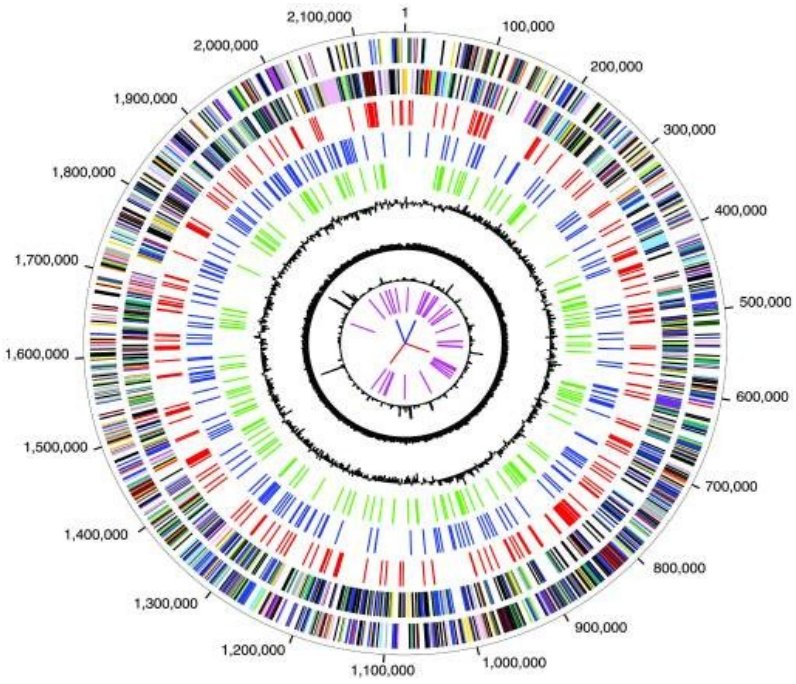
A la recerca del genoma fosc en bacteris: anàlisi del genoma de *Chlorobaculum tepidum*



Guillem García Martínez

Bell-lloc del Pla, 21 de novembre de 2011

A la recerca del genoma fosc en bacteris.
Anàlisi del genoma de *Chlorobaculum tepidum*

Alumne:
Guillem García Martinez

Tutor:
Sr. Manel Montoliu

Girona, curs 2011-2012

El que és vàlid pel bacteri ho és per l'elefant

Jacques Monod (1910-1976)

RESUM

En genètica, es coneix com a matèria fosca, la part del genoma la qual no es coneix la seva funció. Fins ara només es coneix la seva existència en organismes eucariotes. En aquest treball es vol esbrinar si els procariotes (bacteris) també tenen matèria fosca o quelcom similar.

Hem analitzat i comparat les estructures secundàries dels ARN missatgers de 6 grups de gens o COGS (5 grups funcionals i un grup de funció hipotètica --HYP) a partir del genoma complet de *Chlorobaculum tepidum,* en base a tres paràmetres d'estabilitat: longitud dels dúplex autocomplementaris, nombre de braços o "*stems*", i energia específica de formació (EEF) calculats online al servidor RNAfold Webserver.

Els paràmetres d'estabilitat dels gens hipotètics (HYP) o no classificats, comparats amb els altres grups, són significativament diferents (amb $P<0,05$) respecte als dels altres cinc grups funcionals. A més, es destacable que la longitud dels gens HYP és gairebé la meitat ($P=0,000$) a la dels gens funcionals.

Per tant, es podria dir que els ARNm derivats dels gens HYP del bacteri *Chlorobaculum tepidum* són inestables i curts amb la qual cosa es podria pensar que la seva funció no és específicament la de codificar proteïnes, almenys tal com les coneixem, i per aquest motiu apareixen com a no classificats en les llistes de categories del genoma. Aquest ADN diferent a la resta en bacteris podria ser un indici de l'existència de material genètic equivalent o similar a la matèria fosca dels eucariotes.

Índex de continguts

1. INTRODUCCIÓ

Els darrers anys s'ha avançat molt en el coneixement del genoma dels éssers vius. S'ha obtingut la seqüència del genoma de molts animals, plantes, microorganismes, etc. Una part d'aquest genoma codifica per proteïnes estructurals i funcionals. La resta, no codificant, es coneix com el "genoma fosc". Descobriments recents indiquen que aquest genoma fosc no és inservible com fins ara es pensava, i possiblement tingui un important paper en la regulació gènica dels éssers vius.

Els procariotes (bacteris i arquees) són organismes amb un sol cromosoma circular y covalentment tancat (en anglès: cccDNA) on aparentment tota la matèria genètica és informativa. En conseqüència, tot fa pensar que els procariotes no presentarien "regions fosques" al seu ADN.

L'objectiu d'aquest treball és analitzar el genoma complert d'un bacteri (*Chlrorobaculum tepidum*) per tal de veure si tots els seus gens presenten les mateixes característiques o pel contrari existeixen parts susceptibles de ser considerades com "genoma fosc"

En aquest treball farem servir eines bioinformàtiques per analitzar l'ARN produït en la transcripció d'aquest genoma per estudiar-ne les principals característiques i utilitzar la estadística per intentar trobar les diferències entre el genoma fosc i el funcional. L'objectiu d'aquest treball és investigar la funció de l'ARN no codificant (ncRNA) mitjançant eines bioinformatiques i intentar aportar una nova informació per a la biologia molecular.

1.1. L'ADN

1.1.1. Definició

L'ADN és una de les macromolècules més complexes del nostre cos. Està situada al nucli de les cèl·lules eucariotes o lliure en el citoplasma en forma de molècula circular en les cèl·lules procariotes. La seva seqüència de nucleòtids conté tota la informació genètica de l'ésser viu.

1.1.2. Estructura

La seva estructura en doble hèlix va ser descoberta per els científics James Watson i Francis Crick al 1953. Dos filaments de nucleòtids situats de forma antiparal·lela de tal manera que la direcció de una és de 5' a 3' (aquests darrers nombres indiquen en quina direcció van les bases nitrogenades i de quina manera estan col·locades) i l'altre de 3' a 5'. Entre les bases nitrogenades es formen ponts d'hidrogen que donen consistència a la molècula. Quant als filaments, són complementaris: Adenina-Timina dos ponts d'hidrogen; Citosina-Guanina tres ponts d'hidrogen. L'estructura de l'ADN en les cèl·lules eucariotes és lineal, al contrari que les cèl·lules procariotes, que és circular.

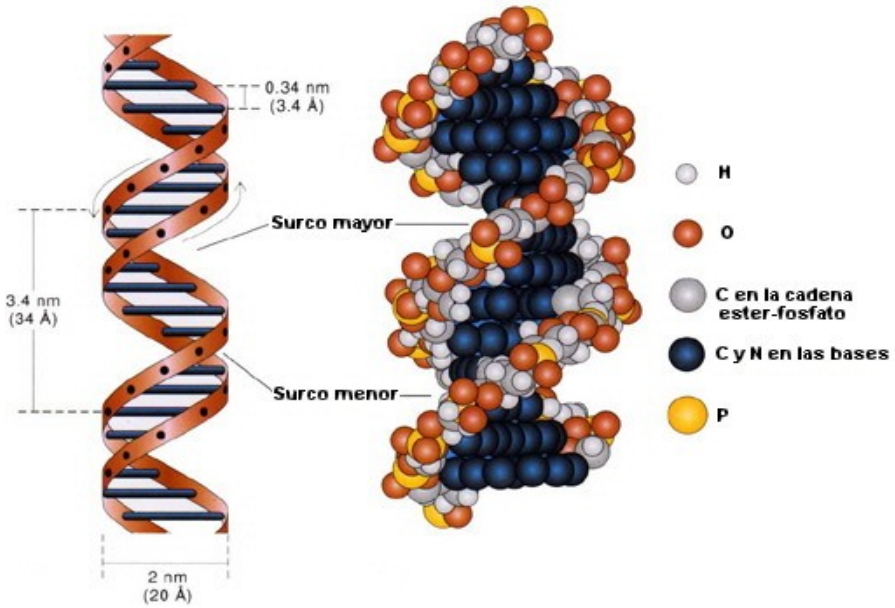


Figura1. Esquema de l'estructura en doble hèlix de l'ADN

1.2. L'ARN

1.2.1. Definició

L'ARN és un àcid nucleic format per una cadena de ribonucleòtids. Està present tant en les cèl·lules eucariotes com les procariotes i és l'únic material genètic d'alguns virus.

1.2.2. Estructura

És lineal i de cadena senzilla. Té des de 75 a 200 nucleòtids. Tant està en el nucli com al citoplasma. L'ARN s'obté del proces anomenat transcripció, on l'ADN es transforma en ARN. Estructura secundària de l'ARN és un plegament de l'estructura primària amb auto complementarietat de les bases, que al plegar-se allibera energia, anomenada energia de formació

1.2.3. Tipus

Hi han diferents tipus d'ARN. Els tipus més coneguts d'ARN són: No només hi ha aquests dos tipus d'ARN, també existeix

- L'ARNm porta la informació sobre la seqüència d'aminoàcids de la proteïna des de l'ADN , lloc en ha estat inscrita, fins el ribosoma, lloc on es sintetitzen les proteïnes de la cèl·lula. Per tant, és una molècula intermediària entre l'ADN i la proteïna. En el procés de desenvolupament de l'ARNm, hi participa un complex format per 5 ribonucleoproteïnes, anomenat espliceosoma, que té la funció d'eliminar els introns dels precursors de l'ARNm.

- L'ARNt (ARN de transferència) són curts polímers de uns 80 nucleòtids que transfereixen un aminoàcid especific al polipèptid en creixement; s'uneixen a llocs específics del ribosoma durant la traducció. Tenen un lloc específic per a la fixació del aminoàcid(extrem 3') un anticodó format per un triplet de nucleòtids que s'uneixen al codó complementari de l'ARNm mitjançant ponts d'hidrogen.

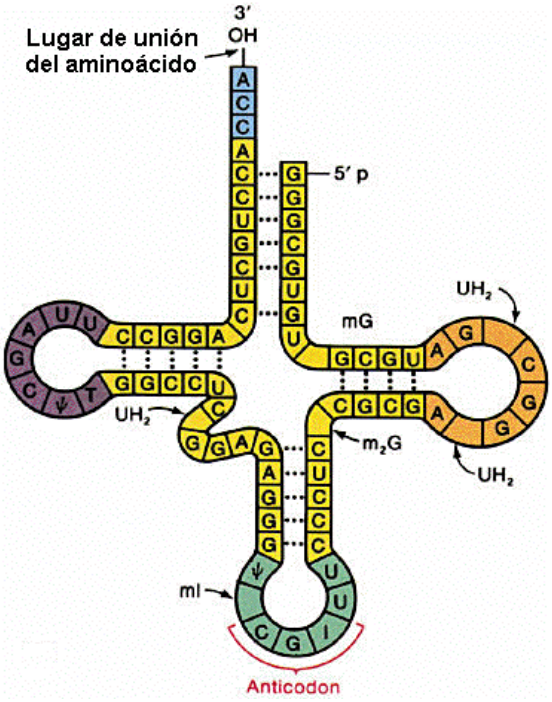


Figura 2. Estructura secundària de l'ARN de transferència mostrant els braços o *stems* amb regions autocomplementàries

- L'ARNr (ARN ribosòmic) es troba combinat amb proteïnes per formar ribosomes, on representa una 2/3 parts de ells mateixos. L'ARNr és molt abundant y representa el 80% de l'ARN trobat al citoplasma de les cèl·lules eucariotes.
- L'ARNhn (ARN heterogeni nuclear) és el precursor de l'ARNm. La seva funció consisteix en eliminar les seqüències sense sentit.
- L'ARNn (ARN nuclear) és una petita molècula d'àcid ribonucleic, sintetitzada i localitzada en el nuclèol de les cèl·lules eucariotes, a partir de la transcripció de l'ADN, format per una curta seqüència d'entre 100 a 300 nucleòtids, i que es precursor i indispensable per a la síntesi de part de l'ARN ribosòmic.
- ARN petit nuclear (snoRNA) es troben en “los cuerpos de Cajal”[1]. Dirigeixen la modificació de nucleòtids d'altres ARN. Transforma alguna base nitrogenada (“A-T-G-C”) en alguna altra base nitrogenada, per

1Els “cuerpos de Cajal” són llocs d'ensamblatge o modificació de la maquinària de transcripció del nucli.

tenir la complementarietat necessària.

- Els ARN no codificant (ncRNA) són un grup gran d'ARN que es transcriuen, però no és tradueix en proteïnes. Els ncRNA poden produir molècules d'ARN funcional en el lloc de la codificació de proteïnes i s'ha trobat que tenen un paper en una gran varietat de processos cel·lulars, com la regulació de la transcripció, el processament de l'ARN i la seva modificació, l'estabilitat de l'ARNm, i fins i tot la degradació de proteïnes. En un recent estudi de l'institut Wistar, situat a Filadèlfia, Estats Units, han descobert la "habilitat" de llargues cadenes de ncRNA (ARN no codificant) per promoure l'expressió de gens, el que significa que encara que ells no són capaços de traduir a proteïnes, les seves transcripció com ncRNA actua com un potenciador de l'expressió de proteïnes per part d'un organisme viu. El ncRNA és una molècula d'ARN funcional que no es transcriu en una proteïna. Aquest ncRNA va ser descobert per primera vegada al 1868 per Friedrich Miescher, biòleg suis i un dels descobridors dels àcids nucleics, i al 1939 ja s'havia implicat l'ARN a la síntesi proteica.

Dècades més tard, Francis Crick, famós científic que va ser un dels descobridors de la doble hèlix de l'ADN, va preveure un component funcional de l'ARN que controlava la traducció de ADN a ARN.

- **Material genètic fosc**: El material genètic fosc, també conegut com a *non-coding DNA* (ADN no codificant), és aquell el qual no codifica a proteïna. En una cèl·lula eucariota, només codifica el 5% del seu material genètic i l'altre 95% és aquest material genètic no codificant.
  - **Exons i introns**: Els exons són les regions d'un gen que no són separades durant el proces anomenat *splicing* (procés en el qual actua els espliceosomes) o empalmament de l'ARN, i per tant es mantenen a l'ARNm. En els gens que codifiquen una proteïna, són els exons els que contenen la informació per produïr la proteïna codificada en el gen. En

aquests casos, cada exó codifica una porció específica de la proteïna completa, de manera que el conjunt d'exons formi la regió codificant del gen. En cèl·lules eucariotes, els exons d'un gen estan separats per regions llargues d'ADN, anomenades introns, que no codifiquen. Aquests introns només han sigut localitzats en les cèl·lules eucariotes. En les procariotes, no existeixen o no han estat identificats fins al dia d'avui.

- **Material genètic fosc en cèl·lules procariotes:** El genoma dels eucariotes normalment és més complexe i més llarg que el genoma de les procariotes. Només es coneix el genoma complet d'algunes cèl·lules procariotes, i d'altres que només alguna part del genoma. D'aquesta part que no es coneix s'anomena material genètic fosc. El bacteri que nosaltres hem triat per estudiar té un 30% de matèria fosca en el seu genoma, cosa que ens facilita l'estudi d'aquest material.

1.3.-*Chlorobaculum tepidum*

El *Chlorobaculum tepidum* és un bacteri que pertany al grup "phylum" *Chlorobia,* un bacteri verd del sofre ("green sulfur"). És un bacteri verd del sofre perquè en la seva fotosíntesis ha canviat l'oxigen, utilitzat per la majoria, per el sofre.

$$H_2S + CO_2 \xrightarrow{h\nu} S^0 + <CH_2O>_n$$

Aquest ésser viu en unes condicions estrictament anaeròbiques sota la superfície de l'aigua. Solen formar poblacions a l'hipolímnion de llacs eutròfics[2]. .

Hem triat el *Chlorobaculum tepidum* per raons molt senzilles: els bacteris tenen el genoma més senzill que les cèl·lules eucariotes, per la raó dels introns, ja que al ser una seqüència sense sentit ens dificultaria la feina, ja que només estan en les cèl·lules eucariotes, perquè els bacteris són el model més utilitzat quan parlem del camp de la microbiologia i també perquè és un dels únics bacteris que té més d'un

2Els llacs eutròfics es caracteritzen per contenir molta matèria orgànica, la qual esgota l'oxigen del fons (hipolímnion) en sedimentar tot creant condicions anaeròbiques on aquests bacteris s'hi desenvolupen si tenen prou llum.

30% del seu genoma desconegut, que no es coneix la seva funció.

El genoma complert d'aquesta bacteri va ser seqüenciat i publicat al 2002 per Donald Bryant[3]. El seu genoma complet consta de 2,5 megabases (Mb). A la pàgina web del GenBank, ens mostra primerament el seu genoma en aminoàcids; per tant, aquesta pàgina ens dona una opció per si volem veure el seu genoma en bases nitrogenades, per veure el seu CDS (Coding DNA Sequence).

- **Comparació entre genomes:**

Si comparéssim el genoma d'un bacteri amb tot el genoma conegut com ara l'*Escherichia coli,* amb el bacteri escollit per nosaltres, el *Chlorobaculum tepidum,* hi podríem veure algunes diferències importants que ajuden a explicar les raons d'escollir aquest darrer com a model per al nostre estudi.

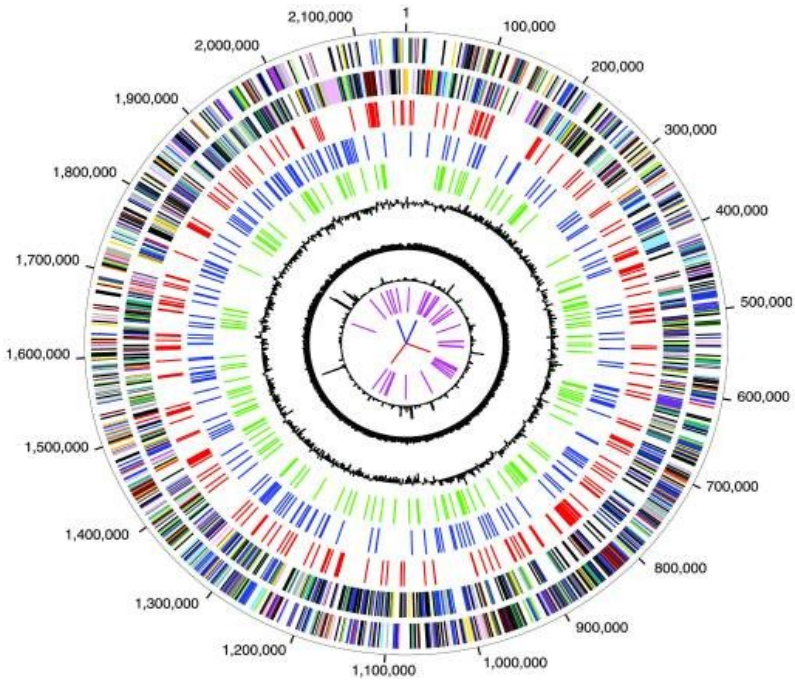


Figura 3. Imatge esquematitzada del genoma complert del bacteri *Chlorobaculum tepidum* amb 2,5 Mbases

3Donald Bryant ocupa la plaça de professor “Ernest C. Pollard” en Biotecnologia a la Universitat de Pennsylvania on és Catedràtic de Bioquímica i Biologia Molecualr.
http://bmb.psu.edu/directory/dab14

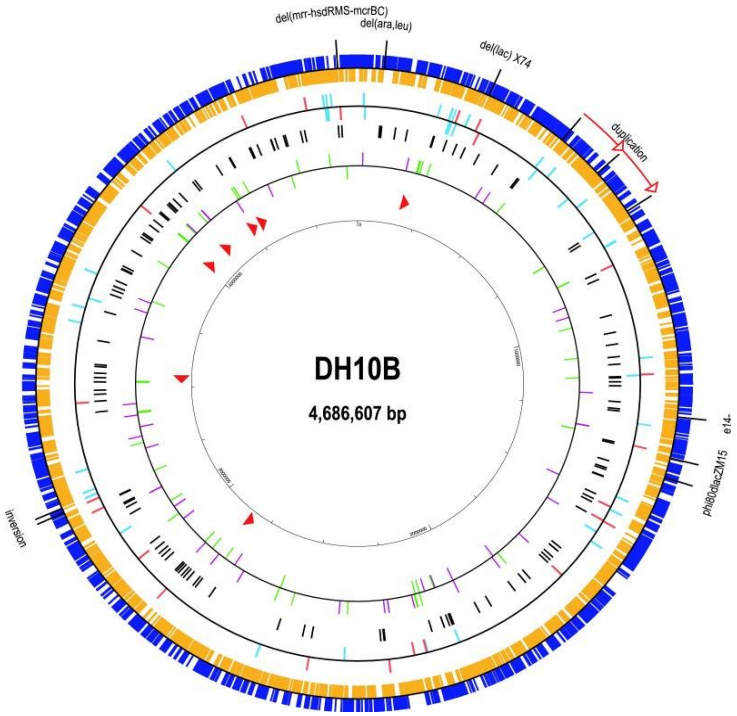


Figura 4. Imatge esquematitzada del genoma complert del bacteri *Escherichia coli* amb 4,7 Mbases

El genoma del *Chlorobaculum tepidum* és circular com correspon ales cèl·lules procariotes. En la imatge superior, si ens hi fixem, podem veure que la línia més superficial no és compacte, sinó que te petits forats. Aquests petits forats és el lloc on es situen els gens no classificats, és a dir, els gens que no tenen funcions conegudes. La raó de que hi hagi tant d'espai en aquesta línia és perquè el 25% del genoma d'aquest bacteri no ha pogut ésser classificat. Però si ens fixem en la imatge inferior dels genomes, podem veure el genoma complert de l'*Escherichia coli.* En la línia superficial els espais són molt menor, comparativament. Els espais que hi han, són substituïts per altres “barres” que representen els gens. Això corrobora que en el genoma de *Escherichia coli* el tant per cent de gens no classificats és molt baix.

2. MATERIALS I MÈTODES:

2.1.- Anàlisi dels gens:

Per començar a treballar i a analitzar aquest bacteri, necessitem el genoma complert del *Chlorobaculum tepidum* però només el podem aconseguir en una pàgina web anomenada GenBank (http://www.ncbi.nlm.nih.gov). El GenBank és la base de dades de seqüències genètiques de l'institut nacional de Salut (NIH) dels Estats Units; un recull anotat de totes les seqüències d'ADN a disposició pública. Un cop arribat al lloc on ens mostra una imatge del genoma,i també ens surt el seu genoma complert esquematitzat en un quadre mostrant-nos els gens del bacteri classificats per funcions. En la figura de baix ens mostra els grups de gens que hem seleccionat al atzar: 5 grups funcionals i 1 grup amb funció desconeguda.

Code	COGs	Description	(% in sequence)	(% in genome)	(% in genus)	(% in Bacteroidetes/Chlorobi group)	(% in Bacteria)
J	138	Translation	5.6212	5.6212	5.8463	4.8172	4.2968
A	0	RNA processing and modification	0.0000	0.0000	0.0000	0.0076	0.0148
K	66	Transcription	2.6884	2.6884	2.8704	4.3028	5.9643
L	105	Replication, recombination and repair	4.2770 ●	4.2770	4.3267	5.1871	4.8261
B	0	Chromatin structure and dynamics	0.0000	0.0000	0.0000	0.0101	0.0269
D	20	Cell cycle control, mitosis and meiosis	0.8147 ●	0.8147	0.8231	0.7893	0.7650
Y	0	Nuclear structure	0.0000	0.0000	0.0000	0.0000	0.0000
V	35	Defense mechanisms	1.4257 ●	1.4257	1.4141	1.6522	1.2590
T	47	Signal transduction mechanisms	1.9145	1.9145	2.1739	3.2258	3.9952
M	148	Cell wall/membrane biogenesis	6.0285	6.0285	6.6484	6.6467	4.3931
N	2	Cell motility	0.0815	0.0815	0.1477	0.2623	1.4575
Z	0	Cytoskeleton	0.0000	0.0000	0.0000	0.0152	0.0116
W	0	Extracellular structures	0.0000	0.0000	0.0211	0.0051	0.0154
U	35	Intracellular trafficking and secretion	1.4257	1.4257	1.6252	1.4583	1.7757
O	77	Posttranslational modification, protein turnover, chaperones	3.1365	3.1365	3.5036	2.8622	2.9290
C	175	Energy production and conversion	7.1283	7.1283	7.4926	4.8476	4.8217
G	77	Carbohydrate transport and metabolism	3.1365	3.1365	3.2925	4.2052	4.9545
E	133	Amino acid transport and metabolism	5.4175	5.4175	5.7408	5.4633	7.2726
F	53	Nucleotide transport and metabolism	2.1589 ●	2.1589	2.3217	1.9385	1.7769
H	126	Coenzyme transport and metabolism	5.1324	5.1324	5.1499	3.5590	2.9700
I	57	Lipid transport and metabolism	2.3218	2.3218	2.5116	2.2211	2.8606
P	117	Inorganic ion transport and metabolism	4.7658	4.7658	5.0232	4.5764	4.5898
Q	37	Secondary metabolites biosynthesis, transport and catabolism	1.5071	1.5071	1.6041	1.4647	2.2675
R	249	General function prediction only	10.1426	10.1426	10.8274	10.1678	10.4609
S	120	Function unknown	4.8880 ●	4.8880	5.4664	5.0072	6.1400
-	638	Not in COGs	25.9878 ●	25.9878	21.1693	25.3072	20.1551

Figura 5. Taula de classificació dels grups dels gen del *Chlorobaculum tepidum*. Els punts vermells indiquen els grups que s'han estudiat en aquest treball.

Al arribar a aquesta pàgina, s'obria una d'aquestes files seleccionades amb el punt vermell a la *Figura 5* on mostrava la seqüència dels gens amb la funció específica d'aquell grup. Tot seguit es prem el link enllaçant de cada gen on t'envia a una plana web on em mostrava la seqüència completa d'aquell gen.

Gens utilitzats: S'han seleccionat al atzar 5 grups específics de gens amb funcions conegudes entre tots els grups o COGs [4] identificats en la pàgina web del GenBank.

Les funcions dels 5 grups seleccionats estan recollides a la taula següents

Identificació COG	**Funció**
D	control del cicle cel·lular,
F	Transport i metabolisme de nucleòtids
L	Replicació, recombinació i reparació
S	Proteïnes que s'assemblen molt a les que, un cop traduïdes, tenen funció
V	Mecanisme de defensa

A part d'aquests 5 grups escollits, s'ha analitzat l'únic grup de gens que no tenen funció

4COG de l'anglès Cluster of Ortlhologous Groups of proteins. Fa refeència a gens que codifiquen per proteïnes amb un mateix origen evolutiu, amb funcions relacionades.

hypothetical protein CT0005 [Chlorobium tepidum TLS]

NCBI Reference Sequence: NP_660911.1

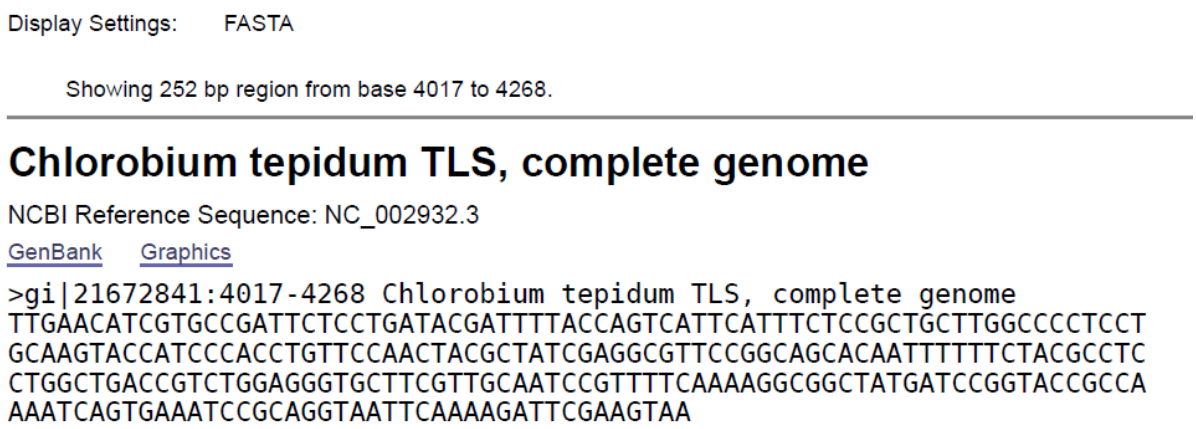
FASTA Graphics

Go to:

```
LOCUS       NP_660911               83 aa            linear   BCT 28-OCT-2011
DEFINITION  hypothetical protein CT0005 [Chlorobium tepidum TLS].
ACCESSION   NP_660911
VERSION     NP_660911.1  GI:21672846
DBLINK      Project: 57897
DBSOURCE    REFSEQ: accession NC_002932.3
KEYWORDS    .
SOURCE      Chlorobium tepidum TLS
  ORGANISM  Chlorobium tepidum TLS
            Bacteria; Chlorobi; Chlorobia; Chlorobiales; Chlorobiaceae;
            Chlorobaculum.
REFERENCE   1  (residues 1 to 83)
  AUTHORS   Eisen,J.A., Nelson,K.E., Paulsen,I.T., Heidelberg,J.F., Wu,M.,
            Dodson,R.J., Deboy,R., Gwinn,M.L., Nelson,W.C., Haft,D.H.,
            Hickey,E.K., Peterson,J.D., Durkin,A.S., Kolonay,J.L., Yang,F.,
            Holt,I., Umayam,L.A., Mason,T., Brenner,M., Shea,T.P., Parksey,D.,
            Nierman,W.C., Feldblyum,T.V., Hansen,C.L., Craven,M.B., Radune,D.,
            Vamathevan,J., Khouri,H., White,O., Gruber,T.M., Ketchum,K.A.,
            Venter,J.C., Tettelin,H., Bryant,D.A. and Fraser,C.M.
  TITLE     The complete genome sequence of Chlorobium tepidum TLS, a
            photosynthetic, anaerobic, green-sulfur bacterium
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 99 (14), 9509-9514 (2002)
   PUBMED   12093901
REFERENCE   2  (residues 1 to 83)
  AUTHORS   Eisen,J.A., Nelson,K.E., Paulsen,I.T., Heidelberg,J.F., Wu,M.,
            Dodson,R.J., Deboy,R., Gwinn,M.L., Nelson,W.C., Haft,D.H.,
            Hickey,E.K., Peterson,J.D., Durkin,A.S., Kolonay,J.L., Yang,F.,
            Holt,I., Umayam,L.A., Mason,T., Brenner,M., Shea,T.P., Parksey,D.,
            Nierman,W.C., Feldblyum,T.V., Hansen,C.L., Craven,M.B., Radune,D.,
            Vamathevan,J., Khouri,H., White,O., Gruber,T.M., Ketchum,K.A.,
            Venter,J.C., Tettelin,H., Bryant,D.A. and Fraser,C.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-APR-2002) The Institute for Genomic Research, 9712
            Medical Center Dr, Rockville, MD 20850, USA
REFERENCE   3  (residues 1 to 83)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (08-APR-2002) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AAM71253.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..83
                     /organism="Chlorobium tepidum TLS"
                     /strain="TLS"
                     /db_xref="taxon:194439"
     Protein         1..83
                     /product="hypothetical protein"
                     /calculated_mol_wt=9356
     Region          1..68
                     /region_name="DUF37"
                     /note="Domain of unknown function DUF37; cl00506"
                     /db_xref="CDD:186043"
     CDS             1..83
                     /locus_tag="CT0005"
                     /coded_by="NC_002932.3:4017..4268"
                     /transl_table=11
                     /db_xref="GeneID:1006132"
ORIGIN
        1 mnivpillir fyqsfispll gpsckyhptc snyaieafrq hnffyaswlt vwrvlrcnpf
       61 skggydpvpp ksvksagnsk dsk
//
```

Figura 6. Entrada a al GenBank corresponent al gen escollit.

La figura 6 mostra l'exemple de com és l'entrada al GenBank, d'un gen qualsevol de C. tepidum, amb la seqüència amb aminoàcids del gen en qüestió. Per obtenir la seqüència d'ADN , s'ha de prémer el link de CDS (*Coding DNA Sequence*); per veure la seqüència en bases nitrogenades. Canviar la configuració de la pantalla de GenBank a FASTA[5] text, on es pot copiar la seqüència sencera del gen sense problemes.


Display Settings: FASTA

Showing 252 bp region from base 4017 to 4268.

Chlorobium tepidum TLS, complete genome

NCBI Reference Sequence: NC_002932.3

GenBank Graphics

```
>gi|21672841:4017-4268 Chlorobium tepidum TLS, complete genome
TTGAACATCGTGCCGATTCTCCTGATACGATTTTACCAGTCATTCATTTCTCCGCTGCTTGGCCCCTCCT
GCAAGTACCATCCCACCTGTTCCAACTACGCTATCGAGGCGTTCCGGCAGCACAATTTTTTCTACGCCTC
CTGGCTGACCGTCTGGAGGGTGCTTCGTTGCAATCCGTTTTCAAAAGGCGGCTATGATCCGGTACCGCCA
AAATCAGTGAAATCCGCAGGTAATTCAAAAGATTCGAAGTAA
```

Figura 7. Seqüència de nucleòtids del gen exemple en format FASTA

Per tal de poder comparar els diferents gens, hem escollit paràmetres que tenen a veure amb l'ARN missatger derivat de la transcripció, ja que aquest presenta una estructura secundària que es pot predir amb uns paràmetres d'estabilitat mesurables. Així, per exemple, l'estabilitat dels missatgers es pot mesurar mitjançant les zones d'autocomplementareïtat (duplex), els braços (o *stems*) que aquestes formen, i l'energia de formació mesurada en Kcal/mol. Aquests paràmetres numèrics ens permetran fer comparacions estadístiques entre els diferents grups i tipus de gens.

Per poder analitzar les seqüències dels gens utilitzem eines bioinformàtiques online, que ens ajudaran a estudiar i recol·lectar informació sobre el gen desitjat. De aquests programes online n'hi han molts, però s'ha utilitzat un programa seleccionat i

5**FASTA** és un format de fitxer informàtic basat en text, que es fa sevir per representar seqüències d'àcids nucleics o de pèptids amb codis d'una sola lletra. Aquest format permet incloure el nom de la seqüència. http://es.wikipedia.org/wiki/Formato_FASTA

recomanat per científics que es pot executar amb un navegador des del lloc web de la universitat de Viena: RNAfold Webserver (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

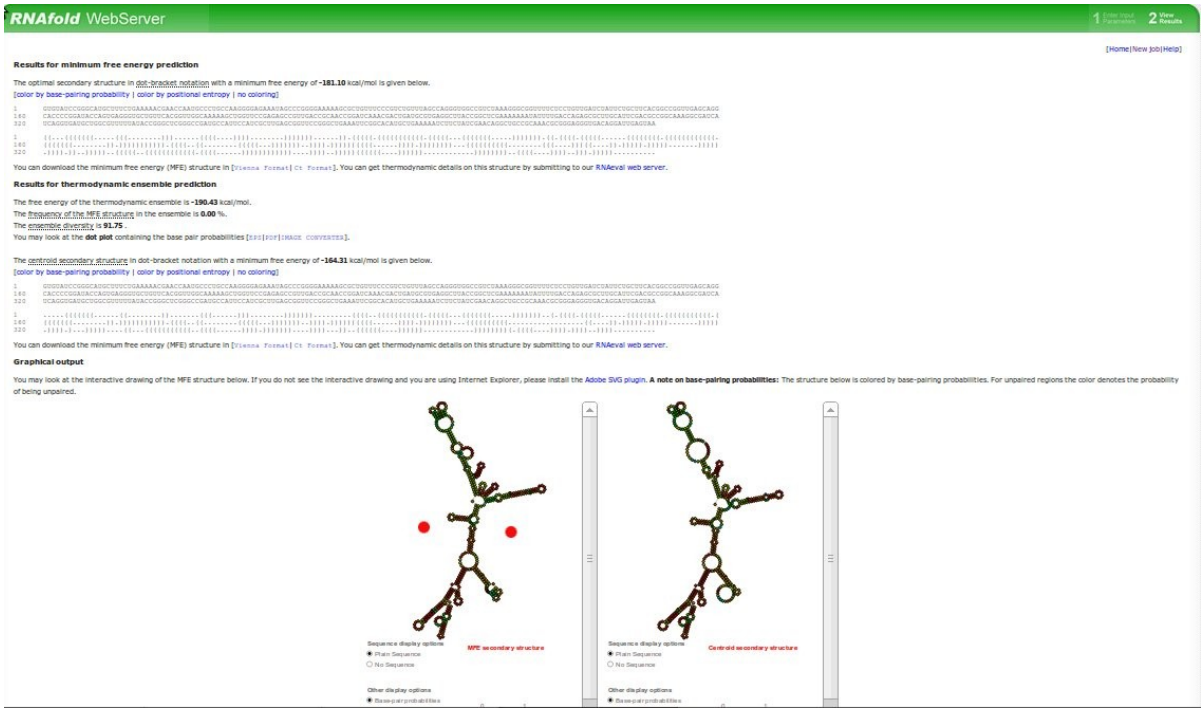


Figura 8a. Entrada de "RNAfolding", mostra del resultat de l'estudi d'un gen.

Un cop mostrat tot el contingut, ens centrem en la imatge marcada en vermell. Aquesta imatge representa l'estructura secundària del gen analitzat. A part d'aquesta imatge hi ha la imatge de la dreta, i no la utilitzem perquè és la imatge de MFE (Minimum Free Energy). Aquesta és l'estructura més estable ja que té la mínima energia de formació.

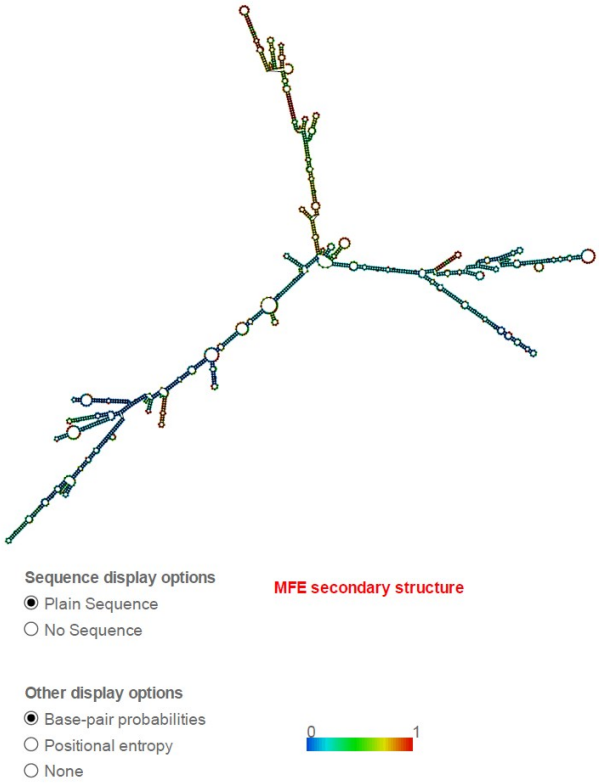


Figura 8b. Detall de l'estructura secundària on es poden veure les zones autocomplementàries de doble cadena (dúplex) i els braços que formen.

Un cop obtinguda l'estructura MFE, comencem a recollir informació sobre l'estructura secundària del gen utilitzat. Les informacions que recollim del gen, s'anomenen els tres paràmetres d'estabilitat, que són: *stems*, Dúplex, EEF (Energia Específica de Formació)

- *stems*: Els *stems* són les branques que tenen els gens en la forma de l'estructura secundària. Són indicadors del nombre de regions auto-complementàries d'un gen. Quants més *stems* més regions complementàries i per tant més punts d'estabilitat (en principi)
- *Dúplex*: El dúplex és la informació sobre la doble cadena de bases nitrogenades més llarga que forma el ARNm del gen en la seva estructura secundària.
- *EEF* (Energia Específica de Formació): és una funció d'estat extensiva amb unitats d'energia que donen la condició d'equilibri i espontaneïtat per una reacció química, a pressió i temperatura constant; on

normalment és negativa perquè sigui estable, també anomenada ΔG o ΔG de reacció (Δgr). Donat que aquest paràmetre és depenent de la longitud del gen, hem normalitzat el seu valor dividint pel nombre de bases del mateix, resultant en un valor d'energia per base nitrogenada.

Finalment s'han recollit tots aquests paràmetres en una taula per a la seva anàlisi estadística.

3. RESULTATS

3.1. Relació entre paràmetres

Abans de fer l'anàlisi estadística, i tenint en consideració tots els paràmetres recollits fins ara, hem representat gràficament la relació existent entre ells. En les següents figures es mostra que, efectivament existeix una forta relació entre tots els paràmetres escollits tant en gens funcionals com en gens no funcionals.

Aquests tres gràfics de la figura 9 es refereixen a la relació que hi ha entre els paràmetres dels gens funcionals. Tot seguit els comparem amb els gràfics que contrasten els mateixos paràmetres però sobre els gens que no tenen funció, els no classificats.

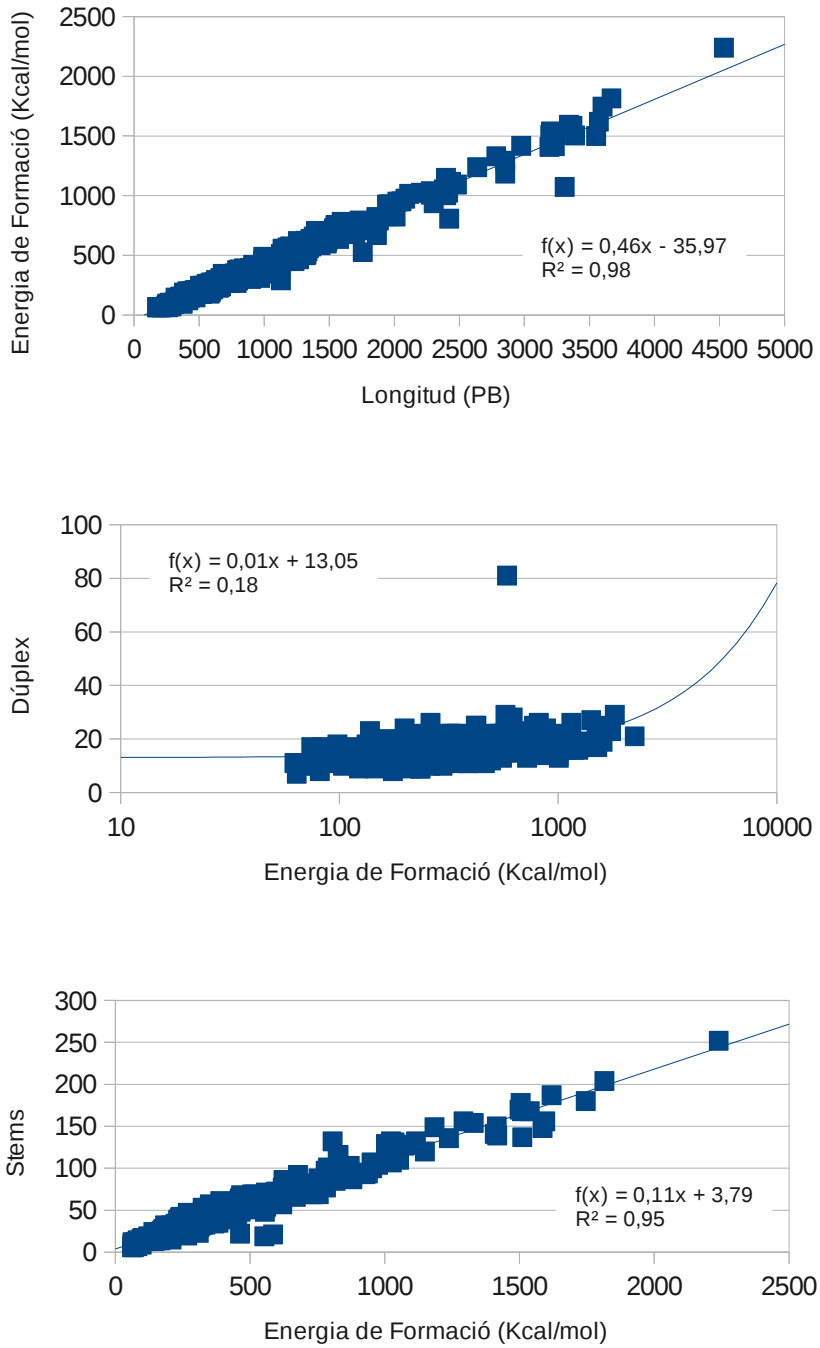


Figura 9. Relació entre els tres paràmetres d'estabilitat dels gens funcionals.

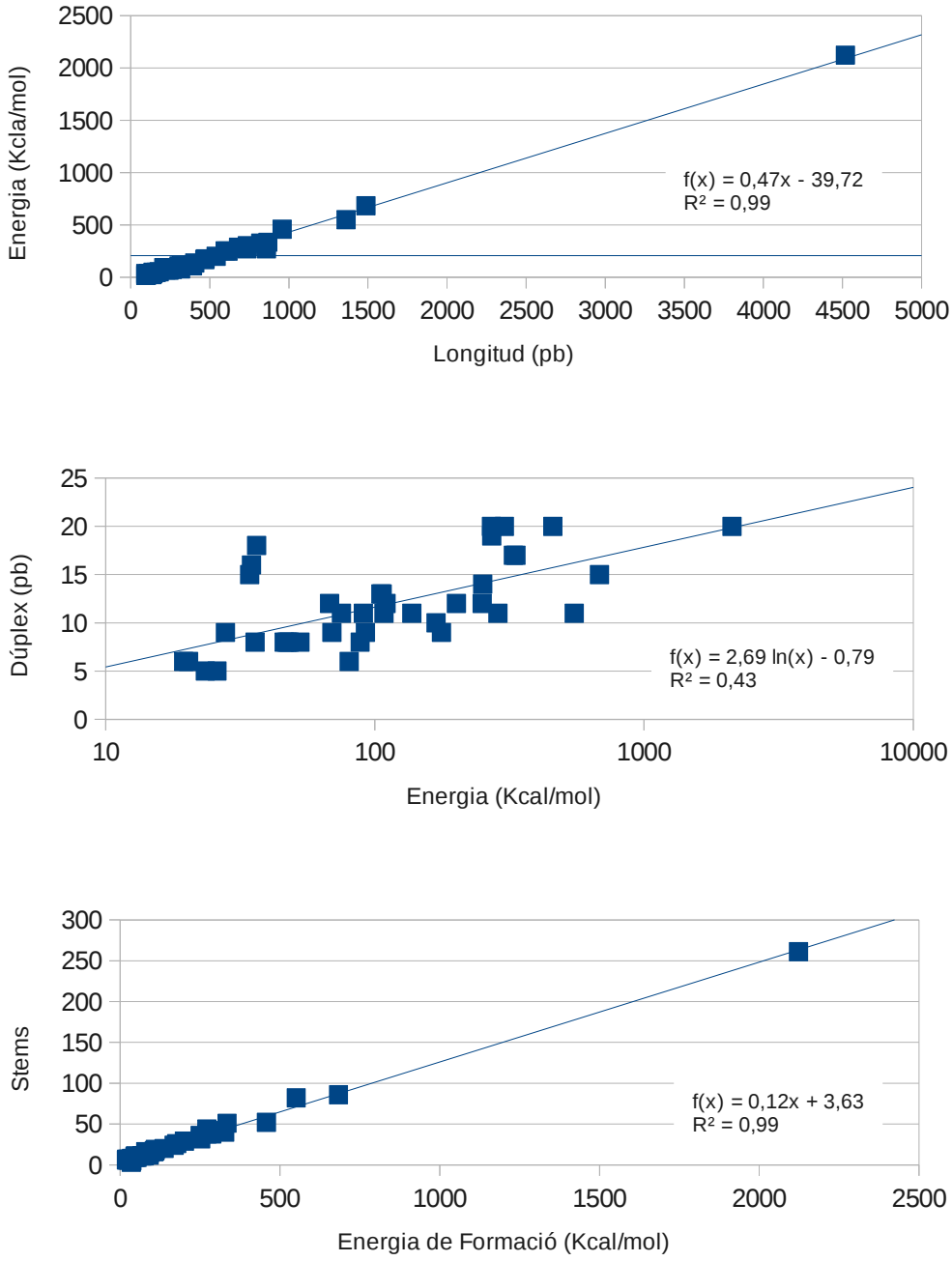


Figura 10: Relació entre els paràmetres d'estabilitat dels gens no funcionals

Un cop tenim els tres primers gràfics dels gens funcionals, s'intenta buscar alguna relació o diferència entre els gens funcionals i els que no ho són. Malgrat la nostre hipòtesi, no em trobat relació alguna entre la informació plasmada en els gràfics. Però no podem afirmar estadísticament que no hi ha diferències; per això introduïm les dades, segons la dependència d'un paràmetre envers l'altre, en la calculadora del

"T test". En l'anàlisi estadística s'utilitza el "T test". El "T test" o "test t Student" és una prova d'estadística que serveix per veure si les mitjanes de dos jocs de dades independents són o no iguals. Existeixen varies calculadores online del "T test", les quals no només em utilitzat aquestes calculadores online, sinó que em utilitzat un programa especialitzat en estadística SPSS[6], amb el qual s'ens és més fàcil fer un anàlisi estadístic sobre la informació recaptada dels gens.

http://studentsttest.com//

http://www.quantitativeskills.com/sisa/statistics/t-thlp.htm

3.2. Anàlisi dels paràmetres de estabilitat dels ARNm segons la classificació dels gens

Tot seguit de recollir la informació sobre l'anàlisi del gen, la hem aplicat a un altre anàlisi, una anàlisi estadística. Exactament n'hem fet 4, d'anàlisis estadístiques: Longitud del gen envers la funció, comparacions entre ID i els tres paràmetres d'estabilitat, anàlisi dels gens HYP segons els tres paràmetres d'estabilitat, anàlisi del gen segons la cadena

En aquesta comparació hem comparat els grups, les diferents categories entre elles per veure si podríem trobar alguna diferència

3.2.1. Comparacions entre els diferents grups de gens (COGs)

En aquest apartat comparem estadísticament tots els grups de gens entre ells segons els tres paràmetres d'estabilitat, per veure si hi ha alguna diferència, i com s'ha explicat anteriorment, ens basarem en el resultat del valor P.

La validesa de la hipòtesi nul·la (que les mitjanes són iguals) ens la indica el valor P obtingut en el test estadístic de comparació de mitjanes representat en les taules que es troben a continuació. El valor P és la probabilitat de que la nostra hipòtesi

6**SPSS:** Programa especialitzat en les anàlisis de dos mitjanes per comprovar si hi ha diferència significativa.

nul·la sigui correcte i per tant ens indica la significació estadística de la comparació. Per aquest treball hem escollit un nivell de significació del 95%. Això vol dir que si el valor de P és igual o inferior a 0,05 la nostra hipòtesi nul·la és incorrecte i podríem afirmar que les dos variables comparades són diferents.

Taula 1. Comparació entre els grups D i F amb els tres paràmetres d'estabilitat

Taula de mitjanes D i F

	ID COG	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	D	19	**16,26**	3.2.1. 3.2.1. 3,813	,875
	F	50	**15,96**	3,736	,528
stems	D	19	**51,11**	36,374	8,345
	F	50	**53,00**	30,722	4,345
ENERGIA ESPECIFICA	D	19	**,397276**	,0506023	,0116090
	F	50	**,427671**	,0355233	,0050238

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	67	,766	,303
	Equal variances not assumed	31,967	,769	,303
stems	Equal variances assumed	67	,829	-1,895
	Equal variances not assumed	28,317	,842	-1,895
ENERGIA ESPECIFICA	Equal variances assumed	67	,006	-,0303953
	Equal variances not assumed	25,050	,024	-,0303953

La taula 1 ens mostra, marcat de color verd, els valor de P, els quals són bastant superiors al mínim per assegurar que és significatiu. Per tant podem afirmar que entre aquests dos grups no hi ha cap diferència significativa pel que fa als dúplex i al nombre de braços. En canvi, en el paràmetre de l'energia de formació, marcat amb color vermell, el valor de P és més petit que 0,05. Per tant podem afirmar que hi ha una diferència significativa entre els grups D i F pel que la energia específica, és a dir, l'energia específica de formació del grup F és major que la del grup D.

Taula 2. Comparació entre els grups D i L amb els tres paràmetres d'estabilitat

Taula de mitjanes D i L

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	D	19	16,26	3,813	,875
	L	92	16,33	4,497	,469
stems	D	19	51,11	36,374	8,345
	L	92	74,18	50,122	5,226
ENERGIA ESPECIFICA	D	19	,397276	,0506023	,0116090
	L	92	,417220	,0520485	,0054264

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	109	,955	-,063
	Equal variances not assumed	29,348	,950	-,063
stems	Equal variances assumed	109	,060	-23,080
	Equal variances not assumed	33,855	,025	-23,080
ENERGIA ESPECIFICA	Equal variances assumed	109	,130	-,0199443
	Equal variances not assumed	26,475	,131	-,0199443

En la comparació entre els grups D i L (taula 2), la hipòtesi nul·la proposada és correcte ja que el valor de P, marcat amb verd, és superior al valor mínim proposat per assegurar les nostres conclusions. Per tant podem assegurar que els tres paràmetres d'estabilitat dels grups D i L no comporten diferències significatives.

Taula 3. Comparació entre els grups D i S amb els tres paràmetres d'estabilitat

Taula mitjanes D i S

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	D	19	**16,26**	3,813	,875
	S	113	**15,67**	7,356	,692
stems	D	19	**51,11**	36,374	8,345
	S	113	**36,72**	26,021	2,448
ENERGIA ESPECIFICA	D	19	**,397276**	,0506023	,0116090
	S	113	**,398776**	,0517035	,0048639

	t-test for Equality of Means

		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	130	,733	,591
	Equal variances not assumed	44,764	,599	,591
stems	Equal variances assumed	130	,038	14,388
	Equal variances not assumed	21,206	,113	14,388
ENERGIA ESPECIFICA	Equal variances assumed	130	,907	-,0014997
	Equal variances not assumed	24,751	,906	-,0014997

En referència a la taula de la comparació entre els grups D i S, podem dir el nombre d'*stems* del grup D és significativament superior al valor del grup S.

Taula 4. Comparació entre els grups D i V amb els tres paràmetres d'estabilitat

Taula mitjanes D i V

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	D	19	**16,26**	3,813	,875
	V	33	**17,06**	4,387	,764
stems	D	19	**51,11**	36,374	8,345
	V	33	**59,21**	36,158	6,294
ENERGIA ESPECIFICA	D	19	**,397276**	,0506023	,0116090
	V	33	**,439759**	,0300433	,0052299

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	50	,512	-,797
	Equal variances not assumed	42,130	,496	-,797
stems	Equal variances assumed	50	,441	-8,107
	Equal variances not assumed	37,483	,443	-8,107
ENERGIA ESPECIFICA	Equal variances assumed	50	,000	-,0424827
	Equal variances not assumed	25,458	,003	-,0424827

En la comparació de la taula 4 es pot veure com la diferència entre els paràmetres dúplex i *stems* no existeix, ja que el valor de P, marcat en verd, és més gran que 0,05. En canvi entre l'energia específica, marcat en vermell, dels dos grups hi ha una diferència significativa, essent el valor de V major que el de D

Taula 5. Comparació entre els grups D i NON amb els tres paràmetres d'estabilitat

Taula mitjanes D i NON

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	D	19	**16,26**	3,813	,875
	NON	40	**11,83**	4,466	,706
stems	D	19	**51,11**	36,374	8,345
	NON	40	**28,83**	42,491	6,718
ENERGIA ESPECIFICA	D	19	**,397276**	,0506023	,0116090
	NON	40	**,337538**	,0746983	,0118108

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	57	,000	4,438
	Equal variances not assumed	41,055	,000	4,438
stems	Equal variances assumed	57	,054	22,280
	Equal variances not assumed	40,955	,044	22,280
ENERGIA ESPECIFICA	Equal variances assumed	57	,003	,0597376
	Equal variances not assumed	49,882	,001	,0597376

En diferència de les altres anàlisis, aquest es contrasten les informacions d'un grup funcional i d'un grup no classificat com és el grup NON. En la taula 5 podem veure un diferència significativa en els tres paràmetres d'estabilitat[7], perquè els dos paràmetres dúplex i l'energia específica, marcats en vermell, el valor està per sota del 0,05. El valor de P per als *stems* està lleugerament sobre d'aquest 0,05, però supera aquest valor per molt poc. Segurament amb un valor de “N” (tamany de mostra) més gran assoliríem el P<0,05

Taula 6. Comparació entre els grups F i L amb els tres paràmetres d'estabilitat

Taula mitjanes F i L

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	F	50	**15,96**	3,736	,528
	L	92	**16,33**	4,497	,469
stems	F	50	**53,00**	30,722	4,345

7El valor de P=0,054 es pot considerar per arrodoniment P=0,05 i per tant es pot considerar significatiu

	L	92	**74,18**	50,122	5,226
ENERGIA ESPECIFICA	F	50	**,427671**	,0355233	,0050238
	L	92	**,417220**	,0520485	,0054264

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	140	,624	-,366
	Equal variances not assumed	117,364	,605	-,366
stems	Equal variances assumed	140	,007	-21,185
	Equal variances not assumed	137,909	,002	-21,185
ENERGIA ESPECIFICA	Equal variances assumed	140	,207	,0104510
	Equal variances not assumed	132,743	,160	,0104510

En aquesta comparació no hi ha cap diferència entre els dos grups F i L en els paràmetres marcats en color verd, però hi ha una diferència significativa en el paràmetre marcat en vermell , els *stems*, perquè el seu valor P és inferior a 0,05.

Taula 7. Comparació entre els grups F i S amb els tres paràmetres d'estabilitat

Taula mitjanes F i S

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	F	50	**15,96**	3,736	,528
	S	113	**15,67**	7,356	,692
stems	F	50	**53,00**	30,722	4,345
	S	113	**36,72**	26,021	2,448
ENERGIA ESPECIFICA	F	50	**,427671**	,0355233	,0050238
	S	113	**,398776**	,0517035	,0048639

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	161	,794	,287
	Equal variances not assumed	157,953	,742	,287
stems	Equal variances assumed	161	,001	16,283
	Equal variances not assumed	81,455	,002	16,283
ENERGIA ESPECIFICA	Equal variances assumed	161	,000	,0288956

	Equal variances not assumed	132,847	,000	,0288956

En aquesta taula 7 de comparacions entre els dos grups F i S, els *stems* i l'energia específica dels dos grups són completament diferents, amb uns valors P 0,001 i 0 respectivament. Mentre que els dúplex, marcat en verd, dels dos grups no tenen cap diferència.

Taula 8. Comparació entre els grups F i V amb els tres paràmetres d'estabilitat

Taula mitjanes F i V

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	F	50	**15,96**	3,736	,528
	V	33	**17,06**	4,387	,764
stems	F	50	**53,00**	30,722	4,345
	V	33	**59,21**	36,158	6,294
ENERGIA ESPECIFICA	F	50	**,427671**	,0355233	,0050238
	V	33	**,439759**	,0300433	,0052299

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	81	,224	-1,101
	Equal variances not assumed	60,859	,241	-1,101
stems	Equal variances assumed	81	,403	-6,212
	Equal variances not assumed	60,751	,420	-6,212
ENERGIA ESPECIFICA	Equal variances assumed	81	,111	-,0120874
	Equal variances not assumed	76,027	,100	-,0120874

En la taula 8 podem veure igualtat entre els tres paràmetres d'estabilitat dels grups F i V amb uns valors de P > 0,05.

Taula 9. Comparació entre els grups F i NON amb els tres paràmetres d'estabilitat

Taula mitjanes F i NON

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	F	50	**15,96**	3,736	,528
	NON	40	**11,83**	4,466	,706
stems	F	50	**53,00**	30,722	4,345

	NON	40	**28,83**	42,491	6,718
ENERGIA ESPECIFICA	F	50	**,427671**	,0355233	,0050238
	NON	40	**,337538**	,0746983	,0118108

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	88	,000	4,135
	Equal variances not assumed	75,948	,000	4,135
stems	Equal variances assumed	88	,002	24,175
	Equal variances not assumed	68,856	,004	24,175
ENERGIA ESPECIFICA	Equal variances assumed	88	,000	,0901329
	Equal variances not assumed	53,008	,000	,0901329

La comparació entre els grups F (funcional) i NON (no funcional) ens mostra clares diferències entre els dos grups, ja que les comparacions amb els tres paràmetres d'estabilitat han donat un valor ínfim, un valor que és més petit que 0,05. Amb aquestes proves podem afirmar que els valors dels paràmetres d'estabilitat del grup de gens sense funció, els no classificats, són menors (duplex, *stems* i EEF) als del grup F.

Taula 10. Comparació entre els grups L i S amb els tres paràmetres d'estabilitat

Taula mitjanes L i S

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	L	92	**16,33**	4,497	,469
	S	113	**15,67**	7,356	,692
stems	L	92	**74,18**	50,122	5,226
	S	113	**36,72**	26,021	2,448
ENERGIA ESPECIFICA	L	92	**,417220**	,0520485	,0054264
	S	113	**,398776**	,0517035	,0048639

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	203	,457	,654
	Equal variances not assumed	189,320	,435	,654
stems	Equal variances assumed	203	,000	37,468
	Equal variances not assumed	130,225	,000	37,468
ENERGIA ESPECIFICA	Equal variances assumed	203	,012	,0184446

	Equal variances not assumed	194,142	,012	,0184446

Com mostra la taula, els grups L i S tenen més diferències significatives que característiques comunes, ja que 2 dels 3 paràmetres d'estabilitat, marcats en vermell, dels dos grups tenen un valor P inferior a la marca de 0,05, indicant la diferència significativa dels dos grups envers els *stems* i l'energia específica. El dúplex dels dos grups no presenta cap diferència.

Taula 11. Comparació entre els grups L i V amb els tres paràmetres d'estabilitat

Taula mitjanes L i V

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	L	92	**16,33**	4,497	,469
	V	33	**17,06**	4,387	,764
stems	L	92	**74,18**	50,122	5,226
	V	33	**59,21**	36,158	6,294
ENERGIA ESPECIFICA	L	92	**,417220**	,0520485	,0054264
	V	33	**,439759**	,0300433	,0052299

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	123	,419	-,735
	Equal variances not assumed	57,780	,416	-,735
stems	Equal variances assumed	123	,118	14,973
	Equal variances not assumed	78,242	,071	14,973
ENERGIA ESPECIFICA	Equal variances assumed	123	,021	-,0225384
	Equal variances not assumed	98,034	,004	-,0225384

En aquesta comparació podem veure que hi ha una diferència significativa en la comparació dels grups L i V envers l'energia de formació, marcat en vermell. Mentre que la comparació envers els altres dos paràmetres no hi ha cap diferència significativa.

Taula 12. Comparació entre els grups L i NON amb els tres paràmetres d'estabilitat

Taula mitjanes L i NON

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	L	92	**16,33**	4,497	,469
	NON	40	**11,83**	4,466	,706
stems	L	92	**74,18**	50,122	5,226
	NON	40	**28,83**	42,491	6,718
ENERGIA ESPECIFICA	L	92	**,417220**	,0520485	,0054264
	NON	40	**,337538**	,0746983	,0118108

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	130	,000	4,501
	Equal variances not assumed	74,742	,000	4,501
stems	Equal variances assumed	130	,000	45,360
	Equal variances not assumed	86,840	,000	45,360
ENERGIA ESPECIFICA	Equal variances assumed	130	,000	,0796819
	Equal variances not assumed	56,131	,000	,0796819

Aquesta taula compara el grup L, un grup funcional, i el grup NON, el grup no funcional. La diferència es total, en tots els aspectes. En els tres paràmetres d'estabilitat hi ha una diferència absoluta, una diferència molt significativa ja que el valor de P és 0.

Taula 13. Comparació entre els grups S i V amb els tres paràmetres d'estabilitat

Taula mitjanes S i V

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	S	113	**15,67**	7,356	,692
	V	33	**17,06**	4,387	,764
stems	S	113	**36,72**	26,021	2,448
	V	33	**59,21**	36,158	6,294
ENERGIA ESPECIFICA	S	113	**,398776**	,0517035	,0048639
	V	33	**,439759**	,0300433	,0052299

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	144	,305	-1,388
	Equal variances not assumed	88,985	,181	-1,388
stems	Equal variances assumed	144	,000	-22,495
	Equal variances not assumed	42,136	,002	-22,495
ENERGIA ESPECIFICA	Equal variances assumed	144	,000	-,0409830
	Equal variances not assumed	91,695	,000	-,0409830

La taula 13 ens diu que la comparació entre els dos grups envers els *stems* i l'energia de formació, hi han diferències significatives, ja que els valors de P dels dos paràmetres són 0 respectivament. En canvi, els dúplex dels dos grups no presenten cap diferència.

Taula 14. Comparació entre els grups S i NON amb els tres paràmetres d'estabilitat

Taula mitjanes S i NON

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	S	113	**15,67**	7,356	,692
	NON	40	**11,83**	4,466	,706
stems	S	113	**36,72**	26,021	2,448
	NON	40	**28,83**	42,491	6,718
ENERGIA ESPECIFICA	S	113	**,398776**	,0517035	,0048639
	NON	40	**,337538**	,0746983	,0118108

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	151	,002	3,848
	Equal variances not assumed	113,454	,000	3,848
stems	Equal variances assumed	151	,170	7,892
	Equal variances not assumed	49,737	,275	7,892
ENERGIA ESPECIFICA	Equal variances assumed	151	,000	,0612373
	Equal variances not assumed	52,821	,000	,0612373

En aquesta comparació del grup NON, no funcional, i el grup S, funcional, es pot

veure marcat de color vermell els dos paràmetres en els quals no hi ha cap relació entre els dos grups. En canvi, els *stems* dels dos grups, al tenir un valor P més gran que 0,05, tenen alguna relació

Taula 15. Comparació entre els grups V i NON amb els tres paràmetres d'estabilitat

Taula mitjanes V i NON

	ID	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	V	33	17,06	4,387	,764
	NON	40	11,83	4,466	,706
stems	V	33	59,21	36,158	6,294
	NON	40	28,83	42,491	6,718
ENERGIA ESPECIFICA	V	33	,439759	,0300433	,0052299
	NON	40	,337538	,0746983	,0118108

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	71	,000	5,236
	Equal variances not assumed	68,827	,000	5,236
stems	Equal variances assumed	71	,002	30,387
	Equal variances not assumed	70,920	,002	30,387
ENERGIA ESPECIFICA	Equal variances assumed	71	,000	,1022203
	Equal variances not assumed	53,296	,000	,1022203

Les diferències entre els grups V, funcional, i NON, no funcional, envers els tres paràmetres d'estabilitat són absolutes. El valor P de totes elles és 0, per tant, no hi ha cap relació.

Dúplex més llarg					
	D	F	L	S	V
D					
F	0,766				
L	0,955	0,624			
S	0,733	0,794	0,457		
V	0,512	0,224	0,419	0,305	
NON	0,000	0,000	0,000	0,002	0,000

Stems					
	D	F	L	S	V
D					
F	0,829				
L	0,025	0,002			
S	0,038	0,001	0,000		
V	0,441	0,403	0,071	0,002	
NON	0,054	0,002	0,000	0,170	0,002

Energia específica de formació					
	D	F	L	S	V
D					
F	0,006				
L	0,130	0,160			
S	0,907	0,000	0,012		
V	0,000	0,111	0,004	0,000	
NON	0,001	0,000	0,000	0,000	0,000

Figura 11. Resum de tots els valors de P en totes les comparacions entre grups.

Els valors en vermell de la figura 11 es refereixen als valors P que són inferiors al valor 0,05. Si ens hi fixem, Tots els valors de P referents al grup NON, està per sota del 0,05. Per tant, podríem dir que els gens del grup NON estan un pas en darrera dels demés grups, com si siguessin més petits, tinguessin menys *stems* i la seva energia específica sigués més petita.

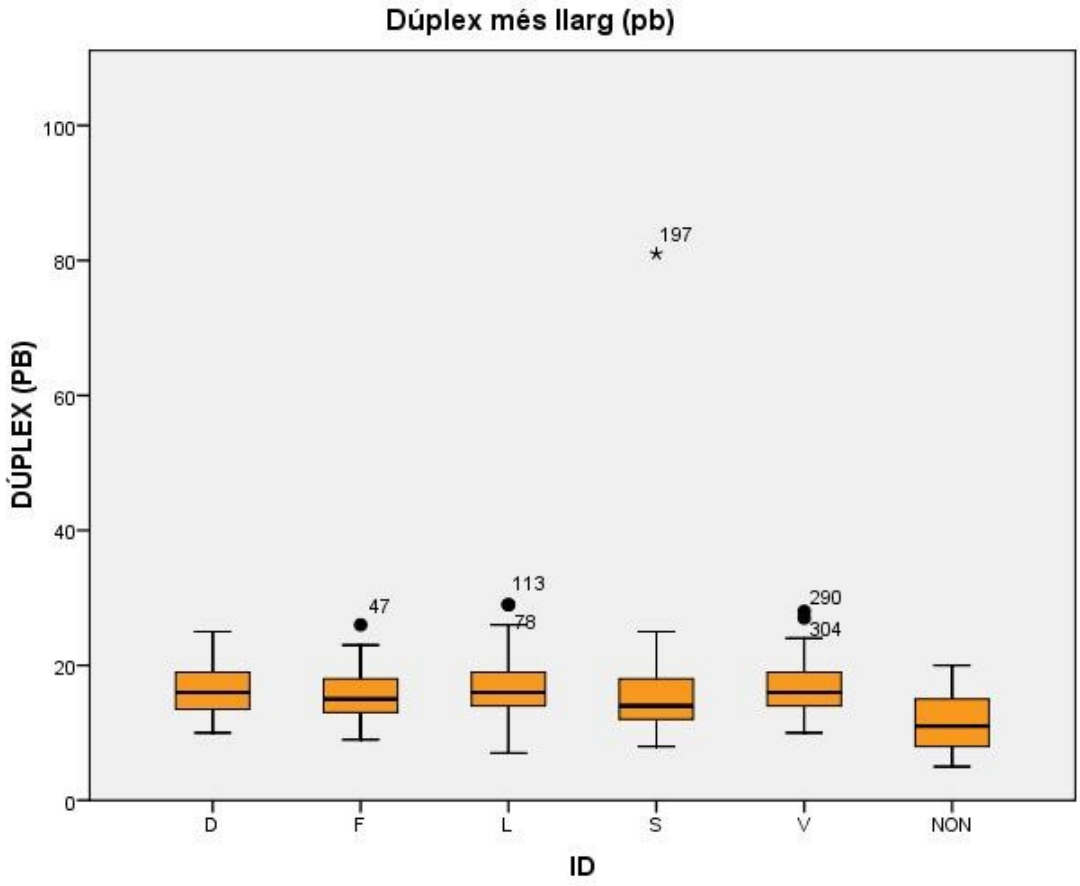


Figura 12. Gràfic resum on ens mostra la mitjana de cada grup grups segons els dúplex

Després de les comparacions, es fa un gràfic de totes les mitjanes dels grups en dúplex. Podem veure en aquest gràfic que totes les mitjanes estan relacionades menys una, la mitjana del grup NON. Al ser més petita la mitjana ens indica que el gen és més petit que els altres. En aquest gràfic, el grup NON és l'únic dels grups que té el valor de P més baix que 0,05, per tant la hipòtesi nul·la que no són diferents és incorrecta. Això ens diu que la longitud del dúplex si depèn de si són classificades o no, depèn de la funció. Els punts fora del gràfic, són punts no considerats per l'anàlisi estadística, no estan dintre de la distribució normal, per tant el programa d'anàlisi estadístic, el SPSS, no els ha inclòs a l'anàlisi. Els que estan marcats amb una estrella, massa grans.

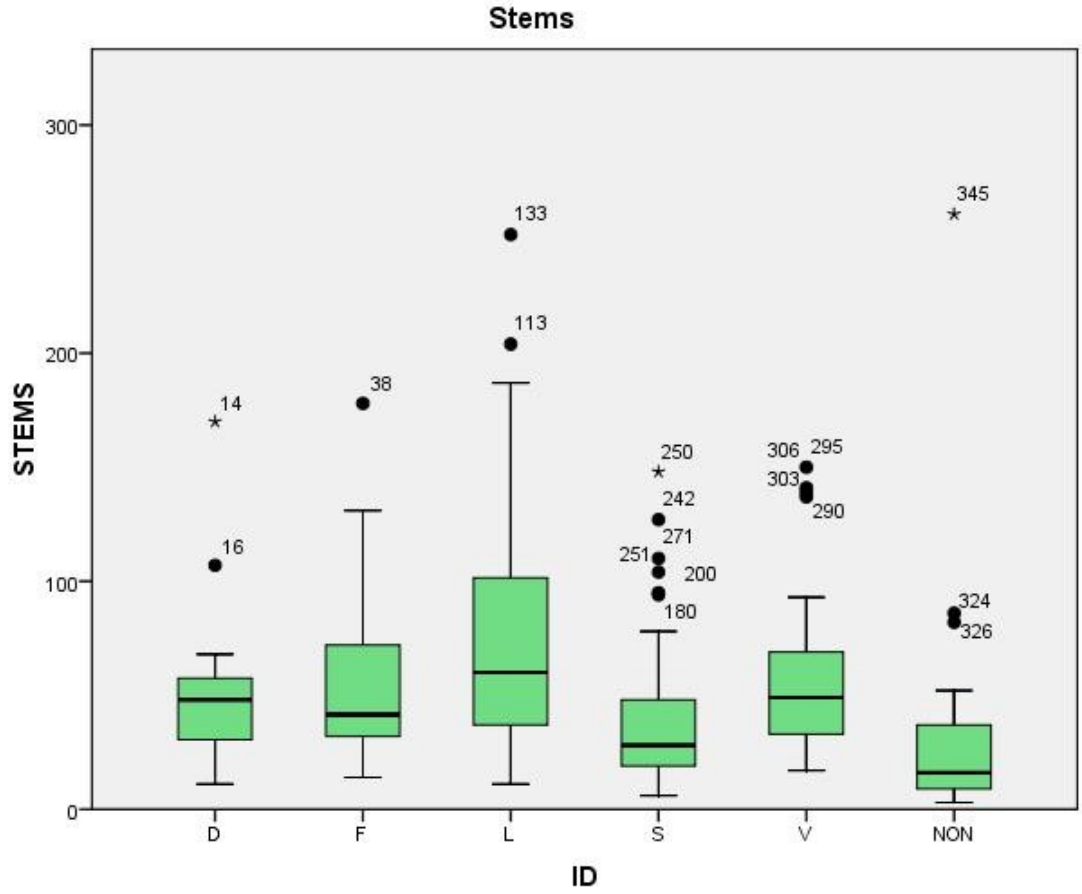


Figura 13. Gràfic que mostra mitjana de "*stems*" de cada un dels grups de gens

La figura 5x ens mostra les mitjanes envers els *stems* de cada grup, i podem veure que el grup NON es queda una altra vegada en darrera. Això podria significar moltes coses, però s'ha interpretat d'una manera. La mitjana dels *stems* en el grup NON és més baixa per dos raons: com hem comprovat abans, els gens del NON són més petites, per tant els *stems* són directament proporcionals, i els *stems* són més reduïts; la segona seria que, com que els *stems* és un paràmetre molt inestable, que és bastant irregular, sigués més baix per coincidència. La proba de que és inestable i irregular són la quantitat de punts i estrelles fora de la mitjana, fora del gràfic.

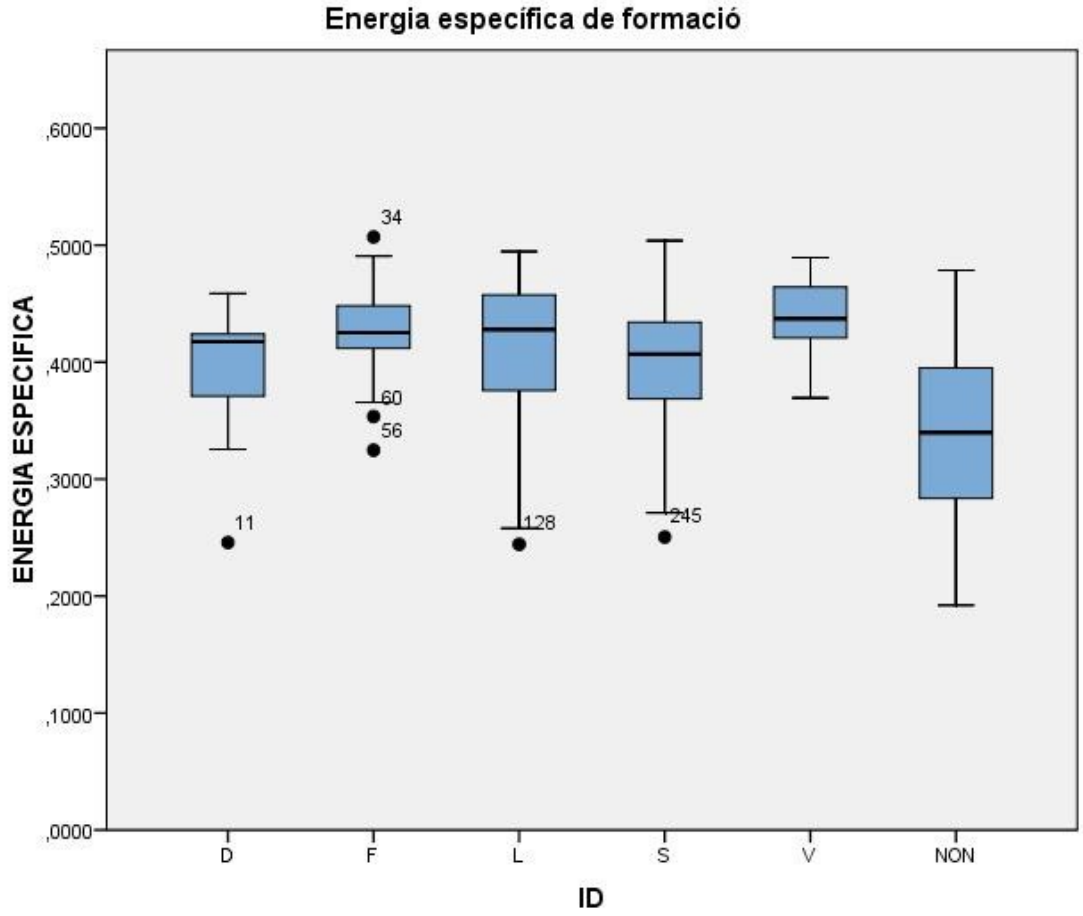


Figura 14. Gràfic representatiu de la energia específica respecte es grups de gens.

El gràfic ens mostra que clarament, una altra vegada els NON tornen a quedar-se endarrere respecta a la mitjana. Com que l'energia específica està formada per la divisió de la energia de formació d'un gen (ΔG) i la longitud d'un gen; que la mitjana estigui per sota de les demés vol dir que, tan com l'energia de formació com la longitud del gen, són més petits si els comparem amb la resta de gens funcionals.

Analitzant les tres gràfiques, podem esmentar que el grup NON té una mitjana més baixa i un valor de P per sota de lo normal dels tres paràmetres d'estabilitat que ens pot portar a una hipòtesi: al tenir una mitjana més baixa l'estructura secundària que pugui formar l'ARN dels gens no classificats serà massa inestable perquè es pugui traduir a proteïna.

Després he agrupat tots els gens classificats i els hem comparat amb els no

classificats. Aquests darrers com que donen lloc a proteïnes hipotètiques els hem anomenat HYP.

3.2.2. Comparacions segons la funcionalitat dels gens (HYP vs. no HYP)

Volíem veure si el conjunt si els gens classificats com a funcionals eren o no diferents per als tres paràmetres als no classificats. Aquest s'han anomenat HYP per generar proteïnes hipotètiques ja que com hem dit la funció és desconeguda.

Es dóna el valor de 1 a les HYP i 0 a les que no ho són

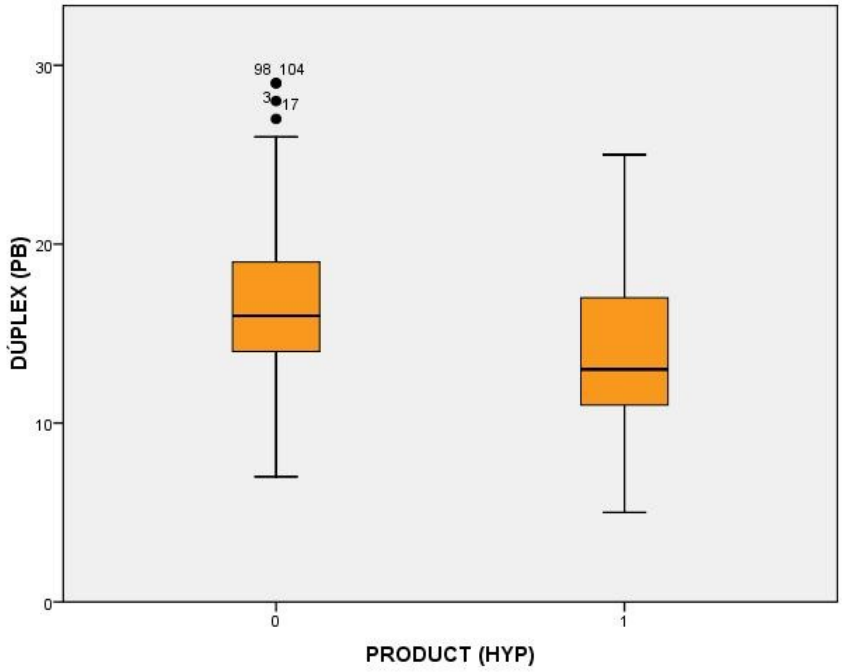
Taula 16. Comparació dels gens HYP amb els no HYP envers els tres paràmetres d'estabilitat

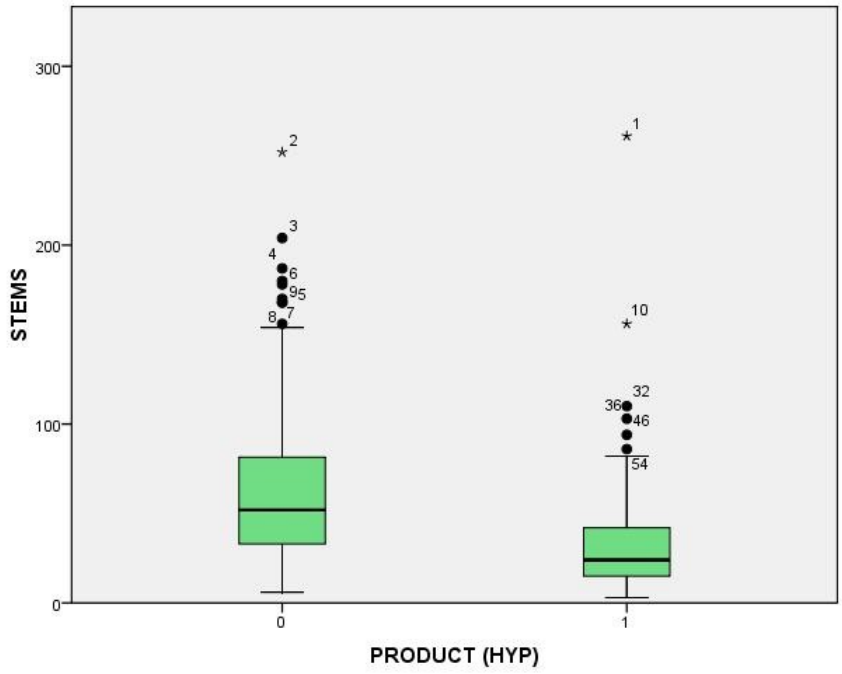
Mitjanes de gens amb funció hipotètica (1) i real (0)

	PRODUCT (HYP)	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	0	203	16,72	6,183	,434
	1	144	14,04	4,279	,357
stems	0	203	63,76	42,825	3,006
	1	144	33,04	30,311	2,526
ENERGIA ESPECIFICA	0	203	,422587	,0454180	,0031877
	1	144	,379210	,0643396	,0053616

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	345	,000	2,678
	Equal variances not assumed	344,812	,000	2,678
stems	Equal variances assumed	345	,000	30,722
	Equal variances not assumed	344,999	,000	30,722
ENERGIA ESPECIFICA	Equal variances assumed	345	,000	,0433768
	Equal variances not assumed	240,675	,000	,0433768

Com es mostra a la taula (color vermell), la diferència entre els gens hipotètics amb els gens no hipotètics és una diferència molt significativa, com indica el valor de P=0,000.


DÚPLEX (PB)
30
20
10
0
98
104
3
17
0
1
PRODUCT (HYP)


STEMS
300
200
100
0
2
3
4
6
95
8
7
1
10
32
36
46
54
0
1
PRODUCT (HYP)

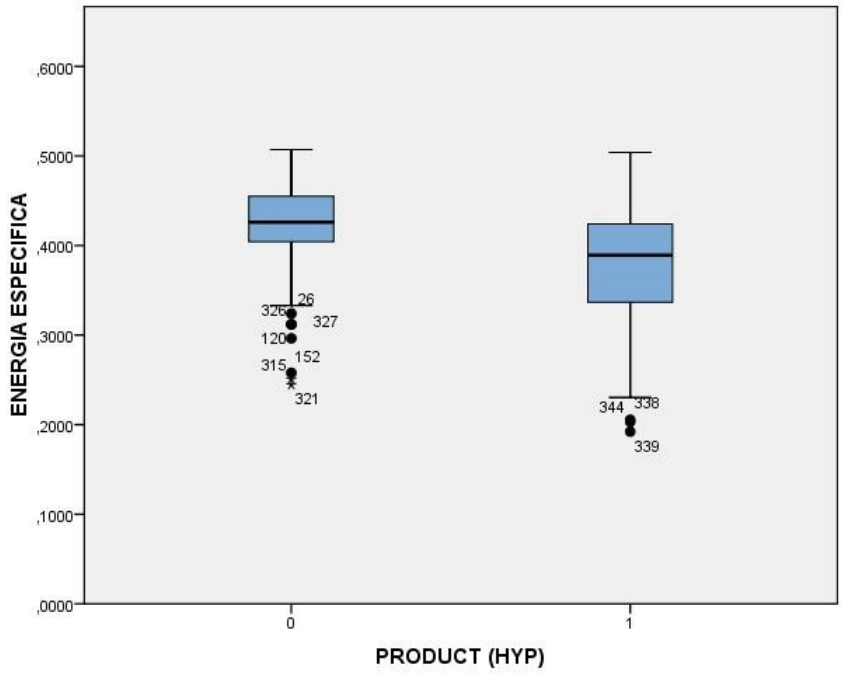


Figura 15. Visualització gràfica de la comapació entre gens codificants per proteines hipotètiques (1)o reals (0)

El primer gràfic ens permet veure que els gens HYP acostumen a tenir la longitud del dúplex més gran que els gens no HYP. Per tant, podem afirmar que la longitud del dúplex depèn de la seva funció.

Pel que fa als *stems*, podem veure que els gens que tenen funció tenen la mitjana més alta, per tant, els gens funcionals acostumen a tenir més *stems* que els no classificats. El valor de P és inferior a 0,05, és a dir, que la hipòtesi nul·la és incorrecte; per tant podem dir que els *stems* dels gens funcionals acostumen a ser més nombrosos que els que tenen funció hipotètica. Els punts i les estrelles fora del gràfic no són considerats dintre de la mitjana.

Finalment, pel que fa a l'energia específica de formació es pot apreciar que és més alta en els gens funcionals. Però no ho podem dir només mirant el gràfic, sinó que ens hem de basar en els resultats de l'anàlisi estadística per confirmar que la nostre hipòtesi nul·la, que l'energia específica no depèn de la funció del gen, és incorrecte. Així ens ho mostra l'anàlisi amb el valor de P, ja que és inferior a 0,05. Els punts i les estrelles fora del gràfic no són considerats dintre de la mitjana.

Per tant, després de veure els resultats de l'anàlisi dels gens que són o no són hipotètics, podem dir que els tres paràmetres d'estabilitat, sí depenen de la funció del gen, que no tingui una funció hipotètica.

3.2.3. Comparacions segons la cadena codificant (+ o -)

Amb el "T test" de cadena volem veure si els tres paràmetres d'estabilitat (*stems*, Longitud del dúplex i Energia específica; dos estructurals i una energètica) són depenent del "Strand".

En el genoma dels bacteris, els gens es poden transcriure indistintament a partir de una o altra cadena, segons el cas. La cadena codificant més freqüent és la (-) ja que dóna lloc al ARNm. La cadena (+) és la que serà traduït a proteïna.

Taula 17. Mitjanes dels gens segons la seva posició en la cadena o "strand"

	STRAND	N	Mean	Std. Deviation	Std. Error Mean
DÚPLEX (PB)	+	198	15,46	6,344	,451
	-	149	15,81	4,506	,369
stems	+	198	52,88	40,901	2,907
	-	149	48,54	41,117	3,368
ENERGIA ESPECIFICA	+	198	,401489	,0580466	,0041252
	-	149	,408702	,0580630	,0047567

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
DÚPLEX (PB)	Equal variances assumed	345	,572	-,346
	Equal variances not assumed	343,928	,553	-,346
stems	Equal variances assumed	345	,329	4,342
	Equal variances not assumed	318,005	,330	4,342
ENERGIA ESPECIFICA	Equal variances assumed	345	,253	-,0072131
	Equal variances not assumed	318,843	,253	-,007

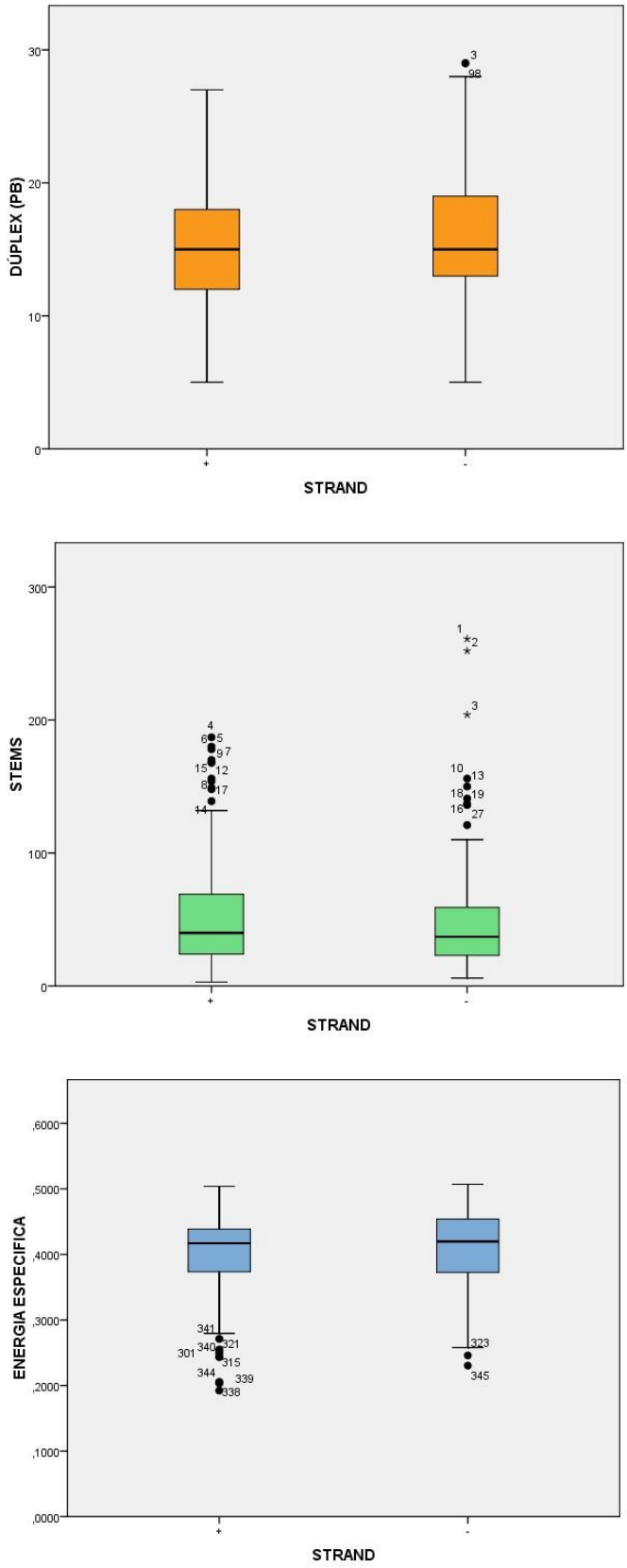


Figura 16. Comparació gràfica de les mitjanes dels tres paràmetres (Duplex, *stems* i EEF) en funció de si la cadena codificant és + o -.

Amb els resultats estadístics obtinguts a la taula 17, podem dir directament que no hi ha cap diferència entre la posició a la cadena o “strand” dels gens envers els tres paràmetres d'estabilitat, marcats en color verd.

Volem comprovar si la longitud del dúplex depèn de la seva ubicació a la cadena, si està a la cadena (+) o a la cadena (-). El gràfic ens mostra que les mitjanes estan una mica desequilibrades, que hi ha més gens en la cadena (-) que a la (+). Podríem treure conclusions d'aquí, però abans hauríem de fer l'anàlisi estadística, amb la qual ens assegurem que el valor de P és major al estimat (0,05) amb un valor superior a aquest anterior. Ens indica que la nostre hipòtesi nul·la que no són diferents és correcte, per tant la longitud del dúplex no depèn de la seva ubicació en la cadena.

En la gràfica dels *stems*, la mitjana dels gens a la cadena (+) amb els gens ubicats a la cadena (-),és casi la mateixa, per tant no podem dir gaire abans de fer l'anàlisi estadística. Per això, al fer l'anàlisi estadística en fixem molt amb el valor de P, el qual és superior a 0,05 (0,329), per tant podem dir que els *stems* no depenen de la ubicació en la cadena

La gràfica de l'EEF (energia específica de formació) ens mostra que els gens en la cadena (-) tenen una mitjana d'energia específica lleugerament més elevada que en la cadena (+). Però aquesta dada no ens indica res, ja que el valor de P després de l'anàlisi estadística, és superior a 0,05 (0,253), i ens mostra que la hipòtesi nul·la, que no hi ha diferència respecte a la ubicació de la cadena, és totalment correcte. Per tant l'energia específica en un gen no depèn de la seva ubicació en la cadena.

Comparant els resultats d'aquests últims gràfics en relació a la cadena i els resultats estadístics, podem afirmar amb seguretat que els tres paràmetres d'estabilitat no depenen de la seva ubicació a la cadena.

Al acabar la comparació ens hem adonat que potser que la longitud del gen depengui de la seva funció, si és HYP o no. Per tant, hem tornat a ajuntar els gens HYP i els gens no classificats respecte la longitud dels gens.

3.2.4. Comparació de la longitud dels grups de gens

Finalment hem analitzat si la longitud del gen varia segons la seva funció. Hem comprovat si la longitud dels gens funcionals, és a dir, els gens classificats, és diferent de la longitud dels gens HYP. La nostra hipòtesi nul·la és que els dos tipus de gens tenen la mateixa longitud.

Taula 18. Comparació de la longitud entre funcionals i no classificats (HYP)

	PRODUCT (HYP)	N	Mean	Std. Deviation	Std. Error Mean
LONGITUD (PB)	0	203	**1262,45**	841,568	59,067
	1	144	**696,27**	586,120	48,843

		t-test for Equality of Means		
		Sig. (2-tailed)	Mean Difference	Std. Error Difference
LONGITUD (PB)	Equal variances assumed	,000	566,182	81,319
	Equal variances not assumed	,000	566,182	76,645

En el cas de l'anàlisi estadística de la longitud del gen, el valor de P mostrat a la taula 18 i marcat en color vermell, és inferior a 0,05. Per tant, podem afirmar que els gens HYP són més curts que els gens funcionals. En el cas de *Chlorobaculum tepidum* podem afirmar que la longitud dels gens HYP és gairebé la meitat envers dels gens classificats. Ja que el valor de P és 0,00, en aquest cas l'afirmació és absolutament certa.

4. CONCLUSIÓ

Els ARNm derivats dels gens HYP del bacteri *Chlorobaculum tepidum* són inestables i curts amb la qual cosa es podria pensar que la seva funció no és específicament la de codificar proteïnes, almenys tal com les coneixem, i per aquest motiu apareixen com a no classificats en les llistes de categories del genoma. Aquest ADN diferent a la resta en bacteris podria ser un indici de l'existència de material genètic equivalent o similar a la matèria fosca dels eucariotes, els introns.

5. BIBLIOGRAFIA I FONTS D'INFORMACIÓ

- http://www.buhardillapodcast.es/el-lado-oscuro-del-genoma/
- http://es.wikipedia.org/wiki/%C3%81cido_ribonucleico#Tipos_de_ARN
- Berg, J.M., Tymoczko, J.L. & Stryer, B.T. Bioquímica (6a. Edició) 2007. Ed. REVERTÉ ISBN 978-84-291-7601-8.
- http://beckerinfo.net/bioinformatics/?p=191
- http://es.wikipedia.org/wiki/Gen_de_ARN
- http://www.ncrna.org/
- http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=247
- http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
- http://www.genebee.msu.su/services/rna2_reduced.html
- http://en.wikipedia.org/wiki/Homology_(biology)#Orthology
- http://ca.wikipedia.org/wiki/Chlorobium
- http://es.wikipedia.org/wiki/Ex%C3%B3n
- Cooper, G.M. & Hausman, R.E. La célula (5a. Edició) 2009. Ed MARBÁN ISBN 978-84-7101-672-0
- http://es.wikipedia.org/wiki/SPSS
- http://es.wikipedia.org/wiki/Espliceosoma
- http://es.wikipedia.org/wiki/Cuerpo_de_Cajal

6. AGRAÏMENTS

En primer lloc agrair al Sr. Manel Montoliu per tota la paciència que ha tingut i per no perdre mai l'esperança que tenia dipositada en mi, i fer que tot això sigui possible.

Donar-li gràcies també a la meva mare, Susi Martínez Planells, de mantenir el meu estat d'ànim perquè no decaigués i pogués continuar sense complicacions.

També agrair-l'hi la seva col·laboració a la meva germana i estudiant de medicina, Cristina García Martínez, al proporcionar-me informació i ajudar-me a solucionar molts dubtes i al meu germà, en Quico, per aguantar les meves enrabiades.

I per últim, i el més important, m'agradaria donar-li tots els meus agraïments i més al meu pare i professor de microbiologia, Jesús García Gil, que ha sigut el pilar i l'ànima del treball. Molt difícilment hagués aconseguit arribar fins on he arribat sense la seva ajuda.

Moltes gràcies a tots, ja que sense vosaltres aquest treball no hagués estat possible.