STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Damian ZAPART, Tomasz WALLER, Magdalena TKACZ
Instytut Informatyki, Uniwersytet Śląski

ANALIZA WYDAJNOŚCI METOD DOSTĘPU DO DANYCH POCHODZĄCYCH Z EKSPERYMENTÓW MIKROMACIERZOWYCH W ZALEŻNOŚCI OD SPOSOBU ICH PRZECHOWYWANIA

Streszczenie: W niniejszej pracy porównano (pod względem wydajności dostępu do danych) dwie metody gromadzenia plików zawierających dane z eksperymentu mikromacierzowego z wykorzystaniem mechanizmów relacyjnych baz danych. Pomiar czasu dostępu do zawartości pliku miał na celu wytypowanie tej metody, która po zaimplementowaniu w systemie informatycznym (którego jedną z funkcjonalności było przechowanie i udostępnianie danych mikromacierzowych) cechowałaby się lepszymi paramatrami dostępu do danych podlegających następnie dalszemu przetwarzaniu.

Słowa kluczowe: mikromacierze DNA, systemy zarządzania bazami danych, SQL Server 2008, bioinformatyka

EFFICIENCY OF MICROARRAY DATA ACCESS MECHANISMS IN COMPARISION WITH DATA STORING METHODS

Summary: The main goal of this article was to select the most optimal method of storing files from microarray experiment in database. The paper compares two methods of storing the data using the database: storing the files from microarray experiment directly in database or storing the reference to the files from the microarrays experiment saved on hard disk.

Keywords: DNA microarray, database management system, SQL Server 2008, bioinformatics

1. Wstęp

Wzrost popularności wykorzystania różnego rodzaju systemów informatycznych przekłada się bezpośrednio na spotęgowanie starań mających przede wszystkim na celu

poprawę wydajności ich pracy. Wiele spośród tych systemów często powiązanych jest z silnikiem bazodanowym, którego jedną z funkcji jest zapewnienie efektywnego mechanizmu gromadzenia i udostępniania danych.

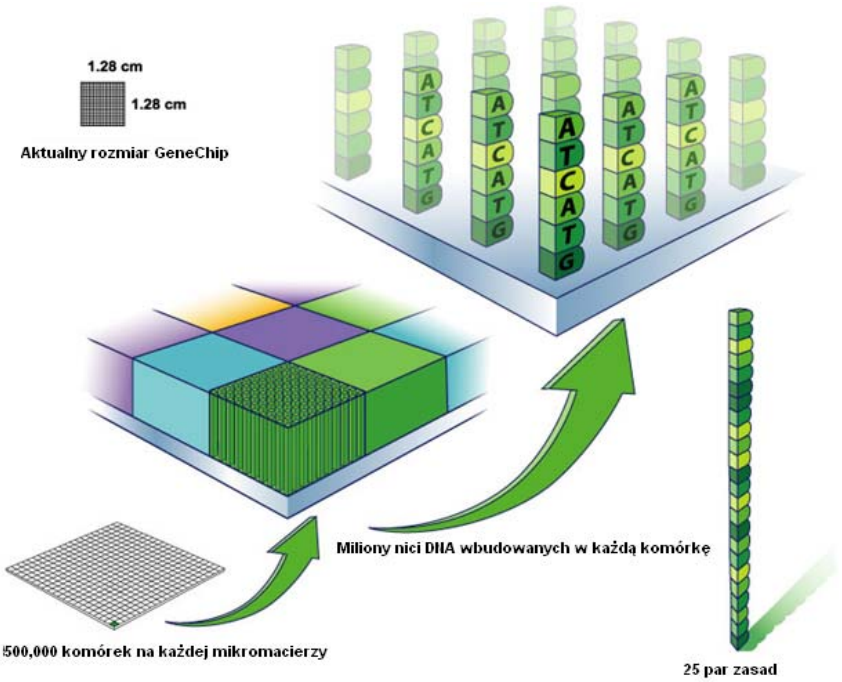
Problem wyboru miejsca przechowywania zbiorów danych jest powszechnie znany w informatyce. Istnieją liczne prace [1, 2, 3, 4, 5] poświęcone tej tematyce. Ogólnie przyjęto zasadę, że pliki niewielkiego rozmiaru powinny być przechowywane w tabelach bazy danych. Pliki z danymi z eksperymentów mikromacierzowych są jednak specyficzne i trudno określić je jako „małe". Nie można tu również z góry przy projektowaniu systemu założyć (z niewielkim marginesem błędu), jakiego rzędu wielkości pliki trzeba będzie zgromadzić. Dlatego też celem niniejszej pracy jest porównanie dwóch metod gromadzenia danych z eksperymentów mikromacierzowych (plików zawierających znormalizowane dane pochodzące z eksperymentów mikromacierzowych) z wykorzystaniem mechanizmów bazy danych.

2. Mikromacierze

Mikromacierze DNA wykorzystywane są przez biologów molekularnych do badania ekspresji wielu tysięcy genów równocześnie. W eksperymentach naukowych najczęściej stosowane są 2 rodzaje mikromacierzy, których budowa i sposób wykorzystania są odmienne. Są to mikromacierze oligonukleotydowe (ang. oligonucleotide microarrays), które charakteryzują krótkie sondy DNA oraz mikromacierze cDNA (ang. cDNA microarrays) zbudowane z sond odpowiadających pełnym badanym sekwencjom genów. Wspólną cechą obu typów jest wykorzystanie efektu hybrydyzacji materiału badanego z sondami mikromacierzy do określenia aktywności poszczególnych genów. Na potrzeby eksperymentu opisanego w niniejszym artykule użyto danych pochodzących z badań, w których wykorzystano mikromacierze oligonukleotydowe typu HG-U133A firmy Affymetrix. Na powierzchni mikromacierzy HG-U133A znajduje się ponad 500 000 pojedynczych pól, zwanych spotami, na które naniesione zostały fragmenty łańcucha DNA o długości ok. 25 zasad azotowych (sondy). Sekwencje te są odcinakami reprezentatywnymi dla danego genu, informacje o nich mogą być wyszukiwane w publicznie dostępnych bioinformatycznych bazach danych.

Eksperyment mikromacierzowy rozpoczyna się w momencie naniesienia materiału - to jest RNA, pobranego i wyizolowanego od danego osobnika, oznakowanego uprzednio znacznikiem fluorescencyjnym na płytkę mikromacierzy. W procesie hybrydyzacji (łączenia) następują trwałe połączenie badanego materiału z krótkimi sekwencjami na mikromacierzy zgodnie z zasadami komplementarności. Następnie mikromacierz zostaje wypłukana od nadmiaru niezwiązanych molekuł oraz zeskanowana, a otrzymany obraz poddany analizie ilościo-

wej. Celem doświadczenia przeprowadzonego z wykorzystaniem mikromacierzy jest zmierzenie intensywności poziomu fluorescencji dla poszczególnych pól mikromacierzy, będącego zarazem reprezentacją aktywności danego transkryptu w badanej tkance. Im wartość fluorescencji dla danego pola jest większa, tym większy poziom ekspresji wykazuje dany transkrypt.



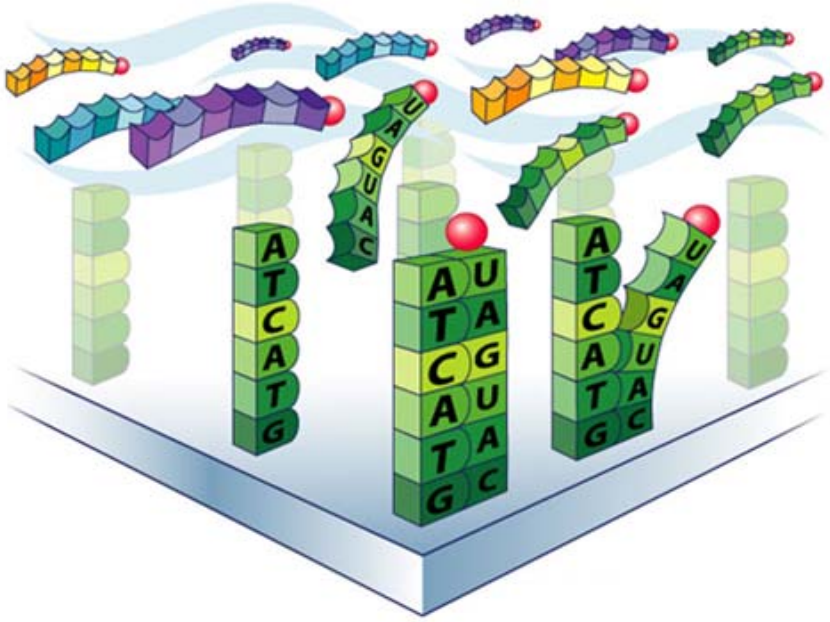
Rys. 1. Schemat budowa mikromacierzy oligonukleotydowej [7]
Fig. 1. Diagram of the construction oligonucleotide microarray [7]

Mikromacierze oligonukleotydowe są z założenia jednorazowego użytku, co oznacza, że po naniesieniu na mikromacierz materiału genetycznego jednego osobnika nie nadaje się ona do powtórnego wykorzystania - użycia w innym eksperymencie. Konsekwencją tego jest fakt, że nie jest możliwe porównywanie poziomów ekspresji genów dla wielu osobników (lub próbek) jednocześnie (w trakcie jednego eksperymentu, wykorzystując jedną mikromacierz). W celu przeprowadzenia eksperymentu mikromacierzowego na większej liczbie osobników lub też powtórzenia eksperymentu konieczne jest wykonanie całego doświadczenia od początku z wykorzystaniem nowych mikromacierzy.

3. Charakterystyka eksperymentu z wykorzystaniem mikromacierzy oligonukleotydowych

Każda z zastosowanych w eksperymencie mikromacierzy jest skanowana, a uzyskany z niej obraz jest przetwarzany specjalnymi algorytmami w celu określenia poziomu ekspresji badanych na niej genów. Proces przygotowania danych z eksperymentu do analizy porównawczej określany jest mianem analizy niskiego poziomu i obejmuje m.in. takie

zagadnienia, jak korekcję tła i normalizację wyników. Rozmiar uzyskanego po przetworzeniu zbioru danych zależy od liczby zastosowanych w nim mikromacierzy (liczba ta przekłada się na liczbę kolumn) i typu mikromacierzy (od jej rodzaju zależy liczba wierszy). W pracy tej wykorzystano do doświadczenia zbiory o rozmiarach 22 283 wierszy (genów) x 25 próbek, 22283 wierszy x 50 próbek, 22283 wierszy x 100 próbek oraz 22 283 wierszy x 150 próbek. Przykładowy fragment zbioru danych po przeprowadzonej analizie niskiego przedstawiono w tabeli 1.



Rys. 2. Schemat procesu hybrydyzacji z uwzględnieniem zasad komplementarności na mikromacierzy [7]
Fig. 2. Flowchart for hybridization to the principles of complementarity in the microarray [7]

Tabela 1

Fragment zbioru danych uzyskanych z eksperymentu mikromacierzowego. Przedstawione liczby dotyczą danych rzeczywistych (dane z eksperymentu o id: E-TABM-15 z pubicznego repozytorium ArrayExpress). Kolumny zaetykietowane Kontrola 1 i Kontrola 2 dotyczą danych uzyskanych od pacjentów z grupy kontrolnej, będącej punktem odniesienia, kolumny zaetykietowane Pacjent I, Pacjent II dotyczą pomiaru ekspresji genów w próbkach pobranych z tkanki płuc zaatakowanej przez nowotwór złośliwy, w których chcemy określić zmianę ekspresji genów.

ID GENU	KONTROLA I	KONTROLA II	PACJENT I	PACJENT II
RUNX2	6,622876	7,186759	6,798659	6,561131
SRC	2,290999	2,851007	2,568154	2,601464
PACAP	3,824643	3,740529	3,725303	3,506553
RNASEL	3,389659	3,407762	3,889823	3,396017
ULBP2	9,039343	9,257492	9,035255	8,958395

4. Problem gromadzenia danych

W wielu ośrodkach badawczych najprostszym i zarazem najpopularniejszym sposobem jest przechowywanie danych z eksperymentu mikromacierzowego bezpośrednio na dysku twardym komputera wraz z ewentualnym podziałem na podkatalogi, które reprezentują poszczególne badania. Podejście to nie jest optymalne, ma ono liczne niedogodności, takie jak:

- Utrudnione wykonywanie kopii zapasowych,
- Problem udostępniania danych poza sieć lokalną,
- Brak ewidencjonowania dostępu do danych,
- Utrudnienie wyszukiwania danych czy opisów eksperymentu.

Rozwiązanie to, choć pozornie wydaje się być wystarczające, warto zastąpić systemem do gromadzenia danych z eksperymentów mikromacierzowych. Można to wykonać budując system, który wspierany mechanizmami zarządzania bazami danych realizowałby funkcjonalność gromadzenia i późniejszego udostępniania plików danych badaczom, tworząc tym samym lokalne repozytorium danych ośrodka naukowego.

5. Metodyka

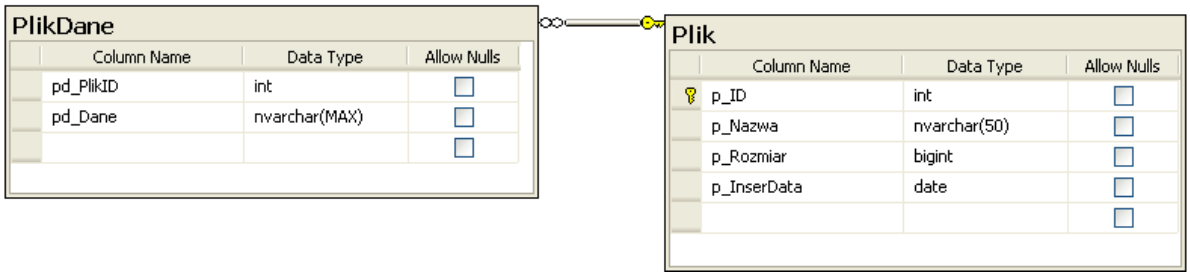
W niniejszym rozdziale zaproponowano i opisano dwie propozycje rozwiązania problemu przechowywania i udostępniania plików. Rozwiązanie testowano przy wykorzystaniu darmowego silnika bazy danych SQL Server 2008 w wersji Express Edition oraz darmowego środowiska obliczeniowego The R Project. Badaniu podlegała efektywność różnych sposobów gromadzenia plików w bazie danych. Metody gromadzenia podzielono na dwie grupy:

- Pliki trzymane są bezpośrednio w tabelach/tabeli bazy danych (import danych do tabel).
- Pliki są fizycznie przechowywane poza bazą, w bazie natomiast umieszczono tylko referencję do pliku z danymi.

Kryterium oceny efektywności był czas mierzony od momentu otwarcia połączenia do bazy, wykonywanie zapytania typu SELECT aż do chwili, kiedy dane zostały wczytane do zmiennej typu `data.frame` (typ reprezentujący tabelę) w środowisku The R Project, w którym odbywa się dalsza analiza danych.

Przeprowadzenie tego typu testów wydaje się być uzasadnione. Często praktykowane w tworzeniu rozwiązań bazodanowych podejście polega na przyjęciu założenia, że treść plików wczytana do systemu zostaje zapisana w mechanizmach bazy danych. Przy takim założeniu zalecanym zwykle rozwiązaniem jest utworzenie dwóch tabel (rys. 3.), z czego w wierszach pierwszej przechowywane są dane opisujące plik, takie jak np.: nazwa, rozmiar,

data wstawienia do bazy, a w drugiej tylko rzeczywiste treści. Struktura taka zapewnia większą wydajność w odróżnieniu od jednotabelowego rozwiązania, w którym pliki danych zgromadzone są tylko w jednej z wielu kolumn w tabeli, ponieważ nie wymaga wczytywania całej zawartości pliku do pamięci RAM (na przykład w trakcie wykonywania zapytania SQL np. o listę nazw plików zgromadzonych w bazie).



Rys. 3. Przykładowy schemat relacji pomiędzy tabelami.
Fig. 3. Example diagram of the relationship between the tables.

Rozwiązanie dwutabelowe ma następujące zalety:

- Dane związane są za pomocą relacji – brak osieroconych lub nieistniejących rekordów relacji.
- Brak jest konieczności obsługi strumieni wejścia i wyjścia, dane z serwera SQL są zwracane automatycznie przez obiekt klasy SqlCommand.
- Można utworzyć kopię zapasową bazy danych.

Ale także wady, a mianowicie:

- Szybki rozrost rozmiaru bazy danych.
- Uszkodzenie tabeli wiąże się z utratą integralności i danych.
- Przed wykorzystaniem danych konieczne jest dokonanie zapisu strumienia informacji na dysk do pliku tymczasowego.

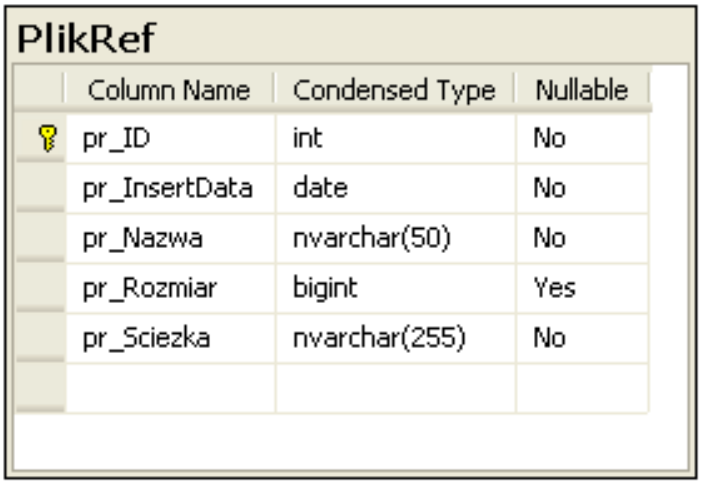
Drugim z rozwiązań jest przechowywanie plików danych na dysku serwera, natomiast w tabeli bazy przechowywana jest tylko referencja (odwołanie) do pliku z danymi. Przykładowy schemat struktury bazy danych (rys. 4) zakłada istnienie tylko jednej tabeli powiązanej z przechowywaniem plików.

Rozwiązanie to cechują następujące zalety:

- Tabela referencji do plików ma niewielki rozmiar.
- Wydajność wykonania poleceń SQL jest większa.
- Brak jest konieczności ponownego zapisu danych na dysk przed użyciem.

Wadami tego rozwiązania są natomiast:

- Brak możliwości utworzenia relacji SQL do pliku na dysku.
- Nie można wykonać wspólnej kopii zapasowej dla bazy danych i plików.
- Możliwa jest utrata integralności (powstawanie rekordów z pustą referencją do pliku oraz plików na dysku, które nigdy nie zostaną wybrane).



PlikRef

	Column Name	Condensed Type	Nullable
🔑	pr_ID	int	No
	pr_InsertData	date	No
	pr_Nazwa	nvarchar(50)	No
	pr_Rozmiar	bigint	Yes
	pr_Sciezka	nvarchar(255)	No

Rys. 4. Tabela w, której znajdują się referencje do plików na dysku.
Fig. 4. The with a reference to the files on disk

Środowisko The R Project [6], z którego korzystano również podczas przeprowadzania opisanego tu eksperymentu, daje możliwość wykorzystania pakietu bibliotek bioinformatycznych Bioconductor [7]. Z ich użyciem można wczytać, a później dokonać analizy danych pochodzących z eksperymentu z wykorzystaniem mikromacierzy.

Dodatkowo, poprzez wykorzystanie biblioteki RODBC środowisko R może zostać połączone (poprzez dostawcę ODBC) bezpośrednio do silnika bazy danych, przez co możliwe jest wykonywanie - z poziomu The R Project – zapytań dowolnego typu do bazy danych.

Dla potrzeb eksperymentu środowisko The R Project zostało – przy użyciu pakietu RODBC – połączone z silnikiem bazy danych Microsoft SQL Server 2008. Po nawiązaniu i otwarciu połączenia w zależności od typu testowanej metody gromadzenia plików w bazie danych zadawano zapytania dotyczące danych zawartych w określonej tabeli. W zależności od metody gromadzenia danych wynik zwrócony przez zapytanie SQL był bezpośrednio zapisywany do zmiennej typu data.frame w środowisku The R Project lub w przypadku, kiedy zapytanie zwróciło referencje do pliku, wówczas do zmiennej wczytywany (funkcja read.table) był plik z otrzymanej ścieżki.

Przed przystąpieniem do eksperymentu wygenerowano (poprzez powielenie kolumn danych z rzeczywistego eksperymentu mikromacierzowego) cztery pliki zawierające rzeczywiste dane pochodzące z eksperymentu z wykorzystaniem mikromacierzy. Liczba wierszy w każdym z plików była stała i wynosiła 22 284 wiersze (liczba probset Id (sond) na mikromacierzy HG-U133A + nagłówek kolumny). Zwiększała się natomiast liczba kolumn, co miało symulować wzrost liczby eksperymentów mikromacierzowych. Tym samym zwiększał się rozmiar pliku i dla kolejnych plików liczba kolumn wynosiła odpowiednio: 25, 50, 100, 150.

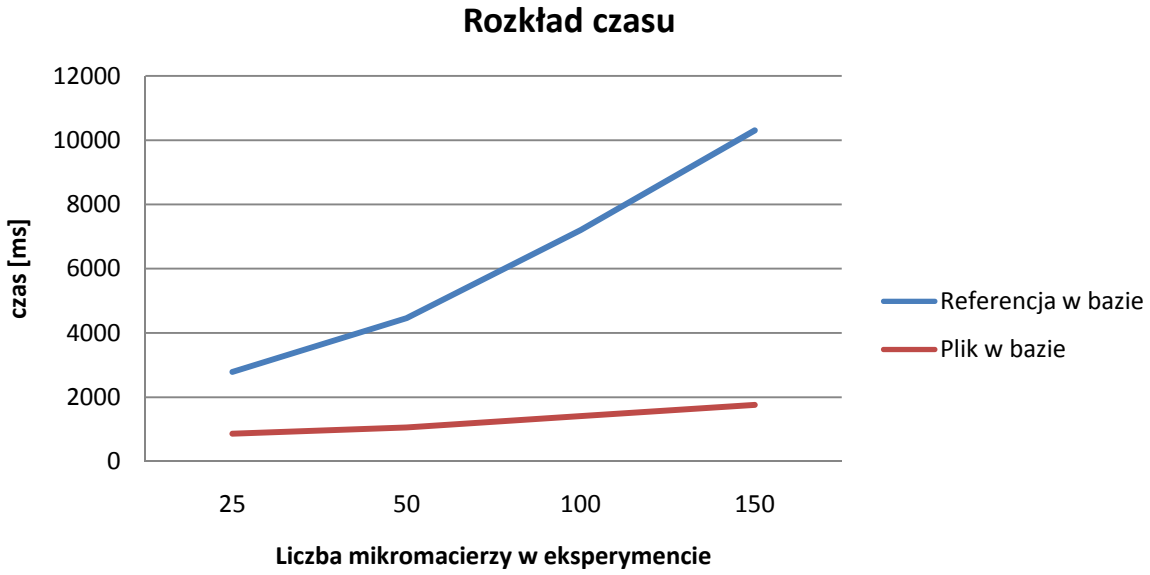
Przy każdorazowym wykonaniu opisanych kroków, składających się z inicjalizacji połączenia do serwera R(D) COM, nawiązania i otwarcia połączenia do bazy danych oraz

pobrania z niej danych przy użyciu klauzuli SELECT i wczytania wyników do zmiennej, mierzony był czas jego wykonania.

6. Wyniki

Czas potrzebny na wczytanie do zmiennej w The R Project danych przechowywanych fizycznie poza bazą danych był kilkakrotnie dłuższy niż w przypadku danych przechowywanych bezpośrednio w bazie (uprzednio zaimportowanych do bazy).

Ponadto, różnica pomiędzy czasem potrzebnym na wczytanie danych z wykorzystaniem referencji do pliku w bazie a czasem potrzebnym na załadowanie danych zapisanych w bazie znacząco rośnie wraz ze wzrostem liczby mikromacierzy wykorzystanych w eksperymencie (odpowiednio dla 25, 50 100 i 150 mikromacierzy jest to 321%, 420%, 510% i 585% zmiana - wzrost czasu). Na rysunku 5 przedstawiona została zależność pomiędzy czasem potrzebnym na przekazanie do środowiska obliczeniowego R danych w zależności od proponowanych w pracy metod składowania i rozmiaru pliku z danymi (liczbą danych z eksperymentów).



Rys. 5. Rozkład czasu dostępu do danych z poziomu środowiska obliczeniowego R-Project w zależności od wybranej metody gromadzenia danych w silniku bazodanowym

Fig. 5. Schedule time access to data from the computing environment R-Project, depending on the chosen method of data collection in the database engine

7. Wnioski

W pracy jednoznacznie wykazano, że metoda bezpośredniego przechowywania danych z eksperymentów mikromacierzowych w bazie danych jest bardziej efektywna niż metoda, w której w DBMS przechowywane byłyby wyłącznie odwołania do plików na dysku.

Różnice w czasie dostępu do danych w przypadku rozwiązania, w którym dane przechowywane są w bazie, stają się znaczące wraz ze wzrostem wielkości zbioru danych z eksperymentów. W rozwiązaniu tym należy co prawda uwzględnić czas potrzebny na import danych do bazy, jednakże jest to czynność wykonywana jednorazowo dla każdego eksperymentu i nie będzie miała wpływu na wzrost czasu niezbędnego do wykonywania analiz danych. Co więcej, w przypadku wielokrotnych analiz i wykonywania wielu zapytań do bazy czas związany z jednorazowym importem danych do bazy można uznać za pomijalny.

BIBLIOGRAFIA

1. Storing Uploaded Files in a Database or in the File System with ASP.NET 2.0 http://imar.spaanjaars.com/QuickDocId.aspx?quickdoc=414
2. Storing Binary Files Directly in the Database Using ASP.NET 2.0 http://www.4guysfromrolla.com/articles/120606-1.aspx
3. To BLOB or Not To BLOB: Large Object Storage in a Database or a Filesystem http://research.microsoft.com/apps/pubs/default.aspx?id=64525
4. A Method of Microarray Data Storage Using Array Data Type http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2709412/
5. Oficjalna strona środowiska The R Project http://www.r-project.org/
6. Oficjalna strona bioinformatycznego pakietu Bioconductor http://www.bioconductor.org/
7. http://www.affymetrix.com/about_affymetrix/media/image-library.affx

Recenzenci: Prof. dr hab. inż. Andrzej Polański
Prof. dr hab. inż. Andrzej Świerniak

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

DNA microarrays are used to measure the expression of thousands of genes simultaneously. During the typical DNA microarray experiment the large number of data is acquired. Final dataset size depends on the microarray used number and their type. The data from the experiments are usually stored in files on the hard disk (for each experiment there is a separate file with the experiment results), which is not the effective solution. In the case of storing the data from successive experiments, the management of these files is getting more and more difficult. In order to avoid such situation, it is possible to substitute manual storing of the data with a suitably projected and implemented information technology system using the Database Management Systems. Thanks to using such system in microarrays laboratory the researchers gain the access to the large number of functionalities facilitating the work with data from many experiments.

The main goal of this article was to select the most optimal method of storing files from microarray experiment in database. The analysis was performed with the use of MS SQL SERVER 2008 Express Edition. In this paper two methods of storing the data using the database: storing the files from microarray experiment directly in database and storing the reference to the files from the microarrays experiment saved on hard disk. The authors tested the time needed to load the data from microarray experiment to R-project computing environment using the proposed methods of storing the data with the use of DBMS.

We proved that storing the files from microarray experiment directly in database is more efficient than storing the reference to the files from the microarrays experiment saved on hard disk.

Adresy

Magdalena Tkacz: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39, 41-200 Sosnowiec, Polska. magdalena.tkacz@us.edu.pl

Damian Zapart: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39, 41-200 Sosnowiec, Polska. Student II roku studiów MU na kierunku informatyka. damian.zapart@gmail.com.

Tomasz Waller: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39, 41-200 Sosnowiec, Polska. Student II roku studiów MU na kierunku informatyka. tomek@bioinformatyka.com.pl