Abstrakt

Tandemová hmotnostní spektrometrie je známá metoda pro identifikaci proteinových a peptidových sekvencí ze vzorků biologického materiálu. Hmotnostní spektrometr generuje desetitisíce spekter, která musí být následně anotována peptidovými sekvencemi. Za tímto účelem lze využít podobnostní vyhledávání v databázích teoretických spekter generovaných z databází známých proteinových sekvencí. Vzhledem k tomu, že objem těchto databází každoročně narůstá téměř exponenciálním tempem, je zapotřebí hledat nové způsoby pro jejich indexování. V této práci se zaměřujeme na využití (ne)metrických přístupových metod jako databázových indexů pro rychlé a aproximativní podobnostní vyhledávání v databázích spekter. Navržená metoda identifikace peptidových sekvencí dosahuje více než 100-násobného zrychlení oproti sekvenčnímu průchodu celé databáze, přičemž je správně anotováno přes 90% spekter. V současnosti je metoda vhodná zejména pro malé směsi proteinů. Pro komplexní směsi proteinů využíváme indexovací metodu založenou na prekurzorovém hmotnostním filtru, která má při použití s modifikací parametrizované Hausdorffovy vzdálenosti vyšší rychlost i přesnost vyhledávání než běžně používané metody. Navržené metody jsou implementovány v aplikaci SimTandem, kterou lze použít pro dávkové zpracování ve frameworku TOPP založeném na knihovně OpenMS.