

## **The landscape of microbial phenotypic traits and associated genes**

Doprinos istraživanju u području molekularne biologije

Prokariotski mikroorganizmi, što uključuje bakterije i arheje, imaju vrlo visoku mogućost prilagodbe raznim okolišima, zahvaljujući fleksibilnom metabolizmu i robusnosti na razne vrste kemijskih i fizikalnih stresova. Velika količina informacija o ovakvim fenotipovima bakterija dostupna je u nestrukturiranom tekstu (npr. znanstveni članci, Wikipedia), međutim baza koja bi sadržavala sistematsku anotaciju mikrobijalnih fenotipova do sada nije postojala. U radu je razvijen računalni sustav nazvan ProTraits koji omogućuje automatsku i sistematičnu anotaciju fenotipova bakterija i arheja koristeći algoritme strojnog učenja. Sustav generira predikcije iz javno dostupnih znanstvenih i popularnih tekstova koji opisuju prokariotske vrste te, dodatno, komparativne analize genomskih podataka istih mikroorganizama. U radu su genomski podaci predstavljeni koristeći pet različitih reprezentacija, od čega su dvije (gensko susjedstvo i učinkovitost translacije) po prvi put korištene za predviđanje fenotipova. Osim generiranja predikcija za 311 fenotipova, ProTraits sustav također definira 113 novih fenotipskih koncepata iz teksta koristeći algoritam ne-negativne matrične faktorizacije. Razvijeni sustav sadrži ~545,000 novih anotacija za 424 fenotipa i 3046 vrsta bakterija i arheja. Predikcije su slobodno dostupne na poveznici: <http://protraits.irb.hr/>.

Važna primjena našeg sustava leži u otkrivanju gen-fenotip asocijacija, ili drugim riječima pronalaganju gena čiji su homolozi statistički nadzastupljeni u organizmima koji iskazuju neki fenotip, a podzastupljeni u organizmima koji ga nemaju. U radu smo pokazali da koristeći naš skup anotacija uspijevamo pronaći 57,088 asocijacija s visokom pouzdanošću, što je 7 puta povećanje u odnosu na broj anotacija koje smo uspjeli pronaći koristeći isključivo prijašnji skup anotacija. Primjerice, promatranjem gena statistički asociiranih s mogućnosti bakterija da inficiraju čovjeka i uzrokuju bolest, prethodno dostupne anotacije pronašle bi 57 takvih gena. Pomoću našeg ProTraits sustava, sada je moguće pronaći 1,187 gena vezanih uz patogenost; u ovom slučaju, povećanje od 21 puta. Ovi geni predstavljaju nove kandidate za istraživanje mehanizama raznih bolesti. Također je pronađeno 824 gena asociiranih specifično s bolestima probavnog sustava te 641 gen koji se češće pronalaze u patogenima dišnog sustava.

Rad „The landscape of microbial phenotypic traits and associated genes“ je objavljen je 24. listopada 2016. u časopisu Nucleic Acids Research izdavača Oxford Journals (faktor odjeka u 2015. godini: 9.2; NAR je 18 od 289 časopisa po ISI listi; gonjih 6%). Prvi autor na radu je Maria Brbić, a dopisni autor je dr. sc. Fran Supek sa Zavoda za elektroniku. Ostali autori rada su: Matija Piškorec, dr. sc. Vedrana Vidulin, dr. sc. Anita Kriško te dr. sc. Tomislav Šmuc. Rad je interdisciplinaran i pokazuje kako iskoristiti algoritme iz područja računarske znanosti za nova otkrića u biologiji.