Generowanie długich otwartych ramek odczytu w genomach prokariotycznych

Rozprawa doktorska wykonana w Zakładzie Genomiki

Krystian Bączkowski

Streszczenie

Powszechnie przyjmuje się, że powstawanie nowej informacji genetycznej w organizmach odbywa się poprzez modyfikacje informacji już istniejącej, np. przez tasowanie eksonów, różnicowanie się genu po jego duplikacji, boczny transfer genów oraz fuzje (łączenie się) i fizje (dzielenie się) genów. Kod genetyczny i sekwencje kodujące białko posiadają jednak pewne właściwości, które umożliwiają pozyskanie zupełnie nowej informacji genetycznej. Jest to ułatwione dzięki zmniejszeniu prawdopodobieństwa powstawania kodonów stop w alternatywnych fazach odczytu sekwencji kodujących białko. Wynika to z małej częstości występowania kodonów zaczynających się od TA w sekwencjach kodujących białko, co sprawia, że na nici komplementarnej (w piątej fazie odczytu genu) stosunkowo rzadko będą generowane inne kodony zaczynające się od tych dwunukleotydów, w tym kodony stop TAA i TAG. Podobny efekt powoduje zmniejszenie używalności niektórych kodonów kodujących serynę i leucynę, które są komplementarne do kodonów stop w fazie czwartej. Dzięki temu mogą być generowane stosunkowo długie otwarte ramki odczytu (ang. Open Reading Frames, ORF) w alternatywnych fazach istniejących genów kodujących białko. Mimo zaproponowania tych mechanizmów generowania nie przeprowadzono dotychczas rozległych analiz umożliwiających znalezienie odpowiedzi na pytania: jak bardzo znacząca jest rola tych mechanizmów w generowaniu ORFów w genomach oraz jakie właściwości genomów szczególnie sprzyjają powstawaniu nowej informacji genetycznej.

W celu znalezienia odpowiedzi na powyższe pytania przeprowadzono szereg analiz na ponad milionie adnotowanych niezachodzących ORFów pochodzących z 457 genomów prokariotycznych. Analizy częstości występowania kodonów stop we wszystkich alternatywnych fazach odczytu badanych ORFów wykazały, że jest możliwe utworzenie stosunkowo długich ramek w tych fazach. W prawie jednej trzeciej genomów średnia odległość między kodonami stop dla niektórych faz alternatywnych przekraczała 300 nukleotydów. Jednakże, aby ramki stały się kodujące nie wystarczy, aby były długie, muszą również nieść pewną informację o kodowanych produktach. Przeprowadzone analizy wskazały, że ponad 8000 alternatywnych odczytów adnotowanych ORFów wykazuje wysokie prawdopodobieństwo kodowania szacowane w programie GeneMark. W jeszcze większym stopniu podobieństwo do rzeczywistych sekwencji wykazały potencjalne produkty

alternatywnych odczytów (średnio kilkadziesiąt procent w danej fazie) ze względu na skład aminokwasowy, średnią hydrofobowości, polarności, punkt izoelektryczny oraz tendencję do tworzenia struktur drugorzędowych i regionów transbłonowych. Jeśli badane sekwencje zostały wygenerowane w alternatywnych fazach odczytu, to po przeczytaniu ich w odpowiedniej fazie, powinny one wykazywać podobieństwo do ramek je generujących. Rzeczywiście udało się odnaleźć takie istotne podobieństwa dla ponad 200 tysięcy badanych sekwencji. Wśród nich było prawie 60 tysięcy sekwencji kodujących białko, co świadczy o możliwości pozyskania funkcji przez nowo wygenerowane sekwencje. Dalsze badania wykazały, że opisane konserwatywne elementy funkcjonalne i strukturalne białek, takie jak domeny czy motywy, są także często zaangażowane w proces generowania. Przeprowadzone analizy pozwoliły częściowo wyjaśnić zagadkę ORFanów, tzw. sierocych ORFów, które są specyficzne dla danego genomu lub co najwyżej blisko spokrewnionych genomów. Duża część takich ramek hipotetycznych została najprawdopodobniej wygenerowana przez geny kodujące białka w fazach alternatywnych. W pracy przedstawiono również kilka scenariuszy generowania ramek na konkretnych przykładach adnotowanych sekwencji w poszczególnych genomach.

Wszystkie przeprowadzone badania wskazują, że siła generowania zależy od fazy odczytu sekwencji generującej. Nowe ramki najczęściej są generowane w fazie czwartej i piątej adnotowanych ORFów. Ramki odczytane w tych fazach charakteryzowały się najmniejszą liczbą kodonów stop, największym potencjałem kodowania wyliczonym w programie GeneMark i dominowały pod względem liczby znalezionych istotnych homologów. Dodatkowo potencjalne produkty odczytane w czwartej fazie adnotowanych ORFów najbardziej przypominały rzeczywiste sekwencje kodujące białko zwłaszcza w tendencji do tworzenia α helis i regionów transbłonowych oraz w rozkładzie punktu izoelektrycznego.

Wykazano, że intensywność generowania ORFów jest silnie dodatnio skorelowana z globalną zawartością guaniny i cytozyny w badanych organizmach, a największymi możliwościami generującymi charakteryzują się przedstawiciele grup bakterii: Acidobacteria, δ-Proteobacteria, Deinococcus-Thermus i β-Proteobacteria. Uzyskane wyniki wskazują, że stwierdzone mechanizmy generowania mogą działać na duża skalę, a sekwencje (nazwane alterlogami) powstałe w wyniku alternatywnego odczytu innych sekwencji mogą łatwo uzyskiwać nową informację genetyczną zapewniającą szybką ewolucję posiadającym je organizmom.

Wszystkie otrzymane dane zintegrowano w relacyjnej bazie danych, która może być dalej rozbudowywana i aktualizowana. Można również za jej pomocą przeprowadzać dodatkowe analizy: określać sposoby zachodzenia poszczególnych sekwencji w genomie, identyfikować błędnie adnotowane ramki, poszukiwać zjawiska fuzji i fizji genowych oraz genów, miedzy którymi doszło do tzw. parowania translacyjnego.