

# Formatos de Registros en Bases de Datos Biológicas

Curso de Bioinformática

Luis E. Garreta U

Pontificia Universidad Javeriana – Cali  
Facultad de Ingeniería - Carrera de Biología

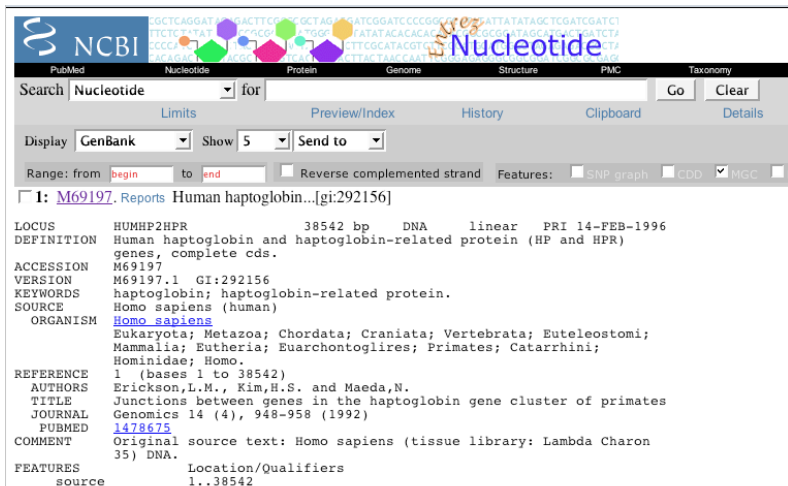
13 de agosto de 2018

# BDs Primarias

- ▶ Contienen datos biológicos originales
- ▶ Secuencias crudas o datos estructurales sometidos por la comunidad científica
- ▶ Estas son:
  - ▶ **GenBank**: mantenida por el NCBI (National Center for Biotechnology Information)
  - ▶ **EMBL**: mantenida por el EBI (European Bioinformatics Institute)
  - ▶ **DDBJ**: DNA Database of Japan

# Formato de los registros GenBank (GBFF)

- <https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html> explica los registros GBFF en detalle.



The screenshot shows the NCBI Nucleotide search results for the Human haptoglobin gene (M69197). The interface includes a search bar with 'Nucleotide' selected, and various filters and options like 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The search results are displayed in a table format, showing the locus, definition, accession number, version, keywords, source, organism, reference, authors, title, journal, PubMed ID, comment, and features.

**NCBI Nucleotide**

Search **Nucleotide** for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display **GenBank** Show **5** Send to [ ]

Range: from **begin** to **end** ☐ Reverse complemented strand Features: ☐ SNP graph ☐ CDD ☒ MGC ☐ H

☐ 1: [M69197](#). Reports Human haptoglobin...[gi:292156]

LOCUS	HUMHP2HPR	38542 bp	DNA	linear	PRI 14-FEB-1996
DEFINITION	Human haptoglobin and haptoglobin-related protein (HP and HPR) genes, complete cds.				
ACCESSION	M69197				
VERSION	M69197.1 GI:292156				
KEYWORDS	haptoglobin; haptoglobin-related protein.				
SOURCE	Homo sapiens (human)				
ORGANISM	<a href="#">Homo sapiens</a> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 38542)				
AUTHORS	Erickson,L.M., Kim,H.S. and Maeda,N.				
TITLE	Junctions between genes in the haptoglobin gene cluster of primates				
JOURNAL	Genomics 14 (4), 948-958 (1992)				
PUBMED	<a href="#">1478675</a>				
COMMENT	Original source text: Homo sapiens (tissue library: Lambda Charon 35) DNA.				
FEATURES	Location/Qualifiers				
source	1..38542				

# Formato de los registros GenBank (GBFF)

- ▶ El formato de información GBFF utilizado en GenBank es también compartido por otras bases de datos como EMBL y DDBJ.
- ▶ Difiere ligeramente del formato FASTA que actualmente se ha convertido en estándar en el campo de la bioinformática.

NCBI Nucleotide

Search: Nucleotide for [Go] [Clear]

Display: GenBank Show: 5 Send to: [ ]

Range: from [ ] to [ ] Reverse complemented strand [ ] Features: [ ] [ ] [ ] [ ]

[ ] I: M69197 Reports Human haptoglobin [gi:292156]

LOCUS HUMH2BFR 38542 bp DNA linear PRI 14-FEB-1996  
DEFINITION Human haptoglobin and haptoglobin-related protein (HP and HPR)  
genes, complete cds.  
ACCESSION M69197  
VERSION M69197.1 GI:292156  
KEYWORDS haptoglobin; haptoglobin-related protein.  
SOURCE Homo sapiens (human)  
ORGANISM Homo sapiens  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;  
Hominidae; Homo.  
REFERENCE 1. (bases 1 to 38542)  
Erickson, L.M., Kim, N.S. and Maeda, H.  
TITLE Junctions between genes in the haptoglobin gene cluster of primates  
JOURNAL Genomics 14 (4), 948-958 (1992)  
COMMENT Original source text: Homo sapiens (tissue library: Lambda Charon  
35) DNA.  
FEATURES  
source Location/Qualifiers  
1..38542  
/organism="Homo sapiens"  
/mol\_type="genomic DNA"  
/db\_xref="taxon:9606"  
/map="16p22.1"  
/tissue lib="Lambda Charon 35"  
434-3382

# Secciones Registro GBFF

Los registros de GenBank contienen tres secciones diferenciadas:

- Cabecera (header): que contiene identificadores, versión, fuente biológica, referencia, etc.
- Características (features): que encada sección de la secuencia contiene información sobre el comienzo, fin, longitud, tipo, etc.
- Secuencia: que contiene la cadena que representa la secuencia en sí misma.

```

LOCUS      SCUM8845      5028 bp      RNA      PLN      21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (Axl21 and Rev7p [REV7] genes, complete cds.
ACCESSION  U08845
VERSION   U08845.1  GI:1293613
KEYWORDS
SOURCE     Saccharomyces cerevisiae (baker's yeast)
ORGANISM   Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomycetes.
REFERENCE  1 [bases 1 to 5028]
AUTHORS   Torpey,L.E., Giblin,P.E., Nelson,J. and Lawrence,C.W.
TITLE      Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL    Yeast 10 (11): 1503-1509 (1994)
PUBMED     7871890
FEATURES   Location/Qualifiers
            source          1..5028
                        /organism="Saccharomyces cerevisiae"
                        /db_xref="taxon:4932"
                        /chromosome="IX"
                        /map="9"
                        /cmap="1..206"
                        /codon_start=3
                        /product="TCP1-beta"
                        /protein_id="AA08865.1"
                        /db_xref="GI:1293614"
                        /translation="SCIVMGISTGSLDHMGITAMRDLIVSGLKRAVVSAGSEA
            gene            687..2158
                        /gene="AXL2"
            CDS             687..2158
                        /note="plasma membrane glycoprotein"
                        /codon_start=1
                        /function="required for axial budding pattern of S.
                        cerevisiae"
                        /product="Axl2p"
                        /protein_id="AA08866.1"
                        /db_xref="GI:1293615"
                        /translation="MTOLGISLLLTATISLLHLVATPYEAYPIKGYPPYRWNESE
            ORIGIN
            1 gctctccat atacaacgt atctccact caggtttaga tctcaaac acggtatg
            61 ccgacatga acagtaggt atcgtcaga gttaacaagt aaacagca gtagtcagt
            121 ctgcattga accgctgaa gtctactaa ggggtgata catcatcgt gcaagacca
            181 gaaccgcaa tagacaacat atgatacata tttagatat acctcgaaa taataaacg
            241 ccacactgc attatttaa ttgaacag aacgaaaa tatcacta tataattga
  
```

# Cabecera (header)

**Locus** Identificador de la secuencia.

**Definition** Breve descripción de la secuencia.

**Accession** Identificador único de entrada, no varía aunque se modifique la secuencia.

**Version** Número de versión de la secuencia.

**GI** Identificador único de la secuencia, cambia con las modificaciones.

**Keywords** Palabras clave que describen la secuencia.

**Organism/Source** Nombre científico del organismo. Nombre común opcional en SOURCE. Taxonomía opcional en ORGANISM.

**Reference** Publicaciones relacionadas.

□ 1: Z92910. Homo sapiens HFE ...[gi:1890179]

[Related Sequences, OMIM,](#)

```

1 LOCUS      1a HSHFE                      1b 12146 bp  1c DNA   1d linear 1e PRI 23-JUL-1999
2 DEFINITION Homo sapiens HFE gene.
3 ACCESSION  Z92910
4 VERSION    Z92910.1 5GI:1890179
6 KEYWORDS   haemochromatosis; HFE gene.
7 SOURCE     human.
8 ORGANISM   Homo sapiens
              Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
              Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
9 REFERENCE  1 (bases 1 to 858)
  AUTHORS    Albig,W., Drabent,B., Burmester,N., Bode,C. and Doenecke,D.
  TITLE      The haemochromatosis candidate gene HFE (HLA-H) of man and mouse is
              located in syntenic regions within the histone gene cluster
  JOURNAL     J. Cell. Biochem. 69 (2), 117-126 (1998)
  MEDLINE    98208340

```

# Información contenida en la sección Locus (ejemplo)

	nombre	longitud	tipo	división GenBank	fecha modificación
LOCUS	LISOD	756 bp	DNA	linear	BCT 30- JUN-1993

1: Z92910. Homo sapiens HFE ...[gi:1890179]

[Related Sequences, OMIM, F](#)

1 LOCUS 1a HSHFE 1b 12146 bp 1c DNA 1d linear 1e PRI 23-JUL-1993  
 2 DEFINITION Homo sapiens HFE gene.  
 3 ACCESSION Z92910  
 4 VERSION Z92910.1 5 GI:1890179  
 6 KEYWORDS haemochromatosis; HFE gene.  
 7 SOURCE human.  
 8 ORGANISM [Homo sapiens](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.  
 9 REFERENCE  
 1 (bases 1 to 858)  
 AUTHORS Albig, W., Drabent, B., Burmester, N., Bode, C. and Doenecke, D.  
 TITLE The haemochromatosis candidate gene HFE (HLA-H) of man and mouse is  
 located in syntenic regions within the histone gene cluster  
 JOURNAL J. Cell. Biochem. 69 (2), 117-126 (1998)  
 MEDLINE [98208340](#)

# Acerca del campo Accession y el campo GI

- ▶ Generalmente los identificadores de acceso (Accession) son combinación de una letra o varias letras junto con números.
- ▶ U12345 ó AF123456 son ejemplos de este tipo de identificadores.
- ▶ Una vez asignado un identificador de acceso a un registro este número nunca cambiará, incluso si la información del registro se modificara, (por ejemplo, cambiando el registro para hacer la secuencia más completa).
- ▶ En cambio el número GI (GenInfo) sirve para identificar cada una de las nuevas versiones o modificaciones que recibe un registro determinado.
- ▶ **Ejemplo:** Cuando se añade una nueva entrada a GenBank se le da un número de acceso (p.e. AF000001). Dado que esta es la primera versión, se le añadirá al identificador un "0.1", quedando de esta manera: AF000001.1. De forma análoga el registro recibe un número GI (p.e. 1234567).



# Características (features)

Es la sección más importante de los registros, ya que contiene la representación directa de la información biológica de los datos en el sistema.

La característica “**source**” (origen) siempre debe aparecer obligatoriamente junto con la localización, **organism** y **db\_xref** (la referencia a su id taxonómico).

```
FEATURES
    source
        Location/Qualifiers
            1..12146
            /organism="Homo sapiens"
            /mol_type="genomic DNA"
            /db_xref="taxon:9606"
            /chromosome="6"
            /map="6p"
            /clone="ICRFy901D1223"
            /clone_lib="ICRF YAC-library"
    gene
    exon
    CDS
        join(1249..1324,4652..4915,5125..5400,6494..6769,
        6928..7041,7995..8035)
        /gene="HFE"
        /function="iron metabolism"
        /note="haemochromatosis candidate gene"
        /codon_start=1
        /protein_id="CABO7442.1"
        /db_xref="GI:1890180"
        /db_xref="GOA:Q30201"
        /db_xref="UniProt/Swiss-Prot:Q30201"
        /translation="MGPRARPALLLLMLLQTAVLQGRLLRSHSLHYLFMGASEQDLGL
        SLFEALGYVDDQLFVFDHESRRVEPTPVVSSRISSQNLQLSQSLKGWDHMFVTVD
        WTIMENHNHNSKESHTLQVILGCEMQEDNSTEGYWKYGYDGGDHLEFCPDTLDWRAAEP
        RAWPTKLEVERHKIRARQNRAYLERDCAQLQQLLELGRGVLDQQVPPPLVKVTHHVT
        SVTTLRCRALNYYPCNITMKWLKDKQPMDAKEFEKPDVLPNGDGTQGVITLAVPPGE
        EQRYTCQVEHPGLDQPLIVIWEPSPSGTLVIGVISGIAVFVVILFIGILFIILRKRG
        SRGANGHYVLAERE"
    intron
    polyA_signal
```

# Características: otros atributos

- ▶ **/organism**: nombre del organismo de la secuencia.
- ▶ **/gene**: nombre del gen relacionado con la secuencia.
- ▶ **/product**: producto génico de la secuencia.
- ▶ **/direction**: dirección de la replicación del ADN.
- ▶ **/codon\_start**: primera base del primer codón completo (1,2 ó 3).

```

FEATURES                     Location/Qualifiers
    source                    1..12146
                                /organism="Homo sapiens"
                                /mol_type="genomic DNA"
                                /db_xref="taxon:9606"
                                /chromosome="6"
                                /map="6p"
                                /clone="ICRFy901D1223"
                                /clone_lib="ICRF YAC-library"
    gene                      1028..10637
                                /gene="HFE"
    exon                      1028..1324
                                /gene="HFE"
                                /number=1
    CDS                       join(1249..1324,4652..4915,5125..5400,6494..6769,
                                6928..7041,7995..8035)
                                /gene="HFE"
                                /function="iron metabolism"
                                /note="haemochromatosis candidate gene"
                                /codon_start=1
                                /protein_id="CABO7442.1"
                                /db_xref="GI:1890180"
                                /db_xref="GOA:Q30201"
                                /db_xref="UniProt/Swiss-Prot:Q30201"
                                /translation="MGPRARPALLLLMLLQTAVLQGRLLRSHSLHYLFMGASEQDLGL
                                SLFEALGYVDDQLFVFDHESRRVEPTPTVWSSRISSQNLQLSQSLKGWDMHMTVDV
                                WTIMENHNHNSKESHTLQVILGCEMQEDNSTEGYWKYGYDGGDHLEFCPDTLDWRAAEP
                                RAWPTKLEVERHKIRARQNRAYLERDCAQLQQLLELGRGVLDQQVPPPLVKVTHHVT
                                SVTTLRCRALNYPQNTIMKWLKDKQPDAAKEFEFKDVLPGDGTQYQGWITLAVPPGE
                                EQRYTCQVEHPGLDQPLIVIEWPSPSGTLVIGVISGIAVFVVILF IGILF IILKRKQG
                                SRGAMGHYVLAERE"
    intron                    1325..4651
                                /gene="HFE"
                                /number=1
    polyA_signal              10617..10622
                                /gene="HFE"

```

# Características: otros atributos de la secuencia

**CDS** (Coding sequence) contiene información de las secuencias codificantes para proteínas.

**Exon** Región codificante.

**Intro** Intrón.

**gene** Nombre asignado a la región que se transcribe.

**mat\_peptide** Secuencia de la proteína después de sufrir modificaciones postraduccionales.

**misc\_feature** Características que no encajan en otras entradas.

FEATURES	Location/Qualifiers
<b>source</b>	1..12146 /organism="Homo sapiens" /mol_type="genomic DNA" /db_xref="taxon:9606" /chromosome="6" /map="6p" /clone="ICRFy901D1223" /clone_lib="ICRF YAC-library" 1028..10637 /gene="HFE" 1028..1324 /gene="HFE" /number=1 join(1249..1324,4652..4915,5125..5400,6494..6769,6928..7041,7995..8035) /gene="HFE" /function="iron metabolism" /note="haemochromatosis candidate gene" /codon_start=1 /protein_id="CABO7442.1" /db_xref="GI:1890180" /db_xref="GOA:Q30201" /db_xref="UniProt/Swiss-Prot:Q30201" /translation="MGPRARPALLLLMLLQTAVLQGRLLRSHSLHYLFMGASEQDLGLSLFEALGYVDDQLFVYDHSRRVEPRTFVSSRISSQMWLQLSQSLKGWDMHFTVDFWTIMENHNHSHKESHTLQVILGCEMQEDNSTEGYWKYGYDGGDHLEFCPDTLDWRAAEPRAWPTKLEVERHKIRARQNRAYLERDCPAQLQQLLELGRGVLDDQVPPPLVKVTHHVTSVTTLRCLALNYYPNITMKWLKDKQPMDAKEFEFKDVLPGNDGTQGYGUITLAVPPGEQRYTCQVEHPGLDQPLIVIEWPSPSGTLVIGVISGIAVFVVILFIGILFIILRKRGSRGANGHYVLAERE" 1325..4651 /gene="HFE" /number=1 10617..10622 /gene="HFE"
<b>gene</b>	
<b>exon</b>	
<b>CDS</b>	
<b>intron</b>	
<b>polyA_signal</b>	

# Secuencia

La última parte de los registros GenBank es la secuencia misma.

La línea BASE COUNT contabiliza el número de cada una de las bases. Ejemplo:

```

BASE COUNT      1510 a   1074 c   835 g   1609 t
ORIGIN
    1 gatcctccat atacaacggt atctccacct caggtttaga totcaacaac ggaaccattg
   61 ccgcacatgag acagtttagt atcgtcgaga gttacaagct aaaacagaga gtatgcagct
  121 ctgcacctga agcccggtgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
  181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaacgc
  241 ccacactgtc attattataa ttagaanaac aacgcacaaa ttatccacta tataattcaa
  301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
  361 attttggcaa cttatgtttc ctcttcgagc agtactcgag cccgtctcca agaatgtaat
  421 aataccatcc gtaggtagtg ttaagatagc catctccaca acctcaaacg tccttgccga
  481 gagtcgcctt cctttgtcga gtaattttca cttttcatat gagaacttat ttctttatc
  541 ttactctcca catcctgtag tgattgacac tgcacacgcc accatcacta gaagacaga
  601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga ttctcgtctt ccaacatcta
  661 cgtatatcaa gaagcattca cttaccatga cacagctcca gatttcatta ttgctgacag
  721 ctactatata actactccat ctagtatggg ccacgcctta tgaggcatat cctatcgaaa
  781 aacataaccc cccagtgcca agagccaatg aatggttac attcgaatt tcacatgata
  841 cctataaato gtctgtagac aagacagctc aaataacata caattgtctc gacttaccca
  901 gctggctctc gtttgactct agttctagaa cgtctccagg tgaaccttat tctgactaac
  961 tacctgatgc gaacacccag ttgtatttca atgtataact cggaggtcaag gactctgcoy
 1021 aacgcacgtc tttgaacaat acataccaat ttgtgttac aaacgcctca tccatctcgc
 1081 tacctcgaga ttcaatcta ttggcgttgt taataaacta tggttatact aacgcacaaa
 1141 acgctctgaa actagatcct aatgaagctc tcaacgtgac ttgtgacgt tcaatgttca
 1201 ctacgaaga atccatttg tctattacg gacgttcca gttgtataat gccgcgttac
 1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggaaggca ccggtgataa
 1321 aetcgggcat tgctccagaa acgaagctaca gttttgtcat catcgctaca gacattgaag
 1381 gatttttctc cgttgaggta gaattcgaaat tegtctcgg ggctccacag taaactactt
 1441 ctattcaaaa tagtttgata atcaacgtta ctgcacacag taacgtttca tatgacttac
 1501 cttcaacta tgtttatctc gatgacgac ctattttctc tgataaatg ggtctataa

```