Curs 13.

Baze de date biologice și platforme/biblioteci pentru bioinformatică .

Biblio:

Cap 2,3 din “Essential Bioinformatics”, Jin Xiong

Cuprins

- Baze de date biologice
- Platforme/biblioteci pentru bioinformatică

Baze de date biologice

- Particularități:
  - Stochează informații de natură biologică (secvențe ADN sau de aminoacizi)
  - Sunt adnotate cu informații de natură bibliografică
  - Trebuie să permită identificarea secvențelor ce se “potrivesc” cu o anumită secvența de interogare. Acest proces este similar cu:
    - Interogarea clasică a bazelor de date
    - Extragerea cunoștințelor din baze de date (ex: identificarea unor șabloane sau a secvențelor similare)

Baze de date biologice

- BD biologice pot fi clasificate in 3 categorii principale:
  - Primare: conțin datele biologice principale = arhive de secvențe furnizate de cercetătorii din domeniu
    - Exemple: GenBank, Protein Data Bank
  - Secundare: conțin informație prelucrată manual sau automat pornind de la BD primare (ex: informații privind funcțiile proteinelor corespunzătoare secventelor).
    - Exemple: SwissProt, Protein Information Resources
  - Specializate: conțin informații specifice anumitor organisme sau tipuri particulare de date
    - Exemple: Flybase, HIV sequence database, Ribosomal Database Project

Baze de date biologice

- BD primare cu secvente ADN (http://www.ncbi.nlm.nih.gov/genbank/)
  - GenBank
  - EMBL (European Molecular Biology Laboratory Database)
  - DDBJ (DNA DataBank of Japan)

- Obs: sunt integrate și împreună formeaza: International Nucleotide Sequence Database Collaboration – transfer zilnic de informații între ele

- Caracteristici:
  - Accesibile gratuit
  - Adnotare minimala a informatiilor
  - Formatele difera usor intre ele

Baze de date biologice

Modalități de acces la GenBank:

- Căutare pe bază de identificator sau informații adnotate prin Search Entrez Nucleotide (http://www.ncbi.nlm.nih.gov/nucleotide/), care conține 3 componente principale:
  - CoreNucleotide (colecția principală de nucleotide) - http://www.ncbi.nlm.nih.gov/nuccore/
  - dbEST (Expressed Sequence Tags – secvențe scurte utile în evaluarea expresiei genice) http://www.ncbi.nlm.nih.gov/nucest/
  - dbGSS (Genome Survey Sequences - secvențe scurte fără adnotări) http://www.ncbi.nlm.nih.gov/nucgss/.
- Căutare pe bază de fragmente de secvențe folosind
  - BLAST (Basic Local Alignment Search Tool).
- Descărcare de secvențe și căutarea în manieră programatică (din alte aplicații) folosind
  - NCBI e-utilities

Baze de date biologice

Modalități de acces la GenBank:

- Căutare pe bază de fragmente de secvențe folosind
  - BLAST (Basic Local Alignment Search Tool) http://blast.ncbi.nlm.nih.gov/Blast.cgi.

- Descărcare de secvențe și căutarea în manieră programatică (din alte aplicații) folosind
  - NCBI e-utilities http://www.ncbi.nlm.nih.gov/books/NBK25501/
  - 8 aplicatii de tip server ce pot fi accesate prin postarea unui URL E-utility URL către NCBI și returnează un răspuns XML
  - pot fi accesate din orice limbaj de programare (Perl, Python, Java, C++)

Baze de date biologice

- **EInfo (database statistics)**
  eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi

  - furnizează numărul de înregistrări indexate în baza de date, data ultimei actualizări, link-uri către alte baze de date
- **ESearch (text searches)**
  eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi

  - procesează interogări de tip text și returnează lista cu identificatorii corespunzători
- **EPost (UID uploads)**

eutils.ncbi.nlm.nih.gov/entrez/eutils/epost.fcgi

- Acceptă o listă de identificatori, o stochează în History Server, și returnează o cheie de identificare pentru setul uploadat.

Baze de date biologice

- **ESummary (document summary downloads)**

  eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi

  - Returnează sumarul documentelor corespunzătoare unei liste de identificatori.
- **EFetch (data record downloads)**

  eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

  - Returneaza înregistrările corespunzătoare listei cu identificatori (în formatul specificat)
- **ELink (Entrez links)**

  eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi

  - Returnează o listă de identificatori corelați cu cei transmiși ca parametru

Baze de date biologice

- **EGQuery (global query)**
  eutils.ncbi.nlm.nih.gov/entrez/eutils/egquery.fcgi
  - Returnează înregistrările ce corespund unei interogări de tip text

- **ESpell (spelling suggestions)**
  eutils.ncbi.nlm.nih.gov/entrez/eutils/espell.fcgi
  - regăsește sugestii de formulare pentru o interogare de tip text

Baze de date biologice

- BD secundare
  - Contin adnotări referitoare la:
    - rolul funcțional
    - structura
    - asocieri cu maladii
    - similarități cu alte secvente
    - referințe bibliografice
  - adnotările sunt avizate de catre specialisti in domeniu
  - UniProt=SWISS-PROT+TrEMBL+PIR
- Tendința curentă: interconectarea tuturor bazelor de date
  - Dificultate: incompatibilități între formate (unele sunt constituite din fișiere text nerelaționate, altele sunt relaționale sau orientate obiect)
  - Soluții: XML, CORBA

Baze de date biologice

Databases and Retrieval Systems	Brief Summary of Content	URL
AceDB	Genome database for *Caenorhabditis elegans*	www.acedb.org
DDBJ	Primary nucleotide sequence database in Japan	www.ddbj.nig.ac.jp
EMBL	Primary nucleotide sequence database in Europe	www.ebi.ac.uk/embl/index.html
Entrez	NCBI portal for a variety of biological databases	www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
ExPASY	Proteomics database	http://us.expasy.org/
FlyBase	A database of the *Drosophila* genome	http://flybase.bio.indiana.edu/
FSSP	Protein secondary structures	www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html
GenBank	Primary nucleotide sequence database in NCBI	www.ncbi.nlm.nih.gov/Genbank
HIV databases	HIV sequence data and related immunologic information	www.hiv.lanl.gov/content/index
Microarray gene expression database	DNA microarray data and analysis tools	www.ebi.ac.uk/microarray
OMIM	Genetic information of human diseases	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
PIR	Annotated protein sequences	http://pir.georgetown.edu/pirwww/pirhome3.shtml
PubMed	Biomedical literature information	www.ncbi.nlm.nih.gov/PubMed
Ribosomal database project	Ribosomal RNA sequences and phylogenetic trees derived from the sequences	http://rdp.cme.msu.edu/html
SRS	General sequence retrieval system	http://srs6.ebi.ac.uk
SWISS-Prot	Curated protein sequence database	www.ebi.ac.uk/swissprot/access.html
TAIR	Arabidopsis information database	www.arabidopsis.org

Baze de date biologice

- Probleme (potențiale):
  - Prezența erorilor în datele primare
    - In cazul secvențelor ADN acestea sunt cauzate in principal de secvențiere (uneori sunt “contaminate” cu secvențe provenind de la vectorii de clonare) – apar în special la secvențele înregistrate înainte de 1990
  - Redundanță mare (în special in BD primare)
    - Cauza: management deficitar al BD
    - Varianta neredundantă: RefSeq (NCBI) – secvențele identice provenite de la același organism sunt combinate
  - Prezența erorilor în adnotări (ex: aceeași genă referită prin nume diferite)
    - Cauze: opinii contradictorii ale cercetătorilor, erori la tehnoredactare
    - Soluție: dezvoltarea unui sistem de asignare consistentă si neambiguă a numelor (ex; GeneOntology)

Baze de date biologice

- Sisteme de regăsire a informației
  - Scop: asigură acces facil la informațiile din bazele de date biologice
  - Exemple:
    - ENTREZ (NCBI – National Center for Biotechnology Information)
      - Permite: căutare pe bază de text (inclusiv la nivelul adnotărilor)
      - Realizează o integrare a informațiilor provenite de la diverse baze de date (ex: pe pagina de la secvențe ADN se găsesc link-uri către secvența de aminoacizi corespunzătoare sau către literatura din PubMed - http://www.ncbi.nlm.nih.gov/pubmed)
    - Sequence Retrieval Systems (SRS)
- Motoarele de căutare corespunzătoare se bazează pe:
  - Algoritmi specifici căutării de informații de tip text
  - Algoritmi specifici identificării potrivirii între secvențe (ex: BLAST)
- Exemplu: FindZebra – pt boli rare (http://findzebra.compute.dtu.dk/) – R. Dragusin, P. Petcu

Baze de date biologice

- Formate de reprezentare a datelor
  - GenBank
  - FASTA

- GeneBank: fișierele cu informații conțin trei secțiuni:
  - Header: descrie originea secvenței și identificarea organismului
  - Features: conține adnotări despre genă și despre semnificația biologică a regiunii corespunzătoare secvenței
  - Sequence entry

Baze de date biologice

GenBank Header

```
LOCUS       Q9ZGE9                     440 aa            linear   BCT 15-JUN-2002
DEFINITION  Light-independent protochlorophyllide reductase subunit N (LI-POR
            subunit N) (DPOR subunit N).
ACCESSION   Q9ZGE9
VERSION     Q9ZGE9  GI:18203677
DBSOURCE    swissprot: locus BCHN_HELMO, accession Q9ZGE9;
            class: standard.
            created: Oct 16, 2001.
            sequence updated: Oct 16, 2001.
            annotation updated: Jun 15, 2002.
            xrefs: gi: 3820536, gi: 3820556
            xrefs (non-sequence databases): InterProIPR000510, PfamPF00148
KEYWORDS    Photosynthesis; Bacteriochlorophyll biosynthesis; Oxidoreductase.
SOURCE      Heliobacillus mobilis
  ORGANISM  Heliobacillus mobilis
            Bacteria; Firmicutes; Clostridia; Clostridiales; Heliobacteriaceae;
            Heliobacillus.
REFERENCE   1  (residues 1 to 440)
  AUTHORS   Xiong,J., Inoue,K. and Bauer,C.E.
  TITLE     Tracking molecular evolution of photosynthesis by characterization
            of a major photosynthesis gene cluster from Heliobacillus mobilis
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 95 (25), 14851-14856 (1998)
  MEDLINE   99061957
   PUBMED   9843979
  REMARK    SEQUENCE FROM N.A.
COMMENT     ---------------------------------------------------------------
            This SWISS-PROT entry is copyright. It is produced through a
            collaboration between the Swiss Institute of Bioinformatics and
            the EMBL outstation - the European Bioinformatics Institute.
            The original entry is available from http://www.expasy.ch/sprot
            and http://www.ebi.ac.uk/sprot
            ---------------------------------------------------------------.
            [FUNCTION] Uses Mg-ATP and reduced ferredoxin to reduce ring D of
            protochlorophyllide (Pchlide) to form chlorophyllide a (Chlide) (By
            similarity). This reaction is light-independent.
            [PATHWAY] Light-independent bacteriochlorophyll biosynthesis.
            [SUBUNIT] Protochlorophyllide reductase is thought to be composed
            of three subunits; bchL, bchN and bchB. Could form a heterotetramer
            of two bchB and two bchN subunits.
            [SIMILARITY] BELONGS TO THE BCHN / CHLN FAMILY.
```

Baze de date biologice

```
FEATURES             Location/Qualifiers
     source          1..440
                     /organism="Heliobacillus mobilis"
                     /db_xref="taxon:28064"
     gene            1..440
                     /gene="BCHN"
     Protein         1..440
                     /gene="BCHN"
                     /product="Light-independent protochlorophyllide reductase
                     subunit N"
                     /EC_number="1.18.-.-"
ORIGIN
        1 merverengc fhtfcpiasv awlhrkikds fflivgthtc ahfiqtaldv mvyahsrfgf
       61 avleesdlvs aspteelgkv vqqvvdewhp kvifvlstcs vdilkmdlev sckdlstrfg
      121 fpvlpastsg idrsftqged avlhallpfv pkeapavepv eekkprwfsf gkesekekae
      181 parnlvliga vtdstiqglq welkqlglpk vdvfpdgdir kmpvineqtv vvplqpylnd
      241 tlatirrerr akvlstvfpi gpdgtarfle aiclefgldt srikekeaqa wrdlepqlqi
      301 lrgkkimflg dnllelplar fltscdvqvv eagtpyihsk dlqqelellk erdvrivesp
      361 dftkqlqrmq eykpdlvvag lgicnpleam gfttawsief tfaqihgfvn aidliklftk
      421 pllkrqalme hgwaeagwle
//
```

Baze de date biologice

- FASTA:
  - unul dintre cele mai simple și populare formate de descriere a secvențelor
  - Ușor de prelucrat și recunoscut de majoritatea aplicațiilor de analiză a secvențelor biologice (inclusiv de către sisteme de software științific cum este MatLab sau Mathematica)
  - Structura:
    - Antet: linie care începe cu simbolul > urmată de o secvența de nume
    - Conținut: linii cu 60-80 caractere
  - Dezavantaj: nu reprezintă toate informațiile care adnotează secvența

Baze de date biologice

```
>gi|18203677|sp|Q9ZGE9|BCHN
MERVERENGCFHTFCPIASVAWLHRKIKDSFFLIVGTHTCAHFIQTALDVMVYAHSRFGFAVLEESDLVS
ASPTEELGKVVQQVVDEWHPKVIFVLSTCSVDILKMDLEVSCKDLSTRFGFPVLPASTSGIDRSFTQGED
AVLHALLPFVPKEAPAVEPVEEKKPRWFSFGKESEKEKAEPARNLVLIGAVTDSTIQQLQWELKQLGLPK
VDVFPDGDIRKMPVINEQTVVVPLQPYLNDTLATIRRERRAKVLSTVFPIGPDGTARFLEAICLEFGLDT
SRIKEKEAQAWRDLEPQLQILRGKKIMFLGDNLLELPLARFLTSCDVQVVEAGTPYIHSKDLQQELELLK
ERDVRIVESPDFTKQLQRMQEYKPDLVVAGLGICNPLEAMGFTTAWSIEFTFAQIHGFVNAIDLIKLFTK
PLLKRQALMEHGWAEAGWLE
```

Cautarea in BD biologice

- Problema:
  - pornind de la o secvența de interogare să se determine toate secvențele din baza de date cu care este “suficient de similara”
- Motivație:
  - Permite să se identifice rolul unor secvențe (gene) nou descoperite
- Tehnici:
  - Exacte: (bazate pe programare dinamică) – ineficiente
  - Euristice: bazate pe algoritmi de potrivire aproximativă și modele statistice

Cautarea in BD biologice

- Caracteristici ale unei metode de căutare:
  - Sensitivitate: capacitatea de a identifica cât mai multe dintre potrivirile reale (“true positive cases”)
  - Specificitate: capacitatea de a elimina potrivirile false (“false positive cases”)
  - Eficiența: furnizarea rezultatelor în timp util -> necesitatea utilizării tehnicilor euristice
- Obs:
  - caracteristicile sunt conflictuale (o creștere a sensitivității conduce de obicei la o scădere a specificității și invers)

Cautarea in BD biologice

- Tehnici euristice:
  - Explorează doar o parte a spațiului de cautare => mai rapide decât programarea dinamica
  - Nu garantează gasirea solutiei optimale

- Exemple:
  - FASTA
  - BLAST

- Ideea de bază (reminder):
  - Pornesc de la potriviri ale unor subsecvențe de lungime mică (cuvinte)
  - Incearca să extindă potrivirile
  - Reunesc regiunile “adiacente” cu scor mare de potrivire

FASTA vs. BLAST

- Similarități:
  - folosesc aceeași idee euristica de a porni de la potrivirea unor secvențe mici și de a le extinde
- Diferențe:
  - In faza de identificare a secvențelor de pornire (“seeds”):
    - BLAST folosește of matrice scor => potrivirea nu trebuie să fie neapărat exactă
    - FASTA folosește o tabelă de hashing => potrivire exactă pt. k-tuple
  - FASTA are sensitivitate mai mare decât BLAST dar BLAST are specificitate mai mare decât FASTA
  - BLAST este mai rapid decât FASTA
  - BLAST poate furniza mai multe potriviri de scor mare; FASTA returneaza doar alinierea finală

Platforme/biblioteci pt programare in bioinformatica

- Extensii ale limbajelor de programare
  - BioJava
  - BioPython
  - BioPerl
  - BioC#
  - BioRuby
- Pachete/ extensii ale tool-urilor pentru calcul științific, statistică sau analiza datelor
  - Matlab toolbox
  - Bioconductor (pachet R)
  - BioWeka (extensie Weka)

BioJava

- http://biojava.org/wiki/Main_Page
- BioJava = framework Java pentru prelucrarea datelor biologice care permite facilitarea dezvoltării rapide a aplicațiilor bioinformatice
- Caracteristici:
  - Proiect open-source găzduit de Open Bioinformatics Foundation (http://www.open-bio.org) – similar cu BioPerl, BioPython, BioRuby și Emboss
  - Inițiat în 2000 și la care au participat peste 60 de dezvoltatori
  - BioJava 3.0 constă din module independent dezvoltate folosind Maven (http://maven.apache.org)
- Conține module pentru
  - parsarea principalelor formate de fișiere utilizate în bioinformatică
  - Aliniere de perechi de secvențe și aliniere multiplă
  - Analiza proprietăților aminoacizilor
  - Detectarea modificărilor în structura proteinelor

BioJava

- Caracteristici BioJava 3.0
- Secvențele sunt definite ca interfețe generice dar există și clase specifice pentru tipurile comune de secvențe
- Conține module de conversie între diferite tipuri de secvențe care încorporează elemente specifice și detalii de natură biologică
- Pentru a minimiza consumul de memorie stocarea se bazează pe conceptul de proxy storage
- Module pentru reprezentarea și manipularea structurilor biomoleculare tridimensionale
- Module pentru alinierea secvențelor bazate pe algoritmi eficienți
- Module pentru accesarea serviciilor Web pentru bioinformatică prin protocoale de tip REST (ex: NCBI Blast prin the Blast URLAPI și HMMER prin http://hmmer.janelia.org/)

BioPython

- http://biopython.org/wiki/Main_Page
- Biopython = set de instrumente pentru prelucrarea datelor biologice
- Proiect găzduit de Open Bioinformatics Foundation
- Permite:
  - Parsarea principalelor formate de fișiere
  - Operații cu secvențe (conversii +alinieri)
  - Interfețare cu Blast (NCBI) și Clustalw
  - Integrarea cu BioSQL = a schema de baze de date pentru secvențe (suportată și de către BioJava și BioPerl)
  - Clasificarea datelor (folosind kNN, Naive Bayes, Support Vector Machines)

BioPerl

- http://www.bioperl.org/wiki/Main_Page
- BioPerl = toolkit pentru prelucrarea datelor biologice
- Caracteristici:
  - Centrat pe manipularea (conversia) datelor (cu accent mai puțin pe algoritmii de prelucrare)
  - Module pentru preluare date, analize statistice simple, identificare de șabloane descrise prin expresii regulate, conectare la baze de date

- http://biohaskell.org/

- BioHaskell = bibliotecă implementată în Haskell pentru analiza datelor biologice
- Caracteristici:
  - Funcții pentru alinierea secvențelor
  - Funcții pentru estimarea structurii secundare a ARN

BioPHP

- http://genephp.sourceforge.net/

- BioPHP= extensie PHP pentru bioinformatică
- Caracteristici:
- Citire date biologice în formatele: GenBank, Swissprot, Fasta, Clustal ALN
- Asigură navigare prin date stocate în mai multe fișiere
- Prelucrări simple asupra secvențelor (conversii, căutare de șabloane descrise prin expresii regulate, construire secvențe consensuale etc.)
- Aliniere simplă și multiplă
- Interfațare cu alte programe (ex: Clustal)

BioRuby

- http://www.bioruby.org/
- BioRuby = set de instrumente și biblioteci free pentru bioinformatică și biologia moleculară
- Conține componente pentru:
  - Analiza secvențelor
  - Analiza filogenetică
  - Modelarea proteinelor
  - Analiza căilor de reglare metabolică
- Conține suport pentru:
  - Majoritatea formatelor utilizate în bioinformatică
  - Accesarea bazelor de date biologice
  - Accesarea serviciilor web publice (BLAST, KEGG, GenBank, MEDLINE and GO).

Bio C#, .NET Bio

- http://www.kofler.or.at/bioinformatics/biosharp.html
- Bio C# = bibliotecă de clase pentru bioinformatică
- Bio C# conține clase pentru:
  - Preluare date în format FASTA
  - Căutare folosind Blast
  - Aliniere locală și globală alignments
  - Prelucrări statistice

- http://bio.codeplex.com/
- .NET Bio = bibliotecă open-source de clase pentru bioinformatică
- .NET Bio conține:
  - Parsare diferite tipuri de fișiere
  - Conectori la servicii web specifice (NCBI BLAST)
  - Algoritmi standard pentru compararea/ alinierea secventelor

Bioclipse

- http://www.bioclipse.net/

- Bioclipse = platformă open source workbench științele vieții.
- Caracteristici:
  - Bazată pe Eclipse Rich Client Platform (RCP)
  - Moștenește arhitectura de plugin-uri, funcționalitatea și interfețele vizuale din Eclipse
  - Oferă prelucrări specifice pentru: chemoinformatică, bioinformatică, web semantic, analiză spectrală, design de medicamente etc.

Bio4J

- http://bio4j.com/

- Bio4j = bază de date orientată pe structuri de tip graf care permit descrierea structurii proteinelor
- Include date disponibile în UniProt KB (SwissProt + Trembl), Gene Ontology (GO), UniRef (50,90,100), RefSeq, NCBI taxonomy și Expasy Enzyme DB.
- Framework pentru accesarea si gestiunea informatiilor despre proteine
- Datele sunt reprezentate astfel încât structura proteinelor sa fie descrisa în manieră semantică

BioConductor, BioWeka

- http://www.bioconductor.org/
- Bioconductor = pachet R (open-source) pentru analiza datelor genomice
- http://sourceforge.net/projects/bioweka/
- BioWeka = extensie pentru bioinformatică a pachetului Weka (data mining)
- Caracteristici:
  - Permite citirea formatelor standard
  - Conține filtre specifice pentru prelucrarea datelor biologice (inclusive adnotări)
  - Are implementați algoritmi de aliniere (inclusiv BLAST)