Metody ograniczenia przestrzeni poszukiwań w modelowaniu nieznanych struktur białkowych

Instytut Inżynierii Biomedycznej i Pomiarowej
Politechnika Wrocławska

Małgorzata Kotulska

Witold Dyrka, Bogumił Konopka, Monika Kurczyńska, Joanna Weżgowiec, Paweł Woźniak,

Motywacja biologiczna

CEL: Przewidywanie struktury przestrzennej białek wyłącznie z sekwencji aminokwasów.

OBIEKT: Białka o znanej sekwencji i nieznanej strukturze

Jak wielu białek dotyczy ten problem ?

UniProtKB/Swiss-Prot – 0,5 mln sekwencji białkowych

UniProtKB/TrEMBLt - 20 mln sekwencji białkowych przetłumaczonych z DNA

PDB – 90 tys. trójwymiarowych struktur białkowych

Źródło: UniProt i PDBe

Klasyczne metody modelowania struktury trzeciorzędowej

- Ab-initio / de Novo
- Porównawcze:
  - Homologiczne
  - Nawlekanie
- Statystyczne (uczenie maszynowe)

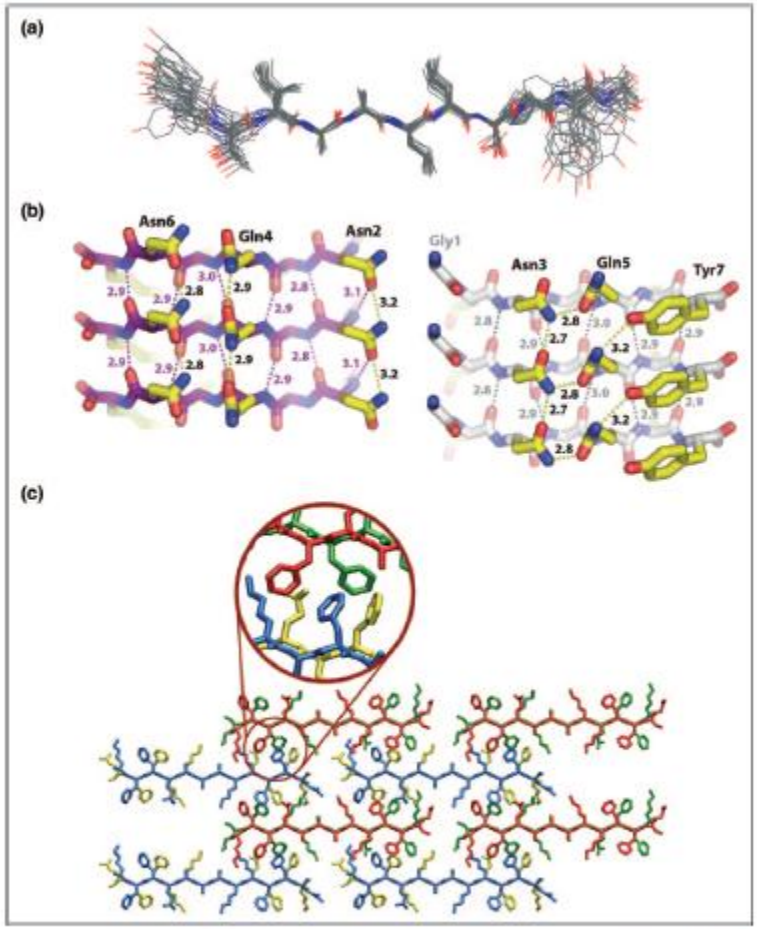
Nasze kierunki badań

- Ocena jakości modeli w oparciu o relację **struktura-funkcja**
  - **charakterystyka przepływowa** kanałów jonowych
  - zastosowanie **ontologii genów** - GO
- Zastosowanie miejsc kontaktowych białek do modelowania ich struktury – **uczenie maszynowe**. **Ontologia miejsc kontaktowych.**
- **Amyloidy** - różne struktury białek o tej samej sekwencji. **Uczenie maszynowe** a **modelowanie fizykochemiczne**

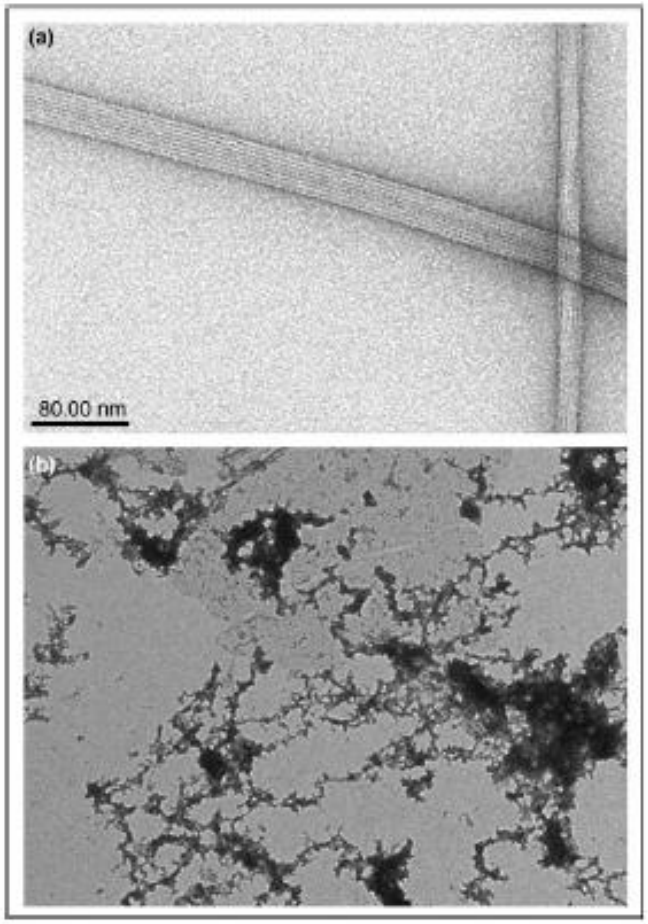
How proteins forming amyloids can be recognized computationally

Amyloid structure

„steric zippers”



High-resolution structures of peptides in amyloid fibrils. **(a)** Structure derived from magic angle spinning NMR of fibrils grown from a peptide fragment of transthyretin [22••]. **(b)** Crystallographic model of a fibril comprising a peptide sequence derived from the yeast prion Sup35 [26
(c) High-resolution model from crystallography and electron microscopy of fibrils of a synthetic peptide sequence [25••].



Electron micrographs of the macromolecular morphology of protein aggregates and amyloid fibrils. **(a)** Amorphous aggregates (of the fusion proteins between the NusA protein and the Alzheimer's β-peptide) and **(b)** amyloid fibers (of the hexapeptide STVIIE [23]). Both aggregates have a high content of β-structure, but dramatically different morphologies when seen through an electron microscope.

Russeau 2006

Exemplary diseases

- Alzheimer disease (amyloβ-42, τ)
- Parkinson disease (α-synuclein)
- Diabetes type II (islet amyloid polypeptide - IAPP, amylin)
- Amylolateral Sclerosis ALS (SOD)
- Huntington disease (huntington)
- Creutzfeld-Jacobs (prion, e.g. sup-35)
- ……

Hot-spots of hexapeptides

Table 1
Amyloid fibril-forming proteins and amyloidogenic determinants experimentally investigated [10 and references therein], and predicted according to [10,14,8]

Amyloidogenic protein	Experimentally determined amyloidogenic regions [10]	Predicted high packing density peptides [10][a]	Conformational switches according to SecStr [14][a] (numbering)	Conformational switches according to SecStr [14] (peptide sequence)	Pattern proposed in [8] according to PATTINPROT [22]
Prion, 253, 22 P04156	132–160	136–141 147–152			
	178–193	174–178 180–184	175–183*	FVHDCVNIT	DCVNIT
		211–217 232–236	209–215*	VVEQMCI	EQMCIT
		240–250	242–251*	LLISFLIFLI	LIFLIV
Amyloid-β, 770, 17 672–713 = > Aβ 42 672–711 = > Aβ 40 P05067	686–695 702–710	687–692 703–708	686–691* 718–721	QKLVFF ITLV	QkLvFF, KLVFFA ITLvmL
Cystatin-C, 146, 26 P01034		73–77 87–92 121–131	73–77* 89–94* 120–123*	LQVVR FLDVEL KAFC	LQVvrA NYFLdV KAFcsF
Amylin, 89, 22 P10997	36–41 42–49	46–51	44–49* 74–79	RLANFL NAVEVL	RLaNFL NAVEvL
Calcitonin, 141, 25 P01258	40–44	48–54 90–95	49–59*	LLLAALVQDYV	LVQdYV
Atrial natriuretic factor, 153, 25 P01160		116–120	37–44 116–119*	LMDFKNLL RALL	MDFknL RALLtA

Why modeling?

Non-informed experimental tests

(variations with repetition)

$20^6 = 64\,000\,000$

Methods in modeling

Physico-chemical methods in modeling

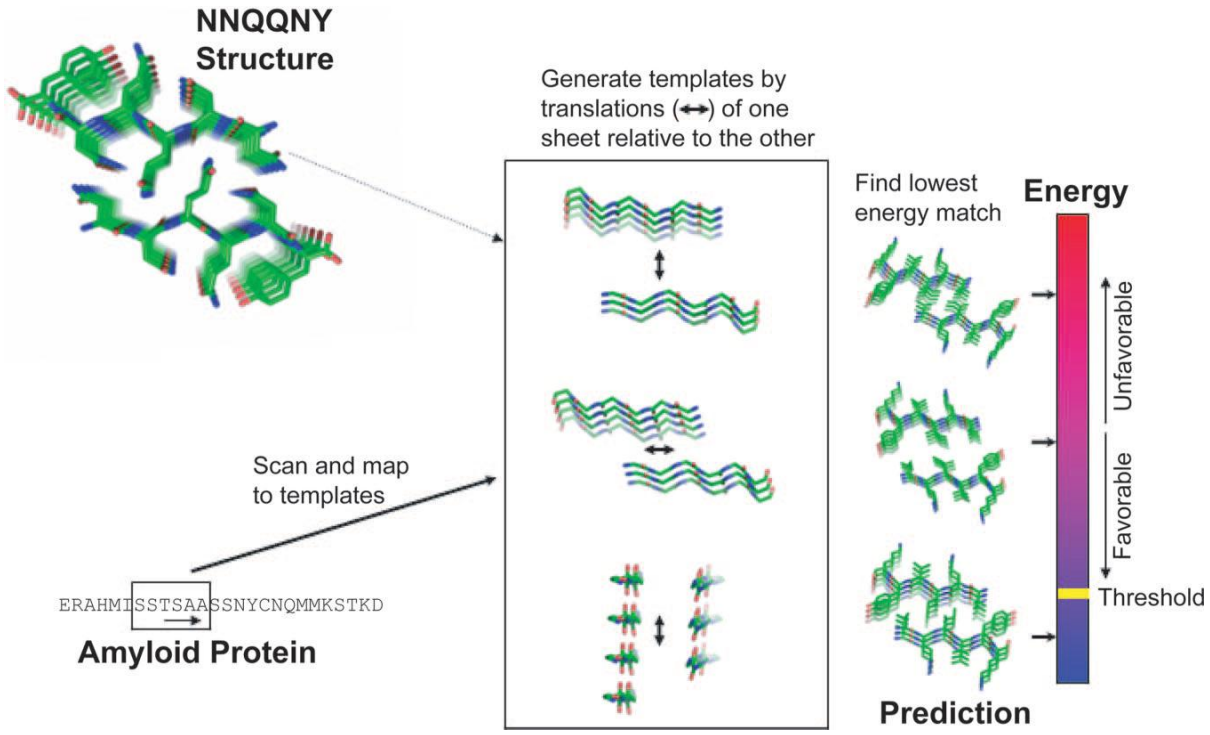
- Concept of β- aggregates
- Packing density (contact sites) – **FoldAmyloid** (Galzitskaya OV, Garbuzynskiy SO, Lobanov MY ,*PLoS Comp.Biol. 2006)*
- Concept of α-β switch (conversion from α-helix to β-sheet)
- Stability of supersecondary structure eg. **Profile 3D**

Statistical methods

- **Waltz** (Maurer-Stroh 2010) – frequency characteristics
- Bayess, Decission Tree (David 2010), bacterial strands

- **Pafig** (Tian 2009), SVM, machine learning based on AAindex features (**Erroneous dataset!**)

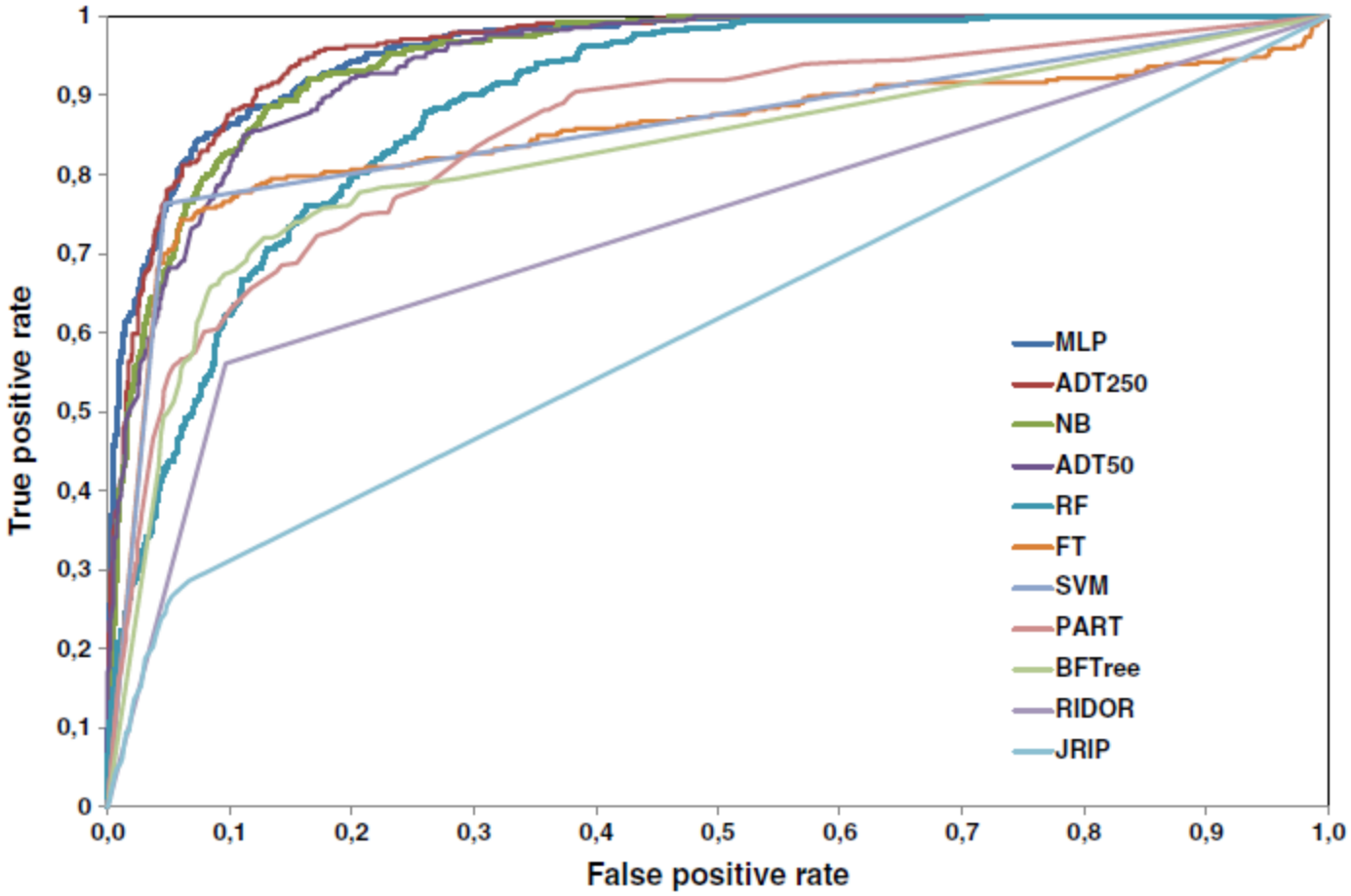
Machine learning methods for hexapeptides



Simplified **3D profile** method gives 93.5% of identity with the original method of **ZipperDB**

Thompson et al. PNAS 2006

Machine learning methods



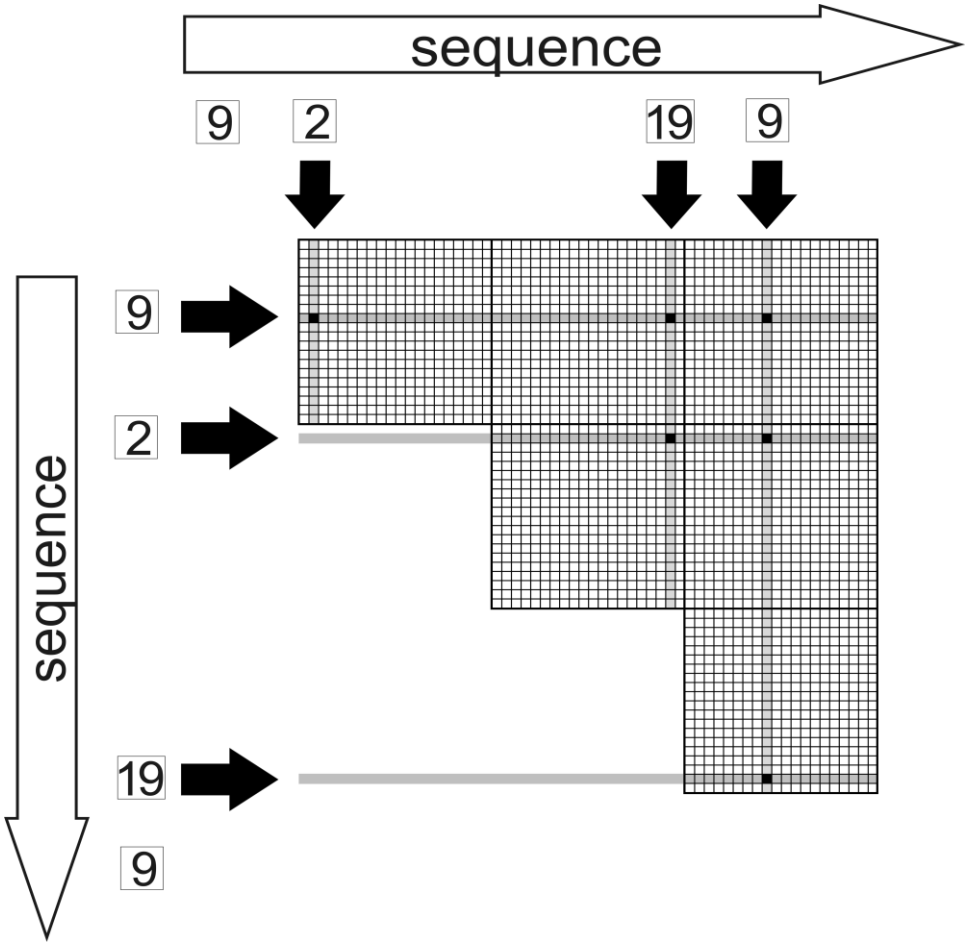
Classification efficiency

- **18–20 CPU-hours** full profile 3D / 1 hexapeptide
- **0.5 CPU-hours** simplified profile 3D
- **Seconds** machine learning

J. Stanislawski, M. Kotulska, O. Unold, Machine learning methods can replace 3D profile method in classification of amyloidogenic hexapeptides, **BMC Bioinformatics** 2013

Recognition of hot-spots based on site specific aminoacid pairwise co-occurrence

Recognition of hot-spots based on site specific aminoacid pairwise co-occurrence

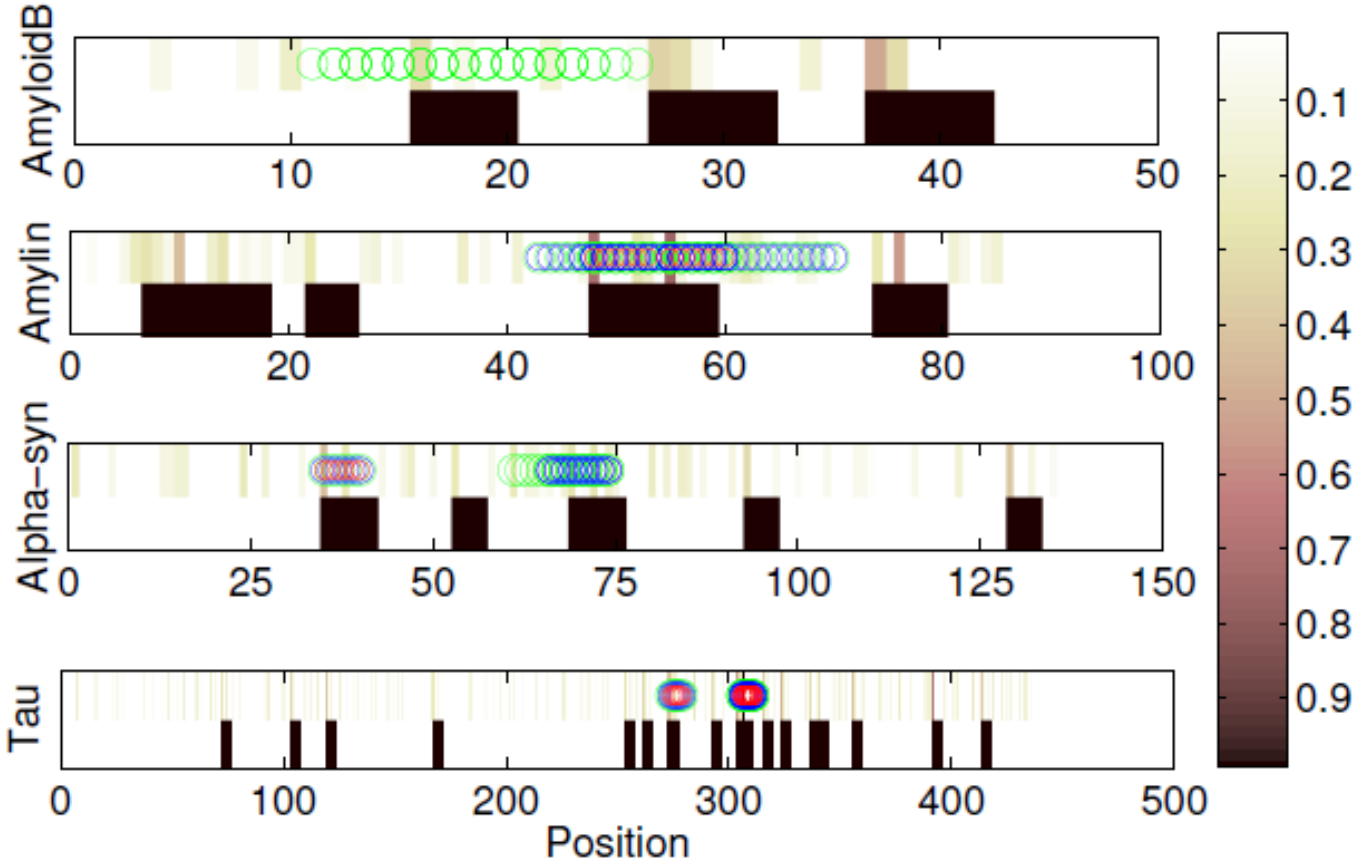


		2	3	4	5
TG	1-	F-F, F-T	F-A, I-L	F-E, Y-Y,I-I	F-D, Y-Y,
	2-		F-A, F-L	F-E, L-Y, V-I	F-D, V-V
	3-			A-E, L-Y, L-I	A-D, Q-V, T-V
	4-				E-D, I-V
AmyHeX	1-	...-T, S-..	V-A, F-V, N-V, Q-V	...-I	S-I, V-A,
	2-		I-V, E-A, ...-V	E-L	E-A, F-I
	3-			A-L, V-T, V-I	V-I
	4-				I-I, L-A
Waltz	1-	S-F, Q-F, G-V, Y-V	Y-N, Q-I	Q-I, S-I, I-I, F-I, G-Y,	G-F, F-Q, Q-S
	2-		V-L, V-N, F-L, F-N,	F-I, V-Y, V-F, V-I	V-F, V-V, V-I
	3-			L-I, N-Y, E-Y	L-V, N-F, L-F
	4-				I-V, I-M, I-F, F-S
All + sup35	1-	S-F, G-V	S-V, Y-N	S-I, Y-Y, L-Y, I-I	S-I, Y-Q
	2-		F-L, T-V	V-I, T-I, V-Y, F-I	R-R, T-I, V-I
	3-			V-I, L-I, N-Y	V-I, L-V
	4-				I-I, I-F, Y-R, Y-I
3D profile	1-	A-I, T-V	S-V, T-I	S-I, V-I, T-I	S-I, V-F
	2-		T-V, V-I,	T-I, V-I, L-I	T-I,
	3-			V-I, I-I, V-N	V-I, L-I, V-V
	4-				I-I, T-C

Correlated pairs of aminoacids

The result of machine learning on each of 4 training datasets and their combination with the window of length 5. The correlated pairs of aminoacids, increasing the chance of amyloidogenicity, are shown with their most probable locations.

Finding hot spots in full length proteins

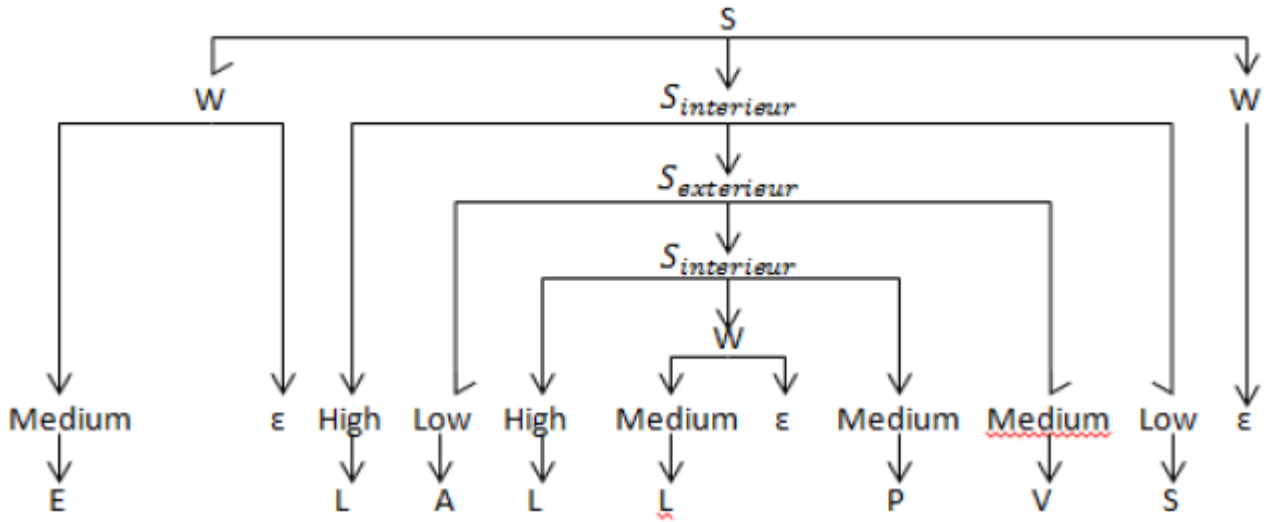


Black blocks location of amyloidogenic segments with $w_l = 0.14$, (Sp=60% on Waltz set). The brown - different w_l values assumed. The circles experimentally by different groups, working on protein fragments of various lengths (green – above 16, blue -11, red - 7).

FISH Amyloid

Gasior P, Kotulska M, FISH Amyloid – a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids, **BMC Bioinformatics** 2014, in print

Exemplary grammar (hydrophobicity)

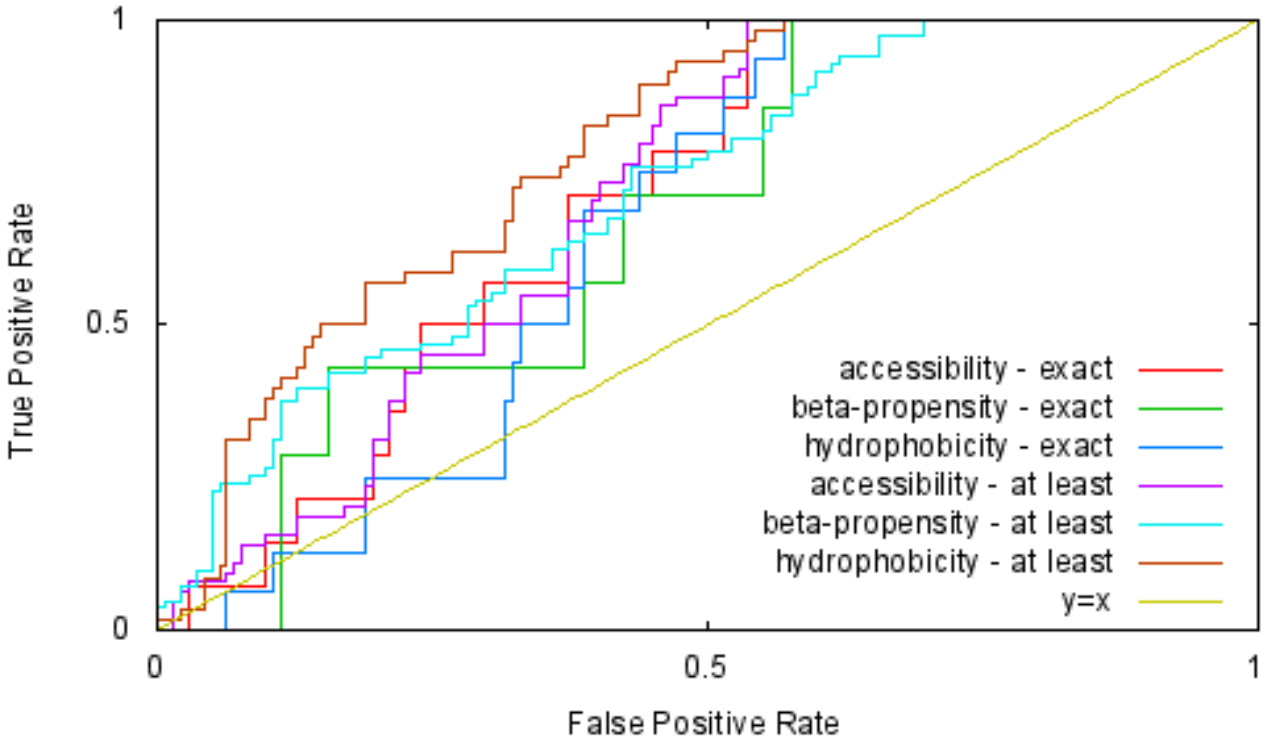


True dependencies according to DSSP

Dependencies found by the grammar

Florence Thirion MS thesis 2013

Recognition of double strands with PCFGs
The grammar designed for β-barrels



ROC curves of grammars trained on Sawaya dataset (experimental double strands) and tested on double Waltz and double Waltz negative

Grammar scheme for barrels: Waldispühl J, Berger B, Clote P, Steyaert JM (2006): Predicting Transmembrane beta-barrels and interstrand residue interactions from sequence. Proteins, 65, 61-74.

SUMMARY

- Amyloidogenicity may be attributed to short segments in sequences, called hot-spots
- Very limited experimenta data
- Hot spots can be recognized with various computational methods
- Transmembrane channels, probably formed by amyloidogenic segments can be a modeling goal

Publikacje

Stanislawski Jerzy, Kotulska Małgorzata, Unold Olgierd: Machine learning methods can replace 3D profile method in classification of amyloidogenic hexapeptides / Jerzy Stanislawski, Malgorzata Kotulska, Olgierd Unold. ***BMC Bioinformatics*** [Dokument elektroniczny]. 2013, vol. 14, [art.] 21,

Kotulska Małgorzata, Unold Olgierd: On the amyloid datasets used for training PAFIG--how (not) to extend the experimental dataset of hexapeptides / Małgorzata Kotulska, Olgierd Unold. ***BMC Bioinformatics*** 2013, vol. 14, [art.] 351.

Gasior Paweł, Kotulska Małgorzata, FISH Amyloid – a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids, **BMC Bioinformatics** 2014, in print

Dyrka Witold, Nebel Jean-Christophe, Kotulska Małgorzata: Probabilistic grammatical model for helix-helix contact site classification / Witold Dyrka, Jean-Christophe Nebel, Małgorzata Kotulska. ***Algorithms Mol Biol.*** 2013,

Dyrka Witold, Bartuzel Maciej, Kotulska Małgorzata, Optimization of 3D Poisson-Nernst-Planck model for fast evaluation of diverse protein channels, **PROTEINS: Structure, Function, and Bioinformatics**. 81(10):1802-22, 2013

Woźniak Paweł P., Kotulska Małgorzata: Characteristics of contact sites with regard to protein topology, [W:] Postępy Inżynierii Biomedycznej, pod red. Lucyna Leniowska, Zbigniew Nawrat, Rzeszów, Uniwersytet Rzeszowski, 2013. s. 81-90, ISBN 978-83-63151-02-7

Woźniak Paweł P.: Influence of the MSA source and the contact site definition on the contact sites prediction accuracy with correlated mutation algorithms, [W:] 11th Students' Science Conference, 03-06 październik 2013 - Będlewo. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2013. s. 365-370

Kurczyńska Monika: Structural, energy and functional changes in different conformations of the potassium channel KcsA, Postępy Inżynierii Biomedycznej, pod red. Lucyna Leniowska, Zbigniew Nawrat, Rzeszów, Uniwersytet Rzeszowski, 2013. s. 101-112, ISBN 978-83-63151-02-7

Kurczyńska Monika, Cichowski Wojciech: Ion transport simulation of transmembrane protein, Biocybernetyka i inżynieria biomedyczna [Dokument elektroniczny]: XVIII krajowa konferencja naukowa, 10-12 października 2013 - Gdańsk / red. Adam Bujnowicz, Jerzy Wtorek. [Gdańsk : Wydział Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej, 2013]. s. 1-9

Monika Kurczyńska: Diffusion coefficient and eletrostatic potential in brownian dynamics simulations, [W:] 11th Students' Science Conference, 03-06 październik 2013 - Będlewo. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2013. s. 475-480.

Cichowski Wojciech, Kurczyńska Monika: Ion Current Analyser (ICA) - narzędzie do analizy wyników symulacji transportu jonów przez kanały jonowe, Otwarta Innowacja. 2013, nr 1/2, s. 16-25.

Konopka Bogumił M.: A procedure for automated contact map-based reconstruction of protein structural models, Postępy Inżynierii Biomedycznej, pod red. Lucyna Leniowska, Zbigniew Nawrat, Rzeszów, Uniwersytet Rzeszowski, 2013. s. 63-73, ISBN 978-83-63151-02-7

Grupa Biofizyki i Bioinformatyki Nanoporów

Witold Dyrka

Małgorzata Kotulska

Paweł Woźniak

Bogumil Konopka

Monika Kurczyńska

Joanna Weżgowiec

AKN BioNanopor

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

Probabilistic Context Free Grammars (PCFG)

Only physicochemical properties of residues

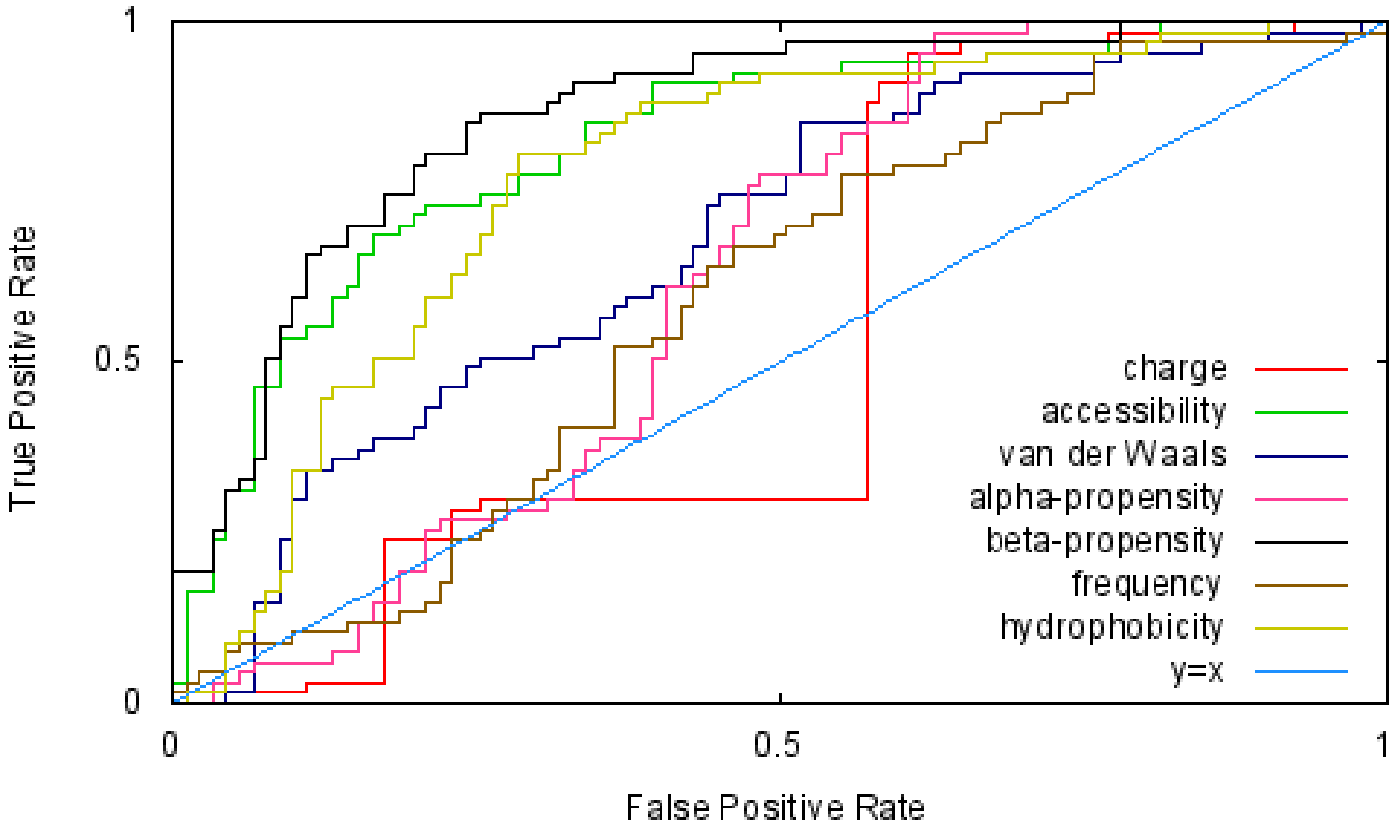
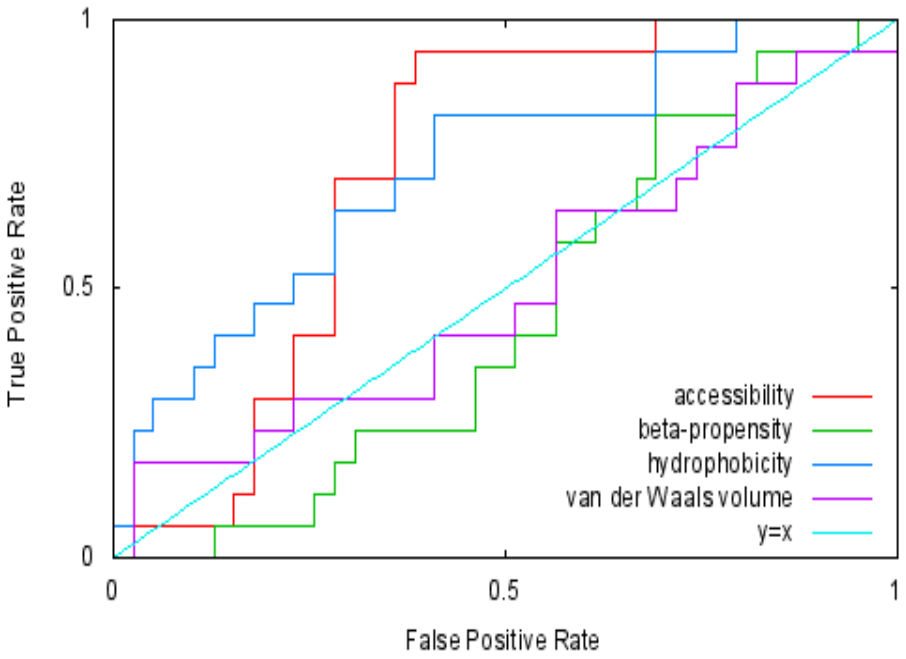
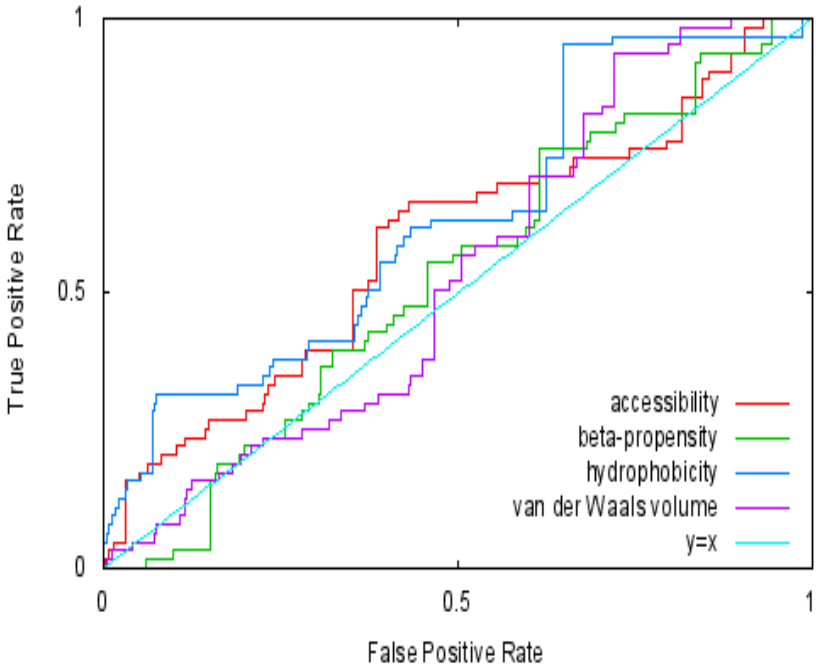


Figure 5.1.1: ROC curves of grammars trained on **Waltz positive** dataset and tested on **AmylHex** datasets.

Beta-aggregates versus amyloids



ROC curves of the grammars trained on Waltz and tested on Tango datasets (beta-aggregates versus amyloids).



ROC curves of the grammars trained on Waltz and tested on Tango datasets (beta-aggregates versus amyloids), when using a **scanning window of 6 residues**