Algorytm wyznaczania *q*-wartości

Mascot jest systemem identyfikacji opartym na modelu statystycznym wykorzystującym empiryczny rozkład częstości występowania jonów fragmentacyjnych o danej masie, pochodzących z peptydów będących wynikiem podziału białek z bazy danych. Podobnie jak większość systemów identyfikacji, grupuje on wyniki w hierarchiczną strukturę, u podstawy której leżą przypisania sekwencji do widm fragmentacyjnych (PSM - *Peptide Spectrum Match*). Pojedynczy peptyd może być reprezentowany przez wiele PSM, pochodzących od jonów o różnym stopniu naładowania lub poddanych sekwencjonowaniu w różnych skanach przebiegu LC-MS/MS. Ostatnim poziomem hierarchii są białka zidentyfikowane na podstawie jednego, lub większej liczby peptydów.

Używaną przez system Mascot miarą jakości przypisania sekwencji do widma fragmentacyjnego jest prawdopodobieństwo p uzyskania obserwowanego dopasowania widm, teoretycznego i eksperymentalnego, w sposób losowy. Dla wygody wyrażane jest w postaci logarytmicznej, jako *score*:

$$score = -10\log(p) \quad .$$

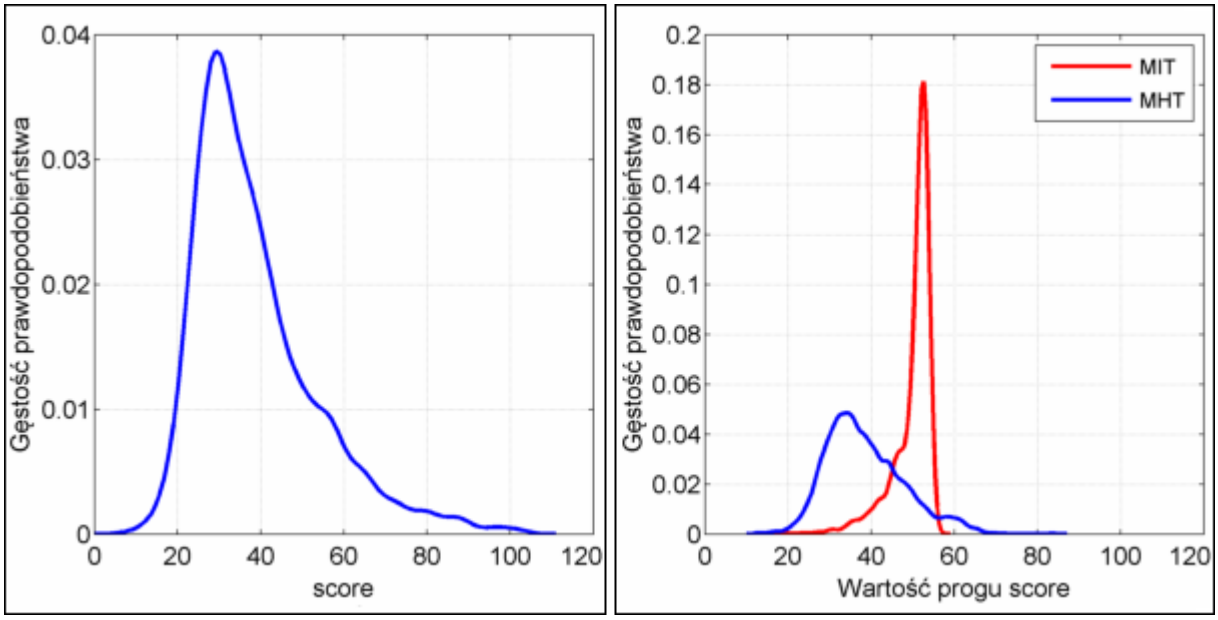
Przykładowy rozkład wartości score zaprezentowany został na rysunku 1.a.

Dla każdego PSM Mascot wyznacza próg istotności wartości *score* określany jako Mascot Identity Threshold (MIT) i dany zależnością:

$$MIT = -10\log\left(20\frac{\alpha}{N}\right) \quad ,$$

gdzie α przyjmuje domyślnie wartość 0,05, a N jest liczbą sekwencji kandydackich, o masach mieszczących się w zadanym przedziale tolerancji wokół masy jonu macierzystego. Występujący we wzorze parametr α ma sens poziomu błędu typu I jedynie przy założeniu pełnej losowości sekwencji peptydów z bazy danych. Ponieważ założenie to w ogólnym przypadku nie jest spełnione, podawany jest również drugi próg, nazywany Mascot Homology Threshold (MHT), będący empiryczną miarą odstępstwa *score* od rozkładu wartości wyznaczonego na podstawie wszystkich sekwencji kandydackich. Niestety, dokładna definicja tego progu nie została przez producenta opublikowana, podobnie zresztą jak i wszelkie szczegóły dotyczące stosowanego modelu statystycznego i sposobu obliczania wartości *score*. Oba progi stosowane były w literaturze jako wartości odniesienia dla *score*, choć często można spotkać się z użyciem arbitralnie wybranej wartości jako kryterium decydującego o wyborze zbioru peptydów.

a) *b)*



Rys. 1. Przykładowe rozkłady wartości: a) miary dopasowania *score*; b) towarzyszących mierze *score* progów istotności *MIT* i *MHT*

W programie MScan stosowane są zmodyfikowane wartości *mscore,* określone jako:

$$mscore = score - MMT = score - min(MIT, MHT) \quad ,$$

gdzie *MIT* i *MHT* są wartościami progów Mascota dla PSM o danej wartości *score*. Miarą jakości identyfikacji białka jest $mscore_B$, wyznaczane na podstawie wartości *mscore* PSM o wartościach pochodzących z danego białka:

$$mscore_B = \sum_{i=1}^{N_B} mscore_i + \overline{MMT} \quad ,$$

gdzie N_B jest liczbą PSM identyfikujących białko, a $\overline{MMT}$ średnią progów istotności użytych do określenia wartości $mscore_i$.

Wybór progu MMT jako odniesienia dla wartości *score* podyktowany jest obserwowanymi rozkładami wartości progów MIT i MHT, których przykłady zostały przestawione na rysunku 1.b. Wartość progu MIT zależy jedynie od liczby sekwencji kandydackich, która dla znacznej części widm jest zbliżona. Skutkuje to wąskim rozkładem progów MIT, sugerującym, że użycie ich do zmodyfikowania *score* tylko w nieznacznym stopniu będzie się różniło od odjęcia arbitralnie przyjętej wartości i w efekcie może nie prowadzić do wzrostu informacji niesionej przez nową miarę. Empiryczny próg MHT, który jest zależny zarówno od widma, jak i sekwencji, charakteryzuje się większą specyficznością, a rozkład jego wartości jest zbliżony w kształcie do rozkładu *score* (rysunek 1.a). Z drugiej jednak strony może on przyjmować nierealistycznie wysokie wartości w szczególnych przypadkach widm, dla których mała liczba sekwencji kandydackich uniemożliwia prawidłową estymację rozkładu *score*.

Miara *mscore* nie jest wykorzystywana w sposób bezpośredni, a służy jedynie do uporządkowania PSM pod względem jakości identyfikacji w celu przypisania im *q*-wartości. Pojęcie *q*-wartości wprowadzone zostało przez Storey'a i Tibshiraniego w kontekście analizy wyników badań ekspresji genów przy użyciu mikromacierzy i jest definiowane jako minimalny FDR, dla którego dana cecha może zostać uznana za istotną statystycznie. Jest więc sposobem przeniesienia właściwości całego zbioru wyników, jaką jest FDR, na poziom pojedynczych cech.

W przypadku wyników sekwencjonowania do wyznaczenia *q*-wartości można wykorzystać metodę przeszukiwania bazy zawierającej zarówno rzeczywiste sekwencje białek, jak i ich odwrócone wersje (baza *target/decoy*). Procedura zaczyna się od posortowania zbioru wszystkich PSM zgodnie z wartościami miary *mscore*. Liczba fałszywie pozytywnych identyfikacji związanych z *i*-tą pozycją posortowanego zbioru szacowana jest jako podwojona liczba PSM o odwróconych sekwencjach znajdujących się na pozycjach nie większych od *i*. Związany z tą pozycją FDR wyznaczany jest jako liczba fałszywych identyfikacji odniesiona do numeru pozycji. Przy znajomości wartości FDR dla kolejnych pozycji zbioru, określenie *q*-wartości sprowadza się do wymuszenia monotoniczności tych pierwszych:

$$q_i = \begin{cases} FDR_i & \text{dla } \; i = N \\ min\left(FDR_i, q_{i+1}\right) & \text{dla } \; i = N-1,\ldots,1 \end{cases},$$

gdzie *N* jest liczebnością zbioru wszystkich PSM.

Filtracja wyników sekwencjonowania odbywa się poprzez odrzucenie wszystkich PSM o *q*-wartościach nie większych od zadanego progu. Dodatkowym warunkiem jest wymóg, aby białka identyfikowane były na podstawie co najmniej dwóch peptydowo różnych sekwencjach.