Bioinformàtica (20428)

Titulació/estudi: Grau en Biologia Humana
Curs: 4t.
Trimestre: 1r.
Nombre de crèdits ECTS: 6 crèdits
Hores de dedicació de l'estudiant: 150 hores
Llengua o llengües de la docència: Català/Anglès
Professorat: Roderic Guigó, Robert Castelo, Cedric Notredame i Toni Gabaldon

1. Presentació de l'assignatura

Els eixos sobre els quals s'articula l'assignatura Bioinformàtica són els següents:

1. El contingut de l'assignatura se centra exclusivament en l'anàlisi computacional de seqüències biològiques. Altres àrees molt importants de la bioinformàtica –com ara l'anàlisi de microarrays, modelització de xarxes metabòliques o de regulació o *data-mining* de la literatura científica– en són explícitament excloses del temari. A més, a l'hora d'establir el temari concret, s'ha tingut en compte que les assignatures Evolució (anterior en la llicenciatura) i Biologia Estructural (posterior) inclouen continguts propis de l'anàlisi de seqüències: construcció de filogènies moleculars i predicció de l'estructura de les proteïnes. Aquests continguts han estat també explícitament exclosos del temari.

2. L'assignatura no consisteix en una mera descripció de receptes algorítmiques per tal de resoldre determinats problemes, sinó que el nucli fonamental de les classes teòriques consisteix en la justificació formal d'alguns dels algoritmes més utilitzats en bioinformàtica. En particular, la programació dinàmica (en què es basen els programes d'alineament de seqüències), les taules "Hash" (que permeten l'acceleració de les recerques de similaritat en les bases de dades), i els mètodes markovians –incloent-hi els models de Markov Ocults (per a la identificació de patrons en seqüències)– són tots introduïts, –i fem èmfasi en el terme *introduïts*– amb el rigor matemàtic necessari.

3. L'assignatura no pretén embotir els estudiants amb un gran nombre de coneixements (programes, servidors i bases de dades diversos). Aquesta és una temptació en la qual, en bioinformàtica, es pot caure fàcilment, donada la impressionant quantitat de recursos que hi ha a Internet. Tanmateix, una temptació que, donat el caire extraordinàriament canviant dels problemes en bioinformàtica, pot ser molt poc profitosa: els problemes rellevants avui ho deixen de ser demà. Contràriament –i d'acord amb l'orientació general de la llicenciatura– l'assignatura, més que informar els estudiants sobre com resoldre problemes que avui existeixen, pretén capacitar-los perquè siguin capaços de fer front a problemes, alguns dels quals, potser avui encara no existeixen. És per això, que el nucli de les classes pràctiques és constituït per una introducció a la

programació. El llenguatge escollit és Perl, un llenguatge interpretat. Això fa que les hores dedicades a programació siguin suficients per tal que els estudiants més motivats puguin implementar programes força sofisticats.

4. Els mètodes bioinformàtics no es desenvolupen aïlladament, sinó que ho fan en resposta a determinats problemes biològics. Aquests, al seu torn, són sovint motivats per desenvolupaments tecnològics. Per exemple, els algoritmes de comparació de seqüències es desenvolupen quan algú s'adona que d'aquesta comparació s'obté informació biològica molt rellevant. L'obtenció de seqüències de manera massiva, que dóna sentit en aquesta comparació però que depèn del perfeccionament de les tècniques de seqüenciació. En aquest sentit, l'assignatura intenta introduir els mètodes bioinformàtics en el context històric del progrés de la biologia que els fa necessaris.

Dels eixos en els quals s'organitza l'assignatura, es dedueix que l'objectiu principal és **proporcionar als estudiants la capacitat i les habilitats per tal que siguin capaços de resoldre (nous) problemes computacionals en biologia**. Aquesta capacitació s'obté ensenyant els fonaments teòrics d'alguns dels algoritmes més importants en bioinformàtica i ensenyant a programar. En concret els objectiu generals de l'assignatura són:

1. Educar els estudiants de biologia en la comprensió i la utilització dels mètodes d'anàlisi computacional de les seqüències biològiques.
2. Introduir els estudiants en el camp dels algoritmes i de la computació.
3. Introduir els estudiants en el sistema operatiu UNIX/Linux i en la programació en el llenguatge Perl.

2. Competències que s’han d’assolir

A continuació es descriuen els objectius de l’assignatura Bioinformàtica amb més detall. Els objectius han estat classificats en:

Essencials: allò que tots els estudiants han de saber o saber fer quan acaben la carrera.

Aconsellables: allò que és recomanable que un estudiant sàpiga o sàpiga fer quan acaba la carrera.

Especialitzats: aquells aspectes específics de la bioinformàtica que no són essencials /aconsellables per a altres àrees de coneixement, però sí que són essencials/aconsellables si es pretén fer una especialització de postgrau.

CONEIXEMENTS QUE S'ADQUIRIRAN A L'ASSIGNATURA BIOINFORMÀTICA

ESSENCIALS

1. Conèixer la mitja dotzena de bases de dades més importants en la recerca en biomedicina avui en dia
2. Concepte d'alineament i similaritat de seqüències
3. Distingir alineament local, global i múltiple
4. Concepte de puntuació d'un alineament
5. Conceptes de gap, identitat, substitució conservativa
6. Concepte de Matriu de Substitució
7. Distinció entre alineament òptim mitjançant programació dinàmica, i alineament aproximat
8. Concepte de patró de seqüència
9. Interpretació probabilística de les matrius de pesos posicionals (PWMs)
10. Regularitats estadístiques en les seqüències de DNA que correlacionen amb dominis funcionals. El concepte de periodicitat de la seqüència de DNA
11. Visió global dels mètodes de descobriment de patrons en seqüències
12. Concepte de computació amb DNA
13. Models probabilístics de les seqüències
14. Models de Markov
15. Introducció als Hidden Markov Models
16. Concepte d'algoritme
17. Concepte de programació dinàmica
18. Concepte de hash table

ACONSELLABLES

1. Estructura bàsica d'una base de dades
2. Algoritme de Needleman-Wuncsh
3. Algoritme de Smith-Waterman
4. Interpretació evolutiva de les matrius PAM
5. Concepte de penalització afí dels gaps
6. Obtenció manual d'una alineament de seqüències, amb construcció de la matriu de programació dinàmica
7. Esquematització de l'algoritme bàsic en què es basen els programes BLAST i FASTA
8. Interpretació probabilística de les puntuacions de l'alineament del BLAST
9. Distinció conceptual dels diferents programes de la família BLAST
10. Tractament dels elements repetitius en recerques de similaritat a les bases de dades
11. El concepte de contingut informatiu d'una patró i de sequence logo
12. Relació entre alineament múltiple i patró de seqüència
13. Elements de seqüència característics dels senyals de splicing canònics
14. Bases de dades d'elements promotors de matrius de factors de transcripció
15. Enumeració d'alguns programes de predicció de gens
16. Predicció comparativa de gens
17. Anàlisi comparativa de genomes
18. Mètodes per a la reconstrucció de la història dels genomes

ESPECIALITZATS

1. Enumeració de les bases de dades més utilitzades en biologia i coneixement de la seva estructura
2. Definició formal d'una base de dades. Esquema
3. Equacions de recurrència dels algoritmes de Needleman-Wunsch i Smith-Waterman
4. Concepte de complexitat d'un algoritme
5. Construcció manual d'un alineament múltiple
6. Relació entre alineament múltiple i arbre filogenètic
7. Derivació de la probabilitat d'un hsp en el programa BLAST
8. Reconstrucció del procés d'obtenció d'una matriu PAM i una matriu BLOSUM
9. Descripció de la relació entre els diferents estadístics codificants
10. Transformació de Fourier per al càlcul de la periodicitat d'una seqüència
11. Hidden Markov Models per a la predicció de gens
12. Double Hidden Markov Models
13. Text mining
14. Construcció i anàlisi de xarxes
15. Utilització d'ESTs per a la reconstrucció del patró de splicing alternatiu dels gens
16. Algoritmes per a la reconstrucció de la història del genoma: Hannehahilli-Pevner
17. Mètode d'Addleman per a la computació amb DNA
18. Estructures de dades i flux de control en programació

HABILITATS PRÀCTIQUES QUE S'ADQUIRIRAN A L'ASSIGNATURA BIOINFORMÀTICA

ESSENCIALS	ACONSELLABLES	ESPECIALITZATS
1. Fer ús de programes centralitzats d'accés a les bases de dades: SRS i ENTREZ	1. Familiaritzar-se en la utilització de bases de dades integratives com ara GENECARDS	1. Tenir coneixements bàsics de MYSQL
2. Obtenir un alineament global/-local entre dues seqüències	2. Obtenir alineaments entre seqüències de proteïnes i seqüències de DNA amb GENEWISE	2. Derivar una matriu de substitució a partir d'un conjunt d'alineaments
3. Obtenir un alineament múltiple de seqüències amb CLUSTAL o T-COFFEE	3. Utilitzar programes d'alineament múltiple de genomes complets com ara MAVID	3. Fer servir HMM per identificar una seqüència en una base de dades de perfils
4. Fer recerques de similaritat en bases de dades utilitzant BLAST i FASTA	4. Utilitzar la base de dades TRANSFAC per a la caracterització de les regions promotores dels gens	4. Fer servir HMM per identificar un perfil de HMM entre una base de dades de seqüències
5. Utilitzar una matriu de pesos per a la recerca de motius en seqüències de DNA: llocs de splicing, motius promotors	5. Utilitzar el programa MEME pel descobriment de motius en seqüències funcionalment relacionades	5. Dissenyar un HMM per identificar dominis diferents en seqüències de DNA
6. Utilitzar el programa REPEATMASKER per localitzar repeticions en seqüències de DNA	6. Utilitzar programes d'alineament de DNA amb CDNA: ESTGENOME, SIM4, etc..	6. Implementar un servidor DAS
7. Utilitzar programes de predicció de gens: GENEID, GENSCAN, FGENES...	7. Utilitzar programes diferents per tal de caracteritzar amb detall una regió del genoma	7. Implementar un programa en PERL de complexitat moderada (per exemple un programa basat en Smith-Watermann per l'alineament de seqüències)
8. Accedir als repositoris d'informació genòmica: ENSEMBL , GENOME BROWSER, NCBI	8. Identificar SNPs en seqüències de DNA	8. Utilitzar de forma avançada el sistema UNIX/LINUX
9. Utilitzar servidors per a l'anàlisi de dades de microarrays	9. Utilitzar de forma avançada el programa GENEID	9. Dissenyar un servidor web com a interfície d'algun programa d'anàlisi de seqüències
10. Utilitzar programes de visualització comparativa de genomes: PIPMAKER, VISTA	10. Utilitzar el programa GAZE	
11. Implementar un programa senzill en PERL (per exemple, per tal de determinar si un nombre donat és primer o no)	11. Implementar un programa de més complexitat en PERL (com ara un programa per calcular el biaix en la utilització de codons en una seqüència de DNA)	
12. Utilitzar de forma bàsica el sistema operatiu UNIX/LINUX	12. 13.Familiaritzar-se amb el sistema operatiu UNIX/LINUX, i m algunes de les seves comandes: sh, awk, sed, sort, join, comm, cat, ...	
	13. Elaborar una pàgina web	

3. Continguts

TEORIA

T1. Introducció a la bioinformàtica.

T2. Comparació de seqüències en el context evolutiu.

T3. Comparació de seqüències en el context evolutiu.

T4. Matrius de substitució.

T5. Introducció a la programació dinàmica.

T6. Introducció a la programació dinàmica.

T7. Recerques de similaritat en bases de dades (BLAST).

T8. Recerques de similaritat en bases de dades (BLAST).

T9. Alineament múltiple de seqüències.

T10. Alineament múltiple de seqüències.

T11. Alineament múltiple de seqüències.

T12. Reconeixement de patrons en seqüències.

T13. Reconeixement de patrons en seqüències.

T14. Reconeixement de patrons en seqüències.

T15. Estadístics codificants.

T16. Predicció de gens.

T17. Recerca de selenoproteïnes en organismes eucariotes .

T18. Recerca de selenoproteïnes en organismes eucariotes.

SEMINARIS

S1. Unix I. Introducció al sistema operatiu UNIX**.**

S2. Introducció als algorismes (I)**.**

S3. Unix II. Comandes i manipulacions bàsiques d'arxius a UNIX**.**

S4. Introducció als algorismes (II).

S5. Problemes d'algoritmes (I).

S6. Unix III. Comandes i manipulacions avançades d'arxius a UNIX.

S7. Problemes d'algoritmes (II).

S8. Perl I. Perl bàsic.

S9. Perl II. Perl bàsic.

S10. Perl III. Vectors i processament de múltiples línies en Perl.

S11. Perl IV. Programació dinàmica.

S12. Elaboració de pàgines web.

S13. Recerques de similaritat en bases de dades (BLAST).

S14. Anotació de genomes (I).

S15. Anotació de genomes (II).

S16. Genome Browsers.

S17. Anàlisi de dades produïdes per instruments de seqüenciació de nova generació (I).

S18. Anàlisi de dades produïdes per instruments de seqüenciació de nova generació (II).

PRÀCTIQUES

P1. Introducció al món Viquipèdia. Supervisió de projectes.

P2. Supervisió de projectes.

P3. Supervisió de projectes.

P4. Supervisió de projectes.

P5. Supervisió de projectes.

P6. Supervisió de projectes.

---

4. Avaluació

L'avaluació dels aprenentatges dels estudiants té dos components: un examen teòric i un treball pràctic. A l'assignatura, donem una gran importància al treball pràctic; no només com a mecanisme d'avaluació, sinó també, i potser sobretot, com a part integral del procés d'aprenentatge. És durant el treball pràctic que els estudiants poden aplicar els coneixements i desplegar les habilitats que hem intentat transmetre'ls durant les classes de Bioinformàtica –tant les teòriques com les pràctiques–. És per això que reservem 10 hores de classes pràctiques per tal que els estudiants duguin a terme, sota la supervisió de professors de l'assignatura, una part del seu treball pràctic. Els treballs es desenvolupen en grups de quatre o cinc persones, s'han de presentar en una pàgina web i han de tenir format d'article científic.

El 60% de la qualificació final correspon a l'examen teòric i el 40%, al treball pràctic.

Criteris de superació i qualificacions qualitatives
Per superar l'assignatura, l'estudiant ha d'obtenir una nota de 5 o superior. La realització del treball pràctic és, en qualsevol cas, imprescindible per superar l'assignatura. La superació del 70% dels objectius implica la qualificació de notable i la superació del 90% dels objectius, la d'excel·lent. Hi haurà un nombre de matrícules d'honor proporcional al nombre total d'alumnes matriculats, i es lliuraran a les millors notes finals, sempre que superin el llindar d'excel·lent.

Recuperació de l'assignatura
Aquells alumnes que hagin suspès l'assignatura a l'avaluació ordinària podran recuperar-la al mes de juliol. La recuperació constarà d'un únic assaig amb 4 preguntes de resposta curta i un PEM de 20 preguntes, que es valorarà de la mateixa manera que l'examen ordinari. La realització del treball pràctic és imprescindible, però aquest només pot ser presentat una vegada.

5. Bibliografia i recursos didàctics

5.1. Bibliografia bàsica

- **Bioinformatics**, David M. Mount, Cold Spring Harbour Laboratory Press, 2001.
- **Bioinformatics Computer Skills**, Cynthia Gibas and Per Jambeck, O'Reilly, 2001.

- **Bioinformatics: A practical guide to the analysis of genes and proteins,** Andreas D. Baxevanis and B.F. Francis Oullete eds., John Wiley & Sons, 2005.
- **Bioinformatics for dummies,** Jean-Michel Claverie and Cedric Notredame, Wiley, 2003.

5.2. Bibliografia complementària

5.3. Recursos didàctics

Cada sessió teòrica o pràctica té material audiovisual associat de suport a la docència. Aquests materials són accessibles a través de la web de l'assignatura –que és actualitzada cada any–. L'adreça d'aquesta web és http://bioinformatica.upf.edu i conté informació general sobre l'assignatura i el programa de l'assignatura, a partir del qual s'accedeix als continguts de suport a la docència per a cada lliçó teòrica o pràctica.

El material docent de suport a les lliçons de teoria inclou el guió de la classe, les figures i les taules que seran utilitzades i les referències (enllaços) a altres documents d'interès. En alguns casos inclou documents interactius elaborats per nosaltres. La documentació per a la majoria de les lliçons és en format HTML –i, per tant, accessible des d'Internet amb qualsevol navegador–; però, en alguns casos, en els quals la complexitat matemàtica és substancial, hem preferit utilitzar LaTeX i generar els documents en PDF. En altres, el material docent de suport a la docència és una presentació en PowerPoint.