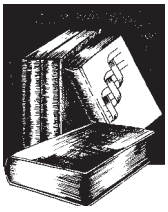


Zagadka sierocych otwartych ramek odczytu (ORF-anów)

Paweł Mackiewicz, Krystian Bączkowski, Maciej Sobczyński, Stanisław Cebrat

Zakład Genomiki, Instytut Genetyki i Mikrobiologii, Uniwersytet Wrocławski, Wrocław

Mystery of orphan Open Reading Frames (ORFans)

Summary

In spite of an increase of a number of sequenced genomes, 20-30% of Open Reading Frames (ORFs) do not show detectable sequence similarity to other sequences in databases and functions of their products have not been recognized. These ORFs are called ORFans and their abundance has been referred to as a "mystery". A large fraction of ORFans are probably spurious non-coding ORFs generated by protein coding sequences or repeat sequences. Some ORFans are probably quickly evolving genes specific to narrow phylogenetic lineages. The most interesting group of ORFans are *de novo* generated genes with new functions and structures which may find some applications in biotechnology and medicine. It is also possible that some ORFans code for RNA regulatory molecules.

Key words:
ORFan, hypothetical gene, microbial genomics, evolution.

Adres do korespondencji

Paweł Mackiewicz,
Zakład Genomiki,
Instytut Genetyki
i Mikrobiologii,
Uniwersytet Wrocławski,
ul. Przybyszewskiego 63/77,
51-148 Wrocław;
e-mail:
pamac@microb.uni.wroc.pl

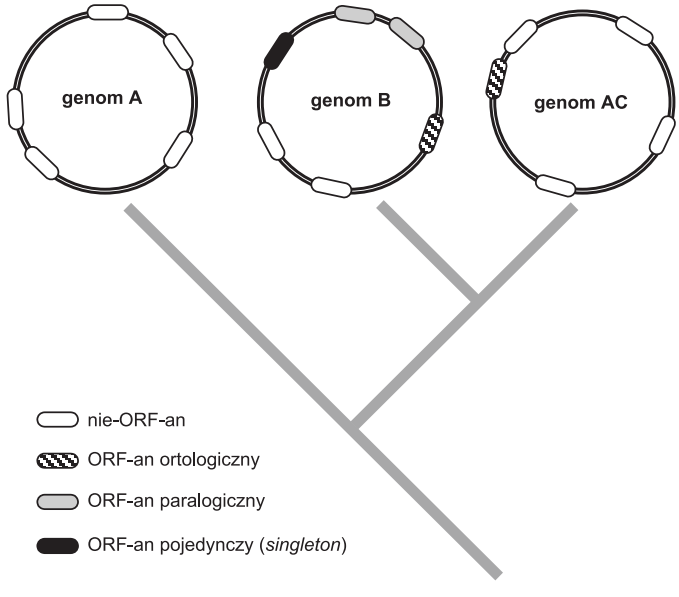
biotechnologia
3 (70) 45–56 2005

1. ORF-y hipotetyczne

Jedną ze standardowych analiz przeprowadzanych po etapie sekwencjonowania genomu jest identyfikowanie w nim potencjalnych sekwencji kodujących białko i poszukiwanie podobnych do nich sekwencji występujących w innych genomach – homologów. Jednak pomimo dynamicznego przyrostu liczby kompletnie zsekwencjonowanych genomów i intensywnego rozwoju narzędzi bioinformatycznych, około 30% sekwencji potencjalnie ko-

dujących białka (tzw. ORF-ów – otwartych ramek odczytu, ORF, *Open Reading Frame*) nie posiada przypisanych funkcji. Określa się je jako geny hipotetyczne. Jeżeli takie ORF-y lub ich potencjalne produkty białkowe wykazują istotne podobieństwo do innych genów obecnych w genomach pochodzących z różnych linii filogenetycznych, to określane są one terminem konserwatywne geny hipotetyczne. Stanowią one około 10% wszystkich adnotowanych potencjalnych genów w genomach mikroorganizmów. Najprawdopodobniej są to sekwencje kodujące białko, a dla wielu z nich z czasem udaje się przypisać za pomocą metod eksperymentalnych i komputerowych istotne dla komórki funkcje (1).

Istnieje jednak bardzo intrygująca grupa ORF-ów hipotetycznych, które nie tylko nie posiadają przypisanej funkcji, ale również nie mają odpowiadających sekwencji homologicznych w odległych genomach. Generalnie określa się je jako ORF-y sieroce, czyli ORF-any (2,3). Stanowią one 20-30% wszystkich ORF-ów w nowo sekwencjonowanych genomach (3-9), a w niektórych (*Plasmodium*) nawet 60% (10). Stanowią one duże wyzwanie dla genomiki. ORF-y sieroce są bardzo niejednorodną grupą (11) – rysunek 1. ORF-any nie posiadające żadnych homologów nazywane są singletonami lub ORF-anami pojedynczymi. Te, które posiadają homologi tylko w swoim genomie są ORF-anami paralogicznymi, a te, które wykazują pewne podobieństwo do ORF-ów występujących w innych bardzo blisko spokrewnionych genomach (gatunkach) – są ORF-anami ortologicznymi. Aby uwzględnić ich pewne podobieństwo do bardzo wąskiej grupy ORF-ów stosowany jest na ich określenie również termin PCOs (*Poorly Conserved ORFs*, 11).



Rys. 1. Klasyfikacja sierocych otwartych ramek odczytu – ORF-anów.

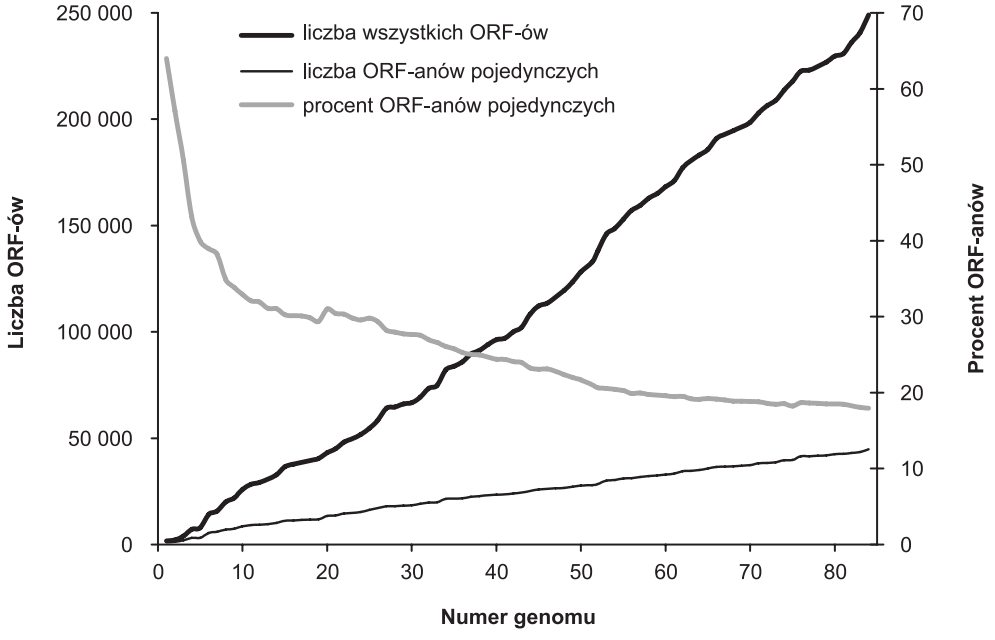
2. ORF-any

Wydaje się, że wraz ze wzrostem liczby nowo odkrywanych genów i rozwojem projektów sekwencjonowania genomów liczba ORF-anów powinna maleć, natomiast obserwuje się jej wzrost (2,12). Zjawisko to określono tzw. tajemnicą lub paradoksem ORF-anów. W analizach przeprowadzonych na 127 genomach mikroorganizmów wykazano obecność 66 846 ORF-anów specyficznych gatunkowo, tzn. nie posiadających homologów poza swoim gatunkiem (13). Średnio w każdym genomie występuje 14,4% takich ORF-anów, przy czym najwięcej z nich (52,3%) znaleziono w genomie *Aeopyrum pernix*, a najmniej u *Buchnera* sp. APS (0,2%). Obecność wielu ORF-anów u *A. pernix* najprawdopodobniej związana jest z dużą liczbą adnotowanych w tym genomie przypadkowych ramek zachodzących na kodujące geny. W celu dokładniejszej analizy tej grupy sekwencji stworzono bazę danych ORFanage poświęconą tylko ORF-anom (14; www.bioinformatics.buffalo.edu/ORFanage). Baza ta zawiera 31 850 pojedynczych ORF-anów stanowiących 13% wszystkich ORF-ów zidentyfikowanych w 84 genomach.

Istnienie ORF-anów, czyli sekwencji bez homologów jest trudne do wytłumaczenia, zakładając powszechnie znane mechanizmy ewolucji genów przez duplikacje innych genów. W każdym z tych przypadków powinny istnieć podobne sekwencje u innych organizmów lub w tym samym genomie. Nawet przyjmując powszechne poziome przenoszenie informacji genetycznej powinno się znajdować podobne sekwencje w gatunkach, które były dawcami nowych genów (tzw. ksenologów). Zaproponowano wiele różnych hipotez tłumaczących istnienie tej dziwnej grupy ORF-ów.

2.1. Ubogie bazy danych

W najprostszym wytłumaczeniu zakłada się, że obecność ORF-anów wynika ze zbyt małej liczby sekwencji obecnych w bazach danych (2,3,12), dlatego nie jest możliwe odnalezienie dla nich sekwencji podobnej. Jednakże, pomimo pewnego zmniejszenia tempa ich przyrostu, całkowita liczba ORF-anów rośnie wraz ze wzrostem liczby poznawanych genomów zamiast spadać (8,9,11). Tempo wzrostu nowo zsekwencjonowanych genomów wzrasta szybciej, tj. o 17,6% niż maleje liczba ORF-anów, tj. o 7,6% (13). Każdy nowo dodawany genom przyczynia się do spadku średnio tylko o 0,1-0,2% udziału procentowego wszystkich ORF-anów (8). Przy dodaniu kolejno zsekwencjonowanego genomu rośnie całkowita liczba ORF-ów i ORF-anów w bazie, jednocześnie maleje procent ORF-anów w całym zbiorze ORF-ów (rys. 2). Spadek jest jednak coraz mniejszy i stabilizuje się przy kilkunastu procentach. Obserwowany pewien spadek frakcji ORF-anów związany jest m.in. z dodawaniem genomów bardzo blisko spokrewnionych do już istniejących, np. szczepów tego samego gatunku (8), co powoduje tylko zmianę kategorii ORF-anu pojedynczego na ortologiczny (specyficznego dla gatunku lub wąskiej grupy).



Rys. 2. Dynamika przyrostu ORF-anów pojedynczych (singletonów). Przy dodawaniu kolejno zsekwencjonowanych genomów (oś X) rośnie całkowita liczba ORF-ów i ORF-anów w bazie (lewa oś Y), jednocześnie maleje procent ORF-anów w całym zbiorze ORF-ów (prawa oś Y). Spadek jest jednak coraz mniejszy i stabilizuje się przy kilkunastu procentach. Dane zaczerpnięto z bazy ORFanage (14; www.bioinformatics.buffalo.edu/ORFanage).

2.2. Masowa utrata genów

Występowanie ORF-anów można wytłumaczyć utratą genów, które początkowo występowały w wielu organizmach-przodkach, a w czasie ewolucji zanikły w wielu potomnych liniach filogenetycznych, pozostając tylko w niektórych z nich (3,11). W tym przypadku ORF-any odpowiadałyby pojedynczym potomkom dawnych białek, które zachowały się tylko w niektórych liniach filogenetycznych lub organizmach. Zjawisko masowej utraty genów zostało obserwowane u wielu organizmów (15-18). Jednak w tym przypadku należałoby założyć, że genom postulowanego przodka musiałby być ogromny i zawierać bardzo dużą liczbę genów zupełnie dziś nieznanych, albo, że niektóre geny żyją bardzo krótko.

2.3. Niekodujące artefakty

Skoro ORF-any nie mają homologów, a zatem biorą się znikąd, to może są artefaktami, błędnie rozpoznanymi i adnotowanymi ORF-ami, które nie są rzeczywistymi genami kodującymi białko. Liczba przypadkowych, niekodujących ORF-ów wzra-

sta wykładniczo wraz ze spadkiem ich długości (19-22). Dokładność metod rozpoznających geny zmniejsza się właśnie w przypadku krótkich sekwencji, dlatego wiele błędnie adnotowanych ORF-ów powinno być krótszych niż znane sekwencje kodujące. Rzeczywiście wiele ORF-anów prawdopodobnie niekodujących jest ORF-ami krótkimi, co stwierdzono zarówno u drożdży *Saccharomyces cerevisiae*, jak i w genomach prokariotycznych (3,8,13,23-26). Według bazy danych ORFanage 63,5% ORF-anów pojedynczych stanowią ramki krótsze niż 150 kodonów. Co ciekawe, liczba ORF-anów krótszych niż 150 kodonów rośnie szybciej niż długich i znika dwa razy wolniej niż długich (8,11).

Analizując skład G+C wykazano, że ORF-any specyficzne gatunkowo posiadają mniejszą zawartość G+C niż nie-ORF-any, zbliżając się do składu sekwencji międzygenowych, co sugerowałoby, że są one ORF-ami niekodującymi wygenerowanymi w przestrzeniach międzygenowych. Jednak nie można wykluczyć, że one kodują, a tylko podlegają mniejszej presji selekcyjnej, co upodabnia je do sekwencji międzygenowych (13). Chociaż prawdopodobieństwo wygenerowania długich ORF-ów w sekwencji losowej (międzygenowej?) jest bardzo małe (19,22,27,28), to stosunkowo długie ORF-y mogą być stosunkowo łatwo generowane w sekwencjach kodujących. Rzeczywiście, istnieje wiele ORF-ów zachodzących na siebie w genomach wirusów, archeonów, bakterii i *Eukaryotów* (np. 29-40). Tylko w nielicznych przypadkach oba ORF-y zachodzące są kodujące. Zaproponowano kilka mechanizmów generujących ramki zachodzące na geny (41), takie jak: wysoka zawartość par GC i używalność kodonów GNC i RNY (np. 30,31,34,42-46) oraz specyficzne właściwości kodu genetycznego i sekwencji kodujących (35,47). Łatwość generowania długich niekodujących ORF-ów nasuwa przypuszczenie, że część ORF-anów mogła powstać w wyniku dawnych duplikacji całych lub części sekwencji kodujących niosących w sobie niekodujące zachodzące ORF-y (23). Zduplikowane sekwencje akumulowały mutacje, które ostatecznie wyeliminowały właściwe ramki odczytu genów pozostawiając ORF-y generowane. Rzeczywiście, dla około 700 ORF-anów adnotowanych w genomie drożdży *Saccharomyces cerevisiae* przeczytanych w fazach niekodujących znaleziono istotne podobieństwo do innych znanych sekwencji (23).

2.4. Błędy sekwencjonowania lub pseudogeny

Zasugerowano również, że obecność ORF-anów jest efektem błędów popełnionych w trakcie sekwencjonowania genomów, powodujących przesunięcie ramki odczytu (2,3). W tym przypadku ich poprawienie (np. na podstawie porównania z innymi genomami) odtworzyłoby właściwą sekwencję kodującą. Jednakże tego typu zmiany w sekwencjach mogą dotyczyć również niefunkcjonujących już genów – pseudogenów, które nabyły mutacje zmieniające ich strukturę. W tym przypadku zmiany mogą być na tyle duże, że może okazać się niemożliwe odnalezienie sekwencji podobnej. Pseudogeny są sekwencjami występującymi zarówno w geno-

mach prokariotycznych (15,48-51), jak i eukariotycznych (52-56). W genomach *Ricketsia* stwierdzono, że około 80% ORF-anów jest pseudogenami i wykazuje skrócenie oraz przesunięcie ramki odczytu (57,58). Jednakże te genomy są bardzo specyficzne i charakteryzują się szczególną tendencją do redukcji swoich rozmiarów (59,60), dlatego założenie, że wszystkie ORF-any lub ich duża frakcja są pseudogenami może nie dotyczyć wielu innych genomów.

2.5. Szybko ewoluujące geny

ORF-any są wyróżniane głównie na podstawie braku istotnych podobieństw do innych sekwencji, dlatego zasugerowano, że stosowane algorytmy do poszukiwania sekwencji podobnych są za mało czułe, aby poprawnie rozpoznać ich potencjalne homologi. Może to dotyczyć wielu ORF-anów zwłaszcza krótkich, ponieważ krótkie sekwencje dają z reguły niskie wartości opisujące stopień podobieństwa (8). Jednakże wiele ORF-anów jest wystarczająco długich i powinny dać wysokie wartości podobieństwa, dlatego przyjęto hipotezę, że mogą one reprezentować geny szybko ewoluujące w danej linii (grupie) filogenetycznej, których sekwencja zmieniła się na tyle, że uniemożliwia znalezienie sekwencji podobnych (2,3,13,18,61-63).

Każdy z sekwencjonowanych szczepów *Escherichia coli*, pomimo niedawnej dywergencji, posiada swój własny zestaw specyficznych dla danego szczepu ORF-anów (64), co sugeruje, że ta grupa ORF-ów musiała powstać w bardzo krótkim czasie (65). U *Drosophila* wykazano, że około 1/3 eksprymowanych genów podlegała szybkiej (lub nawet neutralnej) ewolucji (66,67). Jednak część z nich nie wykazuje jawnego fenotypu, co sugeruje, że mogą one już nie być sekwencjami kodującymi. Szybkie tempo ewolucji wykazano dla ORF-anów specyficznych dla *Escherichia coli* oraz *Salmonella enterica* i uznanych za kodujące (65). Fischer i Eisenberg (3) zaproponowali, że wiele ORF-anów może kodować białka błonowe, które zachowując swoją strukturę mogą bardziej zmieniać swoją sekwencję uniemożliwiając znalezienie homologów standardowymi algorytmami. Wiadomo, że struktura przestrzenna białek ewoluuje wolniej niż sekwencja aminokwasowa (struktura pierwszorzędowa).

Przyjmując koncepcję ORF-anów jako szybko ewoluujących genów, kodowałyby one polipeptydy będące odległymi członkami już znanych rodzin białek i posiadałyby podobne struktury przestrzenne i funkcje (3,9). Rzeczywiście, w przypadku niektórych ORF-anów stwierdzono, że ich produkty posiadają zidentyfikowane struktury przestrzenne i złożenia białek występujące w już znanych rodzinach białkowych (68). Aby uzyskać więcej informacji na temat szybko ewoluujących ORF-anów konieczne byłoby przeprowadzenie bardziej zaawansowanych analiz komputerowych opartych na profilach sekwencji, eksperymentalnych badań biochemicznych lub eksperymentalnych i obliczeniowych analiz strukturalnych.

2.6. Geny nabyte w wyniku horyzontalnego transferu

W ostatnich analizach wskazano, że duża cześć ORF-anów mogła zostać nabyta od bakteriofagów w wyniku horyzontalnego transferu. Za takim pochodzeniem ORF-anów może świadczyć to, że są one krótkie, bogate w A+T i podlegają szybkiej ewolucji podobnie jak geny fagowe, a ponadto występują w grupach w okolicach miejsc, do których integrują się fagi (13,65,69). Rzeczywiście, Shmuely i wsp. (70) wykazali, że białko archebakteryjnego faga PhiCh1 jest podobne do ORF-anu paralogicznego występującego u *Halobacterium* sp., natomiast Daubin i Ochman (65) zidentyfikowali istotne podobieństwo między prawie 50 ORF-anami specyficznymi dla niektórych przedstawicieli gamma-proteobakterii a genami bakteriofagów. Brak homologów w przypadku wielu innych ORF-anów może wynikać ze słabej znajomości genomów wirusów i fagów oraz z dużej dywergencji sekwencji należących do tych grup.

Chociaż funkcji większości genów bakteriofagowych nie poznano, to przyjmuje się, że część z nich odpowiada za utrzymanie profaga w genomie bakteryjnym przez dostarczanie gospodarzowi pewnych korzyści. Wyjaśniałoby to, dlaczego takie geny (ORF-any) nabyte od bakteriofagów są u bakterii zachowywane po wyeliminowaniu przyległych sekwencji fagowych. Kilka ORF-anów z genomu *E. coli* powstałych prawdopodobnie w ten sposób zaangażowanych jest w procesy translacji i replikacji (65). Wiele z nich ulega ekspresji w czasie stresu lub głodu, co może wskazywać na stare powiązanie między tymi genami a samolubnymi elementami (*selfish*) mobilizowanymi właśnie w tym czasie.

2.7. Całkowicie nowe geny

W związku z tym, że wśród znanych sekwencji nie znajduje się żadnych homologów do ORF-anów, być może tworzą one nowe, wygenerowane *de novo*, nie znane jeszcze rodziny białek posiadające nowe struktury przestrzenne i być może nowe funkcje nie zaobserwowane w już istniejących rodzinach białek (3,9,11,71,72). Przyjmując takie założenie niemal każdy ORF-an (lub grupy ORF-anów paralogicznych i ortologicznych) musiałby być przedstawicielem nowej nadrodziny białek. Liczba tych nadrodzin powinna być wielokrotnie większa niż do tej pory przyjmowano (73,74) i, co więcej, nadrodziny te powinny być bardzo zróżnicowane między sobą i odległe ewolucyjnie od już znanych (3). Dla trzech ORF-anów stwierdzono, że posiadają one nowe, nie znane wcześniej struktury przestrzenne i złożenia białek (68).

Niewykluczone, że ORF-any tego typu zostały wygenerowane w innych fazach wewnątrz już istniejących sekwencji kodujących, które charakteryzują się łatwością generowania stosunkowo długich ORF-ów. Wielu autorów zwróciło uwagę na taką możliwość tworzenia nowych genów w ewolucji (31,33,41,44,47,75). Znaleziono

kilka przykładów tak wygenerowanych genów badając sekwencje drożdży (35) oraz człowieka i wirusów (76,77).

W generowaniu nowych genów mogą odgrywać również rolę sekwencje powtórzone. Przykładem genów wygenerowanych *de novo* przy ich udziale jest rodzina genów zwana *morpheus* występująca u naczelnych (78), gen *LQK1* (79) oraz gen *AFGP* kodujący glikoproteinę zapobiegającą zamarzaniu ryb antarktycznych (80).

2.8. Niekodujący RNA

Nie można wykluczyć, że niektóre ORF-any, szczególnie u organizmów eukariotycznych, mogą reprezentować RNA niekodujące białek, które odgrywają dużą rolę u tych organizmów w regulacji ekspresji genów, organizacji materiału genetycznego i ochronie przed pasożytami przez eliminowanie defektywnego RNA i RNA transpozonów oraz wirusów (81-83). Jest to stosunkowo niedawno odkryte zjawisko zwane wyciszaniem genów lub interferencją RNA (RNAi – RNA *interference*). Transkrypty niektórych ORF-anów (wygenerowanych w antysensie genów) mogłyby pełnić funkcję regulatorową, hybrydyzując z transkryptami genów kodujących białka regulując ich ekspresję. Gdyby okazało się, że ORF-any kodują funkcjonalne RNA, to byłyby one bardzo istotnym składnikiem materiału genetycznego.

2.9. Czy ORF-any są sekwencjami kodującymi?

Najważniejszym problemem związanym z ORF-anami jest pytanie czy są to sekwencje kodujące? Jedna z metod określających czy dana sekwencja koduje białko opiera się na wynikach analiz potencjalnej liczby podstawień nukleotydowych w miejscach synonimicznych, w których zmiany nie wpływają na kodowany aminokwas i niesynonimicznych, związanych ze zmianą kodowanego aminokwasu. Sekwencje kodujące białka powinny charakteryzować się większą liczbą substytucji w miejscach synonimicznych (Ks) niż w miejscach niesynonimicznych (Ka). W przeprowadzonych analizach ORF-anów *Drosophila* (63) oraz ORF-anów specyficznych dla linii *Escherichia* i *Salmonella* (65) wykazano, że ewoluują one podobnie do sekwencji kodujących białko i posiadają stosunek Ka/Ks mniejszy niż jeden. Na podstawie tej metody wykazano również, że wiele krótkich ORF-ów w genomach mikroorganizmów przypomina pod tym względem sekwencje kodujące białko, a zatem może kodować (84).

W wielu przypadkach wykazano doświadczalnie, że ORF-any ulegają ekspresji (18,61,70,85, patrz również 65). Jednak sam fakt transkrypcji nie może być wystarczającym dowodem na kodowanie białka przez daną sekwencję. Wiele transkryptów może być „fałszywych", syntetyzowanych z sekwencji niekodujących położonych blisko promotorów. Stwierdzono ponadto, że nie ma korelacji między ilością mRNA

i ilością kodowanego przez nie białka w komórkach drożdży (86), co może być m.in. związane z obecnością niekodujących transkryptów. Jednakże, brak zidentyfikowanych transkryptów dla niektórych ORF-ów niekoniecznie musi świadczyć, że te ORF-y nie kodują białek. Mogą one ulegać ekspresji w innych warunkach niż zastosowane w doświadczeniu, albo ich poziom ekspresji może być bardzo niski i wobec tego niewykrywalny przez stosowane metody.

W analizach komputerowych potencjalnych białek kodowanych przez ORF-any specyficzne gatunkowo wykazano, że w porównaniu do reszty białek charakteryzują się one nieco mniejszą hydrofobowością, są mniejsze, i mają wyższy punkt izoelektryczny, co sugeruje, że mogą one być czynnikami transkrypcyjnymi lub innymi białkami regulatorowymi wiążącymi się do kwasów nukleinowych (13,62).

Niektórym ORF-anom, po dokładniejszych analizach udało się przypisać różne kategorie funkcjonalne, w tym rolę w transkrypcji i translacji (87,88, patrz również 68). Przykładowo, uzyskano ekspresję produktu genu *ykfE* z genomu *Escherichia coli* oznaczonego wcześniej jako ORF-an i poddano go krystalizacji (88,89). Okazało się, że produktem tego genu jest białko homodimeryczne będące inhibitorem lizozymu typu C. Określono również trójwymiarową strukturę tego białka i odszukano białko o podobnej strukturze przestrzennej u *Pseudomonas aeruginosa*. Wcześniej funkcja tego białka była także nie znana. Jednak po analizach okazało się, że białko to jest również inhibitorem lizozymu typu C. Dzięki przeprowadzonym analizom udało się zatem określić nową rodzinę białek bakterii. Rodzina ta jest bardzo ciekawa ze względu na wysoki stopień dywergencji sekwencji do niej należących, oraz wysoką konserwatywność ich funkcji.

3. Zakończenie

Widać, że żadna z przedstawionych hipotez dotyczących ORF-anów nie jest do końca przekonująca. Najprawdopodobniej ta grupa ORF-ów jest bardzo niejednorodna zarówno pod względem pochodzenia, jak i roli w komórce. Obecność znacznej frakcji ORF-anów nie można tłumaczyć małą liczbą poznanych genomów, bo mimo wzrostu tej liczby, frakcja ORF-anów nie maleje zgodnie z oczekiwaniami. Tylko bardzo niewielka część ORF-anów wynika z błędów sekwencjonowania, ponieważ dokładność odczytów sekwencji jest obecnie wysoka. Część ORF-anów, zwłaszcza krótkich, może być sekwencjami niekodującymi, wygenerowanymi w przestrzeniach międzygenowych, w sekwencjach powtórzonych i przede wszystkim w sekwencjach kodujących białko. Osobną grupę mogą reprezentować szybko ewoluujące geny specyficzne dla wąskiej linii filogenetycznej. W tym przypadku sekwencje tych białek uległy tak dużej dywergencji, że staje się niemożliwe zidentyfikowanie dalszych homologów na poziomie struktury pierwszorzędowej. Niewykluczone, że wiele z tych ORF-ów nie podlega selekcji i staje się pseudogenami. Pewien udział w tworzeniu ORF-anów może mieć masowa utrata genów w różnych liniach filoge-

netycznych oraz horyzontalny transfer genów, zwłaszcza od bakteriofagów. Najciekawszą grupę ORF-anów mogą stanowić ORF-y kodujące małocząsteczkowe RNA i całkowicie nowe geny kodujące białko o nowych funkcjach i strukturach.

Poznanie natury ORF-anów jest ważne dla poznania pełnego zróżnicowania białek kodowanych w organizmie. Niektórym ORF-anom udało się przypisać już funkcje i poznać struktury przestrzenne ich produktów. Niewykluczone, że wśród ORF-anów znajdują się istotne dla funkcjonowania komórki białka o całkowicie nowych, ciekawych strukturach, które mogą znaleźć zastosowanie w medycynie i biotechnologii. Szczególnie interesujące mogą się okazać te geny, które są związane z patogennością lub wirulencją.

Literatura

1. Galperin M. Y., Koonin E. V., (2004), Nucleic Acids Res., 32, 5452-5463.
2. Dujon B., (1996), Trends Genet., 12, 263-270.
3. Fischer D., Eisenberg D., (1999), Bioinformatics, 15, 759-762.
4. Doolittle R. F., (1997), Nature, 388, 515-516.
5. Fraser C. M., Eisen J. A., Salzberg S. L., (2000), Nature, 406, 799-803.
6. Wren B. W., (2000), Nature Rev. Genet., 1, 30-39.
7. Boucher Y., Nesbo C. L., Doolittle W. F., (2001), Curr. Opin. Microbiol., 4, 285-289.
8. Siew N., Fischer D., (2003), Proteins: Struct. Funct. Genet., 53, 241-251.
9. Siew N., Fischer D., (2003), Structure (Camb), 11, 7-9.
10. Gardner M. J., Hall N., Fung E., White O., Berriman M., Hyman R. W., Carlton J. M., Pain A., Nelson K. E., Bowman S., et al., (2002), Nature, 419, 498-511.
11. Siew N., Fischer D., (2003), Comp. Funct. Genomics, 4, 432-441.
12. Casari G., de Druvar A., Sander C., Shneider R., (1996), Trends Genet., 12, 244-255.
13. Charlebois R. L., Clarke G. D., Beiko R. G., Jean A., (2003), FEMS Microb. Lett., 225, 213-220.
14. Siew N., Azaria Y., Fischer D., (2004), Nucleic Acids Res., 32, D281–D283.
15. Andersson J. O., Andersson S. G. E., (1999), Curr. Opin. Gen. Dev., 9, 664-671.
16. Aravind L., Watanabe H., Lipman D. J., Koonin E. V., (2000), Proc. Natl. Acad. Sci. USA, 97, 11319-11324.
17. Petrov D. A., Sangster T. A., Johnston J. S., Hartl D. L., Shaw K. L., (2000), Science, 287, 1060-1062.
18. Zdobnov E. M., von Mering C., Letunic I., Torrents D., Suyama M., Copley R. R., Christophides G. K., Thomasova D., Holt R. A., Subramanian G. M., et al., (2002), Science, 298, 149-159.
19. Termier M., Kalogeropoulos A., (1996), Yeast, 12, 369-384.
20. Das S., Yu L, Galtatzes C., Rogers R., Freeman J., Bienkowska J., Adams R. M., Smith T. F., (1997), Nature, 385, 29-30.
21. Basrai M. A., Hieter P., Boeke J. D., (1997), Genome Res., 7, 768-771.
22. Gierlik A., Mackiewicz P., Kowalczuk M., Dudek M. R., Cebrat S., (1999), Int. J. Modern Phys. C., 10, 635-643.
23. Mackiewicz P., Kowalczuk M., Gierlik A., Dudek M. R., Cebrat S., (1999), Nucleic Acids Res., 27, 3503-3509.
24. Skovgaard M., Jensen L. J., Brunak S., Ussery D., Krogh A., (2001), Trends Genet., 17, 425-428.
25. Mackiewicz P., Kowalczuk M., Mackiewicz P., Nowicka A., Dudkiewicz M., Łaszkiewicz A., Dudek M. R., Cebrat S., (2002), Yeast, 19, 619-629.
26. Mira A., Klasson L., Andersson S. G., (2002), Curr. Opin. Microb., 5, 506-512.
27. Senapathy P., (1986), Proc. Natl. Acad. Sci. USA, 83, 2133-2137.
28. Fickett J. W., (1994), Comput. & Chem., 18, 203-205.
29. Barrell B. G., Air G. M., Hutchison C. A., (1976), Nature, 264, 34-41.

30. Ikehara K., Okazawa E., (1993), Nucleic Acids Res., 21, 2193-2199.
31. Merino E., Balbas P., Puente J. L., Bolivar F., (1994), Nucleic Acids Res., 22, 1903-1908.
32. de Antonio A., D'Angelo M., Dal Pero F., Sartorello F., Pandolfo D., Pallavicini A., Lanfranchi G., Valle G., (1997), Yeast, 13, 261-266.
33. Pavesi A., de Iaco B., Granero M. I., Porati A., (1997), J. Mol. Evol., 44, 625-631.
34. Rother K. I., Clay O. K., Bourquin J.P., Silke J., Schaffner W., (1997), Biol. Chem., 378, 1521-1530.
35. Cebrat S., Mackiewicz P., Dudek M. R., (1998), Biosystems, 42, 165-176.
36. Wang L. F., Park S. S., Doi R. H., (1999), J. Bacteriol., 181, 353-356.
37. Iwabe N., Miyata T., (2001), Gene, 280, 163-167.
38. Behrens M., Sheikh J., Nataro J. P., (2002), Infect. Immun., 70, 2915-2925.
39. Rogozin I. B., Spiridonov A. N., Sorokin A. V., Wolf Y. I., Jordan I. K., Tatusov R. L., Koonin E. V., (2002), Trends Genet., 18, 228-232.
40. Yelin R., Dahary D., Sorek R., Levanon E. Y., Goldstein O., Shoshan A., Diber A., Biton S., Tamir Y., Khosravi R., et al., (2003), Nat. Biotechnol., 21, 379-386.
41. Boldogkoi Z., (2000), J. Mol. Evol., 51, 600-606.
42. Boldogkoi Z., Murvai J., (1994), Virus Genes, 9, 47-51.
43. Boldogkoi Z., Murvai J., Fodor I., (1995), Trends Genet., 11, 125-126.
44. Ikehara K., Amada F., Yoshida S., Mikata Y., Tanaka A., (1996), Nucleic Acids Res., 24, 4249-4255.
45. Silke J., (1997), Gene, 194, 143-155.
46. Boldogkoi Z., Barta E., (1999), Biosystems, 51, 95-100.
47. Cebrat S., Dudek M. R., (1996), Trends Genet., 12, 12.
48. Mira A., Ochman H., Moran N. A., (2001), Trends Genet., 17, 589-596.
49. Lawrence J. G., Hendrix R. W., Casjens S., (2001), Trends Microbiol., 9, 535-540.
50. Homma K., Fukuchi S., Kawabata T., Ota M., Nishikawa K., (2002), Gene, 294, 25-33.
51. Liu Y., Harrison P.M., Kunin V., Gerstein M., (2004), Genome Biol., 5, R64.
52. Harrison P. M., Echols N., Gerstein M. B., (2001), Nucleic Acids Res., 29, 818-830.
53. Harrison P., Kumar A., Lan N., Echols N., Snyder M., Gerstein M., (2002), J. Mol. Biol., 316, 409-419.
54. Zhang Z., Harrison P., Gerstein M., (2002), Genome Res., 12, 1466-1482.
55. Harrison P. M., Milburn D., Zhang Z., Bertone P., Gerstein M., (2003), Nucleic Acids Res., 31, 1033-1037.
56. Torrents D., Suyama M., Zdobnov E., Bork P., (2003), Genome Res., 13, 2559-2567.
57. Ogata H., Audic S., Renesto-Audiffren P., Fournier P. E., Barbe V., Samson D., Roux V., Cossart P., Weissenbach J., Claverie J. M., Raoult D., (2001), Science, 293, 2093-2098.
58. Amiri H., Davids W., Andersson S. G. E., (2003), Mol. Biol. Evol., 20, 1575-1587.
59. Andersson J. O., Andersson S. G. E., (1999), Mol. Biol. Evol., 16, 1178-1191.
60. Andersson J. O., Andersson S. G. E., (2001), Mol. Biol. Evol., 18, 829-839.
61. Malpertuy A., Tekaia F., Casaregola S., Aigle M., Artiguenave F., Blandin G., Bolotin-Fukuhara M., Bon E., Brottier P., de Montigny J., et al., (2000), FEBS Lett., 487, 113-121.
62. Wood V., Rutherford K. M., Ivens A., Rajandream M.-A., Barrell B., (2001), Comp. Funct. Genom., 2, 143-154.
63. Domazet-Loso T., Tautz D., (2003), Genome Res., 13, 2213-2219.
64. Welch R. A., Burland V., Plunkett III G., Redford P., Roesch P., Rasko D., Buckles E. L., Liou S. R., Boutin A., Hackett J., et al., (2002), Proc. Natl. Acad. Sci., 99, 17020-17024.
65. Daubin V., Ochman H., (2004), Genome Res., 14, 1036-1042.
66. Schmid K. J., Tautz D., (1997), Proc. Natl. Acad. Sci. USA, 94, 9746-9750.
67. Schmid K. J., Aquadro C. F., (2001), Genetics, 159, 589-598.
68. Siew N., Fischer D., (2004), J. Mol. Biol., 342, 369-373.
69. Daubin V., Lerat E., Perriere G., (2003), Genome Biol., 4, R57.
70. Shmuely H., Dinitz E., Dahan I., Eichler J., Fischer D., Shaanan B., (2004), Bioinformatics, 20, 1248-1253.
71. Vitkup D., Melamud E., Moult J., Sander C., (2001), Nat. Struct. Biol., 8, 559-566.
72. Lee D., Grant A., Buchan D., Orengo C., (2003), Curr. Opin. Struct. Biol., 13, 359-369.

73. Chothia C., (1992), Nature, 357, 543-544.
74. Orengo C. A., Jones D. T., Thornton J. M., (1994), Nature, 372, 631-634.
75. Keese P. K., Gibbs A., (1992), Proc. Natl. Acad. Sci. USA, 89, 9489-9493.
76. Facchiano A., (1995), J. Mol. Evol., 40, 570-577.
77. Facchiano A., Facchiano F., van Renswoude J., (1993), J. Mol. Evol., 36, 448-457.
78. Johnson M. E., Viggiano L., Bailey J. A., Abdul-Rauf M., Goodwin G., Rocchi M., Eichler E. E., (2001), Nature, 413, 514-519.
79. Lipovich L., Hughes A. L., King M. C., Abkowitz J. L., Quigley J. G., (2002), Gene, 286, 203-213.
80. Chen L., DeVries A. L., Cheng C. H., (1997), Proc. Nat. Acad. Sci. USA, 94, 3811-3816.
81. Eddy S. R., (2001), Nat. Rev. Genet., 2, 919-929.
82. Wagner E. G., Flardh K., (2002), Trends Genet., 18, 223-226.
83. Mattick J. S., (2003), BioEssays, 25, 930-939.
84. Ochman H., (2002), Trends Genet., 18, 335-337.
85. Alimi J. P., Poirot O., Lopez F., Claverie J. M., (2000), Genome Res., 10, 959-966.
86. Gygi S. P., Rochon Y., Franza B. R., Aebersold R., (1999), Mol. Cell. Biol., 19, 1720-1730.
87. Hutchison C. A. III, Peterson S. N., Gill S. R., Cline R. T., White O., Fraser C. M., Smith H. O., Venter J. C., (1999), Science, 286, 2165-2169.
88. Monchois V., Abergel C., Sturgis J., Jeudy S., Claverie J. M., (2001), J. Biol. Chem., 276, 18437-18441.
89. Abergel C., Monchois V., Chenivesse S., Jeudy S., Claverie J. M., (2000), Acta Cryst., D56, 1694-1695.