Biolog w cyberprzestrzeni

Wojciech Makałowski
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, USA

1. Wprowadzenie

Do niedawna jedynym źródłem informacji na którym biolodzy mogli polegać były materiały publikowane w drukowanych czasopismach i książkach. Aczkolwiek media te pozostają nadal podstawowym źródłem informacji biologicznej, to biolodzy obecnie muszą być w stanie korzystać z informacji przechowywanych elektronicznie w niezliczonej liczbie różnych baz danych rozproszonych na całym świecie, w różnych punktach internetu. Taka forma „publikacji" ma wiele zalet. Wśród nich, najważniejsze to łatwość dostępu, szybkość i efektywność wyszukiwania informacji, oraz stosunkowo niskie koszty. Artykuł ten stanowi wprowadzenie do internetu dla zagubionych lawiną napływającej informacji biologów. Przedstawione zostaną związane z tym pewne koncepcje, narzędzia pozwalające na poruszanie się w *cyberprzestrzeni*, oraz wybrane miejsca interesujące biologów. Należy jednak pamiętać, że wraz z rozwojem internetu, ulega też on często głębokim przemianom. Na przykład, średni półokres życia adresu w internecie wynosi około czterech lat. Może się zatem zdarzyć, że zasoby opsane tutaj już nie istnieją. Bardziej szczegółowe opisy i podręczniki na ten temat są szeroko dostępne w różnych komercyjnych publikacjach. Znalezienie tych źródeł nie powinno stanowić większych problemów po przeczytaniu tego artykułu.

2. Internet

Internet nie jest pojedynczą siecią komputerową, jakby sugerowała sama nazwa. Tak naprawdę jest to sieć sieci, łączących ogromną liczbę fizycznych sieci komputerowych na całym globie. Komputery, które są podłączone do internetu komunikują się ze sobą za pomocą specjalnych protokołów i standardów komunikacyjnych, pozwalających komputerom na odnalezienie siebie nawzajem i wymianę informacji pomiędzy nimi. Samo działanie internetu wykracza poza ramy tego opracowania, jednakże bardziej szczegółowe informacje na ten temat można znaleźć w wielu podręcznikach poświęconych

internetowi (1,2). Szacuje się, że obecnie ponad siedem milionów komputerów jest bezpośrednio podłączonych do internetu.

Początkowo komunikacja pomiędzy komputerami ograniczała się do prostych metod, takich jak poczta elektroniczna (e-mail) do wymiany listów pomiędzy użytkownikami i ftp (*file transfer protocol*) do przesyłania dużych zbiorów pomiędzy komputerami. Z chwilą wzrostu zapotrzebowania na wymianę informacji, nowe sytemy przekazu danych zostały rozwinięte, w znacznym stopniu ułatwiając dostęp do informacji. Najbardziej znanym z tych systemów jest gopher, powstały na Uniwersytecie w Minnesocie. Nie będzie on jednak tu omawiany, z tego względu, że w ostatnim czasie został on praktycznie wyparty przez WWW. Wraz z rozwojem technik multimedialnych nastąpiło zapotrzebowanie na systemy pozwalające na wymianę również tego typu informacji. Zbiegło się ono z rozwojem „światowej pajęczyny" (WWW — *World Wide Web*) i narzędzi z nią związanych.

3. Poczta elektroniczna

Poczta elektroniczna jest standardowym narzędziem użytkownika internetu i stała się głównym środkiem komunikacji pomiędzy biologami na całym świecie. Poczta elektroniczna jest nie tylko sposobem wymiany korespondencji, ale także sposobem komunikowania się z komputerami, które prowadzą automatyczne serwisy, takie jak porównywanie sekwencji nukleotydowych i aminokwasowych, różnego rodzaju analizy danych, czy też wyszukiwanie i automatyczne przesyłanie informacji z baz danych. Główną zaletą serwerów działających poprzez pocztę elektroniczną jest jej łatwość użycia i szeroki dostęp. Wystarczy odpowiednio sformułować list i żądana usługa zostaje wykonana przez nierzadko odległy od nas komputer, a wyniki przesłane wprost do naszej skrzynki. Poczta elektroniczna zwalnia też użytkownika z instalowania i utrzymywania własnych kopii różnych, nieraz dużych i skomplikowanych, zestawów oprogramowania oraz baz danych. Minusem tego rodzaju usług jest to, że nie są one interaktywne, a produkowane wyniki przesyłane są do nas wyłącznie w fromie tekstowej (nie graficznej). Zbiór serwerów dostępnych za pomocą poczty elektronicznej, szczególnie użytecznych dla biologów molekularnych, jest przedstawiony w tab. 1. Najbardziej aktualna lista takich serwerów jest utrzymywana przez Amosa Bairocha z Uniwersytetu w Genewie. Można ją otrzymać poprzez ftp pod adresem *expasy.hcuge.ch* (zbiór *serv_ema.txt*, w katalogu */databases/info*). Poza nielicznymi wyjątkami, wysłanie listu ze słowem „help" (bez cudzysłowu) na podany w tab. 1 adres, spowoduje automatyczne przesłanie szczegółowej instrukcji użycia danego serwera.

4. FTP (*file transfer protocol*)

Innym ograniczeniem poczty elektronicznej jest to, że przesyłanie dużych zbiorów i programów jest utrudnione lub wręcz niemożliwe. O wiele wygod-

niejszym sposobem przesyłania dużych ilości informacji jest użycie anonimowego ftp (*anonymous file transfer protocol*). System ten pozwala na podłączenie się do „obcego" komputera (hosta) bez posiadania ważnego konta i hasła na danym komputerze. Użytkownik „wlogowuje" się do komputera używając jako nazwy konta „anonymous" i swojego adresu elektronicznego jako hasła. Protokół ten pozwala nastepnie na poruszanie się po różnych „publicznych" katalogach aby odnaleźć interesujące użytkownika zbiory i przerzucenie ich na jego własny komputer. Wiele grup ogłasza dostępność swoich danych przez anonimowe ftp poprzez podanie adresu komputera i katalogu, na którym dane te są przechowywane. Obecnie system ftp jest integralną częścią przeglądarek WWW, co umożliwia łatwe przeglądanie zasobów obcych komputerów bez znajomości katalogów na których dane są przechowywane. Wybór miejsc ftp, szczególnie interesujących dla biologów molekularnych, zawarte są w tab. 2.

5. „Światowa pajęczyna" (*World Wide Web* — WWW)

Zarówno poczta elektroniczna, jak i ftp są w pewnym sensie statycznym sposobem komunikowania się z obcymi komputerami. Wymagają one też stosunkowo dużej ilości informacji na temat: czego konkretnie szukamy, i gdzie jest to dostępne (wymagana jest znajomość adresów i skomplikowanych niekiedy nazw zbiorów i usług dostępnych w sieci). Pierwszą próbą przełamania tego stanu było wprowadzenie systemu *gopher*. Był on jednym z pierwszych narzędzi nawigacyjnych internetu, umożliwiających stosunkowo łatwe i bezwysiłkowe podróżowanie po internecie. System ten jednak, aczkolwiek wciąż w użyciu, okazał się niewystarczający w okresie rewolucji multimedialnej i nie będzie tu bliżej dyskutowany. Zainteresowanych odsyłam do najnowszej wersji FAQ (*Frequently Asked Questions*), którą można uzyskać poprzez anonimowe ftp, łacząc się z komputerem *rtfm.mit.edu* (katalog */pub/usenet/news.answers*, zbiór *Gopher-faq*), lub porzez pocztę elektroniczną, wysyłając list o treści „send usenet/news.answers/Gopher-faq" na adres: *mail-server@rtfm.mit.edu*.

Prawdziwą rewolucję w rozwoju internetu przyniosło wprowadzenie „światowej pajęczyny" (WWW) przez Europejskie Centrum Badań Jądrowych (CERN), którą dalej będę nazywał po prostu pajęczyną. Pajęczyna, podobnie zresztą jak gopher, oparta jest na zasadzie klient-serwer, tzn. program umieszczony na komputerze użytkownika (tzw. klient) oddziałuje z programem umieszczonym na komputerze w innym punkcie sieci (tzw. serwer). Podstawową nowością w działaniu pajęczyny jest jej organizacja: umożliwia ona bezpośredni dostęp do poszczególnych miejsc poprzez „kliknięcie" na tzw. hiperpołączenia (*hyperlinks*). Hiperpołączenie, to po prostu zbiór wyrazów wyróżniony od reszty tekstu kolorem i (lub) podkreśleniem (tzw. hipertekst), który po „naciśnięciu" przenosi użytkownika w nowe miejsce sieci. Użytkownik znajduje żądaną informację poprzez przenoszenie się z miejsca na miejsce w procesie adekwatnie nazwanym „serfowaniem" (*Web-surfing*). Do wybrane-

go miejsca można też dotrzeć „tradycyjnie" poprzez wypisanie jego adresu, który w pajęczynie jest zwany URL (*uniform resource locator*). Właśnie te adresy, niewidoczne dla użytkownika w czasie „serfowania" z użyciem hiperpołączeń, pozwalają nie tylko na dotarcie do innych miejsc pajęczyny, ale też na połączenie się z dowolnym miejscem ftp lub z dowolnym gopherem.

Biolodzy z chwilą powstania pajęczyny uczynili z niej jedno z podstawowych źródeł wymiany informacji. Chyba nie ma działu biologii, który by nie był na niej reprezentowany. Szczególnie upodobali ją sobie biolodzy molekularni umieszczając w niej dosłownie wszystko od przepisów laboratoryjnych, po gotowe wyniki doświadczeń laboratoryjnych, włączając w to pełną dokumentację, jak np. zdjęcia żeli elektroforetycznych, czy obrazy mikroskopowe. Pajęczyna też stała się miejscem uzupełniającym tradycyjne formy publikacji. Zbiory danych wykorzystywanych w badaniach są często zbyt obszerne, aby były drukowane, często też czasopisma ograniczają liczbę dozwolonych w publikacji rycin; wreszcie współczesne formy prezentacji, takie jak np. filmy poglądowe nie mieszczą się w tradycyjnych formach publikacji. To wszystko znajduje swe miejsce na sieci dzięki pajęczynie. Jednym z pierwszych czasopism, które dopuszczało niekonwencjonalne formy dodatków było wydawane przez Cold Spring Harbor Laboratory „Genome Research". Właśnie to czasopismo, jako uzupełnienie do jednego z artykułów umieściło w sieci film wideo, w którym zademonstrowana została nowa metoda mapowania genów (URL: http://207.22.83.2:443/cshl/journals/gr/supplement/samad/vid1.html). Praktycznie każde znaczące czasopismo próbuje zaakcentować swą obecność w pajęczynie. Jedne, takie jak „Nature" czy „Science" prezentują w sieci jedynie spisy treści wraz z wybranymi dodatkowymi informacjami, inne, jak „Journal of Biological Chemistry" publikują „elektronicznie" całe artykuły. W tym drugim przypadku ogólnie dostępne są zazwyczaj spisy treści i streszczenia artykułów, natomiast całe artykuły osiągalne są jedynie dla prenumeratorów lub za wniesieniem odpowiedniej opłaty. Spis wybranych czasopism elektronicznych interesujących dla biologów został zamieszczony w tab. 3. Również indywidualni badacze lub instytucje zamieszczają wyniki swoich badań w pajęczynie (np. autor tego opracowania umieścił napisany przez siebie rozdział książki na serwerze NCBI, URL: http://www.ncbi.nlm.nih.gov/Makalow/sines.html). Wykorzystując możliwości pajęczyny udaje się często wzbogacić oryginalną publikację przez połączenia z odpowiednimi bazami danych. Bardzo popularne jest, np. hiperłączenie referencji zamieszczonych w artykule z bazą literaturową (w wymienionym przykładzie z bazą Medline), co umożliwia czytelnikowi głębsze zaznajomienie się z cytowaną literaturą. Idąc dalej, czytelnik wykorzystując moc rozwijanego w NCBI *Entrez*, może zaznajomić się z innymi artykułami na dany temat. W podobny sposób można połączyć wymienione w artykule sekwencje nukleotydowe lub białkowe z odpowiednią bazą w NCBI i znów wykorzystując potencjał *Entrez* znaleźć dodatkowe informacje, które z kolei mogą doprowadzić do nowych odkryć. Technologia pajęczyny pozwala też na łatwe poruszanie się w obrębie tekstu, np. przenoszenie się z miejsca powołującego się na rycinę, bezpośrednio do niej.

Innym ciekawym aspektem wykorzystania „światowej pajęczyny" jest umieszczanie w niej różnego rodzaju serwisów analizujących dane. Chyba najbardziej znanym tego przykładem jest „pajęcza" wersja rodziny programów BLAST (http://www.ncbi.nlm.nih.gov/BLAST/), pozwalających na wyszukiwanie w bazie danych sekwencji podobnych do przesłanej przez użytkownika. Jedną z największych zalet tej formy wykonywania usług jest jej prostota. Użytkownik nawet z największym „komputerowstrętem" jest w stanie zanalizować sekwencję DNA, czy białka używając serwera BLAST w NCBI. Również w tym przypadku serwer wykorzystuje możliwości WWW dodając nowe właściwości do programu, jak np. bezpośredni dostęp do rekordu z sekwencją, która wykazała podobieństwo do sekwencji przesłanej; dalej można dotrzeć do publikacji z nią związanych, itd.

6. Wyszukiwanie informacji w „światowej pajęczynie"

Ilość informacji prezentowanej w pajęczynie jest tak olbrzymia, że znalezienie interesujących nas miejsc może być kłopotliwe. Aby ułatwić orientację w cyberprzestrzeni redakcje niektórych czasopism zdecydowały się na publikowanie artykułów wprowadzających w tę tematykę (3) oraz na regularne publikowanie miejsc w pajęczynie interesujących daną grupę czytelników (4). Innym sposobem jest rozpoczęcie „serfowania" od wirtualnych bibliotek, czyli miejsc w pajęczynie, które są kolekcją stron WWW, uszeregowanych według tematów. Dwa takie miejsca, szczególnie popularne wśród biologów to: WWW Virtual Library: Bioscience, utrzymywana przez Keitha Robinsona z Uniwersytetu Harwarda (http://golgi.harvard.edu/biopages/all.html) oraz Pedro's Biomolecular Research Tools (http://www.public.iastate.edu/~pedro/research_tools.html) zebrane przez Pedro Coutinho z Uniwersytetu Stanowego Iowa.

Innym bardziej wyszukanym sposobem wynajdywania czegoś w pajęczynie jest używanie tzw. *search engines*. Programy te używają różnych algorytmów do wyszukiwania informacji na podstawie słów kluczowych zawartych w tytułach stron WWW lub całych tekstach. Użytkownicy jednej z najbardziej popularnych przeglądarek pajęczyny Netscape mają bezpośredni dostęp do kilku takich programów przez naciśnięcie guzika „Net Search" w górnej części okna Netscape. Wśród dostępnych programów znajdują się tak popularne jak: Yahoo, Alta Vista, InfoSeek, czy WebCrawler. Ponieważ każdy z tych programów używa nieco innego algorytmu przeszukiwania, otrzymane wyniki trochę się różnią. Dlatego też, zamiast używać tych programów osobno, można wykorzystać jedno z narzędzi, które zrobi to za nas. Na przykład, program SavvySearch (http://guaraldi.cs.colostate.edu:2000/) zbiera wyniki przeszukiwania z wykorzystaniem różnych „search engine", usuwa duplikaty i przesyła zbiór hiperpołączeń do użytkownika. Dodatkowo, narzędzie to jest obecnie dostępne aż w dwudziestu językach (niestety, na wersję polską musimy jeszcze poczekać).

TABELA 1
WYBRANE SERWERY POCZTY ELEKTRONICZNEJ UŻYTECZNE DLA BIOLOGÓW MOLEKULARNYCH

Nazwa	Rodzaj usługi	Adres serwera
BLAST	wyszukiwanie sekwencji homologicznych z baz białkowych i nukleotydowych za pomocą programu BLAST	blast@ncbi.nlm.nih.gov
BLOCKS	wyszukiwanie motywów białkowych	block@howard.fhcrc.org
CENSOR	sprawdzanie nadesłanej sekwencji na obecność ludzkich sekwencji powtarzalnych	server@charon.lpi.org
EBI File Server	otrzymywanie rekordów z baz białkowych (SWISS-PROT) i nukleotydowych (EMBL)	netserv@ebi.ac.uk
GenFinder	przewidywanie struktury genów i struktury drugorzędowej białek	service@bchs.uh.edu
GenMark	przewidywanie rejonów kodujących białka za pomocą metody model łańcuchów Markowa	genmark@ford.gatech.edu
GRAIL	przewidywanie rejonów kodujących białka za pomocą algorytmu sieci neuronowych	grail@ornl.gov
HUGEMAP	przeszukiwanie ludzkiej mapy fizycznej Genethon	hugemap@genethon.fr
nnpredict	przewidywanie struktury drugorzędowej białek	nnpredict@celeste.ucsf.edu
RETRIEVE	wyszukiwanie informacji zawartych w bazach NCBI	

TABELA 2
WYBRANE SERWERY ftp UŻYTECZNE DLA BIOLOGÓW MOLEKULARNYCH

Miejsce	Zasoby	Adres
NCBI	bazy danych: GenBank, SWISS-PROT, PIR, repbase, HOVERGEN i in. programy: BLAST, MACAW, Entrez, Authorin, censor i in.	ncbi.nlm.nih.gov
EBI	bazy danych: EMBL, SWISS-PROT różne programy dla biologii molekularnej na różne komputery	ftp.ebi.ac.uk
ExPASy	bazy danych: Enzyme, EPD, Prosite, SWISS-PROT, SWISS-2DPAGE, SWISS-3DPAGE i in.	expasy.hcuge.ch
GDB	bazy danych: GDB, OMIM	ftp.gdb.org
IuBio	duża kolekcja oprogramowania dla biologii molekularnej, wśród nich PHYLIP (analiza filogenetyczna), readseq (rozpoznawanie i przekształcanie ok. dwudziestu formatów sekwencji nukleotydowych)	ftp.bio.indiana.edu

Literatura

1. Falk, B., (1994), The Internet Roadmap. Sybex, San Francisco.
2. Krol, E., (1994), The Whole Internet User's Guide and Catalog. O'Reilly, Sebastopol.
3. Baxevanis, A.D., et al., (1996), Curr. Opin. Biotechnol., 7, 99-101.
4. Baxevanis, A.D., et al., (1996), Curr. Opin. Biotechnol., 7, 102.

The biologist in the Cyberspace

Summary

The Internet has become one of the main means of information exchange between biologist. The development of the World Wide Web has been especially influential in expanding the use of the Internet. This review is meant as an introductory Internet tutorial for the layman with the emphasis on benefits for molecular biologists. The use of e-mail servers and descriptions of the ftp and WWW sites relevant to the biologists are discussed.

Key words:
WWW, biology.

Adres do korespondencji:
Wojciech Makałowski, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA, fax: 301-480-9241;
e-mail: makalow@ncbi.nlm.nih.gov