Biostatistica si Bioinformatica.

Teme pentru proiecte (2016)

I. Implementarea unor algoritmi.

1. Implementare algoritm pentru Partial Digest Problem (la alegere între variante de tip backtracking/branch and bound/programare dinamică)

   Biblio:curs 4, http://www.bioalgorithms.info/presentations/Ch04_DNA_mapping.pdf + lucrari din directorul proiecte/biblio/PartialDigestProblem)

2. Implementare algoritm branch and bound pentru identificarea șabloanelor

   Biblio: curs 5, http://www.bioalgorithms.info/presentations/Ch04_Motifs.pdf + lucrari din directorul proiecte/biblio/MotifsBranch&Bound

3. Implementare algoritm branch and bound pentru "median string search" (curs 5, http://www.bioalgorithms.info/presentations/Ch04_Motifs.pdf)

4. Implementare algoritm de tip greedy pentru identificare șabloane (CONSENSUS) (curs 5, http://www.hku.hk/bruhk/gcgdoc/consensus.htm, http://www.hku.hk/bruhk/gcgdoc/fitconsensus.html) + lucrari din directorul proiecte/biblio/Consensus)

5. Implementarea unui algoritmi probabilişti pentru identificarea șabloanelor (ex: MEME) (http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18047721) + lucrari din directorul proiecte/biblio/MEME

6. *Aliniere globală.* Implementare algoritm Needleman-Wunsch. Facilități minimale: selecție matrice cu scoruri de substituție (specifica pt nucleotide/aminoacizi), penalizare liniară a gap-urilor. Facilitate suplimentară (pcte bonus): penalizare afină a gap-urilor.

   Biblio: curs 6 + lucrari din directorul proiecte/biblio/NeedlemanWunsch
   http://www.maths.tcd.ie/~lily/pres2/sld013.htm,
   http://www.ludwig.edu.au/course/lectures2005/Likic.pdf,
   http://media.wiley.com/product_data/excerpt/91/04708483/0470848391.pdf)

7. *Aliniere locală.* Implementare algoritm Smith-Waterman. Facilitati minimale: selectie matrice cu scoruri de substitutie (specifica pt nucleotide/aminoacizi), penalizare liniară a gap-urilor. Facilitate suplimentară (pcte bonus): penalizare afină a gap-urilor.

   Biblio: curs 6 + proiecte/biblio/SmithWaterman
   http://www.maths.tcd.ie/~lily/pres2/sld013.htm,
   http://jaligner.sourceforge.net/

8. *Alinierea extremităților (overlap alignment).* Se folosește în principal pentru a determina alinierea dintre o porțiune finală a unei secvențe și o porțiune initială a altei secvențe. Spre deosebire de alinierea globală clasică în matricea de scor corespunzătoare alinierii extremităților prima linie și prima coloană au valoarea 0 (în felul acesta nu se penalizează

gap-urile de la extremități). In plus, daca in alg Needleman Wunsch clasic la construirea alinierii se pornește de la cea mai mare valoare din matricea de scor în alinierea extremităților se pornește de la cea mai mare valoare aflată fie pe ultima coloană fie pe ultima linie. Se cere implementarea unei variante a algoritmului Needleman Wunsch pt cazul alinierii extremităților.

Biblio: proiecte/biblio/OverlapAlignment

9. *Aliniere multiplă.* Extinderea directă a algoritmului Needleman-Wunsch pentru cazul a 3 secvențe de nucleotide + penalizare liniară a gap-urilor (programare dinamica 3D)
   Biblio: curs 8 + lucrări din directorul proiecte/biblio/MultipleAlignment+DP

10. Implementarea unui algoritm de grupare si testarea lui pentru date biologice. Variante:

    - Algoritm partițional (ex: k-means, ISODATA)
    - Algoritm ierarhic aglomerativ + vizualizare dendrograma

Biblio: http://www.bioalgorithms.info/presentations/Ch10_Clustering.pdf
http://www.cs.umd.edu/hcil/bioinfovis/hce.shtml
+ lucrari director proiecte/biblio/Clustering

11. Implementarea unui algoritm pentru biclustering (ex. Algoritmul Cheng-Church)
    Biblio: lucrari director proiecte/biblio/Biclustering

12. Implementarea unui algoritm pentru construirea arborilor filogenetici (la alegere intre algoritmul UPGMA si algoritmul NeighbourJoining).
    Biblio: lucrari director proiecte/biblio/PhyloTrees

II. Analize comparative intre pachete software utilizate in bioinformatica

13. *Aliniere multiplă.* Analiza comparativă a cel puțin două implementări curente la alegere dintre:

    - MAFFT (http://align.bmr.kyushu-u.ac.jp/mafft/software/)
    - MUSCLE (http://www.drive5.com/muscle/)
    - CLUSTALW (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/)
    - DIALIGN (http://www.gsf.de/biodv/dialign.html)
    - T-COFFEE (http://igs-server.cnrs-mrs.fr/%7Ecnotred/Projects_home_page/t_coffee_home_page.html)
    - Kalign (http://msa.cgb.ki.se/)
    - Jalview (editor pentru aliniere multipla) http://www.jalview.org/)
    - MUSCA (http://cbcsrv.watson.ibm.com/Tmsa.html)

14. Studiu comparativ a unor instrumente software pentru identificarea șabloanelor. La alegere două dintre:
    - BLOCKS (Multiple alignments of conserved regions) http://blocks.fhcrc.org/
    - CDD: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

- eMOTIF: http://motif.stanford.edu/emotif/
- Pfam: http://www.sanger.ac.uk/Software/Pfam/
- PRINTS: http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
- ProDom: http://www.toulouse.inra.fr/prodom.html
- PROSITE: http://www.expasy.org/prosite
- ProtoMap: http://protomap.cornell.edu
- Motif Scan: http://hits.isb-sib.ch/cgi-bin/PFSCAN
- MEME: http://meme.sdsc.edu/meme4_3_0/intro.html)
- CONSENSUS: http://bifrost.wustl.edu/consensus/index.html

15. Studiu comparativ al unor instrumente software pentru analiza structurii proteinelor. La alegere două dintre:

    - ASTRAL http://astral.stanford.edu/
    - PDB http://www.pdb.org/
    - SCOP http://scop.mrc-lmb.cam.ac.uk/scop
    - MMDB http://www.ncbi.nlm.nih.gov/Structure

16. Analiza facilităților oferite de (la alegere):
    - BioWeka – varianta pentru bioinformatică a pachetului de data mining Weka (http://Bioweka.sourceforge.net)
    - Bioconductor - componenta pt bioinformatică din pachetul statistic R (http://www.bioconductor.org/) (http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/)

17. Studii comparative între biblioteci pentru bioinformatică din diferite limbaje de programare (la alegere două dintre ele):
    - BioJava (http://www.biojava.org/wiki/Main_Page)
    - BioPerl (http://bio.perl.org/wiki/Main_Page)
    - BioPython (http://biopython.org/wiki/Main_Page)
    - BioC# (http://www.kofler.or.at/bioinformatics/biosharp.html), .NetBio (http://bio.codeplex.com/)
    - BioRuby (http://bioruby.org/)

III. Studii bibliografice + exemple ilustrate utilizând software existent

18. *Algoritmi euristici pentru căutare în baze de date biologice. FASTA*: descriere algoritm, prezentare pachete software care conțin implementarea algoritmului + exemple de utilizare. Obs: pt implementarea algoritmului se obține un bonus de 2 pcte.

    Biblio:
    Lucrari in directorul Proiecte/biblio/FASTA/
    http://www-bimas.cit.nih.gov/fastainfo/fasta_algo
    http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

19. *Algoritmi euristici pentru căutare în baze de date biologice. BLAST*: descriere algoritm, prezentare pachete software care conțin implementarea algoritmului + exemple de utilizare. Obs: pt implementarea algoritmului se obtine un bonus de 2 pcte.

Biblio:
Lucrari in directorul Proiecte/biblio/Blast/
http://pbil.univ-lyon1.fr/alignment.html

20. *Single Nucleotide Polymorphism.* Se referă la variabilitatea în genomul unor indivizi aparținând aceleiași specii manifestată prin valori diferite ale unor nucleotide izolate. Identificarea SNP-urilor și utilizarea lor în extragerea de informații din secvențe face obiectul mai multor cercetări în informatică. Scopul proiectului este trecerea în revistă a unor astfel de studii.

    Biblio:
    Lucrari in directorul Proiecte/biblio/SNP/

21. *HMM – Hidden Markov Models.* Modelele Markov cu stări ascunse permit modelarea structurilor cu caracter dinamic (fiind utilizate pentru recunoașterea vorbirii) putând fi astfel utilizate pentru analiza secvențelor biologice.

    Biblio: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2766791/
    http://compbio.soe.ucsc.edu/ismb99.handouts/KK185FP.html
    Lucrări în directorul proiecte/biblio/HMM

22. *Protein Folding.* Determinarea structurii terțiare a proteinelor pornind de la secvențe de aminoacizi se bazează pe modele geometrice de plasare a componentelor și algoritmi de optimizare având ca scop minimizarea energiei asociată structurii. Scopul proiectului este trecerea în revistă a principalelor modele laticiale și a unor algoritmi aleatori (de tip Monte Carlo) utilizați în protein folding.

    Biblio: Lucrări în directorul Proiecte/biblio/ProteinFolding/

23. *Networks Motifs.* Identificarea motivelor (blocuri structurale) în rețelele de reglare genică reprezintă a problemă de interes curent care necesită aplicarea unor algoritmi de optimizare. Scopul proiectului este trecerea în revistă a problematicii analizei rețelelor complexe și a unor algoritmi de identificare a structurilor în rețele biologice.

    Biblio: Lucrări în directorul Proiecte/biblio/NetworkMotifs/

24. Studiu bibliografic al bazelor de date si instrumentelor software utilizate in analiza expresiei genice (la alegere două):

    - ArrayExpress http://www.ebi.ac.uk/arrayexpress
    - Mouse Atlas and Gene Expression Database: http://genex.hgu.mrc.ac.uk/
    - NetAffx http://www.affymetrix.com/
    - Stanford Microarray Database http://genome-www.stanford.edu/microarray/
    - KEGG http://www.genome.ad.jp/kegg/
    - Klotho http://www.ibc.wustl.edu/klotho/

25. Studiu bibliografic al utilizarii ontologiilor in bioinformatică (la alegere două dintre variante)

- Gene ontology (www.geneontology.org/)
- Ontology based knowledge representation (http://www.cs.man.ac.uk/~stevensr/onto/)
- Bio-ontologies links (http://anil.cchmc.org/Bio-Ontologies.html)
- Ontology for bioinformatics applications (http://bioinformatics.oxfordjournals.org/cgi/reprint/15/6/510)

IV. **Tema la alegere** – singura conditie este sa fie corelata cu problem specifice bioinformaticii si sa fie similara ca ordin de complexitate cu temele anterioare. Bibliografia va fi identificata de catre student (cel putin 3 lucrari recente referitoare la tematica).

Indicații pentru pregătirea proiectului. Fiecare proiect trebuie să conțină:

1. Un **referat** în care este descrisă problema/metoda/algoritmul studiat/analizat/implementat având structura următoare:
   - Titlu, autor
   - Rezumat de 5-15 rânduri în care se descrie pe scurt conținutul referatului
   - Introducere – în care se prezintă problema abordată, se trec în revistă obiectivele urmărite, se prezintă pe scurt ce s-a obținut și se descrie structura referatului
   - Prezentarea detaliată a metodei/algoritmului/instrumentului software
   - Descrierea contribuției personale:
     - In cazul implementărilor – constă dintr-o scurtă descriere a implementării și exemple de rulare
     - In cazul studiilor comparative constă în rezultate obținute rulând algoritmii selectați și comentarii privind comportarea lor; pot fi menționate și observații referitoare la modul de instalare al programului (dacă e cazul)
     - In cazul studiilor bibliografice constă în identificarea particularităților, avantajelor și dezavantajelor metodelor/sistemelor studiate și în trecerea în revistă a problemelor considerate încă deschise
   - Concluzii: se prezintă succinct principalele rezultate obținute
   - Bibliografie: se menționează toate sursele bibliografice folosite (pe lânga cele sugerate în lista de teme trebuie identificate cel puțin 3 lucrări care abordează tematica respectivă). Pentru fiecare lucrare se menționează: autori, titlu, revista/editura, pag, an. Pentru resursele electronice se menționeaza link-ul. Toate lucrările menționate în bibliografie trebuie referite în text în contextul adecvat.
2. Parte aplicativă:
   - In cazul implementărilor: aplicația (poate fi realizată în orice limbaj de programare, inclusiv utilizând R)
   - In cazul studiilor comparative: datele de test utilizate și rezultatele rulărilor
   - In cazul studiilor bibliografice: un exemplu de utilizare a tehnicii/sistemului studiat
3. Slide-uri corespunzătoare unei prezentări de cca 10 minute.

Notare:

Proiect+activitate laborator: 80%
Test scris tip grila (cu acces la materialele bibliografice): 20%

Punctajul maxim pentru un proiect este:

- 10p – pentru proiecte care conțin implementări personale
- 9p – pentru studii comparative
- 8p – pentru studii bibliografice

Proiectul va fi prezentat la examen. Este necsar sa pregătiți cateva slide-uri (de exemplu in PowerPoint) care să fie suportul prezentarii.

Restul de puncte (până la 10) se poate obține din implementarea algoritmilor analizați și/sau din exercițiile/temele de la laborator. Exercițiile/temele de la fiecare laborator valoreaza 0.5p. Cei care nu au fost la laborator pot să îmi trimită rezolvările (ca fisiere R).