STUDIA INFORMATICA 2008
Volume 29 Number 4A (80)

Andrzej MASTEJ, Dariusz MROZEK
Politechnika Śląska, Instytut Informatyki

PROTEIN MOLECULAR VIEWER – WIZUALZIACJA STRUKTUR PRZESTRZENNYCH BIAŁEK ZAPISANYCH W FORMACIE PDBML

Streszczenie. Wizualizacja struktur przestrzennych białek ma ogromne znaczenie dla analizy funkcji i aktywności tych złożonych cząstek biologicznych. Narzędzia wizualizacji umożliwiają bowiem graficzne przedstawienie złożonej, wewnętrznej konstrukcji białek, zbudowanych z setek lub tysięcy atomów, połączonych wzajemnie wiązaniami kowalencyjnymi. W niniejszym artykule przedstawiono najważniejsze możliwości autorskiego programu Protein Molecular Viewer (PMV), który pozwala na prezentację struktur białkowych pobieranych z popularnej bazy danych Protein Data Bank. Możliwość wizualizacji struktur zapisanych w formacie PDBML czyni program PMV jednym z nielicznych na świecie, które posiadają taką funkcję.

Słowa kluczowe: bioinformatyka, białka, struktura przestrzenna, wizualizacja

PROTEIN MOLECULAR VIEWER – THE VISUALIZATION OF PROTEIN MOLECULAR STRUCTURES STORED IN THE PDBML FORMAT

Summary. Visualization of protein molecular structures is a very important part of the analysis of protein function and activity. Molecular viewers allow to represent graphically the complex construction of proteins and give us an idea of the biological molecules built up with hundreds or thousands of atoms linked to each other by covalent bonds. In the chapter we present the most interesting features of our Protein Molecular Viewer (PMV). PMV is a molecular visualization tool developed at the Silesian University of Technology in Gliwice, which is used to show protein structures loaded from the well-known Protein Data Bank. With the possibility of loading and presenting protein structures from the PDBML data format the PMV becomes one of a few tools in the world having this unique function.

Keywords: bioinformatics, proteins, spatial structure, visualization, molecular viewer

1. Wprowadzenie

Aplikacje wizualizacji struktur przestrzennych białek i innych związków biologicznych należą do szerokiej grupy narzędzi analizy molekularnej używanych w biochemii, proteomice i biologii systemów. Funkcjonowanie organizmów żywych w ujęciu biologicznym jest bowiem ściśle związane z istnieniem i aktywnością białek. Białka są niezwykle ważnymi cząsteczkami, pełniącymi kluczową rolę we wszystkich reakcjach biochemicznych zachodzących w komórkach organizmów. Pośredniczą one w wielu ważnych funkcjach komórkowych, takich jak: transport i magazynowanie, oddychanie komórkowe, tworzenie podtrzymujących struktur mechanicznych, ochrona immunologiczna, wrażliwość na bodźce, kontrola wzrostu i różnicowania i in. [1, 2].

Biorąc pod uwagę ogólną budowę białek, można powiedzieć, że są to makrocząsteczki, o masie cząsteczkowej powyżej 10 kDa (1 Da = $1{,}66\times10^{-24}$g) zbudowane z aminokwasów (>100 aminokwasów) połączonych w łańcuchy wiązaniami peptydowymi [3]. W budowie białek wyróżnia się cztery poziomy opisu lub reprezentacji: strukturę pierwszo-, drugo-, trzecio- i czwartorzędową. Trzy ostatnie poziomy definiują tzw. konformację białka lub jego strukturę przestrzenną [3]. Analizę biochemiczną białek prowadzi się zazwyczaj na jednym z wybranych poziomów.

Strukturę pierwszorzędową białek określa tzw. sekwencja białka, czyli kolejność aminokwasów w łańcuchu białkowym [4], stąd też często zamiast pojęcia struktura pierwszorzędowa używa się pojęcia sekwencja. Przykład sekwencji białek mioglobiny i hemoglobiny przedstawiono na rys. 1. Poszczególne aminokwasy w liniowym łańcuchu są reprezentowane przez litery alfabetu.

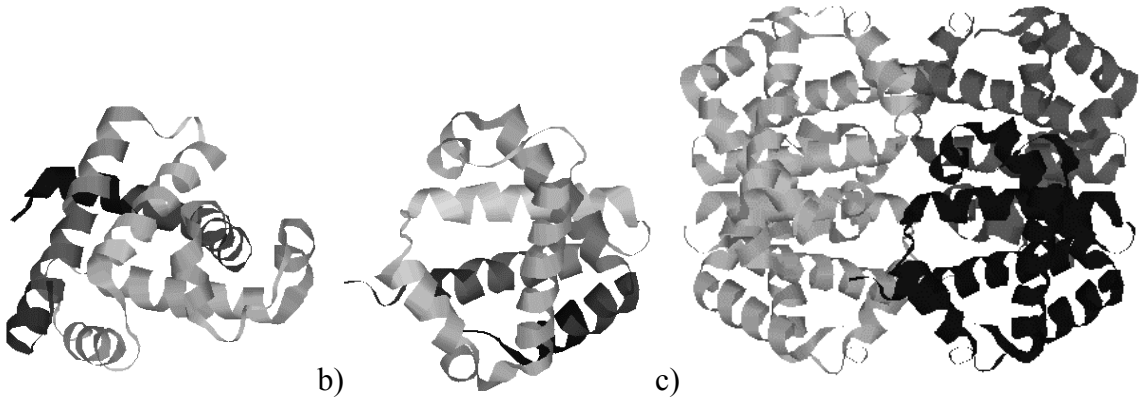
```
>1MBN:A|PDBID|CHAIN|SEQUENCE
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILK
KKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG

>4HHB:A|PDBID|CHAIN|SEQUENCE
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMP
NALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

Rys. 1. Sekwencje białek: mioglobiny (PDB ID: 1MBN) i hemoglobiny (PDB ID: 4HHB, łańcuch A) w formacie FASTA

Fig. 1. Amino acid sequences of myoglobin (PDB ID: 1MBN) and hemoglobin (PDB ID: 4HHB, chain A) in the FASTA format

Struktura drugorzędowa opisuje wzajemne przestrzenne ułożenie reszt aminokwasowych, sąsiadujących ze sobą w sekwencji liniowej. Ten poziom opisu wyróżnia w strukturze przestrzennej pewne charakterystyczne, zwykle regularnie pofałdowane regiony. Przykładem struktury drugorzędowej jest *helisa* α i *harmonijka* β [3, 4]. **Struktura trzeciorzędowa** odnosi się do powiązań przestrzennych i wzajemnego ułożenia reszt aminokwasowych zarówno tych oddalonych od siebie w sekwencji liniowej, jak i sąsiadujących ze sobą (rys. 2a i 2b) [4].

Opisuje zatem ukształtowanie struktury spowodowane dodatkowymi wewnętrznymi oddziaływaniami elektrostatycznymi, wodorowymi oraz ewentualnymi kowalencyjnymi mostkami dwusiarczkowymi[1]. Ten poziom opisu oddaje biologicznie aktywną przestrzenną konformację białka [3]. **Strukturę czwartorzędową** wyróżniamy w białkach składających się z więcej niż jednego łańcucha polipeptydowego (rys. 2c). Struktura ta opisuje wzajemne ułożenie podjednostek i rodzaj ich kontaktu, który może mieć charakter kowalencyjny lub niekowalencyjny [4].



Rys. 2. Struktury: a) trzeciorzędowa mioglobiny (PDB ID: 1MBN), b) trzeciorzędowa hemoglobiny (PDB ID: 4HHB, tylko łańcuch A), c) czwartorzędowa hemoglobiny (PDB ID: 4HHB, wszystkie łańcuchy), w każdym przypadku widoczne są spiralne helisy α będące elementami struktury drugorzędowej

Fig. 2. Structures: a) tertiary of myoglobin (PDB ID: 1MBN),b) tertiary of hemoglobin (PDB ID: 4HHB, chain A), quaternary of hemoglobin (PDB ID: 4HHB, all chains), in all cases the secondary structures of α helices are visible

Analiza struktury przestrzennej białka jest niezwykle istotna z punktu widzenia funkcji białka, jego aktywności i reakcji, w które wchodzi dane białko. Tego typu analiza, poparta zazwyczaj obserwacjami struktury, obejmuje nie tylko samą sekwencję aminokwasów, ale również cechy geometryczne badanej cząsteczki. Nie ulega wątpliwości, że struktury nawet małych białek są bardzo skomplikowane – białka zbudowane są z setek aminokwasów, a zatem tysięcy atomów różnych pierwiastków chemicznych. Aby analizować tak złożone struktury, bardzo przydatne okazują się narzędzia, które umożliwiają oglądanie i badanie najdrobniejszych szczegółów budowy przestrzennej [5]. Powszechnym celem programów wizualizujących struktury jest prezentacja ogólnej struktury białek, ogólnego kształtu, a także umożliwienie uproszczenia struktury poprzez wyodrębnienie struktur drugorzędowych. Dzięki temu użytkownik ma możliwość przestudiowania kształtu białka lub jego wybranych fragmentów i porównania go z innym białkiem, a przez to subiektywnej oceny podobieństwa i różnic. Od wczesnych lat osiemdziesiątych naukowcy korzystali z coraz szerszej wiedzy na temat struk-

[1] Mostki dwusiarczkowe są wiązaniami kowalencyjnymi wytworzonymi przez grupy sulfhydrylowe (-SH), pełnią funkcję stabilizacyjną dla konstrukcji białka [3].

tury i funkcji białek do przebudowywania istniejących białek, a od niedawna też do projektowania całkowicie nowych białek. Z tego powodu programy wizualizacji struktury znajdują zastosowanie w inżynierii leków – pomagają m.in. w tworzeniu skutecznych leków. Osiągnięcia współczesnej farmacji są imponujące, ale często nie można zbudować leku, dopóki nie zostanie poznana struktura chorobotwórczego białka. Programy do wizualizacji mogą wspomagać projektowanie inhibitorów – związków hamujących nadmierną aktywność niektórych białek, w szczególności enzymów [6]. Prężną gałęzią współczesnej biochemii, biologii molekularnej i biotechnologii jest również przewidywanie struktury białek, które otwiera wiele możliwości praktycznych zastosowań w medycynie i przemyśle. Narzędzia wizualizacji struktury mają także istotne znaczenie w patologii molekularnej, w której analizuje się wpływ drobnych mutacji w budowie białka na jego aktywność biologiczną [7, 8].

W badaniach prowadzonych bezpośrednio przez autorów niniejszego artykułu, narzędzia wizualizujące struktury białkowe pozwalają na weryfikację wyników procesu poszukiwania podobieństwa strukturalnego białek. W procesach poszukiwania podobieństwa strukturalnego, dla zadanej struktury białkowej poszukuje się struktur o identycznej lub zbliżonej budowie przestrzennej. Dzięki narzędziom poszukiwania podobieństwa strukturalnego możliwa jest identyfikacja funkcji nowo powstałych lub odkrytych białek na podstawie ich struktur przestrzennych, przez ich porównanie z innymi, znanymi białkami. Możliwe jest także poszukiwanie białek posiadających pewne charakterystyczne, ważne biologicznie regiony struktury przestrzennej. Autorzy posiadają w tym obszarze własne osiągnięcia w postaci opracowanego algorytmu EAST [9, 10]. Narzędzia wizualizacji stanowią wówczas ostatnie stadium kontroli poprawności rezultatów wyszukiwania.

Wizualizacja struktur przestrzennych białek odbywa się zwykle na podstawie danych opisujących struktury, w postaci współrzędnych atomów (x, y, z), przechowywanych w bazach danych. Dane dotyczące struktury przestrzennej określonego białka otrzymuje się dzięki krystalografii rentgenowskiej (rentgenografii strukturalnej) lub obserwacji z wykorzystaniem nuklearnego rezonansu magnetycznego (spektroskopii NMR). Do najbardziej znanych baz struktur należą Protein Data Bank (PDB) [11] i NCBI Molecular Modeling DataBase (MMDB) [12]. Bazy te umożliwiają wymianę swoich danych w określonych formatach, a co za tym idzie, wizualizację struktur molekularnych na komputerach użytkowników.

W niniejszym artykule przedstawiono program *Protein Molecular Viewer (PMV)* do wizualizacji struktur molekularnych białek zapisanych w postaci zbiorów w formacie PDBML [13]. Program PMV jest autorskim rozwiązaniem w obszarze wizualizacji struktur przestrzennych i powstał z myślą o weryfikacji wyników działania algorytmów poszukiwania podobieństwa białek. Program ten posiada kilka unikalnych cech odróżniających go od dostępnych na świecie rozwiązań i jest na co dzień używany w badaniach nad strukturą

białek, prowadzonych przez autorów. Możliwości programu *Protein Molecular Viewer* zostały przedstawione w podrozdziale 4. Tytułem wprowadzenia w podrozdziale 2 krótko scharakteryzowano powszechne formaty wymiany danych strukturalnych białek z bazy PDB. Popularne narzędzia wizualizacji białek zostały natomiast opisane w podrozdziale 3.

2. Formaty danych makromolekularnych bazy PDB

Struktury trójwymiarowe makrocząsteczek, które określono technikami NMR lub krystalografii rentgenowskiej są archiwizowane w postaci zbiorów danych. Struktury przestrzenne białek zapisywane są zwykle w plikach tekstowych w zdefiniowanych formatach. Istnieje wiele formatów opisu struktur makrocząsteczek. W niniejszym artykule skoncentrowano się tylko na tych, które są bezpośrednio związane z bazami danych przechowującymi dane molekularne. Najbardziej znanymi i popularnymi formatami opisującymi struktury białek są: PDB [14], mmCIF [15] i PDBML [13]. Zbiory w tych formatach zawierają m.in. ogólny opis cząsteczki, informacje o naukowcach, którzy odczytali/odkryli strukturę cząsteczki, szczegóły i parametry badania, sekwencję aminokwasową białka. Najważniejszą częścią tych zbiorów są jednak dane precyzyjnie lokujące każdy z atomów cząsteczki w przestrzeni trójwymiarowej.

2.1. Format PDB

Jednym ze sposobów przechowywania danych eksperymentalnych o strukturach przestrzennych białek jest format PDB [14]. Jest to format tekstowy oparty na zapisie kolumnowym. Każdy zbiór PDB jest złożony z tzw. rekordów, na które składa się jeden bądź więcej wierszy, zawierających po 80 kolumn. Każdy rekord jest opisany za pomocą identyfikatora (6 pierwszych znaków każdego wiersza) i jest podzielony na pola. Rekordy mogą przechowywać różne informacje, dotyczyć różnych aspektów struktury przestrzennej. Zbiór PDB zawiera zwykle kilka różnych typów rekordów uporządkowanych we właściwy sposób, by opisać strukturę molekuły biologicznej. Na przykład, współrzędne prostokątne (x, y, z) atomów tworzących strukturę przestrzenną molekuły są zapisane w rekordach ATOM oraz HETATM wraz z danymi o: numerze atomu, oznaczeniu pierwiastka chemicznego, nazwie aminokwasu (w kodzie 3-literowym), oznaczeniu łańcucha, numerze aminokwasu w sekwencji oraz dodatkowymi informacjami (rys. 3). Rekordy HETATM gromadzą dane o współrzędnych atomów tzw. grup niestandardowych, w stosunku do aminokwasów i nukleotydów. Do grup niestandardowych można zaliczyć np. cząstki wody.

Szczegółowa dokumentacja formatu PDB znajduje się na stronie internetowej bazy Protein Data Bank [16].

```
ATOM   1058  N   ARG A 141      -6.399  12.034 -10.391  7.00 17.59      4HHB1262
ATOM   1059  CA  ARG A 141      -8.000  12.137 -10.191  6.00 24.58      4HHB1263
ATOM   1060  C   ARG A 141      -8.327  13.610  -9.572  6.00 44.44      4HHB1264
ATOM   1061  O   ARG A 141      -7.492  14.016  -8.882  8.00 21.81      4HHB1265
ATOM   1062  CB  ARG A 141      -8.478  10.914  -9.869  6.00 33.40      4HHB1266
ATOM   1063  CG  ARG A 141      -8.068  10.650  -8.145  6.00 17.28      4HHB1267
ATOM   1064  CD  ARG A 141      -9.053   9.446  -7.867  6.00 14.66      4HHB1268
ATOM   1065  NE  ARG A 141      -8.372   9.269  -6.610  7.00 22.73      4HHB1269
ATOM   1066  CZ  ARG A 141      -9.233   8.420  -5.781  6.00 26.88      4HHB1270
ATOM   1067  NH1 ARG A 141     -10.147   7.455  -6.079  7.00 23.24      4HHB1271
ATOM   1068  NH2 ARG A 141      -8.672   8.328  -4.506  7.00 33.34      4HHB1272
ATOM   1069  OXT ARG A 141      -9.474  13.682  -9.742  8.00 31.52      4HHB1273
TER    1070      ARG A 141                                              4HHB1274
...
...
HETATM 4561  P   PO4     1      -6.147 -21.111  -3.332 47.00 31.17   1  4HHB4765
HETATM 4562  P   PO4     2       5.931 -21.573   3.319 47.00 32.97   1  4HHB4766
HETATM 4563  O   HOH     3        .061  -2.494 -16.397  8.00 18.64      4HHB4767
HETATM 4564  O   HOH     4      -2.661  -3.608   9.261  8.00 21.26      4HHB4768
HETATM 4565  O   HOH     5        .111  13.200  11.373  8.00 21.54      4HHB4769
HETATM 4566  O   HOH     6      27.864    .667    .584  8.00 21.15      4HHB4770
HETATM 4567  O   HOH     7      -8.041 -19.581 -16.153  8.00 26.28      4HHB4771
HETATM 4568  O   HOH     8        .093  -2.470  16.222  8.00 21.64      4HHB4772
HETATM 4569  O   HOH     9      10.459   5.072 -14.201  8.00 29.96      4HHB4773
HETATM 4570  O   HOH    10     -16.704  12.691  -9.201  8.00 23.30      4HHB4774
HETATM 4571  O   HOH    11       2.825  -3.769  -9.602  8.00 23.08      4HHB4775
```

Rys. 3. Fragment pliku w formacie PDB opisującego hemoglobinę (PDB ID: 4HHB), na którym zaprezentowano sekcję danych dotyczącą położenia atomów tworzących strukturę przestrzenną molekuły

Fig. 3. A part of the file in the PDB format describing hemoglobin (PDB ID: 4HHB). The presented section describes location of atoms in the spatial structure of the whole molecule

2.2. Format mmCIF

Format mmCIF (ang. *MacroMolecular Chemical Interchange Format*) [14, 15] jest to nowszy niż PDB format opisu białek, których struktury można pobrać z bazy Protein Data Bank. Podstawą organizacji danych w rekordzie mmCIF (rys. 4) jest zbiór tablic odnośnikowych – przed każdą grupą danych opisujących strukturę, np. przed danymi opisującymi współrzędne poszczególnych atomów struktury umieszczone są nazwy odnośnikowe, określające znaczenie kolejnych pozycji. Znaczenie tych odnośników możemy sprawdzić w słowniku mmCIF [15]. Same dane zapisane są natomiast podobnie jak w formacie PDB w kolejnych wierszach, w których we właściwych polach znajdują się odpowiednie informacje opisowe. Format mmCIF wprowadza język opisu danych DDL (ang. *Data Description Language*), co umożliwia łatwe przeprowadzenie rozkładu struktury pliku przez komputer. Umożliwia to również kontrolę poprawności danych opisujących struktury białkowe oraz ułatwia zapisywanie danych w relacyjnej bazie danych [17].

```
loop_
_atom_site.id
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.auth_seq_id
_atom_site.cartn_x
_atom_site.cartn_y
_atom_site.cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.label_entity_id
_atom_site.label_seq_id
  1    O5*   G A   1       7.231  -2.196  -5.399  1.00 22.25   1 1
  2    C5*   G A   1       6.950  -3.464  -4.723  1.00 15.86   1 1
  3    C4*   G A   1       8.299  -4.018  -4.302  1.00 15.20   1 1
  4    O4*   G A   1       9.257  -3.779  -5.318  1.00 10.85   1 1
```

Rys. 4. Fragment pliku w formacie mmCIF, na którym zaprezentowano przykładowy opis kilku atomów w strukturze białka

Fig. 4. A part of a file in the mmCIF format showing a description of several atoms in a protein structure

2.3. Format PDBML

Format PDBML (ang. *Protein Data Bank Markup Language*) [13] to najnowszy format opisu struktur białek za pomocą języka XML. Format XML jest użytecznym sposobem zapisu i wymiany danych, nadając zbiorom danych pewną strukturę. Większość plików XML zawiera dane, które nie są przeznaczone do bezpośredniej prezentacji użytkownikowi. Właściwe dane muszą zostać wcześniej wyekstrahowane z tzw. znaczników XML (ang. *XML tags*). Wcześniejsze opracowanie słownika mmCIF znacznie ułatwiło utworzenie formatu XML oraz związanego z nim słownika opisu XML Schema dla makromolekularnej struktury białek. Użycie słownika XML Schema usprawnia kontrolę poprawności danych zapisanych w formacie PDBML. Zaletą języka XML jest to, iż jest on bardzo popularnym formatem wymiany danych między aplikacjami, a współczesne relacyjne bazy danych zapewniają wsparcie dla zbiorów XML. Najnowszy format danych strukturalnych PDBML posiada wiele zalet, jednak nie jest pozbawiony wad. Użycie znaczników XML prowadzi do zwiększenia rozmiarów plików w formacie PDBML. W porównaniu z formatem mmCIF objętość tego samego zbioru w formacie PDBML może być blisko dziesięciokrotnie większa [17]. Fragment zbioru w formacie PDBML zaprezentowano na rys. 5. Fragment ten zawiera opis tylko jednego, wybranego atomu w strukturze cząsteczki. Porównując ten sposób zapisu i wymiany informacji do wcześniej omawianych formatów PDB i mmCIF, można zaobserwować duży narzut informacji opisowej w stosunku do samych danych.

```
<PDBx:atom_siteCategory>
  <PDBx:atom_site id="1">
    <PDBx:group_PDB>ATOM</PDBx:group_PDB>
    <PDBx:type_symbol>N</PDBx:type_symbol>
    <PDBx:label_atom_id>N</PDBx:label_atom_id>
    <PDBx:label_alt_id xsi:nil="true" />
    <PDBx:label_comp_id>ILE</PDBx:label_comp_id>
    <PDBx:label_asym_id>A</PDBx:label_asym_id>
    <PDBx:label_entity_id>1</PDBx:label_entity_id>
    <PDBx:label_seq_id>1</PDBx:label_seq_id>
    <PDBx:Cartn_x>17.399</PDBx:Cartn_x>
    <PDBx:Cartn_y>48.509</PDBx:Cartn_y>
    <PDBx:Cartn_z>17.807</PDBx:Cartn_z>
    <PDBx:occupancy>7.00</PDBx:occupancy>
    <PDBx:B_iso_or_equiv>19.92</PDBx:B_iso_or_equiv>
    <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
    <PDBx:auth_comp_id>ILE</PDBx:auth_comp_id>
    <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
    <PDBx:auth_atom_id>N</PDBx:auth_atom_id>
    <PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_model_num>
  </PDBx:atom_site>
</PDBx:atom_siteCategory>
```

Rys. 5. Fragment pliku w formacie PDBML, na którym zaprezentowano opis tylko jednego atomu cząsteczki

Fig. 5. A part of a file in the PDBML format showing a description of one atom in a protein structure

3. Dostępne aplikacje wizualizacji białek

W niniejszym rozdziale zostaną przedstawione i krótko omówione wybrane, istniejące popularne programy do wizualizacji struktur białkowych oraz ich najważniejsze możliwości.

3.1. Program RasMol

RasMol [18] jest najpopularniejszym programem do wizualizacji struktur molekularnych oraz związków chemicznych. Narzędzie to jest bardzo rozbudowane, udostępnia wiele form wizualizacji struktur białkowych. Uzyskane struktury można powiększać i obracać, a wybrane atomy wyróżniać kolorami. W programie RasMol istnieje możliwość eksportu uzyskanych obrazów do wielu znanych formatów graficznych. Program został napisany w języku C, dzięki czemu jest on bardzo wydajny. Umożliwia on jednak odczyt zbiorów danych strukturalnych białek zapisanych tylko w formatach PDB i mmCIF.

3.2. Program Jmol

Jmol [19] jest narzędziem napisanym w języku Java, które wizualizuje struktury białek i innych związków chemicznych zapisanych w formacie PDB. Funkcjonalnością program

Jmol dorównuje programowi RasMol. Silnik graficzny, z którego korzysta Jmol, został napisany całkowicie w języku Java (nie korzysta z bibliotek Java3D, OpenGL lub jakiejkolwiek innej biblioteki wspierającej akcelerację sprzętową) i został zoptymalizowany pod kątem wizualizacji cząsteczek. Wydajność tego programu jest zaskakująco dobra, pomimo że został on całkowicie napisany w języku Java. Jego wadą jest jednak brak możliwości wczytywania struktur zapisanych w formacie PDBML. Program jest dostępny jako aplikacja Java oraz jako aplet Java.

3.3. Program jV

jV [20] to program do wizualizacji struktur 3D białek napisany w języku Java z wykorzystaniem biblioteki JOGL (ang. *Java bindings for OpenGL*), wspierającej akcelerację sprzętową [21]. Narzędzie to, jako jedno z niewielu, umożliwia wczytywanie struktur z najnowszego formatu opisującego białka, tj. PDBML. Program zapewnia bardzo szybką wizualizację. Funkcjonalność jest zbliżona do programu RasMol. Program jest dostępny w Internecie w postaci aplikacji Java oraz w formie apletu Java.

3.4. Program Cn3D

Cn3D [22] jest to bardzo wydajny program do wizualizacji białek napisany w jezyku C++. Program wykorzystuje technologię OpenGL. Jakość wyświetlanej grafiki jest bardzo wysoka przy zachowaniu dużej płynności wyświetlania. Program umożliwia wyświetlanie białek umieszczonych w bazie MMDB (ang. *Molecular Modeling DataBase*). Baza MMDB zawiera jednak zbiory zapisane we własnym formacie ASN.1 [23] i takie zbiory wizualizuje program Cn3D. Bardzo wygodnym rozwiązaniem jest jednoczesne wyświetlanie struktury molekularnej białka na ekranie oraz wyświetlanie w drugim oknie jego sekwencji aminokwasowej. Dzięki temu użytkownik może bardzo łatwo zaznaczyć wybraną grupę aminokwasów, która od razu jest wyróżniana w strukturze przestrzennej białka.

3.5. Program BioClipse

Bioclipse [24] jest to ciekawe narzędzie do wizualizacji struktur przestrzennych cząstek biologicznych dla chemików i bioinformatyków napisane w języku Java na podstawie Eclipse Rich Client Platform (RCP) [25]. Bioclipse bazuje na interfejsie oraz funkcjonalności środowiska Eclipse. Funkcjonalność narzędzia może być rozszerzana przez zainstalowanie dostępnych wtyczek (ang. *plugins*). Za wizualizację 3D cząsteczek odpowiada wtyczka Jmol-plugin. Jednak funkcjonalność tej wtyczki została znacząco ograniczona w porównaniu do aplikacji Jmol przedstawionej wcześniej. Program umożliwia

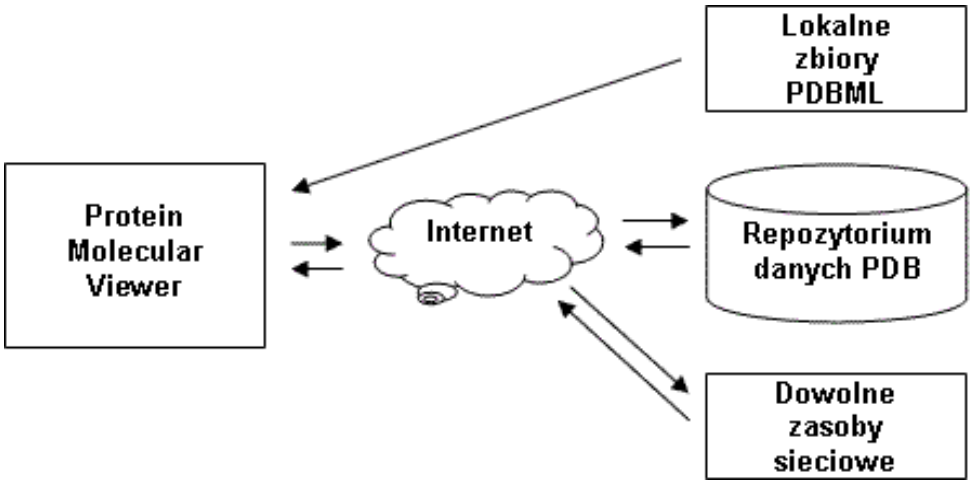
wizualizację różnych związków chemicznych, a także ich wzorów strukturalnych. Możliwa jest edycja plików opisujących cząsteczki oraz graficzna edycja struktur 2D związków.

4. Program Protein Molecular Viewer

Protein Molecular Viewer (PMV) jest autorskim, udostępnionym nieodpłatnie, programem do wizualizacji struktur przestrzennych białek. Program został napisany w języku Java. Dzięki temu program PMV może pracować jako aplikacja lub jako aplet na stronie www. PMV potrafi odczytywać dane molekularne pochodzące z bazy PDB zapisane wyłącznie w formacie PDBML. Starsze formaty, takie jak PDB i mmCIF, nie są obsługiwane. Dzieje się tak, ponieważ PMV jest skoncentrowany na technologie sieciowe i popularny format XML. Natomiast możliwość odczytu danych zapisanych w formacie PDBML daje programowi PMV przewagę nad innymi programami do wizualizacji struktur, omówionymi w poprzednim rozdziale. Taką możliwość posiadał tylko program jV. Do wizualizacji struktur przestrzennych molekuł biologicznych używana jest biblioteka Java3D, która musi być zainstalowana przed uruchomieniem programu PMV. Program PMV posiada dość specyficzną architekturę pracy, która zostanie przedstawiona w podrozdziale 4.1. W kolejnych podrozdziałach zostaną przedstawione natomiast najważniejsze możliwości programu PMV.

4.1. Architektura wymiany danych podczas wizualizacji

Program PMV pracuje ze zbiorami danych, które mogą mieć charakter lokalny lub być rozproszone w sieci Internet (rys. 6). Użytkownik może zatem wizualizować struktury zapisane w postaci plików PDBML na dysku twardym swojego komputera lub pobierać struktury do wizualizacji z sieci Internet. W tym drugim przypadku struktury mogą zostać pobrane bezpośrednio z repozytorium XML bazy PDB ze Stanów Zjednoczonych – użytkownik podaje tylko identyfikator PDB ID białka, które chce wizualizować – lub z dowolnego miejsca, które może być identyfikowane przez adres HTTP lub FTP (np. *http://www.pdb.org/pdb/files/2abx.xml*). Dzięki temu możliwe jest wizualizowanie struktur molekularnych, będących elementami różnych zasobów sieciowych. Jedynym warunkiem jest tylko to, by struktury miały postać zbiorów w formacie PDBML. Przed wczytaniem struktury użytkownik musi podać unikalny identyfikator PDB ID cząsteczki w bazie PDB.

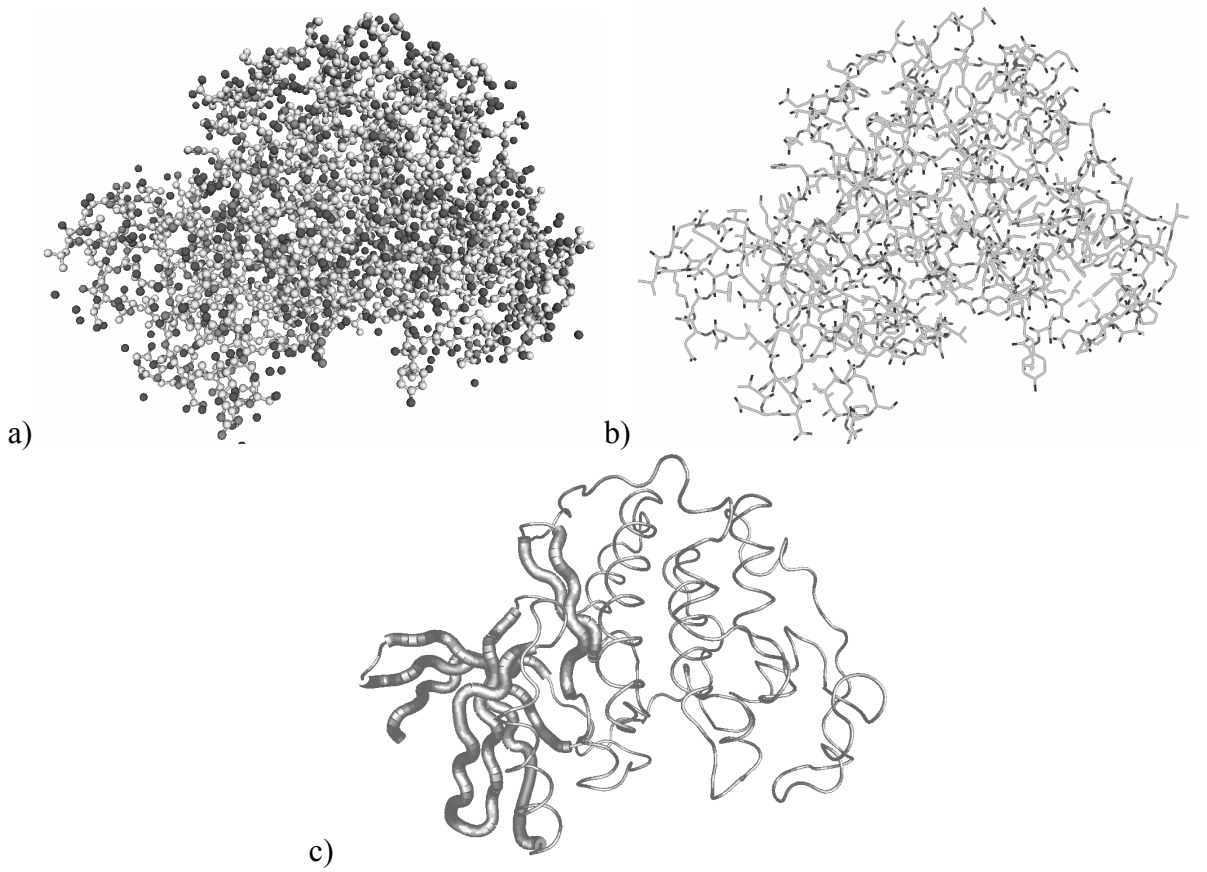


Rys. 6. Architektura wymiany danych makromolekularnych podczas wizualizacji struktur cząsteczek w programie PMV
Fig. 6. Architecture of the macromolecular data exchange during the visualization of molecule structures in the PMV

4.2. Tryby wizualizacji molekuł biologicznych

Podczas wizualizacji białek i innych cząstek lub związków chemicznych użytkownik może mieć swoje preferencje odnośnie do sposobu wizualizacji. Zwykle zależy to od celu wizualizacji, jaki postawi sobie użytkownik, np. czy jest to dogłębna analiza struktury z uwzględnieniem najdrobniejszych jej szczegółów, czy może tylko ogólny przegląd struktury. Z tego powodu program PMV udostępnia trzy tryby wizualizacji struktur (rys. 7):

- atomowy (ang. *atom*) – jest to najbardziej szczegółowy tryb wizualizacji, w którym prezentowane są wszystkie atomy struktury oraz wiązania pomiędzy atomami (rys. 7a); główny akcent prezentacji jest położony na atomy, co może być przydatne m.in. przy pomiarach odległości pomiędzy atomami lub obserwacjach drobnych odkształceń struktury, np. na skutek zajścia określonej reakcji chemicznej; duża liczba szczegółów sprawia natomiast, że widoczne są głównie elementy pierwszego planu;
- szkieletowy (ang. *sticks*) – akcentuje szkielet konstrukcyjny wizualizowanej cząsteczki utworzony przez kowalencyjne wiązania międzyatomowe (rys. 7b); same atomy są oznaczone symbolicznie odpowiednimi kolorami na końcach wiązań; jest to tryb mniej szczegółowy niż tryb atomowy; szczególnie przydatny przy obserwacji odkształceń konformacyjnych zachodzących na skutek różnych czynników;
- wstążkowy (ang. *ribbon*) – ujawnia elementy struktury drugorzędowej w budowie białek (rys. 7c); tryb najmniej szczegółowy, polecany do analizy ogólnej budowy cząstek biologicznych.



Rys. 7. Struktura kinazy CDK2 (PDB ID: 1B38) wizualizowana za pomocą programu PMV – reprezentacje: a) atomowa, b) szkieletowa, c) wstążkowa

Fig. 7. Spatial structure of the CDK2 kinase (PDB ID: 1B38) visualized in the PMV – different representations: a) atom, b) sticks, c) ribbon

4.3. Wyróżnianie fragmentów struktury

Jedną z interesujących cech programu PMV jest możliwość wyróżniania fragmentów struktury przez zmianę sposobu kolorowania lub zaznaczenie wybranej grupy atomów. W programie udostępniono dwa podstawowe sposoby kolorowania struktury: kolorowanie atomów (opcja domyślna) i kolorowanie łańcuchów.

Tabela 1

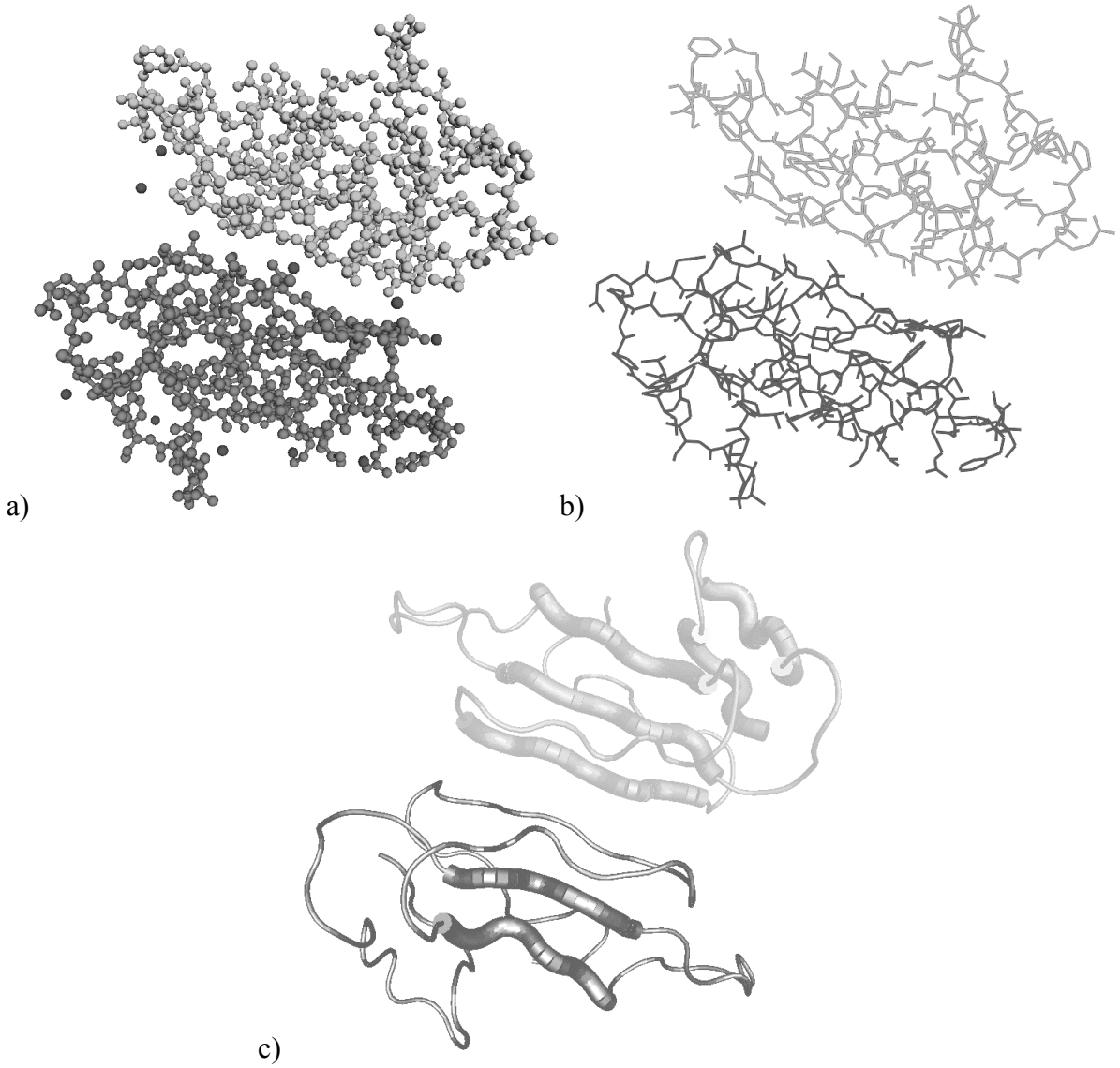
Sposób kolorowania atomów poszczególnych pierwiastków chemicznych

Pierwiastek	Symbol	Kolor	Pierwiastek	Symbol	Kolor
Tlen	O	czerwony	Siarka	S	żółty
Wodór	H	biały	Fosfor	P	pomarańczowy
Azot	N	niebieski	Inne	-	zielony
Węgiel	C	jasny szary			

Kolorowanie atomów jest widoczne tylko w trybach atomowym i szkieletowym wizualizacji. Ma ono na celu wyróżnienie atomów w zależności od pierwiastka chemicznego. Przy-

jęto reguły kolorowania jak w tabeli 1. Kolorowanie atomów, jako domyślne ustawienie, można było zaobserwować na przedstawionych wcześniej rys. 7a i 7b.

Kolorowanie łańcuchów jest możliwe w każdym trybie wizualizacji i ma na celu wyróżnienie każdego z łańcuchów w strukturze czwartorzędowej białka. Kolorowanie łańcuchów w różnych trybach wizualizacji zaprezentowano na rys. 8.

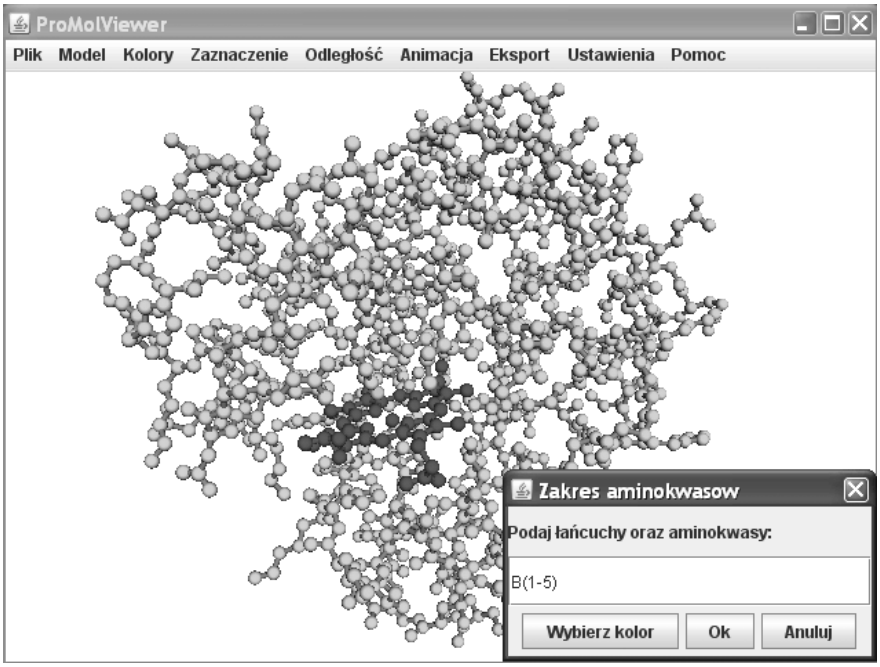


Rys. 8. Kolorowanie łańcuchów cząstki toksyny Alpha-Bungarotoxin (PDB ID: 2ABX) w programie PMV: a) tryb atomowy, b) tryb szkieletowy, c) tryb wstążkowy
Fig. 8. Chain painting of the Alpha-Bungarotoxin molecule (PDB ID: 2ABX) in the PMV: a) atom mode, b) sticks mode, c) ribbon mode

Istnieje również możliwość zaznaczenia tylko wybranej grupy atomów. Pozwala to użytkownikowi wyróżnić tylko wskazane atomy w prezentowanej strukturze (rys. 9). Wybrane atomy zostają oznaczone określonym kolorem. Jest to najbardziej zaawansowana forma kolorowania struktury, przydatna podczas analizy budowy określonych regionów białka. Jeśli

na przykład użytkownik chciałby obejrzeć, jak zbudowany jest aktywny obszar katalityczny enzymu, może wybrać tylko te aminokwasy białka, które tworzą ten obszar i oznaczyć je kolorem. Znacznie łatwiej jest wówczas zidentyfikować właściwe atomy w skomplikowanej strukturze widocznej na ekranie, a następnie zbliżyć je i przyjrzeć się im dokładnie lub zmierzyć wzajemne odległości.

Zaznaczając atomy użytkownik podaje łańcuchy i zakresy aminokwasów, w skład których wchodzą kolorowane atomy, np. zapis A(1-15) oznacza, że użytkownik chce wyróżnić atomy aminokwasów o numerach od 1 do 15 w łańcuchu A cząsteczki białka. Możliwe jest także tworzenie bardziej skomplikowanych wyrażeń, np. A(1-15); B(10); C(20-35), dzięki czemu można zaznaczyć rozłączne grupy atomów.



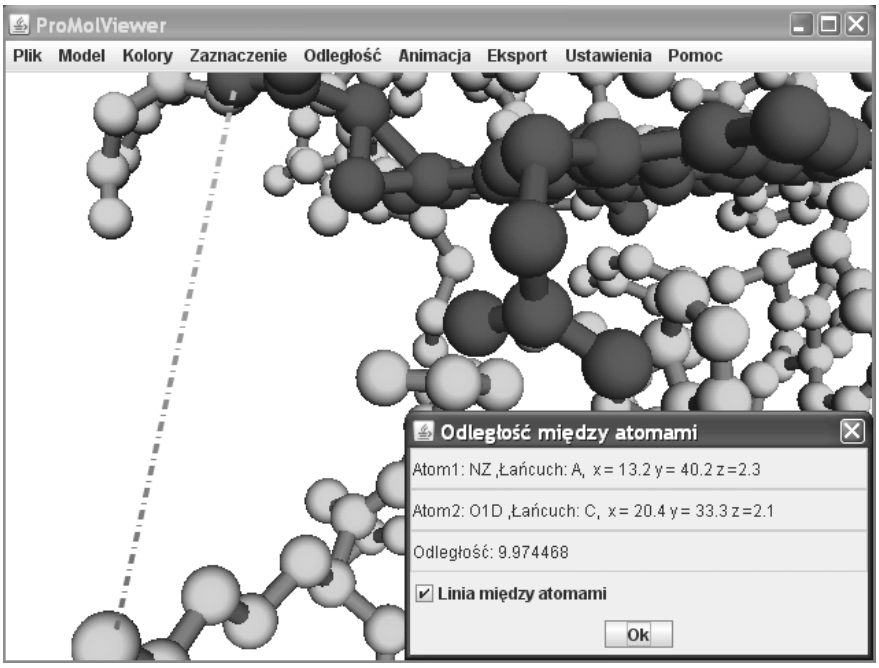
Rys. 9. Wyróżnienie grupy atomów kolorem w programie PMV – zaznaczono fragment struktury mioglobiny (PDB ID: 1MBN) odpowiedzialny za wiązanie tlenu
Fig. 9. Structure painting by marking a group of atoms - part of the structure of the myoglobin (PDB ID: 1MBN) responsible for oxygen binding

4.4. Dodatkowe opcje programu PMV

Program PMV posiada jeszcze kilka dodatkowych, interesujących możliwości. W programie PMV można dokonywać pomiaru odległości pomiędzy wskazaną parą atomów. Użytkownik może tego dokonać tylko w trybie atomowym wizualizacji. Informacja o odległości pojawia się wówczas w dodatkowym oknie informacyjnym, a pomiędzy wskazanymi atomami zostaje narysowana linia (rys. 10). Tego typu pomiary odległości mogą być bardzo pomocne podczas szczegółowej analizy wybranych fragmentów struktury.

W trakcie pracy użytkownik może manipulować obrazem 3D za pomocą myszki, m.in.: obracać strukturę, przesuwać strukturę, powiększać/pomniejszać strukturę.

Jeśli zachodzi potrzeba, użytkownik ma także możliwość eksportu prezentowanej struktury w jej bieżącym układzie do jednego z popularnych formatów graficznych. Możliwy jest eksport obrazu do formatu JPG, PNG, BMP, GIF. Jeżeli jakość obrazu jest niewystarczająca, możliwa jest zmiana rozdzielczości obrazu oraz koloru tła.



Rys. 10. Okno programu PMV – wyróżniony i powiększony fragment struktury mioglobiny (PDB ID: 1MBN) odpowiedzialny za wiązanie tlenu oraz pomiar odległości pomiędzy atomami

Fig. 10. The PMV: Marked and enlarged part of the myoglobin structure (PDB ID: 1MBN) responsible for oxygen binding with the line evaluating the distance between two selected atoms

4.5. Dostępność programu PMV

Protein Molecular Viewer jest programem darmowym, dostępnym dla wszystkich, którzy potrzebują narzędzia do wizualizacji danych z bazy Protein Data Bank zapisanych w formacie PDBML. Program został napisany w dwóch wersjach językowych: polskiej i angielskiej. Podczas pracy programu można z łatwością zmienić język interfejsu GUI. PMV można pobrać wraz z niezbędnymi bibliotekami ze strony Protein Molecular Viewer *http://zti.polsl.pl/dmrozek/pmView.htm*

5. Podsumowanie

Protein Molecular Viewer, pomimo iż wciąż znajduje się w fazie rozwoju, posiada unikalne cechy spośród programów do wizualizacji struktur przestrzennych białek pochodzących z bazy PDB. Za najważniejszą zaletę należy uznać możliwość odczytu danych zapisanych w najnowszym formacie PDBML. Jak dotąd potrafią to tylko nieliczne narzędzia na świecie. W programie PMV zaimplementowano ponadto najważniejsze możliwości spotykane w innych narzędziach, takie jak: kolorowanie struktury, różne tryby prezentacji, możliwość pomiaru struktury, możliwość eksportu obrazu do plików graficznych, prezentacja informacji opisowych dotyczących wczytanej struktury, obracanie, przesuwanie, zmiana rozmiaru i animacja struktury i in. Są to funkcje, które znacznie wspomagają proces analizy strukturalnej białek. Do istotnych cech programu PMV należy zaliczyć również niezwykle rzadką zdolność pobrania i prezentacji struktury białka bezpośrednio z repozytorium Protein Data Bank lub z dowolnego zasobu sieciowego, z wykorzystaniem protokołów HTTP i FTP. Program PMV został napisany w języku Java, dzięki czemu jest niezależny od platformy systemowej i sprzętowej. Ponadto, może pracować jako aplet, który może być elementem każdej strony www.

Program PMV nie jest również pozbawiony drobnych wad. Za główną wadę należy uznać spadek szybkości prezentacji podczas wizualizacji bardzo dużych białek, co może prowadzić do wzrostu zużycia pamięci operacyjnej. Najmniej obciążający jest wówczas model szkieletowy, natomiast najbardziej obciążający model atomowy. Powodem spadku wydajności jest mocno rosnąca złożoność budowanej sceny 3D z wykorzystaniem obiektów biblioteki Java 3D. Rozwiązanie tego problemu będzie jednym z elementów przyszłych prac związanych z rozbudową programu PMV.

LITERATURA

1. Fersht A: Enzyme Structure and Mechanism. 2nd ed. W.H. Freeman & Co., NY 1985.
2. Dickerson R.E., Geis I.: The Structure and Action of Proteins. 2nd ed. Benjamin/Cummings, Redwood City, Calif. Concise 1981.
3. Bańkowski E.: Biochemia. Podręcznik dla studentów uczelni medycznych. Wydawnictwo Medyczne Urban & Partner, Wrocław 2004.
4. Hames B.D., Hooper N.M., Houghton J.D.: Krótkie wykłady – Biochemia. Wydawnictwo Naukowe PWN, Warszawa 2000.
5. Alberts B., i in.: Podstawy biologii komórki. PWN, Warszawa 2005.
6. Stepkiewicz O., Sokalski A.: Cząsteczki na zamówienie. CHIP 07/2000.

7. Murray R.K., Daryl K.G., Mayes P.A., Rodwell V.W.: Biochemia Harpera. Wydawnictwo Lekarskie PZWL, Warszawa 1995.
8. Creighton T.E.: Proteins: Structures and molecular properties. 2nd ed. Freeman, San Francisco 1993.
9. Mrozek D., Małysiak B., and Kozielski S.: EAST: Energy Alignment Search Tool, LNAI, 2006, Vol. 4223, s. 696÷705.
10. Mrozek D., Małysiak B.: Searching for Strong Structural Protein Similarities with EAST. Journal of Computer Assisted Mechanics and Engineering Sciences, 2007, Vol. 14, s. 681÷693.
11. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., and Bourne P.E.: The Protein Data Bank. Nucleic Acids Res., 2000, Vol. 28, s. 235÷242.
12. Marchler-Bauer A., Addess K.J., Chappey C., Geer L., et al.: MMDB: Entrez's 3D structure database. Nucleic Acids Res., 1999, Vol. 27(1), s. 240÷3.
13. Wesbrook J., Ito N., Nakamura H., Henrick K., Berman H.M.: PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics, 2005, Vol. 21(7), s. 988÷992.
14. Westbrook J.D., Fitzgerald P.M.D.: The PDB Format, mmCIF Formats, and Other Data Formats. Structural Bioinformatics, Volume 44, John Wiley & Sons, Inc. 2003.
15. Bourne P.E., Berman H.M., Watenpaugh K., et al.: The macromolecular Crystallographic Information File (mmCIF). Methods Enzymol., 1997, Vol. 277, s. 571÷590.
16. Witryna internetowa Protein Data Bank (2008), http://www.wwpdb.org/documentation/
17. Baxevanis A.D., Ouellette B.F.F.: Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, Inc. 2001.
18. Sayle R.: RasMol, Molecular Graphics Visualization Tool. Biomolecular Structures Group, Glaxo Welcome Research & Development, Stevenage, Hartfordshire 1998.
19. Steinbeck Ch., Han Y., Kuhn S., Horlacher O., et al.: The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. Journal of Chemical Information and Computer Sciences, 2003, Vol. 43(2), s. 493÷500.
20. Kinoshita K., Nakamura H.: jV Users Guide. http://www.pdbj.org/jV/Help.html (2008).
21. Dokumentacja techniczna biblioteki JOGL (2008), https://jogl.dev.java.net/
22. Hogue CW.: Cn3D: a new generation of three-dimensional molecular structure viewer. Trends Biochem Sci. 1997, Vol. 22(8), s. 314÷6.
23. Ohkawa H., Ostell J., Bryant S.: MMDB: an ASN.1 specification for macromolecular structure. Proc Int Conf Intell Syst Mol Biol. 1995, Vol. 3, s. 259÷267.
24. Spjuth O., Helmus T., Willighagen E.L., Kuhn S., et al.: Bioclipse: an open source workbench for chemo- and bioinformatics. BMC Bioinformatics. 2007, Vol. 22, s. 59÷60.

25. Dokumentacja techniczna RCP, http://wiki.eclipse.org/index.php/Rich_Client_Platform (2008).

Recenzent: Dr inż. Sławomir Nowak

Wpłynęło do Redakcji 19 września 2008 r.

Abstract

Applications that visualize spatial structures of proteins and other biological compounds belong to the wide group of tools of molecular analysis used in biochemistry, proteomics and system biology. Functioning of living organisms in biological aspect is tightly related with the existence and activity of proteins. Proteins are important molecules that play a key role in all biochemical reactions in organisms' cells. They are involved in many processes, e.g.: reaction catalysis, energy storage, signal transmission, maintaining of cell mechanical structure, immune response, stimuli response, cellular respiration, transport of small bio-molecules, regulation of cell growth and division [1, 2].

Analyzing their general construction proteins are macromolecules with the molecular mass above 10 kDa (1 Da = 1.66×10^{-24}g) built up with amino acids (>100 amino acids, aa). Amino acids are linked in linear chains by peptide bonds [3]. In the construction of proteins we can distinguish four description (or representation) levels: primary structure, secondary structure, tertiary structure and quaternary structure. The last three levels define the protein conformation or protein spatial structure [3, 4]. The biochemical analysis is usually carried on one of the description levels.

The analysis of protein spatial structure is very important from the viewpoint of protein function, protein activity and reactions the protein is involved in. This type of analysis supported by the observations of protein structure, include not only a sequence, but also geometrical features of studied molecule. There is no doubt, the structures of even small molecules are very complex – proteins are built up of hundreds of amino acids, and then thousands of atoms. Visualization tools, which allow to display and to study spatial structures in the finest details, are useful to explore such complex structures [5]. The common purpose of the tools is a presentation of the general atomic structure of proteins, general spatial shape, and presentation of the simplified construction by the extraction and revealing of secondary structures. As a result, a user can study the shape of a protein or its specific regions and compare it to other proteins, and thus evaluate the similarities and differences. Since the early

eighties scientists have made a use of growing knowledge about various protein structures and functions to rebuild existing proteins and to design completely new molecules. For this reason, protein structure viewers are often applied in drug engineering - they help in the design of effective drugs. The achievements of the modern pharmacy are impressive. However, it is impossible to construct a new drug until the structure of the pathogenic, malfunction protein is found out. Visualization tools are then supportive in the design of inhibitors – substances that decreases excessive activity of some proteins, especially enzymes [6]. One of a vibrant branch of the modern biochemistry, molecular biology and biotechnology became the prediction of protein structures, since it gives many possibilities to the medicine and industry. This process cannot be performed without the insight into protein internal arrangement. Finally, visualization tools are indispensable in the molecular pathology, where scientists investigate the influence of small mutations in protein structures on the protein activity [7, 8].

The visualization of protein structures is performed on the basis of structural data, which have a form of Cartesian coordinates (x, y, z). These coordinates can be retrieved from appropriate databases. Therefore, the visualization of proteins is possible on personal computers. The most popular databases are Protein Data Bank (PDB) [11] and NCBI Molecular Modeling DataBase (MMDB) [12]. They contain data obtained as a result of X-ray crystallography or NMR spectroscopy. These repositories make the data available for an exchange in appropriate formats, like: PDB [14], mmCIF [15] and PDBML [13] for the PDB repository, and ASN.1 [21] for the MMDB repository. The PDBML [13] is the newest format, which benefits from the strength of the XML technology and it will certainly become the main exchange format of structural data for the Protein Data Bank (PDB).

In the chapter, we present our newly developed Protein Molecular Viewer (PMV) – a tool, which visualize protein structures stored in the PDBML data sets. The PMV has several unique features that distinguish it from other viewers. We use the Protein Molecular Viewer to verify results of the protein structural similarity searching processes carried out with the use of our EAST algorithm [9, 10].

Adresy

Andrzej MASTEJ: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-101 Gliwice, Polska, andrewus@interia.pl .
Dariusz MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-101 Gliwice, Polska, Dariusz.Mrozek@polsl.pl .