Moduł bioinformatyczny bazy metanogenów

Projekt badawczy Doktoranckiego Koła Naukowego Bioinformatyki

Michał Burdukiewicz, Przemysław Gagat

1 Założenia i cel projektu badawczego

Metanogeny to grupa archebakterii, które w procesie oddychania produkują metan. Ze względu na istotną rolę w różnych środowiskach beztlenowych, takich jak przewód pokarmowy czy gleba, metanogeny są obiecującym obiektem badań naukowych, zarówno pod względem zastosowań przemysłowych jak i analiz filogenetycznych.

Baza Methanogens (Jabłoński et al., 2015) (`http://metanogen.biotech.uni.wroc.pl/`) jest największym istniejącym zbiorem danych dotyczących warunków hodowlanych metanogenów. Nie zawiera jednak dodatkowych informacji, które umożliwiłyby przeprowadzenie analiz filogenetycznych.

Zadaniem badawczym jest dodanie do bazy Methanogens zestawu narzędzi umożliwiających bioinformatyczną analizę metanogenów. Wymaga to zarówno stworzenia narzędzi dostępnych jako web servers, a także uzupełnienia bazy o dane istotne z punktu widzenia biologii obliczeniowej. Dodatkowo, ponieważ obecne narzędzia wizualizacyjne bazy są w tej chwili dość ograniczone, zostaną poszerzone w trakcie trwania projektu.

2 Obecny stan badań

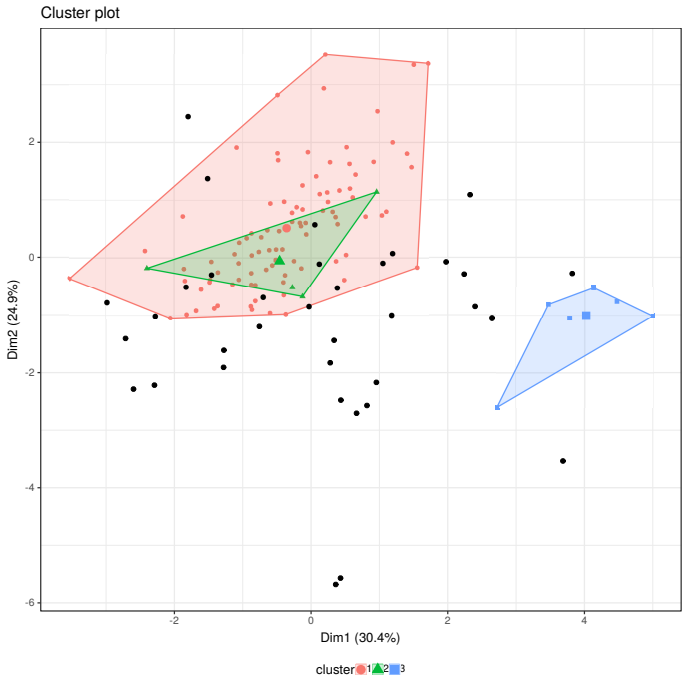
Trzonem informacji zgromadzonych w bazie Methanogens są wyczerpujące dane na temat warunków hodowlanych metanogenów. Informacje te zostały pozyskane ręcznie z różnych źródeł i wystandaryzowane. Dodatkową zaletą bazy Methanogens są narzędzia wizualizacyjne pozwalające na interaktywne porównywanie ze sobą różnych metanogenów pod względem ich cech fenotypowych.

Jakkolwiek istniejące funkcje wizualizacyjne są bardzo pomocne, to obecnie brak narzędzi umożliwiających jednoczesną wizualizację więcej niż dwóch zmiennych. To ograniczenie wydaje się szczególnie dotkliwe biorąc pod uwagę wielowymiarowość danych zgromadzonych w bazie Methanogenes.

Istotnym brakiem bazy Methanogenes jest nieobecność w niej sekwencji nukleinowych i aminokwasowych metanogenów. Taka informacja jest cenna we wszelkiego rodzaju metodach biologii obliczeniowej, w tym badaniach filogenetycznych. Wzbogacenie bazy metanogenów o odpowiednie sekwencje nukleino-

we i aminokwasowe z pewnością uczyni ją przydatniejszą dla szerszego grona użytkowników pod warunkiem dodania do bazy oprogramowania umożliwiające efektywne przeszukiwanie zbioru sekwencji.

3 Metody



Rys. 1: Przykładowa analiza skupień danych zebranych w bazie Methanogens.

3.1 Pozyskanie i przeszukiwanie sekwencji metanogenów

Sekwencje nukleotydowe i aminokwasowe metanogenów zostaną pozyskane z baz NCBI, odpowiednio Nucletide i Protein. W celu automatyzacji procesu, sekwencje zostaną pozyskane z użyciem EUtils API udostępnionym przez NCBI (`http://www.ncbi.nlm.nih.gov/books/NBK25500/`).

W celu zagwarantowanie większej jakości pozyskanych rekordów, sekwencje zostaną poddane ręcznej kuracji.

Zgromadzony zbiór sekwencji będzie można przeszukiwać dzięki lokalnej instancji algorytmu Blast (Altschul et al., 1990), który zostanie udostępniony użytkownikom bazy.

3.2 Analiza częstościowa

Baza zostanie uzupełniona o dodatkowy moduł, web server odpowiedzialny za wizualizację sekwencji za pomocą analizy częstościowej. Typowa analiza występowania pojedynczych reszt nukletydowych lub aminokwasowych zostanie rozszerzona o analizę n-gramów, ciągłych lub nieciągłych podsekwencji długości n. Obliczenia zostaną wykonane dzięki narzędziom z pakietu *biogram* (Burdukiewicz et al., 2017), a web server zostanie zrealizowany w technologii *shiny*.

3.3 Analiza skupień

Web server dedykowany analizie skupień danych zgromadzonych w bazie zostanie oparty o algorytm klasteryzacji gęstościowej (Ester et al., 1996). Metodę tę wybraną ze względu na odporność na zaszumienie danych i szybkość działania. Analiza będzie w pełni interaktywna, co oznacza, że użytkownik będzie mógł wybrać interesujące go cechy, a następnie określić parametry działania algorytmu. Wynikiem analizy nie będą tylko dane liczbowe, ale również graficzna reprezentacja znalezionych klastrów 1.

4 Czas realizacji projektu

Projekt będzie realizowany w okresie 1.06.2017 do 30.11.2017.

5 Planowane wydatki

Łączny koszt projektu badawczego to 9 560,00 zł. Ze względu na możliwość zmiany cen poszczególnych produktów, Doktoranckie Koło Naukowe Bioinformatyki zwraca się o przyznanie 10 000,00 zł.

Tab. 1: Kosztorys projektu badawczego.

Nazwa	Koszt
Akcesoria niezbędne w realizacji zadań badawczych	5560,00 zł
Wyjazdy konferencyjne	4 000,00 zł
Łącznie	9 560,00 zł

5.1 Akcesoria niezbędne w realizacji zadań badawczych

Właściwe zrealizowanie projektu badawczego wymaga dokupienie akcesoriów i sprzętu komputerowego (Tab. 2), takich jak klawiatury i monitor niezbędne do wykorzystania stanowisk komputerowych udostępnionych przez Zakład Genomiki. Konieczna jest również rozbudowa tych stanowisk poprzez dokupienie pamięci DDR. Wykonanie części zadań badawczych nie byłaby możliwa gdyby nie komputery przenośne udostępnione członkom Koła przez Zakład Genomiki.

Tab. 2: Koszty akcesoriów niezbędnych w realizacji zadań badawczych.

Nazwa	Koszt jednostkowy	Liczba sztuk	Łączny koszt
pamięć DDR3 8GB	250	12	3000,00 zł
Monitor 27'	1000	1	1000,00 zł
kabel DVI-D 1m	200	1	200,00 zł
kabel HDMI 5m	160	1	160,00 zł
klawiatura USB	400	2	800,00 zł
torba na laptopa	200	1	200,00 zł
plecak na laptopa	200	1	200,00 zł
			5 560,00 zł

Bezpieczny transport otrzymanego sprzętu wymaga zakupu specjalnych plecaków na laptopy.

5.2 Wyjazdy konferencyjne

Utworzone web servery zostaną zaprezentowane podczas konferencji International Society for Computational Biology w Pradze (21-26.07.2017). Dofinansowane zostaną dwa wyjazdy na konferencję.

Tab. 3: Kosztorys wyjazdów konferencyjnych.

Nazwa	Cena (szt.)	Liczba	Łączna cena
Dofinansowanie wyjazdu	2 000 zł	2	4 000 zł
		Łącznie:	4 000,00 zł

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990, October). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Burdukiewicz, M., Sobczyk, P., & Lauber, C. (2017). *biogram: analysis of biological sequences using n-grams.* Retrieved from `http://CRAN.R-project.org/package=biogram` (R package version 1.5)

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Portland, Oregon: AAAI Press. Retrieved 2017-05-18, from `http://dl.acm.org/citation.cfm?id=3001460.3001507`

Jabłoński, S., Rodowicz, P., & Łukaszewicz, M. (2015, April). Methanogenic archaea database containing physiological and biochemical characteristics. *International Journal of Systematic and Evolutionary Microbiology*, *65*(Pt 4), 1360–1368. doi: 10.1099/ijs.0.000065