PROMOTER SIMILARITY

MÍRIAM SUBIRATS

NEUS XIVILLÉ

Agraïments

Volem agrair la col·laboració de totes aquelles persones que han fet possible la realització d'aquesta projecte.

En primer lloc als directors del projecte. Al Dr. Xavier Messeguer i Peypoch, tutor d'aquest, que va brindar-nos la possibilitat de realitzar un projecte de bioinformàtica, orientat a la investigació. També li volem agrair el suport ofert al llarg de la realització del mateix. A Domènec Farré per ajudar-nos en tot el que ha estat al seu abast, tant en explicacions biològiques, com en entrebancs de programació que ens han sorgit en el transcurs del projecte, com pels seus valuosos comentaris, revisions i suport al llarg de tot aquest temps.

A Jose Marcos López Caravaca, integrant del Laboratori de Càlcul de la Facultat, per la seva col·laboració i disposició en tot el que fa referència al dia a dia del projecte.

A Agustí Costa, Lluís Sintes, Anna Montraveta, Jordi Munmany, Patricia Cardona, Pilar Cortada, i a la família de cadascuna de les integrants, pel seu amor i suport incondicional i continu.

I a tots aquells que d'una manera o altra ens han ajudat, encara que el seu nom no figuri de forma explícita en aquestes línies; l'ajut de qualsevol sempre és benvingut.

ÍNDEX

1. INTRODUCCIÓ

1.1. MOTIVACIÓ DEL PROJECTE 7
1.2. ESTRUCTURA DEL PROJECTE 7
1.3. ESTRUCTURA DE LA MEMÒRIA 8

2. CONTEXT BIOLÒGIC

2.1. ÀCID DESOXIRRIBONUCLEIC 10
2.2. EXPRESSIÓ GÈNICA 12

2.2.1 *Síntesi de les proteïnes* 12
1. Transcripció 12
2. Traducció 13
2.2.2. Sistema de regulació de l'expressió d'un gen 16

2.3. ALINEAMENT DE SEQÜÈNCIES 18

2.3.1. AND, ARN I proteins com seqüències 18
2.3.2. *Concepte d'alineament i zones de similitude* 18
2.3.3. Procés d'alineament entre seqüències 19
2.3.4. Tipus d'alineament entre seqüències 19

3. INTRODUCCIÓ AL PROJECTE

3.1. OBJECTIUS DEL PROJECTE 21
3.2. NOM DEL PROGRAMA: "PromoterSimilarity" 24

4. PROMOTER SIMILARITY 25

4.1. EMMAGATZEMAMENT DE LA INFORMACIÓ NECESSÀRIA 27

4.1.1. Tractament de les dades 27
4.1.2. Emmascarament de seqüències 34
4.1.3. *Alineament de seqüències* 38
4.1.4. *Factors de transcripció* 44
4.1.5. Factors conservats 51
4.1.6. Conjunts de factors conservats 54

5. CONCLUSIONS 61

6. GLOSSARI 62

7. BIBLIOGRAFIA 63

1. INTRODUCCIÓ

1.1. MOTIVACIÓ DEL PROJECTE

Aquest ha estat un projecte que no tan sols ha tingut per objectiu ésser entregat com a Projecte de Final de Carrera, sinó també saciar les il·lusions de les dues integrants d'aquest.

Ambdues teníem cert interès en fer un projecte destinat a fins biològics, ja que ens agradava l'àrea en qüestió. Així que ens vam posar en contacte amb el Dr. Xavier Messeguer i Peypoch que, finalment, ens va presentar l'oportunitat de realitzar un projecte en el camp de la bioinformàtica.

Aquest projecte no tan sols ens ha servit per aplicar els coneixements adquirits al llarg de la carrera, sinó també per descobrir la complexitat del món de la investigació. Tal i com s'anirà explicant al llarg d'aquesta memòria, han estat necessaris molts esforços per aconseguir superar les traves trobades durant aquest últim any i, potser no sempre els resultats han estat tan satisfactoris com ens hagués agradat.

Tot i així, ambdues estem satisfetes de la feina feta, pel que hem après en genètica i per la quantitat de coneixements que hem hagut d'aplicar per desenvolupar l'eina. Però sobretot, per l'opció que se'ns ha brindat de crear una aplicació novedosa i sense cap tipus d'antecedent registrat. Potser no hem aconseguit arribar a tots els resultats que ens hagués agradat, però em aconseguit marcar un precedent en aquest camp d'investigació i em aplanat el camí per a la persona que en un futur no molt llunyà, decideixi reprendre'l.

1.2 ESTRUCTURA DEL PROJECTE

La memòria d'aquest projecte presenta una estructura poc habitual degut a dos grans motius. Per una banda, existeix la peculiaritat d'haver estat dut a terme per dues estudiants, enlloc d'haver seguit el model més convencional d'un sol estudiant per projecte. I per altra banda, hi ha la dificultat afegida d'haver escollit un tema d'investigació, en el que els requisits inicials no són tan clars com ho acostumen a ser en altres tipologies de projectes, sinó que s'han anat perfilant durant el desenvolupament del mateix.

Això ha fet que fos difícil traçar una línea divisòria clara en els objectius del projecte, que el separés en dos parts d'igual complexitat i que es poguessin desenvolupar paral·lelament per les dues integrants.

Per tant, es va decidir dividir desenvolupament del projecte per tasques i no per objectius. D'aquesta manera, totes les decisions del projecte així com el disseny en la seva totalitat, ha estat realitzat integrament per ambdues estudiants. El desenvolupament de l'aplicació ha estat efectuat de manera seqüencial, i l'assignació de les tasques s'ha realitzat a mesura que les anteriors s'anaven enllestint.

Degut a la normativa de la Facultat ha estat necessari repartir la descripció del projecte en dues memòries. Hem cregut convenient distribuir els continguts per facilitar la comprensió al lector i per obtenir dues memòries de similar complexitat.

La primera de les memòries (volum I) es centra en el tractament de dades conegudes, la realització de càlculs, i l'emmagatzemament dels resultats obtinguts.

La segona memòria (volum II) descriu el servei prestat per la nostra aplicació. S'ofereix consultar els resultats generats en el procés de la Memòria I i obtenir-ne de nous a partir de dades introduïdes per l'usuari.

1.2 ESTRUCTURA DE LA MEMÒRIA

Aquesta memòria representa la primera part del nostre projecte. Es pretén que el lector tingui un primer contacte amb el projecte desenvolupat i n'entengui el seu funcionament intern i el procés d'obtenció de resultats. Per tal d'aconseguir-ho s'ha dividit en les següents parts:

1. **Introducció.** Presentant un projecte poc convencional com ho és aquest, hem cregut necessari explicar i justificar, primerament, els motius que ens van portar a realitzar un projecte d'aquesta índole. I, en segon lloc, justificar el perquè de com han estat distribuïdes les dues memòries i de quina estructura s'ha seguit en cadascuna d'elles.

2. **Context biològic.** En aquest punt, s'intenta fer una petita aproximació al camp de la genètica entre el que ens hem vist immerses durant aquest últim any. S'expliquen els

coneixements biològics bàsics per poder entendre els objectius d'aquest projecte i la terminologia emprada per explicar-los.

3. **Introducció al projecte.** Normalment, aquest punt és troba al principi d'una memòria però, en el nostre cas, era impossible explicar els objectius del projecte abans d'haver entès perfectament, tots i cadascun, dels conceptes biològics necessaris.

4. **Promoter Similarity.** Un cop explicat l'entorn biològic i plantejats els diferents objectius, ja es pot explicar com s'han dut a terme. Hem dividit l'explicació, separant per cadascun dels mètodes importants desenvolupats per a generar els resultats desitjats. En aquest apartat ens hem centrat molt en les diferents estructures de dades creades per emmagatzemar tota la informació necessària per al funcionament del nostre programa. Això és degut, a que des del inici del projecte, hem apostat per crear un programa amb la major rapidesa possible, per a que un usuari no s'hagués d'esperar gaire per obtenir resultats, i les estructures creades eren bàsiques per assolir aquest objectiu.

2. CONTEXT BIOLÒGIC

2.1. ÀCID DESOXIRRIBONUCLEIC

El genoma constitueix tot el material genètic de les cèl·lules d'un organisme. La major part del material genètic, en animals i plantes, es localitza en el nucli cel·lular, organitzat en cromosomes que són estructures cel·lulars formades per àcid desoxirribonucleïc (ADN).

La funció principal de l'ADN és la fabricació de totes les proteïnes necessàries per al funcionament de l'organisme i la codificació de les instruccions essencials per a formar un ésser viu.

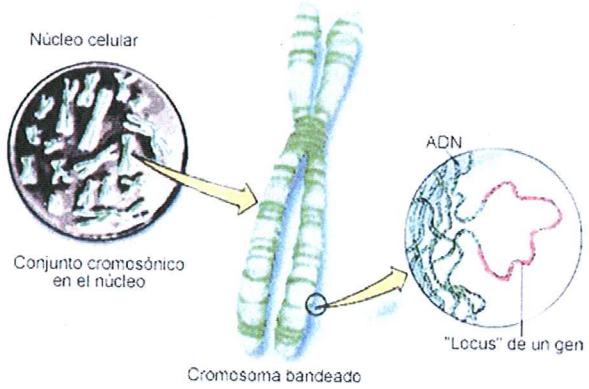

Figura 1

Observem que el nucli està format per un conjunt de cromosomes, i que cadascun d'aquests està format per àcid desoxirribonucleic.

L'ADN és un polinucleòtid, és a dir, està format per la unió de nucleòtids. Cada nucleòtid conté un grup fosfat, a través del qual es realitzen les unions entre nucleòtids, una desoxiribosa (sucre) i una base nitrogenada, que permet diferenciar els nucleòtids en quatre tipus i classificar-los en dos grups de bases: dos puríniques anomenades adenina (A) i guanina (G) i dos pirimidíniques anomenades citosina (C) i timina (T).

L'estructura de l'ADN és una parella de llargues cadenes de nucleòtids enrotllades una al voltant de l'altra, formant una doble hèlix. Aquest parell de cadenes resten unides mitjançant les bases nitrogenades que les formen seguint sempre un mateix model, l'adenina sempre s'enfronta a la timina (A-T) i la guanina a la citosina (G-C). Dues cadenes unides seguint aquesta estructura s'anomenen complementàries.

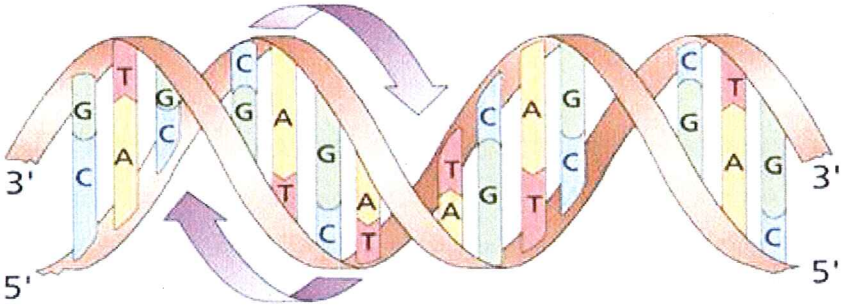

Figura 2

Estructura en hèlix de les dues cadenes d'àcid desoxirribonucleic, unides per complementarietat de bases.

Degut a l'estructura química del nucleòtids, els dos extrems d'una cadena d'àcid nucleic són diferents (5', 3'). L'adhesió de les dues cadenes d'àcid nucleic no es produeix per enllaç covalent (enllaç químic fort) sinó mitjançant un tipus d'enllaç feble: els ponts d'hidrogen. L'adenina (A) s'uneix a la timina (T) mitjançant dos ponts d'hidrogen, i la citosina (C) s'uneix a la guanina (G) amb tres ponts d'hidrogen. Aquestes unions tenen lloc de forma que ambdues cadenes són antiparal·leles, ja que l'extrem 3' d'una s'enfronta a l'extrem 5' de l'altra. Per convenció, una cadena d'ADN sempre es representa en sentit de 5' a 3', que de fet és el sentit en que es llegeix la seva informació.

Aquesta estructura és molt important, ja que cadascuna de les cadenes es pot utilitzar de motlle per a reproduir l'altra, que és el que succeeix durant el procés de replicació. En aquest procés es desenrotlla l'hèlix, mitjançant la interacció d'un enzim anomenat ADN polimerassa, obtenint-se així dues cadenes independents, capaces de sintetitzar la seva cadena complementària. L'exemple més clar en que es produeix la replicació, és en la divisió d'una cèl·lula, moment en el que cal duplicar l'ADN per obtenir parts idèntiques.

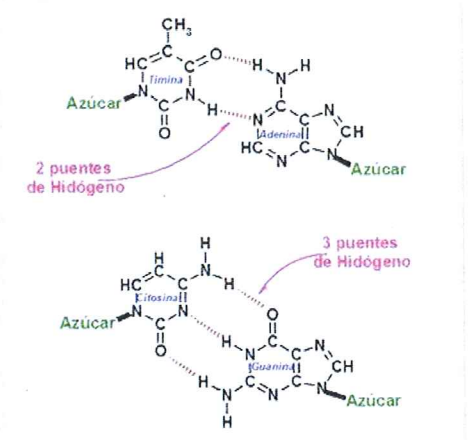

Figura 3

Enllaç químic entre les bases nitrogenades.

2.2. EXPRESSIÓ GÈNICA

2.2.1. Síntesi de proteïnes

Les cèl·lules necessiten proteïnes pròpies pel seu desenvolupament i funcionament. Per això existeix un mecanisme en l'interior de la cèl·lula que construeix les proteïnes en funció de les seves necessitats. Aquest és el que anomenem procés de síntesi de proteïnes (o síntesi proteica).

La síntesi de proteïnes consta de dues etapes: la transcripció, que es produeix en el nucli de les cèl·lules, i la traducció, que té lloc al citoplasma, en uns orgànuls anomenats ribosomes formats per 2 subunitats (la gran i la petita).

En el transcurs de l'evolució, tots els organismes s'han assegurat de que la informació necessària per a sintetitzar les seves proteïnes es trobés present en les seves cèl·lules i en la seva descendència. Químicament aquesta informació resideix en certes regions de l'ADN, conegudes com a gens, que constitueixen un 3% del total del genoma. I amb la divisió cel·lular i la replicació del codi genètic , la transmissió està assegurada.

1. Transcripció

S'entén per transcripció la còpia d'informació d'una regió de l'ADN de doble cadena a una molècula d'àcid ribonucleïc (ARN) de cadena simple anomenat ARN-missatger (ARNm). L'ARN és un polinucleòtid que, a diferència de l'ADN, el sucre dels nucleòtids és ribosa (i no desoxirribosa) i la base timina (T) és substituïda per l'uracil (U). La transcripció és catalitzada per un enzim anomenat ARN-polimerassa. El procés s'inicia amb la separació d'una regió de les dues cadenes d'ADN. Una d'aquestes regions és utilitzada com a motlle per l'ARN-polimerassa per a generar una nova cadena complementària a la utilitzada com a motlle. L'única diferència consisteix en que la timina (T) de l'ADN inicial és substituïda per l'uracil (U) en l'ARNm. Així, per exemple, una seqüència ATGCAT de la cadena motlle, produiria una seqüència UACGUA.

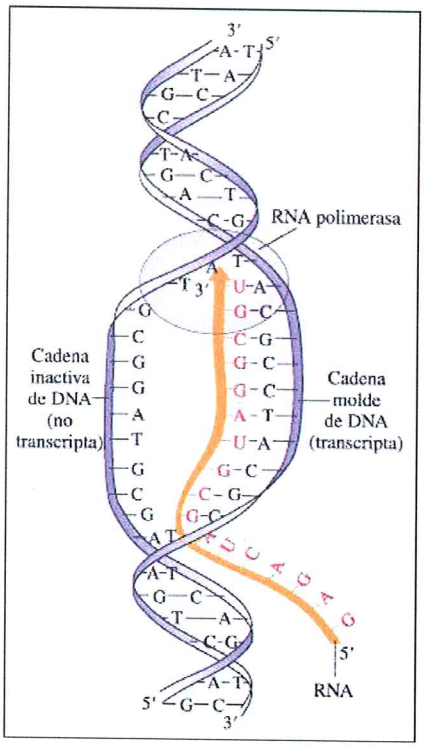

Figura 4
Fase de transcripció, en que es genera una cadena simple de RNA, a partir d'una de les cadenes de l'ADN.

Aquesta cadena d'ARNm és la que dirigeix al citoplasma per a continuar amb el procés de síntesi.

2. Traducció

La traducció és la fase on realment es creen les proteïnes. Per aquest motiu, molts cops la segona fase del procés, rep el nom de síntesi de proteïnes.

Quan l'ARNm arriba al citoplasma cel·lular, cada tripleta de nucleòtids consecutius (codó) especifica un aminoàcid. Donat que l'ARNm conté 4 bases, el nombre de combinacions possibles de grups de 3 és de 64, nombre més que suficient per a codificar els 20 aminoàcids existents. De fet, un aminoàcid pot ser codificat per diversos codons.

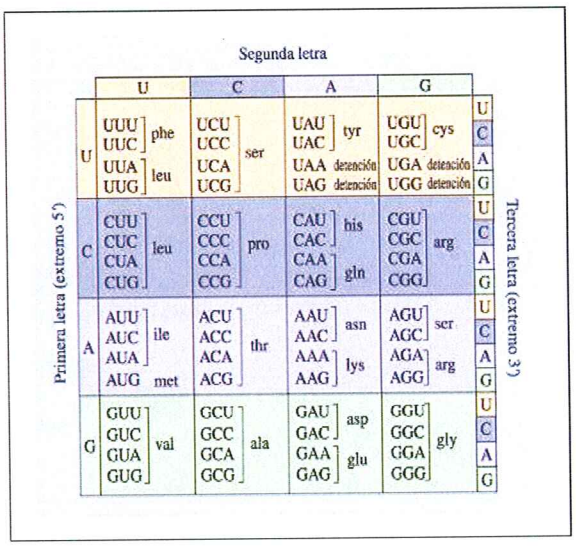

Figura 5
Taula dels 64 aminoàcids possibles.

Existeix un altre tipus d'ARN, l'ARN de transferència (ARNt), en l'estructura del qual destaquen 3 bases nitrogenades que reben el nom d'anticodó. Per cada aminoàcid existeix com a mínim un ARNt (un anticodó específic). Per poder sintetitzar les proteïnes els codons de l'ARNm han de ser reconeguts pels ARNt's, aparellant-se codó i anticodó per complementarietat de bases.

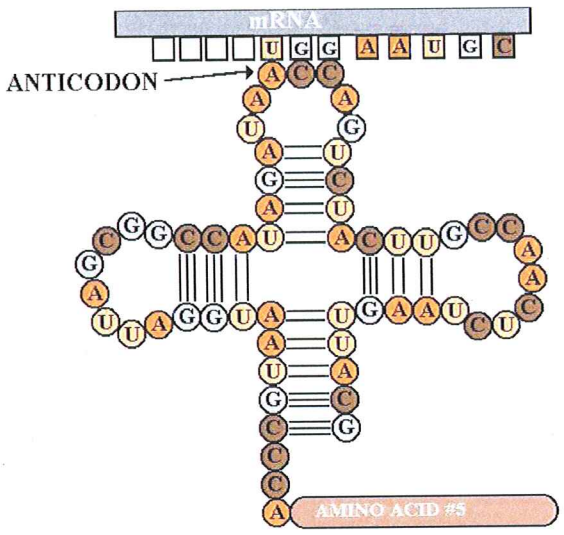

Figura 6
Reconeixement entre en ARNt i el ARNm, en la fase de traducció.

La primera etapa de la síntesi proteica comença quan la unitat més petita del ribosoma s'inserta en l'extrem 5' de l'ARNm, exposant el primer codó, que sempre és AUG al primer anticodó UAC de l'ARNt. La

traducció de l'ARNm és fa per tant sempre en sentit de 5' a 3'.

A partir d'aquest moment es continua llegint l'ARNm i els aminoàcids dels anticodons reconeguts es van adherint a la proteïna en formació.

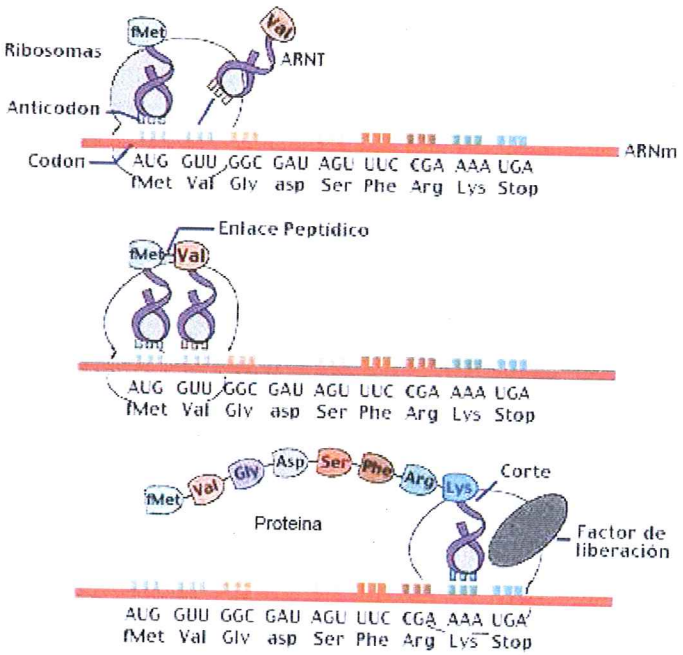

Figura 7

Observem el procés de formació de la proteïna en la fase de traducció del procés de síntesi proteica.
Els codons (ARNm) i els anticodons (ARNt) es reconeixen i en conseqüència es van adherint els aminoàcids existents en els anticodons, formant la proteïna.

En el moment en que es llegeix el codó de finalització, que pot ser UUA, UGA o UAG, no s'hi uneix cap anticodó, i és el moment en que es dona per finalitzada la proteïna.

Finalment, la proteïna creada s'allibera de l'últim ARNt, que també es separa de l'ARNm, desassociant-se de les subunitats ribosòmiques.

D'aquesta manera, tots aquest elements queden lliures per a ser reutilitzats de nou. De fet, és molt freqüent que abans de que finalitzi la síntesi d'una proteïna ja s'iniciï la d'una altra, fent així, que una mateixa molècula d'ARNm sigui emprada per diversos ribosomes simultàniament.

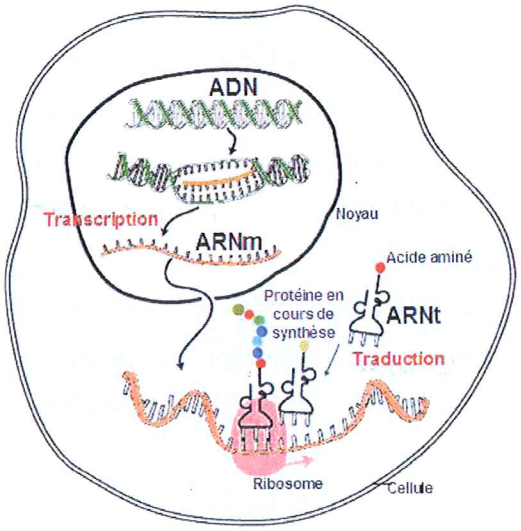

Figura 8

Representació de la síntesi de proteïnes. La fase de transcripció es produeix al nucli, i el ARNm surt del nucli, moment en que es realitza la traducció.

2.2.2. Sistema de regulació de l'expressió d'un gen

El gen és la unitat mínima d'emmagatzemament d'informació en l'ADN i la unitat d'herència ja que es transmet a la descendència.

El gen es divideixen en tres parts seguint un criteri de funcionalitat:

- ***Regió codificant***: Fragment d'ADN que determina la seqüència d'aminoàcids de la proteïna que codifica. Així doncs, és la part que es copia a ARNm durant la fase de transcripció del procés de síntesi de proteïnes. Fins al moment, aquesta ha estat la part considerada més important, degut a que codifica la proteïna, i és per aquest motiu que a aquesta regió se l'anomena gen.
- ***Promotor***: Seqüència específica d'ADN, formada per un nombre variable de bases nitrogenades anteriors a la regió codificant. La seva funció és determinar el moment en que s'ha de manifestar aquella regió codificant, és a dir, el moment en que s'inicia la síntesi de les proteïnes.
- ***Terminador***: Part de l'ADN format per les últimes tres bases del gen que indiquen el final del procés de síntesi de la proteïna en qüestió.

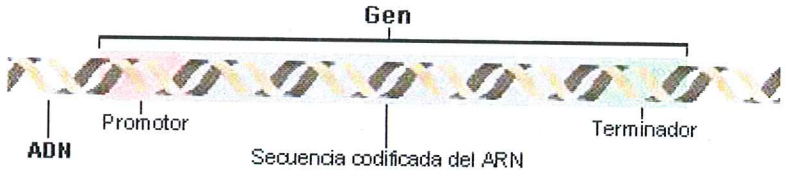

Figura 9
Seqüència d'ADN on apareix un gen amb les seves tres parts: Promotor, regió codificant i terminador.

Al llarg del procés de síntesi de les proteïnes, les cèl·lules han desenvolupat un sistema de regulació que controla quina proteïna és necessària per l'organisme, en quina quantitat i en quin moment cal sintetitzar-la.

El sistema més habitual és regulant la quantitat d'ARNm que es produeix. La quantitat d'ARN que produeix un gen en un moment donat, en el procés de transcripció, depèn de la facilitat amb que l'ARN-polimerassa pot unir-se amb el promotor i iniciar la còpia. Aquesta unió pot veure's facilitada o dificultada per la unió de certes proteïnes a les regions del promotor. Aquestes

proteïnes són el que s'anomenen factors d'inici de la transcripció, o factors de transcripció. La seva presència o capacitat d'unir-se al promotor depèn de factors com estímuls externs, hormones, nutrients, llum... Amb tot s'aconsegueix una regulació de l'expressió del gen i en conseqüència una regulació en la síntesi de la proteïna.

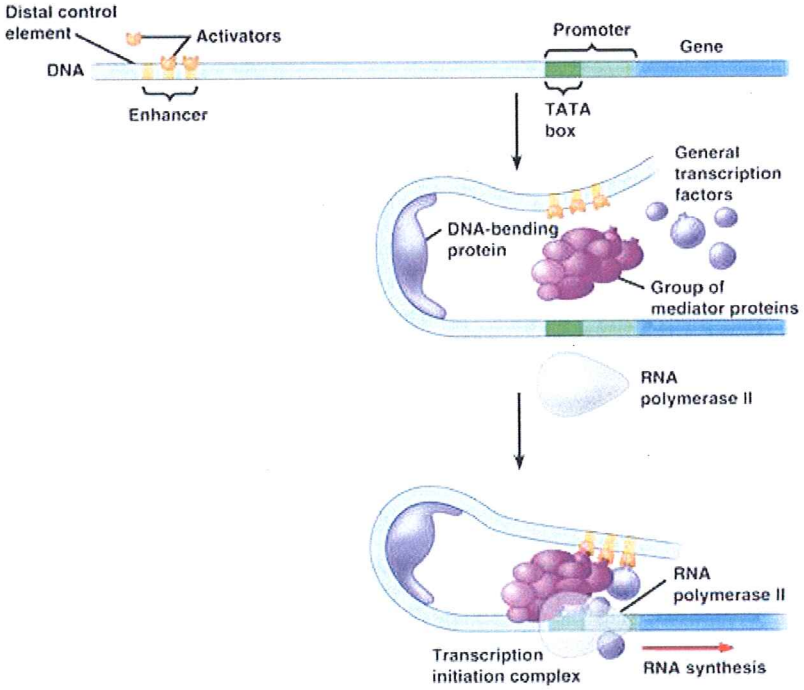

Figura 10

En la transcripció d'un gen es veuen involucrats diversos components que en faciliten o dificultenel procés.

Podem veure com els activadors, juntament amb grups mediadors de proteïnes, l'ARN-polimerassa i els factors de transcripció intervenen en la lectura del promotor i en la posterior generació de l'ARNm.

2.3. ALINEAMENT DE SEQÜÈNCIES

2.3.1. ADN, ARN i proteïnes com seqüències

Es poden interpretar les cadenes de nucleòtids (ADN o ARN) com seqüències formades per 4 lletres (A, C, G i T/U). Així un gen, seria una seqüència d'As, Cs, Gs i Ts que conté la informació per a la síntesi d'una proteïna. Una proteïna també es pot reduir a una seqüència de símbols d'un alfabet de mida 20 (els 20 aminoàcids).

2.3.2. Concepte d'alineament i zones de similitud

L'alineament és una forma de ressaltar possibles zones de similitud entre seqüències. Si aquestes zones codifiquen gens o proteïnes, les similituds existents podrien indicar relacions funcionals i/o evolutives entre ambdues seqüències.

Per tant, l'aparició de zones que presenten similitud entre seqüències pot produir-se o bé perquè ambdues seqüències tenen una funcionalitat semblant, o bé per la possible evolució dels gens d'ambdues seqüències a partir d'un mateix gen original. Els canvis evolutius dels gens (i les seves proteïnes) desencadenen variacions en els organismes que poden donar lloc a noves espècies.

El fet de que diverses posicions de les seqüències es mantinguin invariables, és a dir, es puguin alinear, ens indica que aquestes zones tenen una especial importància per al manteniment de l'estructura i la funció de la proteïna i la seva modificació no ha estat tolerada al llarg de l'evolució.

Així podem afirmar que la comparació entre seqüències és una forma de descobrir quines parts d'aquestes són més importants (estan més conservades) i de descobrir quines proteïnes tenen un origen comú.

2.3.3. Procés d'alineament entre seqüències

Per dur a terme l'alineament, es disposen les dues seqüències a alinear una damunt de l'altra de manera que en una mateixa columna s'hi pot observar una base nitrogenada de cadascuna i es realitza una comparació de bases entre ambdues seqüències per trobar regions comunes.
Aquestes regions han de disposar-se l'una al damunt de l'altra i per aconseguir-ho s'insereixen gaps (espais sense informació genètica) en una o en ambdues seqüències, obtenint-se d'aquesta manera, una fàcil visualització de les regions comunes en ambdues seqüències.

Un cop realitzat l'alineament es poden distingir diverses zones, la més important pel cas que ens ocupa és la regió comuna, també anomenada arrel. El grau de similitud entre les bases nitrogenades d'aquesta zona és el que s'interpreta com a mesura de conservació de la seqüència. Les regions no coincidents de seqüències que comparteixen zones comunes poden interpretar-se com a punts de mutació i els buits, com a "indels" (mutacions d'inserció o eliminació –delete en anglès-) introduïdes en alguna de les dues seqüències.

```
Seqüència 1   TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
Seqüència 2   --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK
```

Figura 11
Exemple d'alineament (FALTA REFER L'EXPLICACIÓ I POSAR UNA IMATGE MÉS ADDIENT).

2.3.4. Tipus d'alineament entre seqüències

Existeixen diversos tipus d'alineament entre seqüències:

- **Alineament global:** Focalitzat en alinear les seqüències en tota la seva extensió, essent de gran utilitat quan les seqüències inicials són similars i de la mateixa llargada aproximadament.

- **Alineament local:** Útil per a seqüències diferenciades en les que es creu que existeixen regions molt similars o motius de seqüències similars en un context més gran.

- **Alineament híbrid o semiglobal:** Intenta localitzar el millor alineament possible que inclogui l'inici d'una de les dues seqüències i el final d'alguna de les dues. Aquest tipus d'alineament pot ser especialment útil quan el final d'una seqüència se solapa amb el final de l'altra, cas en que

l'alineament local i el global són totalment desaconsellables. Un alineament global intentaria forçar l'alineació, estenent-se més enllà de la regió de solapament, mentre que l'alineament local no cobriria totalment la regió solapada.

El problema d'alinear dues seqüències diferents es redueix a trobar quin és el millor de tots els alineaments possibles. Per a obtenir la millor solució, s'ha desenvolupat un sistema de puntuació ("scores"), en el que, per a cada possible alineament, se suma un cert valor per cada resultat positiu (identitat o similaritat entre nucleòtids) i es redueix en un altre valor diferent per a les substitucions o gaps. D'acord amb aquest criteri, el millor dels alineaments serà el que tingui una puntuació més elevada.

S'han desenvolupat diversos algorismes per calcular el millor alineament possible. Els més coneguts són:

- **Neddeleman-Wunsch:** En aquest cas l'alineament òptim ha d'estendre's des del principi fins al final d'ambdues seqüències. Per aquest motiu, aquest algorisme és el que utilitzen els programes d'alineament global.
- **Smith-Waterman:** Modificació de l'anterior, que permet obtenir el millor alineament local (no és necessari que la similaritat s'estengui fins als extrems de la seqüència). En aquest cas, es considera que un alineament és el millor possible si la seva puntuació no es pot incrementar allargant l'alineament per qualsevol dels dos extrems.
- **SIM:** És una modificació recent de l'anterior que detecta també els alineaments subòptims, és a dir, altres alineaments amb puntuació inferior. Aquest mètode genera una llista amb els alineaments locals detectats, començant des de el que té major puntuació.

3. INTRODUCCIÓ AL PROJECTE

3.1. OBJECTIU DEL PROJECTE

Actualment hi ha moltes eines que estudien el comportament, l'estructura i la funcionalitat dels gens. Ara bé, les zones anteriors als gens, zones que inhibeixen o desinhibeixen el seu caràcter, ja que intervenen directament en la determinació de com, quan i quina proteïna cal generar, són entitats poc estudiades i de les que queda molt per descobrir segons els experts genetistes.

És en aquesta zona, en el promotor, on es troba l'entrellat de l'expressió genètica i la clau de l'evolució de les espècies, i és en la que cal incidir més els estudis genètics ja que és la més desconeguda.

Ens disposem a realitzar una eina que permeti determinar el grau de similaritat entre seqüències promotores d'una espècie, visualitzar els factors de transcripció que s'hi localitzen i mostrar quins són els factors que hi apareixen conservats, per tal de poder concloure possibles relacions funcionals i/o evolucions gèniques en les espècies estudiades.

Aquest és l'objectiu final del programa, l'objectiu general, però en realitat aquest s'assoleix amb la interacció d'una sèrie d'objectius més petits detallats a continuació:

1. **Tres espècies a estudiar:** Es pretén poder observar el grau de similaritat entre seqüències promotores de tres espècies concretes: gallina, humà i ratolí. Aquestes han estat escollides per l'usuari final del nostre programa, investigadors del *Parc de Recerca Biològica de Barcelona (PRBB)*, degut a que són tres de les espècies amb les seqüències de promotors més estudiades i de les quals es creu que poden concloure resultats genètics més enriquidors i favorables.

2. **Modificació de la informació de les espècies:** El programa disposa d'un conjunt de fitxers d'entrada per cada espècie on existeixen les seqüències promotores conegudes i amb les que juntament amb una seqüència d'entrada (a elecció de l'usuari) són processades pel programa per la futura generació de resultats biològics. Aquest conjunt de fitxers pot modificar-se en qualsevol moment a petició de l'administrador del programa i el programa utilitzarà la nova informació en el seu funcionament. Això implica la durabilitat del projecte, ja que si en un futur pròxim es descobrissin més seqüències promotores per a alguna de les espècies o alguna de les ja existents sofrís alguna modificació, el programa no

quedaria obsolet, ja que l'administrador podria introduir la nova informació i obtenir resultats actualitzats.

3. **Generació de resultats de similaritat:** L'objectiu primordial d'aquest projecte és la generació de resultats que seran mostrats a l'usuari mitjançant una interfície, però per tal d'obtenir aquests resultats són imprescindibles unes dades d'entrada. Per una banda, els fitxers amb la informació de les seqüències promotores conegudes per a les tres espècies escollides pels investigadors i per altra, una seqüència promotora que escollirà l'usuari i que serà comparada amb les seqüències de l'espècie que prefereixi l'usuari. Existeixen dues possibilitats per a la seqüència promotora escollida per l'usuari, amb les que apareixen dos nous objectius:

   3.1. **Generació de resultats de similaritat per una seqüència introduïda per l'usuari:** La seqüència promotora o query pot ser qualsevol seqüència de 2000 bases nitrogenades que l'usuari desitgi. Així, els resultats obtinguts serien els generats per la comparació d'aquesta nova seqüència introduïda per l'usuari amb la resta de seqüències promotores conegudes per a l'espècie que escolleixi l'usuari d'entre les tres de les que disposa.

   3.2. **Generació de resultats de similaritat per una seqüència existent en alguns dels fitxers d'entrada:** L'usuari pot escollir com a seqüència promotora o query qualsevol de les seqüències existents per a alguna de les tres espècies. Així, els resultats obtinguts serien els generats per la comparació d'aquesta seqüència escollida entre les existents amb la resta de seqüències promotores conegudes de l'espècie de la que forma part.

4. **Factors conservats:** Un dels resultats obtinguts de la comparació d'una seqüència escollida per l'usuari, a la que anomenarem query, amb totes les seqüències de promotors d'una espècie, és la visualització dels factors conservats entre la seqüència query i la resta. És a dir, els factors de transcripció que es troben en zones alineades entre la query i alguna de les altres seqüències i que, a més, es troben en la mateixa posició respecte l'inici de l'alineament en el que estan inclòs.

5. **Estadístics i alineaments.** Un cop l'usuari ha visualitzat els factors de transcripció conservats entre la seqüència escollida i totes les de l'espècie indicada, se li ofereix l'opció de visualitzat, per una banda, la informació de tots els alineaments produïts en el procés de càlcul dels resultats. Per altra banda, també pot consultar, per cada seqüència resultant, la proporció dels seus factors de transcripció que han aparegut en zones alineades, en zones no alineades o en conservades.

6. **Arbres de factors:** Aquest objectiu ha sorgit a meitat del projecte en substitució d'un altre. En un primer moment, es va pensar definir com a objectiu la millora de l'eficiència del programa PROMO, programa del grup *alggen* que és utilitzat per la nostra aplicació per obtenir certs resultats, tal i com s'explicarà més endavant. Però a mesura que anava avançant el projecte, es va pensar que era una millor opció que, un cop mostrats els resultats dels factors de transcripció a l'usuari, aquest pogués escollir-ne un conjunt. Llavors, se li mostrarien a l'usuari només les seqüències en les que apareguessin tots els factors de transcripció del conjunt com a conservats.

7. **Disseny de l'aplicació:** Finalment és necessària una interfície gràfica que permeti prestar la quantitat de serveis mencionats anteriorment a l'usuari

En aquesta memòria, s'expliquen aquestos objectius des de la vessant del servidor. Es pretén generar tots els resultats corresponents als esmentats entre els punts de l'1 al 6 per a totes les seqüències de promotors coneguts per a tres espècies (gallina, humà i ratolí).

3.2. NOM DEL PROGRAMA: "PromoterSimilarity"

El programa, tal i com es descriu a l'apartat anterior, pretén determinar el grau de similaritat entre seqüències promotores d'una espècie, un cop coneguts els factors de transcripció que s'hi localitzen i sabent quins apareixen conservats. Amb aquests resultats serà llavors quan els experts genetistes podran concloure possibles relacions funcionals o possibles evolucions genètiques.

Els experts genetistes, usuaris finals del nostre programa, han decidit anomenar al programa "**PromoterSimilarity**".

Figura 12
Logo de l'aplicació Promoter Similarity