Bazy Danych i Usługi Sieciowe

Bioinformatyczne bazy danych i usługi sieciowe

Paweł Daniluk

Wydział Fizyki

Jesień 2012



Bioinformatyka

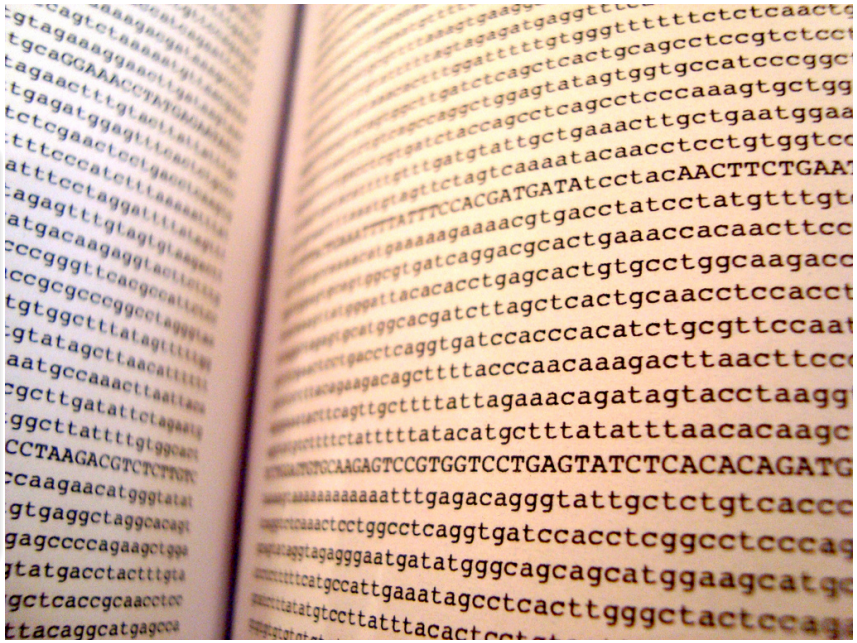


Zagadnienia

- katalogowanie informacji biologicznych
- analiza sekwencji DNA
- analiza sekwencji genomów, porównywanie genomów
- ustalanie ewolucyjnych relacji pomiędzy zbiorami sekwencji / organizmów (drzewa filogenetyczne)
- genotypowanie
- analiza ekspresji genów
- analiza sekwencji białek – proteomika
- katagolowanie funkcji genów/białek, analiza dróg metabolicznych oraz szlaków sygnałowych
- modelowanie układów biologicznych
- wirtualne dokowanie (ang. virtual docking)
- komputery DNA
- morfometria / analiza obrazu

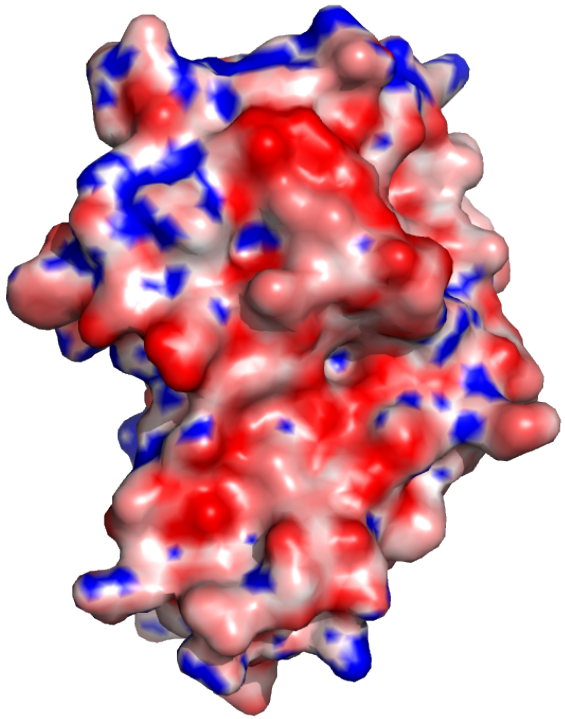
Dane

- sekwencje (nukleotydów, reszt aminokwasowych)



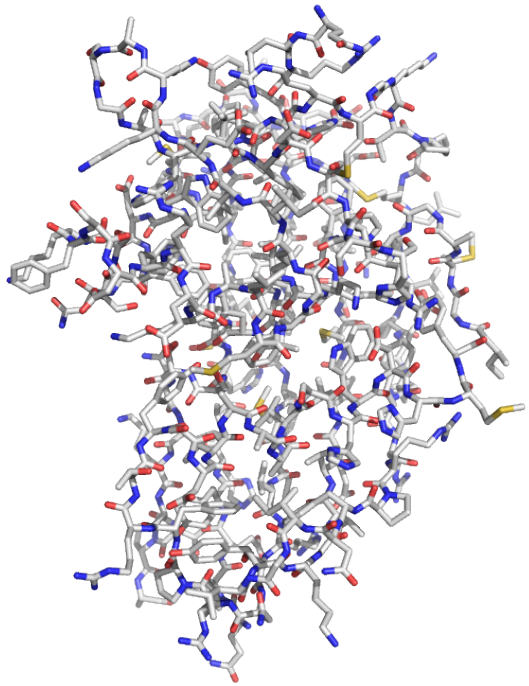
Dane

- sekwencje (nukleotydów, reszt aminokwasowych)
- struktury przestrzenne molekuł



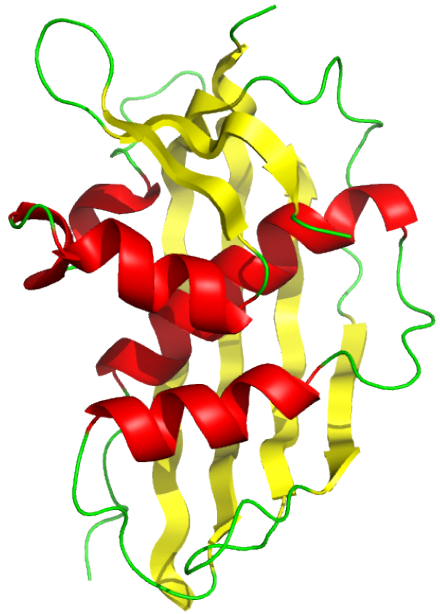
Dane

- sekwencje (nukleotydów, reszt aminokwasowych)
- struktury przestrzenne molekuł



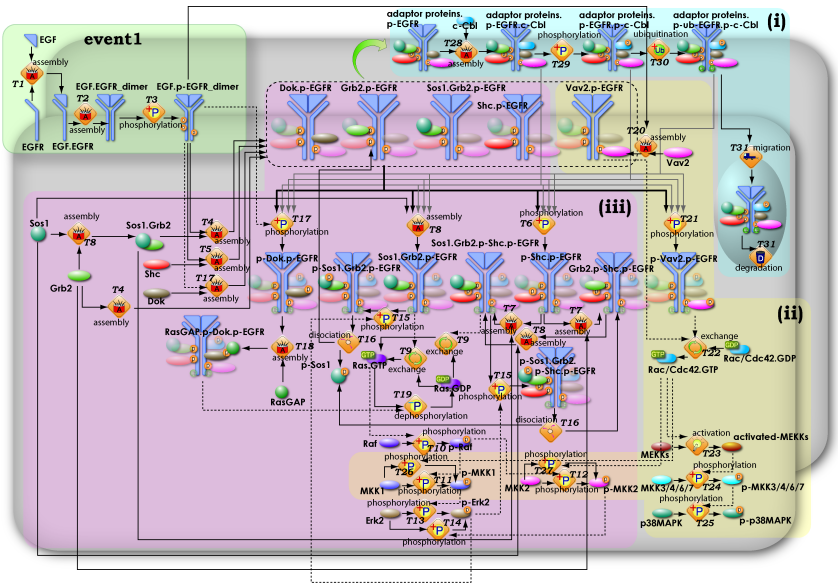
Dane

- sekwencje (nukleotydów, reszt aminokwasowych)
- struktury przestrzenne molekuł



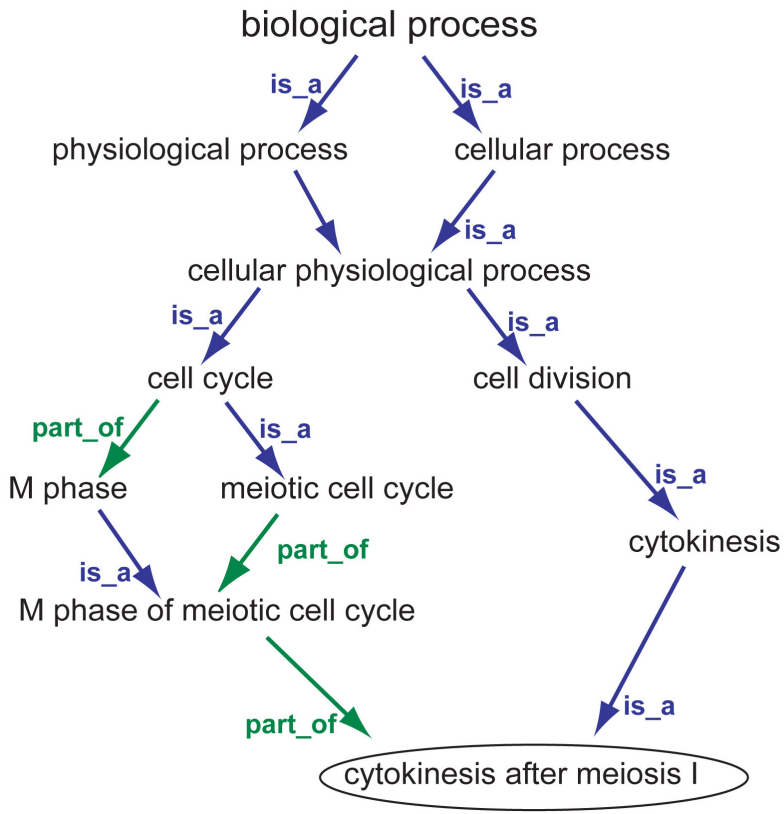
Dane

- sekwencje (nukleotydów, reszt aminokwasowych)
- struktury przestrzenne molekuł
- sieci



Dane

- sekwencje (nukleotydów, reszt aminokwasowych)
- struktury przestrzenne molekuł
- sieci
- ontologie



Reprezentacja danych

- pliki tekstowe
- dane semi-strukturalne (XML)
- tabele

Reprezentacja danych

- pliki tekstowe
- dane semi-strukturalne (XML)
- tabele

FASTA - sekwencja reszt aminokwasowych

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

Reprezentacja danych c.d.

PDB - struktura przestrzenna białka

```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

Reprezentacja danych c.d.

PDBML - struktura przestrzenna białka

(a) Example of a fully marked-up PDBML atom record

```
<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_comp_id>ASP</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>23.482</PDBx:Cartn_x>
         <PDBx:Cartn_y>-0.621</PDBx:Cartn_y>
         <PDBx:Cartn_z>-1.419</PDBx:Cartn_z>
         <PDBx:occupancy>1.00</PDBx:occupancy>
         <PDBx:B_iso_or_equiv>35.27</PDBx:B_iso_or_equiv>
         <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
         <PDBx:auth_comp_id>ASP</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>
```

(b) Example of the alternative simplified XML PDB atom record

```
<category_atom_record>
  <atom_record id="1">
   ATOM 1 A A 1 1 ? . ASP ASP N N N 23.482 -0.621 -1.419 1.00 35.27
  </atom_record>
</category_atom_record>
```

Usługi bioinformatyczne

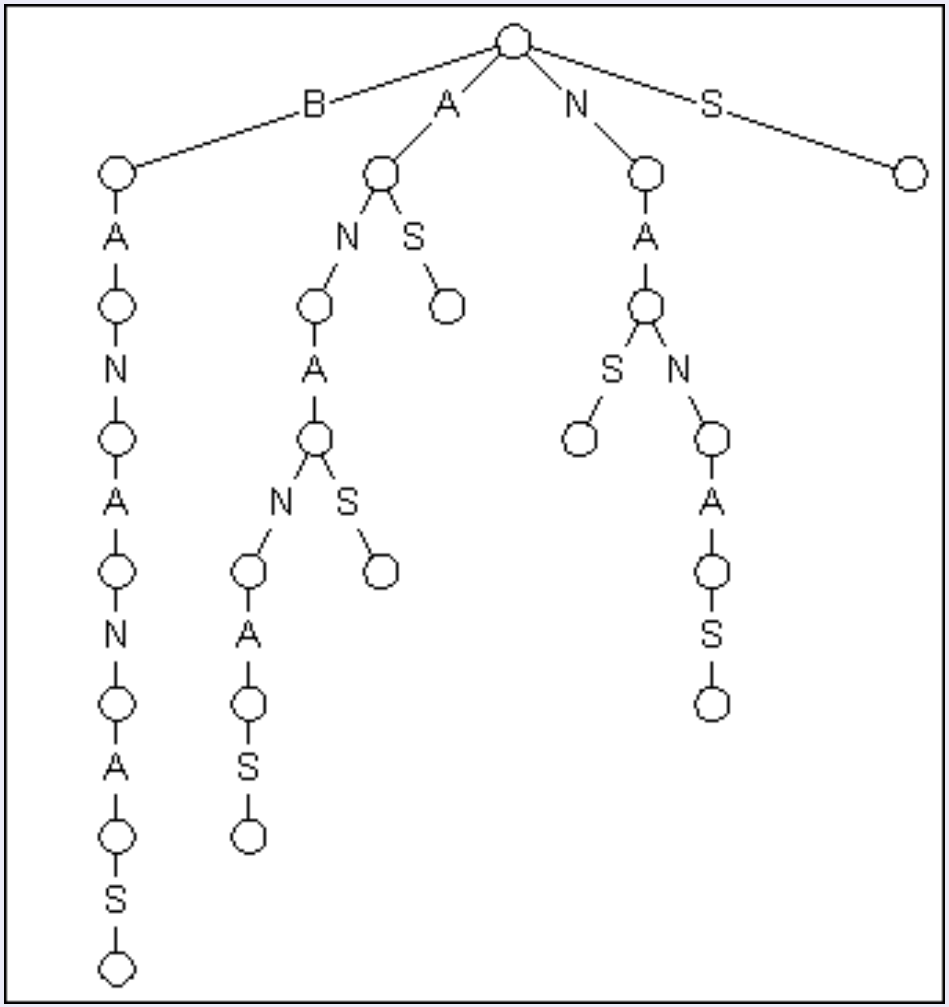
- porównywanie (sekwencji, struktur)
- wyszukiwanie w zbiorze danych
- analiza sekwencji
- mnóstwo innych

Bioinformatyczne bazy danych

- duże ilości danych
- brak spójnego schematu
- skomplikowane kryteria wyszukiwania (często wykonuje się wstępne obliczenia)

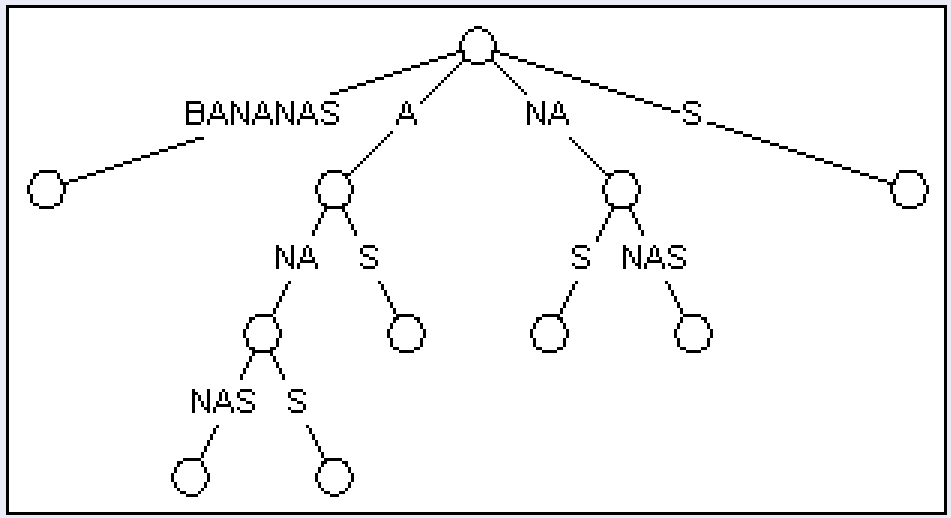
Przeszukiwanie

Drzewa sufiksowe



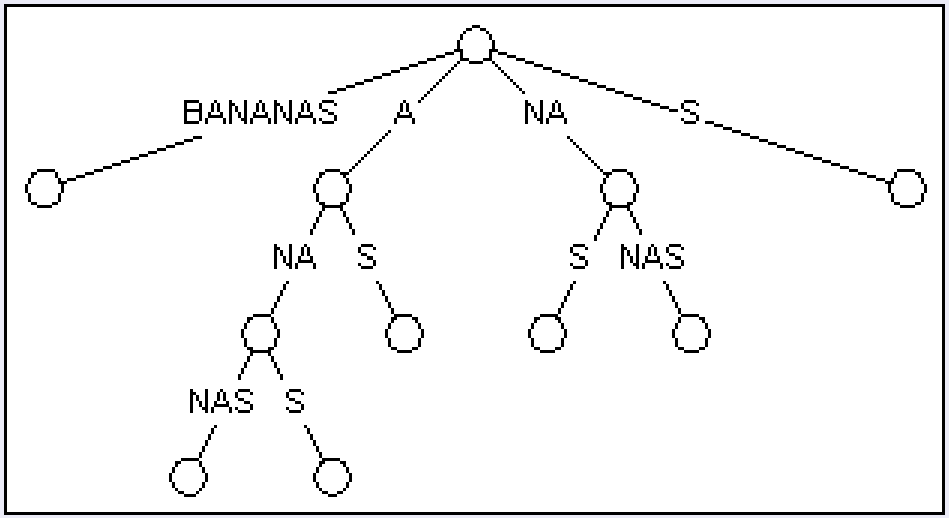
Przeszukiwanie

Drzewa sufiksowe



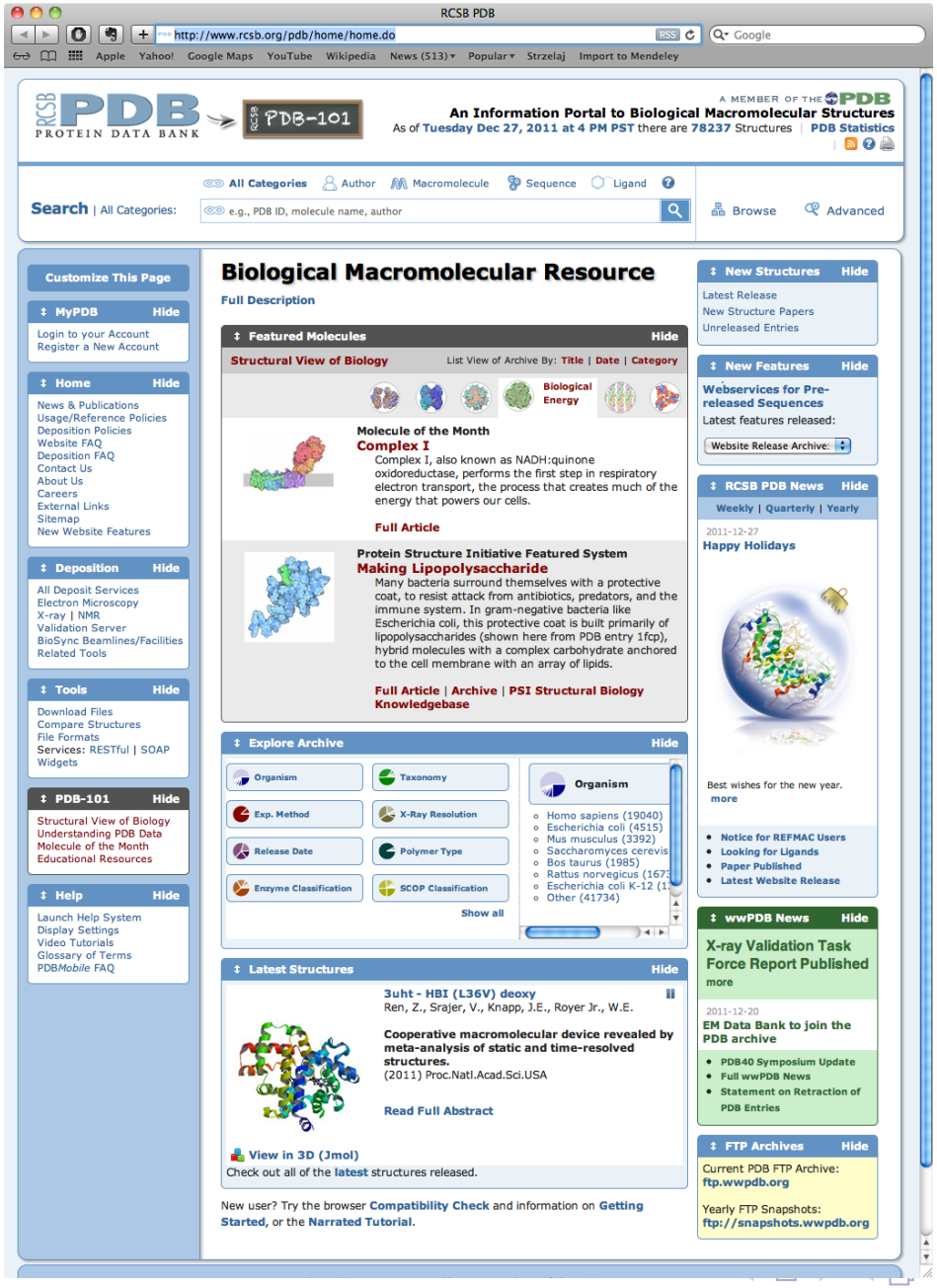
Przeszukiwanie

Drzewa sufiksowe



- Konieczne jest wstępne obliczenie.
- Możliwe jest szybkie stwierdzenie, czy istnieje sufiks zaczynający się od dowolnego słowa.
- Drzewo sufiksowe można zbudować dla wielu słów na raz.

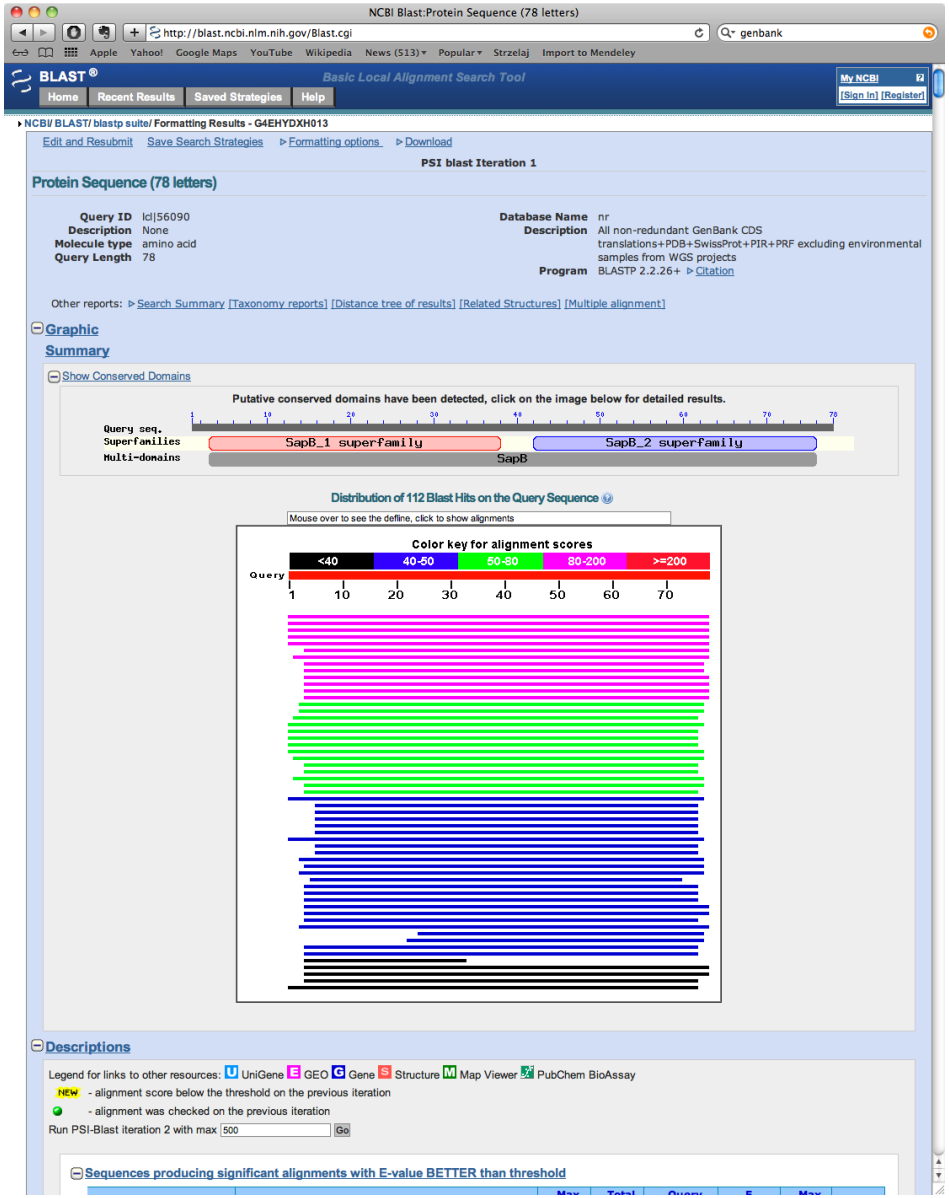
Protein Data Bank



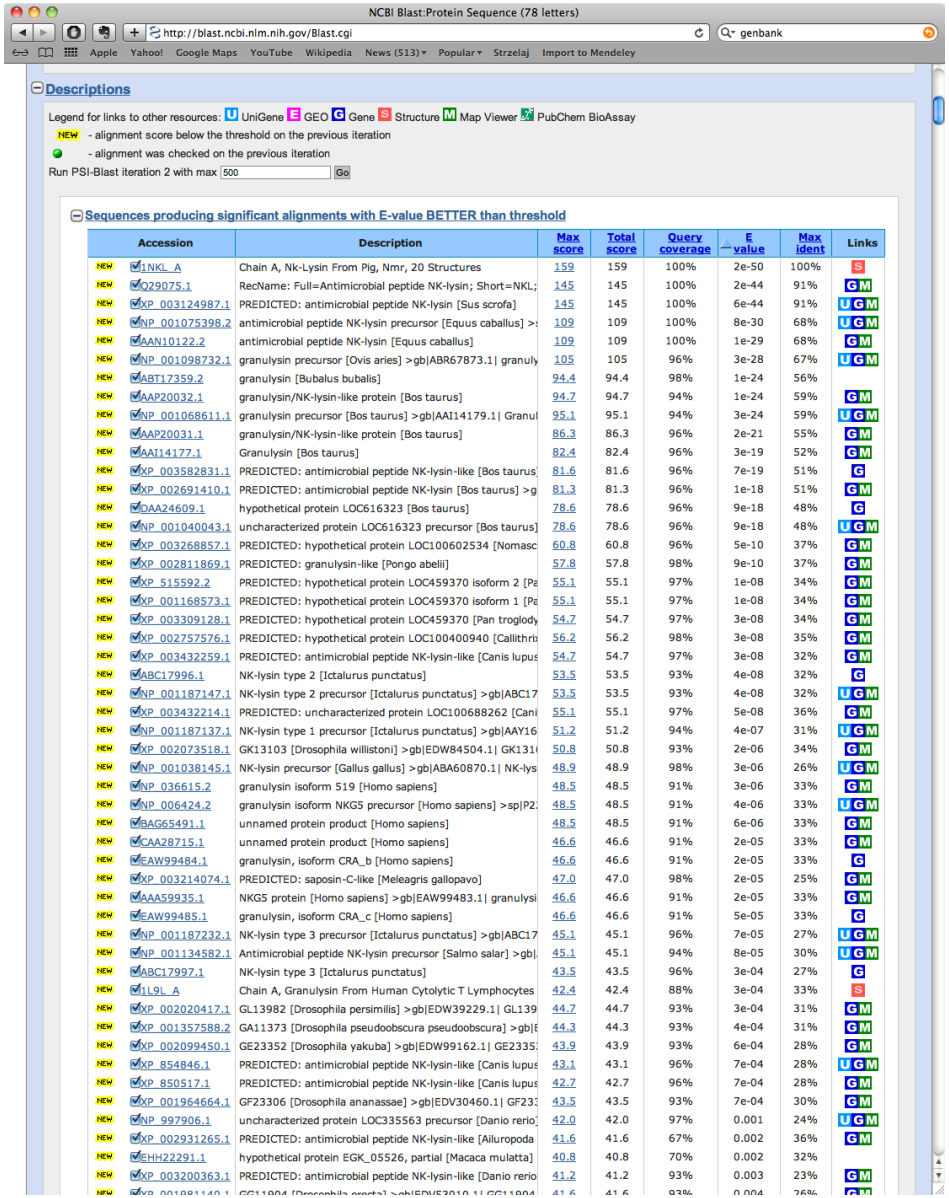
Protein Data Bank



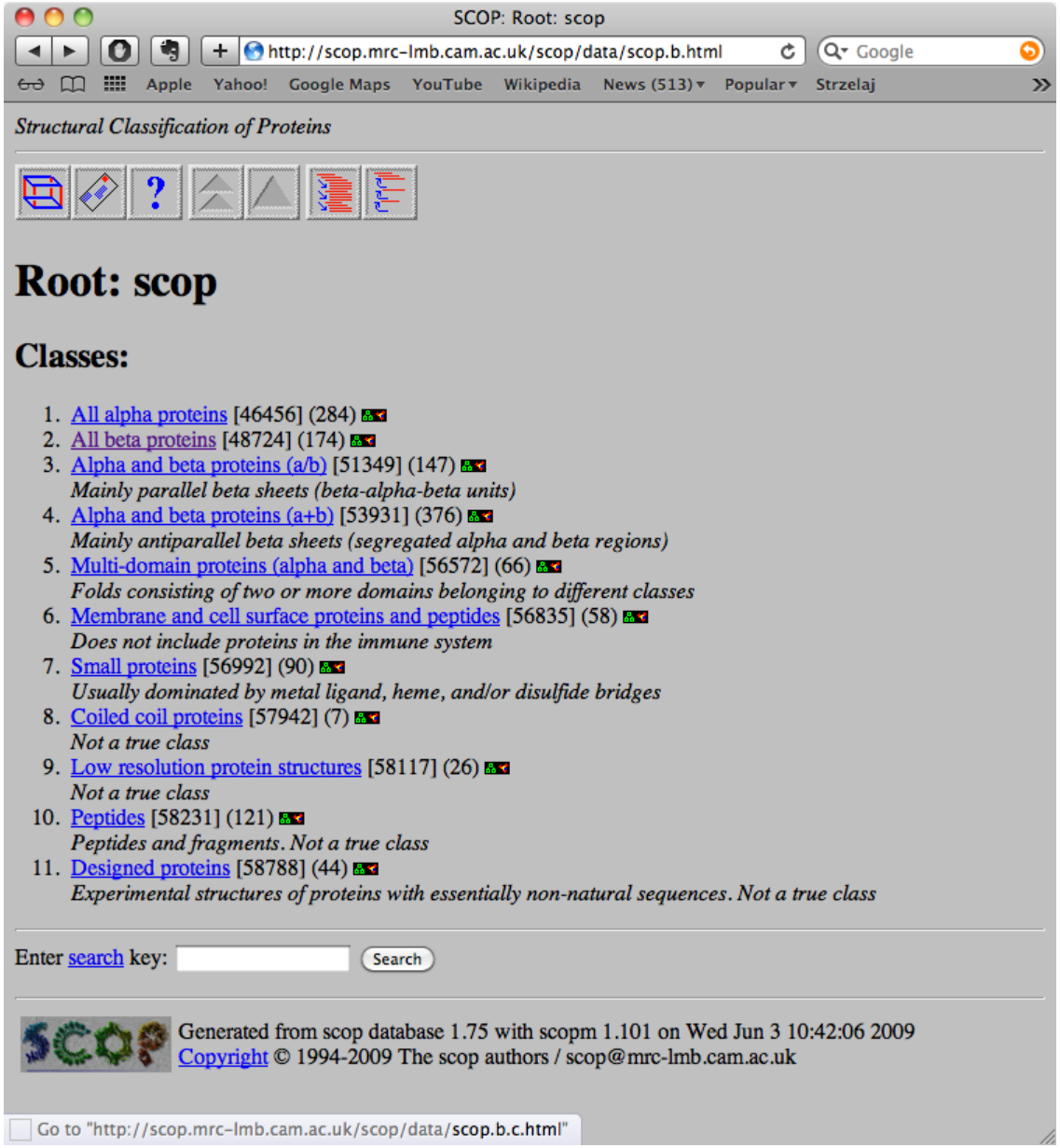
GenBank


NCBI Blast:Protein Sequence (78 letters)
http://blast.ncbi.nlm.nih.gov/Blast.cgi
genbank
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite/ Formatting Results - G4EHYDXH013
Edit and Resubmit
Save Search Strategies
Formatting options
Download
PSI blast Iteration 1
Protein Sequence (78 letters)
Query ID lcl|56090
Description None
Molecule type amino acid
Query Length 78
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.26+ Citation
Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment]
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
Superfamilies
Multi-domains
SapB_1 superfamily
SapB_2 superfamily
SapB
Distribution of 112 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1 10 20 30 40 50 60 70
Descriptions
Legend for links to other resources: UniGene GEO Gene Structure Map Viewer PubChem BioAssay
NEW - alignment score below the threshold on the previous iteration
- alignment was checked on the previous iteration
Run PSI-Blast iteration 2 with max 500 Go
Sequences producing significant alignments with E-value BETTER than threshold

GenBank


NCBI Blast:Protein Sequence (78 letters)
http://blast.ncbi.nlm.nih.gov/Blast.cgi
genbank
Apple Yahoo! Google Maps YouTube Wikipedia News (513) Popular Strzelaj Import to Mendeley
Descriptions
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay
NEW - alignment score below the threshold on the previous iteration
- alignment was checked on the previous iteration
Run PSI-Blast iteration 2 with max 500 Go
Sequences producing significant alignments with E-value BETTER than threshold
Accession | Description | Max score | Total score | Query coverage | E value | Max ident | Links
1NKL_A | Chain A, Nk-Lysin From Pig, Nmr, 20 Structures | 159 | 159 | 100% | 2e-50 | 100% | S
Q29075.1 | RecName: Full=Antimicrobial peptide NK-lysin; Short=NKL; | 145 | 145 | 100% | 2e-44 | 91% | G M
XP_003124987.1 | PREDICTED: antimicrobial peptide NK-lysin [Sus scrofa] | 145 | 145 | 100% | 6e-44 | 91% | U G M
NP_001075398.2 | antimicrobial peptide NK-lysin precursor [Equus caballus] | 109 | 109 | 100% | 8e-30 | 68% | U G M
AAN10122.2 | antimicrobial peptide NK-lysin [Equus caballus] | 109 | 109 | 100% | 1e-29 | 68% | G M
NP_001098732.1 | granulysin precursor [Ovis aries] >gb|ABR67873.1| granuly | 105 | 105 | 96% | 3e-28 | 67% | U G M
ABT17359.2 | granulysin [Bubalus bubalis] | 94.4 | 94.4 | 98% | 1e-24 | 56% |
AAP20032.1 | granulysin/NK-lysin-like protein [Bos taurus] | 94.7 | 94.7 | 94% | 1e-24 | 59% | G M
NP_001068611.1 | granulysin precursor [Bos taurus] >gb|AAI14179.1| Granul | 95.1 | 95.1 | 94% | 3e-24 | 59% | U G M
AAP20031.1 | granulysin/NK-lysin-like protein [Bos taurus] | 86.3 | 86.3 | 96% | 2e-21 | 55% | G M
AAI14177.1 | Granulysin [Bos taurus] | 82.4 | 82.4 | 96% | 3e-19 | 52% | G M
XP_003582831.1 | PREDICTED: antimicrobial peptide NK-lysin-like [Bos taurus] | 81.6 | 81.6 | 96% | 7e-19 | 51% | G
XP_002691410.1 | PREDICTED: antimicrobial peptide NK-lysin [Bos taurus] >g | 81.3 | 81.3 | 96% | 1e-18 | 51% | G M
DAA24609.1 | hypothetical protein LOC616323 [Bos taurus] | 78.6 | 78.6 | 96% | 9e-18 | 48% | G
NP_001040043.1 | uncharacterized protein LOC616323 precursor [Bos taurus] | 78.6 | 78.6 | 96% | 9e-18 | 48% | U G M
XP_003268857.1 | PREDICTED: hypothetical protein LOC100602534 [Nomasc | 60.8 | 60.8 | 96% | 5e-10 | 37% | G M
XP_002811869.1 | PREDICTED: granulysin-like [Pongo abelii] | 57.8 | 57.8 | 98% | 9e-10 | 37% | G M
XP_515592.2 | PREDICTED: hypothetical protein LOC459370 isoform 2 [Pa | 55.1 | 55.1 | 97% | 1e-08 | 34% | G M
XP_001168573.1 | PREDICTED: hypothetical protein LOC459370 isoform 1 [Pa | 55.1 | 55.1 | 97% | 1e-08 | 34% | G M
XP_003309128.1 | PREDICTED: hypothetical protein LOC459370 [Pan troglody | 54.7 | 54.7 | 97% | 3e-08 | 34% | G M
XP_002757576.1 | PREDICTED: hypothetical protein LOC100400940 [Callithri | 56.2 | 56.2 | 98% | 3e-08 | 35% | G M
XP_003432259.1 | PREDICTED: antimicrobial peptide NK-lysin-like [Canis lupus | 54.7 | 54.7 | 97% | 3e-08 | 32% | G M
ABC17996.1 | NK-lysin type 2 [Ictalurus punctatus] | 53.5 | 53.5 | 93% | 4e-08 | 32% | G
NP_001187147.1 | NK-lysin type 2 precursor [Ictalurus punctatus] >gb|ABC17 | 53.5 | 53.5 | 93% | 4e-08 | 32% | U G M
XP_003432214.1 | PREDICTED: uncharacterized protein LOC100688262 [Cani | 55.1 | 55.1 | 97% | 5e-08 | 36% | G M
NP_001187137.1 | NK-lysin type 1 precursor [Ictalurus punctatus] >gb|AAY16 | 51.2 | 51.2 | 94% | 4e-07 | 31% | U G M
XP_002073518.1 | GK13103 [Drosophila willistoni] >gb|EDW84504.1| GK131 | 50.8 | 50.8 | 93% | 2e-06 | 34% | G M
NP_001038145.1 | NK-lysin precursor [Gallus gallus] >gb|ABA60870.1| NK-lys | 48.9 | 48.9 | 98% | 3e-06 | 26% | U G M
NP_036615.2 | granulysin isoform 519 [Homo sapiens] | 48.5 | 48.5 | 91% | 3e-06 | 33% | G M
NP_006424.2 | granulysin isoform NKG5 precursor [Homo sapiens] >sp|P2 | 48.5 | 48.5 | 91% | 4e-06 | 33% | U G M
BAG65491.1 | unnamed protein product [Homo sapiens] | 48.5 | 48.5 | 91% | 6e-06 | 33% | G M
CAA28715.1 | unnamed protein product [Homo sapiens] | 46.6 | 46.6 | 91% | 2e-05 | 33% | G M
EAW99484.1 | granulysin, isoform CRA_b [Homo sapiens] | 46.6 | 46.6 | 91% | 2e-05 | 33% | G
XP_003214074.1 | PREDICTED: saposin-C-like [Meleagris gallopavo] | 47.0 | 47.0 | 98% | 2e-05 | 25% | G M
AAA59935.1 | NKG5 protein [Homo sapiens] >gb|EAW99483.1| granulysi | 46.6 | 46.6 | 91% | 2e-05 | 33% | G M
EAW99485.1 | granulysin, isoform CRA_c [Homo sapiens] | 46.6 | 46.6 | 91% | 5e-05 | 33% | G
NP_001187232.1 | NK-lysin type 3 precursor [Ictalurus punctatus] >gb|ABC17 | 45.1 | 45.1 | 96% | 7e-05 | 27% | U G M
NP_001134582.1 | Antimicrobial peptide NK-lysin precursor [Salmo salar] >gb| | 45.1 | 45.1 | 94% | 8e-05 | 30% | U G M
ABC17997.1 | NK-lysin type 3 [Ictalurus punctatus] | 43.5 | 43.5 | 96% | 3e-04 | 27% | G
1L9L_A | Chain A, Granulysin From Human Cytolytic T Lymphocytes | 42.4 | 42.4 | 88% | 3e-04 | 33% | S
XP_002020417.1 | GL13982 [Drosophila persimilis] >gb|EDW39229.1| GL139 | 44.7 | 44.7 | 93% | 3e-04 | 31% | G M
XP_001357588.2 | GA11373 [Drosophila pseudoobscura pseudoobscura] >gb|E | 44.3 | 44.3 | 93% | 4e-04 | 31% | G M
XP_002099450.1 | GE23352 [Drosophila yakuba] >gb|EDW99162.1| GE2335 | 43.9 | 43.9 | 93% | 6e-04 | 28% | G M
XP_854846.1 | PREDICTED: antimicrobial peptide NK-lysin-like [Canis lupus | 43.1 | 43.1 | 96% | 7e-04 | 28% | U G M
XP_850517.1 | PREDICTED: antimicrobial peptide NK-lysin-like [Canis lupus | 42.7 | 42.7 | 96% | 7e-04 | 28% | G M
XP_001964664.1 | GF23306 [Drosophila ananassae] >gb|EDV30460.1| GF23 | 43.5 | 43.5 | 93% | 7e-04 | 30% | G M
NP_997906.1 | uncharacterized protein LOC335563 precursor [Danio rerio] | 42.0 | 42.0 | 97% | 0.001 | 24% | U G M
XP_002931265.1 | PREDICTED: antimicrobial peptide NK-lysin-like [Ailuropoda | 41.6 | 41.6 | 67% | 0.002 | 36% | G M
EHH22291.1 | hypothetical protein EGK_05526, partial [Macaca mulatta] | 40.8 | 40.8 | 70% | 0.002 | 32% |
XP_003200363.1 | PREDICTED: antimicrobial peptide NK-lysin-like [Danio rerio | 41.2 | 41.2 | 93% | 0.003 | 23% | G M

SCOP


SCOP: Root: scop
http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html
Google
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (513)
Popular
Strzelaj
Structural Classification of Proteins
Root: scop
Classes:
1. All alpha proteins [46456] (284)
2. All beta proteins [48724] (174)
3. Alpha and beta proteins (a/b) [51349] (147)
Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (376)
Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (66)
Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (58)
Does not include proteins in the immune system
7. Small proteins [56992] (90)
Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (7)
Not a true class
9. Low resolution protein structures [58117] (26)
Not a true class
10. Peptides [58231] (121)
Peptides and fragments. Not a true class
11. Designed proteins [58788] (44)
Experimental structures of proteins with essentially non-natural sequences. Not a true class
Enter search key:
Search
Generated from scop database 1.75 with scopm 1.101 on Wed Jun 3 10:42:06 2009
Copyright © 1994-2009 The scop authors / scop@mrc-lmb.cam.ac.uk
Go to "http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.c.html"

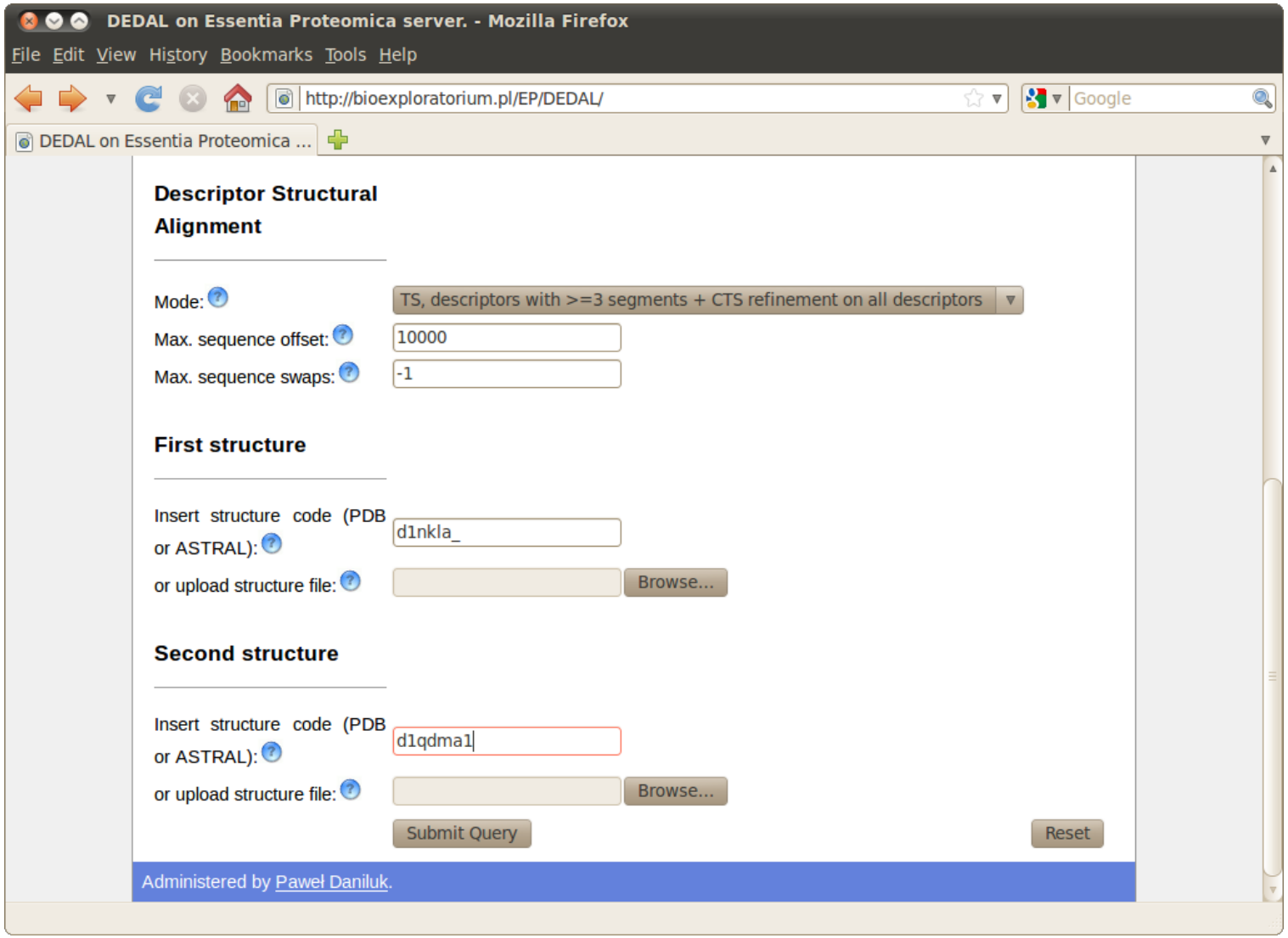
Usługi bioinformatyczne

Większość dobrych czasopism wymaga, aby algorytmy opisywane w publikowanych artykułach miały publicznie dostępne implementacje.

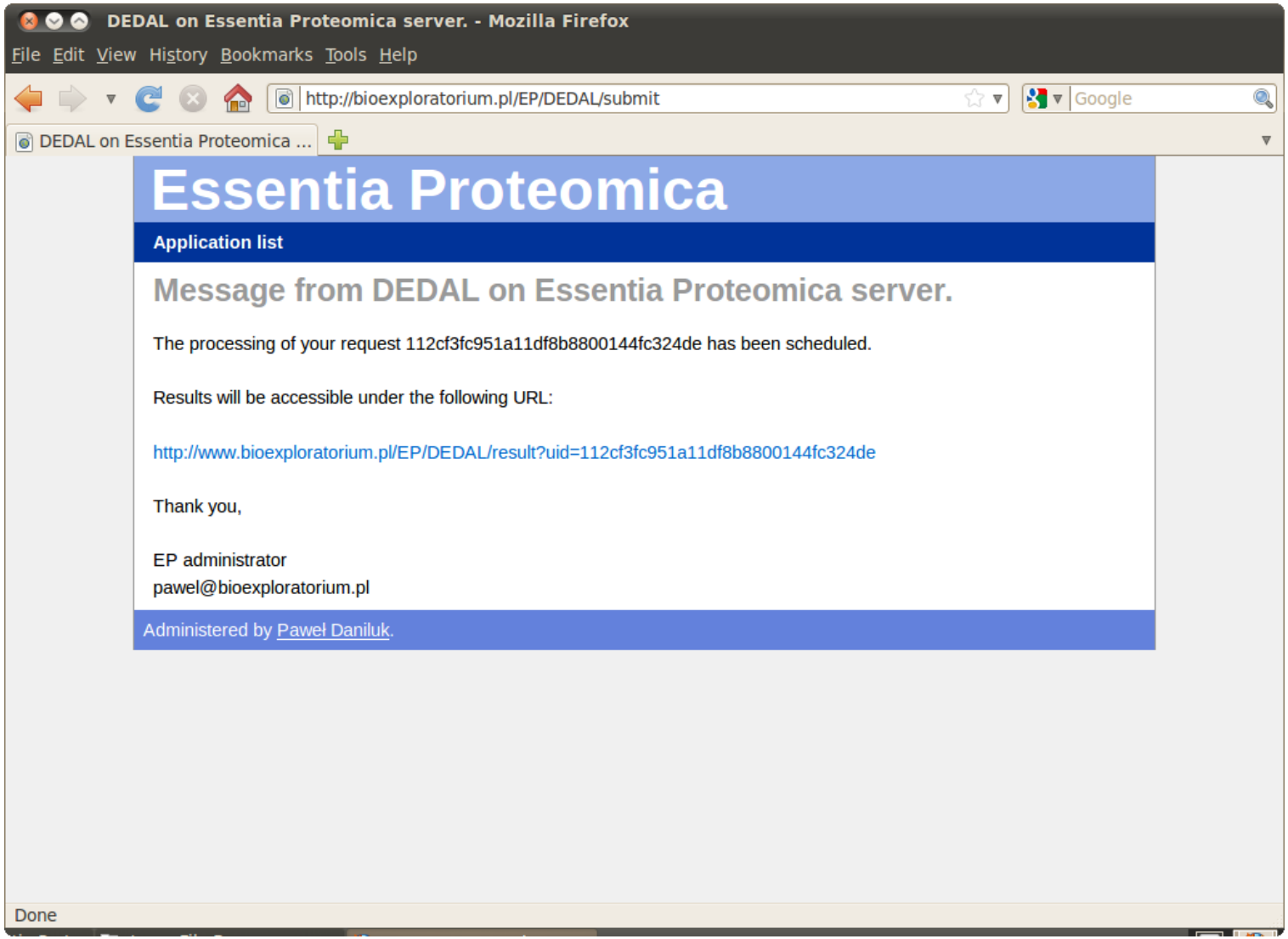
Typowy scenariusz

1. Użytkownik otwiera stronę z formularzem.
2. Użytkownik podaje parametry zapytania, adres e-mail i wysyła zapytanie.
3. Zapytanie jest rejestrowane w systemie i umieszczane w kolejce.
4. Wyświetlana jest informacja o przyjęciu zapytania, jego identyfikatorze i ew. przybliżonym czasie oczekiwania.
5. Zapytanie jest wykonywane.
6. Przygotowywana jest strona z wynikiem.
7. Użytkownik otrzymuje e-mail z linkiem do wyniku.

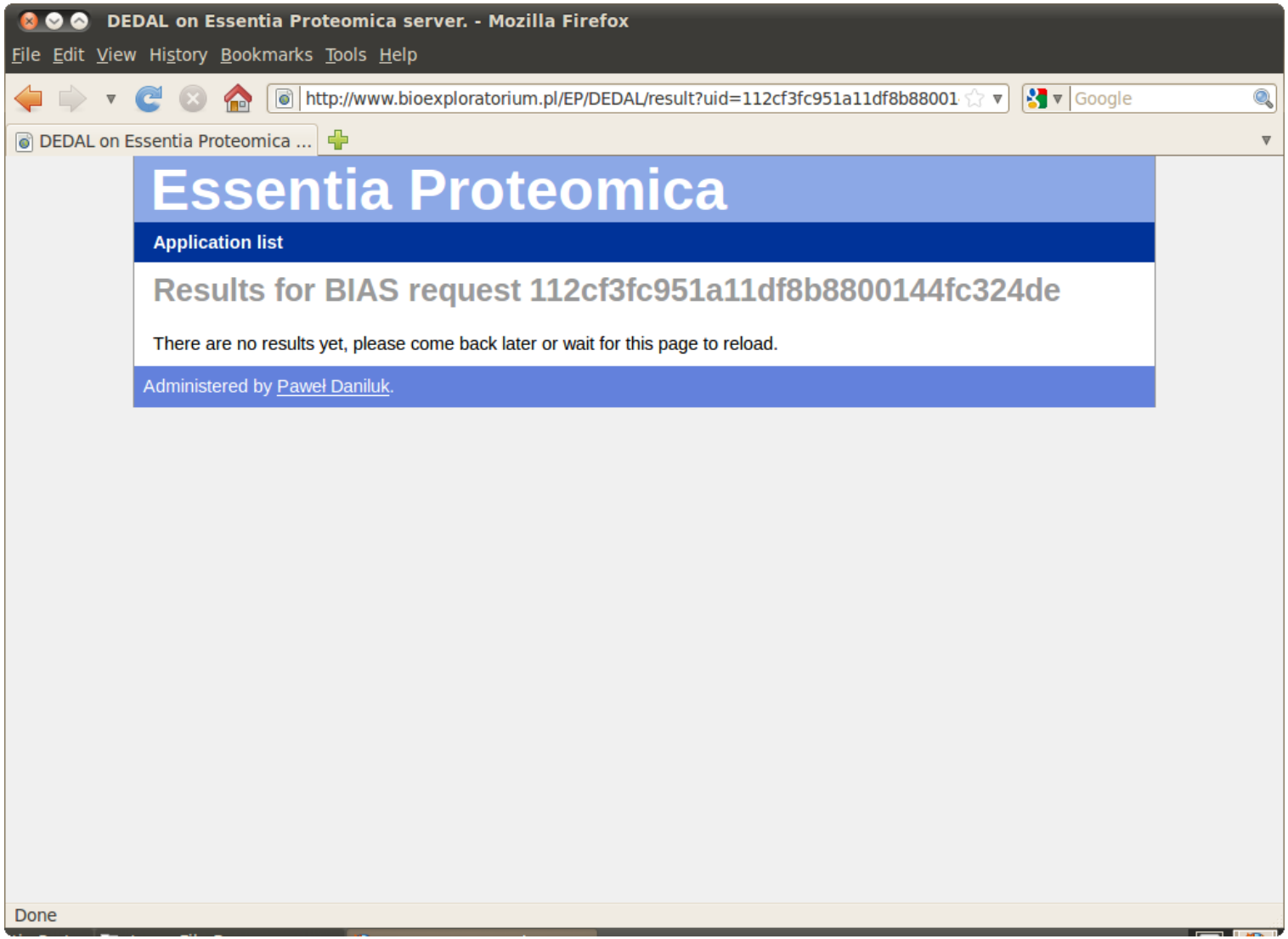
Usługi bioinformatyczne – przykład



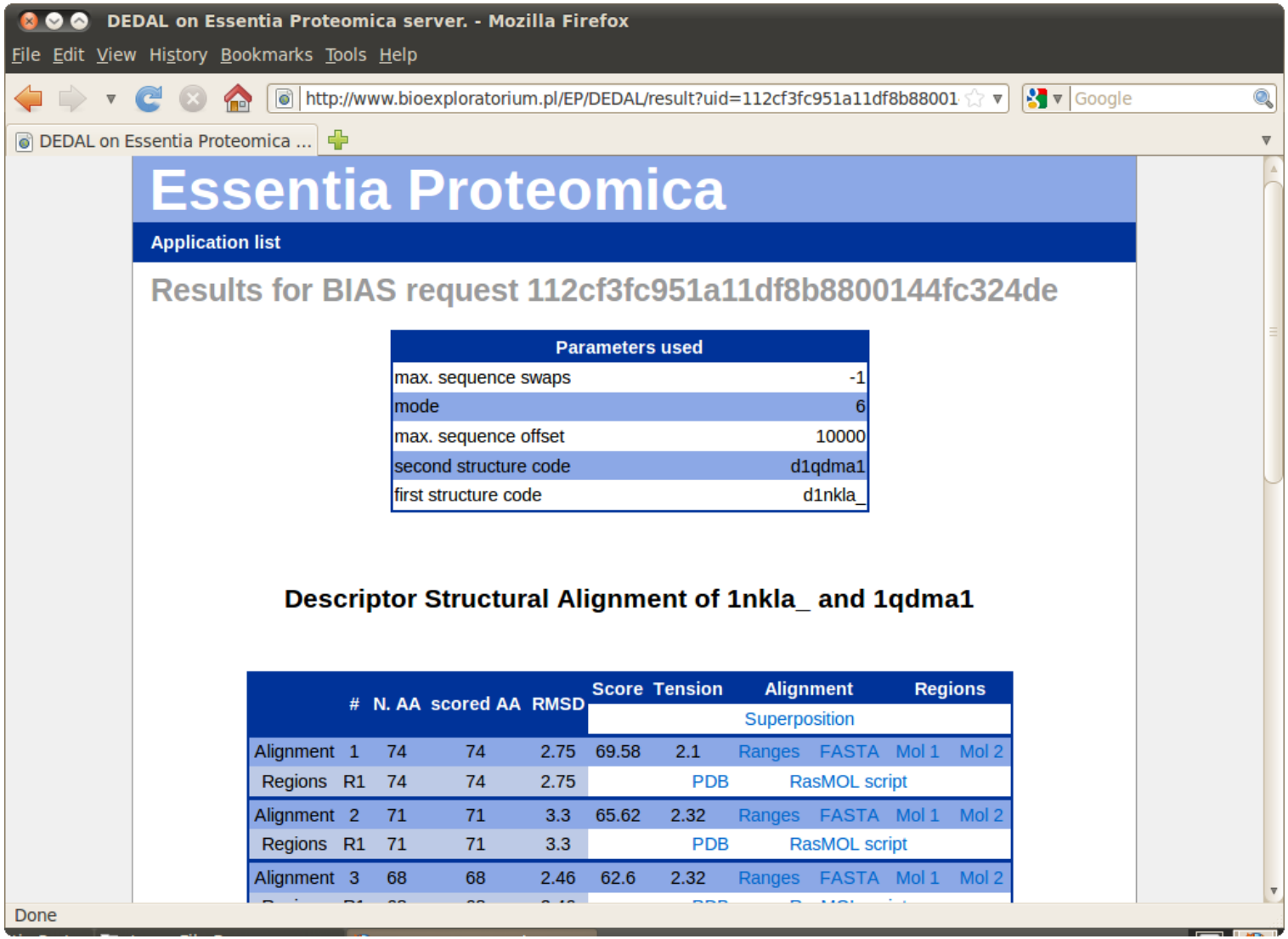
Usługi bioinformatyczne – przykład



Usługi bioinformatyczne – przykład



Usługi bioinformatyczne – przykład



Taverna Workbench

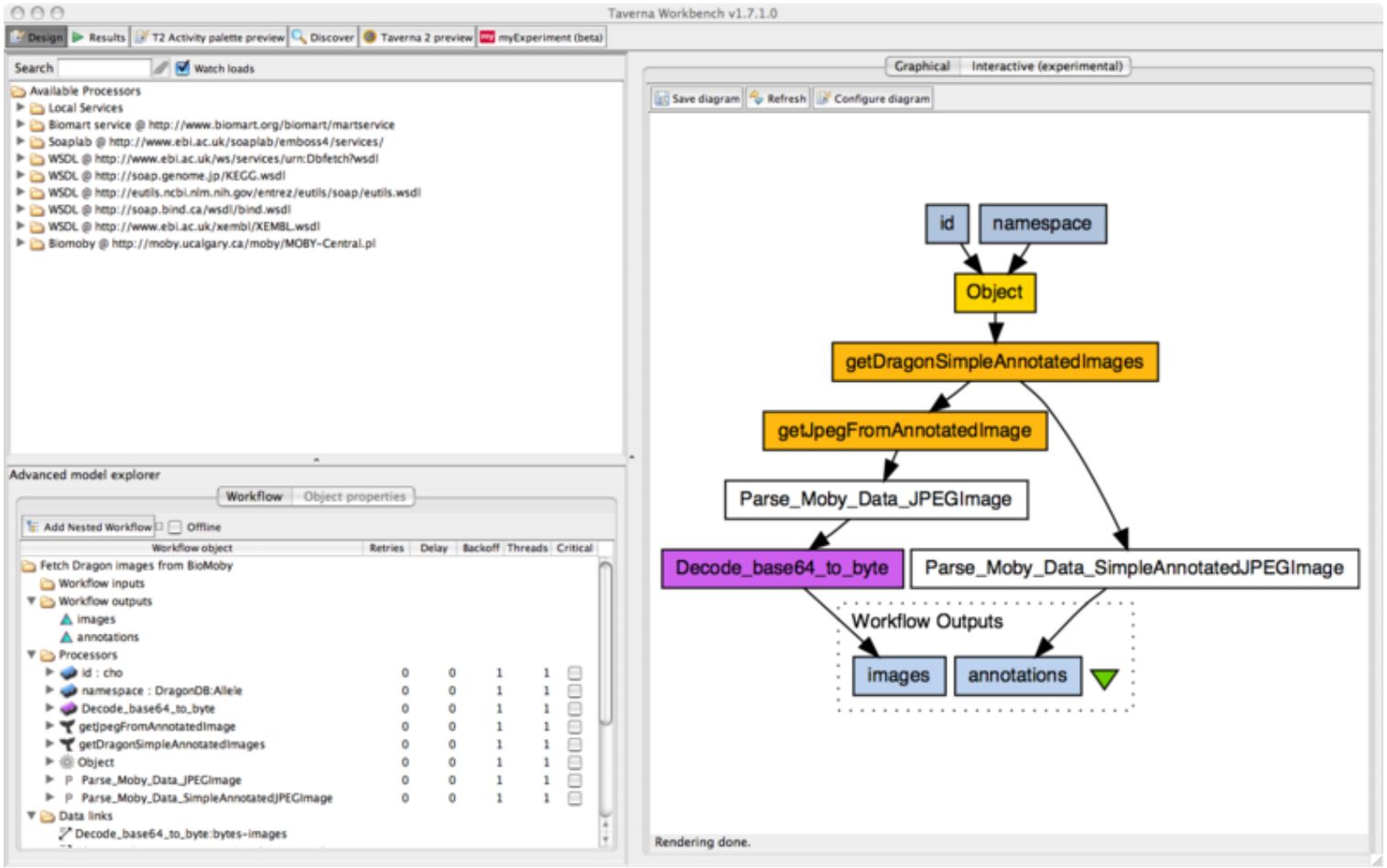
Podejście z formularzem jest przyjazne dla użytkownika, ale niezbyt użyteczne do automatyzacji.

Protokoły wywoływania usług i ich opisu

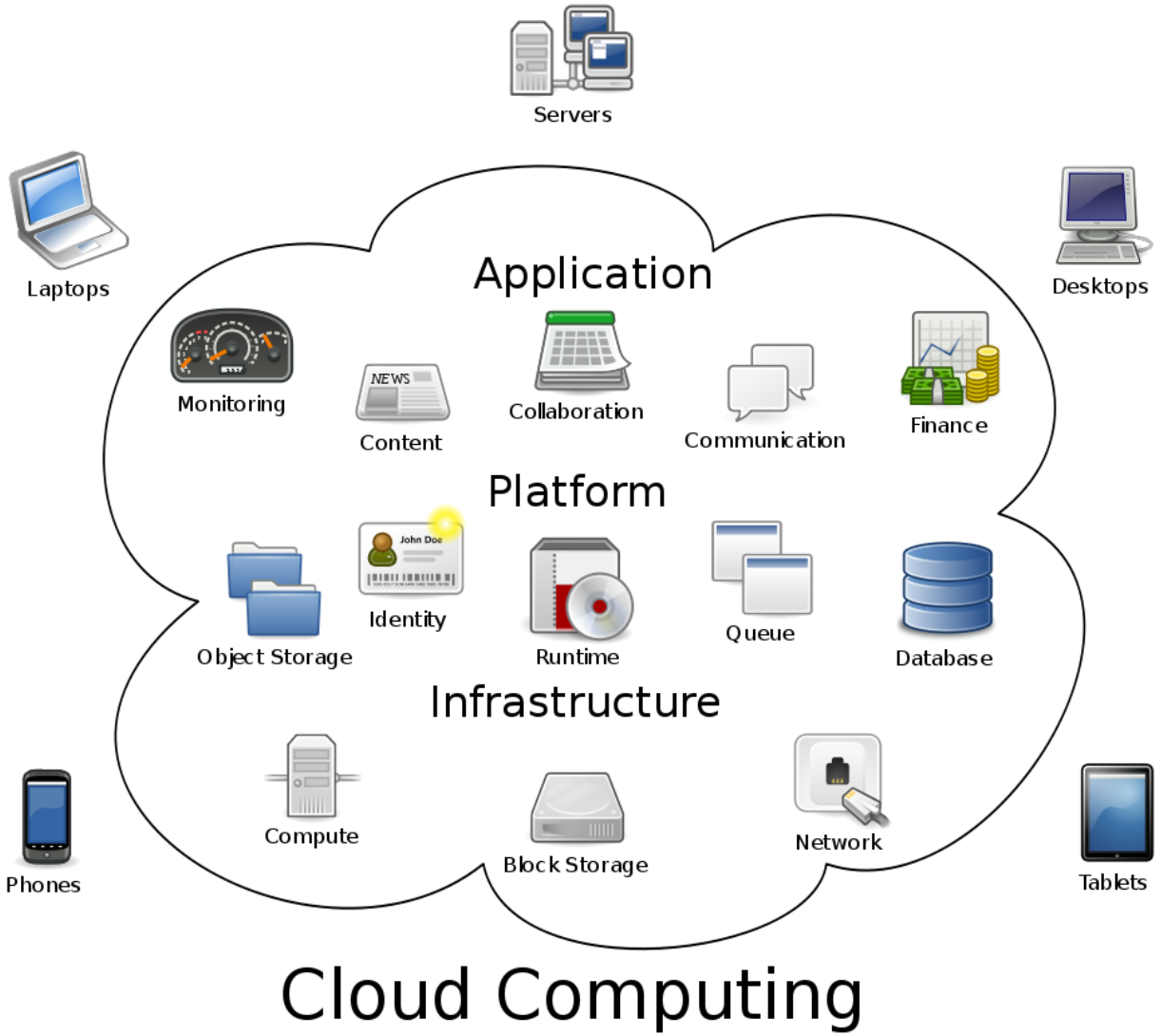
- SOAP (pierwotnie Simple Object Access Protocol)
- Web Services Description Language (WSDL, pronounced 'wiz-dul')

Możliwość łączenia wielu usług (dostępnych na różnych serwerach) w większe schematy.

Taverna Workbench – c.d.



Cloud computing



Cloud computing – c.d.

Warstwy

- Klient
- Usługi (SaaS – Software as a Service) – np. Google Apps, iCloud
- Platformy (PaaS – Platform as a Service) – np. Google App Engine, serwisy hostingowe
- Infrastruktura (IaaS – Infrastructure as a Service) – np. Amazon EC2
- Serwery

Cloud computing – wady i zalety

Zalety

- skalowalność
- niezależność od miejsca
- niższy koszt (efekt skali, inna kwalifikacja kosztu)
- łatwość konserwacji
- bezpieczeństwo
- daje większe możliwości użytkownikom końcowym

Cloud computing – wady i zalety

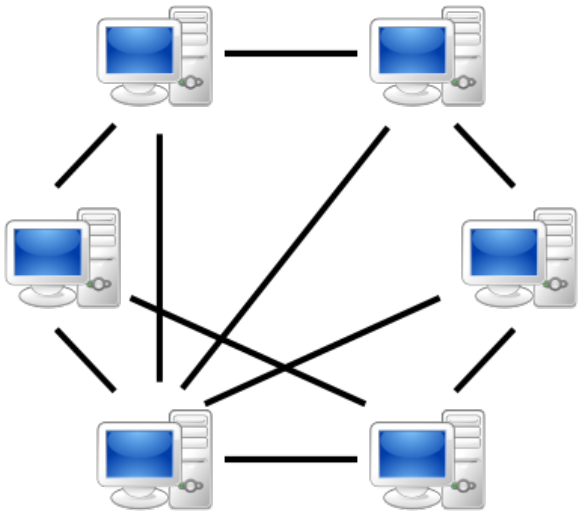
Zalety

- skalowalność
- niezależność od miejsca
- niższy koszt (efekt skali, inna kwalifikacja kosztu)
- łatwość konserwacji
- bezpieczeństwo
- daje większe możliwości użytkownikom końcowym

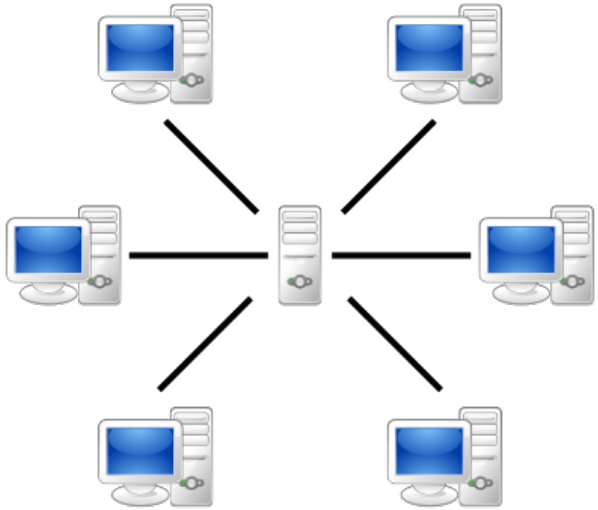
Wady

- brak kontroli nad danymi (data escrow)
- mniejsza możliwość dostosowania do szczególnych wymagań
- wielość standardów
- problemy prawne
- bezpieczeństwo

Peer to Peer (P2P)



vs.



Peer to Peer (P2P) – c.d.

P2P

- podejście czyste – wszystkie węzły są jednakowe
- podejście hybrydowe – superwęzły
- podejście scentralizowane

Peer to Peer (P2P) – c.d.

P2P

- podejście czyste – wszystkie węzły są jednakowe
- podejście hybrydowe – superwęzły
- podejście scentralizowane

Zalety

- lepsze wykorzystanie zasobów (przepustowości)
- łatwość konfiguracji
- niezawodność

Peer to Peer (P2P) – c.d.

P2P

- podejście czyste – wszystkie węzły są jednakowe
- podejście hybrydowe – superwęzły
- podejście scentralizowane

Zalety

- lepsze wykorzystanie zasobów (przepustowości)
- łatwość konfiguracji
- niezawodność

Wady

- niebezpieczeństwo "wstrzyknięcia" błędnych treści
- brak kontroli nad zawartością
- problemy prawne (cenzura internetu)
- blokowanie przez ISP (net neutrality?)

Strona wykładu

http://bioexploratorium.pl/wiki/
Bazy_Danych_i_Usługi_Sieciowe_-_2012z