

Dariusz MROZEK, Bożena MAŁYSIAK-MROZEK, Stanisław KOZIELSKI,
Krzysztof KRAL, Marcin MOCZAŁA, Łukasz OLEŚ, Tomasz SUWAŁA,
Adam SZECÓWKA
Politechnika Śląska, Instytut Informatyki

EDB – BAZA DANYCH ROZKŁADÓW ENERGII POTENCJALNEJ BIAŁEK*

Streszczenie. Analiza cech strukturalnych i energetycznych białek może być kluczem do zrozumienia, w jaki sposób białka oddziałują ze sobą w reakcjach komórkowych. Podczas badania złożonych procesów, w których uczestniczą białka, niezwykle pomocne mogą być rozkłady energii potencjalnej na poszczególnych atomach struktury. Baza Energy Distribution Data Bank (EDB, <http://edb.aei.polsl.pl>) przechowuje rozkłady energii różnych typów dla struktur białkowych pobranych ze znanej amerykańskiej bazy Protein Data Bank. W niniejszym artykule opisujemy cel zbudowanej przez nas bazy EDB, możliwe sposoby jej wykorzystania w badaniach naukowych, możliwości wyszukiwania właściwej informacji i plany dalszego rozwoju.

Słowa kluczowe: bioinformatyka, struktura białka, pola siłowe, profile energetyczne, bazy danych

EDB – ENERGY DISTRIBUTION DATA BANK

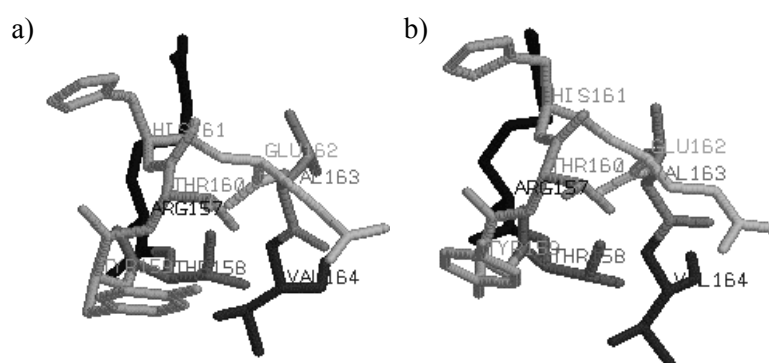
Summary. The analysis of structural and energy features of proteins can be a key to understand how proteins work and interact to each other in cellular reactions. The distributions of energy over each atom in protein structures can be very supportive for the studies of the complex processes proteins are involved in. The Energy Distribution Data Bank (EDB, <http://edb.aei.polsl.pl>) stores a variety of energy distributions for protein molecular structures retrieved from the well-known Protein Data Bank. In the paper, we describe the purpose of the EDB, a possible use of the information stored in it, query possibilities, and plans for future development.

Keywords: bioinformatics, chemoinformatics, protein structure, force field, energy profiles, databases

* Praca naukowa finansowana ze środków na naukę w latach 2008-2010 jako projekt badawczy

1. Wprowadzenie

Funkcje biologiczne białek są zwykle wynikiem procesu fałdowania białek podczas procesu ich syntezy [1, 2, 3]. Właściwie uformowane cząstki biologiczne mogą właściwie spełniać swoją rolę komórkową. Niewłaściwa konstrukcja może prowadzić do poważnych schorzeń, takich jak choroba Alzheimera. Tego typu wady konstrukcyjne w strukturze białek są najczęściej wynikiem mutacji genetycznych na wczesnym etapie formowania się białka [4]. Konformacja białek jest kluczowym czynnikiem decydującym o aktywności i funkcjach, jakie białka pełnią w komórkach organizmów żywych [1]. Konformacja białek jest określona przez organizację i położenie atomów w strukturze przestrzennej białka. Działanie i aktywność białek są często regulowane przez środowisko komórkowe lub przez inne cząstki biologiczne, np. ligandy, które wchodząc w interakcję z danym białkiem powodują zmianę jego aktywności przez zmianę konformacji przestrzennej. Niektóre białka spontanicznie zmieniają swoją konformację z jednej na drugą i jest to element ich naturalnego cyklu komórkowego. Pracują one wówczas jako tzw. molekularne przełączniki kontrolujące lub regulujące odpowiedni proces komórkowy, np. mogą być jednym z wielu elementów ścieżek sygnałowych, gdzie zmieniają swoją konformację w zależności od nadchodzących bodźców [5], [6], [7]. Na rys. 1 zaprezentowano przykład zmiany konformacji białka ludzkiej kinazy CDK2 (Human Cyclin-Dependent Kinase 2) spowodowany fosforylacją aminokwasu treoniny 160 (Thr¹⁶⁰). Na rys. 1a zaprezentowano fragment molekuly CDK2 przed fosforylacją, natomiast na rys. 1b zaprezentowano ten sam region struktury po fosforylacji. W celu lepszej wizualizacji przełączenia konformacji przedstawiono tylko fragmenty struktur molekularnych kinazy CDK2 – osiem aminokwasów w otoczeniu Thr¹⁶⁰.



Rys. 1. Fragmenty struktur przestrzennych kinazy CDK2 (aminokwasy 157-164): a) przed fosforylacją, b) po fosforylacji Thr¹⁶⁰

Fig. 1. Parts of molecular structures of the CDK2 molecule (amino acids 157-164): a) before phosphorylation, b) after phosphorylation on the Thr¹⁶⁰

W procesach modelowania konformacji białek lub modelowania niektórych zjawisk, w których biorą udział cząstki białkowe, stosuje się często metody mechaniki molekularnej [8], [9]. Metody te pozwalają wyznaczać zbiory właściwości energetycznych dla zadanych

struktur białkowych. Mechanika molekularna pełni tu niezwykle istotną rolę ze względu na fakt, iż mechanika kwantowa jest czasem zbyt złożona obliczeniowo, aby modelować procesy komórkowe dla tak dużych cząstek, jakimi są białka. Na przykład, znacznie prościej jest zastosować metody bazujące na tzw. polach siłowych (ang. *force fields*) do modelowania struktury białka niż metod mechaniki kwantowej.

Energia potencjalna struktury białka jest własnością tej struktury lub funkcją organizacji, wzajemnego oddziaływania, typów i położenia atomów w tej strukturze. Energia potencjalna może być wyznaczona na podstawie $3N-6$ wewnętrznych współrzędnych lub $3N$ współrzędnych kartezjańskich, gdzie N jest liczbą atomów w strukturze [9]. Większość obliczeń wykonywanych podczas modelowania procesów molekularnych wiąże się z wyznaczeniem energii potencjalnej, np. w celu znalezienia minimów energetycznych związanych ze stabilnymi stanami struktury białka.

Podczas gdy struktury białek są przechowywane w specjalnych repozytoriach, takich jak Protein Data Bank (PDB) [10], ich właściwości energetyczne muszą być wyznaczone za każdym razem, gdy modeluje się proces komórkowy, w którym dane białko uczestniczy. Takie działanie jest w niektórych sytuacjach kłopotliwe i zwykle powoduje zwiększenie złożoności całego procesu modelowania, ze względu na dodatkowy koszt związany z procesami obliczeniowymi. Z tego powodu autorzy niniejszego artykułu postanowili zaprojektować i zbudować repozytorium przechowujące rozkłady energii potencjalnej, wyznaczone dla znanych struktur białkowych z bazy Protein Data Bank. Powstała baza o nazwie Energy Distribution Data Bank (EDB) udostępnia dane szerokiej społeczności międzynarodowej przez swoją witrynę Internetową EDB. Możliwości wyszukiwania informacji przez witrynę EDB zostały zaprezentowane w rozdziale 4. Podstawy teoretyczne związane z procesem wyznaczania rozkładów energii potencjalnej zostały przedstawione w rozdziale 2. Sposób wyznaczania rozkładów, integracji danych z informacjami z bazy PDB oraz ładowanie danych do bazy EDB zostały opisane w rozdziale 3. Bank rozkładów energii EDB może mieć kilka potencjalnych zastosowań, które zostały przedstawione i opisane w rozdziale 5 niniejszego artykułu.

2. Profile energetyczne struktur molekularnych

Baza Energy Distribution Data Bank przechowuje zarówno dane molekularne białek w postaci współrzędnych kartezjańskich kolejnych atomów struktury przestrzennej, jak i profile energetyczne białek, będące rozkładami energii różnych typów na poszczególnych atomach struktury. W procesach obliczeniowych nad wyznaczeniem profili energetycznych zostało użyte pole siłowe Amber94 [15]. Umożliwia ono modelowanie struktur molekularnych zgodnie z następującą formułą [9]:

$$E_T(r^N) = E_{BS} + E_{AB} + E_{TA} + E_{VDW} + E_{CC} \quad (1)$$

gdzie $E_T(r^N)$ jest całkowitą energią potencjalną struktury białka. Energia ta jest funkcją pozycji (r) N atomów struktury. Pozycje atomów (r) są opisane przez kartezjańskie współrzędne (x, y, z). Dla struktury r^N można wyznaczyć kilka różnych typów energii składowych [9]:

- energię naprężeń wiązań kowalencyjnych E_{BS} (ang. *bond stretching*)

$$E_{BS}(r^N) = \sum_{i=1}^{bonds} \frac{k_i}{2} (d_i - d_i^0)^2 \quad (2)$$

gdzie: k_i jest stałą sprężystości wiązania, d_i jest odległością pomiędzy parą atomów (rzeczywistą długością wiązania), d_i^0 jest optymalną długością wiązania;

- energię odkształceń kątów pomiędzy wiązaniami kowalencyjnymi E_{AB} (ang. *angle bending*)

$$E_{AB}(r^N) = \sum_{i=1}^{angles} \frac{k_i}{2} (\theta_i - \theta_i^0)^2 \quad (3)$$

gdzie: k_i jest stałą deformacyjną, θ_i wartość i -tego kąta pomiędzy parą wiązań, θ_i^0 optymalna wartość tego kąta;

- energię kątów torsyjnych E_{TA} (ang. *torsional angle*)

$$E_{TA}(r^N) = \sum_{i=1}^{torsions} \frac{V_n}{2} (1 + \cos(n\omega - \gamma)) \quad (4)$$

gdzie: V_n jest wysokością n -tej bariery torsyjnej (ang. *torsional barrier*), n to okresowość (ang. *periodicity*), ω jest wartością kąta torsyjnego, γ jest współczynnikiem przesunięcia fazowego (ang. *phase factor*);

- energię oddziaływań van der Waalsa E_{VDW}

$$E_{VDW}(r^N) = \sum_{i=1}^N \sum_{j=i+1}^N \left[4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \right] \quad (5)$$

gdzie: r_{ij} jest odległością pomiędzy atomami i oraz j , σ_{ij} jest odległością między atomami, przy której siły odpychania i przyciągania się równoważą (ang. *collision diameter*), ε_{ij} jest głębokością studni potencjału odpowiadającą minimum energii potencjalnej (ang. *well depth*);

- energię oddziaływań elektrostatycznych E_{CC} (ang. *electrostatic*)

$$E_{CC}(r^N) = \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad (6)$$

gdzie: q_i, q_j są ładunkami elektrycznymi atomów i oraz j , r_{ij} jest odległością pomiędzy oddziałującymi atomami i oraz j , ε_0 jest stałą dielektryczną.

Całkowita energia potencjalna $E_T(r^N)$ jest sumą pięciu energii składowych przedstawionych wcześniej (zgodnie z wyrażeniem 1). Energie składowe natomiast sumują drobne potencjały wynikające z oddziaływań międzyatomowych (wyrażenia 2-7). Na rys. 2 przedstawiono wyniki obliczeń energii potencjalnej dla przykładowego białka. Do obliczeń wykorzystano pakiet mechaniki molekularnej TINKER [16]. Obliczenia, których wyniki przedstawiono na rys. 2, zostały wykonane dla całej struktury białka, a wartości zostały pogrupowane względem typu energii i zsumowane dla wszystkich atomów struktury. Tego typu obliczenia prowadzi się często przy poszukiwaniu stabilnej struktury białka, przez minimalizację jej energii potencjalnej.

W badaniach prowadzonych przez autorów obliczane są nie tylko całkowite wartości poszczególnych typów energii (sumaryczne wartości energii danego typu dla wszystkich atomów struktury), ale również rozkłady energii na poszczególne atomy w strukturze r^N . Interesują nas zatem najdrobniejsze wartości energii, wynikające z oddziaływania dwóch, trzech lub czterech atomów. Zestaw rozkładów różnych typów energii na poszczególne atomy struktury jest nazywany *profilem energetycznym* struktury. Na rys. 3 przedstawiono rozkłady różnych typów energii dla poszczególnych atomów struktury tego samego przykładowego białka co w poprzednim przypadku (z rys. 2). Wartości energii potencjalnej odpowiednich typów dla poszczególnych atomów można odczytać z właściwych pól (EB, EA, ET, EV, EC) kolejnych rekordów. Również w tym przypadku do obliczeń wykorzystano pakiet mechaniki molekularnej TINKER [16].

W badaniach autorów profile energetyczne są obliczane na podstawie struktur molekularnych białek pobranych z popularnej, amerykańskiej bazy Protein Data Bank. W obecnej chwili baza EDB przechowuje profile energetyczne wielu struktur białkowych z bazy PDB.

Total Potential Energy :	16414.4791 Kcal/mole	
Energy Component Breakdown :	Kcal/mole	Interactions
Bond Stretching	15780.6439	5244
Angle Bending	3273.7577	9546
Improper Torsion	132.6853	1218
Torsional Angle	1885.2275	13946
Van der Waals	1573.8932	13476625
Charge-Charge	-6231.7285	13476625

Rys. 2. Wyniki obliczeń energii potencjalnej dla przykładowej struktury białkowej
 Fig. 2. Results of the potential energy calculations for a sample protein structure

```

Total Potential Energy :                16414.4791 Kcal/mole

Potential Energy Breakdown over Atoms :

Atom      EB      EA      EBA      EUB      EAA
          EOPD     EID     EIT     ET      EPT
          ETT     EV      EC      ECD     ED
          EP      ER      ES      ELF     EG

  1      0.0554   0.0001   0.0000   0.0000   0.0000
          0.0000   0.0000   0.0000   0.0000   0.0000
          0.0000  -0.4665  -21.7710  0.0000   0.0000
          0.0000   0.0000   0.0000   0.0000   0.0000

  2      0.7041   0.5509   0.0000   0.0000   0.0000
          0.0000   0.0000   0.0208   0.6207   0.0000
          0.0000  -0.2624  27.2493   0.0000   0.0000
          0.0000   0.0000   0.0000   0.0000   0.0000

  3      0.4255   0.0000   0.0000   0.0000   0.0000
          0.0000   0.0000   0.0000   0.0000   0.0000
          0.0000   0.3543  -44.4122   0.0000   0.0000
          0.0000   0.0000   0.0000   0.0000   0.0000

```

Rys. 3. Wyniki obliczeń rozkładów energii potencjalnej na poszczególnych atomach przykładowej struktury białkowej

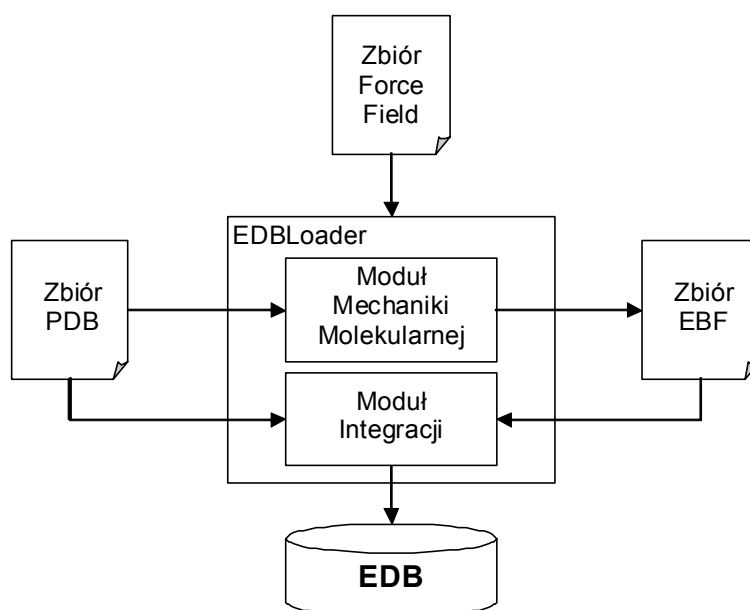
Fig. 3. Calculation of the potential energy breakdown over particular atoms for a sample protein structure

3. Wyznaczanie profili energetycznych struktur molekularnych

Profile energetyczne struktur molekularnych białek są wyznaczane przez autorski program o nazwie EDBLoader, którego sposób działania przedstawiono na rys. 4. Program EDBLoader i jego Moduł Mechaniki Molekularnej (3M) pobierają na wejściu plik w formacie PDB zawierający dane molekularne białek. W wyniku procesu obliczeniowego na wyjściu programu zostaje wygenerowany plik EBF (Energy Breakdown File) z rozkładami energii dla zadanej struktury wejściowej. Plik EBF ma podobną postać do pliku zaprezentowanego na rys. 3. Pojedynczy plik EBF zawiera dane dla pojedynczej struktury, zadanej w postaci wejściowego pliku PDB. Podczas obliczania rozkładów energii moduł 3M używa pakietu TINKER i wywołuje niektóre z jego procedur. Ponieważ w bazie EDB wraz z profilem energetycznym przechowywane są również dodatkowe informacje dotyczące danego białka i jego struktury (np. współrzędne kartezjańskie atomów, sekwencja aminokwasów, informacje opisowe), dane ze zbioru EBF są integrowane z odpowiednimi rekordami zbioru PDB. Scalenia danych dokonuje Moduł Integracji (IM), który następnie ładuje kompletne dane do bazy EDB.

Proces obliczania profili energetycznych dla białkowych struktur przestrzennych wymaga podania dodatkowego zbioru parametrów fizykochemicznych, określającego pole siłowe (ang. *force field*). Znanych jest kilka lub kilkanaście pól siłowych, które mogą zostać użyte

do modelowania molekuł biologicznych. Umożliwiają one wyznaczenie różnych typów energii potencjalnej. W prowadzonych przez autorów badaniach i w zaimplementowanym programie EDBLoader wykorzystywane jest jedno z najbardziej znanych pól Amber94 [15]. W przyszłości planowane jest również użycie innych zbiorów parametrów, takich jak Charmm [17] i Amoeba [18], a przez to poszerzenie zbioru możliwych do wyznaczenia rozkładów energetycznych.



Rys. 4. Architektura i sposób działania programu EDBLoader
Fig. 4. The architecture of the EDBLoader

4. Baza danych Energy Distribution Data Bank

Baza danych *Energy Distribution Data Bank* (EDB) przechowuje profile energetyczne struktur białkowych, dodatkowe informacje dotyczące struktur, informacje o wyodrębnionych strukturach drugorzędowych, a także opisy białek. Podobnie jak w przypadku bazy Protein Data Bank, dostęp do zdeponowanych w bazie informacji jest darmowy dla wszystkich. Użytkownicy mogą uzyskać potrzebne informacje ze strony bazy EDB (pod adresem <http://edb.aei.polsl.pl>). Na stronie tej dostępny jest prosty mechanizm wyszukiwania danych. W bazie EDB jest obecnie przechowywanych 32 229 profili energetycznych struktur białkowych (stan na dzień 2009-02-03).

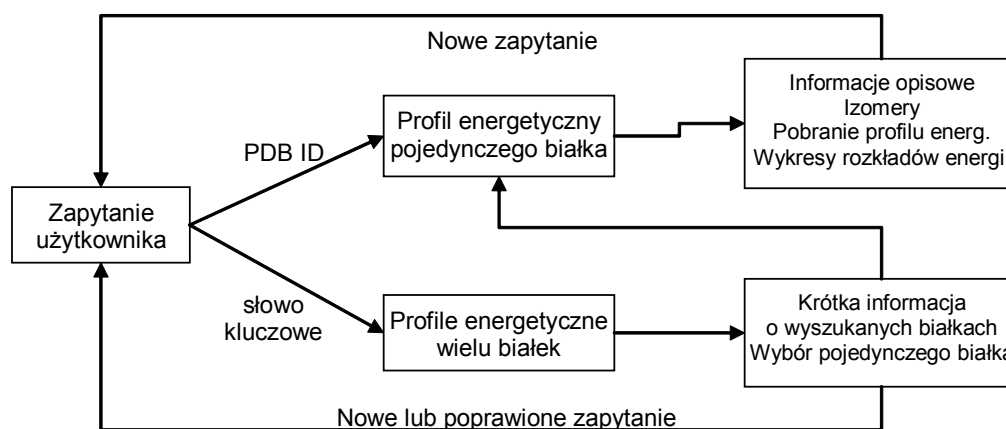
4.1. Możliwości wyszukiwania danych na stronie bazy EDB

Witryna internetowa bazy EDB zapewnia dostęp do danych przechowywanych w bazie szerokiej społeczności użytkowników. Użytkownicy mogą uzyskać informacje o profilach

energetycznych konkretnych białek przechowywanych w bazie EDB na dwa różne sposoby (rys. 5).

W pierwszym ze sposobów użytkownicy mogą podać identyfikator PDB ID białka, którego profil chcą otrzymać na stronie www (rys. 6, pole *PDB ID*). PDB ID jest unikalnym identyfikatorem struktur białek przechowywanych w bazie danych Protein Data Bank. PDB ID jest także stosowany w bazie EDB w celu zunifikowanej identyfikacji molekuł i profili energetycznych. Proces wyszukiwania informacji w bazie EDB ze względu na PDB ID jest przeznaczony głównie dla użytkowników, którzy znają mechanizm wyszukiwania w bazie danych PDB. Jako wynik wyszukiwania użytkownicy otrzymują zbiór informacji o danej molekuł, listę izomerów, możliwość pobrania profilu energetycznego białka w formacie EDML oraz możliwość wygenerowania wykresów zagregowanych rozkładów dla wskazanych typów energii.

W drugim ze sposobów użytkownicy mogą podawać słowa kluczowe (rys. 6, pole *keyword*), które są używane do wyszukiwania względem informacji opisowych białek w bazie EDB. W wyniku takiego wyszukiwania użytkownicy otrzymują listę molekuł, które spełniają kryteria wyszukiwania (rys. 7). Wybierając jedną molekułę z listy, użytkownicy otrzymują dokładny opis molekuły jak w pierwszym podejściu.



Rys. 5. Możliwości wyszukiwania informacji na stronie EDB

Fig. 5. Query options available at the EDB website

Istnieje także trzeci sposób wyszukiwania profilu energetycznego przechowywanego w bazie EDB. Jest on ściśle związany z poszukiwaniem podobieństwa białek z zastosowaniem rozkładów energetycznych. Sposób ten jest zaimplementowany w autorskiej metodzie EAST [11], [12]. Metoda EAST wspomaga wyszukiwanie strukturalnych wzorców o specyficznych cechach energetycznych, np. aktywnych obszarów w enzymach. Niestety, obecnie nie ma jeszcze możliwości użycia metody EAST za pośrednictwem witryny EDB, ponieważ metoda ta została zaimplementowana jako program komputerowy do własnych badań autorów (nie jako aplikacja www). W przyszłości ulegnie to jednak zmianie – planowana jest powtórna implementacja metody EAST jako apletu www.

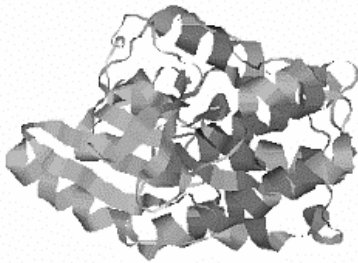
As of Tuesday 2008-12-09 there are 32229 energy profiles

EDB

ENERGY DISTRIBUTION DATA BANK

Home About EDB

Energy Distribution Data Bank (EDB) is a database, which consists of distributions of the atomic charges over all atoms of protein structures. There are different types of energies stored in the EDB: bond stretching, angle bending, torsional angle, improper torsion, van der Waals, and electrostatic. These energies (charges) are calculated on the basis of protein structures taken from the well-known [Protein Data Bank \(PDB\)](#). The computations were done with the use of the **Amber94** force field according to the art of molecular mechanics. The EDB enables users to exchange energy distributions data free of charge. The data of each molecule can be transferred through the **EDML** data sets. Moreover, the distributions of the energy can be visualized at the EDB website in the form of charts. However, in the case, charges are aggregated for each amino acid in a protein polypeptide chain producing so called energy characteristics.



How to use the website:
Simply, put a PDB ID identifier in the PDB ID search field (e.g. 4HHB) or a keyword in the Keyword search field (e.g. hemoglobin) and press the Search button. You can also use general keywords, like: human or mouse. Molecules meeting your search criteria will appear as a result. Then, you can choose, which molecule you want to see in more details, you can download an EDML energy distribution file for a chosen molecule, or generate charts of aggregated energy characteristics of different types of energy.

Website management: [Dariusz Mrozek](#), Institute of Informatics, Silesian University of Technology, Gliwice, Poland

PDB ID:

Keyword:

News

2008-10-29
[EDB has been started!](#)

Rys. 6. Główna strona witryny EDB
Fig. 6. The EDB website – main page

4.2. Zagregowane charakterystyki energetyczne

Profile energetyczne pozwalają prowadzić analizę różnych procesów molekularnych, np. wykrywanie drobnych zmian w strukturze białek. Rozkłady energetyczne mogą być wizualizowane na stronie EDB w formie wykresów. Ponieważ przedstawienie wielowymiarowej informacji w przestrzeni 3D jest trudne, wartości energii są agregowane dla każdego aminokwasu w łańcuchu polipeptydowym, tworząc dla każdego typu energii tzw. charakterystyki energetyczne. Pierwsze wersje wspomnianej metody EAST wykorzystywały charakterystyki energetyczne w procesie poszukiwania podobieństwa białek [18]. Charakterystyki energetyczne są rozkładami energii wzdłuż łańcucha aminokwasów (rys. 8).

As of Tuesday 2008-12-09 there are 32229 energy profiles

EDB

ENERGY DISTRIBUTION DATA BANK

Home About EDB

PDB ID	Name	Description	Protein Class
1XZY	ALPHA-HEMOGLOBIN STABILIZING PROTEIN	SOLUTION STRUCTURE OF THE P30-TRANS FORM OF ALPHA HEMOGLOBIN STABILIZING PROTEIN (AHSP)	CHAPERONE
1W09	ALPHA-HEMOGLOBIN STABILIZING PROTEIN	SOLUTION STRUCTURE OF THE CIS FORM OF THE HUMAN ALPHA-HEMOGLOBIN STABILIZING PROTEIN (AHSP)	CHAPERONE
1W0A	ALPHA-HEMOGLOBIN STABILIZING PROTEIN	SOLUTION STRUCTURE OF THE TRANS FORM OF THE HUMAN ALPHA-HEMOGLOBIN STABILIZING PROTEIN (AHSP)	CHAPERONE
1W0B	ALPHA-HEMOGLOBIN STABILIZING PROTEIN	SOLUTION STRUCTURE OF THE HUMAN ALPHA-HEMOGLOBIN STABILIZING PROTEIN (AHSP) P30A MUTANT	CHAPERONE
1RTX	CYANOGLOBIN	CRYSTAL STRUCTURE OF SYNECHOCYSTIS HEMOGLOBIN WITH A COVALENT HEME LINKAGE	OXYGEN STORAGE/TRANSPORT
1369	CYANOGLOBIN	THE X-RAY STRUCTURE OF THE CYANOBACTERIA SYNECHOCYSTIS HEMOGLOBIN "CYANOGLOBIN" WITH CYANIDE LIGAND	OXYGEN STORAGE/TRANSPORT
136A	CYANOGLOBIN	THE X-RAY STRUCTURE OF THE CYANOBACTERIA SYNECHOCYSTIS HEMOGLOBIN "CYANOGLOBIN" WITH AZIDE LIGAND	OXYGEN STORAGE/TRANSPORT

PDB ID:

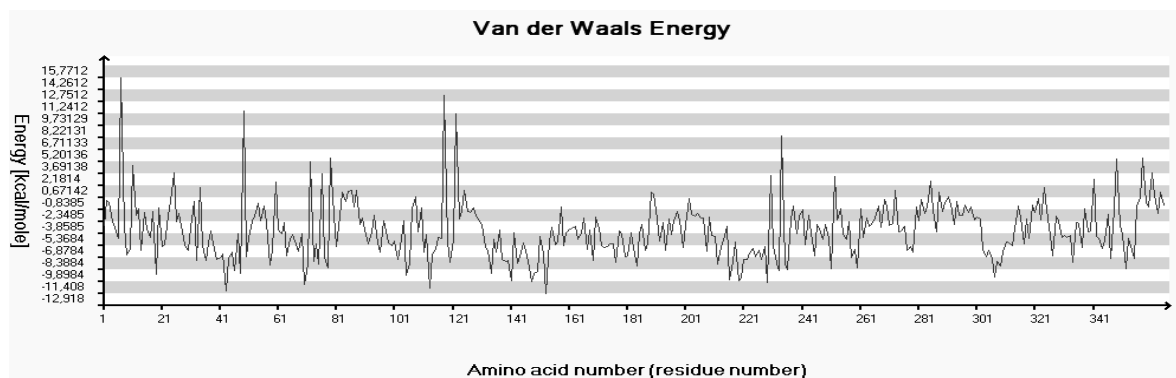
Keyword:

News

2008-10-29

[EDB has been started!](#)

Rys. 7. Częściowe wyniki wyszukiwania w bazie EDB dla słowa kluczowego 'hemoglobin'
 Fig. 7. The EDB website – partial results of the query process for the 'hemoglobin' keyword



Rys. 8. Charakterystyka energetyczna oddziaływań van der Waalsa dla molekuly 2Q8F
 Fig. 8. Van der Waals energy characteristics for molecule 2Q8F

4.3. Format wymiany danych EDML

Profile energetyczne wybranych molekuł mogą być pobierane z bazy EDB w postaci zbiorów w formacie EDML (Energy Distribution Markup Language). Pojedynczy plik EDML zawiera opis białka, informacje o strukturze drugorzędowej, współrzędne kartezyjańskie atomów i rozkłady różnych składowych energii dla wszystkich atomów struktury białka. Na rys. 9 przedstawiono fragment zbioru EDML dla molekuly o PDB ID = 2Q8F (Pyruvate Dehydrogenase Kinase Isoform 1).

```
<?xml version="1.0" encoding="UTF-8"?>
<edb:EnergyProfile xmlns:edb="http://edb.aei.polsl.pl/edml"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="edb.aei.polsl.pl/edml edml.xsd"
  forcefield="Amber94" pdbid="2Q8F">
  <edb:Header>
    <edb:PDBID>2Q8F</edb:PDBID>
    <edb:Isomer>A</edb:Isomer>
    <edb:Name>[PYRUVATE DEHYDROGENASE] KINASE ISOZYME 1</edb:Name>
    <edb:Description>PYRUVATE DEHYDROGENASE KINASE ISOFORM 1
    </edb:Description>
    <edb:ProteinClass>TRANSFERASE</edb:ProteinClass>
    <edb:Source>
      <edb:OrganismScience>HOMO SAPIENS</edb:OrganismScience>
      <edb:OrganismCommon>HUMAN</edb:OrganismCommon>
    </edb:Source>
    <edb:Sequence Chain="A" code="3">GLY VAL PRO GLY GLN VAL ASP PHE
      TYR ALA ARG PHE SER PRO SER ... </edb:Sequence>
  </edb:Header>
  <edb:Energy Chain="A">
    <edb:Peptide aminoacid="GLY" residueno="1" xsi:type="AmberType">
      <edb:Atom name="N" atomno="1">
        <edb:Coordinates>
          <edb:Cartn_x>15,3590002059937</edb:Cartn_x>
          <edb:Cartn_y>55,6310005187988</edb:Cartn_y>
          <edb:Cartn_z>-11,798999786377</edb:Cartn_z>
        </edb:Coordinates>
        <edb:ETotalEnergy>9,49699964879983</edb:ETotalEnergy>
        <edb:EBondStretch>0,223499998450279</edb:EBondStretch>
        <edb:EAngleBend>9,9999974737875E-05</edb:EAngleBend>
        <edb:ETorisonalAngle>0,001799996908009</edb:ETorisonalAngle>
        <edb:EVanDerWaals>-0,652300000190735</edb:EVanDerWaals>
        <edb:EChargeCharge>9,92389965057373</edb:EChargeCharge>
        <edb:EImproperTorsion>0</edb:EImproperTorsion>
      </edb:Atom>
      <edb:Atom name="CA" atomno="2">
        <edb:Coordinates>
          <edb:Cartn_x>13,9099998474121</edb:Cartn_x>
          <edb:Cartn_y>55,7799987792969</edb:Cartn_y>
          <edb:Cartn_z>-12,1809997558594</edb:Cartn_z>
        </edb:Coordinates>
        <edb:ETotalEnergy>-0,898499988717958</edb:ETotalEnergy>
        <edb:EBondStretch>0,259600013494492</edb:EBondStretch>
        <edb:EAngleBend>0,0979999974370003</edb:EAngleBend>
        <edb:ETorisonalAngle>0,001799996908009</edb:ETorisonalAngle>
        <edb:EVanDerWaals>-0,484600007534027</edb:EVanDerWaals>
        <edb:EChargeCharge>-0,773299992084503</edb:EChargeCharge>
        <edb:EImproperTorsion>0</edb:EImproperTorsion>
      </edb:Atom>
      ...
    </edb:Peptide>
    ...
  </edb:Energy>
</edb:EnergyProfile>
```

Rys. 9. Fragment zbioru w formacie EDML dla białka 2Q8F

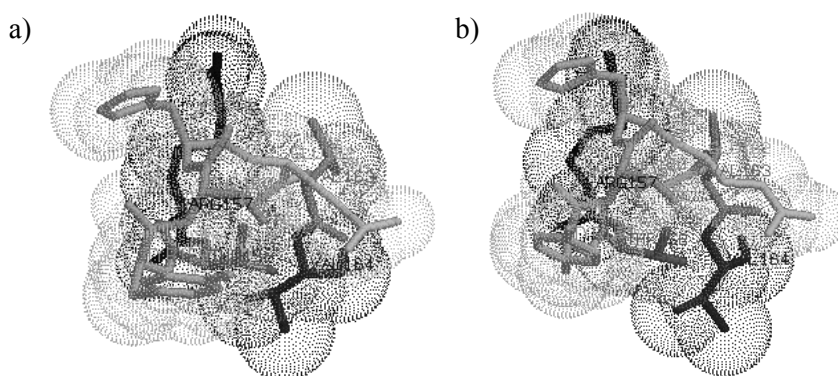
Fig. 9. Part of the EDML file for the molecule 2Q8F

5. Zastosowania bazy EDB

Baza Energy Distribution Data Bank (EDB) udostępnia swoje dane społeczności międzynarodowej nieodpłatnie. Zamiarem autorów jest spopularyzowanie bazy EDB, dzięki czemu będzie ona mogła być wykorzystana do wielu różnorodnych badań naukowych. W obecnej chwili w bazie prowadzone są poszukiwania białek podobnych strukturalnie w oparciu o przechowywane rozkłady energii potencjalnej. Poszukiwanie podobieństwa bazującego na rozkładach energii było tematem badań autorów już w poprzednich latach. W wyniku badań powstał algorytm EAST (ang. *Energy Alignment Search Tool*) [11, 12] poszukiwania podobieństwa białek przez dopasowania ich rozkładów energetycznych.

Istnieją również inne potencjalne zastosowania bazy EDB. Baza EDB przechowuje rozkłady energetyczne, które mogą być traktowane jako szablony rozkładu potencjału i wykorzystane w procesach przewidywania struktur białkowych jako szablony referencyjne.

Rozkłady energii pozwalają na obserwację odkształceń konformacyjnych białek, np. drobnych zmian w budowie będących rezultatem zmiany stanu cząsteczki z aktywnej na nieaktywną lub wynikiem mutacji powodującej dysfunkcję białka [14]. Wykorzystanie rozkładów energii z bazy EDB może stanowić jeden z elementów strukturalnej analizy białek i wspomagać obserwacje konformacji prowadzone za pomocą programów trójwymiarowej wizualizacji, takich jak np. Protein Molecular Viewer [19]. Na przykład, na rys. 10 przedstawiono te same fragmenty struktury ludzkiej kinazy CDK2 co na rys. 1 z dodatkowymi powierzchniami van der Waalsa wokół kolejnych atomów. Zmiana konformacji spowodowana fosforylacją Thr¹⁶⁰ powoduje zmiany rozkładu potencjału w strukturze, co można zaobserwować na rys. 10.



Rys. 10. Fragmenty struktur przestrzennych kinazy CDK2 (aminokwasy 157-164) z powierzchniami van der Waalsa: a) przed fosforylacją, b) po fosforylacji Thr¹⁶⁰

Fig. 10. Parts of molecular structures of the CDK2 molecule (residues 157-164) with the van der Waals potential surfaces: a) before phosphorylation, b) after phosphorylation on the Thr¹⁶⁰

Ponieważ baza EDB przechowuje również dane molekularne o położeniu atomów oraz dane o sekwencji aminokwasów białka (struktury pierwszorzędowe), możliwe jest poszukiwanie podobnych fragmentów struktury białka, np. obszarów aktywnych enzymu lub miejsc

energetycznie korzystnych, na poziomie sekwencji, struktury przestrzennej oraz na poziomie energetycznym.

Informacje o rozkładach energii różnych typów, m.in. oddziaływań elektrostatycznych i oddziaływań van der Waalsa mogą zostać wykorzystane w badaniach nad wzajemnym oddziaływaniem białek (dokowaniem białek, ang. *protein docking*) oraz w modelowaniu reakcji komórkowych.

6. Podsumowanie

Baza EDB jest uniwersalnym repozytorium profili energetycznych wyznaczonych dla struktur białkowych. W badaniach nad funkcjami i aktywnością białek można wyróżnić wiele procesów, w których informacje pobrane z bazy EDB mogą być użyteczne, np. analiza strukturalno-energetyczna, wykrywanie zmian strukturalnych, wspomaganie przewidywania struktur przestrzennych białek, poszukiwanie podobieństwa białek na poziomie strukturalno-energetycznym. Z całą pewnością istnieje także wiele innych obszarów, które nie zostały tutaj przytoczone.

Obecnie w bazie EDB jest przechowywanych 32 229 profili energetycznych. Dla porównania, w bazie PDB jest zgromadzonych 54 466 struktur białek (stan na 03.02.2009). W przyszłości planuje się wygenerowanie i zapisanie w bazie EDB profili energetycznych dla pozostałych molekuł z bazy PDB. Ponieważ liczba struktur białek gromadzonych w PDB każdego roku przyrasta wykładniczo, liczba profili energetycznych w bazie EDB będzie równocześnie rosła. Należy jednak zauważyć, że pomimo, iż w bazie EDB nie jest przechowywanych zbyt wiele informacji opisowych na temat białek (które są zapisane w bazie PDB), baza ta osiągnęła już rozmiar około 90 GB, który stale się powiększa.

BIBLIOGRAFIA

1. Branden C., Tooze J.: *Introduction to Protein Structure*. Garland 1991.
2. Allen J.P.: *Biophysical Chemistry*. Wiley-Blackwell 2008.
3. Cantor C.R., Schimmel P.R.: *Biophysical Chemistry*. W.H. Freeman 1980.
4. Lodish H., Berk A., Zipursky S.L., et al.: *Molecular Cell Biology*. Fourth Edition. W. H. Freeman and Company, NY 2001.
5. Dickerson R.E., Geis I.: *The Structure and Action of Proteins*. 2nd ed. Benjamin/Cummings, Redwood City, Calif. Concise 1981.

6. Creighton T.E.: Proteins: Structures and molecular properties. 2nd ed. Freeman, San Francisco 1993.
7. Berridge M.J.: The Molecular Basis of Communication within the Cell. Scientific American, 1985, No. 253 (4), p. 142÷152.
8. Burkert U., Allinger N.L.: Molecular Mechanics. American Chemical Society, Washington D.C. 1980.
9. Leach A.: Molecular Modelling: Principles and Applications. 2nd Edition. Pearson Education EMA, UK 2001.
10. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., et al.: The Protein Data Bank. Nucleic Acids Res., 2000, No. 28, p. 235÷242.
11. Mrozek D., Małysiak B.: Searching for Strong Structural Protein Similarities with EAST. Journal of Computer Assisted Mechanics and Engineering Sciences, 2007, No. 14, s. 681÷693.
12. Małysiak B., Momot A., Kozielski S., Mrozek D.: On Using Energy Signatures in Protein Structure Similarity Searching. In: Rutkowski L., et al. (eds.) Artificial Intelligence and Soft Computing, LNAI, Springer, Heidelberg 2008, vol. 5097, s. 939÷950.
13. Mrozek D., Małysiak B., Kozielski S.: Energy Profiles in Detection of Protein Structure Modifications, In: IEEE International Conference on Computing and Informatics, Kuala Lumpur 2006, s. 1÷6.
14. Cornell W.D., Cieplak P., et al.: A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J.Am. Chem. Soc., 1995, No. 117, s. 5179÷5197.
15. Ponder J.: TINKER – Software Tools for Molecular Design, User's Guide. Dept. of Biochemistry & Molecular Biophysics, Washington University, School of Medicine, St. Louis 2001.
16. MacKerrell A.D. Jr., et al.: All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B, 1998, No. 102, s. 3586÷3616.
17. Ponder J.W., Case D.A.: Force Fields for Protein Simulation. Adv. Prot. Chem., 2003, No. 66, s. 27÷85.
18. Mrozek D., Mastej A., Małysiak B.: Protein Molecular Viewer for visualizing structures stored in the PDBML format. Advances in Soft Computing, vol. 47, Springer Verlag GmbH, 2008, s. 377÷386.

Recenzent: Dr Ewa Romuk

Abstract

Protein conformation is a key factor deciding about protein activity and protein functions in living cells. The conformation is determined by the arrangement of atoms in a protein structure which forms the general shape of the molecule in the 3-dimensional space [1, 2, 3]. Protein biological functions are usually the result of the folding process during the protein synthesis. Appropriately folded molecules can play their biological roles correctly. Misfolded molecules can cause disease states such as Alzheimer's. These states are usually results of genetic disorders at the very early stage of the protein formation [4]. Protein functioning and activity is often regulated by the cell environment or by binding to other molecules such as ligands. This usually results in a protein conformation switching. Some proteins spontaneously switch from one conformational state to another as a part of their normal life cycle. They can be involved in many cellular processes, like cell signaling where they switch their conformations according to incoming stimuli [5, 6, 7].

Modeling of protein conformations and modeling of some phenomena the proteins are involved in regularly make use of molecular mechanics, which produces a set of energy properties for particular molecular structures [8, 9]. The molecular mechanics plays an important role here, since quantum mechanics is sometimes too computationally complex to model molecular processes for such huge molecules, like proteins. For example, it is much easier to use force field methods to model protein structures in particular processes instead of quantum methods. The potential energy is one of the property or function of a protein structure. It is a product of $3N-6$ internal coordinates or $3N$ Cartesian coordinates, where N is a number of atoms [9]. All computations performed during the modeling of molecular processes involve calculations of potential energy, e.g. to find energy minima related to the stable points of the structure.

While protein structures are stored in dedicated repositories, like Protein Data Bank (PDB) [10], their energy properties have to be calculated any time the structure is used in modeling processes. This can be troublesome in some situations and can generate additional computational costs. For these reasons, we have developed a data repository to store information associated to energy distributions, which are calculated for known protein structures using molecular mechanics methods. The repository is called Energy Distribution Data Bank (EDB, <http://edb.aei.polsl.pl>) and it has several main purposes. In the paper, we describe the purpose of the EDB, a possible use of the information stored in it, query possibilities, and plans for the future development.

Adresy

Dariusz MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, dariusz.mrozek@polsl.pl.

Bożena MAŁYSIAK-MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, bozena.malysiak@polsl.pl.

Stanisław KOZIELSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, Stanislaw.kozielski@polsl.pl.

Krzysztof KRAL: student Politechniki Śląskiej, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska.

Marcin MOCZAŁA: student Politechniki Śląskiej, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska.

Łukasz OLEŚ: student Politechniki Śląskiej, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska.

Tomasz SUWAŁA: student Politechniki Śląskiej, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska.

Adam SZECÓWKA: student Politechniki Śląskiej, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska.