KTH Computer Science
and Communication

Data Analysis and Next Generation Sequencing : Applications in Microbiology.

NICOLAS INNOCENTI

Doctoral Thesis
Stockholm, Sweden 2015

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i fysik fredagen den 30 oktober 2015 klockan 14.00 i AlbaNova Universitetscentrum, Kungl Tekniska högskolan, Roslagstullsbacken 21, Stockholm, rum FA32.

TRITA-CSC-A 2015:15 • ISSN-1653-5723 • ISRN-KTH/CSC/A–15/15-SE • ISBN 978-91-7595-699-2

Sammanfattning

Nästa Generations Sekvensering (NGS) är en ny teknik som i grunden har förändrat hur man kan studera levande organismer. Medan man tidigare endast kunde undersöka ett fåtal gener samtidigt så ger NGS möjlighet att utföra mätningar på ett helt genom, på en gång. Priset att betala är att tekniken genererar en så stor mängd data att det krävs nya bearbetnings- och analysmetoder för att få fram användbar information.

Denna avhandlings huvudbidrag är utvecklingen av en ny experimentell metod tagRNA-seq som kombinerar 5'tagRACE, en tidigare utvecklad teknik, med RNA-sekvensering. I korthet kan man med tagRNA-seq identifiera 5'-RNA-ändar i RNA från bakterier och att bestämma om de är primära, dvs som de syntetiserades i cellen, eller processade, dvs om de senare ändrats. Detta sker med hjälp av korta RNA-sekvenser, taggar, som ligeras till 5'-RNA-ändarna, en tagg för de primära och en tagg för de processade. Metoden har använts för att bestämma transkriptionsstarter och processeringspunkter i två bakteriearter, *Escherichia coli* och *Enterococcus faecalis*. Metoden har också använts för att studera polyadenylering i E. coli, där direkta och indirekta regulatoriska effekter kan separeras om man kan särskilja processade RNA-molekyler, samt för att visa hur data från tagRNA-seq kan användas för att hitta antisensetranskript i bakterier, vilka annars kan vara svåra att skilja från experimentella felkällor i RNA-sekvensering. En detaljerad analys visade på oscillationer i signalen från RNA-seq mot genernas 3'-ändar vilka har förklarats och kvantifierats med hjälp av Kolmogorovs brutna-stav modell. Belägg presenteras också för cirkularisering av några RNA-transkript, i egna såväl som i offentligt tillgängliga data.

Det praktiska problemet att välja ut vilka taggar som kan användas i tagRNA-seq ledde till en teoretisk frågeställning om ord som saknas i en text. En teori har utvecklats för att beskriva fördelningen, beroende på längden, av minimala saknade ord (minimal absent words - MAWs) i en slumpmässig text. Huvuddelen av de minimala saknade orden i levande organismers genom följer samma fördelning. Med undantag för en en del virus skiljer sig emellertid fördelningarna från den i slumpmässiga texter för tillräckligt långa ord. Minimala saknade ord från denna svans av fördelningarna är är nära relaterade till befintliga sekvenser i genomen som mestadels återfinns i närheten av regioner med viktiga regulatoriska funktioner.

I avhandlingen presenteras också en ny metod att bestämma artsammansättningen av blandade bakterieprov, baserat på tekniken komprimerad sensning (compressive sensing). En ny algoritm presenteras vilken är jämförbar med andra nyligen utvecklade metoder inom området.

TRITA-CSC-A 2015:15 • ISSN-1653-5723 • ISRN-KTH/CSC/A–15/15-SE • ISBN 978-91-7595-699-2