Redukcja redundancji wyników identyfikacji poprzez klasteryzację białek.

Podczas analizy próbek o nieznanym składzie białkowym zwykle używana jest możliwie jak najobszerniejsza baza danych sekwencji aminokwasowych, co pozwala zminimalizować ryzyko braku identyfikacji na skutek niekompletności tej ostatniej. Dodatkowo, poprawia to skuteczność działania algorytmu identyfikacji w przypadku widm o słabej jakości, na podstawie których nie jest możliwe odtworzenie pełnej sekwencji peptydów. Z drugiej strony, duże bazy, takie jak NCBI, charakteryzują się silną redundancją i częstymi zmianami identyfikatorów odpowiadających najbardziej aktualnym wersjom sekwencji. Skutkuje to niepożądanym wydłużeniem listy białek zidentyfikowanych na podstawie tych samych peptydów oraz utrudnia automatyzację przetwarzania wyników identyfikacji.

Aby uwzględnić wspomniane wyżej zjawisko, MScan umożliwia analizowanie nie tylko poszczególnych białek, ale także całych rodzin o zbliżonej sekwencji. Grupowanie białek w rodziny odbywa się na podstawie wyników aglomeracyjnej klasteryzacji hierarchicznej. Miarą podobieństwa pary białek jest procent identyczności F, równy procentowi identycznych reszt aminokwasowych zajmujących odpowiadające sobie pozycje w ich dopasowanych globalnie sekwencjach. Dopasowanie globalne sekwencji wykonywane jest za pomocą opartego na programowaniu dynamicznym algorytmu Needlemana-Wunsha z afinicznym modelem kar za przerwy i wybraną macierzą substytucji reszt aminokwasowych. Możliwe jest również ominięcie kosztownego obliczeniowo procesu wyznaczania dopasowania sekwencji i użycie przybliżonej miary podobieństwa, opartej na zliczaniu liczby K-merów, czyli subsekwencji aminokwasów o długości K . Podobne do siebie sekwencje będą charakteryzować się większą liczbą wspólnych K-merów i tym samym wyższą wartością miary podobieństwa, która dla dwóch sekwencji S_1 i S_2 o długościach, odpowiednio, L_1 i L_2 dana jest zależnością:

$$F_{Kmer} = \frac{\sum_{\Xi_K} min(N_{\xi 1}, N_{\xi 2})}{min(L_1, L_2) - K + 1} \quad ,$$

gdzie Ξ_K oznacza zbiór wszystkich K-merów o długości K, a $N_{\xi 1}$ i $N_{\xi 2}$ to liczby wystąpień K-meru ξ w sekwencjach S_1 i S_2.