SECVENTIEREA ARN (RNA-seq)

- Secventierea transcriptomului (totalitatea speciilor moleculare de ARN sintetizate pe baza informatiei genetice dintr-un genom) este realizata prin tehnologiile NGS.

- RNA-seq permite cuantificarea concentratiilor relative de ARN, modificarile nivelelor de expresie genica (variatii ale ratelor de transcriptie genica) in diferite conditii patologice sau experimentale, expresia diferentiala a genelor in timp si spatiu, precum si analiza fenomenului de *splicing* alternativ dintr-un anumit tip de celula sau tesut.

- De asemenea, RNA-seq permite adnotarea functionala a genelor. De exemplu, daca pentru o gena identificata *in silico* poate fi secventiat un ARNm care corespunde acestei gene, cel mai probabil acesta gena este functionala, intrucat a fost transcrisa *in vivo*.

ASAMBLAREA GENOMURILOR SECVENTIATE - CALITATEA ASAMBLARII

În etapa de asamblare, se pornește de la *reads*-uri (*paired* si *unpaired reads*), cu care sunt construite *contig*-uri care evident, nu au aceeasi lungime. Cu cât *contig*-urile de mari dimensiuni sunt mai numeroase (sunt in proportie mai mare), cu atât este mai buna calitatea secvențierii și a asamblarii. Calitatea asamblarii este exprimata cu ajutorul parametrului/valorii **N50**.

De exemplu, o valoare N50 = 4000, înseamnă ca jumatate (50%) din totalitatea nucleotidelor ordonate in *contig*-uri se afla intr-un grup de *contig*-uri de dimensiuni diferite, dintre care cel mai scurt are 4000. Daca presupunem ca o gena are in medie 2kb, va fi foarte dificil sa identificam gene in numeroasele *contig*-uri scurte care au doar cateva sute de nucleotide. Evident, calitatea secventierii/asamblarii este mult mai buna atunci cand parametrul **N50** = 4000 decat atunci cand **N50** = 350.

Din *contig*-uri sunt construite *scaffold*-uri (colecție de *contig*-uri și *gap*-uri din ACELASI cromozom), iar dupa ce sunt rezolvate *gap*-urile de secvențiere este obținut *CONTIG*-UL IDEAL (SUPER-*CONTIG*), care acoperă integral molecula de ADN a unui cromozom.

De mentionat ca si pentru *scaffold*-uri poate fi utilizat parametrul calitativ **N50**, cu aceeasi semnificatie ca si in cazul *contig*-urilor. De asemenea, numarul si dimensiunea gap-urilor din *scaffold*-uri reprezinta un parametru calitativ, un *scaffold* de secventa este de o calitate mai buna daca are gap-uri mai putine si mai scurte.

In functie de calitatea secventierii si asamblarii *reads*-urilor rezulta:

- **Genomuri secvențiate complet** (*finished sequence*), ceea ce înseamnă că există un *contig* ideal pentru fiecare dintre cromozomii in care este partajat respectivul genom;

- Secvenţe draft, alcătuite dintr-o colectie de *contig*-uri şi *scaffold*-uri de mari dimensiuni.

Cand conditiile financiare si logistice permit, e recomandabil ca un genom sa fie secventiat prin doua strategii alternative de secventiere, de exemplu prin WGSS si WGS-NGS. De asemenea, colectiile de *reads*-uri obtinute trebuie asamblate alternativ cu software diferite, intrucat nu exista software de asamblare perfecte. Rezultatele asamblarilor obtinute sunt ulterior comparate in vederea obtinerii unei secvente consens a genomului respectiv.

Asamblarea genomurilor poate fi realizata relativ usor daca exista o secventa de referinta cu care pot fi comparate *contig*-urile obtinute. Acest avantaj exista doar in cazul speciilor care au fost deja secventiate si pentru care exista deja un genom de referinta disponibil in bazele de date internationale. De exemplu, asamblarea *contig*-urilor unei linii noi de *Drosophila melanogaster* sau de *Saccharomyces cerevisiae* poate fi realizata relativ usor, intrucat exista deja genomurile de referinta ale acestor specii.

In cazul in care nu exista un genom de referinta, asamblarea se face *de novo*, utilizand software specializate de bioinformatica.

Genomurile asamblate sunt depuse in baze de date internationale precum GenBank (https://www.ncbi.nlm.nih.gov/genbank) si, de obicei, sunt publice.

Catena de referinta a fiecarui cromozom din genom este simbolizata + (plus), insa reprezinta o notiune relativa, intrucat este aleasa in mod arbitrar de catre autorii secventerii. In bazele de date, catena complementara, simbolizata si catena - (minus), nu este scrisa in mod explicit, dar este mereu luata in considerare in analizele bioinformatice.

Multe software de bioinformatica sunt instruite sa genereze revers-complementul catenei de referinta.

ADNOTAREA GENOMURILOR SECVENTIATE

Secventa asamblata a unui genom reprezinta, de fapt, un text de mari dimensiuni, scris cu doar patru litere: A, C, G si T. Pentru a intelege ce contine acest text, este necesar sa il adnotam, adica sa ii asociem note explicative care sa ne permita sa intelegem informatiile cuprinse in acest text.

- Prin adnotare structurala se intelege aplicarea unor tehnici de bioinformatica pentru a identifica *in silico* elemente structurale care intra in alcatuirea genomurilor.

- Prin adnotare functionala se intelege aplicarea unor tehnici moleculare pentru a demonstra functionalitatea elementelor structurale identificate *in silico*.

ADNOTAREA STRUCTURALA A GENOMURILOR

Prima etapa consta in identificarea *in silico* a unor elemente structurale precum:

- gene care codifica pentru ARNm (si implicit, pentru proteine);
- gene care codifica pentru alte tipuri de ARN (ARNr, ARNt, micro ARN, etc.);
- transpozoni (elemente genetice mobile);
- elemente reglatoare *in cis*.

Cea mai importanta etapa a adnotarii structurale este identificarea de gene care codifica pentru proteine. In acest scop, au fost dezvoltate software care identifica structuri numite **ORF** (ORF = Open Reading Frames) in interiorul unei secvente de ADN de interes (secventa *query*).

Definitie ORF: o secventa de nucleotide cuprinsa intre un codon start (de obicei, ATG) si un codon stop din avalul codonului start. La bacterii, unde nu exista introni, notiunea de ORF este identica cu notiunea de gena. In schimb, echivalarea unei gene cu un ORF este complicata de existenta intronilor.

Pentru un fragment de ADN oarecare exista 6 **cadre de citire** (*reading frames*) distincte, respectiv cate 3 variante diferite de a imparti cele doua catene in succesiuni de codoni (multipli de trei nucleotide consecutive).

Concret, daca impartim fiecare catena in codoni, obtinem cate trei cadre de citire posibile pentru fiecare catena. Se observa ca, in cazul cadrului de citire 1 al primei catene ADN din figura de mai jos, primul codon este ATG, in cazul cadrului de citire 2 – TGA, iar pentru cadrul de citire 3 – GAT. Evident, daca fiecare cadru de citire ar fi tradus in secventa de aminoacizi respectiva, ar rezulta trei peptide distincte.

Aceeasi situatie este valabila si pentru cele trei cadre de citire corespunzatoare catenei complementare, care este scrisa sub forma de revers-complement si marcata cu culoarea gri. Rezulta ca, teoretic, pot fi sintetizate 6 peptide distincte.

Exemplu de cadre de citire:

```
5’ATGATCGATCGATCGTAACCC3’
  - - - 1
   - - - 2
    - - - 3
5’GGGTTACGATCGATCGATCAT3’
  - - - 4
   - - - 5
    - - - 6
```

Se observa ca, in exemplul prezentat aici, chiar daca exista 6 cadre de citire, exista doar un ORF in cadrul de citire 1, format din 6 codoni. Acest ORF incepe cu codonul start ATG (care

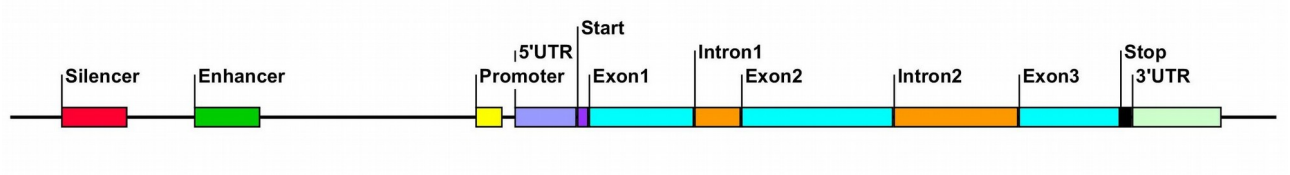
codifica pentru aminoacidul metionina) si se termina cu codonul stop TAA (am utilizat codul genetic ADN, nu cel ARN).

EXERCITIU: CARE ESTE SECVENTA DE AMINOACIZI A PEPTIDEI CODIFICATE DE CADRUL DE CITIRE 1?

Ca regula empirica, daca pentru un fragment de ADN de cateva kb sunt identificate mai multe ORF-uri virtuale, ORF-ul cel mai lung este cel real. Caracterul autentic, functional al unui ORF nu poate fi demonstrat decat pe cale experimentala, prin tehnici moleculare precum qRT-PCR sau *microarray*.

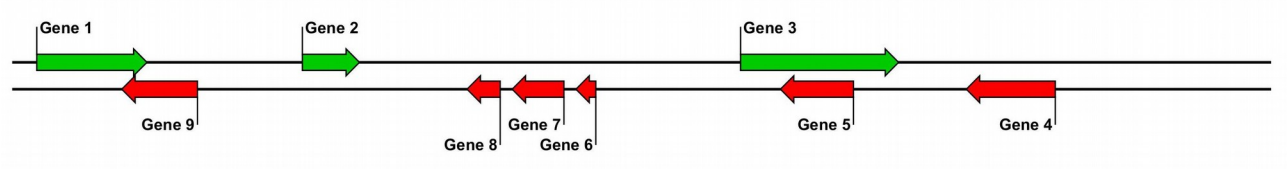
De aceea, genele identificate in urma adnotarii structurale primare se mai numesc si ***gene ipotetice, gene in silico, computed genes,* sau *putative genes***.

Exista software de bioinformatica (de exemplu: *Genie*) capabile sa recunoasca in textul genomului module de sub-secvente care sunt specifice pentru structura genelor. De exemplu, identificarea modulului Promotor – 5'UTR – Codon Start – Exon - Intron – Exon – Intron - Exon – Codon Stop – 3'UTR este sinonima cu identificarea unei gene *in silico*. In exemplul de mai jos, este descrisa structura generala a genelor la eucariote:



Genele au **CATENA SENS** fie in catena de referinta a genomului, fie in catena minus a genomului. Se numeste catena sens doarece contine efectiv codonii din ARNm, doar ca in loc de uracil contine, evident, timina. Catena sens a unei gene oarecare este acea catena care NU e folosita ca matrita pentru transcriptie/sinteza ARNm.

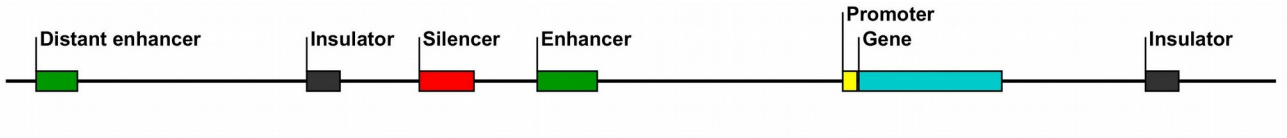
In mod conventional, in bazele de date sau in articole stiintifice este scrisa doar catena sens a unei gene. In browserele genomice, genele care au catena sens in catena de referinta a genomului sunt simbolizate cu o sageata cu varful orientat spre dreapta. Genele care au catena sens in catena minus a genomului sunt simbolizate cu o sageata cu varful orientat spre stanga.



- **Elementele reglatoare *in cis*** (*cis-acting*) sunt fragmente de ADN de mici dimensiuni (cateva zeci sau sute de nucleotide), localizate de regula in amonte de genele pe care le regleaza. Se cunosc si cazuri de elemente reglatoare *in cis* localizate in avalul genelor reglate sau chiar in interiorul acestora.

- Totalitatea elementelor reglatoare *in cis* dintr-un genom poarta denumirea de **CISTROM**.

Exemple de elemente reglatoare *in cis*: ***promotorii*** (initiaza transcriptia genica), ***enhancers*** (activeaza transcriptia genica), ***silencers*** (inhiba/represeaza expresia genica) si ***insulators*** (protejeaza gena de efectul unor *enhanceri* sau *silenceri* mai indepartati, altii decat cei care controleaza specific o anumita gena). Numele de elemente reglatoare *in cis* vine de la faptul ca aceste fragmente de ADN sunt localizate in aceasi molecula de ADN cu gena a carei transcriptie o regleaza.



Exista si elemente reglatoare *in trans*, care pot regla activitatea unor gene localizate in diferite molecule de ADN (cromozomi). Elementele reglatoare *in trans* nu sunt secvente de ADN, ci **proteine** implicate in reglarea fenomenului de transcriptie (factorii de transcriptie).

Secventele reglatoare *in cis* sunt reprezentate adesea ca secvente CONSENS, deoarece pentru fiecare secventa reglatoare *in cis* exista mai multe variatiuni de secventa pe aceasi tema. De aceea, este mai convenabil ca diferitele variante ale unei anumite secvente reglatoare sa fie reprezentate ca o secventa unica, ideala. De exemplu, atunci cand intalnim in literatura de specialitate informatia conform careia **secventa TATA** (TATA BOX) este o componenta a promotorului genelor la eucariote, este vorba de o generalizare. De fapt, TATA este o secventa consens formata din 7-8 nucleotide, in care predomina nucleotidele A si T si care incepe cu secventa TATA. In mod real, genele au diferite variante ale acestei secvente consens. Ideea de baza a stabilirii unei secvente consens este alinierea cat mai multor secvente **echivalente** de aceeasi lungime, apoi nucleotida predominanta in fiecare pozitie este scrisa in secventa consens.

De exemplu, presupunem ca intr-un genom exista secventa reglatoare ipotetica consens TCGTC, care intervine in modularea transcriptiei unei gene G, conservata la 4 specii de plante. De

fapt, fiecare dintre cele 4 specii are, in mod concret, o varianta distincta a acestei secvente. Se observa insa un *pattern*, in sensul ca in pozitiile 1 si 4 exista o nucleotida care participa la formarea a doua punti de hidrogen (A sau T), iar in pozitiile 2, 3 si 5 exista o nucleotida care participa la formarea a trei punti de hidrogen (C sau G). Secventa consens este 5′TCGTC3′, iar in mod intamplator specia S2 are chiar varianta “ideala” a acestei secvente consens.

```
S1:      5′TGGTG3′
S2:      5′TCGTC3′
S3:      5′ACGAC3′
S4:      5′TCCTC3′
Consens: 5′TCGTC3′
```

Transpozonii - elemente genetice mobile, sunt prezenti in toate genomurile procariote si eucariote. In general, cu cat o specie este plasata mai sus pe scara evolutiva, cu atat ponderea transpozonilor in genom este mai mare. Astfel, la mamifere transpozonii pot reprezenta aproximativ 40-50% din genom, in timp ce genele care codifica proteine reprezinta doar circa 2%. In replica, la bacterii genele reprezinta un procent relativ mult mai mare.

Transpozonii de clasa I - sunt mai numerosi si nu sunt mobilizati propriu-zis (de exemplu retrotranspozoni de tip SINE, LINE, etc.) Acesti transpozoni sunt convertiti prin revers-transcriere in molecule de ADNc (ADN complementar), care este inserat in alte situsuri din genom, ceea ce conduce la o crestere neta a numarului copiilor/genom. Acest mecanism de transpozitie se numeste *copy-and-paste*.

Transpozonii de clasa II - transpozeaza efectiv, printr-un mecanism ce consta in excizia transpozonului din locatia originala, urmata de mobilizarea si reinsertia acestuia intr-o alta locatie in genom. Transpozitia acestor elemente mobile este catalizata de transpozaza, o enzima codificata chiar de catre transpozonii de clasa II. Elementul mobil P de la *Drosophila melanogaster* este un transpozon de clasa II.

Adnotarea structurala se bazeaza pe principiul similaritatii de secventa. Astfel genele ipotetice si/sau proteinele codificate teoretic de catre acestea sunt analizate comparativ *versus* genele/proteinele confirmate experimental si care pot fi gasite in baze de date dedicate. Principiul este simplu: similariatea de secventa reflecta adesea inrudirea functionala a genelor/proteinelor.

De exemplu daca analiza bioinformatica indica faptul ca o gena *in silico* are secventa asemanatoare cu a unor gene care codifica kinaze la diferite specii, este foarte probabil ca respectiva gena ipotetica sa faca parte din familia genelor care codifica kinaze.

Genele pentru care nu exista gene omoloage in alte genomuri, nu fac parte din nici o familie si sunt denumite **gene orfane** (*orphan genes*). Este foarte probabil ca genele orfane sa fie specifice pentru o singura specie.

O categorie speciala de elemente componente ale genomurilor sunt **PSEUDOGENELE**. Asa cum le spune si numele, pseudogenele sunt gene false, in sensul ca au caracteristicile structurale ale unei gene (5'UTR, promotor, codon start, alternanta de exoni si introni, codon stop, secventa care codifica pentru regiunea poliA din ARNm si 3'UTR) insa nu sunt functionale, din cauza unor mutatii. Pseudogenele sunt mai frecvente la eucariote.

Daca doua gene sunt intr-adevar inrudite, adica au un stramos molecular comun (evolutia are loc initial la nivel molecular), acestea sunt OMOLOAGE. Daca e vorba de gene inrudite care apartin unor genomuri (specii) diferite, acestea sunt ORTOLOAGE, iar daca apartin aceluiasi genom se numesc PARALOAGE.

De exemplu, genele care codifica hemoglobina la om si la cal sunt ORTOLOAGE (apartin unor genomuri/specii diferite), in timp ce genele care codifica hemoglobina si respectiv mioglobina la om sunt gene PARALOAGE (apartin aceluiasi genom/specie).

De retinut faptul ca procesul de adnotare structurala este continuu, adica in mod constant au loc readnotari ale genomului. Motivul consta in faptul ca, pe masura ce sunt dezvoltate noi software de adnotare, acestea sunt "capabile" sa identifice in genomuri elemente structurale care au fost omise de "scanarea" efectuata de softwarele mai vechi.

In articolul "*A beginner's guide to eukaryotic genome annotation*" este prezentata o lista cu diferite software utilizate in procesul de adnotare a genomurilor, precum si notiuni generale legate de adnotarea genomurilor.