{j=1}^6 \frac{D_i + D_j}{h_j} \left[1 - \frac{ez}{2kT} (\Phi_j - \Phi_i) \right]}$$

$$\Phi_i = \frac{\sum_{j=1}^6 \epsilon_j \Phi_j / h_j^2 + \sum_{\nu} z_{\nu} e n_{\nu i} / \epsilon_0 + \rho_i / \epsilon_0 V}{\sum_{j=1}^6 \epsilon_j / h_j^2}$$

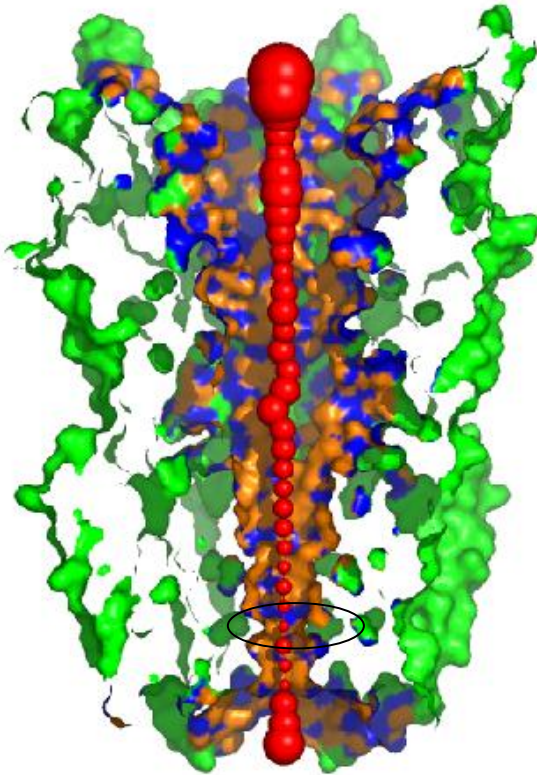
Jak dobrze PNP modeluje przepływ?

TABLE XIX. Summary of results

Channel	$\varnothing[\text{\AA}]$	Conductance	Rectification	Selectivity
α -hemolysin	14	+	+	+
GLIC2	5.4	$\uparrow$	+	+
KcsA	3	$\downarrow$	+	+
MthK	≤ 2	$\downarrow$	-	+
Aquaporin-1	3	+	?	+
ClC	≤ 2	+	?	+

Czy model przepływowy jest
wrażliwy na pojedyncze mutacje?

GLIC2



5x GLU222

- Neutralization of 5 glutamate residues:
- Change of the protein net charge from $-5e$ to $0e$ at $\text{pH}=5$
- Selectivity change – predicted by 3D PNP:
cation selective $\rightarrow$ anion selective
- Consistent with Brownian Dynamics simulation

Table 9.7: Current-voltage characteristic of GLIC2 channel model in symmetrical 140 mM NaCl solution at $\text{pH}=5$, dielectric coefficients $\epsilon_p = 4$, $\epsilon_{nw} = 40$, partition coefficient $\xi = 0.5$. In (Song & Corry 2010): $\epsilon_p = 2$, $\epsilon_{nw} = 60$

Model	D	net charge[e]	V = -100 mV		V = +100 mV		I(+)/I(-)
			I[pA]	I ₊ /I ₋	I[pA]	I ₊ /I ₋	
Our PNP, $\Delta=3\text{\AA}$	unified	-5 (WT)	-5.8	1.6:1	8.5	1.4:1	1.47
		neutral	-7.3	1:1.2	7.1	1:1.2	0.98
Song'10 BD	ion-spec.	-5 (WT)	-1.4	∞ :1	2.2	∞ :1	1.57
	ion-spec.	neutral	-	1:13	-	1:13	-

GLIC – pojedyncza mutacja

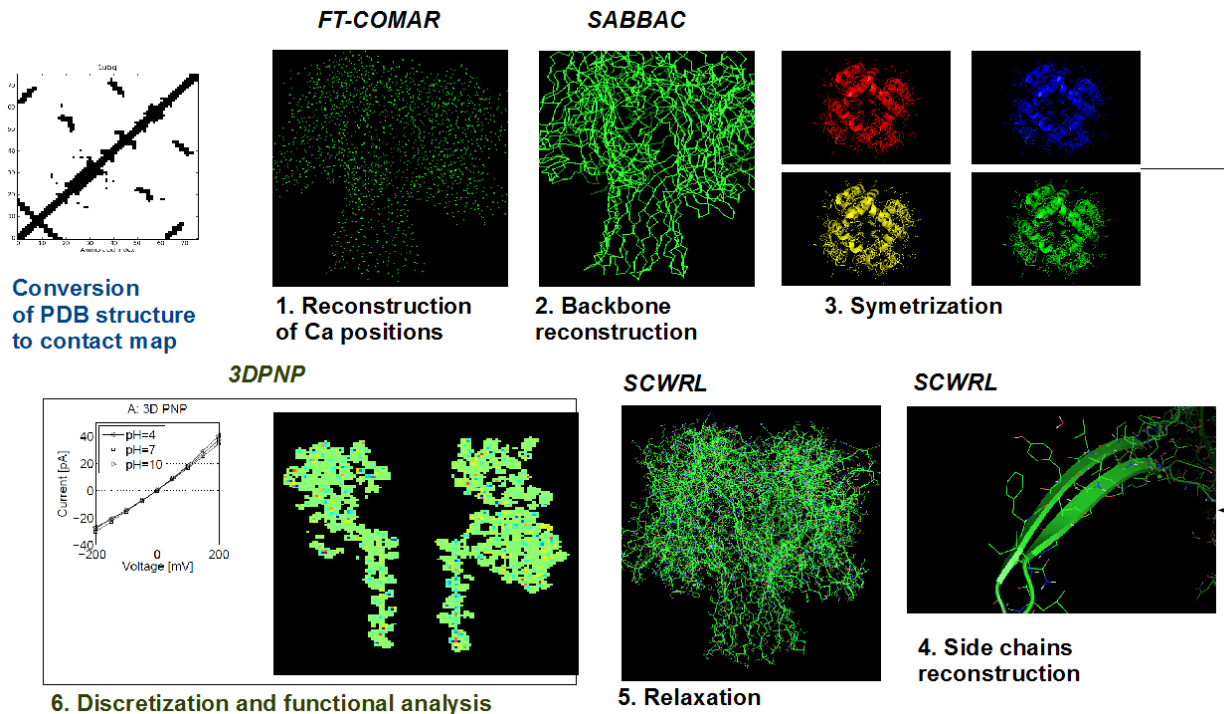
- Neutralizację GLU 222 można potraktować jako model substytucji glutaminianu przez neutralny aminokwas
- Zmiana selektywności i linearyzacja I-V, jako wynik mutacji oraz modelu
- Podobny efekt jak w 3D PNP został uzyskany przez dynamikę Browna (BD)

Ocena funkcjonalna modelu strukturalnego

Jak dobrze można odtworzyć strukturę kanału z kontaktów?

Struktura kanału - odtwarzanie z kontaktów

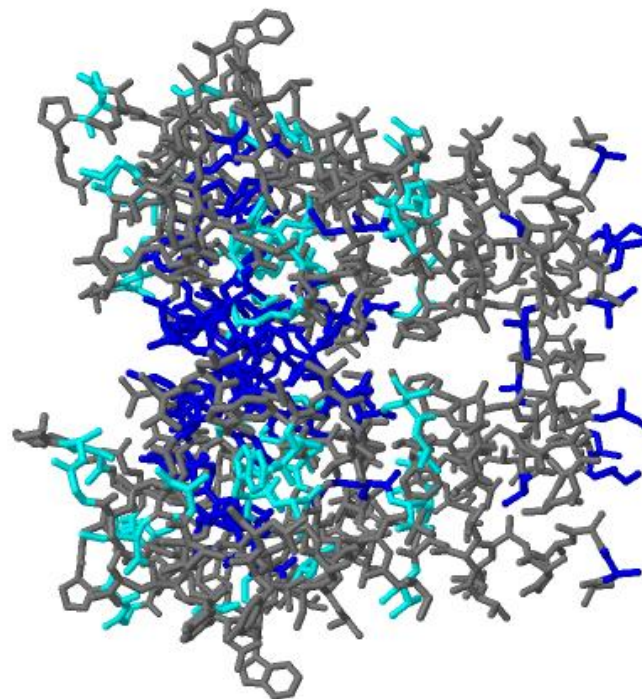
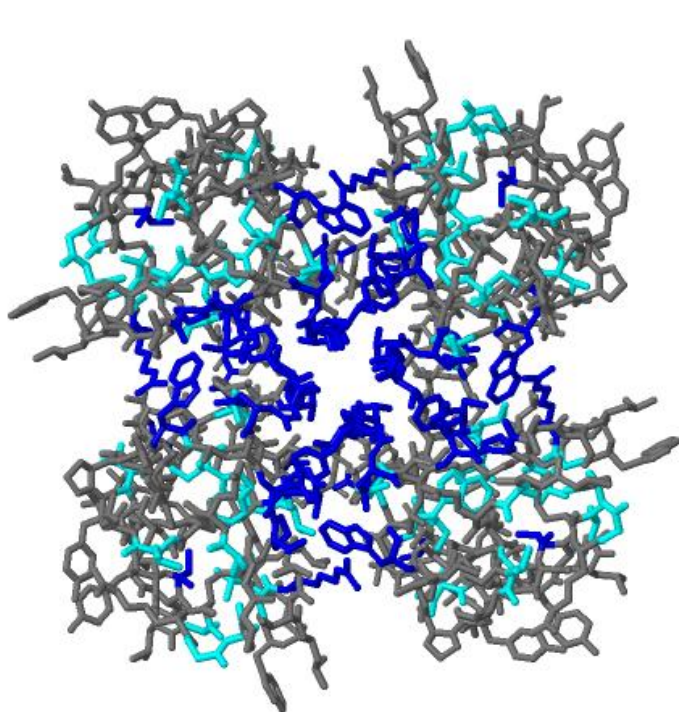
From contact map
through channel structure
to I-V curve



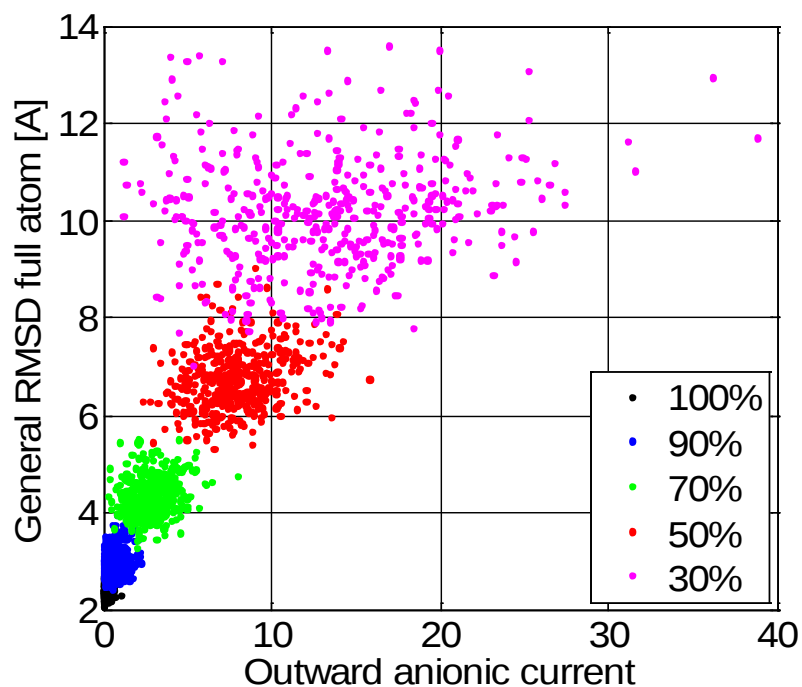
Korelacje pomiędzy RMSD aminokwasu a odchyleniem w funkcjonalności kanału

Kolor wskazuje miejsce z istotną korelacją ($p\text{val} \leq 0.01$) pomiędzy aminokwasem a

- Przynajmniej 2 (**niebieski**)
- 1 (**błękitny**) cechą funkcjonalną

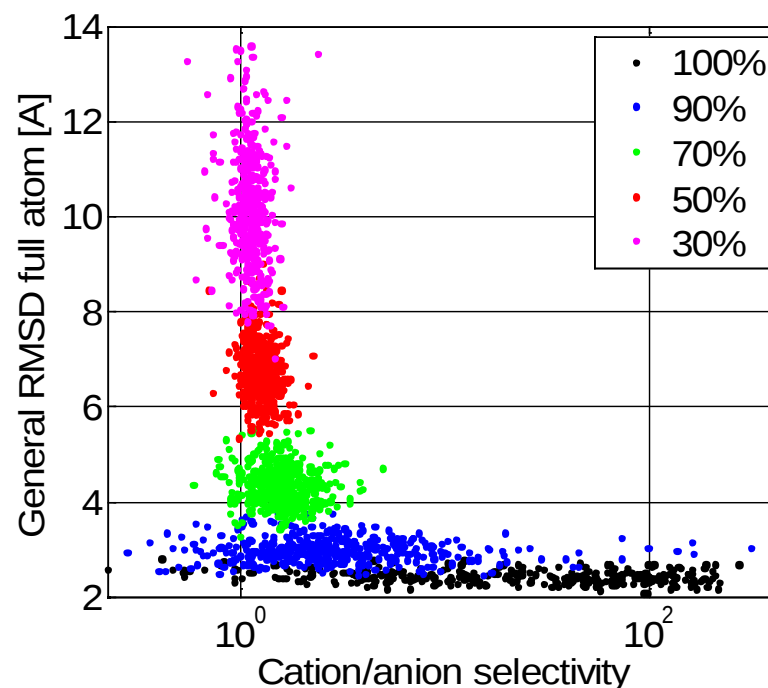


Korelacja pomiędzy całkowitym RMSD modelu a jego funkcjonalnością



Korelacje Kendalla

100%: $\tau = 0.15$, $pval = 3.8e-5$
30-90%: $\tau = 0.71$, $pval \approx 0$
90%: $\tau = 0.10$, $pval = 3.4e-3$
70%: $\tau = 0.15$, $pval = 4.5e-6$
50%: $\tau = 0.18$, $pval = 2.0e-8$
30%: $\tau = 0.09$, $pval = 3.7e-3$



Korelacje Kendalla

100%: $t = -0.15$, $pval = 4.4e-5$
30-90%: $\tau = -0.47$, $pval = 2.3e-184$
90%: $t = -0.03$, $pval = 4.2e-1$
70%: $t = -0.09$, $pval = 6.5e-3$
50%: $t = -0.13$, $pval = 4.3e-5$
30%: $t = -0.08$, $pval = 9.2e-3$

Dynamika Browna

Przepływ jonów w oparciu o równanie Langevina

$$m_i \frac{dv_i}{dt} = -m_i \gamma_i v_i + R_i + F_i,$$

$i=1,\dots,N$

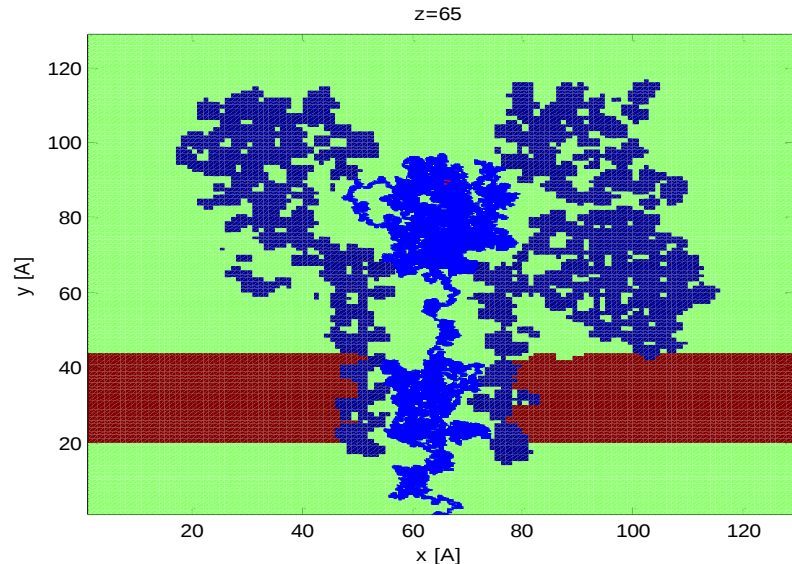
m_i – masa

v_i – prędkość

γ_i – współczynnik tarcia

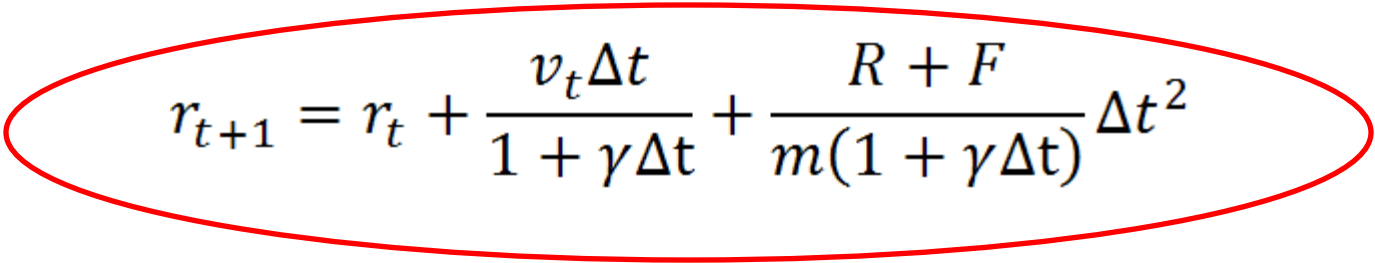
R_i – siła losowa

F_i – siła systematyczna



ROZWIĄZANIE NUMERYCZNE

$$\left\{ \begin{array}{l} m\ddot{r}_{t+1} = -m\gamma\dot{r}_{t+1} + R + F \\ \ddot{r}_{t+1} = \frac{\dot{r}_{t+1} - \dot{r}_t}{\Delta t} \\ \dot{r}_{t+1} = \frac{r_{t+1} - r_t}{\Delta t} \end{array} \right. \quad \frac{r_t - r_{t-1}}{\Delta t} = v_t$$


$$r_{t+1} = r_t + \frac{v_t \Delta t}{1 + \gamma \Delta t} + \frac{R + F}{m(1 + \gamma \Delta t)} \Delta t^2$$

r_t – położenie jonu w chwili t

Δt – krok czasowy

m – masa

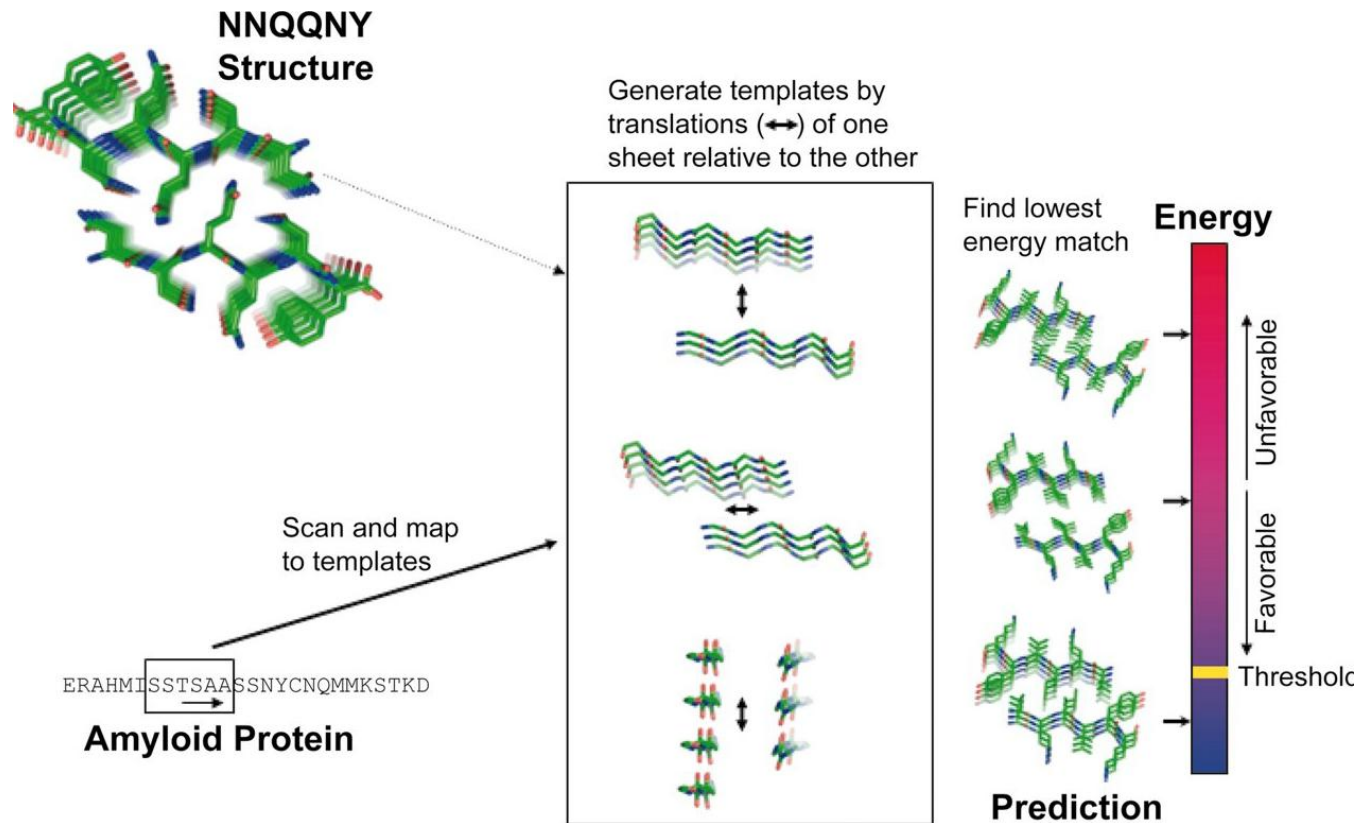
γ – współczynnik tarcia

R – siła losowa

F – siła systematyczna

Białka amyloidogenne

Przewidywanie nietypowych kontaktów w strukturach typu beta. Budowa bazy danych

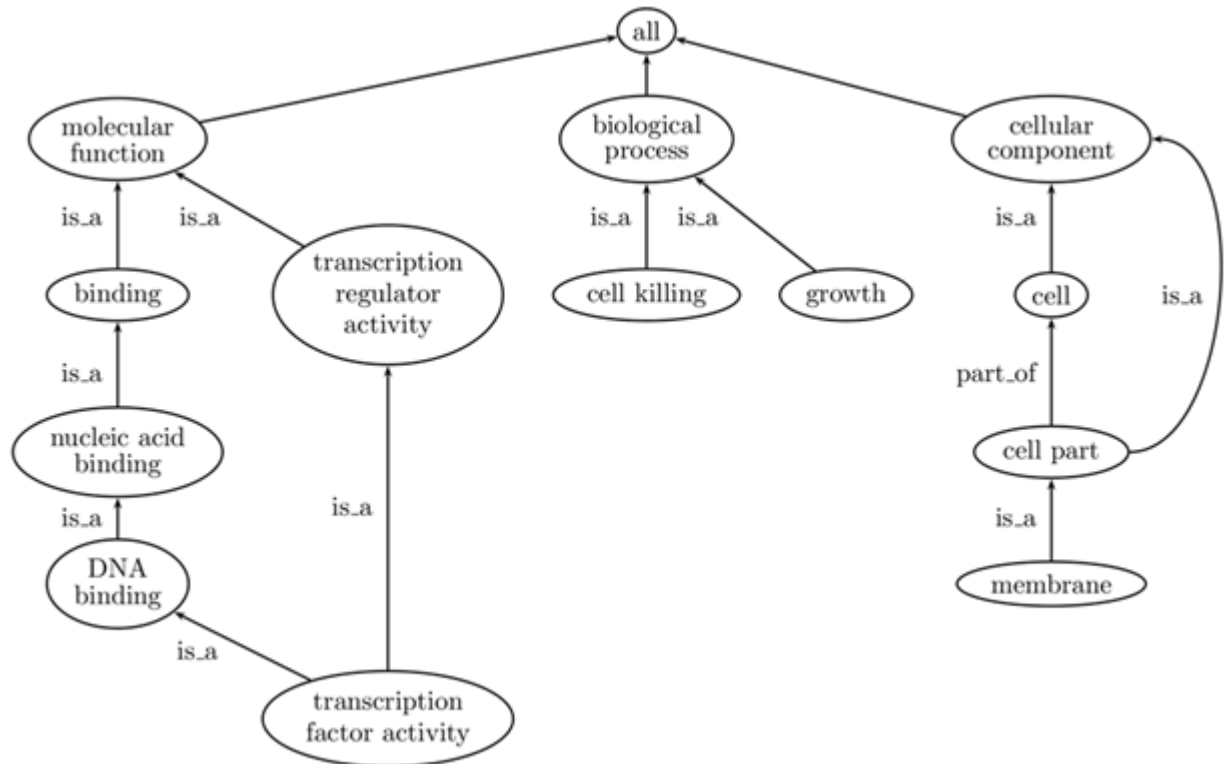


Ontologia

Podobieństwo

Semantyczne – Miara określająca stopień podobieństwa dwóch wyrażeń. Do wyliczania wykorzystuje się algorytmy oparte o badanie ontologii.

Powiązania białek z terminami ontologii
seqdata.dat –
udostępniona przez
serwis ebi.ac.uk
zawiera przypisanie
białek z bazy UNIPROT



Ontologia GO (Gene Ontology)

Wykorzystywane algorytmy

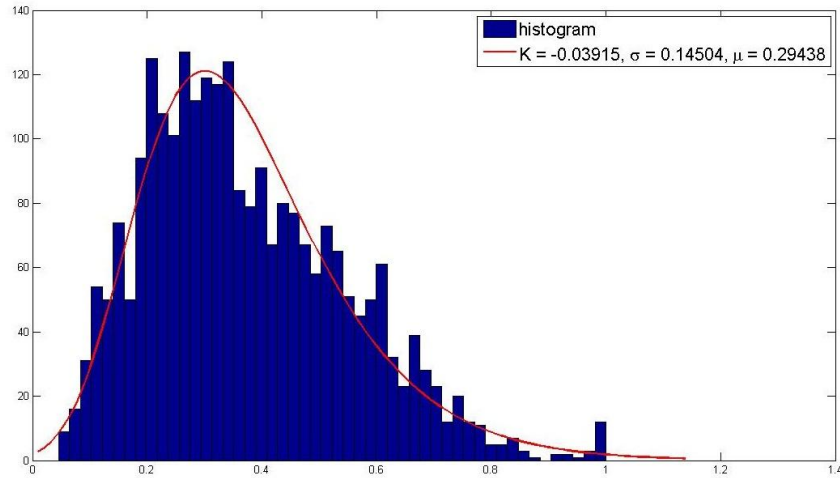
Algorytm Wanga

- Algorytm hybrydowy
- Wykorzystuje strukturę drzewa ontologii
- Bada stosunek ilości wspólnych przodków badanych terminów w odniesieniu do sumy wszystkich przodków obu terminów

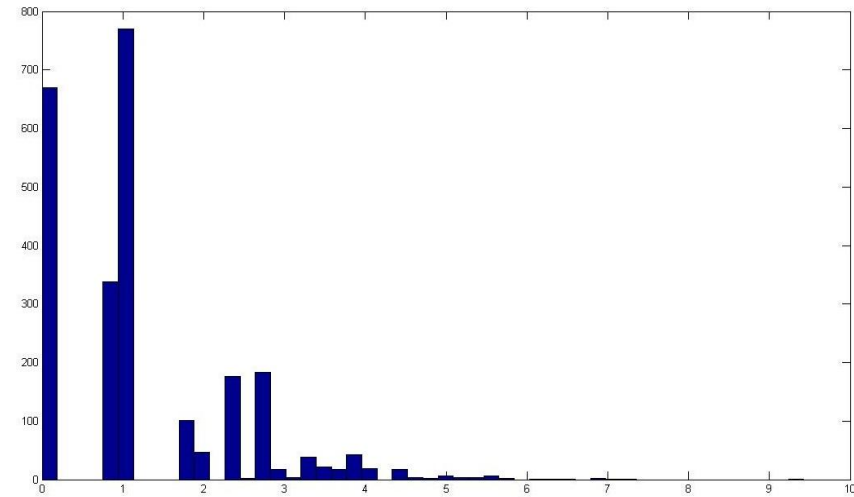
Algorytm Resnika

- Algorytm bazujący na węzłach
- Wykorzystujący teorię informacji
- Bada częstotliwość występowania terminu oraz jego szczegółowość

Rozkład podobieństwa losowo wybranych par białek



Algorytm Wanga



Algorytm Resnika

Rozkład Fishera –Trippetta
Rozkład wartości ekstremalnych

Grupa biofizyki i bioinformatyki nanoporów



Witold
Dyrka



Bogumil
Konopka



Joanna
Basałyga



<http://www.ibp.pwr.wroc.pl/KotulskaLab/>