STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Krzysztof ŚWIDER, Bartosz JĘDRZEJEC, Damian JAKOBUS
Politechnika Rzeszowska, Katedra Informatyki i Automatyki

INTEGRACJA I WIZUALIZACJA DANYCH STRUKTURALNYCH NA PRZYKŁADZIE BAZ SIECI BIOLOGICZNYCH

Streszczenie. Do najważniejszych zadań *biologii systemowej* należy badanie sieci biologicznych, celem ich lepszego poznania i praktycznego wykorzystania. W artykule przedstawiono zintegrowane środowisko BiNArr do wstępnego przetwarzania oraz wizualizacji danych, pochodzących z wybranych baz sieci biologicznych. Zaproponowano jednolitą grafową reprezentację struktur pozyskanych z oryginalnych zasobów oraz przygotowano moduły do ich wizualizacji i edycji. Przewidziano także możliwość eksportu grafów w formatach wymaganych przez aplikacje drążenia grafów. Do prezentacji wybranych funkcji systemu posłużyły – udostępnione w bazach KEGG – mapy szlaków metabolicznych oraz sieci oddziaływań białko-białko, pozyskane z zasobów DIP.

Słowa kluczowe: dane strukturalne, grafy, wstępne przetwarzanie danych, edycja, wizualizacja, standard XML, sieci biologiczne

INTEGRATION AND VISUALIZATION OF STRUCTURED DATA REPRESENTING BIOLOGICAL NETWORKS

Summary. The investigation of biological networks for their better understanding and making available for practical use is currently the important task in *systems biology*. The paper presents an integrated environment BiNArr aimed to perform some data preparation operations as well as visualization of the network data stored in biological databases. We proposed the unified graph representation for the structures extracted from original resources and developed the modules for their visualization and edition. Another important feature is the automatic coding of the resulting graphs in several formats required by different graph mining applications. In order to present some capabilities of the application, the structures from example databases representing metabolic pathways (KEGG) as well as protein-protein interactions (DIP) were used.

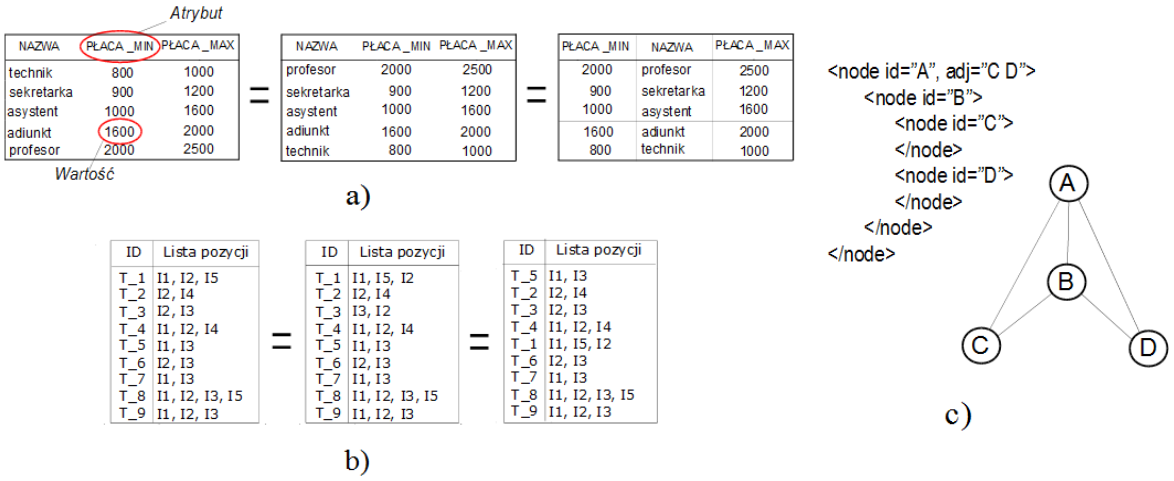
Keywords: structured data, graphs, data preparation, edition, visualization, XML standard, biological networks

1. Wprowadzenie

Do najważniejszych cech współczesnych baz danych, obok niespotykanego wcześniej tempa wzrostu ich objętości, należy z pewnością rosnący udział danych, których natura wykracza poza klasyczne formy typu *atrybut-wartość* czy tzw. *transakcje*, rozumiane jako podzbiory pozycji należących do pewnej kategorii. Dane te określane są jako *strukturalne* i występują w wielu dziedzinach, spośród których najczęściej wymienia się: chemię, bioinformatykę, użytkowanie sieci Web, analizę dokumentów XML, rozpoznawanie obrazów oraz zjawiska socjologiczne. Z uwagi na to, że dane strukturalne nie są opisywane z góry zadanym schematem, ich przetwarzanie i analiza wymagają z reguły zastosowania odrębnych metod oraz opracowania specjalnych narzędzi. Dotyczy to w szczególności problemów pozyskiwania oraz wstępnego przetwarzania danych z baz biologicznych rozważanych w tym artykule.

1.1. Dane strukturalne

Pojęcie danych strukturalnych w sposób intuicyjny zilustrowano na przykładach pokazanych na rys.1. Podstawową cechą danych strukturalnych jest zdolność do opisywania złożonych relacji pomiędzy atrybutami, stąd do ich reprezentowania są zwykle wykorzystywane bardziej złożone struktury, takie jak: sekwencje, drzewa i grafy.



Rys. 1. Przykładowe formy danych: a) tabele atrybut-wartość; b) transakcje; c) dane strukturalne
Fig. 1. Typical data in databases: a) the attribute-value tables; b) transactions; c) structured data

W ostatnim okresie można zaobserwować wyraźny wzrost zainteresowania m.in. drążeniem danych strukturalnych, a w rezultacie prowadzonych w wielu ośrodkach prac uzyskano wiele interesujących wyników, zarówno o charakterze podstawowym, jak i praktycznym [2]. Tradycyjne metody eksploracji danych, jak: segmentacja, klasyfikacja, wyszukiwanie częstych wzorców czy indeksowanie rozszerzono tak, aby mogły być zastosowane dla danych strukturalnych. Znacząca liczba prac dotyczy analizy sekwencji oraz drzew różnych typów,

jednak szczególną uwagę badaczy w ostatnim dziesięcioleciu skupiły struktury modelowane za pomocą grafów.

1.2. Bazy sieci biologicznych

System biologiczny może być opisany przez rozległą sieć biologiczną, złożoną z wielu mniejszych sieci. Ogólnie biorąc, system komórkowy jest reprezentowany za pomocą sieci trzech rodzajów: szlaki metaboliczne, oddziaływania białko-białko oraz sieci regulacji genów. Przedmiotem badań (tutaj) są dane opisujące szlaki metaboliczne (*metabolic pathways*) [8] oraz oddziaływania białko-białko (*protein-protein interactions*) [3].

Przykładem danych przechowywanych w bazach biologicznych są sieci szlaków metabolicznych, które stanowią struktury grupujące reakcje zachodzące w komórkach. Formalnie, sieć taka może być zdefiniowana jako zbiór obiektów oraz powiązań, jakie występują pomiędzy nimi. Obiektami są związki chemiczne (metabolity), reakcje biochemiczne, enzymy i geny. Wynikiem reakcji biochemicznych jest zbiór składający się z jednego lub więcej produktów będących związkami chemicznymi. Produkty są otrzymywane z innego zbioru związków, określanych jako substraty. Zasadniczo, reakcje mogą przebiegać w obu kierunkach, jednak w określonych warunkach fizjologicznych pewne reakcje zachodzą tylko w jedną stronę. W takim przypadku określa się je jako nieodwracalne, o ile inne warunki nie ulegną zmianie. Pewne reakcje wewnątrzkomórkowe przebiegają spontanicznie, jednak większość podlega działaniu katalizatorów w postaci jednego lub większej liczby enzymów, które znacznie przyśpieszają przebieg reakcji. Enzymy są białkami lub kompleksami białkowymi zakodowanymi za pomocą jednego lub większej liczby genów.

Zarówno w przypadku szlaków metabolicznych, jak i innych typów sieci biologicznych do modelowania ich struktur dość powszechnie stosuje się grafy.

1.3. Problematyka artykułu

Ogólnym celem prowadzonych przez autorów prac jest wstępne przetwarzanie danych strukturalnych pochodzących z baz biologicznych. W szczególności opisano własne zintegrowane środowisko BiNArr (*Biological Network Arranger*), potrzebne do transformacji oraz wizualizacji sieci biologicznych.

Edycja oraz interaktywna analiza struktur grafowych stanowią interesujący obszar badań, o czym świadczy powstanie specjalnych pakietów oprogramowania, wspomagających prace w tym zakresie. Jednocześnie nadal aktualnym zadaniem wydaje się adaptacja opracowanych technik i narzędzi dla potrzeb analizy danych z dziedziny bioinformatyki. Przykładem środowiska wspomagającego analizę sieci biologicznych jest pakiet Cytoscape [1], umożliwiający

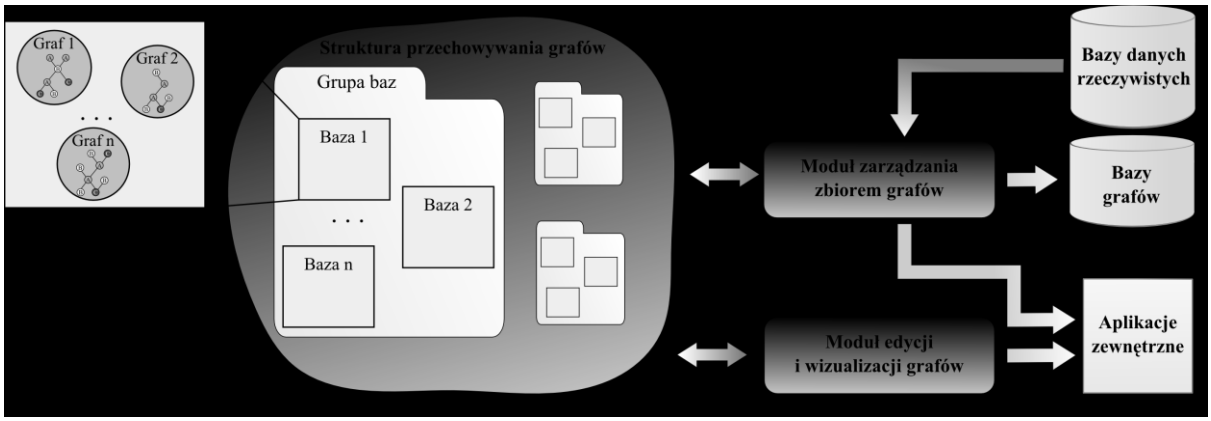
wizualizację, modelowanie oraz analizę sieci interakcji molekularnych i genetycznych. Innym przykładem jest oprogramowanie ProViz [4], służące do interaktywnej wizualizacji sieci interakcji białko-białko. W szczególności możliwa jest nawigacja w dużych grafach oraz wydobywanie istotnych, z biologicznego punktu widzenia, właściwości. Liczba dotychczas opracowanych pakietów jest znacząca, a zestawienie istotnych charakterystyk wybranych narzędzi do analizy sieci zawiera praca [1].

W odróżnieniu od niektórych, bardziej rozbudowanych środowisk, system BiNArr realizuje stosunkowo ograniczoną liczbę funkcji i jest wyraźnie ukierunkowany na przygotowanie danych strukturalnych do analizy. Zaproponowano jednolitą grafową reprezentację struktur, pozyskanych z oryginalnych zasobów oraz przygotowano moduły do ich wizualizacji i edycji. Przewidziano możliwość eksportu grafów w formatach wymaganych przez różne aplikacje drążenia grafów. Ponadto, wykonano testy weryfikujące przydatność aplikacji dla danych rzeczywistych.

2. Opis środowiska BiNArr

2.1. Architektura i funkcje systemu

Zintegrowane środowisko potrzebne do wstępnej obróbki i wizualizacji danych strukturalnych opisujących sieci biologiczne ma budowę modułową pokazaną na rys. 2.



Rys. 2. Ogólny schemat aplikacji BiNArr
Fig. 2. The general schema of BiNArr

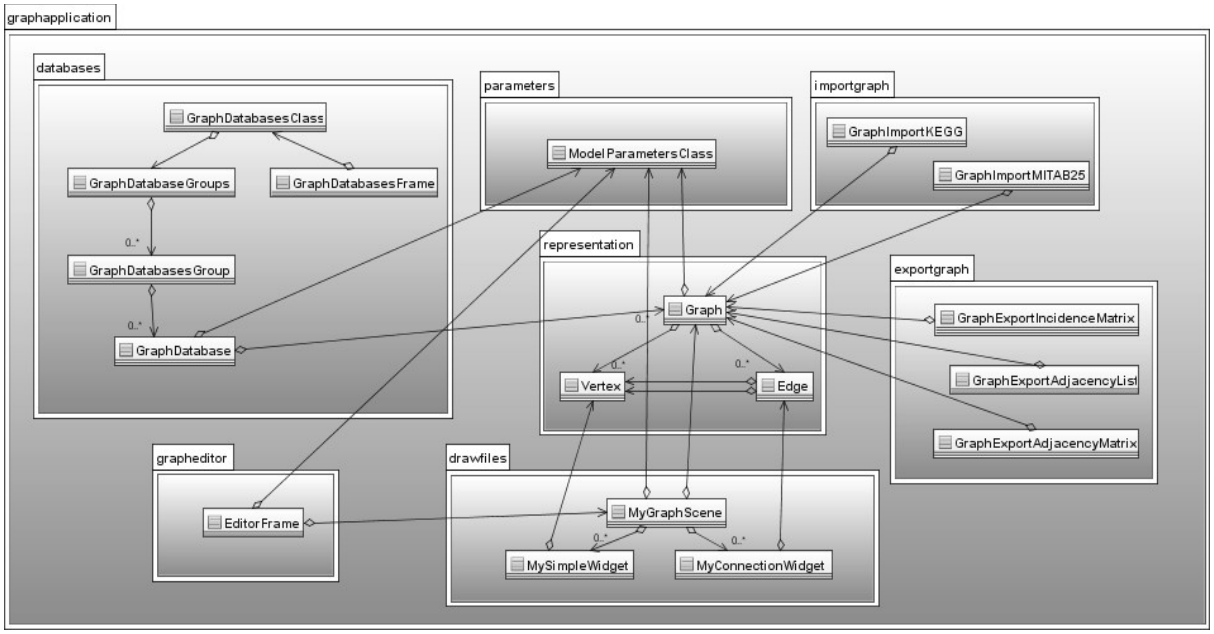
Na aplikację składają się w szczególności: struktura przechowywania grafów, moduł zarządzania zbiorem grafów oraz moduły edycji i wizualizacji. Struktura przechowywania grafów odzwierciedla istniejące w systemie grupy baz danych grafowych, którym przypisywana jest unikalna nazwa. Grupa baz obejmuje jednorodne, według pewnego kryterium, zbiory grafów. Taka organizacja wynika z przewidywanego przetwarzania grafów różnych typów, pochodzących z różnych źródeł. Każda baza grafów charakteryzuje się parametrami określa-

jącymi jakiego typu grafy mogą być w niej przechowywane (np. skierowane/nieskierowane, spójne/niespójne, z krawędziami etykietowanymi/nieetykietowanymi itp.). Do pewnej bazy możliwy jest zapis tylko tych grafów, które spełniają określone kryteria. Struktura zapisu grup i baz grafów na dysk przechowywana jest w pomocniczym pliku XML. W pliku tym, oprócz parametrów istniejących baz grafów, przechowywane są także ścieżki do plików zawierających pojedyncze grafy. Zapis każdego grafu w oddzielnym pliku został podyktowany względami wydajnościowymi.

Plik do przechowania grafu pozwala na elastyczny zapis szerokiej klasy grafów w postaci ujednoliconej struktury. Struktura ta składa się z trzech elementów: parametrów grafu, wierzchołków oraz krawędzi. Parametry obejmują m.in.: nazwę grafu, datę utworzenia, typ, paletę kolorów wykorzystaną do oznaczania wierzchołków, dopuszczalne wagi dla krawędzi itp. Wierzchołki grafów są opisywane za pomocą identyfikatorów, informacji o położeniu oraz ewentualnie takich danych, jak: kolor, kształt czy etykieta. Krawędzie definiowane są za pomocą identyfikatorów, źródłowego i docelowego wierzchołka oraz ewentualnej etykiety. Do realizacji głównych funkcji systemu służą moduły: zarządzania zbiorem grafów oraz edycji i wizualizacji.

2.2. Realizacja komputerowa

System zrealizowano w środowisku Java z wykorzystaniem ogólnie dostępnych bibliotek, jak np. Xerces, JUNG2, FreeHEP VectorGraphics i innych. Przygotowano także własne klasy, realizujące podstawowe zadania systemu. Tworzą one pakiety pokazane na rys. 3.



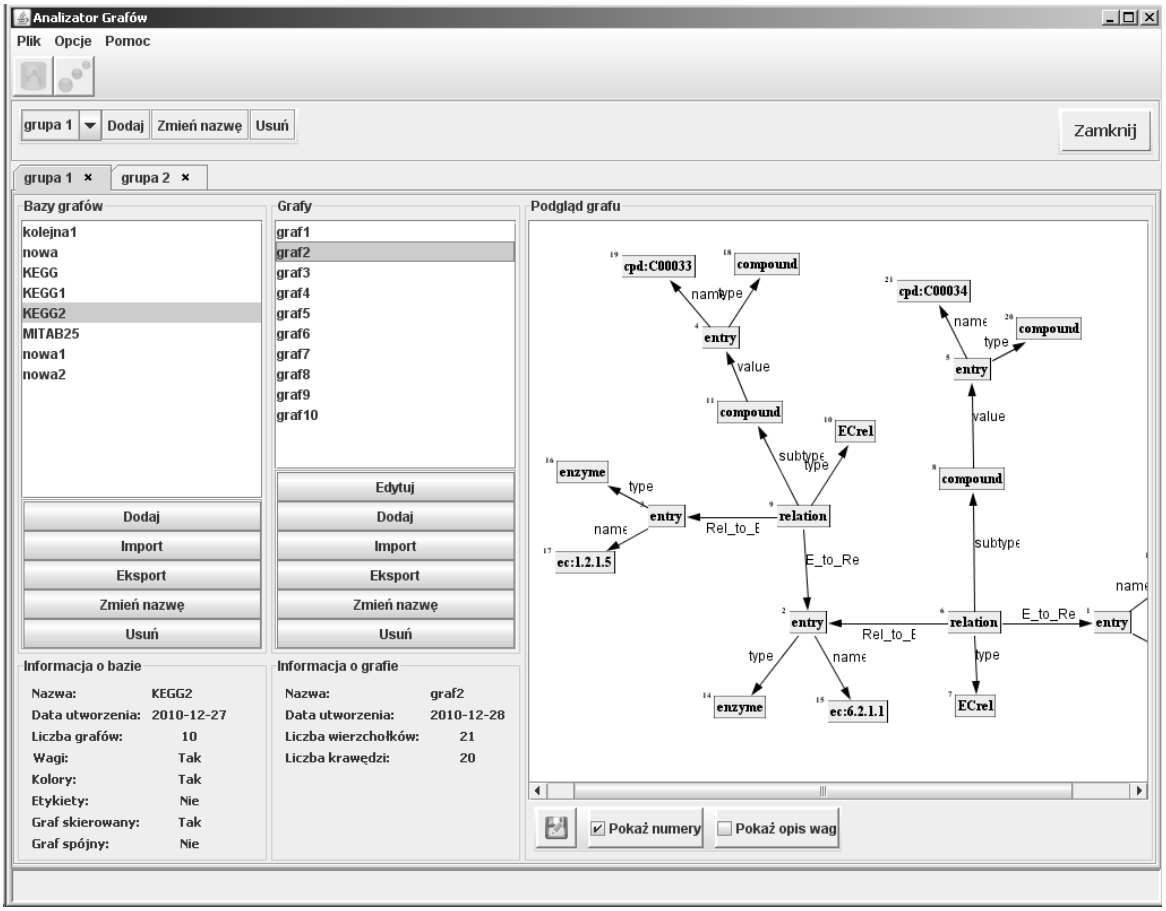
Rys. 3. Diagram pakietów aplikacji
Fig. 3. The package diagram of the aplication

Pakiet *representation* odpowiada za wewnętrzną reprezentację grafów za pomocą wierzchołków oraz krawędzi. Klasa *Graph* wykorzystuje klasę *ModelParametersClass,* która jest częścią pakietu *parameters* i określa typ grafu oraz jego parametry. Pakiet *databases* odnosi się do struktury przechowywania grafów i obejmuje klasy odpowiedzialne za organizację baz grafowych systemu. Na najwyższym poziomie znajduje się klasa *GraphDatabasesClass*, której zadaniem jest odczyt oraz zapis struktury przechowywania grafów z/do plików konfiguracyjnych. Kolejna klasa *GraphDatabasesGroups* jest odpowiedzialna za zbiór wszystkich grup baz grafów i można w niej wyróżnić funkcje umożliwiające dodawanie, usuwanie i modyfikację jej elementów. Podobne funkcje, tylko w odniesieniu do poszczególnych baz grafów, zostały przypisane klasie *GraphDatabsesGroup*. Klasa *GraphDatabase* odpowiedzialna jest za zarządzanie pojedynczą bazą grafów: przechowuje informacje o lokalizacji zbiorów plikowych z poszczególnymi grafami oraz umożliwia ich odczyt oraz zapis. Ostatnią klasą tego pakietu jest *GraphDatabasesFrame*, której zadaniem jest obsługa interfejsu Modułu zarządzania zbiorem grafów. Kolejnym pakietem odpowiedzialnym za import grafów do wewnętrznej struktury grafów jest *importgraph*. Do głównych zadań, które realizują klasy tego pakietu należy odczyt danych ze zbiorów zewnętrznych, przekształcenie ich do postaci grafowej oraz zapis we wskazanej przez użytkownika bazie grafów. W celu zapewnienia komunikacji z innymi aplikacjami do przetwarzania struktur grafowych zaprojektowano klasy tworzące pakiet *exportgraph*, które umożliwiają zapis grafów do popularnych reprezentacji, takich jak: macierze incydencji oraz listy i macierze sąsiedztwa. W miarę potrzeb możliwe jest zaimplementowanie zapisu także do innych formatów. Za pracę edytora grafów odpowiadają klasy pakietu *grapheditor*.

2.3. Interfejs użytkownika

Moduły wizualizacji i edycji grafów oraz zarządzania bazami grafów zostały wyposażone w interfejs, który umożliwia „ręczne” tworzenie i edycję grafów oraz zarządzanie sposobem ich przechowywania na dysku.

W oknie edycji istnieje możliwość zarówno tworzenia nowych grafów (funkcja ta została włączona przede wszystkim dla potrzeb przygotowania prostych przykładowych grafów do celów prezentacyjnych), jak i edycji grafów, zapisanych w bazach systemu. Użytkownik dodaje, zmienia oraz usuwa wierzchołki i krawędzie grafu, przy czym istnieje możliwość edycji parametrów (np. koloru, kształtu, wagi itp.) wielu elementów jednocześnie, po uprzednim ich zaznaczeniu. Jedną z opcji jest eksport grafu do jednego z następujących typów plików graficznych: *eps*, *svg*, *emf*, *pdf* i *png*. Po zakończeniu edycji grafu, dokonane zmiany można zapisać do jednej z wewnętrznych baz grafów lub nadpisać modyfikowany graf.



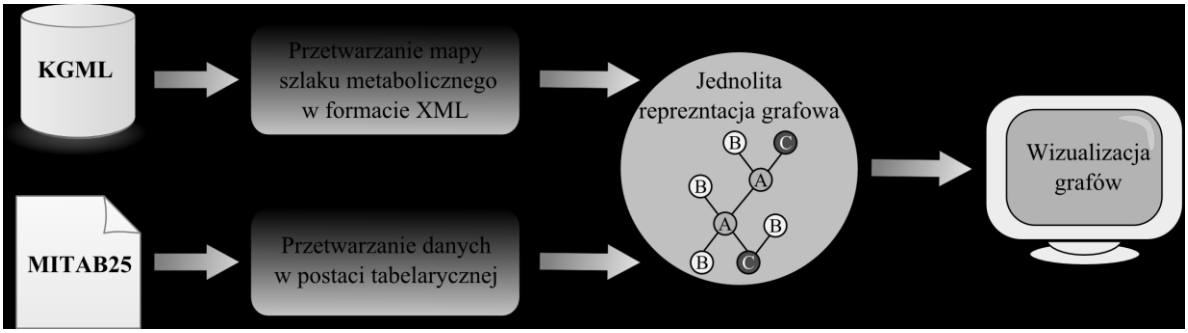
Rys. 4. Interfejs użytkownika: ekran do zarządzania zbiorem grafów
Fig. 4. User interface: graph database management display

Moduł zarządzania zbiorem grafów umożliwia administrowanie wewnętrzną strukturą przechowywania grafów: grupami baz i bazami grafów oraz pojedynczymi grafami (rys. 4). Daje on m.in. możliwość dodawania, zmiany nazwy i usuwania elementów tej hierarchii. Dostęp do poszczególnych grup baz grafów odbywa się przez wskazanie odpowiedniej zakładki lub rozwijaną listę wyboru. Wtedy bazy grafów oraz pojedyncze grafy pojawiają się w odpowiednich oknach.

Jedną z najważniejszych funkcji modułu zarządzania zbiorem grafów jest import z baz zewnętrznych struktur, które są następnie modelowane za pomocą grafów i zapisywane w bazach systemu. Możliwe jest przy tym pobieranie danych z pojedynczych źródeł lub wskazanie kilku źródeł jednocześnie. Oprócz źródła, użytkownik ma możliwość wskazania formatu pobieranego pliku oraz kształtu i układu wierzchołków grafu wynikowego. Podobnie jest z zapisem grafów, który może być wykonany dla pojedynczych grafów, a także dla całych baz. Moduł udostępnia podstawowe informacje o aktualnie używanej bazie grafów oraz przetwarzanym grafie, wyświetlanym w prawej części panelu. Podobnie jak w przypadku opcji edycji, grafy można eksportować do popularnych formatów graficznych zarówno wektorowych, jak i binarnych.

3. Przetwarzanie struktur biologicznych

W tym rozdziale zostaną zaprezentowane dwa przykłady przetwarzania zewnętrznych struktur biologicznych. Są to mapy szlaków metabolicznych w formacie KGML oraz sieci oddziaływań białko-białko w formacie MITAB25. Ogólny schemat przekształceń pokazano na rys. 5.



Rys. 5. Transformacja danych zewnętrznych
Fig. 5. Transformation of external data

Struktury rzeczywiste są przekształcane do postaci jednolitej reprezentacji grafowej i mogą być wyświetlane oraz poddawane edycji w systemie, a także przetwarzane w innych aplikacjach. W kolejnych podrozdziałach przedstawiono więcej szczegółów na temat modelowania sieci biologicznych z wykorzystaniem struktur grafowych oraz wizualizację uzyskanych wyników.

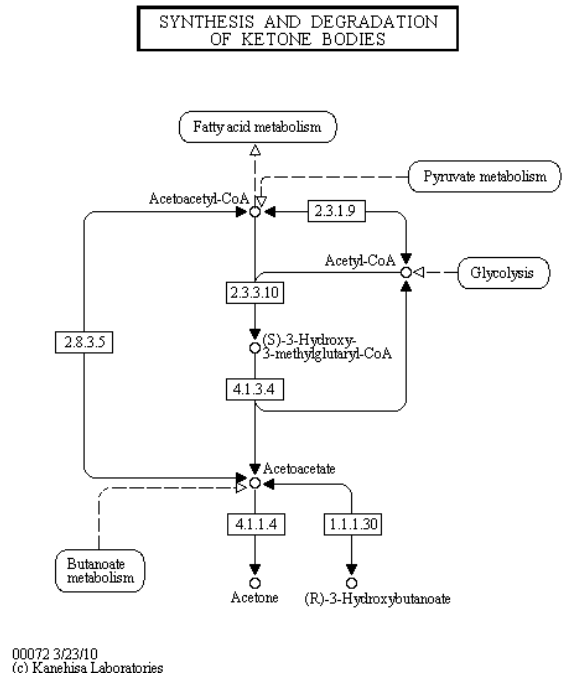
3.1. Transformacja map szlaków metabolicznych

Mapy szlaków metabolicznych z bazy KEGG[1] (*Kyoto Encyclopedia of Genes and Genomes*) [6, 5] są pobierane w opartym na XML formacie KGML. Na rys. 6 pokazano schemat przykładowej mapy szlaków metabolicznych pochodzącej z bazy KEGG. Odpowiadający tej mapie plik XML zawiera 3 główne typy elementów:

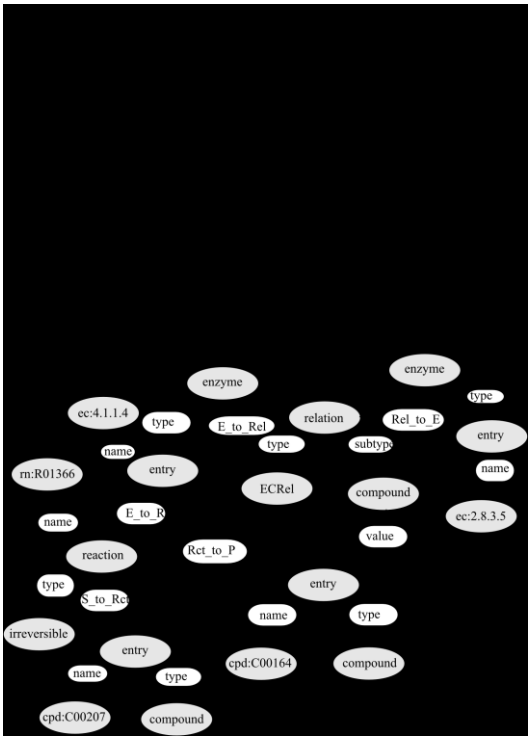
- *entry* – opisuje istniejące węzły grafu (enzymy, związki chemiczne oraz inne mapy);
- *reaction* – opisuje zachodzące reakcje, uwzględniając wszystkie substraty oraz produkty, a także enzymy uczestniczące w reakcji;
- *relation* – opisuje relacje pomiędzy dwoma enzymami a związkiem chemicznym.

Wejściowa mapa szlaków metabolicznych jest najpierw zapisywana w formie pewnego grafu pośredniego, który jest dalej przekształcany przez moduł transformujący do postaci zaproponowanej w [10]. Etykieta każdego węzła grafu jest zmieniana na *entry*, przy czym wprowadzane są dwa dodatkowe węzły, zawierające nazwę (etykietę) oraz typ węzła podstawowego.

[1] http://www.kegg.jp



Rys. 6. Synteza i degradacja związków ketonowych[2]
Fig. 6. Synthesis and degradation of ketone bodies



Rys. 7. Transformacja mapy według [10]
Fig. 7. Map transformation according to [10]

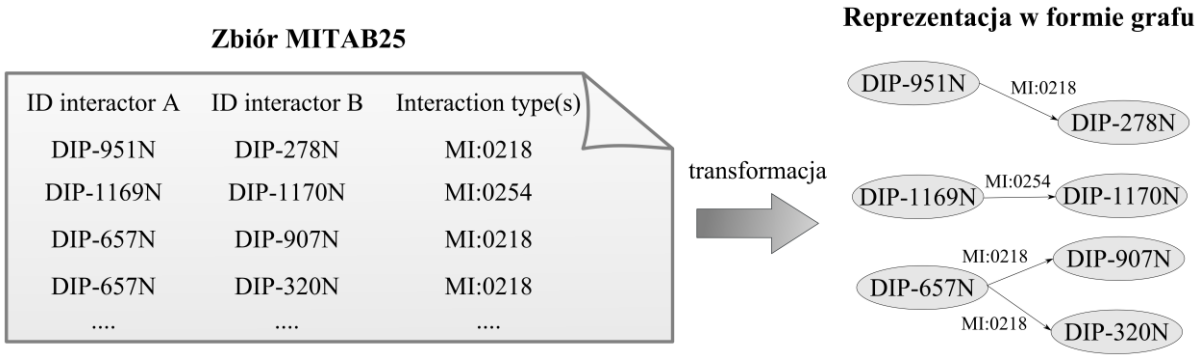
Są one łączone z wierzchołkiem *entry* etykietowanymi krawędziami (odpowiednio *name* oraz *type*). Kolejnym etapem przekształcenia jest przedstawienie połączeń pomiędzy węzłami grafu wejściowego. W grafie wynikowym każdej relacji (*relation*) oraz reakcji (*reaction*) z grafu podstawowego odpowiadają nowe węzły. Podobnie jak w przypadku *entry*, nowy węzeł *reaction* generuje dodatkowe węzły zawierające informację, która się odnosi do nazwy i typu. Natomiast dla węzła odpowiadającego relacji utworzone zostają wierzchołki określające typ oraz podtyp (odpowiednie krawędzie to *type* oraz *subtype*). W dalszej fazie następuje połączenie nowych wierzchołków *reaction* oraz *relation* z odpowiednimi węzłami, które uczestniczą w reakcji bądź relacji, a ostatnim etapem jest rozmieszczenie wierzchołków na palecie. Graf wynikowy ma teraz większą liczbę wierzchołków, które zawierają dodatkowe informacje, odczytane z pliku XML. Fragment mapy wejściowej oraz odpowiadający jej graf wynikowy pokazano na rys. 7.

3.2. Transformacja sieci oddziaływań białko-białko

Baza interakcji białko-białko DIP[3] (*Database of Interacting Proteins*) [9] przechowuje informacje o parach łańcuchów białkowych, które (jak stwierdzono eksperymentalnie) tworzą połączenia między sobą. Baza ta jest podzielona na wiele zbiorów, które zawierają dane doty-

[2] http://www.genome.jp/dbget-bin/www_bget?pathway+map00072
[3] http://dip.doe-mbi.ucla.edu/

czące poszczególnych gatunków i są dostępne w formacie plików MITAB25[4]. Standard MITAB25 jest częścią projektu PSI-MI 2.5 [7] wspieranego przez wiele instytucji zajmujących się oddziaływaniami białko-białko i pozwala na zapis interakcji w plikach tekstowych. Poszczególne wiersze pliku opisują parametry pojedynczego oddziaływania za pomocą 15 pól, zawierających m.in.: unikalne identyfikatory białek wchodzących w skład interakcji, rodzaj interakcji oraz dane dotyczące publikacji, w której to powiązanie zostało opisane. Dla potrzeb generowanej reprezentacji grafowej wykorzystano tylko 3 atrybuty interakcji: identyfikatory pierwszego i drugiego białka oraz rodzaj interakcji (rys. 8).



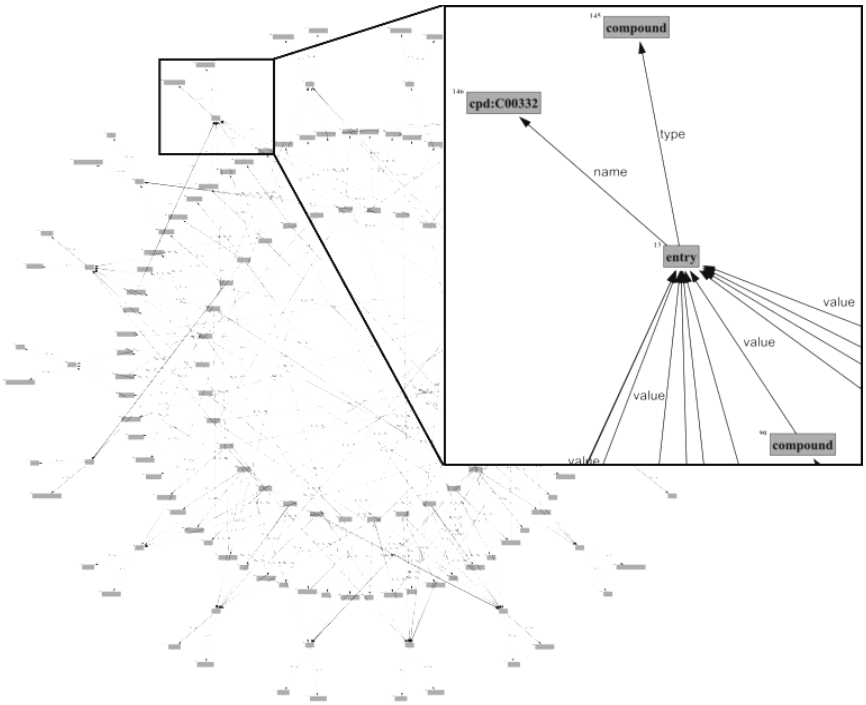
Rys. 8. Transformacja interakcji białko-białko z formatu MITAB25 do postaci grafu
Fig. 8. Transformation of protein-protein interactions from MITAB25 to graph representation

Poszczególne białka wchodzące w skład interakcji są reprezentowane za pomocą wierzchołków grafu, natomiast interakcje pomiędzy nimi za pomocą krawędzi. Typ oddziaływania opisuje etykieta krawędzi.

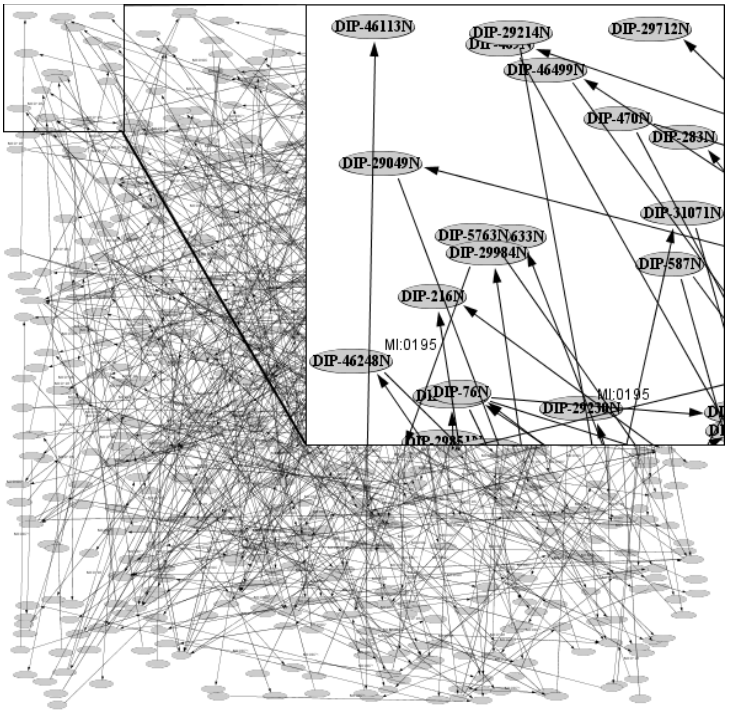
3.3. Wizualizacja

Wynikiem omówionych w poprzednim podrozdziale transformacji sieci biologicznych są struktury grafowe, reprezentujące uzyskane modele. Warto zwrócić uwagę, że grafy odpowiadające danym rzeczywistym to w wielu przypadkach struktury obejmujące tysiące węzłów/krawędzi. Wprawdzie podstawowym zadaniem środowiska BiNArr jest wstępne przetworzenie danych z baz biologicznych tak, aby uzyskać ich jednolitą i spójną postać zapisaną we własnej bazie grafów, niemniej ze względów praktycznych opracowano także moduł do wizualizacji otrzymanych struktur. Funkcja wizualizacji zapewnia zarówno ogólny widok grafu, jak i, w razie potrzeby, pozwala uzyskać bardziej szczegółowy wgląd w jego strukturę, dzięki możliwości powiększania wybranych jego fragmentów. Ma to istotne znaczenie, szczególnie w przypadku rozległych struktur grafowych reprezentujących dane rzeczywiste.

[4] http://code.google.com/p/psicquic/wiki/MITAB25Format



a)



b)

Rys. 9. Grafy: a) mapy szlaków metabolicznych *Synteza i degradacja związków ketonowych*, b) interakcji białko-białko dla gatunku *mysz domowa*

Fig. 9. Graphs: a) representation of metabolic pathway *Synthesis and degradation of ketone bodies*, b) representation of core protein-protein interactions for *house mouse* species

Na rys. 9a i 9b pokazano przykłady wizualizacji grafów uzyskanych w wyniku transformacji odpowiednio: mapy szlaków metabolicznych z bazy KEGG oraz danych opisujących interakcje białko-białko z bazy DIP.

4. Podsumowanie

Artykuł prezentuje aktualny stan projektu, realizowanego przez autorów, zmierzającego do opracowania zintegrowanego środowiska do transformacji oraz wizualizacji danych opisujących sieci biologiczne. Ogólnym celem prowadzonych badań jest opracowanie metod i narzędzi do wstępnego przetwarzania danych strukturalnych, pochodzących z ogólnodostępnych baz biologicznych. W odróżnieniu od niektórych istniejących środowisk, prezentowany system realizuje stosunkowo ograniczoną liczba funkcji i jest wyraźnie ukierunkowany na przygotowanie danych strukturalnych do analizy; np. z wykorzystaniem algorytmów drążenia grafów. Dla struktur biologicznych, pozyskiwanych z oryginalnych zasobów, zaproponowano jednolitą, grafową reprezentację oraz przygotowano moduły do ich wizualizacji i edycji. Przewidziano także możliwość eksportu grafów w kilku formatach do wykorzystania przez aplikacje zewnętrzne. W ramach dotychczasowych prac wykonano testy weryfikujące przydatność aplikacji dla rzeczywistych danych opisujących sieci biologiczne. Ich wyniki wskazują na możliwości praktycznego wykorzystania w badaniach nad algorytmami drążenia grafów lub w innych zastosowaniach.

W ramach przyszłych prac nad systemem planuje się jego rozszerzenie m.in. o możliwość pracy z danymi, w których do opisu sieci biologicznych zastosowano inne formaty. Ponadto, wydaje się, że system, stosunkowo niewielkim nakładem sił, mógłby być przystosowany do przetwarzania danych strukturalnych z innych dziedzin.

BIBLIOGRAFIA

1. Cline M. S. et al.: Integration of biological networks and gene expression data using Cytoscape. Nature Protocols, t. 2, No. 10, 2007, s. 2366÷2382.
2. Cook D. J., Holder L. B..: Mining Graph Data. John Wiley and Sons Inc., 2007.
3. De Las Rivas J., Fontanilio C.: Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. PLoS Computational Biology, t. 6, No. 6, e1000807, 2010.

4. Iragne F. et al.: ProViz: Protein Interaction Visualization and Exploration. Bioinformatics, Bioinformatics Advance Access published September 3, 2004, available via: bioinformatics.oxfordjournals.org.
5. Kanehisa M. et al.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research, t. 34, Database-Issue, 2006, s. 354÷357.
6. Kanehisa M., Goto S.: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, t. 28, No. 1, 2000, s. 27÷30.
7. Kerrien S. et al.: Broadening the horizon-level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biology, z. 5, 44, 2007.
8. Lacroix V., Cottret L., Thébault P., Sagot M.-F.: An Introduction to Metabolic Networks and Their Structural Analysis. IEEE/ACM Trans. Comput. Biol. Bioinformatics, t. 5, No. 4, 2008, s. 594÷617.
9. Xenarios I., Rice D. W., Salwinski L., Baron M. K., Marcotte E. M., Eisenberg D.: DIP: the Database of Interacting Proteins. Nucleic Acids Res, z. 28(1), 2000, s. 289÷291.
10. You C., Holder L., Cook D.: Substructure Analysis of Metabolic Pathways by Graph-based Relational Learning. Biomedical Data and Applications, t. 224, s. 237÷261.

Recenzenci: Dr inż. Bożena Małysiak-Mrozek
Dr inż. Dariusz Mrozek

Wpłynęło do Redakcji 26 stycznia 2011 r.

Abstract

An important feature of modern information systems is the continuously increasing rate of data overcoming the commonly used *attribute-value* or *transaction* representations. Such data is commonly referred to as (semi-)structured because of the highly complex and irregular structure, which, in fact, contains a considerable part of its semantics (see Fig.1). The structured data are usually modeled as sequences, trees and graphs, and commonly used in chemistry, bioinformatics, pattern recognition, Web usage, XML documents analysis, software engineering and social processes.

The investigation of biological networks for their better understanding and making available for practical use is currently the important task in *systems biology*. The authors developed an integrated environment BiNArr (*Biological Network Arranger*) aimed to perform some data preparation operations as well as visualization of the network data stored in biolog-

ical databases. Dissimilar to the existing tools like Cytoscape [1] the functionality of our application is rather limited and strictly oriented for transforming structured data from real databases into graphs. This allows to perform further processing, e.g. with use of graph mining algorithms. The general schema of BiNArr is depicted in Fig.2. We proposed the unified graph representation for the structures extracted from original resources and developed the modules for their visualization and edition. Another important feature is the automatic coding of the resulting graphs in several formats required by different applications. In order to present some capabilities of BiNArr we used the biological structures representing metabolic pathways extracted from KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [6,5] as well as protein-protein interactions provided in DIP (*Database of Interacting Proteins*) [9]. The Figure 9a shows the resulted complex graphs representing models of the input biological structures produced and visualized by BiNArr.

Adresy

Krzysztof ŚWIDER: Politechnika Rzeszowska, Wydział Elektrotechniki i Informatyki, ul. W. Pola 2, 35-959 Rzeszów, Polska, kswider@prz-rzeszow.pl.
Bartosz JĘDRZEJEC: Politechnika Rzeszowska, Wydział Elektrotechniki i Informatyki, ul. W. Pola 2, 35-959 Rzeszów, Polska, bartoszj@prz-rzeszow.pl.
Damian JAKOBUS: Politechnika Rzeszowska, Wydział Elektrotechniki i Informatyki, ul. W. Pola 2, 35-959 Rzeszów, Polska, kswider@prz-rzeszow.pl.