МАТЕМАТИЧКА ГИМНАЗИЈА

Матурски рад

Из предмета

Програмирање и програмски језик

Предвиђање функција протеина: Примена на MYBL2, ген укључен у настанак и развој рака

Ученик:

Катарина Станковић

Ментори:

др Бранислава Гемовић

проф. Милан Чабаркапа

Београд, јун 2018. године

Садржај

1. Увод

1.1. Зашто предвиђати функције протеина?

Познавање функција протеина, односно процеса у које је протеин укључен, омогућује одређивање мета у поступцима лечења болести.

На примеру рака: када знамо да је протеин Х укључен у процесе који су поремећени у ћелијама рака (процеси везани за деобу ћелија или ћелијску смрт, нпр. пролиферација, апоптоза и слично), онда знамо да **управо тај протеин Х** може бити мета за терапију која ће да нпр. блокира његову генску експресију или да утиче на његове интеракције са другим протеинима.

Такође, различити процеси могу да доведу до настанка рака. Ако се експериментално нађе да је протеин Х мутиран или има промењену генску експресију у ћелијама рака, а знамо да је он укључен у неки ћелијски процес, онда можемо да претпоставимо да је **управо тај процес** кључан за настанак тог типа рака. Тада можемо специјализованим лековима да делујемо на тај тип рака, нпр. антипролиферацијским лековима.

Предвиђање функција протеина је постало неопходно због бројности протеина (20000 хуманих) и њихових функција. Зато је потребно филтрирати функције које ће онда научници да експериментално испитају, уместо да испитују сваку функцију на сваком протеину, узалудно трошећи време и новац.

1.2. Биоинформатичка анализа

Биоинформатика је научна дисциплина која истражује биолошке информације уз примену компјутерских алата. Ова дисциплина обухвата широки спектар истраживања. Укључује методе за прикупљање, организовање, представљање и анализу података. Подаци који се најчешће анализирају биоинформатичким методама су ДНК и протеинске секвенце, структуре макромолекула, као и генске експресије. Циљ биоинформатике је да дође до нових сазнања користећи биолошке податке и рачунарске анализе.

1.3. Циљ истраживања

Циљ овог истраживања је развој програма за екстракцију функција протеина предвиђених помоћу алгоритма претходно развијеног у Институту за нуклеарне науке Винча и примена овог програма на биолошке функције гена MYBL2.

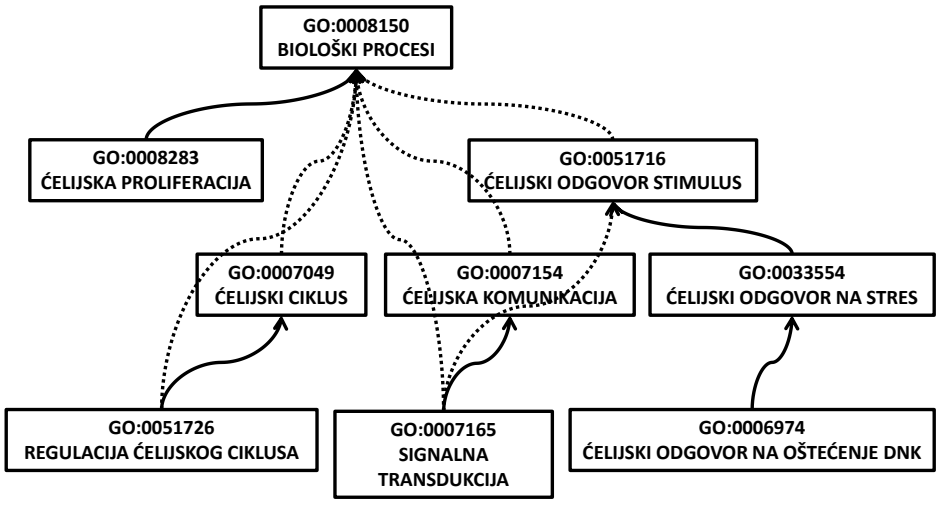
2. Функције протеина

2.1. Онтологија гена

GO (енг. *Gene Ontology*) је база свих описаних функција које протеин може да има. Организована је у 3 дисјунктне под-онтологије:

I) ћелијска компонента (ЋК)
II) молекуларна функција (МФ)
III) биолошки процес (БП)

Свака под-онтологија је директни ациклични граф, где је сваки GO термин чвор, а везе између термина су ивице између чворова. GO је хијерархијски организована, при чему су *потомачки* термини више специјализовани у односу на *родитељске* термине. Сваки чвор у једној под-онтологији директно или индиректно, је повезан са *кореном*, где корен представља чвор са GO термином саме под-онтологије (Слика 1). GO термин је начин обележавања функције. Свака функција је представљена GO термином облика 'GO:XXXXXXX', где је X цифра од 0 до 9.



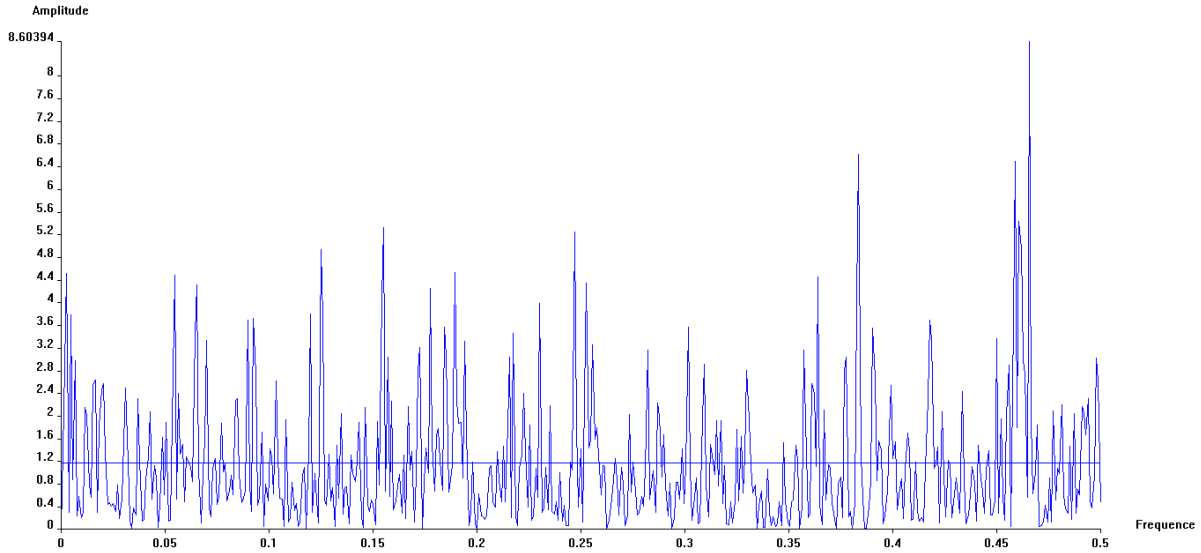
Слика 1 – Схематски приказ дела стабла под-онтологије биолошки процеси за MYBL2

2.2. Алгоритам за предвиђање функција протеина

2.2.1. ISM

ISM (енг. *Informational Spectrum Method*) је метода за анализу протеинских секвенци заснована на превођењу протеинске секвенце у нумеричку секвенцу и Фуријеовој трансформацији добијеног низа бројева.

Нумеричка секвенца се добија кодирањем сваке аминокиселине бројем који представља табличну вредност електрон-јон потенцијал интеракција (енг. *Electron-Ion Interaction Potential*). Тако добијени низ бројева се преводи у информациони спектар коришћењем Фуријеове трансформације (Слика 2).



Слика 2 – Информациони спектар протеина MYBL2

2.2.2. Како ради алгоритам?

Рад алгоритма при предвиђању функције протеина X се састоји из 3 корака.

1) Селекција свих протеина који задовољавају услов заснован на ISM.
   Услов селекције је поклапање првих пикова спектра протеина X и упоређеног протеина. Први пик је пик са највећом амплитудом спектра, а поклапање првих пикова подразумева да на истој фреквенцији имају своје прве пикове.
2) На основу селектованих протеина, идентификују се GO термини enrichment анализом. Enrichment анализа је статистичка метода којом се утврђују засићени GO термини у анализираном сету протеину.
3) Индентификоване GO термине приписујемо протеину X.

Алгоритам примењен на све хумане протеине даје резултат у виду текстуалног фајла, састављен од три колоне (Слика 3).

```
1120002  Q9BSD3  GO:0051174  1.00
1120003  Q9BSD3  GO:0009893  1.00
1120004  Q9BSD3  GO:0010604  1.00
1120005  Q9BSD3  GO:0006793  1.00
1120006  Q9BSD3  GO:0065009  1.00
1120007  Q9BSD3  GO:0019222  1.00
1120008  Q9BSD3  GO:0060255  1.00
1120009  Q9BSD3  GO:0080090  1.00
1120010  Q9BSD7  GO:0006066  1.00
1120011  Q9BSD7  GO:1901615  1.00
1120012  Q9BSD7  GO:0006629  1.00
1120013  Q9BSD7  GO:0044281  1.00
1120014  Q9BSD7  GO:0044710  1.00
1120015  Q9BSD7  GO:0044238  1.00
```

Слика 3 – Исечак резултата алгоритма за предвиђање функција

У једном реду је идентификациони стринг протеина, GO термин функције који одговара протеину из истог реда, а број на крају представља вероватноћу да та функција буде стварна функција овог протеина, међутим, вероватноћа није калибрисана, односно свака вероватноћа износи 1.00, па се та колона неће узимати у обзир.

2.2.3. CAFA

CAFA (енг. *The Critical Assessment of protein Function Annotation algorithms*) изазов је такмичење биоинформатичких група из целог света у прављењу алгоритама за аутоматизовано предвиђање функција протеина. Организатори овог изазова обезбеђују 'тренинг сет', за тренирање нових алгоритама и врше евалуацију предатих предикција. 'Тренинг сет' представља тренутно стање GO, тј. све експерименталне потврђене функције за сваки протеин (Слика 4).

```
1564887  P10244  GO:0090304
1564888  P10244  GO:0018130
1564889  P10244  GO:0019438
1564890  P10244  GO:0044271
1564891  P10244  GO:1901362
1564892  P10244  GO:0051171
1564893  P10244  GO:0080090
1564894  P10244  GO:0006807
1564895  P10244  GO:0006725
1564896  P10244  GO:0034641
1564897  P10244  GO:0044238
1564898  P10244  GO:0046483
1564899  P10244  GO:1901360
```

Слика 4 – Исечак из 'тренинг сета'

Процене перформанси у току процеса развоја алгоритма се заснивају на поређењу предвиђања и 'тренинг сета'. При поређењу овог фајла са фајлом резултата алгоритма, издвајамо 3 категорије у које можемо сврстати функције једног протеина.

ТП – тачно позитивне – функције које је алгоритам предвидео, а већ су експериментално показане

ЛН – лажно негативне – експериментално показане функције које алгоритам није предвидео

ЛП – лажно позитивне – функције које је алгоритам предвидео, а нису експериментално показане

Управо су ЛП од највећег значаја. Оне су филтриране функције које представљају основу за даља научна истраживања.

Међутим, алгоритам развијен у ИНН Винча као резултат даје фајл са предвиђањима и додатни фајл са статистиком која описује перформансе њиховог алгоритма на сваком протеину (који зависи од броја ТП, ЛП и ЛН), али у овим фајловима нема инфромације о томе које функције припадају којој категорији. Отуда потреба за програмом за екстракцију функција.

3. Програм за екстракцију функција

Програм има два улазна фајла, фајл са предвиђањима (алгоритам из ИНН Винче) и фајл са познатим експериментално доказаним функцијама (CAFA тренинг сет) . Циљ програма је да изврши екстракцију функција, тј. да се добије листа ТП, ЛП и ЛН функција сваког протеина (Слика 5).

```
Protein ID
TP:
GO:xxxxxxx
…
Broj TP
LP:
GO:xxxxxxx
…
Broj LP
LN:
GO:xxxxxxx
…
Broj LN
Protein ID
…
```

Слика 5 – Формат излазног фајла, тј. циљ програма

Основна идеја програма за екстракцију укључује следеће кораке: i) Издвојити одговарајуће функције протеина из оба фајла, ii) пронаћи функције које се налазе у оба фајла (оне представљају ТП) и iii) све што преостане од функција за тај протеин из CAFA фајла су ЛН, а iv) све што преостане од функција из фајла са ИНН Винча су ЛП. Потребно је пратити и број ТП, ЛП и ЛН за сваки протеин због статистике. Статистика се састоји из истих параметара који су у ИНН Винча поставили да се мере перформансе алгоритма у односу на сваки протеин, управо да би се проверило да ли је успешно изведено екстраковање функција, уместо ручне провере. Јако је важно да се приликом прављења програма, због величине улазних података, тежи меморијској оптимизацији.

У реализацији ове идеје потребно је мало детаљније погледати улазне фајлове. Запажа се како су протеини у блоковима, што значи да након континуитета идентификационог стринга дуж колоне, престаје се са тражењем функција тог протеина, јер се неће наћи нигде другде у фајлу. Такође, морамо узети у обзир могућност да ће постојати протеини који се налазе у само једном од два фајла, беспотребно продужујући време извршавања. Зато пре свега овога, прави се једна мапа протеина, тј. текстуални

фајл који садржи пресек идентификационих стрингова протеина из полазних фајлова, настао другим програмом у C.

Коришћене су две структуре у C. Једна структура описана је информационим делом у коме смештамо идентификацију протеина и који има два показивача на другу структуру. Она је описана GO термином и једним показивачем на исту структуру. Прва структура узима идентификацију протеина из мапе протеина.

Леви показивач прве структуре показује на листу GO термина који одговарају протеину, а извучени из фајла ИНН Винча, а десни на све GO термине извучене из CAFA фајла који одговарају протеину (Схема 1).

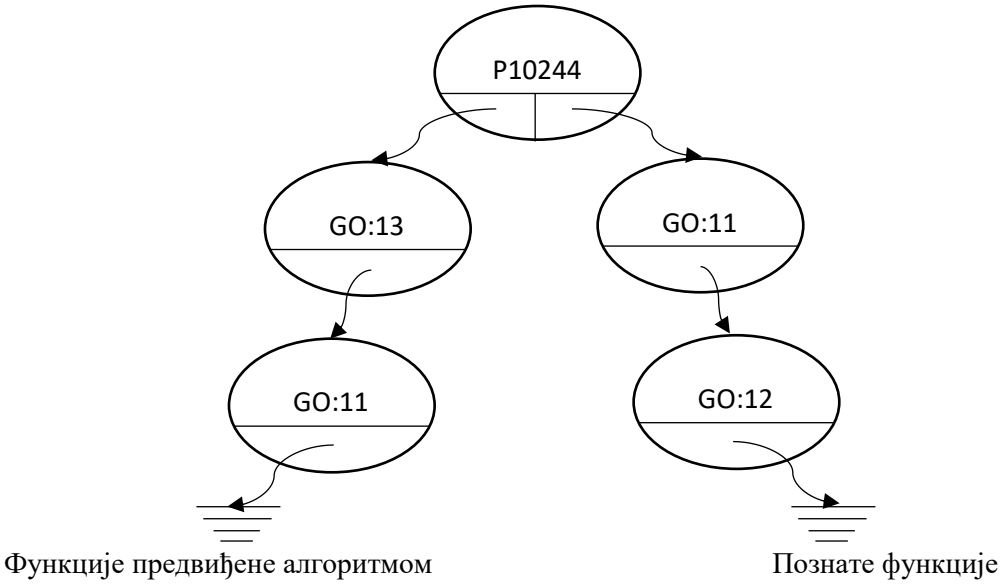


Схема 1 – Упрошћен приказ алгоритма за екстракцију функција

Редом за сваки протеин из мапе протеина, стварају се чворови са одговарајућим информацијама из оба фајла. Помоћним показивачима се пореде сваки GO термин из леве гране са сваким из десне гране. Ако се нађе исти GO термин у десној грани, уписује се у излазни фајл као ТП, бришу се оба чвора и следећи чвор у левој грани се испитује. Када се дође до краја леве гране, сви преостали у тој грани представљају ЛП, а сви преостали у десној грани представљају ЛН, па се такви уписују у фајл и бришу се сви чворови. Понавља се поступак за сваки протеин из мапе. Меморијска оптимизација је постигнута тиме што се само један протеин обрађује у једном тренутку, тј. из фајлова се доводе информације само за један протеин, па одмах следи обрада.

3.1. Резултати програма за екстракцију функција

Програм је примењен на све људске протеине и успешно је вратио 2 излазна фајла, фајл 'Статистика' и фајл 'Екстраковане функције', где се фајл 'Статистика' поклапа са фајлом статистике из ИНН Винча.

3.2. Предвиђене функције MYBL2

Програм је екстраковао 16 ЛП функција протеина MYBL2 (Табела 1).

Табела 1 – Термини груписани у кластере на основу веза у GO под-онтологији Биолошки Процеси (БП)

ЛП термин	**Функција**	**Референце**
GO:0006259	Процеси укључени у метаболизам ДНК	1°, 2°
GO:0006260	Репликација ДНК	
GO:0006091	Генерисање прекурсорских метаболита и енергије	
GO:0006974	Ћелијски одговор на оштећење ДНК	1°, 3°, 6°, 7°
GO:0033554	Ћелијски одговор на стрес	
GO:0051716	Ћелијски одговор на стимулус	
GO:0050896	Одговор на стимулус	
GO:0006950	Одговор на стрес	
GO:0008283	Ћелијска пролиферација	1°, 4°
GO:0007165	Сигнална трансдукција	1°, 5°
GO:0023052	Сигналинг	
GO:0007154	Ћелијска комуникација	
GO:0044700	Редундантан термин	
GO:0006383	Транскрипција са промотора РНК полимеразе III	Транскрипција
GO:0045945	Позитивна регулација транскрипције са промотора РНК полимеразе III	
GO:0006359	Регулација транскрипције са промотора РНК полимеразе III	

Референце: 1° Musa et al. (2017), 2° Lorvellec et al. (2010), 3° Liu et al. (2004), 4° Ness et al. (2003), 5° Seong et al. (2011), 6° Mannefeld et al. (2009), 7° Ahlbory et al. (2005)

Користећи претраживач литературе PubMed и алат QuickGO, можемо детаљније анализирати добијене ЛП функције:

- Један термин се показао као редундантан.
- Три од петнаест функција су блиско везане за транскрипцију, што је већ раније позната функција протеина MYBL2.
- За једанаест од петнаест предвиђених функција постоје референце које описују експерименталан доказ ових функција.
- За само једну од петнаест предвиђених функција нису нађене референце, али то не значи да она није функција протеина MYBL2, ово је прави пример функције који бисмо дали научнику да испита, јер би тада знао шта да испита.

4. Закључак

У овом раду, улога компјутерских алата приказана је на примеру гена MYBL2. Програм за екстракцију је успешно вратио све категоризоване функције свих протеина, уз меморијску оптимизацију. За преко 93% ЛП протеина MYBL2 нађене су референце које описују експериментален доказ ових функција. Једна од потврђених ЛП је ћелијска пролиферација, важан процес за патогенезу рака.

Ови резултати указују на важан допринос програма за екстракцију функција, који је омогућио препознавање функција које су предвиђене рачунарским алатом, а пре тога нису укључене у GO базу података. Примена овог програма, заједно са алгоритмом за предвиђање функције протеина, омогућава експериментално тестирање само оних ћелијских улога протеина које су највероватније и релевантне за одређени тип болести, као што је рак.

5. Литература

1. V. R. Perović, doktorska studija “Razvoj multifunkcionalne bioinformatičke platforme zasnovane na potencijalu electron-jon interakcije bioloških molekula”, (2013)

2. Musa J, Aynaud MM, Mirabeau O, Delattre O, Grünewald TG. MYBL2 (B-Myb): a central regulator of cell proliferation, cell survival and differentiation involved in tumorigenesis. Cell Death Dis. 2017;8(6):e2895.

3. Jiang Y, Oron TR, Clark WT, Bankapur AR, D’Andrea D, Lepore R, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016;17(1):184.

4. B. S. Gemović, doktorska disertacija „Bioinformatička analiza proteina uključenih u patogenezu mijeloidnih maligniteta", (2015).

5. Ashburner et al. Gene ontology: tool for the unification of biology (2000) Nat Genet 25(1):25-9. Online at Nature Genetics

6. GO Consortium, Nucleic Acids Res., 2017

7. https://www.ncbi.nlm.nih.gov/pubmed/

8. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. (2009) QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics. 2009; 25(22):3045-6.

9. Lorvellec, M. , Dumon, S. , Maya-Mendoza, A. , Jackson, D. , Frampton, J. and García, P. (2010), B-Myb is Critical for Proper DNA Duplication During an Unperturbed S Phase in Mouse Embryonic Stem Cells . STEM CELLS, 28: 1751-1759. doi:10.1002/stem.496

10. 71. Liu DX, Biswas SC, Greene LA. B-myb and C-myb play required roles in neuronal apoptosis evoked by nerve growth factor deprivation and DNA damage. J Neurosci 2004; 24: 8720–8725.

11. Ness SA. Myb protein specificity: evidence of a context-specific transcription factor code. Blood Cells Mol Dis 2003; 31: 192–200.

12. Seong H-A, Manoharan R, Ha H. B-MYB positively regulates serine-threonine kinase receptor-associated protein (STRAP) activity through direct interaction. J Biol Chem 2011; 286: 7439–7456.

13. Mannefeld M, Klassen E, Gaubatz S. B-MYB is required for recovery from the DNA damage-induced G2 checkpoint in p53 mutant cells. Cancer Res 2009; 69: 4073–4080.

14. Ahlbory D, Appl H, Lang D, Klempnauer K-H. Disruption of B-myb in DT40 cells reveals novel function for B-Myb in the response to DNA-damage. Oncogene 2005; 24: 7127–7134.