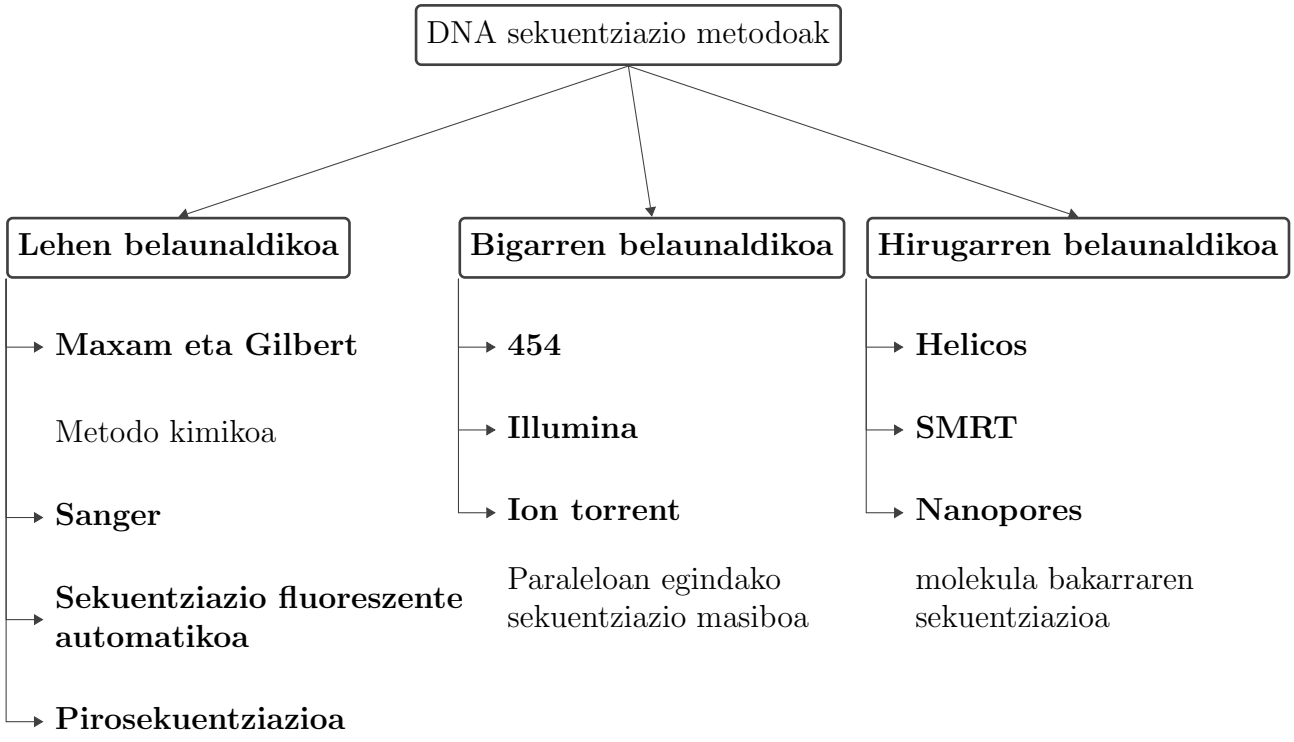
BIOINFORMATIKA LABURPENA

Egilea: Laura I. Sarasola

1 DNA sekuentziazio metodoak

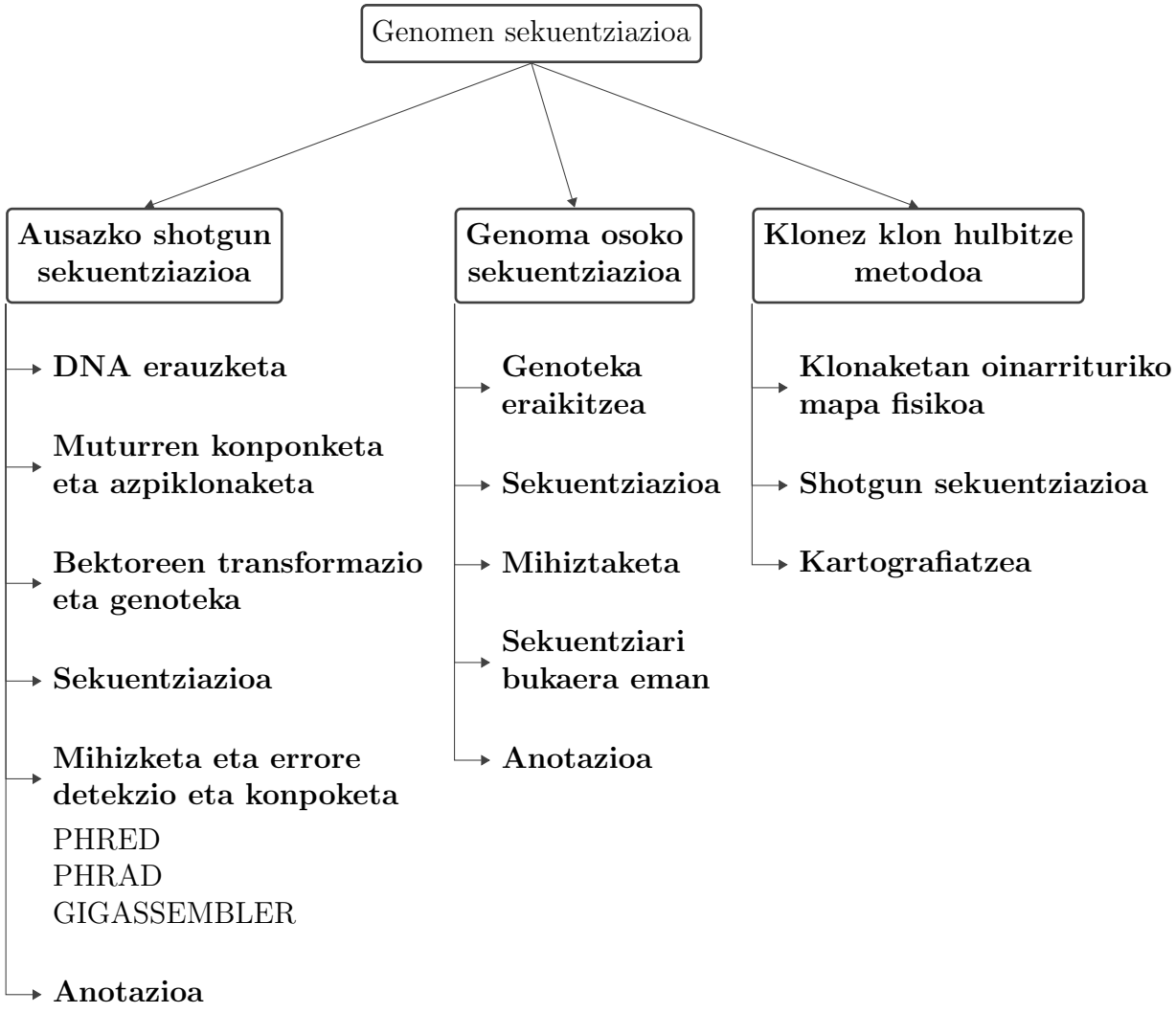


Sequencing by Synthesis (SBS): Sintesian oinarrituriko sekuentziazioa DNA polimerasaren aktibitateaz baliatuz, jatorrizko harizpiaren erreplikazioan oinarrituriko sekuentziazio metodoa da. Jatorrizko sekuentzia sekuentziaturiko harizpiaren osagarria izango da.

Zubi bidezko PCR: Euskarri solidoan genotekaren fragmentuak anplifikatzen dira multzokaturik, anplifikaziorako hasleak euskarrian immobilizaturik daudelarik. Illumina plataforman erabiltzen da.

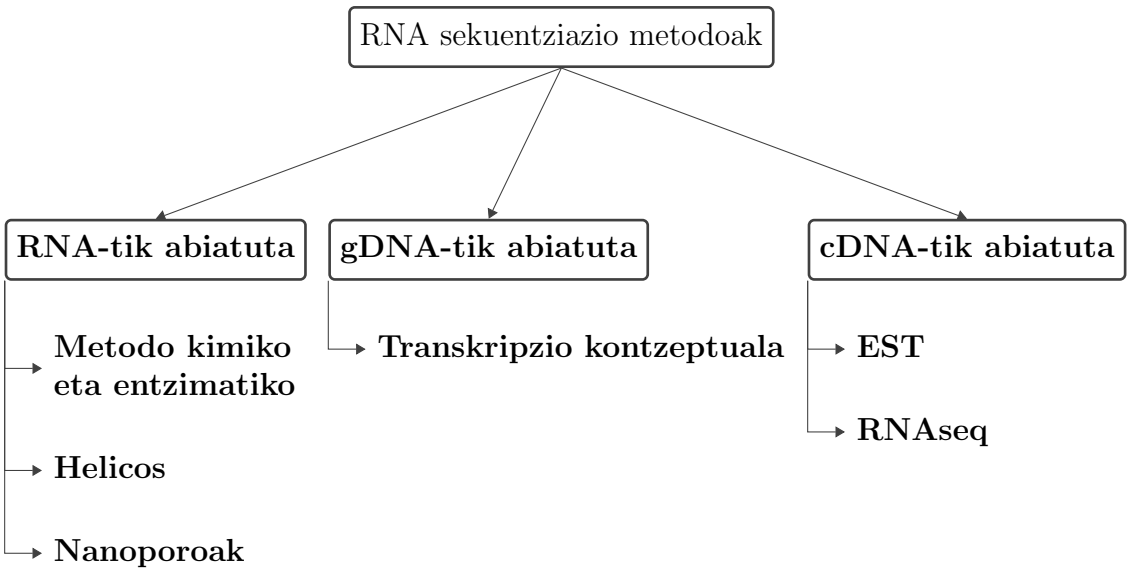
Emultsio PCR: Fase urtsuko tanta batean anplifikaziorako osagaiak dituzten bihiak eta genoteka fragmentuak emultsionatzen dira. Fragmentuen anplifikazioa bihiaren azaleran gertatzen da, eta sekuentziaziorako Picotiter plaka erabiltzen da. 454 Roche plataforman erabiltzen da.

2 Genoma sekuentziazioa



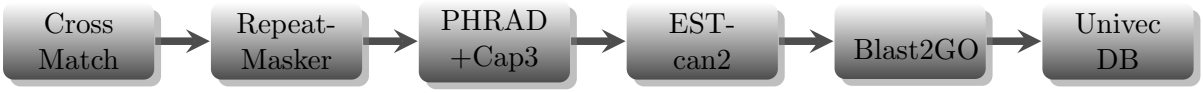
Giza Genoma Proiektua: 1990. urtean sortua, giza genoma proiektuak gizakiaren eukromatina DNA sekuentzia ezagutzea zuen helburu. Giza genomaren sekuentziazioa erakunde publiko (Internatinal Human Genome Sequencing Consortium, IHGSC) eta CELERA enpresa pribatuak burutu zuten.
IHGSC erakundeak klonez klon hurbilketa metodoa erabili zuen bitartean, CELERA enpresak genoma osoko shotgun sekuentziazioa erabili zuen.[1]

3 RNA sekuentziazioa



EST: EST-ak sekuentziaturiko mRNA harizpien adierazpen partzialak dira, DNA liburutegiko ausazko klonen sekuentziazio erreakzioetatik lorturikoak. EST proiektuak genoma proiektuak osatzeko edo geneak modu ekonomikoagoan aurkitzeko erabili ohi dira.
EST-en sekuentziazioa organismo edo ehun jakin baten mRNA purifikazioarekin hasten da. Alderantzizko transkripzioaren bitartez, mRNA cDNA bilakatu eta bektoreetan txertatzen dira genoteka eraikitzeko. Ausaz hautaturiko klonetatik abiatuta sekuentziazio erreakzioak burutzen dira hasle unibertsalak erabiliz.[2]

EST prozesamendu bioinformatikoa:



RNA-Seq: sekuentziazio teknologia berrietan oinarrituriko teknika da. RNA-Seq metodoan cDNA liburutegia eraikitzen da eta anplikaziorik gabe sekuentziatu egiten da; hortaz, mRNA adierazpen mailak eta transkriptoen egitura aztertzeko egokia da. Besteak beste, Helicos eta Illumina plataformak erabili daitezke sekuentziaziorako.[3]

Hibridazioa: Azido nukleiko zunden bitartez osagarritasun printzipioan oinarrituriK RNA sekuentzia espezifikoak detektatzeko teknika. Mota askotako teknikak egon daitezke, hala nola Northen Blot, mikromatrize eta mikrotxipak.

4 Proteinen sekuentziazioa

Proteinen amaierako egitura eta funtzioa finean aminoazido sekuentzian oinarritzen da, hots egitura primarioan. Hiru dira proteina sekuentziaziorako teknika nagusiak: Edman degradazioa, tandem masa espektrometria eta bioinformatikan oinarrituriko DNA edo RNA transkripzioa. Proteina sekuentziazioan proteina purua beharrezkoa da eta homogeneotasuna lortzeko kromatografia likidoa erabiltzen da.

Edman degradazioan proteinen N-terminaleko eraldaketa eta muturreko aminoazidoaren askapena erabiltzen da, askaturiko aminoazidoa kromatografiaz identifikatzeko.

Oinarrian, N-muturra Edman erreaktiboari esker eraldatzen da, eta medioa baldintza azidoetara aldatuz muturreko aminoazidoaren askapena lortzen da. Medioaren erredukzioz proteina muturra Edman erreaktiboarekin berriz eraldatu daiteke eta prozesua errepikatu. Teknika honek proteina kantitate oso txikiak eskatzen ditu (pikogramoak) eta proteina zuzenean aztertu daiteke, aurretiko liseriketarik gabe. Bestalde, proteinaren sekuentzia laburrak (50 aa) lortzen dira, eta prozesu motela du.

Masa esprektrometrian oinarrituriko sekuentziazioa *de novo* sekuentzioa da. Tandem masa espektrometrian liserituriko proteinen masa espektroen interpretazioz aminoazido sekuentzia ondorioztatu daiteke, zatiki bakoitzean liseriketa patroi bereizgarria gertatzen delarik.

Tandem masa espektrometriari esker, egitura eta funtzio analisiak, biomarkatzaile identifikazio, zuhaitz filogenetikoen eraikitzea eta proteina paralogo eta ortologoen bereizketa lortu daiteke.[4]

5 DNA anilisirako metodo estatistikoak

Bioinformatikan DNA sekuentziak sinbolo bereizgarrien (A, G, C, T/U) segidaren moduan ulertzen da; fasta formatua erabiltzen da ordenagailuan sekuentzia jaso eta kudeatzeko: sekuentziaren deskripzioa > zeinuaren ostean adierazten da hasieran eta jarraian sekuentzia hutsunerik gabe.

DNA sekuentzien analisirako hainbat programa bioinformatiko:

Simgene: PCRrako hasleen diseninua.

Webcutter: DNA sekuentzian entzimen errestrikzio guneak adierazten ditu.

REBsites: DNA sekuentzia errestrikzio entzimekin lixeritu eta elektroforesi emaitza erakusten du.

Bioinformatikan DNA sekuentzia S bezela adierazten da, s_i karaktere dituena. Azpisekuentziak adierazteko $s(n : n + m)$ erabiltzen da. Sekuentziak aztertzeko modelo estatistiko deserdinak erabili daitezke:

Modelo multinomiala: Nukleotidoak independente kontsideratzen ditu, eta agertzeko probabilitatea bere maiztasunarengatik soilik dago baldintzatuta.

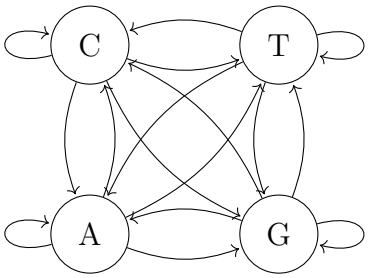
$$P(s) = \sum_{i=1}^{n} p(s_i)$$

Markov-en modeloa: Nukleotido baten presentzia aurretik dauden nukleotidoengatik dago baldintzaturik; magnitude desberdinetako koerlazioak eraiki daitezke, kontuan hartzeko diren aurretiko nukleotido kopuruaren arabera: 1. ordenekoa, 2. ordenekoa... Markov-en 0. ordeneko modeloa modelo multinomialarekin bat dator.

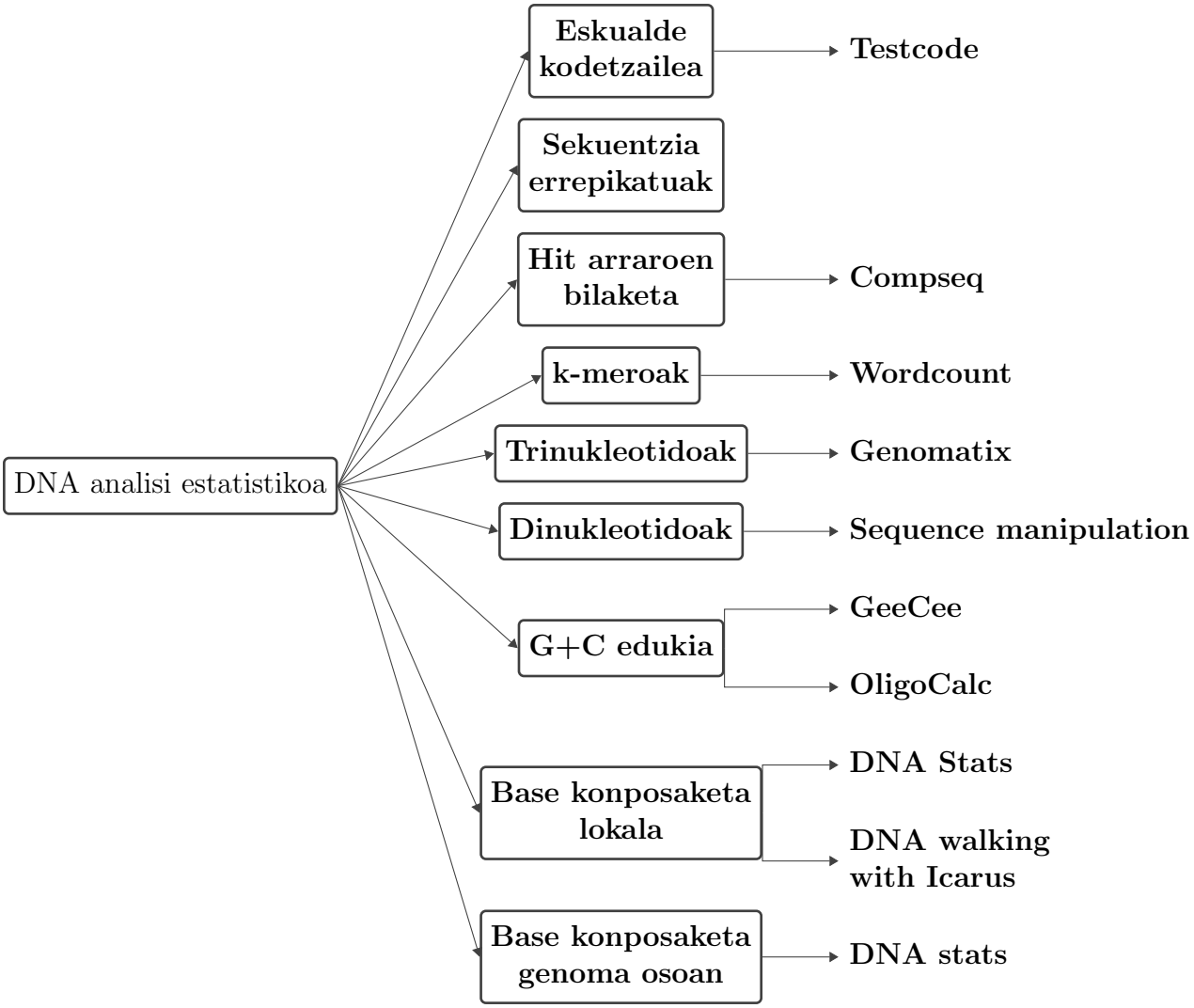
$$logP(s) = log\pi(s_i) + \sum_{i=2}^{n} p(s_i, s_{i-1})$$

Markov-en modeloa osatzeko, dinukleotidoen maiztasun matrizea eraikitzen da:

$$\begin{pmatrix} p(AA) & p(AC) & p(AG) & p(AT) \\ p(CA) & p(CC) & p(CG) & p(CT) \\ p(GA) & p(GC) & p(GG) & p(GT) \\ p(TA) & p(TC) & p(TG) & p(TT) \end{pmatrix}$$



6 DNA analisi estatistikoa

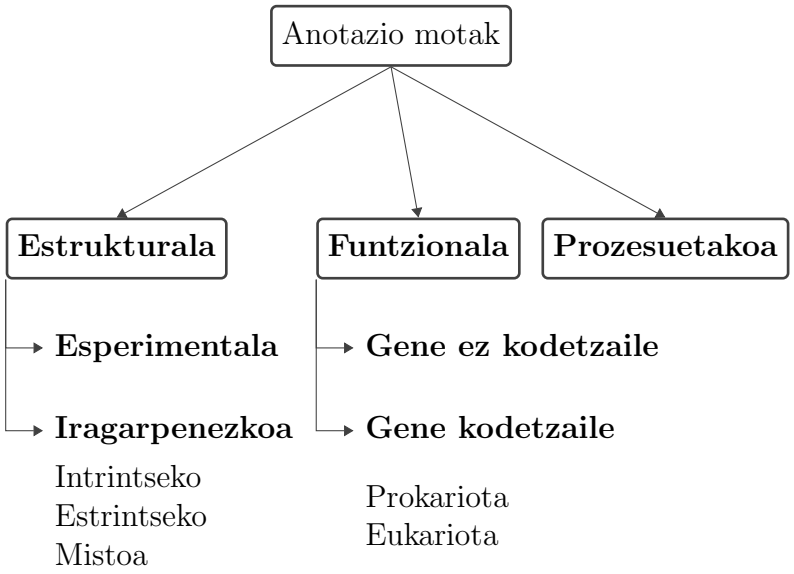


Base konposaketa: Genoma osoko base konposaketa espezie identifikazioarako erabili daiteke, eta base konposaketa lokala gene eta elementu desberdinen identifikaziorako. Base konposaketa lokala aztertzerakoan, lehioaren tamainak aldakortasun eta fluktuazioak eragin ditzake nukleotidoen maiztasunean.

G+C edukia: G+C edukia DNA sekuentzien bereizgarrian den ezaugarria da, eta informazio biologiko adierazgarria ematen du. Chargaff-en legean oinarriturik, harizpi zuzeneko A eta C edukiak jarraitzaileko T eta Gren berdinak dira; ondorioz, G+C eta A+T edukien baturak 1 ematen du.

Dinukleotidoak: dinukleotidoak azpiadierazita daudela kontsideratzen da maiztasun erlatiboa $p_{XY} \leq 0.78$ denean, eta gainadierazita $p_{XY} \geq 1.23$ denean. Oro har, dinukletiodoen maiztasunak sekuentzia osoan zehar mantentzen dira.

k-meroak: bi nukleotido baino gehiagoko sekuentziak dira eta lehio mugikorraren bitartez aztertu daitezke. Hitz luzeen kontaketan, maiztasun erlatiboa ez berdin 1 bada, hitz arraro kontsideratzen da: gainadierazita izan ohi dira elementu errepikakorrak, elementu erregulatzaileak edo funtzio biologiko jakina dutenak, eta azpiadierazita transkripzio faktoreen batura guneak, errestrikzio guneak birus genoman edo bakterio immune sistemarekin bateragarriak ez diren sekuentziak.



ENCODE: DNA elementuen entziklopedia. Giza genomaren eskualde funtzional guztien identifikazioa jasotzen du (eskualde kodetzaile, elementu erregulatzaile...). Anotazioa automatizatua edo eskuzkoa izan daiteke, azken hau fidagarriagoa izanik.

Anotazio estrukturala: DNA egitura edo sekuentzia espezifikoetan oinarriturik, DNA elementu desberdinen identifikazioa lortzea. Elementu desberdinak identifikatzeko esperimentalki aztertu daitezke edo iragarpenezko metodoak erabili: metodo intrintsekoek sekuentzia beraren informazioa (nukleotido maiztasuna, seinale sekuentziak, CG edukia, kodoi erabilera) kontsideratzen dira, metodo estrintsekoek sekuentzien arteko konparaketa egiten duten bitartean.

Anotazio funtzionala: DNA elementuen funtzio biologikoa iragartzen saiatzen da. Anotazio funtzionalean gene ez kodetzaile (tRNA, rRNA, snRNA) eta gene kodetzaileak (prokarioto edo eukarioto) aztertu daitezke.

ORF egiazkotasuna: (Prokariotoetan) DNA sekuentzian ORF jakin bat zorizkoa den beztertzeko, p balio edo egiazkotasuna determinatzen da kodoien agertzeko probabilitatea kontuan izanik. Ekiprobableak diren kodoientzako:

$$p = (1/64)(61/64)^k(1/64)$$

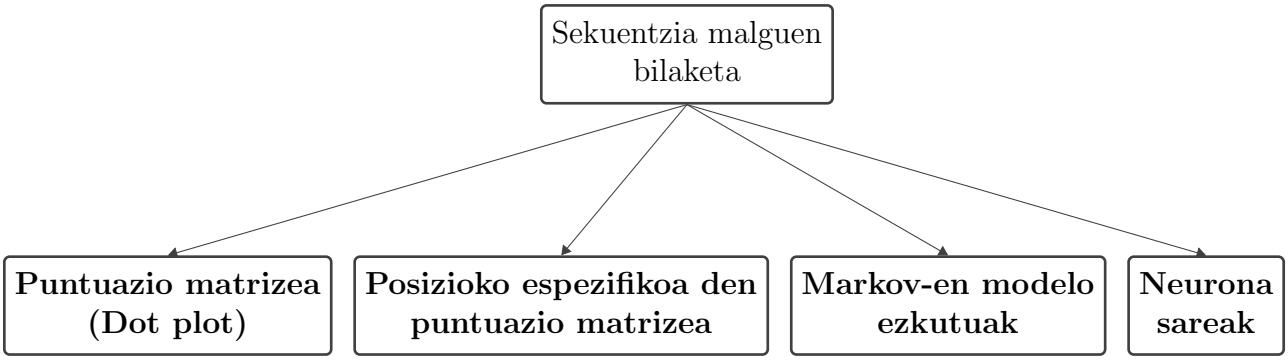
k ORF-ren luzera izanik, Start eta Stop kodoiak baztertuz. Atari estatistiko moduan, α esangura maila erabiltzen da: % 95-ko adierazkortasuna ezartzen denean, $\alpha = 0.05$ eta beraz ORF agertzeko probabilitatea p$< \alpha$ izan behar du adierazkorra izan dadin.

I.motako errorea: ORF-ak zoriz agertzeko probabilitatea determinatzerakoan, positibo faltsuen (ORF ez diren sekuentziak onartzea) kopurua adierazgarria izango da α balioa handitu ahala.

II. motako errorea: ORF-ak zoriz agertzeko probabilitatea determinatzerakoan, negatibo faltsuen (ORF diren sekuentziak baztertzea) kopurua adierazgarria izango da α balioa murriztu ahala.

7 Geneen bilaketarako metodo konputazionalak

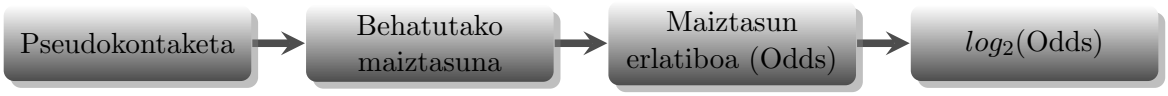
Metodo konputazionalen bitartez, DNA sekuentzien informazio kantitate handiak kudeatu daitezke gene eta DNA elementuen anotazioa lortzeko.



Sekuentzia malguak: Adierazpen erregular edo sekuentzia malguan posizio jakin batean nukleotido bat baino gehiago onartzen duen sekuentziak dira.

Puntuzko matrizea: Bi sekuentzia perpendikularki alderatzen dira, posizio desberdinen arteko koinzidentzia puntu batez adieraziz (Dot plot). Puntuen diagonalak matrizean sekuentzien arteko kointzidentzia onartzen da.

Posizio espezifikoko puntuazio matrizea (PSSM): Sekuentzia homologoen lerrokatze anizkoitzetik profil edo patroiak adierazten dituzte. Oinarrian, patroiak datubaseko sekuentziekin lerrokatzen dira, posizio bakoitzean sinbolo bakoitzarentzako probabilitate bat lortuz. Probabilitateak grafikoki adierazi daitezke LOGO-aren bitartez.[5]PSSM-a eraikitzeko urratsak:



Markov-en modelo ezkutuak (HMM): *Markov-en modelo ezkutuak* gertaera ikusgarrien eboluzioa deskribatzen dute,ikusgarriak ez diren barne faktoreen menpekoak direnak. Ikusgarria den gertaerari 'sinbolo' deritzo, eta barne faktore ezezagunari 'egoera'. Ezkutuko egoerek Markov katea eraikitzen dute, eta sinbolo ikusgarrien probabilitate banaketa (emisio probabilitate) barne egoerek (trantsizio probabilitateak) baldintzatuko dute.[6]Markov-en modeloak patroi edo perfilak sortzea baimentzen du, datu baseetan gorde eta sekuentziekin alderatuko dena. Perfilarekin bat datozen sekuentziak HMM perfilera gehituz, hau eguneratu eta eraginkorragoa bilakatzen da.

Neurona sareak: Neurona sareak, neurona biologikoen prozesamendu teknikan oinarriturik eta abiapuntua eredu multzo bat izanik, patroi ezagumendu eta hurbilketa burutzen duen tresna estatistikoa da.[7]Sareak 'input' informazioa prozesatu eta aldagaien arteko funtzioen parametro edo koefizienteak modifikatzeko gaitasuna du, ondoren iragarpen automatikoak egin ahal izateko

8 Azido nukleikoen datu baseak: GenBank

Genbank azido nukleiko sekuentzien datubasea da, eskuragarri dauden anotaturiko DNA sekuentzien bilduma. GenBank *International Nucleotide Sequence Database Collaboration* elkartearen parte da.

Sekuentzien bidalketa tresna desberdinen bitartez egin daiteke, hala nola BankIt, Sequin eta tbl2asn.GenBank banaketa taxonomiko eta funtzionalak ditu, eta sekuentzia bidaltzerakoan sailkapen bat egokitzen zaio. Sekuentziaren bilaketa egiteko metodo zuzena edo ezzuzena erabili daiteke, Nucleotide datu basea edo BLAST/NCBI tresnak erabiliz hurrenez hurren.

9 Proteina sekuentzien analisia

- **Egitura primarioa:** Proteinetan aminoazidoen segidari deritzo egitura primario; proteinaren parametro fisiko-kimiko, mozketa gune, seinale sekuentzia, transmintz eremu eta egitura bereizgarrien inguruko informazioa eman dezakete.
- **Egitura sekundarioa:** Aminoazido sekuentziak hiru dimentsiotan lortzen duen egitura. Egitura-funtzioari buruzko informazioa lortu daiteke.
- **Egitura tertziarioa:** Egitura sekundarioaren tolesduraz lorturiko egitura; egitura-funtzioari buruzko informazioa lortu daiteke.
- **Egitura kuarternarioa:** Azpiunitateen elkarrekintzek kooperatibitateari buruzko informazioa eman dezakete.
- **Proteina domeinua:** proteinaren sekuentzian egitura tertziario egonkor eta tolesdura independentea duen atala da, oro har egitura tertziario bereizgarria duena.

- **Egitura supersekundarioak:** Luzera laburreko egitura tertziario trinkoa da, ondoz ondoko egitura sekundarioek sorturikoa. Alfa helize egitura supersekundarioen adibideak helize-bira-helize, superkiribilketa helikoidala, helize sorta eta EF eskua dira; beta egitura supersekundarioen artean, β- urkila, β-meandroa, greka grekoa, upel eta sandwich egiturak daude. Bi egitura sekundarioen konbinazioz $\alpha - \beta - \alpha$, TIM upela, Rossman motiboa eta ferra egitura daude.

- **Aminoazidoen ezaugarriak:** Aminoazidoen karga, polaritatea, hidrofobizitatea eta bolumena aztertu daitezke. Venn diagramak aminoazidoak grafikoki sailkatzen ditu aipaturiko ezaugarrien arabera.

- **Itzulpen osteko eraldaketak:** Aminoazidoen eraldaketak (glikosilatzea, fosforilatzea, proteolisatzea...) aminoazido sekuentziak erregulaturik egon ohi da, adibidez fosforilazioa serina, treonina edo tirosina hondarretan gertatzen da, hiru aminoazidotako seinalea duelarik. Glikosilazioa N, O, GPI, C edo fosfoglikosilazio motakoa izan daiteke.

- **Seinale sekuentziak:** proteinen garraioa zuzentzen dute zelularen konpartimentu desberdinetara edota kanpoaldera.

- **Transmintz eremuak:** Bi transmintz eremu bereizten dira nagusiki, egitura sekundarioan oinarriturik, α-helize azaua eta β-upelak. Eremu transmintzen sekuentziek ezaugarri bereizgarriak dituzte, eta honen luzera inklinazioaren menpekoa izango da; α-helizetan 15-30 aa inguru ditu eremu transmintzak, β-upelak 8-12 aa dituen bitartean.

- **Positive inside araua:** Mintz proteinen alde zitoplasmatikora karga positiboko aminoazidoak orientatzen dituela ematen du aditzera arau honek, kanpoaldera karga negatiboko aminoazidoak orientatzen diren bitartean. Era berean, organuluetan karga positiboa zitoplasmara orientatzen da, organuluaren barrualdean karga negatiboa orientatzen den bitartean.

Programa informatikoak

- **ProtParam** programak proteinaren aminoazido konposaketa eta ezaugarri fisiko-kimikoak ematen ditu (Iraungitze koefizientea, erdi bizitza, aa konposaketa...)

- **Peptide cutter** programak mozketa guneak ematen ditu eta proteolisia gertatzeko probabilitatea.

- **PeptideMass** bitartez proteina proteasa batekin mostu eta lortzen diren peptidoen sekuentzia eta masa ematen du.

- **NetPhos** bitartez inguruko aminoazido konposaketan oinarriturik (3 aa normalean) fosforilazio guneak iragartzen ditu eta fosforilatzeko probabilitatea adierazi.

- **NetNGlyc** programak N-glikosilazioak iragartzen ditu eta **NetOGlyc** programak aldiz O-glikosilazioak.

- **Big-PI Predictor** programak GPI guneak iragartzen ditu.
- **Signalpeptide** eta **Proline** datu baseek seinale sekuentziak jasotzen dituzte. Seinale sekuentziak iragartzeko **Compgenomics** web orria erabili daiteke, eta **Signal-CTF** seinale sekuentziaren mozketa gunea iragarri. Beste programa batzuk **PSORT, SignalIP, Predisi, Phobius eta Signal-3L** dira.
- **COILS** erramintak eta **Multicoil** programak supperkiribilketa helikoidalak iragartzen dituzte.
- **2Zip**-ek leuzina kremailerak iragartzen ditu.
- **Disprot** programak IUP-ak (intrinsically unstructured proteins) iragartzeko tresnak ditu. **DisEMBL**, **GeneSilico** eta **IUPred** programek desegituratutako eremuak iragartzen dituzte.
- **eMotif** eta **eMatrix** motiboak iragartzeko tresnak dira.
- **MemType-2L** mintz proteinak iragartzeko programa; 8 mintz proteina mota bereizten ditu.
- **Octopus** programak mintz proteinen topologia iragartzen du.
- **TMHMM** eta **HMMTOP** programek transmintz-helizeak iragartzen dituzte, eta **Tmpred** eta **TopPred** programek eremu transmintzak iragartzen dituzte.
- **Amphipaseek** programak helize anfipatikoak iragartzen ditu.
- **PredTMBB** eta **BOCTOPUS** programek β-upelak iragartzen dituzte.

10 Sekuentzien lerrokatzea

Homologia: Homologia kontzeptua konparaketa azterketetarako erabiltzen da eta jatorri ebolutibo amakomuna ematen du aditzera. Homologiak bi sekuentzien arteko konparaketaren kalitatea adierazi nahi du, bi sekuentzien artean antzekotasun maila edo portzentaia dagoenean.

Antzekotasuna: Antzekotasuna kontzeptu kuantitatiboa da. Sekuentzien lerrokatze azterketan, bat datozen aminoazido edo nukleotido kopurua neurtzen da emaitz kuantitatibo bat emanez. Antzekotasun portzentaiak sekeuntziaren hondar identikoak kontuan izateaz gain, hutsuneak eta hondar antzekotasunak (ezaugarri fisiko-kimiko antzekoak) onartzen ditu.

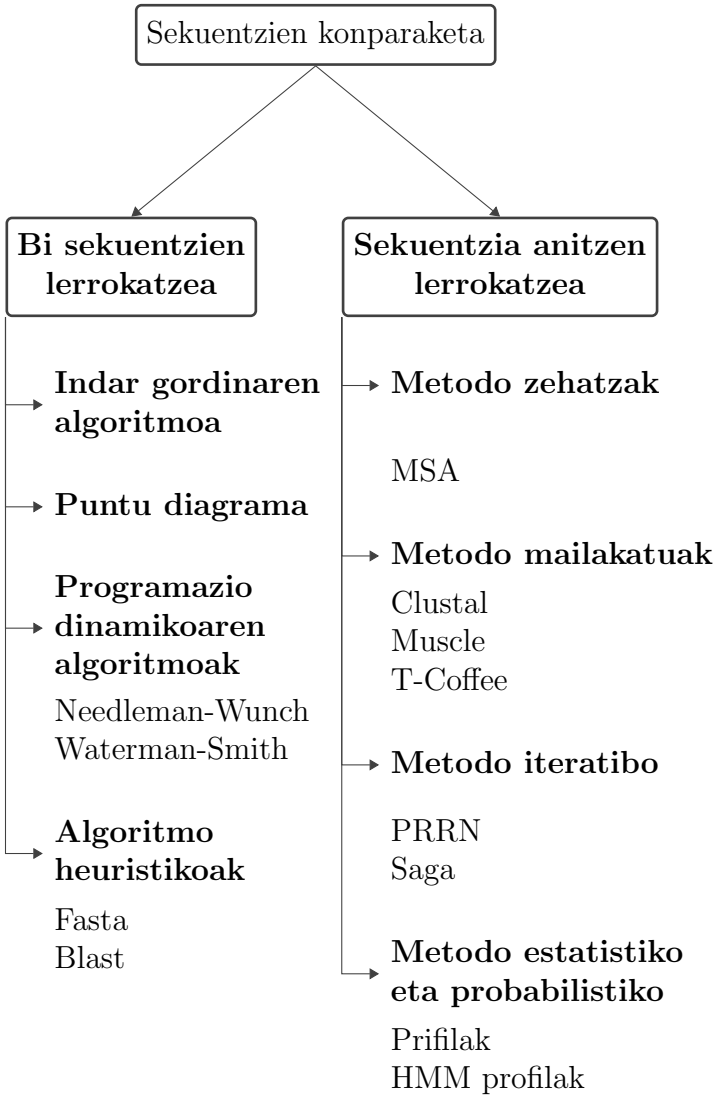
Identitatea: Identitate portzentaiak bi sekuentzien artean hondar identikoen portzentaia ematen du aditzera. Homologia ez da ondorioztatzen identitate datuetatik abiatuta; izan ere, identita maila eboluzio konbergentearen ondorio moduan ulertu daiteke.[8]

Analogo: Antzeko aktibitatea duten gene edo proteina baina jatorri desberdinekoak. Gene eta proteina analogoak eboluzioa konbergentearen ondorio dira.

Paralogoak: Organismo berean duplikazioz sorturiko gene homologoak. Eboluzio dibergentearen ondorioz sortzen dira jatorri berdinetik.

Ortologo: Espeziazio prozesuan, geneen duplikazio eta eboluzio dibergentziaren ondorioz, espezie desberdinetan homologoak diren geneak.

Xenologo: Organismo desberdinetan gene homologoak dira, gene transferentzia horizontalaren ondorioz sortuak. Gene xenologoak detektatzeko G+C edukia edo kodoi erabilera besteak beste erabili daitezke.



10.1 Puntuazio sistemak

Sekuentzien lerrokatze optimoa lortuarren, beharrezkoa da RNA, DNA eta proteina hondarren ordezkapen edota txertaketa/hutsuneen probabilitateak ezagutzea. Enpirikoki lorturiko datuetan oinarriturik mutazio probabilitateen hurbilketak lortzen dira, metodo desberdinen bitartez. Nukleotidoen kasuan, Jukes Cantor eta Kimura modeloak aztertuko dira hemen, eta aminoazidoen kasurako PAM eta BLOSUM matrizeak.

Nukleotidoen ordezkapenak belaunaldiz belaunaldi eta denbora ebolutibo luzeetan zehar gertatzen dira eta prozesu hau Markov-en katearen modeloaren arabera ulertu daiteke. Egoerak nukleotidoak izango dira eta trantsizioak nukleotido ordezkapenak.

Jukes Cantor modeloa: Mutazioen probabilitatea berdina da ordezkapen guztietarako, baita baseen maiztasunak; hots, modelo uniformean oinarritzen da.

$$s_{ij} = log_2(p_i Mij / p_i p_j)$$

s_{ij}-ren arabera, mutazioen probabilitatea zoriz gertatzearen probabilitatearena baina handiago edo txikiago den jakin daiteke.

Kimura modeloa: Kimura modeloan transbersio probabilitateak trantsizio probabilitateetatik bereizten dira, azken hauek probableagoak izanik; trantsizioak hiru aldiz probableagoak dira hain zuzen ere.

Ordezkapen prozesua hondar bakoitzarentzako denbora jarraian gertaturiko Markov prozesu moduan ulertu daiteke. Ordezkapen probabilitateak Q matrizearen bitartez adierazten dira, t denbora ebolutibo jakin baterako. Ordezkapen matrizea eraikitzeko modelo ebolutiboa hartzen da kontuan eta peptidoen lerrokatzerako erabili daitezke.

PAM matrizea (Point accepted mutation): PAM matrizeak hutsunerik gabe lerrokaturiko 71 amino azido blokeen sekuentziekin eraiki zen. Sekuentzia multzoak gutxienez %85-ko antzekotasuna partekatzen zuten eta eta mutazio probabilitateak zuhaitz filogenetikoetatik abiatzen dira. Mutazioak itzulgarriak direla onartzen da; hots, P(A→B) = P(B→A). Hortaz gainera, mutazioak independenteak dira aurreko mutazioekiko. PAM1 matrizea %1 baino desberdintasun gutxiago duten sekuentzien lerrokatze puntuaziorako da optimoa; sekuentzien arteko distantzia handitu ahala, PAM1 matrizearen berreturak erabiltzen dira (PAM160, PAM250...). PAM matrizea berkalkulatu ostean, Gonnet PAM250 eta JTT PAM250 sortu ziren.[**?**]

BLOSUM matrizea: BLOSUM matrizea 2000 bloke multzo baten lerrokatze lokaletik sortu zen (ez da modelo ebolutiboan oinarritzen). Matrizea definitzeko sekuentzien arteko distantzia kontuan hartzen da, hots, kontserbaturiko amino azido proportzioa. BLOSUM62 eta BLOSUM80 matrizeak %62 eta %80-ko gutxieneko antzekotasuna duten sekuentziekin eraikitzen dira, aztergai diren sekuentzien arteko distantziaren arabera. Hortaz, sekuentzien arteko distantzia handiago denean BLOSUM62 erabiltzen da, distantzia laburragoa denean BLOSUM80 erabiltzen den bitartean.[**?**]

10.2 Bi sekuentzien lerrokatzea

Indar gordinaren algoritmoa: Indar gordinaren algoritmoak bi sekuentzien arteko lerrokatzean aukera posible guztiak kontsiderantzen ditu

$$\frac{2^{2n}}{\sqrt{\pi n}} \tag{1}$$

Puntu diagrama: Bi sekuentzia (n eta m luzeretakoak) perpendikularki lerrokatzen dira nxm tamainako matrize bat osatuz. Sekuentzia baten eta bestearen arteko koinzidentzia gertatzen denean, matrizean puntu batez adierazten da, sekuentzien hondar guztietarako. Puntu diagramak sekuentzien arteko lekokatzeak eta egitura desberdinak agerian uzten ditu.

- **Diagonala:** Sekuentzien arteko kointzidentzia.
- **Simetria:** Sekuentzia berdinak lerrokatuz diagonalarekiko simetria lortzen da.
- **Dioagonalarekiko paraleloak:** Sekuentzia berdinen arteko lerrokatzean, tandem errepikapenak.
- **Perpendikularrak:** Alderantzizkaturiko sekuentzia edo sekuentzia palindromikoa; besteak beste, transkripzio faktoreen batura guneak, transposoiak, errestrikzio guneak eta stem-loop egiturak.
- **Puntu ugariko guneak:** Konplexutasun baxuko guneak; k luzerako lehioaren bitartez, emaitzak filtratu daitezke.
- **Desplazamendu bertikal edo horizontalak:** Bi sekuentzia desberdinen arteko lerrokatzean, indelak ematen ditu aditzera. RNA, cDNA edo ESTak DNArekin lerrokatuz, introi eta exoien antolamendua ematen du aditzera.
- **Gainezarpenak:** Sekuentzia desberdinen arteko lerrokatzean tandem errepikapenak.

Programazio dinamikoaren algoritmoa: *Divide and conquer* algoritmoak problema azpiproblema txikiagoetan zatitzen du, ebazteko errezagoak direnak. Azpiproblemen emaitzak kateatutak daude eta konbinatu egin daitezke problema nagusiaren emaitza lortzeko.

- **Needleman-Wunch:** Antzeko konposaketa eta luzera duten sekuentzien arteko lerrokatzea egiten du. Lerrokatze osorako eta erdiglobalerako erabiltzen da.
- **Waterman-Smith:** Luzera eta konposaketa desberdineko sekuentziak lerrokatzeko erabiltzen da. Lerrokatze lokalerako erabiltzen da.

Algoritmoa heuristikoak: Algoritmo heuristikoek problemem optimizaziorako soluzioaren hurbilketa proposatzen dute. Ez dute lerrokatze optimoa bermatzen eta beraz espezifikotasun, sentikortasun eta abiadurarren menpe daude. Oro har, algpritmo heuristikoek sentikortasun baxua izateagatik hainbat lerrokatze esanguratsu galdu daitezke.

Algoritmo heuristikoak *seed-amd-extend* teknikan oinarritzen dira besteak beste. Sekeuntzien lerrokatzerako gehien erabiltzen diren algoritmoak Fasta eta Blast dira.

- **FASTA.** Sentikortasun hobekuntza du beste algoritmoen aurrean, espezifitate eta abiadura gehiegi galdu gabe.
  1. Puntu diagramaren bitartez sekuentzien k-tuploak identifikatu.
  2. k-tuploak puntuatu eta hamar hoberenak hautatu. BLOSUM50 ordezkapen matrizea erabili ohi da aminoazidoentzako, eta identitate matrizea nukleotidoentzako.
  3. *init* parametroa: Hautaturiko sekuentziak beste ordezpekapen matrize batean oinarrituz berpuntuatzen dira (PAM matrizea) eta *init* parametroa lortzen da lerrokatze optimoetarako.
  4. *initn* parametroa: Sekuentzien arteko antzekotasun bilaketarako hutsuneak txertatzen dira, hauek penalizatuz.
  5. Programazio dinamiko bandeatua (Waterman-Smith algortimoa) erabiltzen du lerrokatze osoaren puntuaketa optimoa lortzeko. Opt eta E-value parametroek lerrokatzeen egokitasuna ematen dute aditzera.
- **BLAST.** DNA sekuentzia homologoak aurkitzeko, zuhaitz filogenetikoak eraikitzeko eta sekuentzia berrien anotazioa egiteko erabili daiteke.
  1. Sekuentziaren aurretiko prozesamendua: konplexutasun baxuko eskualdeak identifikatu eta ordeztu egiten dira interferentziarik ez sortzeko.
  2. Hitzen zerrenda sortzen da, hitz tamaina jakin baterako (w): proteinentzako hitzaren tamaina 3 izan ohi da, eta DNArentzako 11.
  3. Hitz bikote bakoitzarentzako (sarrera sekuentzia eta datu baseko sekuentzia) puntuazio bat kalkulatzen da ordezkapen puntuazio matrizeetan oinarrituz (PAM edo BLOSUM). Soilik muga (*threshold*-T-) gainditzen duten hitzak erabiliko dira lerrokatzerako.
  4. Puntuaketa eta hautaketa prozesua intereseko sekuentziaren hitz guztietarako errepikatzen da.
  5. Bi sekuentzien artean hitzen koinzidentzia gertatzen bada, lerrokatzea bi aldeetara hedatzen da koinzidentzia puntutik. Puntuazio altuko bikoteak (HSP, *High Scoring Pairs*) kontsideratzen dira.
  6. HSP-en esangura (E) kalkulatzen da formulen bitartez: E balio baxuek sekuentzia identikoak direla adierazten du, E balio altuek sekuentzien arteko antzekotasunak baztertzen dituen bitartean.

  BLAST programan hitzen tamaina txikitu ahala, sentikortasuna handitu baina abiadura murrizten da. T edo muga parametroa handitu ahala sentikortasuna murriztu baina abiadura areagotzen da.
- **BLAST2.** *two-hit* algoritmoa erabiltzen du: koinzidentzien luzera bikoitza duten hutsuneak onartzen dira hitzen artean. Luzapenaren ondorioz, puntuazio altuko lerrokatzeak lortzen dira.

  Hitz gutxiago luzatzen direnez, sentikortasuna murrizten da eta muga (T) txikiago erabiliz orekatzen da. Algoritmoak lerrokatze optimo bakarra lortzen du.

10.3 Sekuentzia anitzen lerrokatzea

MSA (*Multiple Sequence Alignment*) sortzeko %30-70 arteko antzekotasuna duten 10-15 sekuentzia erabiltzen dira. Sekuentzia partzialak ekiditen dira, eta oro har MSA-ak lerrokatze globaletarako erabiltzen dira. Normalean sekuentzia errepikakorrek arazoak ematen dituzte lerrokatze anizkoitzetan.

Metodo zehatzak: 3 sekuentzien lerrokatzerako 3 dimentsiotako matrizeak erabiltzen dira, eta lauki bakoitzean zazpi eragiketa kontsideratzen dira. *Sum of pairs* metodoa erabiltzen da, sekuentziak binaka lerrokatu eta ostean puntuazioak batuz. Heuristika erabiltzen da sekuentzia laburrak lerrokatzeko.

Metodo mailakatua: Sekuentziak oso antzekoak direnean edo sekuentzia asko lerrokatu behar direnean erabiltzen da.

1. Needleman-Wunch algoritmoen bitartez, sekuentzia guztiak binaka lerrokatzen dira, eta D distantzia parametroa kalkulatzen da (mismatch/match).
2. D parametroaren arabera, zuhaitz filogenetikoa eraikitzen da: D balio baxuek ebolutiboki gertu daudela adierazten dute.
3. MSA antzekotasun handieneko sekuentziarekin hasten da eta zuhaitzaren arabera bikote gehiago gehitu. Hasierako akatsak aurrerantzean anplifikatu egiten dira.

Metodo mailakatua erailtzen duten programak dira ClustalW, T-Coffee eta Muscle.

- **ClustalW.** Zuhaitz filogenetikoak eta Markov-en modelo ezkutuak erabiltzen ditu lerrokatze anizkoitzak burutzeko. Sekuentzia kopuru handiak erabili daitezke eta zehaztasun handia eskeintzen du.[10]
- **T-Coffee.** Programak informazio gehigarria sartzea onartzen du (informazio estruktural, profils edota egitura sekundario) eta emaitzetan akats gutxiago gertatzen dira; ondorioz, programa motelagoa da.[**?**]
- **Muscle.** Ehunka sekuentzia lerrokatzeko erabili daiteke eta abiadura handikoa da. Algoritmoak hiru urrats nagusi ditu: (1) lerrokatze mailakatu zirriborroa, (2) lerrokatze mailakatuaren hobekuntza eta (3) fintze urratsa. D distantzia determinatzeko k-meroen koinzidentziak kontsideratzen dira (match ordez).[12]

Metodo iteratiboa: Oinarrian, algoritmoak azpitaldeak modu errepikakorrean lerrokatzen ditu azpisekuentziak akats kopurua murrizteko eta aurrerantzean ez anplifikatzeko. Metodo heuristikoak erabiltzen ditu lerrokatze optimoaren hurbilketa lortzeko. Metodo iteratiboak erabiltzen dira antzekotasun baxua duten sekuentziekin.

Metodo iteratiboa erabiltzen duten programak dira SAGA, Multialign eta PRRP.[13]

- **SAGA.** Algoritmoa genetikoa erabiltzen du eta lerrokatzea sekuentzia populazio baten hobetze gradualean oinarritzen da.
  1. 100 sekuentzia lerrokatzen dira 0-50 posizioen artean desplazatuz. Muturretan hutsuneak onartzen dira luzera berdina izan dezaten (G_0 hasieraketa)

2. Hasierako lerrokatzeak *sum-of-pairs* metodoaren bidez ebaluatzen dira, ordezkapen matrize eta hutsune penalizazio arruntak erabiliz.
3. Lerrokatze anizkoitzaren bigarren belaunaldia hautatzen da, soilik *fitness* (*sum-of-pairs* edo *objectif function*) hoberena duten lerrokatzeak kontsideratuz (hautespen naturala). Lerrokatze parental optimoek seme optimoak emango dituzte.
4. Ugalketa urratsean, lerrokatze guztietatik puntuazio hoberena duten %50a MSA eraikitzeko zuzenean erabiltzen dira; beste %50ak mutazio (hutsuneak) eta birkonbinazioak jasaten dituzte MSA-n erabili aurretik.
5. Prozesua puntuazioa gehiago hobetu ezin daitekeenean amaitzen da.

Metodo estatistiko edo probabilistikoa: Markov-en modelo ezkutuak erabiltzen ditu. Algoritmo desberdinak erabili daitezke: zehatza, mailakatua, iteratibo/estokastikoa...

* MSA-ak editatzeko tresnak: Jalview eta Boxshade

* Clustal Omega programa erabiliena da eta antzekotasun tarte zabalak onartzen ditu. T-Coffee emaitz zehatzenak lortzen ditu.

11 MSA analisia

Motiboak: sekuentzia luburreko patroi edo adierazpen erregularrak dira, esanahi biologikoa izan ohi dutenak (gune katalitiko, batura guneak talde prostetiko, ioi edo molekulentzako). Adostasun sekuentziak erabiltzen dira sinbolo kontserbatu eta haien arteko distantziak adierazteko; sinbolo mota desberdinak erabiltzen dira adierazpenerako (hizkiak-X-[ ]-{ }/aminoazido kontserbatua-hutsunea-aminoazido posibleak-bazterturiko aminoazidoak).

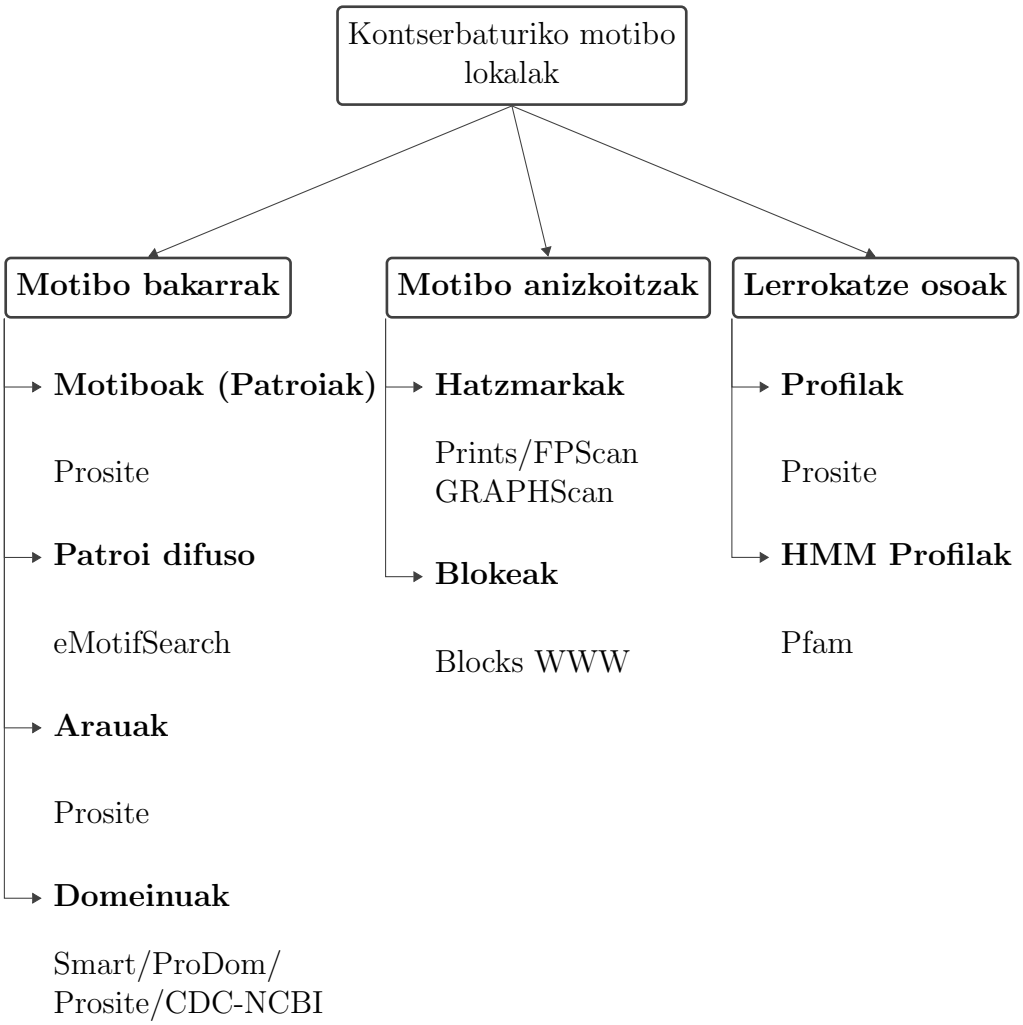
Patroi difusoak: Adostasun sekuentzietan hondarren ezaugarri fisiko kimiko berdinak kontsideratzen dira. Oro har, luzeagoak dira eta esanahi biologikoa hobeto islatzen dute. Sentikorragoak dira, baina espezifitate baxuagoa dute.[14]

Arauak: Patroi laburrak dira eta espezifitate oso baxua izan ohi dute (positibo faltsu ugari); proba biokimikoek gune funtzionalak ziurtatzen dituzte.

Domeinuak: Proteinen unitate funtzional edo estrukturalak. Eboluzio molekularrean unitate bezela aztertzen dira, errekonbinatuz proteina funtzio desberdinak sortzen direlarik.

Hatzmarkak: Motiboen multzoek osatzen dute. Hatzmarkak MSAtik lorturiko maiztasun matrizean oinarritzen da (sistema ez-haztatua erabiliz). Metodo iteratiboen bitartez datu baseetatik sekuentzia berriak sartzen dira eta matrize berria eraikitzen da. Prozesua konbergentziara iritsi arte errepikatzen da eta PRINTS datu basean sarrera egin.

Intereseko sekuentzietan hatzmarka bilaketa bakoitzerako puntuazio bat kalkulatzen da. Hatzmarken bilaketa emaitzetan sekuentziak zenbat motibo dituen adierazten da, baita zenbateko antzekotasuna duten (I edo i bitartez adierazita).



Blokeak: Blokeek lerrokatze anizkoitzetan motibo kontserbatuenak ematen dituzte aditzera, hutsune eta intsertorik gabe. Sistema haztatuen bitartez eraikitzen dira, MSA-ko sekuentziak oso antzekoak direlarik. Blokeen analisirako programa bioinformatikoek blokeen bilaketa, bloke eraikuntza edo bloke informazio bilaketa egin dezakete.[15]

Profilak: Profilak, patroiak ez bezela, lerrokatze globaleko puntuazio tauletatik abiatzen da eta ez dira sekuentzia laburretara mugatzen. Profilak sortzeko lerrokatze anizkoitza eta puntuazio sistema konplexuak erabiltzen dira, metodo haztatuan oinarriturik. Lerrokatzerako homologo urrunak erabiltzen ditu eta eskualde kontserbatu zein ez-kontserbatuak aztertzen ditu. Diseinatzeko zailak dira eta funtzio biologikoa topatzeko desegokiak.

HMM-n oinarrituriko profilak: Markov-en modelo ezkutuan oinarriturik eraikitzen dira, eta MSA-tik abiatuta emisio eta trantsizio probabilitateak kalkulatzen dira.

Erreferentziak

[1] Chial, H. (2008) *DNA sequencing technologies key to the Human Genome Project.* Nature Education 1(1):219

[2] Parkinson J.(2009) *Expressed sequence tags: generation and analisys.* Methods in molecular biology. 533

[3] Wang Z., Gerstein M., Snyder M. (2009) *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet. 10(1): 57-63

[4] Saraswathy N., Ramalignam P. (2011) *Concepts and Techniques in genomics and proteomics.* Biohealthcare publishing (Oxford). 193-196.

[5] M. Michael Grominha (2010) *Protein sequence analisys in protein informatics.*

[6] Yoon B.J. (2009) *Hidden markov models and their applications in biological sequences analisys.* Curr Genomics. 10(6):402-415.

[7] Manning T., Sleator R.D., Walsh P. (2014) *Biollogically inspired intelligent decision making: a commentary on the use of ANN in bioinformatics.* Bioengineered. 5(2): 80-85.

[8] Kanduc D. (2012) *Homology, similarity, and identity in peptide epitope immunodeffinition.* J. Pept. sci.

[9] Polanski A., Kimmel M. (2007) *Bioinformatics.* Springer Berlin Heilderberg New York

[10] Sievers F., Wilm A., Dineen D. eta al. (2011) *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol Syst Biol. 7(1):539

[11] Notredame C., Higgins D.G., Heringa J. (2000) *T-Coffee* J Mol Biol. 302(1):205-17.

[12] Edgar R. (2004) *Muscle: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics. 5: 113

[13] Wallace I.M., Orla S., Higgins D.G. (2005) *Evaluation of iterative alignment algorithms and multiple alignment.* Bioinformatics. 21(8): 1408-1414.

[14] Sigrist C.J.A., Cerutti L., Hulo N. eta al. (2002) *PROSITE: A documented database using patterns and profiles as motif descriptors.* Briefings in bioinformatics. 3 (3): 265-274

[15] Henikoff J.G., Henikoff S. (1996) *Blocks database and its applications.* Elsevier. 266:88-105.