Истраживање образаца у одређивању карактеристика протеина

Улфета Маровац

Беланчевине или протеини су важни биолошки макромолекули полимерне природе (полипептиди), који се састоје од амино киселина и представљају основну градивну јединицу сваке ћелије. У њихов састав улазе 20+3 различите амино киселине због чега се у биолошким базама података представљају као ниске формиране од 23 различита карактера. Протеини се могу класификовати на основу њихове примарне структуре, секундарне структуре, функција које обављају, итд.

Једна од могућих класификација протеина по функцији је према припадности одређеном кластеру ортологих група ЦОГ[1] (Cluster of Orthhologous Groups - COGs). Ова класификација је заснована на претходном поређењу протеина према сличности по примарној структури, која је најчешће последица хомологије, тј. заједничког (еволуционог) порекла. ЦОГ база података је добијена поређењем познатих или предвиђених протеина комплетно секвенционисаних прокариотских (археа и бактеријских) генома и класификацијом према њиховој ортологији. Протеини су класификовани у 25 категорија, које могу бити распоређене у три основне функционалне групе (протеини одговорни за: (1) садржај и обраду информација, (2) ћелијске процесе и (3) метаболизам), или у групу недовољно окарактерисаних протеина. Класификација протеина према припадности одређеној ЦОГ категорији (*KOG* за еукариотске организме) је важна за боље разумевање биолошких процеса, као и различитих патолошких стања код људи и других организама.

У раду је предложен модел за класификацију протеина у ЦОГ категорије на основу аминокиселинских n-грама (ниски дужине n). Скуп података садржи протеинске секвенце генома из 8 различитих таксономских класа бактерија (*Aquificales*, *Bacteroidia*, *Chlamydiales*, *Chlorobia*, *Chloroflexia*, *Cytophagia*, *Deinococci*, *Prochlorales*) за које постоје информације о класификацији по ЦОГ категоријама. Приказана је нова метода заснована на генерализованим системима једначина Булове алгебре, која се користи за издвајање n-грама који карактеришу протеине одговарајуће ЦОГ категорије. Приказаном методом значајно се смањује број n-грама који се обрађују у односу на претходно коришћене методе n-грамске анализе, тако да се добија на уштеди меморијског простора и времена обраде протеина.

До сада познате методе класификације протеина по функционалним категоријама су вршиле поређење сваког новог протеина (коме треба одредити функцију) са скупом свих протеина који су већ класификовани према функцијама ради одређивања групе која садржи протеине који су најсличнији протеину који се класификује. Нова метода за функционалну класификацију протеина одређује ЦОГ категорију протеина без

[1]Користићемо у даљем тексту скраћеницу ЦОГ за кластере ортологих група а не КОГ да не би дошло до мешања термина са класификацијом еукариотских организама (*euKaryote Orthologous Groups-KOG*)

поређења са другим протеинским секвенцама, већ се у протеину траже обрасци (n-грами дужине до 10) који су карактеристични за одговарајућу ЦОГ категорију.
На основу предложене методе релизован је предиктор за класификацију протеина према ЦОГ категоријама. Најмања поузданост предвиђања се задаје као улазни параметар предиктора. При тестирању предиктора постигнути су јако добри резултати са највећом поузданошћу класификације од 99%. Део протеина из скупа изабраних класа прокариота које нису класификоване досадашњим методама је придружен одговарајућим ЦОГ категоријама.
Због својих особина и једноставности конструкције модела, предложена метода може да се примени и на сличним областима у којима се проблем решава преко n-грамске анализе секвенци.
Теореме до којих се дошло решавајући проблем издвајања секвенцијалних образаца представљају значајан допринос теорији генерализованих система Булових једначина. Њима је представљен алгоритам за решавање дисјункције Булових једначина и система Булових неједначина који су до сада били отворени проблеми.
Даљи рад на развоју модела биће усмерен на повећање прецизности модела који се конструише коришћењем протеина из различитих класа, и укључивање могућности за класификацију протеина који се налазе у више од једне ЦОГ категорије.

Кључне речи: карактеристике протеина, класификација, истраживање секвенцијалних образаца, n-грам, Булова алгебра