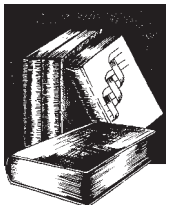


PRACE PRZEGLĄDOWE

Biologiczne bazy danych i ich zastosowanie w funkcjonalnej analizie porównawczej organizmów – wybrane zagadnienia

Ireneusz Ślesak[1,2], Stanisław Karpiński[1]

[1]Instytut Fizjologii Roślin, Polska Akademia Nauk, Kraków
[2]Katedra Genetyki, Hodowli i Biotechnologii Roślin,
Szkoła Główna Gospodarstwa Wiejskiego, Warszawa

Biological databases and their application for functional comparative analysis of organisms – selected issues

Summary

A quickly increasing amount of biological data, such as DNA/RNA and protein sequences, determines fundamental role of biological databases for functional comparative analysis of genomes/metabolic pathways of various organisms. This is neatly illustrated by 'Databases', the new type of scientific journal that has recently appeared. The journal is dedicated to bioinformatic data processing and storage. Biological databases and platforms play a crucial role in the development and progress of the systems biology. Systems biology integrates key elements of an organism functions e.g. gene and protein content, regulation of gene expression, enzymes activities, and specific profiles of different metabolites. Therefore, it is possible to generate a holistic and integrated model of basic biological functions and processes that determine an organism phenotype.

Key words:
biological databases, BLAST, EMBL, metabolic pathways, molecular phylogeny, NCBI, regulatory elements, UniProt/Swiss-Prot.

Adres do korespondencji

Stanisław Karpiński,
Katedra Genetyki,
Hodowli i Biotechnologii Roślin,
Szkoła Główna Gospodarstwa Wiejskiego,
ul. Nowoursynowska 159,
02-776 Warszawa;
e-mail:
stanislaw_karpinski@sggw.pl,
stan_39@o2.pl

biotechnologia
4 (91) 39–52 2010

1. Wprowadzenie – rys historyczny

Od pierwszej połowy lat 80. ubiegłego wieku wzrasta zainteresowanie sposobami gromadzenia i zarządzania szeroko rozumianymi danymi biologicznymi, ze szczególnym uwzględnieniem

genów i białek. Najstarszą bazą danych jest bank struktur białkowych (PDB, ang. *Protein Data Bank*), który powstał już w 1971 r. i do dziś gromadzi ustalone metodami eksperymentalnymi, takimi jak krystalografia rentgenowska i magnetyczny rezonans jądrowy (NMR, ang. *Nuclear Magnetic Resonance*) struktury przestrzenne białek (1,2). W 1981 r. pierwsze sekwencje nukleotydowe zostają zdeponowane w bibliotece EMBL (ang. *European Molecular Biology Laboratory* – Europejskie Laboratorium Biologii Molekularnej), czyli na ok. 8 lat przed formalnym powstaniem sieci Internet i sieciowego systemu WWW (ang. *World Wide Web*). Główne biologiczne bazy danych, takie jak: NCBI – GenBank (ang. *National Center for Biotechnology Information* – Narodowe Centrum Informacji Biotechnologicznej – bank genów), Swiss-Prot (baza z sekwencjami białkowymi z siedzibą w Szwajcarii), DDBJ (ang. *DNA Data Bank of Japan* – bank sekwencji DNA w Japonii) również powstają w latach 80. ubiegłego wieku. Jednakże do gwałtownego rozwoju tych baz, które sukcesywnie zaczynają gromadzić coraz więcej informacji nt. znanych sekwencji kwasów nukleinowych (DNA, RNA), sekwencji aminokwasowych białek oraz publikacji naukowych, dochodzi w latach 90. XX w. Co prawda, już w latach 1986-1988 trzy największe bazy: EMBL, NCBI i DDBJ rozpoczynają współpracę, której celem jest wymiana i ujednolicenie sposobu gromadzenia informacji sekwencyjnych, ale *de facto* dopiero rozwój globalnej sieci komputerowej Internet i protokołu WWW (początek lat 90. XX w.) pozwala na szybki obieg informacji i aktualizację danych sekwencyjnych pomiędzy tymi bazami w ramach konsorcjum INSDC (ang. *International Nucleotide Sequence Database Collaboration*), czyli międzynarodowej współpracy między bazami danych z sekwencjami nukleotydowymi (3). Zatem, warto podkreślić, że użyteczność biologicznych baz danych i rozwój bioinformatyki byłby niewielki, gdyby nie dynamiczny rozwój sieci komputerowej Internet. Możliwość szybkiego łączenia się i transferu informacji pomiędzy dowolnymi bazami sprawia, że zarówno deponowanie jak i uzyskiwanie istotnych dla użytkownika informacji jest szybkie, tanie i proste (1).

Biopolimery takie jak kwasy nukleinowe i białka, które są ciągiem ułożonych w odpowiedniej kolejności (sekwencji), odpowiednio połączonych ze sobą nukleotydów i aminokwasów, stanowią idealny materiał do analizy w kategoriach informatycznych. Z punktu widzenia informatyki np. sekwencje aminokwasowe białek są pewnym ciągiem znaków, które niosą ze sobą określoną informację – w tym przypadku – informację biologiczną o określonej strukturze przestrzennej i funkcji danego białka. Co więcej, informacja zawarta w sekwencji aminokwasowej jest ściśle związana, z poprzedzającą ją w procesie przekazu informacji genetycznej, sekwencją nukleotydów w obrębie określonej cząsteczki DNA/RNA. Taki, a nie inny charakter struktury I-rzędowej białek i kwasów nukleinowych sprawia, że biologia – zwłaszcza część biologii molekularnej – z nauki typowo eksperymentalnej zaczyna powoli ewoluować w kierunku nauki z pogranicza teorii informacji i informatyki.

Niebywały postęp w technikach sekwencjonowania DNA umożliwił stały dopływ danych nt. sekwencji. Najpierw były to genomy mikroorganizmów prokariotycznych i organelli komórkowych, a następnie genomy eukariontów (4), ze sztandarowym

projektem sekwencjonowania genomu człowieka (*Homo sapiens*), który ukończono w 2001 r. (5,6).

Celem tego opracowaniu nie jest systematyczne i szczegółowe omówienie wszystkich baz danych. Przede wszystkim nie pozwalają na to ramy artykułu. Ponadto, w literaturze anglojęzycznej dostępnych jest wiele podręczników i opracowań, w których omawiano biologiczne (bioinformatyczne) bazy danych (1,7-10). Warto także nadmienić, że w ciągu kilku ostatnich lat ukazały się na rynku polskim dwa tłumaczenia podręczników do bioinformatyki: 1) Baxevaniasa i Ouellette'a (11) i 2) Higgsa i Attwood (12), w których bardziej szczegółowo omówiono biologiczne bazy danych.

W artykule tym zaprezentowano ogólny opis biologicznych baz danych. Dwie wybrane bazy/platformy, które mogą być przydatne dla szerokiego grona biologów i biotechnologów – niekoniecznie zawodowo zajmujących się bioinformatyką – przedstawiono bardziej szczegółowo. Zwrócono również uwagę na ograniczenia i błędy jakie mogą być związane z pracą z bazami danych.

2. Definicja bazy danych

Sposób zbierania, katalogowania oraz łatwego wyszukiwania informacji, z już zgromadzonych zasobów, był na początku istotą konstrukcji dobrej bazy. Tworzenie baz danych było jedną z podstaw tworzącej się nowej dziedziny wiedzy, czyli bioinformatyki (3). Spośród wielu definicji bioinformatyki do dnia dzisiejszego wiele z nich uwzględnia podstawową rolę baz danych w rozwoju tej dziedziny wiedzy. W myśl tej definicji bioinformatyka jest ni mniej, ni więcej tylko nauką o przechowywaniu, wyszukiwaniu i analizie informacji na temat ważnych biologicznie cząsteczek, głównie genów i białek, a ostatnio także całej gamy substancji drobnocząsteczkowych. Dokonywanie odkryć w bioinformatyce odbywa się zatem na bazie – dosłownie i w przenośni – już zgromadzonych informacji, które jako takie nie mają w ogóle lub w sposób wystarczający przypisanej interpretacji biologicznej. Na podstawie „genowo-białkowych" baz zaczęły powstawać zbiory danych pozwalające na uzyskiwanie dodatkowych informacji – przede wszystkim o charakterze ewolucyjnych zależności pomiędzy sekwencjami dla wybranych grup organizmów. Praktycznie dowolna możliwość operowania zgromadzonymi w bazach „surowymi" danymi sekwencyjnymi okazała się doskonałym narzędziem analitycznym w filogenetyce molekularnej i genomice, zwłaszcza genomice funkcjonalnej. Zatem czym jest biologiczna (bioinformatyczna) baza danych? Można ją zdefiniować jako odpowiednio zorganizowane zbiory z bardzo dużą ilością danych, głównie sekwencji DNA/RNA i białek. Zazwyczaj są one ściśle powiązane z darmowym oprogramowaniem umożliwiającym przeszukiwanie zasobów i ekstrakcję pożądanej przez użytkownika informacji. Dobra baza danych powinna się charakteryzować następującymi cechami: a) możliwością precyzyjnej i szybkiej – za pomocą minimalnej liczby kroków („kliknięć") – ekstrakcji tylko tych informacji, które są poszukiwane, b) możliwością obróbki informacji, przy zastosowaniu skojarzone-

go z bazą oprogramowania, c) możliwością łatwego przejścia poprzez odpowiednie łącza internetowe do innych baz danych, d) możliwością zdefiniowania przez użytkownika sposobu podglądu uzyskanych danych i ich zapisania i/lub wydrukowania, e) małą redundancją danych, tzn. zgromadzone informacje powinny być unikatowe, a obecność wszelkich nadmiarowych danych powinna być ograniczona do minimum. Cieszące się znaczną popularnością bazy zawierają ogromną ilość danych, które na bieżąco są aktualizowane, a wyszukiwanie informacji rozpoczyna się zazwyczaj w polu, w którym należy wpisać odpowiednie słowo kluczowe (ang. *key word*). Coraz więcej dostępnych baz nie służy tylko do gromadzenia i katalogowania informacji, jak np. znaczna część klasycznych baz: NCBI, EMBL, czy Swiss-Prot. Większość baz jest także udostępnionymi *on-line* zestawami oprogramowania (platformami), które korzysta z informacji zgromadzonych w innych bazach. Stąd też bazy stają się serwisami oprogramowania, które służy rozwiązywaniu konkretnego zagadnienia biologicznego. Przyrost informacji w bazach danych, jest bardzo szybki (1). Przykładowo, analiza danych na przestrzeni 10 lat: 1998-2008 pokazała, że liczba zdeponowanych sekwencji w GenBank (NCBI) i DDBJ podwajała się średnio co dwa lata (13). Coraz wydajniejsze – w związku z postępującą automatyzacją – metody sekwencjonowania genomów – głównie genomów mikroorganizmów – raczej nie zmienią tempa przyrostu danych sekwencyjnych. Wspomniana baza NCBI w tej chwili oferuje 1000 całkowicie zsekwencjonowanych genomów prokariotycznych (www.ncbi.nlm.nih.gov). Biologiczne bazy danych były pierwotnie tworzone głównie przez biologów molekularnych i biochemików. Obecnie obserwuje się w konstrukcji baz tendencję, aby niezależnie od ich merytorycznej zawartości, były one czytelne i zrozumiałe dla niespecjalistów z różnych dziedzin nauk biologicznych, rolniczych i medycznych. W tabeli zamieszczono kilka podstawowych biologicznych baz danych ze wskazaniem jaki główny rodzaj informacji przechowują.

Generalnie liczba baz danych zawierająca informacje z samej biologii molekularnej wynosi obecnie ponad 1000 (14) i wszystko wskazuje na to, że z roku na rok nadal będzie się powiększać.

Tabela

Przykładowe biologiczne bazy danych

Typ bazy danych	Rodzaj gromadzonej informacji	Nazwa bazy danych	Adres www
1	2	3	4
bibliograficzne	literatura naukowa	biblioteka wirtualna (ICM)	http://vls.icm.edu.pl/ (płatna)
		NCBI – PubMed	http://www.ncbi.nlm.nih.gov/pubmed/
		linki do czasopism publikujących prace z szeroko rozumianej biologii roślin	http://www.e-journals.org/botany/

1	2	3	4
taksonomiczne	klasyfikacja organizmów	NCBI – Taxonomy	http://www.ncbi.nlm.nih.gov/guide/taxonomy/
zawierające sekwencje kwasów nukleinowych	informacje o sekwencjach DNA/RNA	NCBI – GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
		EMBL	http://www.ebi.ac.uk/embl/
		DDBJ	http://www.ddbj.nig.ac.jp/
genomiczne	informacje o genach i całych genomach	NCBI – genomy	http://www.ncbi.nlm.nih.gov/sites/genome
		EMBL	http://www.ebi.ac.uk/genomes/
zawierające sekwencje białek	informacje o sekwencjach białkowych	PIR	http://pir.georgetown.edu/
		Swiss-Prot	http://www.expasy.ch/sprot/
		UniProt	http://www.uniprot.org/
zawierające struktury przestrzenne białek	eksperymentalnie zweryfikowane 3-wymiarowe struktury białek	PDB	http://www.rcsb.org/pdb/home/home.do
zawierające rodziny białek, domeny białkowe itp.	klasyfikacja białek i identyfikacja domen białkowych	PROSITE	http://www.expasy.ch/prosite/
		PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php
		Pfam	http://pfam.sanger.ac.uk/
enzymy/szlaki metaboliczne	informacja o szlakach metabolicznych: enzymach oraz ich substratach i produktach	BioCyc	http://biocyc.org/
		KEGG – pathway	http://www.genome.jp/kegg/pathway.html
		BRENDA	http://www.brenda-enzymes.org/
inne specjalistyczne bazy danych np. dotyczące jednej grupy białek, metabolitów drobnocząsteczkowych lub dedykowane jednemu organizmowi	przewidywanie rodzaju dysmutazy ponadtlenkowej (SOD) oraz stopnia jej oligomeryzacji	SODa	http://babylone.ulb.ac.be/SODa/index.php
	zawartość polifenoli w pożywieniu	Phenol-Explorer	http://www.phenol-explorer.eu
	genom *Arabidopsis thaliana*	TAIR	http://www.arabidopsis.org/

2.1. Ograniczenia w korzystaniu z baz

Dane sekwencyjne w bazach nie powinny być traktowane jako absolutnie niezmienne i ostateczne. Warto sobie uświadomić, że każda sekwencja w bazach jest wynikiem eksperymentu, czyli mniej lub bardziej dokładnego procesu sekwencjonowania. Eksperyment jest podstawowym źródłem danych do sekwencyjnych baz danych. Jednakże, czasami zdarza się, że opublikowana (zdeponowana) w bazie sekwencja zawiera błędy lub jest błędnie przypisana do genu/białka. Sytuacja taka może mieć miejsce szczególnie, wtedy gdy sekwencjonowanie przeprowadza się wysoko przepustowymi technikami, np. techniką sekwencjonowania genomów *shot-*

gun („strzału na ślepo"). Problemem są także zanieczyszczenia materiału genetycznego powstałe w trakcie przygotowania DNA do sekwencjonowania. Przykładowo, ostatnio opublikowana sekwencja genomu ogórka (*Cucumis sativus*) zawiera około 1500 sekwencji pochodzenia bakteryjnego i wirusowego (15), a zdeponowana i obecnie przygotowywana do publikacji sekwencja genomu tej rośliny, wykonana przez polskie konsorcjum, nie zawiera takich błędów (S. Karpiński, informacja ustna). Problem błędnych adnotacji dla danych sekwencyjnych zaczyna narastać wraz z lawinowo rosnącą liczbę zsekwencjonowanych genomów, dla których w sposób automatyczny – korzystając z algorytmów obliczeniowych – przypisywane są sekwencje białek o określonej funkcji. Nawet, jeśli genom jest zsekwencjonowany poprawnie, to błędnie mogą zostać wykonane adnotacje, czyli przypisanie danemu białku konkretnej funkcji biologicznej. Na podstawie przeprowadzonych analiz porównawczych dla 37 modelowych rodzin różnych białek enzymatycznych, które zostały dokładnie scharakteryzowane metodami eksperymentalnymi, pokazano, że automatyczna adnotacja funkcjonalna dla sekwencji aminokwasowych tych enzymów, z wykorzystaniem wybranych baz danych, może być błędna nawet na poziomie kilkudziesięciu procent dla sekwencji białkowych w obrębie danej rodziny (16). Ograniczeniem ryzyka związanego z pracą na białkach o źle przypisanej funkcji jest korzystanie z baz, które są aktualizowane ręcznie, na podstawie danych sekwencyjnych zweryfikowanych eksperymentalnie, np. Swiss-Prot (tab. 1), SFLD (ang. *Structure-Function Linkage Database*, http://sfld.rbvi.ucsf.edu/) (16). Korzystanie z nieprawidłowo opisanych sekwencji będzie prowadzić, np. do błędnych wyników analiz o charakterze strukturalno-funkcjonalnym. Przykładowo, przewidywanie *in silico* różnych właściwości białek, takich jak np. lokalizacja białka w określonej organelli komórkowej jest tylko pewną propozycją (przewidywaniem) i zależy od obecności specyficznych sekwencji sygnałowych, które decydują o eksporcie określonego białka do wybranej organelli komórkowej. Jednak nie można takiego wyniku – nawet jeśli sekwencja jest prawidłowa – traktować jako ostatecznego dowodu, że dane białko rzeczywiście jest zlokalizowane, np. w chloroplastach czy w mitochondriach. Jak na razie, przewidywania na podstawie algorytmów bioinformatycznych nie zastąpią eksperymentalnej weryfikacji hipotezy badawczej – w tym przypadku hipotezy o tym, że analizowane białko jest zlokalizowane w chloroplastach lub w mitochondriach. Nie można jednak wykluczyć, że już wkrótce udoskonalone algorytmy stosowane w różnych aplikacjach będą z coraz większą dokładnością odwzorowywać rzeczywistość biologiczną (3).

3. Platforma: phylogeny.fr – narzędzie do filogenetyki molekularnej

Rekonstrukcja historii ewolucyjnej sekwencji DNA/RNA i białek jest obecnie podstawą większości badań w zakresie genomiki funkcjonalnej, przewidywania funkcji nieznanego genu/białka, wykrywania tzw. poziomego transferu genów (ang. *horizon-*

tal gene transfer), czy identyfikacji nowych mikroorganizmów. Obecnie jest kilkadziesiąt programów dostępnych *on-line* lub do pobrania z Internetu, które umożliwiają analizę filogenetyczną. Ich zbiór można znaleźć na stronie Joe Felsensteina: http://evolution.genetics.washington.edu/phylip/software.html).

Poniżej omówiono w zarysie bardzo przydatną platformę do analiz z zakresu filogenetyki molekularnej jaką jest phylogeny.fr (http://www.phylogeny.fr). Właściwie phylogeny.fr jest dostępnym *on-line* oprogramowaniem służącym do analizowania wprowadzonych danych z sekwencjami nukleotydowymi/białkowymi pod kątem filogenetycznym. Podstawowym elementem przy przewidywaniu historii ewolucyjnej określonych sekwencji jest dysponowanie zbiorem sekwencji homologicznych, czyli takich, które pochodzą od wspólnego przodka. Homologia sekwencji pociąga za sobą takie konsekwencje jak podobna struktura i funkcja badanego genu/białka. Jednak jak uzyskać zbiór sekwencji, które z możliwie dużym prawdopodobieństwem są sekwencjami homologicznymi? Najpowszechniej stosowanym sposobem na otrzymanie zbioru takich sekwencji jest zastosowanie programu BLAST (ang. *Basic Local Alignment Serach Tool*) lub jego nowszej wersji PSI-BLAST (ang. *Position-Specific Iterated BLAST*), który podaje listę najlepszych lokalnych dopasowań sekwencji zapytania (ang. *query sequence*) z sekwencjami zgromadzonymi w bazie danych (17,18). Algorytm BLAST jest zaimplementowany w wielu bazach danych, np. UniProt/Swiss-Prot, EMBL, choć pierwotną lokalizacją BLAST'a jest baza NCBI. Warto tutaj podkreślić, że BLAST daje wyniki dotyczące podobieństwa pomiędzy liniową sekwencją liter (nukleotydy lub aminokwasy), a nie zawsze sekwencje najbardziej podobne są rzeczywiście homologiczne i odwrotnie – sekwencje o niskim stopniu podobieństwa mogą faktycznie być homologiczne, bo świadczy o tym np. podobieństwo struktury 3-rzędowej. Zazwyczaj, w praktyce, jeśli przyrównywane sekwencje otrzymują odpowiedni parametr statystyczny tzw. wartość oczekiwaną, czyli E-value (ang. *expected value*), znacznie mniejszą od 1, zwykle $E < 0.0001$ (1e-04) oraz identyczność w przypadku sekwencji kwasów nukleinowych i białek – dla odcinków o długości 100 i więcej liter – jest $> 70\%$, dla sekwencji nukleotydowych i $> 25\%$ w przypadku sekwencji aminokwasowych, to zazwyczaj takie sekwencje można uznać za homologiczne (3,19,20). Warto dodać, że identyczność na poziomie 50% dla odcinków np. 20-40-aminokwasowych może być czysto przypadkowa (19). Inną praktyczną informacją jest to, że lepiej jest porównywać sekwencje aminokwasowe. Sekwencje 20-literowe (20 aminokwasów) niosą więcej informacji, niż 4-literowe (4 nukleotydy). Szansa, że przypadkiem znajdzie się wysoki poziom podobieństwa pomiędzy niespokrewnionymi ewolucyjnie sekwencjami jest wyższa dla sekwencji 4-literowych, niż 20-literowych. Zatem losowe znalezienie identycznych lub bardzo podobnych – a w rzeczywistości niehomologicznych – sekwencji DNA/RNA przy przeszukiwaniu baz jest znacznie bardziej prawdopodobne, niż dla sekwencji aminokwasowych w białkach (20). Platforma phylogeny.fr zawiera również odnośnik do wspomnianego algorytmu BLAST (*BLAST Explorer*) i użytkownik może dokonać przyrównania sekwencji z wybranej bazy do sekwencji zapytania i otrzymać wynik, który

w odpowiednim formacie, np. FASTA może dalej być poddany analizie filogenetycznej (21). Program BLAST występuje w kilku wersjach, a te zaimplementowane na platformie phylogeny.fr to: BlastP, BlastN, TBlastN, BlastX. Algorytm BlastP przeszukuje białkową bazę danych wykorzystując, jako sekwencję zapytania białko (sekwencję aminokwasową); BlastN przeszukuje bazę sekwencji nukleotydowych (DNA/RNA) wykorzystując, jako sekwencję zapytania zadaną sekwencję nukleotydową; TBlastN przeszukuje bazę białek, która powstała w wyniku translacji z sekwencji nukleotydowych, wykorzystując jako sekwencję zapytania białko; BlastX przeszukuje bazę białek wykorzystując jako sekwencję zapytania białko, które powstało w wyniku translacji z sekwencji nukleotydowej. W implementacji BLAST'a na phylogeny.fr użytkownik musi: 1) wprowadzić wybraną sekwencję zapytania w postaci nieobrobionej (ang. *raw*) lub w formacie FASTA, 2) zaznaczyć jaką wersję BLAST'a należy użyć: BlastP, BlastN, TBlastN, czy BlastX (domyślnie zaznaczony jest BlastP), 4) zaznaczyć względem jakiej bazy danych ma być przeszukiwana sekwencja zapytania (domyślnie zaznaczona jest nieredundantna białkowa baza NCBI z podaną datą aktualizacji), 3) zaznaczyć dla jakiej wartości oczekiwanej E mają być znalezione sekwencje (domyślnie jt. wartość 1e-05, czyli 10^{-5}) oznacza to, że BLAST znajdzie i wyświetli sekwencje, których lokalne podobieństwo będzie nie większe niż $E = 10^{-5}$). Po analizie program podaje zestaw sekwencji, które można z dużym prawdopodobieństwem uznać za homologiczne i zastosować je do dalszych analiz filogenetycznych. Podstawową zaletą platformy phylogeny.fr może być wykonanie analiz zarówno przez niespecjalistów przez użycie opcji „jedno kliknięcie" (ang. „*one-click*"), jak również przez bardziej zaawansowanych użytkowników, którzy mogą zmieniać zestawy różnorodnych parametrów, np. w opcji dla zaawansowanych (ang. *advanced*). Zbiór sekwencji wprowadzony w jednym z formatów np. FASTA, może być następnie przyrównany z wykorzystaniem programów najpowszechniej wykorzystywanych do przyrównań wielosekwencyjnych (ang. *multiple alignment*), takich jak: MUSCLE, ProbCons, T-Coffee, 3D-Coffee lub ClustalW. Domyślnie, platforma phylogeny.fr wykorzystuje program MUSCLE (22). Użytkownik powinien też podać swój adres e-mail, na który przesłane będą wyniki. Następnie, po kliknięciu, wykonane przyrównanie wielosekwencyjne jest poddawane obróbce przez zaimplementowany na platformie program Gblocks, który automatycznie usuwa najmniej efektywnie dopasowane fragmenty przyrównanych sekwencji (23). W wariancie bardziej zaawansowanym można zaznaczyć opcję, aby usuwane były wszystkie przerwy w dopasowaniu wielosekwencyjnym. Potem program konstruuje drzewo filogenetyczne na bazie jednej z czterech metod: 1) największej wiarygodności (ML, ang. *Maximum Liklelihood*), 2) parsymonii, 3) przyłączania sąsiada (NJ, ang. *Neighbour Joining*), 4) bayesowską (21, www.phylogeny.fr). W opcji „jedno kilknięcie" drzewo tworzone jest na podstawie algorytmu największej wiarygodności (ML), który obok metody bayesowskiej jest uznawany za najlepszy do przewidywania drzew filogenetycznych (24). W modelu największej wiarygodności zaimplementowanym na platformie phylogeny.fr wykorzystuje się model ewolucyjny substytucji (podstawień)

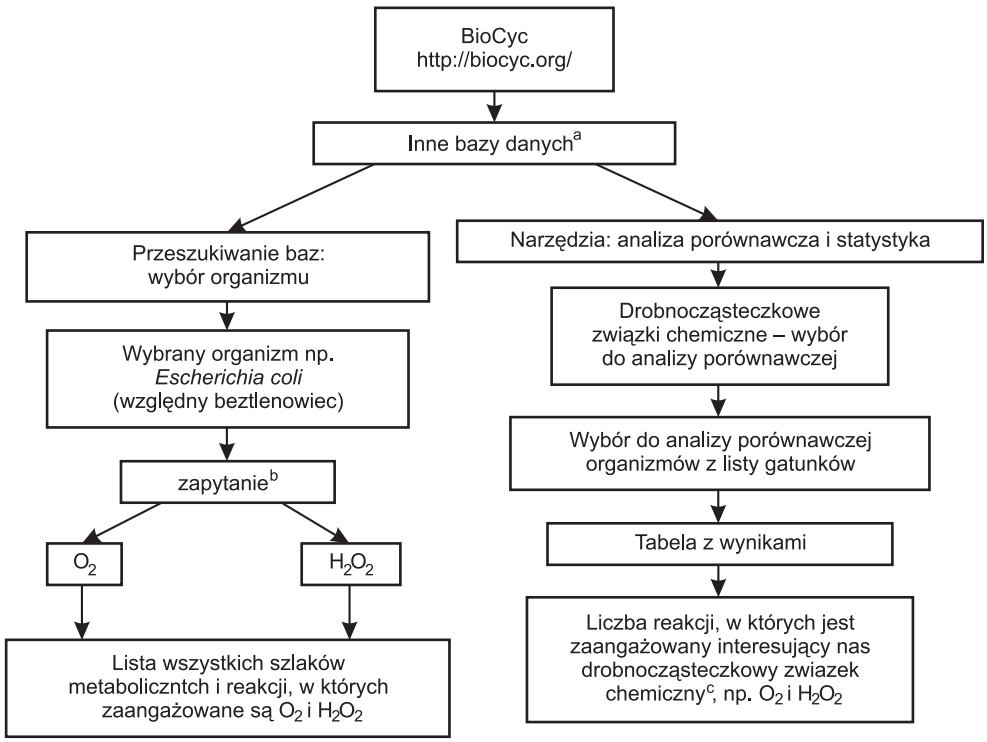
nukleotydów: HKY85 (25) i substytucji aminokwasów: WAG (26). Statystyczna istotność drzewa oceniana jest dwoma metodami tzw. metodą samopróbkowania (ang. *bootstrap*) lub za pomocą tzw. przybliżonego testu ilorazu wiarygodności (aLRT, ang. *approximate Likelihood Ratio Test*) (27, www.phylogeny.fr). Domyślnie ustawiona jest metoda aLRT, która jest znacznie szybsza i nie ustępuje wiarygodnością metodzie samoporóbkowania. W rezultacie, użytkownik otrzymuje na swój adres e-mail ostateczny wynik – a właściwie łącze internetowe do strony phylogeny.fr z wynikiem, którym jest narysowane drzewo filogenetyczne. Czas otrzymania drzewa wynikowego zależy m.in. od ilości i długości przyrównywanych sekwencji, zadanego modelu ewolucyjnego, metody wybranej do statystycznego oszacowania drzewa. Platforma umożliwia podglądnięcie całej procedury tworzenia drzewa od momentu przyrównania wielosekwencyjnego do otrzymania ostatecznego drzewa o określonej topologii. Na tym etapie drzewo może być uwidocznione i przekształcone do postaci wymaganej przez użytkownika, przy użyciu jednego z programów: Drawgram, Drawtree i TreeDyn. W opcji „jedno kliknięcie" drzewo jest przedstawiane w programie TreeDyn, który umożliwia jego edycję i zmianę ukorzenienia drzewa, koloru, grubości, wielkości gałęzi i czcionek (21). Ostateczny obraz uzyskanego drzewa może być wyeksportowany/zapisany w pliku PDF lub PNG, w postaci gotowej do publikacji. Warto podkreślić, że ważnym praktycznym składnikiem platformy phylogeny.fr jest również konwerter formatów powszechnie używanych w bioinformatyce. Użytkownik może wykorzystać wymienioną platformę do przekształcenia dowolnej sekwencji nukleotydowej/aminokwasowej zapisanej w jednym formacie w inny format, który jest wymagany do działania różnych aplikacji bioinformatycznych.

4. BioCyc – uniwersalny zbiór danych dla szlaków metabolicznych i genomów

W organizmach żywych metabolizm opiera się na trzech podstawowych filarach: 1) przemian rozmaitych związków drobno- i wielocząsteczkowych, 2) aktywności enzymów, które katalizują reakcje chemiczne przekształcające wymienione grupy związków chemicznych, 3) aktywności określonych grup genów, które kodują odpowiednie enzymy. Spośród wielu baz danych, które dedykowane są szlakom metabolicznym na szczególną uwagę zasługuje baza BioCyc (http://biocyc.org/), która zawiera zbiór baz szlaków metabolicznych i genomów (PGDBs, ang. *Pathway/Genome Databases*) (28). Baza ta jest *de facto* metabazą, czyli ma charakter nadrzędny wobec wielu innych baz katalogujących szlaki metaboliczne i geny/enzymy. BioCyc zawiera obecnie kolekcję baz wtórnych z w pełni zsekwencjonowanymi genomami dla 505 różnych organizmów, wśród których są przedstawiciele trzech głównych domen życia: *Bacteria*, *Archaea*, *Eucarya*. Najwięcej jest genomów mikroorganizmów, najmniej zwierząt. Głównymi bazami wtórnymi w obrębie BioCyc są: EcoCyc, MetaCyc i Hu-

manCyc. EcoCyc zawiera pełny genom wraz ze szlakami metabolicznymi/enzymami i regulacją transkrypcyjną dla *Escherichia coli* (29). MetaCyc jest największą bazą wtórną w obrębie BioCyc i zawiera ponad 1400 rozmaitych szlaków metabolicznych zweryfikowanych eksperymentalnie dla ponad 1800 organizmów. MetaCyc pozwala na analizę metabolizmu podstawowego i wtórnego wraz z określonymi genami/enzymami i substratami/produktami danych reakcji (30). Z kolei baza HumanCyc stanowi cenne źródło informacji nt. genomu i szlaków metabolicznych/enzymów człowieka (31). Zgodnie z tym, o czym wspomniano, najobszerniejszą bazą jest MetaCyc, co nie zmienia faktu, że wszystkie wymienione bazy są zintegrowane w obrębie metabazy BioCyc i przejście pomiędzy poszczególnymi podbazami jest szybkie i proste. Ponadto BioCyc integruje inne bazy dla genów/enzymów i szlaków metabolicznych, których pełną listę można znaleźć na stronach: 1) http://biocyc.org/otherpgdbs.shtml, 2) http://biocyc.org/biocyc-pgdb-list.shtml#tier2. Bazy mogą być przeszukiwane na kilka sposobów. Tradycyjną metodą jest wpisanie słowa kluczowego w polu wyszukiwarki. Bazę można też przeszukiwać bezpośrednio poprzez podanie skrótu identyfikatora np. dla genu. Jednak najbardziej interesująca jest możliwość porównywania pomiędzy wybranymi z listy grupami organizmów: a) wybranych szlaków metabolicznych, b) map metabolicznych/genomowych. W BioCyc zaimplementowany jest również BLAST, czyli istnieje możliwość znalezienia sekwencji homologicznych DNA/RNA lub białka względem sekwencji zapytania dla wybranego genu/enzymu. Baza podaje również statystykę przy porównywaniu wybranych organizmów dla takich parametrów jak: a) typy reakcji z uwzględnieniem podziału na reakcje z udziałem związków drobnocząsteczkowych, białek, reakcji dotyczących transportu, udziału poszczególnych klas enzymów w określonych grupach reakcji (klasyfikacja enzymów wg kategorii komisji enzymatycznej – E.C., ang. *Enzyme Commission*) itp., b) podział na szlaki metaboliczne np. z uwzględnieniem szlaków biosyntetycznych, szlaków transdukcji sygnału itp., c) udział związków drobnocząsteczkowych jako substratów/produktów, inhibitorów, aktywatorów, czy kofaktorów enzymów, d) wykrywanie białek ortologicznych (ortologów) oraz białek unikatowych dla wybranych organizmów, e) wykrywanie białek biorących udział w procesach transportu przez błonę komórkową, f) identyfikacja genów przypadających na jednostkę transkrypcyjną i ilość operonów przypadającą na szlak metaboliczny. Na rysunku 1 przedstawiono wybrane sposoby wyszukiwania informacji dotyczących związków drobnocząsteczkowych. Szlaki metaboliczne przedstawiane są w czytelnej postaci jako ciąg następujących po sobie reakcji. Jeśli enzymy katalizujące poszczególne reakcje są pokazane pogrubioną czcionką oznacza to, że udział takiego enzymu w katalizowaniu konkretnej reakcji – w przypadku danego organizmu – jest potwierdzony eksperymentalnie.

W BioCyc i głównych podbazach wchodzących w jej skład dla każdego enzymu są odnośniki do innych baz danych, takich jak np. NCBI, UniProt/Swiss-Prot, czy baz dedykowanych konkretnemu organizmowi np. TAIR (ang. *The Arabidopsis Information Resource*) dla *A. thaliana* (tab.) W bazach tych można znaleźć informacje nt. sekwencji



Rys. Dwa schematycznie przedstawione sposoby przeszukiwania metabazy BioCyc (BioCyc.org 13.0 Website) i baz z nią stowarzyszonych pod kątem zaangażowania związków drobnocząsteczkowych: tlenu (O_2) i nadtlenku wodoru (H_2O_2) w szlakach/reakcjach metabolicznych u wybranych grup organizmów. [a]Inne bazy: AraCyc (www.arabidopsis.org/biocyc/index.jsp), CalbiCyc (http://pathway.candidagenome.org/), DictyCyc (http://dictybase.org/Dicty_Info/dictycyc_info.html), EcoCyc (http://ecocyc.org/), HumanCyc (http://humancyc.org/), MedicCyc (http://www.noble.org/MedicCyc/index.html), MetaCyc (http://metacyc.org/), MouseCyc (http://mousecyc.jax.org/), RiceCyc (http://www.gramene.org/pathway/), SolCyc (http://solcyc.sgn.cornell.edu/), Yeast Biochemical Pathways (http://pathway.yeastgenome. org/biocyc/); [b]zapytanie może również dotyczyć nazwy: organizmu, enzymu, genu, szlaku metabolicznego, itp. [c]bazy nie obejmują drobnocząsteczkowych związków chemicznych, które są bardzo rozpowszechnie u wszystkich organizmów, takich jak: H_2O, H^+, ATP, ADP, AMP, fosforany, difosforany, NAD^+, NADH, $NADP^+$, NADPH, utlenionych/zredukowanych akceptorów elektronów (www.biocyc.org).

kodującej cDNA/mRNA, czy sekwencji aminokwasowej poszukiwanego enzymu. Zarówno w przypadku szlaku metabolicznego, enzymu, czy substancji drobnocząsteczkowej, baza podaje zwięzłe informacje związane z tematem zapytania, czyli np. krótką charakterystykę danego szlaku metabolicznego, wraz z podaniem odnośników do literatury.

Główną zaletą bazy BioCyc jest możliwość przewidywania *de novo* szlaków metabolicznych dla organizmu, którego genom został zsekwencjonowany, na podstawie istniejącej już bazy szlaków i enzymów. Przewidywanie szlaków biochemicznych *in silico* jest szczególnie istotne w przypadku obiektów, które trudno poddają się ma-

nipulacjom genetycznym, np. wyłączenie określonego genu/genów jest zbyt skomplikowane lub wręcz niemożliwe. Stąd klasyczna eksperymentalna analiza biochemiczna tego typu organizmów jest znacznie utrudniona. Dobrym przykładem zastosowania takiego podejścia do analizy metabolizmu było zaproponowanie sposobu asymilacji siarki u ekstremofilnej bakterii *Acidithiobacillus ferrooxidans* (32). BioCyc/MetaCyc daje możliwość pobrania z Internetu odpowiedniego oprogramowania narzędziowego tzw. PATHWAY TOOLS (narzędzi do szlaków metabolicznych), które można wykorzystać do konstruowania *in silico* szlaków metabolicznych dla dowolnego organizmu o zsekwencjonowanym genomie. Na podstawie tak przewidywanych szlaków metabolicznych, np. dla organizmu patogenicznego, można typować przemiany metaboliczne/enzymy, które należy zahamować, aby przykładowo nie rozwijał się jakiś proces chorobowy wywoływany przez ten organizm. Przykładem takich badań może być analiza *in silico* metabolizmu zarodźca malarii *Plasmodium falciparum* (33). Innym ważnym aspektem zastosowania BioCyc i powiązanych z nią baz wtórnych jest możliwość przewidywania, w jakim kierunku należy modyfikować szlaki metaboliczne, które są np. związane z biosyntezą ważnych z przemysłowego punktu widzenia związków chemicznych lub degradacją substancji szczególnie toksycznych dla środowiska. Metabaza BioCyc stwarza szerokie pole do analiz metabolizmu *in silico* u wybranych organizmów, które mogą mieć zastosowanie w biotechnologii.

5. Eukariotyczne bazy genomowe – poszukiwanie elementów regulatorowych

Transkrypcyjna kontrola ekspresji genów jest fundamentalnym procesem dla wewnątrz- i międzykomórkowych szlaków sygnałowych. Transkrypcja DNA angażuje aktywacje rożnych białkowych czynników transkrypcyjnych tzw. elementów regulatorowych typu *trans*, które oddziałują z krótkimi specyficznymi sekwencjami DNA w promotorach genów, czyli z elementami regulatorowymi typu *cis*. Zrozumienie i rozpracowanie sieci sygnałowej opartej na tych elementach jest kluczem do poznania zasad funkcjonowania organizmów żywych, np. regulacji rozwoju embrionalnego człowieka, aklimatyzacji roślin do różnorodnych warunków środowiska, wrodzonej odporności na choroby, cyklu komórkowego i wielu innych podstawowych procesów biologicznych (34). Jedne z największych możliwości analizy promotorów genów mamy u *Arabidopsis thaliana*. Internetowa strona arabidopsis.org daje możliwość korzystania z programu 'Patmatch' (ang. *Pattern Matching* – dopasowanie wzorca), który wyszukuje krótkie sekwencje nukleotydowe (< 20 nukleotydów) w promotorach, intronach i w nie transkrybowanych rejonach (NTR, ang. *Non-Transcriptable Regions*) wszystkich genów *A. thaliana*. Dodatkowo, arabidopsis.org daje użytkownikowi do dyspozycji narzędzie 'Motif analysis' (analiza motywów sekwencyjnych), pozwalające na znalezienie sześcioliterowych sekwencji nukleotydowych – prawdopodobnych elementów regulatorowych typu *cis*, które zawierają promoto-

ry genów wybranych do przeszukiwania. Pracę ułatwia także zestaw narzędzi 'Bulk Data Retrieval and Analysis' (odzyskanie i analiza dużej ilości danych), w którym to można znaleźć opisy genów, skopiować wybrane sekwencje promotorów, sekwencje kodujące białka, sekwencje białek, intronów i NTR-ów, jak również analizować eksperymenty mikromacierzowe. Użytkownik ma także łatwy dostęp do wszystkich danych poprzez serwer ftp.

Do analizy promotorów bardzo przydatne są także inne programy wyszukujące elementy regulatorowe typu *cis* w danej sekwencji. Takie możliwości daje nam baza PLACE (ang. *Plant Cis-acting Regulatory DNA Elements*, www.dna.affrc.go.jp/PLACE/), czy PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) dedykowane roślinom i strona SoftBerry (http://linux1.softberry.com/berry.phtml), która dotyczy szeroko rozumianej analizy genomów eukariotycznych.

6. Podsumowanie

Znaczenie biologicznych baz danych, sposobów gromadzenia i katalogowania informacji będzie wzrastać. Niech znakiem wzrastającego znaczenia baz będzie powstanie w 2009 r. nowego czasopisma „Database. The Journal of Biological Databases and Curation", które jest dedykowane biologicznym bazom danych i powiązanemu z nimi oprogramowaniu (35). Biologiczne bazy danych ze zgromadzonymi w nich informacjami mają podstawowe znaczenie dla rozwoju dziedziny wiedzy jaką jest biologia systemów (ang. *systems biology*), która poprzez łączenie wszystkich elementów, takich jak: ekspresja genów, aktywność enzymów, sieć wzajemnie powiązanych szlaków metabolicznych, stawia sobie za zadanie stworzenie holistycznego obrazu organizmu, który znajduje odzwierciedlenie w jego fenotypie (36).

Dziękujemy Panu magistrowi Piotrowi Gawrońskiemu za cenne uwagi dotyczące pracy z eukariotycznymi bazami danych. Praca jest finansowana dzięki projektowi Welcome 2008/1 z Fundacji na rzecz Nauki Polskiej i funduszom strukturalnym z Unii Europejskiej oraz projektowi COST nr: 595/N-COST/2009/0.

Literatura

1. Lesk A. M., (2003), *Introduction to Bioinformatics*, Oxford University Press, 115-159.
2. Selzer P. M., Marhöfer R. J., Rohwer A., (2008), *Applied Bioinformatics. An Introduction*, Springer-Verlag Berlin Heidelberg, 1-30.
3. Xiong J., (2006), *Essential bioinformatics*, Cambridge University Press, 63-74.
4. Mackiewicz P., Zakrzewska-Czerwińska J., Cebrat S., (2005), Biotechnologia, 3, 7-21.
5. Venter J. C., Adams M. D., Myers E. W., et al., (2001), Science, 291, 1304-1351.
6. McPherson J. D., Marra M., Hillier L., et al., (2001), Nature, 409, 934-941.
7. Robbins R. J., (1994), Publishing Research Quarterly, 10, 3-27.
8. Baxevanis A. D., (2003), *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc., 1.1.1-1.1.4.
9. Stein L. D., (2003), Nature Rev. Genet., 4, 337-345.

10. Selzer P. M., Marhöfer R. J., Rohwer A., (2008), *Applied Bioinformatics. An Introduction*, Springer-Verlag Berlin Heidelberg, 45-74.
11. Baxevanis A. D., Ouellette B. F. F., (2004), *Bioinformatyka, Podręcznik do analizy genów i białek*, PWN, Wraszawa, 490.
12. Higgs P. G., Attwood T. K., (2008), *Bioinformatyka i ewolucja molekularna*, PWN, Warszawa, 531.
13. Menlove K. J., Clement M., Crandall K. A., (2009), *Bioinformatics for DNA Sequence Analysis. Series: Methods in Molecular Biology*, 537, Ed. D. Posada, Humana Press, 1-22.
14. Galperin M. Y., Cochrane G. R., (2009), Nucleic Acid Res., 37, D1-D4.
15. Huang S., Ruiqiang Li R., Zhang Z., et al., (2009), Nat. Genet., 41, 1275-1281.
16. Schnoes A. M., Brown S. D., Dodevski I., Babbitt P.C., (2009), Plos Comput. Biol., 5, 1-13.
17. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J., (1990), J. Mol. Biol., 215, 403-410.
18. Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J., (1997), Nucleic Acids Res., 25, 3389-3402.
19. Claverie J-M., Notredame C., (2007), *Bioinformatics for dummies*, 2nd ed., Wiley Publishing Inc., 199-234.
20. Holmes R. M., (2007), *A cell biologist's guide to modeling and bioinformatics*, John Wiley & Sons, Inc., 9-34.
21. Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J. F., Guindon S., Lefort V., Lescot M., Claverie J. M., Gascuel O., (2008), Nucleic Acids Res., 36, W465-469.
22. Edgar R. C., (2004), Nucleic Acids Res., 32, 1792-1797.
23. Castresana J., (2000), Mol. Biol. Evol., 17, 540-552.
24. Hall B. G., (2005), Mol. Biol. Evol., 22, 792-802.
25. Hasegawa M., Kishino H., Yano T., (1985), J. Mol. Evol., 2, 160-174.
26. Whelan A., Goldman N., (2001), Mol. Biol. Evol., 18, 691-699.
27. Anisimova M., Gascuel O., (2006), Syst. Biol., 55, 539-552.
28. Karp P. D., Ouzounis C. A., Moore-Kochlac C., Goldovsky L., Kaipa P., Ahren D., Tsoka S., Darzentas N., Kunin V., Lopez-Bigas N., (2005), Nucleic Acids Res., 33, 6083-6089.
29. Keseler I. M., Bonavides-Martinez C., Collado-Vides J., Gama-Castro S., Gunsalus R. P., Johnson D. A., Krummenacker M., Nolan L. M., Paley S., Paulsen I. T., et al., (2009), Nucleic Acids Res., 37, D464-D470.
30. Caspi R., Foerster, H., Fulcher C. A., Kaipa P., Krummenacker M., Latendresse M., Paley S., Rhee S.Y., Shearer A., Tissier C., Walk T. C., Zhang P., Karp P. D., (2008), Nucleic Acids Res., 36(1), D623-D631.
31. Romero P., Wagg J., Green M. L., Kaiser D., Krummenacker M., Karp P. D., (2004), Genome Biol., 6, R2.
32. Valdés J., Veloso F., Jedlicki E., Holmes D., (2003), BMC Genomics, 15, 51.
33. Yeh I., Hanekamp T., Tsoka S., Karp P. D., Altman R. B., (2004), Genome Res., 14, 917-924.
34. Geisler M., Kleczkowski L. A., Karpinski S., (2006), Plant J., 45, 384-398.
35. Landsman D. , Gentlemanm R., Kelso J., Ouellette B. F. F., (2009), Database, ID bap002, doi:10.1093/database/bap002.
36. de Backer P., Waele D. D., van Speybroeck L., (2010), Acta Biotheor., 58, 15-49.