

# Wykorzystanie sztucznych sieci neuronowych i algorytmu genetycznego do badania proteomów

Maciej Sobczyński

Praca doktorska wykonana w Zakładzie Genomiki  
pod kierunkiem dr hab. Pawła Mackiewicza

Skład aminokwasowy białek może być modelowany zarówno przez presję selekcyjną na ich wymogi funkcjonalne i strukturalne, jak i presję mutacyjną modyfikującą bezpośrednio skład nukleotydowy sekwencji kodujących białka. Żeby zbadać jaki jest wpływ tych presji na skład aminokwasowy oraz jak on zależy od pokrewieństwa ewolucyjnego i właściwości ekologicznych prawie 200 mikroorganizmów prokariotycznych zastosowano sieć neuronową (samoorganizujące się mapy) Kohonena. Metodę tą wybrano, aby skład aminokwasowy białek porównywać kontekstowo oraz szukać wzorców/prototypów składów charakterystycznych dla poszczególnych proteomów, które zostały zakodowane w postaci centroidów wyuczonych neuronów.

Wielkość sieci została odpowiednio dobrana do zbioru danych w oparciu o Bayesowskie kryterium informacyjne, minimalizację błędu topologicznego i przestrzennej autokorelacji. W oparciu o skład aminokwasowy na sieci klasyfikowano białka pochodzące z proteomów organizmów wykazujących określone właściwości środowiskowe. Odległości między grupami proteomów były oparte o frakcje ich białek przypisanych do poszczególnych neuronów. Zastosowana metoda okazała się dobrym narzędziem do rozróżniania proteomów opisanych wieloma parametrami.

Przeprowadzone analizy wykazały wyraźny związek składu aminokwasowego z warunkami życia badanych organizmów. Najbardziej odmiennym składem od innych grup ekologicznych charakteryzowały się organizmy żyjące w wysokich temperaturach (hipertermofile) i bytujące wewnątrz komórek innych organizmów (endopasożyty i endosymbionty). Związek składu aminokwasowego z cechami środowiskowymi był bardziej widoczny w sekwencjach regionów międzydomenowych niż domen białkowych, co może sugerować pewien wpływ presji mutacyjnej na ten skład. Jednak zmiana ich składu może być również adaptacją białek do specyficznych warunków środowiska. Pokazano ponadto, że pary gatunków o *mniej*

liczbie wspólnych cech środowiskowych różnią się składem aminokwasowym *bardziej*, niż by to wynikało z ich pokrewieństwa ewolucyjnego (mierzonego liczbą podstawień w sekwencjach RNA).

Natomiast pary gatunków o *większej* liczbie wspólnych cech ekologicznych różnią się składem aminokwasowym *mniej*, niż należałoby tego oczekiwać z ich pokrewieństwa.

W pracy zastosowano także algorytm ewolucyjny, za pomocą którego szukano podzbioru aminokwasów, które najlepiej rozróżniają porównywane zbiory białek. Aby wyeliminować ewentualny wpływ presji selekcyjnej na skład proteomów, a zbadać efekt presji mutacyjnej, wygenerowano sztuczne białka w oparciu o globalny skład nukleotydowy genomów i porównano je z rzeczywistymi białkami. Analiza pokazała, że dla sztucznych proteomów, w przeciwieństwie do rzeczywistych zbiorów, nie jest możliwe wytypowanie aminokwasów, które rozróżniałyby porównywane zbiory białek, gdyż każdy aminokwas pojawiał się w rozwiązaniu algorytmu z tą samą częstością  $\frac{1}{2}$ . Skutkiem tego nie było także możliwe zwiększenie rozróżnienia badanych zbiorów. Dodatkowo wytypowane aminokwasy nie wykazywały korelacji we współwystępowaniu.

Rezultaty przeprowadzonych badań pozwalają stwierdzić, że jest bardzo mało prawdopodobne, aby skład aminokwasowy białek był wynikiem wyłącznie działania presji mutacyjnej. Istotny wpływ odgrywa tutaj presja selekcyjna. Zgodnie z tym wyeliminowanie efektu selekcji, jako istotnej siły kształtującej skład proteomów, dało wyniki nieobserwowane w naturze.