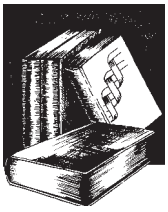


Od sekwencji do funkcji – poszukiwanie genów i ich adnotacje

Krystian Bączkowski, Paweł Mackiewicz, Maria Kowalczuk, Joanna Banaszak, Stanisław Cebrat

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

From sequence to function – looking for genes and their annotations

Summary

Today, we have very powerful and effective machines and methods to sequence and analyze DNA sequences. Almost every week, new genomes are added to sequence databases. However, those data are useless without additional annotations. Genes need to be found and their functions defined. Experimental work is too slow to analyze each sequence of a potential gene but computational methods facilitate such analyses. Here, we review the methodology, potential problems and constraints in genes finding and their annotation. We describe some new approaches including comparative genomics.

Key words:
genomics, genome, gene finding, gene annotation, gene function.

Adres do korespondencji

Krystian Bączkowski,
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) 22–44 2005

1. Wstęp

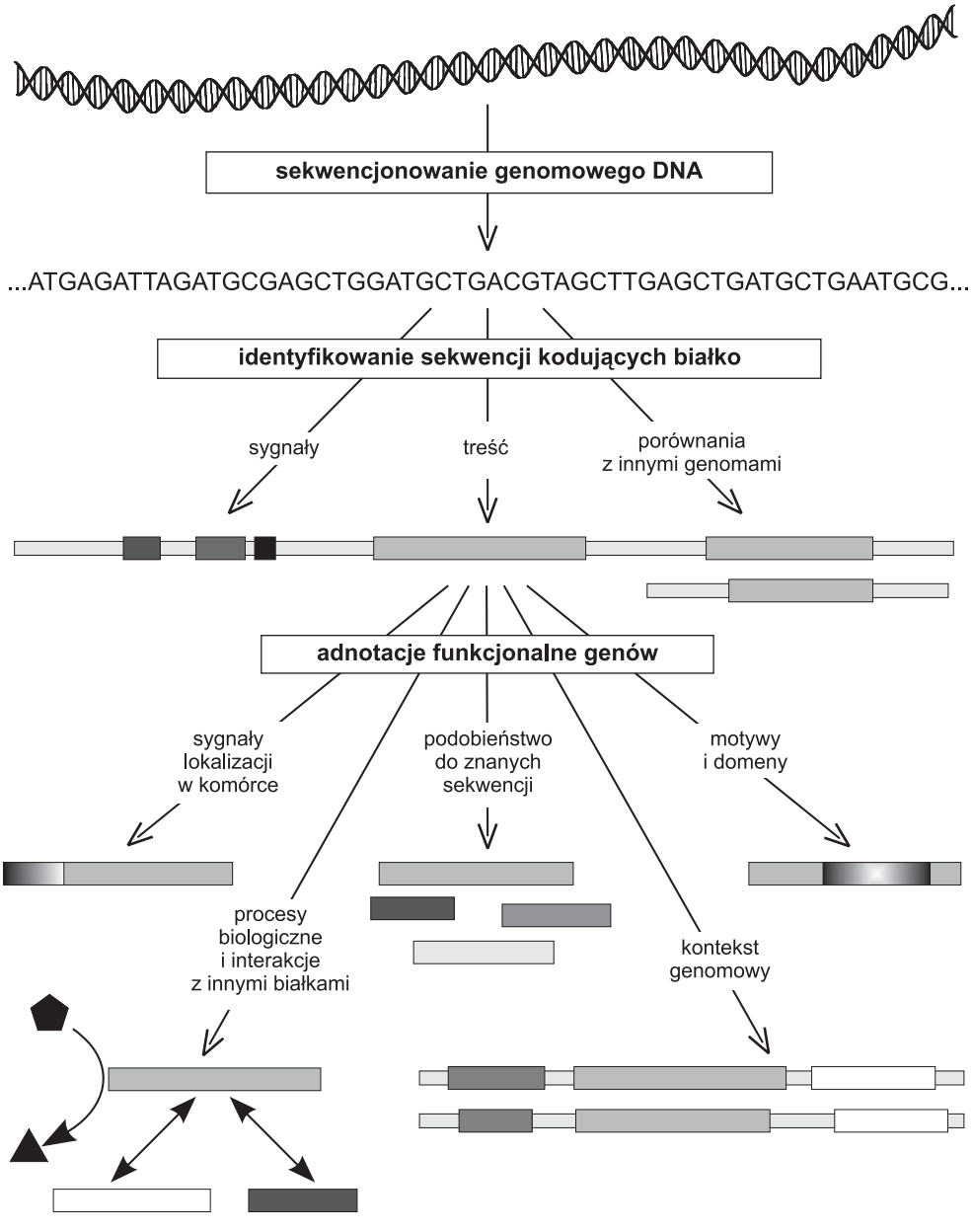
Od czasu pierwszych projektów sekwencjonowania metodą Sangera (1), techniki określania kolejności nukleotydów w sekwencjach bardzo się rozwinęły dzięki automatyzacji i komputeryzacji. Obecnie zsekwencjonowanie genomu bakteryjnego zajmuje tylko kilkadziesiąt dni. Poczynając od 1999 r. liczba poznawanych genomów podwaja się średnio co 15 miesięcy, a co miesiąc publikowane są sekwencje średnio czterech genomów (2). Jednakże uzyskanie sekwencji genomu jest dopiero początkiem w próbach zrozumienia podstaw funkcjonowania organizmu na

poziomie molekularnym. Kolejnym etapem w rozszyfrowywaniu genomu jest odnalezienie w nim sekwencji kodujących i przypisanie kodowanym przez nie produktom funkcji pełnionych w organizmie, czyli dokonanie ich adnotacji na podstawie analiz bioinformatycznych i doświadczalnych. Etapy poszukiwania sekwencji kodujących białko i przypisywania im funkcji należą obecnie do podstawowych i niezbędnych zadań w analizie każdego genomu (rys. 1).

2. Poszukiwanie sekwencji kodujących białko

Poszukiwanie sekwencji kodujących u organizmów prokariotycznych różni się od analiz sekwencji organizmów eukariotycznych. Różnica polega na tym, że w genomie *Prokaryotów* jest mało przestrzeni międzygenowych, w związku z tym geny są upakowane. Natomiast u *Eukaryotów* geny są bardzo rozproszone. Sekwencje kodujące stanowią w genomie bakterii *Escherichia coli* 88% genomu (3), u drożdży – około 70% (4), a u człowieka – 1-2% (5). Ponadto u *Prokaryotów* występują geny ciągłe, nieposiadające intronów, a u *Eukaryotów* geny są podzielone na eksony i introny o różnej liczbie i wielkości. Ponadto eksony mogą być rozproszone na dużym odcinku chromosomu i mogą występować eksony niekodujące w regionie 5'UTR (*UnTranslated Region*). U obu typów organizmów problemem jest identyfikacja krótkich sekwencji kodujących ze względu na małą istotność statystyczną różnic między ich sekwencjami a sekwencjami losowymi. Mogłoby się wydawać, że identyfikowanie genów u *Prokaryotów* jest łatwiejsze, jednak dodatkowe problemy w analizach tych genomów związane są z zachodzeniem na siebie potencjalnych sekwencji kodujących. Pierwotnie przyjmuje się, że w takich układach kodującym jest ORF dłuższy (6). Jednak istnieją przypadki, w których koduje ORF krótszy (7,8), lub kodują obydwa ORF-y (9,10). Właściwe rozpoznanie genów utrudnia również obecność pseudogenów (11), występujących w sekwencjach *Eukaryotów* znacznie częściej niż w sekwencjach *Prokaryotów*.

Jednym z pierwszych sposobów odnalezienia potencjalnych sekwencji kodujących białko, szczególnie u *Prokaryotów* i prostych *Eukaryotów* (np. drożdży), jest odnalezienie otwartych ramek odczytu – ORF-ów (*Open Reading Frame*). Algorytmy programów wyszukujących ORF-y są proste. Ich działanie polega na wyszukiwaniu w sekwencji genomu, czytanego w sześciu fazach, kodonów startu translacji (najczęściej ATG lub GTG, TTG i CTG) oraz jednego z trzech kodonów stopu translacji (TGA,TAA,TAG). Fragment sekwencji znajdujący się między tymi kodonami jest ORF-em i jest traktowany jako potencjalnie kodujący białko. Niekodujące ramki odczytu mogą w sekwencji DNA zostać wygenerowane przypadkowo, a im są krótsze, tym prawdopodobieństwo ich przypadkowego pojawienia się jest większe (7,8,12-15). Z tego powodu w większości analiz wybierane są najczęściej tylko ORF-y dłuższe niż 100 kodonów (16,17).



Rys. 1. Etapy analizy genomowego DNA: sekwencjonowanie, poszukiwanie sekwencji kodujących białko, przypisywanie genom funkcji. Wyróżniono najważniejsze metody w identyfikowaniu genów i dokonywaniu ich adnotacji funkcjonalnych, które opisano w tekście.

Po określeniu zbioru ORF-ów identyfikuje się tzw. elementy „zanieczyszczające", czyli fragmenty sekwencji wirusów zintegrowanych z genomem bakterii, sekwencje wektorów użytych w trakcie sekwencjonowania, ruchome elementy genetyczne, sekwencje powtórzone i sekwencje pseudogenów. Jest to etap trudny, ponieważ struktury te nie są jeszcze dobrze poznane (4).

Istnieją trzy podejścia do rozpoznawania genów w zsekwencjonowanych genomach: na podstawie „treści" (zawartości), sygnałów (kontekstu) i podobieństwa do innych sekwencji. W pierwszym z nich analizowane są specyficzne właściwości DNA związane z kodowaniem białka, jak: skład nukleotydowy, asymetria, preferencyjne wykorzystanie kodonów i innych „słów" DNA (oligomerów). Sekwencje kodujące istotnie różnią się pod tym względem od sekwencji niekodujących. W drugim podejściu wyszukiwane są sygnały związane z ekspresją sekwencji kodującej, powiązane z procesami transkrypcji i translacji. Trzecie podejście obejmuje wyszukiwanie w bazach danych sekwencji podobnych do analizowanego odcinka DNA oraz inne analizy porównawcze genomów (za pomocą programów BLAST, FASTA, Critica i in.).

2.1. Poszukiwanie sekwencji kodujących na podstawie „treści"

W związku z funkcją kodowania białka, sekwencje kodujące charakteryzują się inną strukturą niż sekwencje niekodujące. W sekwencjach kodujących stwierdzono obecność rytmu o okresie 3 pz związanego ze strukturą trójkową kodu genetycznego (18-22). Inną powszechnie znaną właściwością sekwencji kodujących jest to, że zawierają dużo puryn: adeniny i guaniny (18,23-27). Ważną cechą jest to, że skład nukleotydowy jest specyficzny dla danej pozycji w kodonie (18,28-34). Generalnie, w pierwszej pozycji w kodonach dominuje adenina i guanina, natomiast w drugiej pozycji przeważa adenina i cytozyna. Dobrym parametrem identyfikującym sekwencje kodujące jest asymetria DNA opisywana różnicami między komplementarnymi nukleotydami w poszczególnych pozycjach w kodonie (32,35,36). Właściwości kodujące przekładają się również na specyficzną używalność kodonów i aminokwasów (30,37). Zróżnicowana jest używalność kodonów synonimicznych, kodujących ten sam aminokwas (38-39) związana z poziomem szybkości translacji białek (39-41).

Do identyfikowania sekwencji kodujących można stosować różne kryteria (prace przeglądowe: 42-44). Najbardziej popularne i skuteczne okazało się określanie używalności heksamerów (dwukodonów) oraz zależność między pozycjami nukleotydów w sekwencji – modele Markowa (45-47). Jest to metoda statystyczna pozwalająca na odszukiwanie ukrytych wzorców w różnego rodzaju ciągach znaków. W metodzie tej zakłada się, że prawdopodobieństwo pojawienia się w sekwencji danego nukleotydu jest zależne od poprzedzających go nukleotydów. W analizach genomu najczęściej stosuje się pięcio- i sześciorzędowe modele. Prawdopodobieństwo pojawienia się danego nukleotydu jest w nich zależne od odpowiednio poprzedzającego go penta- lub heksameru. W programie Glimmer wprowadzono in-

terpolowane modele Markova (IMM) (48). Tak zmodyfikowana metoda pozwala na uzyskanie bardziej wiarygodnych wyników. W tradycyjnych modelach stosujących tylko jeden typ oligomeru o danej długości (np. pentamery) może zabraknąć danych do obliczenia prawdopodobieństwa ich występowania. Model IMM omija ten problem przez wykorzystywanie oligomerów o różnej długości w zależności od częstości ich występowania.

Algorytmy rozpoznające geny wykorzystujące ukryte modele Markova muszą zostać wcześniej „wytrenowane" na wzorcowym, reprezentatywnym zestawie sekwencji kodujących i niekodujących, ponieważ „uczą się" charakterystycznych właściwości genów i sekwencji niekodujących. Istnieją odmiany programów przystosowane do rozpoznawania genów tylko określonego organizmu. Program GeneMark.HMM został „przetrenowany" na sekwencjach kodujących *E. coli* i najlepiej rozpoznaje geny właśnie tego organizmu (49). Istnieją również algorytmy tego typu służące do wyszukiwania sekwencji nabytych w wyniku poziomego transferu. Program GENMARK Genesis (50) automatycznie grupuje geny analizowanych genomów bakteryjnych i dla każdej utworzonej grupy opracowuje odrębny łańcuch Markova. Za pomocą tego programu stwierdzono, że około 15% genów w genomie *E. coli* pochodzi z bocznego transferu (51). W tabeli 1 zamieszczono listę programów związanych z identyfikowaniem genów i przewidywaniem ich struktury.

Tabela 1

Wybrane programy służące do identyfikowania genów

Nazwa	Organizmy	Stosowane metody	Adres internetowy, http://
AUGUSTUS	*Arabidopsis*, człowiek, *Drosophila*	HMM	augustus.gobics.de/
EasyGene	*Prokaryota*	HMM, H	cbs.dtu.dk/services/EasyGene/
EcoParse	*Prokaryota, E. coli*	HMM	ecoparse@cse.ucsc.edu
ETOPE	ssaki	stosunek liczby podstawień synonimicznych do niesynonimicznych	Nekrut.uchicago.edu/etope
EuGene	*Arabidopsis*	IMM, PD	www.inra.fr/bia/T/EuGene/
FGENEH	*Eukaryota*	HMM, PD, LDA	www.softberry.com/berry.phtml
GeneHacker	*Prokaroyta*	HMM	www-btls.jst.go.jp/GeneHacker/
GeneId3	kręgowce, rośliny	WAM, HMM, PD, AD	www1.imim.es/geneid.html
GeneLang	*Drosophila*, dwuliścienne, kręgowce	metody lingwistyczne, HMM, PD, WAM	arete.ibb.waw.pl/PL/html/gene_lang.html
GeneMark	*Prokaroyta, Eukaryota*	HMM	opal.biology.gatech.edu/GeneMark/genemark24.cgi
Genie	człowiek, *Drosophila*	GHMM, NN, PD, H	www.fruitfly.org/seq_tools/genie.html
GenomeScan	kręgowce	GHMM, PD, H	genes.mit.edu/genomescan.html

GENSCAN	*Arabidopsis,* kręgowce, kukurydza	GHMM, WAM, MDD, PD, H	genes.mit.edu/GENSCANinfo.html
GlimmerHMM	*Eukaryota*	GHMM, IMM, DD	tigr.org/software/GlimmerHMM/index.shtml
GlimmerM	*Arabidopsis, Oryza sativa, Plasmodium falciparum*	IMM, PD	ftp.tigr.org/pub/software/GlimmerM
GrailEXP	*Arabidopsis*, człowiek, *Drosophila*, mysz	NN, PD, H	grail.lsd.ornl.gov/grailexp/
HMMER	*Eukaryota*	HMM, H	hmmer.wustl.edu/
HMMGene	*C. elegans*, kręgowce	HMM	cbs.dtu.dk/services/HMMgene/
MORGAN	kręgowce	DD	www.tigr.org/~salzberg/morgan.html
MZEF	*Arabidopsis*, człowiek, drożdże, mysz	AD	industry.ebi.ac.uk/~thanaraj/MZEF-SPC.html
PROCRUSTES	kręgowce	określa strukturę genów przez porównanie z innymi genami, DP	www-hto.usc.edu/software/procrustes/wwwserv.html
Sgp2	*Eukaryota*	HMM, H, PD	genome.imim.es/software/sgp2/index.html
SNAP	*Eukaryota*	HMM	homepage.mac.com/iankorf/
SORFIND	*Eukaryota*	macierze, miara Fouriera	iubio.bio.indiana.edu:7780/archive/00000124/
TigrScan	*Eukaryota*	GHMM, IMM	tigr.org/software/pirate/tigrscan/TigrScan-menu.html
TWAIN	*Eukaryota*	GPHMM, H	www.tigr.org/software/pirate/twain/twain.html
Twinscan	człowiek, mysz	GHMM, HMM, WAM	genes.cs.wustl.edu/
Veil	kręgowce	HMM, PD	www.tigr.org/~salzberg/veil.html
Xpound	człowiek	HMM	bioweb.pasteur.fr/seqanal/interfaces/xpound-simple.html

AD – analizy dyskryminacyjne, DD – drzewa decyzyjne, GHMM – ogólne modele Markova, H – poszukiwanie homologów, HMM – ukryte modele Markova, IMM – interpolowane modele Markova, MDD, *Maximal Dependence Decomposition*, NN – sieci neuronowe, PD – programowanie dynamiczne, WAM, *Weigth Array Matrix*.

2.2. Wyszukiwanie sygnałów związanych z sekwencjami kodującymi

Istnieje szeroka pula programów (tab. 2) wyszukujących sygnały, czyli charakterystyczne sekwencje nukleotydowe składające się na system regulacji ekspresji genów: procesy inicjacji i regulacji transkrypcji, inicjacji translacji, wiązania rybosomów, wycinania intronów, modyfikacji transkryptów (rys. 2). Niektóre z tych sekwencji są krótkie, jednak są na tyle konserwatywne, że możliwe jest ich rozpoznawanie.

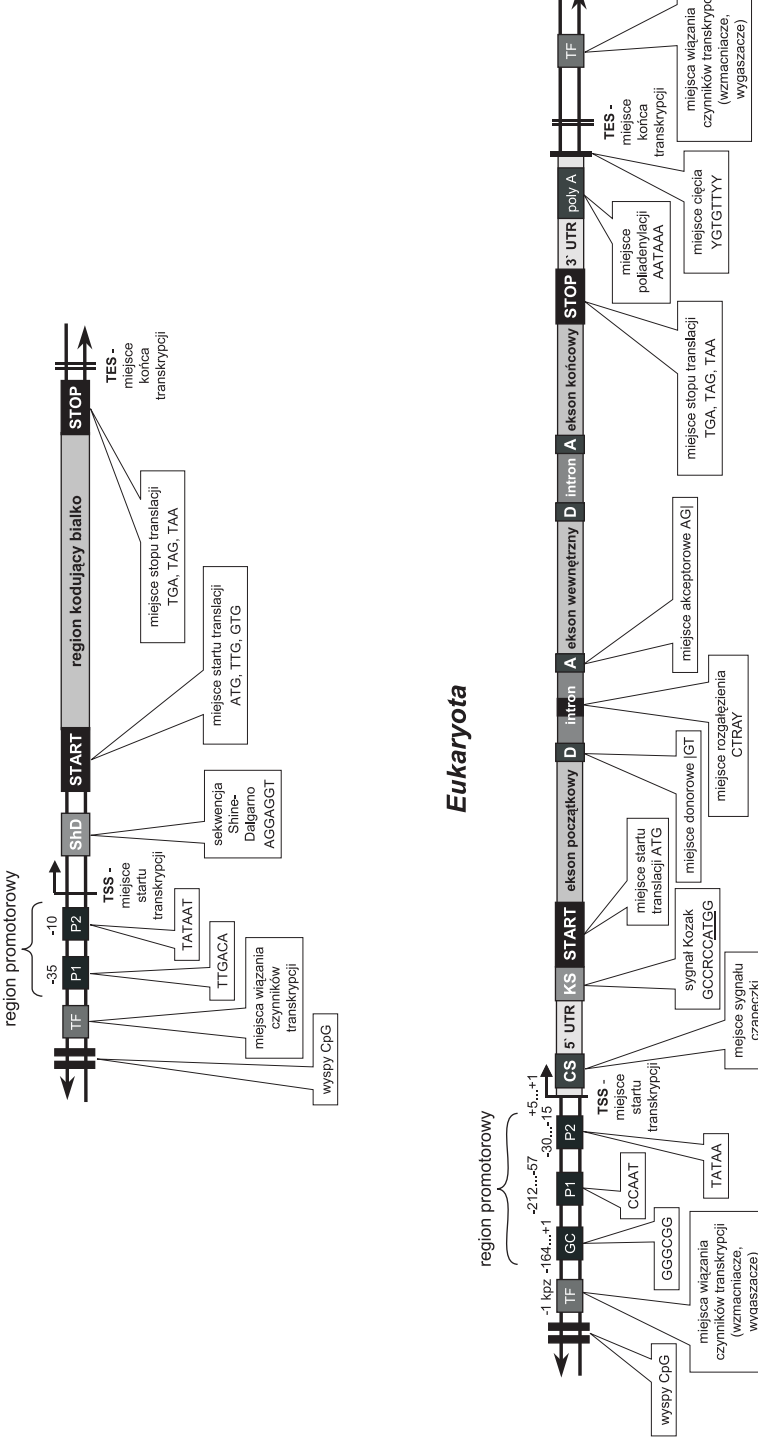
Tabela 2

Programy służące do identyfikowania sygnałów związanych z kodowaniem

Nazwa	Organizmy	Identyfikowany sygnał	Adres internetowy, http://
GeneSplicer	*Arabidopsis*, człowiek, *Drosophila*, *Oryza sativa*, *Plasmodium falciparum*	miejsca łączenia eksonów	www.tigr.org/tdb/GeneSplicer/gene_spl.html
polyadq	*Eukaryota*	miejsca poliadenylacji	rulai.cshl.org/tools/polyadq/polyadq_form.html
RBSFinder	*Prokaryota*	miejsca wiązania rybosomów	www.tigr.org/software/
RepeatMasker	*Eukaryota*	maskuje regiony o słabej złożoności	repeatmasker.org/
Signal Scan	*Prokaryota, Eukaryota*	regiony promotorowe w bazie TRANSFAC i TFD	www-bimas.cit.nih.gov/molbio/signal/
SplicePredictor	*Arabidopsis*, kukurydza	miejsca łączenia eksonów	bioinformatics.iastate.edu/cgi-bin/sp.cgi
TESS	*Eukaryota, Prokaryota*	miejsca wiązania czynników transkrypcyjnych	cbil.upenn.edu/tess/
TransTerm	*Prokaryota*	miejsca terminacji transkrypcji	www.tigr.org/software/

Elementy składowe systemu promotorowego wykazują mniejszą zmienność w organizmach prokariotycznych niż u eukariotycznych, przez co są lepszymi wyznacznikami obecności sekwencji kodujących. Jednakże nawet u *Prokaryotów* najczęściej wykorzystywane sekwencje TATAAT (okno Pribnova) oraz sekwencje TTGACA (52) odnajdywane są tylko w około 70% promotorów (53). W genomach organizmów eukariotycznych sygnały te charakteryzują się większą zmiennością położenia i są bardziej zróżnicowane. Szczególnie zmienne i specyficzne dla poszczególnych genów są miejsca wiązania czynników transkrypcji (wzmacniacze, wygaszacze), dlatego nie mogą być one uniwersalnie stosowane do rozpoznawania genów (54). Oprócz sygnałów związanych z transkrypcją, do wyznaczania początków sekwencji kodujących wykorzystuje się również sygnały translacji, miejsca wiązania rybosomów (53,55). U *Prokaryotów* są to sekwencje Shine-Dalgarno: AGGAGG (56), a u *Eukaryotów* najbardziej znanym sygnałem jest sygnał Kozak charakterystyczny dla kręgowców (57). U różnych grup organizmów sygnały te są jednak dosyć zróżnicowane (58).

Istnieje grupa sygnałów charakterystyczna tylko dla genomów organizmów eukariotycznych i nieposiadająca odpowiedników u prokariotycznych: sygnały czapeczki, poliadenylacji, wzmacniacze i wygaszacze ekspresji oraz sygnały wyznaczające granice między intronami i eksonami (rys. 2). Programy identyfikujące geny *Eukaryotów* biorą pod uwagę cztery rodzaje sekwencji kodujących: eksony inicjujące (od kodonu start do pierwszego miejsca łączenia 5'), eksony wewnętrzne (od miejsca łączenia 3' do miejsca łączenia 5'), eksony końcowe (od miejsca łączenia 3' do



Rys. 2. Struktura genów u organizmów prokariotycznych i eukariotycznych. Na schematach zaznaczono wybrane sygnały wykorzystywane w identyfikowaniu genów i podano konsensusy sekwencji charakterystyczne dla organizmów modelowych lub najczęściej stosowane. Liczby powyżej sekwencji oznaczają jej położenie względem miejsca początku transkrypcji. Nie zachowano proporcji wielkości między poszczególnymi rodzajami sekwencji (np. introny są w rzeczywistości o wiele dłuższe niż eksony).

kodonu stop) oraz sekwencje niepodzielone – bez intronów (53). Problemem w identyfikowaniu granic eksonów jest ich duże zróżnicowanie. Nie zawsze obecne są klasyczne dwunukleotydy miejsc donorowych GT i akceptorowych AG. Również wysoka zmienność miejsc inicjacji transkrypcji TSS (*transcription start site*) powoduje, że identyfikacja eksonów inicjacyjnych jest najtrudniejsza. Do określenia końca genu podzielonego, wykorzystuje się poza kodonem stop, sygnał poliadenylacji, który ma prostszą i bardziej konserwatywną strukturę. Jest heksamerem AATAAA znajdującym się 20-30 pz za sekwencją kodującą genu. Jednak sygnał ten nie jest jednoznacznym wyznacznikiem, ponieważ badania nad sekwencjami EST (*Expressed Sequence Tag*) wykazały, że w ponad połowie wszystkich regionów 3' UTR brak jest tego heksameru (59).

W przeprowadzonych testach, program GENSCAN wykorzystując opisane sygnały określił poprawnie 66% miejsc inicjacji transkrypcji i około 78% kodonów terminacyjnych, ze specyficznością odpowiednio 84 i 91%. Mimo że te wyniki są zadowalające, to jednak identyfikacja eksonów inicjujących i terminacyjnych jest o wiele gorsza niż identyfikacja eksonów wewnętrznych oparta głównie na identyfikacji sygnałów miejsc łączenia (53).

Programy rozpoznające miejsca funkcjonalne zarówno w genomach *Prokaryotów*, jak i *Eukaryotów*, opierają się na poszukiwaniu konsensusu (zgodności) sekwencji, stosują macierze wag pozycji (PWM, *position weight matrix*), np. PromoterScan (60), wykorzystują sieci neuronowe (61-63), np. NetGene (64), analizy lingwistyczne (65), analizę Fouriera (66) oraz modele Markova (45-47).

Grupa wykorzystywanych sygnałów jest bardzo zróżnicowana i zwiększa się w miarę odkrywania kolejnych sposobów regulacji ekspresji genów, co pociąga za sobą konieczność ciągłej modyfikacji używanych narzędzi. Dużym wyzwaniem dla komputerowej identyfikacji genów jest uwzględnienie dodatkowych zjawisk zachodzących u *Eukaryotów*, jak: alternatywna obróbka mRNA, transsplicing, redagowanie RNA, alternatywna transkrypcja i translacja. Procesy te powodują, że przy stosunkowo małej liczbie sekwencji kodujących można uzyskać duże zróżnicowanie proteomu.

2.3. Analizy porównawcze

Dostępność wielu zsekwencjonowanych genomów umożliwia poszukiwanie genów przez porównywanie ich sekwencji. Wykazanie podobieństwa między badaną, a już znaną sekwencją kodującą, może świadczyć o tym, że sekwencja badana koduje. Do algorytmów wyszukujących sekwencje podobne, zdeponowane w bazach danych należy BLAST (67) oraz FASTA (68). Poszukiwania można przeprowadzać na poziomie nukleotydowym, jednak lepsze wyniki uzyskuje się porównując sekwencje aminokwasowe ze względu na ich większą konserwatywność. W tym przypadku badane sekwencje nukleotydowe i te zdeponowane w bazach danych tłumaczy się

w sześciu fazach na sekwencje aminokwasowe za pomocą odpowiednich wersji programów. Takie analizy są utrudnione w przypadku najczęściej analizowanych sekwencji EST, które są często fragmentaryczne, zanieczyszczone pozostałym niekodującym DNA, wykazują błędy wynikające z procedur sekwencjonowania. Alternatywne składanie eksonów również znacznie utrudnia odnalezienie satysfakcjonujących podobieństw między badanym DNA, sekwencjami EST i cDNA. Istnieje również ryzyko nieodnalezienia homologa w bazie EST do analizowanej sekwencji z powodu słabej reprezentacji sekwencji EST wynikającej z ich ekspresji w ściśle określonych warunkach (4). Również elementy powtórzone znacznie utrudniają analizy porównawcze. Ich zawartość w genomach *Eukaryotów* jest znaczna. Szacuje się, że w genomie człowieka, nawet 40% sekwencji DNA stanowią tego typu elementy (69). Pomocne przy identyfikowaniu genów u *Eukaryotów* jest porównywanie sekwencji genomów blisko spokrewnionych, u których ułożenie genów (syntenia) i ich struktura powinny być bardzo podobne. Przy porównywaniu wielu genomów można również o wiele łatwiej identyfikować konserwatywne regiony regulatorowe (70).

Jedną z niedawno wprowadzonych dodatkowych metod pozwalających na rozróżnienie sekwencji kodujących od niekodujących, jest szacowanie potencjalnej liczby podstawień nukleotydowych w miejscach synonimicznych, w których podstawienie nie powoduje zmiany kodowanego aminokwasu i niesynonimicznych, w których zmiana powoduje zmianę kodowanego aminokwasu. W związku z tym sekwencje kodujące powinny charakteryzować się większą liczbą podstawień w miejscach synonimicznych niż niesynonimicznych. Metoda ta znalazła zastosowanie nie tylko do testowania, czy dana sekwencja koduje (71), ale również do identyfikowania nowych eksonów u człowieka (72). Z powodzeniem może być ona stosowana u innych organizmów, również w przypadku krótkich ORF-ów bakterii (73).

2.4. Aplikacje zintegrowane

Stosowanie opisanych metod osobno, nie zawsze daje dobre wyniki przewidywania genów. Aby przeprowadzone analizy były pełniejsze i bardziej wiarygodne, większość używanych współcześnie aplikacji łączy je (Grail, GeneFinder, GeneScan, Genie) – tabela 1. Bardziej zaawansowane systemy tego typu wykorzystują sieci neuronowe (Grail), analizy drzew decyzyjnych (MORGAN), analizy dyskryminacyjne (MZEF), programowanie dynamiczne (GlimmerM), ukryte modele Markova i inne metody. Aplikacje te nie są pozbawione wad. Waga różnych parametrów i sygnałów musi być dokładnie obliczona, ponieważ może się zdarzyć sytuacja, w której pojedyncza, mało znacząca cecha może zadecydować o błędnym zaklasyfikowaniu analizowanej sekwencji. Może to doprowadzić do pomijania genów o pewnych właściwościach odbiegających od właściwości zbioru ogólnego.

2.5. Ocena algorytmów rozpoznających geny

Opracowano dwa parametry, które pozwalają określić poprawność danego algorytmu rozpoznającego geny. Czułość jest zdolnością do znajdowania możliwie jak największej liczby sekwencji kodujących, selektywność natomiast – zdolnością do ignorowania sekwencji niekodujących jako przypadków fałszywie pozytywnych (74). Mimo wielu jeszcze nie rozwiązanych kwestii, efektywność programów odnajdujących geny u *Prokaryotów* jest wysoka. Najlepsze programy odnajdują sekwencje kodujące białko z czułością i specyficznością na poziomie ponad 98% (75). W przypadku *Eukaryotów* wyniki te nie są aż tak zadowalające. Przetestowano dwanaście algorytmów stosujących ukryte modele Markova. Zadaniem testowanych aplikacji było określenie struktury genowej znanego i opisanego odcinka DNA pochodzącego z genomu *Drosophilla melanogaster*. Najlepsze ze sprawdzanych algorytmów wykazały specyficzność i czułość na poziomie 90-95% w przypadku, w którym próbowano określić, czy dany nukleotyd jest częścią eksonu. Jednak, gdy podjęto próbę odnalezienia granic ekson – intron, czułość gwałtownie spadła. W przypadku próby dokładnego określenia struktury podzielonego genu, otrzymane wartości są jeszcze niższe. Czułość spada do 40%, a specyficzność do 30% w najlepszym z badanych algorytmów. Od 5 do 15% genów zostało zupełnie pominiętych (53). Wyniki te uzmysławiają, ile błędów znajduje się w bazach danych sekwencji organizmów eukariotycznych (76). Dane tego typu są prawdopodobnie obarczone o wiele większym poziomem błędu w przypadku genomu ludzkiego, ponieważ wraz ze wzrostem długości rejonów międzygenowych czułość zastosowanych algorytmów spada (77). Błędnie zidentyfikowane geny występują również w genomach prokariotycznych, dlatego zasadne jest ponowne przeprowadzenia ich identyfikacji (78,79).

Z powodu niedoskonałości metod identyfikujących geny, liczby adnotowanych genów w poszczególnych genomach mogą odbiegać od rzeczywistości. Dotyczy to nie tylko dużych genomów eukariotycznych (np. u człowieka zakres szacowanej liczby genów wynosi 30 000-100 000), ale również mniejszych genomów mikroorganizmów. Wynika to z tego, że często przyjmowane kryterium długości (> 100 kodonów) jako wyznacznika ORF-ów kodujących może prowadzić do uwzględniania w bazach ramek niekodujących zwłaszcza o długości 100-150 kodonów (36,80-83). Szacowane liczby genów uzyskane na bazie rozkładów długości i podobieństwa do znanych genów wskazują, że w genomach mikroorganizmów liczba ORF-ów jest zawyżona o 10-30% (81). Przykładowo, w genomie *E. coli* szacowna liczba genów wynosi około 3771 przy całkowitej liczbie ORF-ów w bazie 4289. Natomiast w genomie *Aeropyrum pernix* za geny uznano pierwotnie wszystkie ORF-y dłuższe niż 100 kodonów (84), podczas gdy na podstawie szacowań dowodzi się, że tylko około 50% z nich to sekwencje kodujące. W intensywnie badanym organizmie modelowym – drożdży *Saccharomyces cervisiae* pierwotnie zidentyfikowano 6275 ORF-ów kodujących (6). Natomiast na podstawie szacowanych liczb sekwencji kodujących opartych na analizach porównawczych genomów i wykorzystujących specyficzne

właściwości genów w składzie wskazuje się, że jest ich 5322-5651 (36,81,85-88). Jednakże pozostały jeszcze do odkrycia krótkie ORF-y kodujące, które mogą te liczby nieco podwyższyć (89-91).

3. Adnotacje funkcjonalne

Po zidentyfikowaniu sekwencji kodujących rozpoczyna się etap ich adnotacji, czyli przypisywania odnalezionym genom (a właściwie ich białkowym produktom) funkcji i roli jaką pełnią w komórce. Najlepszą metodą byłyby szczegółowe badania eksperymentalne poszczególnych genów, jednakże są to badania kosztowne i czasochłonne. Na przykład, w modelowym organizmie *E. coli* co roku charakteryzuje się 20-30 genów (92). Zważywszy, że według bazy GenProtEC tylko 2370 genów (56% wszystkich ORF-ów) zostało dokładnie zbadanych eksperymentalnie, pełne scharakteryzowanie pozostałych może zająć jeszcze wiele dekad. W innym, intensywnie badanym genomie modelowym – drożdży *Saccharomyces cerevisiae* scharakteryzowanych jest według bazy SGD 4231 genów (64%). Inne genomy są o wiele słabiej zbadane pod tym względem. W wielu przypadkach, szczególnie wśród organizmów należących do królestwa *Archaea* nie przeprowadzano żadnych badań eksperymentalnych. Sposobem na rozwiązanie tych problemów jest zastosowanie metod obliczeniowych, które mogą przypisać funkcję z dużym prawdopodobieństwem poprawności lub przynajmniej zasugerować potencjalne funkcje i w związku z tym ukierunkować dalsze badania eksperymentalne. Stosuje się różne podejścia w adnotacjach funkcjonalnych, jak poszukiwanie sekwencji podobnych zdeponowanych w bazach danych, identyfikowanie motywów funkcjonalnych i domen strukturalnych w białkach, określanie cech strukturalnych związanych z lokalizacją białka w komórce, przypisywanie funkcji na podstawie kontekstu genomowego i przewidywanie procesu biologicznego (rys. 1).

3.1. Poszukiwanie sekwencji podobnych

Najprostszym podejściem w identyfikowaniu potencjalnej funkcji badanego genu jest przeszukanie baz danych w celu znalezienia sekwencji podobnych. Jeżeli znaleziona sekwencja lub sekwencje posiadają przypisaną już funkcję i wykazują duży stopień podobieństwa do badanej, można by sądzić, że nasz gen również pełni podobną funkcję. Analizy przeprowadza się częściej na poziomie aminokwasowym niż na nukleotydowym ze względu na większą konserwatywność sekwencji białkowych. Standardowo do przeszukiwania baz danych stosuje się program FASTA (68) i BLAST (67). Wykorzystywać można również bardziej zaawansowany PSI-BLAST, który wykonuje wielokrotne przeszukiwania bazy danych. Sekwencje znalezione w danym przeszukiwaniu są wykorzystywane do tworzenia macierzy PSSM (*posi-*

tion-specific scoring matrices) stosowanej w następnym etapie poszukiwań. Zwiększa to znacznie czułość poszukiwań. Jedną z lepszych baz stosowaną w tego typu analizach jest baza SWISS-PROT, która jest bazą pochodną w stosunku do baz pierwotnych – archiwizujących dane (GenBank, EMBL-EBI lub DDBJ). Mimo że zawiera ona mniejszą liczbę sekwencji, dane są w niej nadzorowane i dokładnie opisywane przez ekspertów, dzięki czemu adnotacje funkcjonalne gromadzonych w niej sekwencji są wiarygodne. Jednakże ze względu na czasochłonność takich adnotacji, wiele kompletnie zsekwencjonowanych genomów nie jest w ten sposób jeszcze opisanych.

Mimo starannych adnotacji sekwencji w bazach danych, informacja przenoszona na analizowaną homologiczną sekwencję nie zawsze jest właściwa. Nie ma uniwersalnych reguł mówiących przy jakim stopniu podobieństwa dwóch sekwencji można wnioskować o ich podobieństwie funkcjonalnym. Ponadto historia genów jest o wiele bardziej skomplikowana niż zwykłe pionowe dziedziczenie genów od przodka (ortologi). Wiele genów ulega duplikacji (paralogi), a ich funkcja może ulec zróżnicowaniu, co prowadzi do błędnego przenoszenia informacji z jednego białka na drugie. W związku z tym najbardziej wiarygodne, jak się wydaje, jest wnioskowanie o funkcji tylko na podstawie genów ortologicznych. Chociaż zidentyfikowanie ortologów jest trudne, powstało kilka baz zbierających białka w grupy ortologiczne (tab. 3). Najbardziej popularną jest baza COGs (*Clusters of Orthologous Groups of Proteins database*) – (93,94), która grupuje białka kodowane przez genomy kompletnie zsekwencjonowanych mikroorganizmów w konserwatywne rodziny. Rodziny te są dodatkowo dzielone na kilkanaście nadrzędnych grup funkcjonalnych. Każda grupa ma przypisaną funkcję i zawiera ortologiczne białka z przynajmniej trzech linii filogenetycznych, które najprawdopodobniej wyewoluowały z pojedynczego przodka. Przypisanie funkcjonalne badanego białka sprowadza się do jego klasyfikacji do jednej lub wielu (jeśli jest to białko wielodomenowe) grup ortologicznych na podstawie podobieństwa sekwencji.

Tabela 3

Bazy i narzędzia pomocne przy adnotacjach funkcjonalnych

Nazwa	Rodzaj gromadzonych danych i przeprowadzane analizy	Adres internetowy, http://
1	2	3
COGs	grupy białek ortologicznych pochodzące z kompletnie zsekwencjonowanych genomów mikroorganizmów i genomów eukariotycznych sklasyfikowane w grupy funkcjonalne	www.ncbi.nlm.nih.gov/COG
MBGD	podobnie jak COGs, ale inny algorytm grupowania białek w grupy	mbgd.genome.ad.jp
EGO	grupy genów ortologicznych organizmów eukariotycznych; zawiera również ortologi genów człowieka związanych z chorobami	www.tigr.org/tdb/tgi/ego/
STRING	identyfikuje przyległe grupy genów ortologicznych, fuzje genów, profile filogenetyczne w analizowanych genomach	string.embl.de/

1	2	3
SNAPer	identyfikuje przyległe grupy genów o zadanym podobieństwie w analizowanych genomach	pedant.gsf.de/snapper/
Predictome	identyfikuje przyległe grupy genów ortologicznych, fuzje genów, profile filogenetyczne w analizowanych genomach; uwzględnia również wyniki badań eksperymentalnych	predictome.bu.edu/
AllFuse	identyfikuje fuzje genów w analizowanych genomach	maine.ebi.ac.uk:8000/services/allfuse
FusionDB	identyfikuje fuzje genów w genomach prokariotycznych	igs-server.cnrs-mrs.fr/FusionDB/
KEGG	szlaki metaboliczne i geny ortologiczne kompletnie zsekwencjonowanych genomów	www.genome.ad.jp/kegg
WIT/ERGO	grupy białek ortologicznych powiązanych funkcjonalnie; szlaki metaboliczne; organizacja genów w operony	www-wit.mcs.anl.gov/wit3/ ergo.integratedgenomics.com/ERGO
BioCyc, EcoCyc, MetaCyc	szlaki metaboliczne różnych organizmów	biocyc.org/, metacyc.org/, ecocyc.org/
GO	klasyfikacja funkcji molekularnych, procesów biologicznych i lokalizacji białek w komórce	www.geneontology.org/
BIND	interakcje między białkami	bind.ca/
DIP	interakcje między białkami	dip.doe-mbi.ucla.edu/
MINT	interakcje między białkami	mint.bio.uniroma2.it/mint/

3.2. Identyfikowanie motywów i domen białek

Bardziej precyzyjnym sposobem adnotacji jest zidentyfikowanie w analizowanym białku motywu lub domeny związanych z określoną funkcją. Motywem określa się konserwatywny zestaw reszt aminokwasowych, które są istotne dla funkcji białka i znajdują się w niewielkiej odległości od siebie. Natomiast domena białkowa jest strukturalnie zwarta, tworzy stabilną strukturę przestrzenną i jest ewolucyjnie konserwatywna. Białka mogą posiadać wiele motywów lub domen, co wzbogaca ich adnotacje funkcjonalne. Istnieje wiele baz gromadzących informacje o motywach i domenach, posiadających narzędzia umożliwiające ich przeszukiwanie (95, tab. 4). Motywy i domeny opisywano i wyróżniano najczęściej na podstawie przyrównania znanych już białek o założonym stopniu podobieństwa. Jednakże w różnych bazach stosowano inny poziom automatyzacji i poprawek wprowadzanych przez ekspertów, dlatego adnotacje badanego białka mogą różnić się w zależności od zastosowanej bazy. Często informacje zawarte w wielu bazach pokrywają się zawierając dane o tych samych motywach lub domenach, które są różnie nazwane i sklasyfikowane. Dlatego bardzo użyteczne są bazy integrujące informacje pochodzące z innych baz. Są nimi InterPro i CDD.

Tabela 4

Bazy motywów i domen białkowych

Nazwa	Rodzaj gromadzonych danych	Adres internetowy, http://
BLOCKS	konserwatywne regiony białek (bloki) zebrane z przyrównania wielu sekwencji	blocks.fhcrc.org/blocks/
PROSITE	motywy sekwencji białek o podobnej funkcji biochemicznej	www.expasy.org/prosite/
PRINTS	zbiory krótkich motywów charakteryzujące funkcjonalnie białka na różnych poziomach podobieństwa	umber.sbs.man.ac.uk/dbbrowser/PRINTS/
eMOTIF	motywy pochodzące z PRINTS i BLOCKS	motif.stanford.edu/emotif/
DOMO	przyrównane domeny białek z bazy SWISS-PROT i PIR	www.infobiogen.fr/services/domo/
Pfam	zbiór profili opisujących domeny białek pogrupowanych w rodziny; utworzone z przyrównań wielu sekwencji w oparciu na HMM (ukryte modele Markova)	pfam.wustl.edu/
ProDom	domeny białkowe wygenerowane automatycznie z baz SWISS-PROT i TrEMBL przy wykorzystaniu PSI-BLAST i profili Pfam	prodes.toulouse.inra.fr/prodom/current/html/home.php
SBASE	domeny białkowe opracowane na podstawie literatury, białkowych baz danych i baz genomowych oraz podobieństwa według algorytmu BLAST	hydra.icgeb.trieste.it/sbase/
SMART	mobilne domeny białkowe, architektura domenowa białek	smart.embl-heidelberg.de/
TIGRFAMs	rodziny białek zidentyfikowane za pomocą HMM (ukrytych modeli Markova) zawierające profile Pfam	www.tigr.org/TIGRFAMs/
InterPro	baza integrująca inne bazy: Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs PIRSF, SUPERFAMILY i UniProt.	www.ebi.ac.uk/interpro/
CDD	konserwatywne domeny z bazy SMART, Pfam, COG oraz NCBI opisane przez PSSM (*Position-Specific Scoring Matrices*)	www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

3.3. Określanie sygnałów związanych z lokalizacją białka w komórce

Ważną informacją, która może zasugerować funkcję białka jest określenie jego lokalizacji w odpowiednich przedziałach komórki. Istnieje wiele sygnałów odpowiedzialnych za właściwe adresowanie syntetyzowanych białek i ich umieszczanie w odpowiednich regionach komórki (96). Są to najczęściej krótkie specyficzne motywy związane z kierowaniem białka do jądra, lizosomów lub peroksysomów, peptydy sygnałowe występujące w części N-końcowej (związane z eksportem białek do retikulum endoplazmatycznego lub na zewnątrz komórki), peptydy tranzytowe (kierujące białka do mitochondriów lub plastydów) oraz hydrofobowe regiony transbłonowe, związane z umieszczeniem lub kotwiczeniem białek w błonach. Powstało wiele programów identyfikujących te sygnały (tab. 5, tab. 6), jednak dokładność ich przewidywań nie zawsze jest wysoka i związana jest z dużą zmiennością oraz małą specyficznością niektórych sygnałów.

Tabela 5

Programy określające lokalizację białek w komórce

Nazwa	Rozpoznawana lokalizacja lub sygnał	Metoda	Adres internetowy, http://
PSORT	różna lokalizacja subkomórkowa u *Prokaryotów* i różnych grup *Eukaryotów*	MR, DD	psort.ims.u-tokyo.ac.jp/form.html
PSORTII	różna lokalizacja subkomórkowa u zwierząt i drożdży	k-NN	psort.ims.u-tokyo.ac.jp/form2.html
couple-subloc	różna lokalizacja subkomórkowa u *Prokaryotów* i *Eukaryotów*	SVM	bioinfo.tsinghua.edu.cn/CoupleLoc/
SubLoc	różna lokalizacja subkomórkowa u *Prokaryotów* i *Eukaryotów*	SVM	www.bioinfo.tsinghua.edu.cn/SubLoc/
NNPSL	cytoplazma, zewnętrzne komórki, mitochondria, jądro	NN	www.doe-mbi.ucla.edu/~astrid/astrid.html
iPSORT	SP, mTP, cTP	MR, DD	hc.ims.u-tokyo.ac.jp/iPSORT/
TargetP	SP, mTP, cTP	NN	www.cbs.dtu. dk/services/TargetP
Predotar	SP, mTP, cTP	NN	genoplante-info.infobiogen.fr/predotar/predotar.html
SignalP	SP *Prokaryotów* i *Eukaryotów*	HMM, NN	www.cbs.dtu.dk/services/SignalP/
SigFind	SP	NN	139.91.72.10/sigfind/sigfind.htm
MITOPRED	MTP u zwierząt lub drożdży i roślin	domeny Pfam	mitopred.sdsc.edu/
MitoProt	mTP	AD	ihg.gsf.de/ihg/mitoprot.html
PlasMit	mTP u *Plasmodium falciparum*	NN	gecco.org.chemie.uni-frankfurt.de/plasmit/
ChloroP	cTP	NN	www.cbs.dtu. dk/services/ChloroP
PCLR	cTP	PCA	apicoplast.cis.upenn.edu/pclr/
PATS	aTP u *Plasmodium falciparum*	NN	gecco.org.chemie.uni-frankfurt.de/pats/pats-index.php
PlasmoAP	aTP u *Plasmodium falciparum*	MR, DD	plasmodb.org/restricted/PlasmoAPcgi.shtml
predictNLS	jądro komórkowe	M	cubic.bioc.columbia.edu/predictNLS/
PTS1 predictor	peroksysomalny sygnał PTS1	AD	mendel.imp.univie.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp

aTP – peptyd kierujący do apikoplastu, cTP – chloroplastowy peptyd tranzytowy, mTP – mitochondrialny peptyd kierujący, SP – peptyd sygnałowy; AD – analiza dyskryminacyjna, DD – drzewa decyzyjne, HMM – ukryte modele Markova, k-NN – k-najbliższych sąsiadów, M – poszukiwanie motywów w bazie danych, MR – metoda reguł, NN – sieci neuronowe, PCA – analiza składowych głównych, SVM, *Support Vector Machine*.

Tabela 6

Programy przewidujące regiony transbłonowe

Nazwa	Adres internetowy, http://
1	2
DAS	www.sbc.su.se/ ~miklos/DAS/
HMMTop	www.enzim.hu/ hmmtop/
Memsat2	bioinf.cs.ucl.ac.uk /psipred/
PhDhtm, PhD topology	cubic.bioc. columbia.edu/pp/
SOSUI	sosui.proteome. bio.tuat.ac.jp/

1	2
TMAP	www.mbb.ki.se/ tmap/
TMHMM	www.cbs.dtu.dk/ services/TMHMM
TMpred	www.ch.embnet. org/software/TMPRED _form.html
TopPred2	www.sbc.su.se/ ~erikw/toppred2/

3.4. Identyfikowanie funkcji na podstawie kontekstu genomowego

Przypisywanie funkcji genu na podstawie kontekstu genomowego jest metodą pośrednią i uzupełniającą w stosunku do opisanych metod – opartych głównie na podobieństwie (97-99). Metody przewidujące funkcje genów w kontekście całego genomu rozwinęły się po nagromadzeniu wystarczającej liczby kompletnie zsekwencjonowanych genomów i oferują całkowicie nowe możliwości oparte na genomice porównawczej. Podczas gdy metody oparte na podobieństwie sekwencji przewidują funkcje molekularne białek, metody oparte na kontekście genomowym dają ogólne przewidywania i określają proces biologiczny, w który są one zaangażowane. Wyróżnia się trzy podejścia w tego typu analizach.

Jedno z tych podejść analizuje skład genów w genomie – stara się zidentyfikować geny, które mają tendencję do wspólnego występowania (lub zaniku) w poszczególnych genomach (93,100-102). Podzbiór genomów, w których występuje dany gen nazywany jest profilem lub wzorem filogenetycznym. Zakłada się, że geny, które są ze sobą funkcjonalnie powiązane powinny współwystępować w genomach lub powinno ich nie być. Takie geny powinny posiadać podobne profile filogenetyczne. Przykładem mogą być geny kodujące różne podjednostki tego samego enzymu lub biorące udział w kolejnych etapach tego samego szlaku metabolicznego. Takie podejście ma sens tylko dla kompletnie zsekwencjonowanych genomów. Wnioski wyciągane z tego typu analiz muszą być jednak starannie rozważane, ponieważ na profile filogenetyczne ma wpływ również redundancja funkcji (tzn. występowanie w genomie wielu genów komplementujących swoje funkcje), zjawisko nieortologicznego zastępowania genów (tzn. niespokrewniony lub odległy gen przejmuje funkcję innego genu), utrata genów specyficzna dla danych linii filogenetycznych i horyzontalny transfer genów. Wzory filogenetyczne bardzo łatwo można uzyskiwać na podstawie wielu baz i narzędzi (tab. 3).

Kolejna metoda wykorzystuje domenową strukturę białek kodowanych przez geny i wykorzystuje zjawisko ich łączenia się (fuzji) i rozszczepiania (103-105). Przyjmuje się, że łączenie się domen powinno być akceptowane przez selekcję, gdyby ułatwiało interakcje funkcjonalne, np. centrów katalitycznych uczestniczących kolejno w danym szlaku metabolicznym. Należy zatem oczekiwać, że białka, które uległy połączeniu u niektórych gatunków, mogą ze sobą oddziaływać fizycznie lub przynajmniej funkcjonalnie u innych gatunków, u których występują oddzielnie. Takie podejście zastosowano z powodzeniem na przykład do identyfikowania białek

systemu transdukcji sygnałów u *Prokaryota* (106). Wielodomenowe białka można łatwo identyfikować na podstawie baz danych wymienionych w tabeli 3 oraz bazy SMART uwzględniającej strukturę modułową białek.

Trzecia metoda wymaga analizy otoczenia genu w genomach. U *Prokaryotów* wiele genów powiązanych funkcjonalnie tworzy operony, dlatego sąsiedztwo genu może dostarczyć istotnych wskazówek dotyczących jego funkcji. Mimo wielu rearanżacji i zmian ułożenia genów, nawet wewnątrz operonów (107-109) obecność tych samych przyległych ciągów genów ortologicznych utrzymujących się w kilku genomach można uważać za istotną wskazówkę o potencjalnych wzajemnych funkcjonalnych oddziaływaniach produktów tych genów (110,111). Przykładowo takie białka mogą uczestniczyć w tym samym szlaku metabolicznym, mogą być podjednostkami tworzącymi większe białko lub jedno z nich może być enzymem, a drugie jego regulatorem. Istnieje kilka narzędzi internetowych, które umożliwiają analizy tego typu: STRING, SNAPper oraz narzędzia w bazach: WIT/ERGO, COG, KEGG, MBGD.

3.5. Adnotacje w aspekcie całej komórki

Ostatecznym etapem adnotacji funkcjonalnych badanego genu jest zrozumienie jego znaczenia dla komórki w kontekście innych produktów genów tego genomu, a zatem określenie w jakim procesie biologicznym jest on zaangażowany i z jakimi białkami oddziałuje. W tym celu stworzono bazy zbierające informacje o interakcjach białek i szlakach metabolicznych (tab. 3). Podane informacje są najczęściej przedstawiane w postaci grafów i sieci zależności. Dobrym systemem adnotującym geny jest system Gene Ontology (GO). Tworzy on hierarchiczny system pojęć i klasyfikuje różne aspekty funkcji genów na trzech poziomach: funkcja molekularna, proces biologiczny i lokalizacja w komórce. Na tym etapie adnotacji, poza analizami obliczeniowymi, uwzględnianych jest wiele informacji pochodzących z różnorodnych analiz eksperymentalnych przeprowadzanych w skali całego genomu, jak: kierowana mutageneza i dysrupcja genów, analiza ekspresji genów za pomocą chipów DNA, wyciszanie genów za pomocą interferencyjnego RNA, analiza interakcji między białkami w systemach dwuhybrydowych, lokalizowanie białek w komórce za pomocą różnych znaczników (np. fluorescencyjnych), identyfikowanie miejsc interakcji białek z kwasami nukleinowymi (selex), charakterystyka elektroforetyczna i strukturalna białek oraz identyfikowanie kompleksów białek za pomocą spektrometrii masowej.

3.6. Automatyzacja, błędy i udoskonalanie adnotacji całych genomów

Z powodu swojej złożoności, etap adnotacji jest bardzo żmudnym i czasochłonnym procesem, dlatego utworzono programy automatyzujące go. W większości przypadków przeprowadzają one opisane analizy – głównie oparte na podobień-

stwie sekwencji i poszukiwaniu domen oraz motywów. Do takich programów należy: GeneQuiz (www.sander.ebi.ac.uk/genequiz), PEDANT (pedant.gsf.de), SEALS (www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS), MAGPIE (genomes.rockefeller.edu), ERGO (ergo.integratedgenomics.com/ERGO), IMAGENE (www.abi.snv.jussieu.fr/research). Wiele kompletnie zsekwencjonowanych genomów mikroorganizmów przed zdeponowaniem w publicznych bazach danych jest wstępnie opisywanych właśnie za pomocą takich narzędzi. Automatyzacja procesu adnotacji znacznie ułatwia i przyspiesza dalsze analizy, jednak powoduje również dużo błędów, sięgających nawet 30% (112-117). Najwięcej błędów jest związanych z przypisywaniem adnotacji na podstawie podobieństwa sekwencji.

Najpoważniejszym błędem jest przypisanie badanemu białku funkcji zupełnie innej niż pełni ono w rzeczywistości. Dokonuje się to najczęściej na bazie jego przypadkowego podobieństwa do innego białka, pełniącego odmienną funkcję. Jest to najczęściej spowodowane obecnością regionów o słabej złożoności w porównywanych sekwencjach dających ich fałszywe podobieństwo. Należy je maskować za pomocą odpowiednich programów, np. SEG domyślnie włączonego przy analizach programem BLAST. Przypisanie błędnej funkcji występuje również, wtedy gdy badane białko jest podobne do białka, które zmieniło funkcję (np. paraloga lub koortologa). Dlatego należy, o ile to możliwe, starać się zidentyfikować białka ortologiczne i na nich opierać swoje analizy. Częstym błędem jest bezkrytyczne przeniesienie adnotacji z jednego białka na drugie, co prowadzi do bardzo zaskakujących wniosków, biorąc pod uwagę charakter organizmu posiadającego błędnie opisane białko. Na przykład białko DRA0097 z bakterii *Deinococcus radiodurans* zostało opisane jako „przypuszczalne białko związane z morfogenezą głowy". Błąd ten wziął się stąd, że to białko wykazuje podobieństwo do białka SPP1 bakteriofaga infekującego bakterię *Bacillus subtilis,* zaangażowanego rzeczywiście w tworzenie główki faga. Również wiele białek kodowanych w genomach *Archaea* wykazujących podobieństwo do białek *Eukaryota* posiada adnotacje uwzględniające nazwy struktur lub białek, których archebakterie nie posiadają (np. jąderek, centromerów, mikrotubul, białek MHC). Niektóre z błędów wynikają z nieuwzględniania wielodomenowej budowy białek, co prowadzi do ograniczania lub rozszerzania przypisywanych funkcji. Inne błędy dotyczą bezpośredniego przeniesienia zbyt szczegółowego opisu z jednego białka na drugie, co prowadzi do zbyt ścisłego (*overprediction*) i błędnego oznaczenia funkcji białka. Na przykład oba białka mogą wykazywać duże podobieństwo i być permeazami aminokwasów, ale mogą charakteryzować się różną specyficznością substratową. Na przykład znaleziony homolog może być permeazą histydyny, a białko badane permeazą tryptofanu. W tym przypadku białko analizowane najlepiej oznaczyć jako „permeaza aminokwasowa", a nie „permeaza histydyny", do czasu aż jego funkcja zostanie sprecyzowana. Szczególnie dużym problemem ogarniającym bazy danych jest dalsza propagacja opisów już błędnie adnotowanych białek, które następnie służą do adnotacji nowych genomów.

Procesu adnotacji nie da się do końca zautomatyzować. Aby uniknąć błędów w adnotacjach niezbędna jest interwencja doświadczonego eksperta. Poza wymie-

nionymi już wskazówkami, wielu błędów można uniknąć przeprowadzając analizy oparte na motywach i domenach. Duże znaczenie ma tu również znajomość biologii organizmu, którego genom badamy.

W celu uwiarygodnienia adnotacji genomów mikroorganizmów zapoczątkowano projekt HAMAP (*High-quality Automated and Manual Annotation of Microbial Proteomes*, 118). Ma on na celu zintegrowanie automatycznych i manualnych metod przypisujących funkcje, w celu przyspieszenia procesu nadzorowanych adnotacji, zapewniając jednocześnie ich wysoką jakość na wzór bazy SWISS-PROT. Wyniki tych adnotacji są integrowane z bazą SWISS-PROT, zawierającą sekwencje białkowe dokładnie adnotowane przez ekspertów. Obecnie (stan z początku 2005 r.) kompletnie opisanych w ten sposób zostało 7 ze 180 genomów prokariotycznych. W sumie opracowano 14% wszystkich adnotowanych białek, których jest 526 727.

4. Zakończenie

Genom można porównać do księgi zawierającej tajemnice jego funkcjonowania, zapisanej jednak w zupełnie obcym języku, którego nie jesteśmy w stanie zrozumieć. Identyfikowanie sekwencji kodujących jest ważnym etapem odczytywania tej księgi i można je porównać do poszukiwania pojedynczych słów w ciągu pozostałych znaków. Jednak o zrozumieniu tej księgi możemy mówić dopiero, wtedy gdy nauczymy się rozumieć znaczenie tych słów, czyli w naszym przypadku dokonamy adnotacji funkcjonalnych genów. Jeśli jednak chcemy tę książkę w pełni zrozumieć, nie wystarczy rozumienie znaczenia poszczególnych słów – potrzebna jest znajomość gramatyki i składni, czyli musimy poznać jak produkty genów współdziałają ze sobą w żywych komórkach w sieci wzajemnych zależności. Dzięki intensywnie rozwijającym się metodom rozpoznającym geny i dokonującym adnotacji będziemy mogli w przyszłości wejrzeć w naturę procesów biologicznych zachodzących w żywych organizmach.

Literatura

1. Sanger F., Nicklen S., Coulson A. R., (1977), Proc. Natl. Acad. Sci. USA, 74, 5463-5467.
2. Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Wheeler L. D., (2005), Nucleic Acids Res., 33, 34-38.
3. Rogozin I. B., Makarova K. S., Natale D. A., Spiridonov A. N., Tatusov R. L., Wolf Y. I., Yin J., Koonin E. V., (2002), Nucleic Acid Res., 30, 4262-4271.
4. Stein L., (2001), Nature Rev. Genet., 2, 493-503.
5. Imanishi T., Itoh T., Suzuki Y., O'Donovan C., Fukuchi S., Koyanagi K. O., Barrero R. A., Tamura T., Yamaguchi-Kabata Y., Tanino M., et al., (2004), PLoS BIOLOGY, 2, 0001.
6. Goffeau A., Barrel B. G., Bussey H., Davies R. W., Dujon B., Feldmann H., Galibert F., Hoheisel J. D., Jacq C., Johnston M., et al., (1996), Science, 274, 546-567.
7. Termier M., Kalogeropoulos A., (1996), Yeast, 12, 369-384.
8. Das S., Yu L, Galtatzes C., Rogers R., Freeman J., Bienkowska J., Adams R. M., Smith T. F., (1997), Nature, 385, 29-30.

9. Argentini C., La Sorsa V., Bruni R., D'Ugo E., Giuseppetti R., Rapicetta M., (1999), J. Gen. Virol., 80, 617-626.
10. Stallmeyer B., Drugeon G., Reiss J., Haenni A. L., Mendel R. R., (1999), Am. J. Hum. Genet., 64, 698-705.
11. Rogic S., Mackworth A. K., Ouellette F. B., (2001), Genome Res., 11, 817-832.
12. Basrai M. A, Hieter P., Boeke J. D., (1997), Genome Res., 7, 768-771.
13. Andrade M. A, Daruvar A., Casari G., Schneider R., Termier M., Sander C., (1997), Yeast, 13, 1363-1374.
14. Winzeler E. A, Davis R. W., (1997), Curr. Opin. Genet. Dev., 7, 771-776.
15. Gierlik A., Mackiewicz P., Kowalczuk M., Dudek M. R., Cebrat S., (1999), Int. J. Modern Phys. C., 10, 635-643.
16. Oliver S. G., van der Aart Q. J. M., Agostoni-Carbone M. L., Aigle M., Alberghina L., Alexandraki D., Antoine G., Anwar R., Ballesta J. P. G., Benit P., et al., (1992), Nature, 357, 38-46.
17. Dujon B., Alexandraki D., Andre B., Ansorge W., Baladron V., Ballesta J. P. G., Banrevi A., Bolle P. A., Bolotin-Fukuhara M., Bossier P., et al., (1994), Nature, 369, 371-378.
18. Shepherd J. C., (1981), J. Mol. Evol., 17, 94-102.
19. Fickett J. W., (1982), Nucleic Acids Res., 10, 5303-5318.
20. Silverman B., Linsker R., (1986), J. Theor. Biol., 118, 295-300.
21. Trifonov E. N., (1998), Physica A, 249, 511-516.
22. Lagunez-Otero J., Trifonov E. N., (1992), J. Biomol. Struct. Dyn., 10, 455-464.
23. Smithies O., Engels W. R., Devereux J. R., Slightom J. L., Shen S. H., (1981), Cell, 26, 345-353.
24. Poncz M., Schwartz E., Ballantine M., Surrey S., (1983), J. Biol. Chem., 258, 11599-11609.
25. Dujon B., Alexandraki D., Andre B., Ansorge W., Baladron V., Ballesta J. P. G., Banrevi A., Bolle P.A., Bolotin-Fukuhara M., Bossier P., et al., (1994), Nature, 369, 371-378.
26. Karlin S., Burge C., (1995), Trends Genet., 11, 283-290.
27. Cebrat S., Dudek M. R., Rogowska A., (1997), J. Appl. Genet., 3, 1-9.
28. Wong J. T., Cedergren R., (1986), Eur. J. Biochem., 159, 175-180.
29. Zhang C. T., Zhang R., (1991), Nucleic Acids Res., 19, 6313-6317.
30. Karlin S., Mrázek J., (1996), J. Mol. Biol., 262, 459-472.
31. Cebrat S., Dudek M. R., Mackiewicz P., Kowalczuk M., Fita M., (1997), Microb. & Comp. Genom., 2, 259-268.
32. Cebrat S., Dudek M. R., Mackiewicz P., (1998), Theory in BioSciences, 117, 78-89.
33. Frank G. K., Makeev V. J., (1997), J. Biomol. Struct. Dyn., 14, 629-639.
34. Wang J., (1998), J. Biomol. Struct. Dyn., 16, 51-57.
35. Cebrat S., Dudek M. R., Rogowska, A., (1997), J. Appl. Genet., 3, 1-9.
36. Mackiewicz P., Kowalczuk M., Mackiewicz D., Nowicka A., Dudkiewicz M., Łaszkiewicz A., Dudek M. R., Cebrat S., (2002), Yeast, 19, 619-629.
37. Karlin S., Blaisdell B. E., Bucher P., (1992), Protein Eng., 5, 729-738.
38. Sharp P. M., Li W. H., (1987), Nucleic Acids Res., 15, 1281-1295.
39. Bennetzen J. H., Hall B. D., (1982), J. Biol. Chem., 257, 3026-3031.
40. Ikemura T., (1981), J. Mol. Biol., 151, 389-409.
41. Sharp P. M., Cowe E., (1991), Yeast, 7, 657-678.
42. Fickett J. W., Tung C. S., (1992), Nucleic Acids Res., 20, 6441-6450.
43. Fickett J. W., (1996), Trends Genet., 12, 316-320.
44. Fickett J. W., (1996), Comput. & Chem., 20, 103-118.
45. Bystroff C., Thorsson V., Baker D., (2000), J. Mol. Biol., 301, 173-190.
46. Claverie J.-M., Sauvaget I., Bougueleret L., (1990), Meth. Enzymol., 183, 237-252.
47. Krogh A., Mian S., Haussler D., (1994), Nucleic Acids Res., 22, 22.
48. Salzberg S. L., Delcher A. L., Kasif S., White O., (1998), Nucleic Acids Res., 26, 2.
49. Lukashin A. V., Borodovsky M., (1998), Nucleic Acids Res., 26, 1107-1115.
50. Hayes W. S., Borodovsky M., (1998), Genome Res., 8, 1154-1171.
51. Bult C. J., White O., Olsen G. J., Zhou L., Fleischmann R. D., Sutton G. G., Blake J. A., FitzGerald L. M., Clayton R. M., Gocayne J. D., et al., (1996), Science, 273, 1058-1073.

52. Hawley D. K., McClure W. R., (1983), Nucleic Acids Res., 11, 2237-2255.
53. Burge Ch. B., Karlin S., (1998), Curr. Opin. Struct. Biol., 8, 346-354.
54. Schug J., Overton G. C., (1997), Proc. Int. Conf. Intell. Syst. Mol. Biol., 5, 268-271.
55. Ferreira da Silva O. M., Mendes D. Q., Ferrari L. I., Vasconcelos R. A. T., (2004), Genet. Mol. Biol., 27, 644-650.
56. Shine J., Dalgarno L., (1974), Proc. Natl. Acad. Sci. USA, 71, 1342-1346.
57. Kozak M., (1996), Mamm. Genome, 7, 563-574.
58. Pesole G., Gissi C., Grillo G., Licciulli F., Liuni S., Saccone C., (2000), Gene, 261, 85-91.
59. Claverie J.-M., (1997), Hum. Mol. Genet., 6, 1735-1744.
60. Prestridge D. S., (1995), J. Mol. Biol., 249, 923-932.
61. Demeler B., Zhou G., (1991), Nucleic Acids Res., 19, 1593-1599.
62. Pedersen A. G., Engelbrecht J., (1995), Intelligent Systems Mol. Biol., 3, 292-299.
63. Ogura H., Agata H., Xie M., Odaka T., Furutani H., (1997), Comput. Biol. Med., 27, 67-75.
64. Brunak S., Engelbrecht J., Knudsen S., (1991), J. Mol. Biol., 220, 49-65.
65. Trifonov E. N., (1996), Comp. Appl. Biosci., 12, 423-429.
66. Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R., (1997), Comp. Appl. Biosci., 13, 263-270.
67. Altschul S., F., Gish W., Miller W., Myers E.,W., Lipman D. J., (1990), J. Mol. Biol., 215, 403-410.
68. Pearson W. R., (1990), Methods Enzymol., 183, 63-98.
69. International Human Genome Sequencing Consortium (IHGSC), (2001), Nature, 409, 860-921.
70. Pennacchio L. A., Rubin E. M., (2001), Nat. Rev. Genet., 2, 100-109.
71. Nekrutenko A., Makova D. K., Wen-Hsiung L., (2002), Genome Res., 12, 198-202.
72. Nekrutenko A., Chung W.Y., Li W. H., (2003), Trends Genet., 19, 306-310.
73. Ochman H., (2002), Trends Genet., 18, 335-337.
74. Pavy N., Rombautus S., Dehais P., Mathe C., Ramana D., Leroy P., Rouze P., (1999), Bioinformatics, 15, 887-899.
75. Delcher A. L., Harmon D., Kasif S., White O., Salzberg S. L., (1999), Nucleic Acids Res., 27, 4636-4641.
76. Reese M., Hartzell G., Harris L. N., Ohler U., Abril J. F., Lewis S. E., (2000), Genome Res., 10, 483-501.
77. Guigo R., Agarwal P., Abril J. F., Burset M., Fickett J. W., (2001), Genome Res., 10, 1631-1642.
78. Bocs S., Danchin A., Médigue C., (2002), BMC Bioinf., 3, 5.
79. Ouzounis C. A., Karp P. D., (2002), Genome Biol., 3, comment2001.1-2001.6.
80. Mackiewicz P., Kowalczuk M., Gierlik A., Dudek M. R., Cebrat S., (1999), Nucleic Acids Res., 27, 3503-3509.
81. Skovgaard M., Jensen L. J., Brunak S., Ussery D., Krogh A., (2001), Trends Genet., 17, 425-428.
82. Siew N., Fischer D., (2003), Proteins: Struct. Funct. Genet., 53, 241-251.
83. Charlebois R. L., Clarke G. D., Beiko R. G., Jean A., (2003), FEMS Microb. Lett., 225, 213-220.
84. Kawarabayasi Y., (1999), DNA Res., 6, 83-101.
85. Kellis M., Patterson N., Endrizzi M., Birren B., Lander E. S., (2003), Nature, 423, 241-254.
86. Zhang C. T., Wang J., Zhang R., (2002), Comput. & Chem., 26, 195-206.
87. Wood V., Rutherford K .M., Ivens A., Rajandream M. A., Barrell B., (2001), Comp. Funct. Genom., 2, 143-154.
88. 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.
89. Kumar A., Harrison P. M., Cheung K.-H., Lan N., Echols N., Bertone P., Miller P., Gerstein M. B., Snyder M., (2002), Nature Biotech., 20, 58-63.
90. Oshiro G., Wodicka L. M., Washburn M. P., Yates III J. R., Lockhart D. J., Winzeler E. A., (2002), Genome Res., 12, 1210-1220.
91. Harrison P. M., Carriero N., Liu Y., Gerstein M., (2003), J. Mol. Biol., 333, 885-892.
92. Thomas G. H., (1999), Bioinformatics, 15, 860-861.
93. Tatusov R. L, Koonin E. V., Lipman D. J., (1997), Science, 278, 631-637.
94. Tatusov R., Galperin M., Natale D., Koonin E., (2000), Nucleic Acids Res., 28, 33-36.

95. Ouzounis C. A., Coulson R. M. R., Enright A. J., Kunin V., Pereira-Leal J. B., (2003), Nature Rev. Genet., 4, 508-519.
96. Emanuelsson O., (2002), Briefings in Bioinf., 3, 361-376.
97. Huynen M. A., Snel B., (2000), Adv. Protein. Chem., 54, 345-379.
98. Huynen M., Snel B., Lathe W., Bork P., (2000), Curr. Opin. Struct. Biol., 10, 366-370.
99. Gabaldon T., Huynen M. A., (2004), Cell. Mol. Life Sci., 61, 930-944.
100. Gaasterland T., Ragan M. A., (1998), Microb. Comp. Genomics, 3, 199-217.
101. Huynen M. A., Bork P., (1998), Proc. Natl. Acad. Sci. USA, 95, 5849-5856.
102. Pellegrini M., Marcotte E. M., Thompson M. J., Eisenberg D., Yeates T. O., (1999), Proc. Natl. Acad. Sci. USA, 96, 4285-4288.
103. Enright A. J, Illopoulos I., Kyrpides N. C., Ouzounis C. A., (1999), Nature, 402, 86-90.
104. Marcotte E. M., Pellegrini M., Ng H. L., Rice D. W., Yeates T. O., Eisenberg D., (1999), Science, 285, 751-753.
105. Snel B., Bork P., Huynen M. A., (2000), Trends Genet., 16, 9-11.
106. Galperin M. Y., Nikolskaya A. N., Koonin E. V., (2001), FEMS Microb. Lett., 203, 11-21.
107. Mushegian A. R., Koonin E. V., (1996), Trends Genet., 12, 289-290.
108. Watanabe H., Mori H., Itoh T., Gojobori T., (1997), J. Mol. Evol., 44, 57-64.
109. Hughes D., (2000), Genome Biol., 1, reviews 0006.
110. Overbeek R., Fonstein M., D'Souza M., Pusch G. D., Maltsev N., (1999), Proc. Natl. Acad. Sci. USA, 96, 2896-2901.
111. Wolf Y. I., Rogozin I. B., Kondrashov A. S., Koonin E. V., (2001), Genome Res., 11, 356-372.
112. Bork P., Bairoch A., (1996), Trends Genet., 12, 425-427.
113. Bork P., Koonin E. V., (1998), Nature Genet., 13, 313-318.
114. Doerks T., Bairoch A., Bork P., (1998), Trends Genet., 14, 248-250.
115. Brenner S. E., (1999), Trends Genet., 15, 132-133.
116. Galperin M. Y., Koonin E. V., (1998), In Silico Biol., 1, 7.
117. Devos D., Valencia A., (2001), Trends Genet., 17, 429-431.
118. Gattiker A., Michoud K., Rivoire C., Auchincloss A. H., Coudert E., Lima T., Kersey P., Pagni M., Sigrist C. J. A., Lachaize C., Veuthey A. L., Gasteiger E., Bairoch A., (2003), Comput. Biol. & Chem., 27, 49-58.