n-gramowa analiza białek metanogenów

Projekt badawczy Doktoranckiego Koła Naukowego Bioinformatyki

Michał Burdukiewicz, Przemysław Gagat

1 Założenia projektu badawczego

Metanogeny to zróżnicowana grupa archeonów

Tab. 1: Przykładowy skrócony alfabet aminokwasowy. Sekwencja ADPH w tym alfabecie zostanie zapisana jako 1335.

Numer grupy	Aminokwasy
1	A, G
2	C
3	D, E, K, N, P, Q, R, S, T
4	F, I, L, M, V, W, Y
5	H

Nasze wstępne wyniki sugerują, że w przypadku niektórych problemów przewidywania właściwości białek, zastosowanie skróconego alfabetu aminokwasowego może znacząco ulepszyć precyzję predykcji. Jednakże taki alfabet musi zostać wybrany spośród wielu potencjalnych skróconych alfabetów. Liczba wszystkich możliwych skróconych alfabetów aminokwasowych n_s zawierających k grup jest wyrażona poprzez:

$$n_s = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^{20}$$

Nawet przy założeniu, że optymalna wielkość skróconego alfabetu jest już znana, to liczba potencjalnych możliwości jest bardzo duża (np. wszystkich alfabetów sześcioelementowych jest 4.306079×10^{12}).

2 Obecny stan badań

Baza Methanogenes jest największym zbiorem informacji na temat metanogenów. Jej zaletą, oprócz kompletności zgromadzonych danych, są narzędzia wi-

zualizacyjne 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 doktliwe biorąc pod uwagę ogrom informacji 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 nukleinowe i aminokwasowe z pewnością uczyni ją przydatniejszą dla szerszego grona użytkowników pod warunkiem dodania do bazy oprogramowania umożliwiające efektywne przeszukiwanie zbioru sekwencji.

Sekwencje poszczególnych gatunków metanogenów mogą zostać graficznie opisane za pomocą częstości pojedynczych reszt, aminokwasów lub nukleotydow. Tego rodzaju dane można wykorzystać do analizy skupień lub skonfrontować je z informacjami fenotypowymi, już obecnymi w bazie.

3 Cel badań

Głównym zadaniem badawczym jest opracowanie nowych metod wizualizacji danych w bazie Methanogenes. Ponieważ baza nie zawiera jeszcze wielu informacji istotnych z punktu widzenia biologii obliczeniowej, zostaną one dodane w trakcie trwania projektu.

4 Metody

4.1 Pozyskanie i przeszukiwanie sekwencji metanogenów

Sekwencje nukleotydowe i aminokwasowe metanogenów zostaną pozyskane z baz NCBI, odpowiednio Nucletide i Protein. W celu automatyzacji procesu, sekwencje zamiast ręcznie 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 XXX.

4.2 Analiza częstościowa

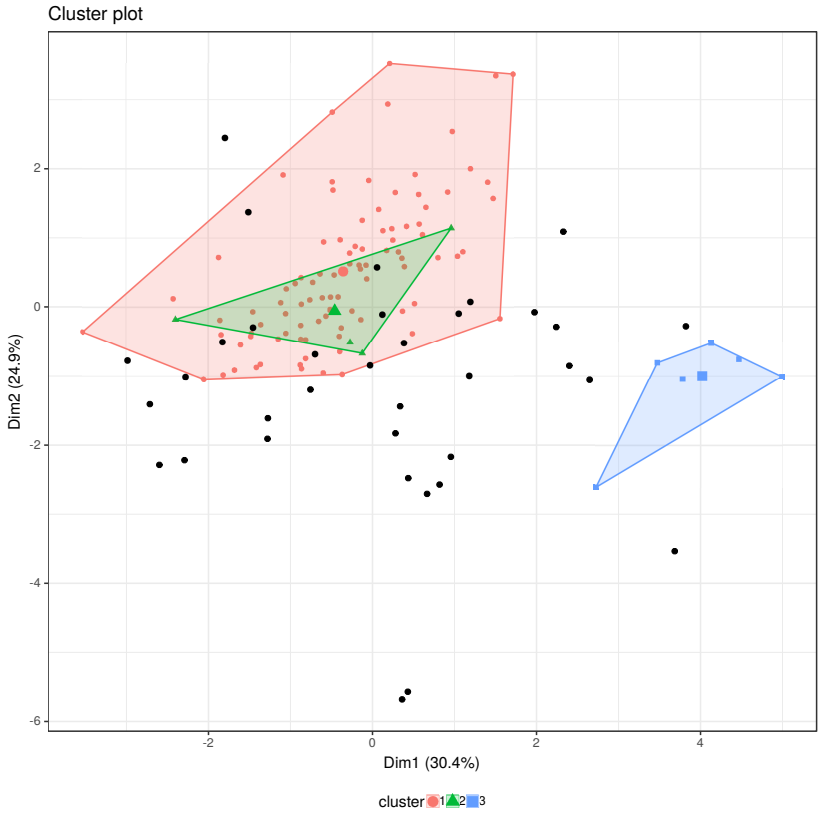
DO bazy zostanie dodany web server odpowiedzialny za wizualizację sekwencji za pomocą analizy częstościowej. Zwyczajowa analiza występowania pojedynczych reszt nukletydowych lub aminokwasowych zostanie rozszerzona o analizę

n-gramów, ciągłych lub nieciągłych podsekwencji długości n. Obliczenia zostaną wykonane dzięki narzędziom z pakietu *biogram* XXX, a web server zostanie zrealizowany w technologii *shiny*.

4.3 Analiza skupień

Web server dedykowany analizie skupień danych zgromadzonych w bazie zostanie oparty o algorytm klasteryzacji gęstościowej (Ester et al. (1996) XXX). 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.



Rys. 1: Wyniki PCA dla częstości aminokwasów w peptydach sygnałowych i dojrzałych białkach *Plasmodiidae* oraz innych eukariontów.

5 Czas realizacji projektu

Projekt będzie realizowany w okresie 1.06.2017 do 30.11.2017.

6 Planowane wydatki

Łączny koszt projektu badawczego to 29 796,00 zł.

Tab. 2: Kosztorys projektu badawczego.

Nazwa	Koszt
Akcesoria niezbędne w realizacji zadań badawczych	760,00 zł
Wyjazdy konferencyjne	4 000,00 zł
Łącznie	29 796,00 zł

6.1 Akcesoria niezbędne w realizacji zadań badawczych

Właściwe zrealizowanie projektu badawczego wymaga również dokupienie akcesoriów (Tab. 3), takich jak pamięci USB niezbędne do przenoszenia duzych objętości danych i słuchawki z mikrofonem do prowadzenia rozmów z zagranicznym uczestnikiem projektu. Wykonanie części zadań badawczych nie byłaby możliwa gdyby nie komputery przenośne udostępnione członkom Koła przez Zakład Genomiki. Bezpieczny transport otrzymanego sprzętu wymaga zakupu specjalnych plecaków na laptopy.

6.2 Wyjazdy konferencyjne

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

Tab. 3: Koszty akcesoriów niezbędnych w realizacji zadań badawczych.

Nazwa	Cena (szt.)	Liczba	Łączna cena
Pendrive USB 3.0 - 32 GB	45,00 zł	4	180,00 zł
Słuchawki z mikrofonem Creative	140,00 zł	2	280,00 zł
Plecak na laptopa	150,00 zł	2	300,00 zł
		Łącznie:	760,00 zł

Tab. 4: Kosztorys wyjazdów konferencyjnych.

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