

Využití báze znalostí pro predikci protein-protein interakčních míst

Jelínek J.¹, Hoksza D.¹

¹ *Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Malostranské nám. 25, 118 00 Praha, jelinek@ksi.ms.mff.cuni.cz*

Vzhledem k náročnosti zkoumání protein-protein interakcí je snaha predikovat interagující oblasti pomocí počítačových metod. Stávající metody však nemají příliš vysokou úspěšnost – na širěji používaných datasetech se hodnota MCC (Matthewův korelační koeficient) obvykle pohybuje kolem 0,2-0,4. Na druhou stranu se, s ohledem na počet již známých proteinových komplexů, otevírá prostor pro metody využívající báze znalostí. Proto zde uvedeme metodu, která pro každou aminokyselinu zkoumaného proteinu hledá v bázi znalostí podobné záznamy s ohledem na sousední aminokyseliny. Do úvahy se bere jak struktura okolí, tak i vybrané fyzikálně-chemické vlastnosti jednotlivých sousedních aminokyselin (typ aminokyseliny; relativní velikost povrchu dostupná solventu). Získané informace se využívají pro inicializaci metody strojového učení zvané CRF (Conditional Random Field). Bázi znalostí jsme vytvořili na základě cca. 60 000 komplexů z PDB (Protein Data Bank), čímž jsme získali informace o cca. 54 milionech aminokyselin a jejich okolí. Použitelnost přístupu závisí na zvoleném způsobu dobývání znalostí, od kterého se vyžaduje dostatečně rychlé porovnávání jednotlivých struktur nebo efektivní odfiltrování nepodobných struktur. Proto se zaměříme na porovnání výhod a nevýhod dvou rozdílných přístupů: kombinaci relační databáze a specializované grafové knihovny, která je rychlá, ale schopná efektivně hledat pouze přesné shody; a metody fingerprintů, která je pomalejší, ale schopná nalézt i mírně odlišné struktury.