Predykcja białek amyloidogennych

Projekt badawczy Doktoranckiego Koła Naukowego Bioinformatyki

Michał Burdukiewicz, Przemysław Gagat

1 Założenia projektu badawczego

Amyloidy to zróżnicowana grupa białek mogących tworzyć zazwyczaj cytotoksyczne kompleksy (Fändrich, 2012). Agregaty amyloid są przyczyną różnych zaburzeń (m.in. choroby Alzheimera, Creutzfelda-Jacoba). Ustalono, że mimo podobieństw w procesie agregacji, białka amyloidogenne są zróżnicowane pod względem długości i składu aminokwasowego. Wszystkie jednak zawierają tzw. *hot-spots*, krótkie sekwencje aminokwasów, które pełnią kluczową rolę w procesie formowania kompleksów amyloid (Breydo & Uversky, 2015).

Celem badań jest utworzenie probabilistycznego modelu *hot-spots*. Opracowany model zostanie zweryfikowany poprzez analizę znanych sekwencji amyloidogennych.

2 Metody

Głównym narzędziem wykorzystywanym w projekcie badawczym jest pakiet *biogram* przeznaczony do analizy n-gramowej. n-gramy (k-mery, k-tuple) to wektory o długości n zawierające znaki z sekwencji wejściowych. Pierwotnie analiza n-gramów rozwijana była na potrzeby analizy języka naturalnego, ale ma również zastosowania w genomice (Fang et al., 2011), transkryptomice (Wang et al., 2014) i proteomice (Guo et al., 2014).

W przewidzianych analizach wykorzystane zostaną zarówno ciągłe jak i nieciągłe n-gramy. Uzyskane zliczenia n-gramów będą przefiltrowane w celu odrzucenia mniej informatywnych n-gramów, a następnie wykorzystane do uczenia lasu losowego (Liaw & Wiener, 2002).

3 Stan badań

Wstępna n-gramowa analiza sekwencji białek uzyskanych z bazy AmyLoadWozniak & Kotulska (2015) została przeprowadzona używając pakietu *biogram* (Burdukiewicz et al., 2015). Stworzony model nazwany roboczo AmyloGram został porównany z najlepszymi istniejącymi predyktorami amoyloidogenności.

Porównanie programów przewidujących amyloidogenność.

Nazwa programu	AUC	Czułość	Specyficzność
AmyloGram	0.8426	0.8054	0.7222
PASTA2 (Walsh et al., 2014)	0.7920	0.7248	0.8593
FoldAmyloid (Garbuzynskiy et al., 2010)	0.7351	0.7517	0.7185

AUC (Area under Curve) to jedna z najpopularniejszych miar jakości klasyfiaktora i zawiera się między 1 (idealna dobra klasyfikacja) i 0 (idealnie zła klasyfikacja). Wartość 0.5 jest typowa dla idealnie losowej predykcji. AmyloGram uzyskując AUC = 0.84 pod względem jakości predykcji przewyższa istniejące programy przewidujące amyloidy.

4 Planowane wydatki

Łączny koszt projektu badawczego to 16 560 zł.

4.1 Ulepszenia istniejącej infrastruktury

Realizacja zaplanowanych zadań badawczych wymaga modyfikacji dostępnego wyposażenia: zakupu nowych dysków twardych oraz baterii do UPS.

Kosztorys ulepszeń istniejącej infrastruktury.

Nazwa	Cena (szt.)	Liczba	Łączna cena
Dysk twardy WD Red Sata 3	1150 zł	6	6900 zł
Bateria APC RBC7	830 zł	2	1660 zł
		Łącznie:	8560 zł

4.2 Wyjazdy zagraniczne

Wyniki badań zostaną zaprezentowane podczas 15th European Conference on Computational Biology (3-7 września 2016, Haga, Holandia). Dofinansowanie umożliwi większej liczbie członków Koła aktywny udział w konferencji i zaprezentowanie nie tylko wyników realizacji zadań badawczych postawionych w tym wniosku, ale również postępów w pracach doktorskich.

Kosztorys wyjazdów zagranicznych.

Nazwa	Cena (szt.)	Liczba	Łączna cena
Dofinansowanie wyjazdu	2000 zł	4	8000 zł
		Łącznie:	8000 zł

5 Współpraca

Projekt jest realizowany przy współpracy z profesor Małgorzatą Kotulską (Politechnika Wrocławska), kuratorem bazy AmyLoad i ekspertem w zakresie analizy sekwencji amyloidogennych, oraz Piotrem Sobczykiem (Politechnika Wrocławska), współtwórcą pakietu *biogram*.

Literatura

Breydo, L., & Uversky, V. N. (2015, July). Structural, morphological, and functional diversity of amyloid oligomers. *FEBS letters*. doi: 10.1016/j.febslet.2015.07.013

Burdukiewicz, M., Sobczyk, P., & Lauber, C. (2015). *biogram: analysis of biological sequences using n-grams*. Retrieved from `http://CRAN.R-project.org/package=biogram` (R package version 1.2)

Fang, Y.-C., Lai, P.-T., Dai, H.-J., & Hsu, W.-L. (2011). Meinfotext 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC Bioinformatics*, *12*(1), 471. Retrieved from `http://www.biomedcentral.com/1471-2105/12/471` doi: 10.1186/1471-2105-12-471

Fändrich, M. (2012, August). Oligomeric Intermediates in Amyloid Formation: Structure Determination and Mechanisms of Toxicity. *Journal of Molecular Biology*, *421*(4–5), 427–440. Retrieved 2015-07-24, from `http://www.sciencedirect.com/science/article/pii/S0022283612000277` doi: 10.1016/j.jmb.2012.01.006

Garbuzynskiy, S. O., Lobanov, M. Y., & Galzitskaya, O. V. (2010). Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, *26*(3), 326-332.

Guo, S.-H., Deng, E.-Z., Xu, L.-Q., Ding, H., Lin, H., Chen, W., & Chou, K.-C. (2014). inuc-pseknc: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, *30*(11), 1522-1529. Retrieved from `http://bioinformatics.oxfordjournals.org/content/30/11/1522.abstract` doi: 10.1093/bioinformatics/btu083

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from `http://CRAN.R-project.org/doc/Rnews/`

Walsh, I., Seno, F., Tosatto, S. C. E., & Trovato, A. (2014, July). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, *42*(W1), W301–W307. Retrieved 2015-07-24, from `http://nar.oxfordjournals.org/content/42/W1/W301` doi: 10.1093/nar/gku399

Wang, Y., Liu, L., Chen, L., Chen, T., & Sun, F. (2014, 01). Comparison of metatranscriptomic samples based on ¡italic¿k-¡/italic¿tuple frequencies. *PLoS ONE*, *9*(1), e84348. Retrieved from

Wozniak, P. P., & Kotulska, M. (2015, June). AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*. doi: 10.1093/bioinformatics/btv375