



**UNIVERSITAT  
ROVIRA i VIRGILI**

**ANÀLISI D'ORTOLOGIA I ORGANISMES MODEL  
PER A L'ESTUDI DE MALALTIES GENÈTIQUES  
HUMANES MITJANÇANT EL DESENVOLUPAMENT  
D'UNA APLICACIÓ BIOINFORMÀTICA**

**TREBALL FINAL DE GRAU BIOTECNOLOGIA**

**Javier Vega Cuadros**

**Doble Grau Biotecnologia i Enginyeria Informàtica**

**Tutor acadèmic:** Santi Garcia-Vallvé

Departament de Bioquímica i Biologia Molecular

E-Mail: [santi.garcia-vallve@urv.cat](mailto:santi.garcia-vallve@urv.cat)



2024/25

Jo, Javier Vega Cuadros, amb DNI 48274138J, soc coneixedor de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guia per estudiants (aprovada el juliol 2017) i afirmo que aquest TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 8 de Gener de 2025

(signatura)

# 1. Agraïments

Primerament, aquest treball de fi de grau està inspirat en la meva anterior estada de pràctiques curriculars. És del meu plaer agrair a l'empresa [Anaxomics S.L](#), amb seu a Barcelona, per la seva amabilitat i confiança de cara a agafar-me com a becari de pràctiques. Gràcies a ells he pogut gaudir d'una estada com a bioinformàtic de veritat. Especialment cal mencionar al meu tutor de pràctiques Pedro Matos per guiar-me d'una manera excel·lent durant tot el procés, i al Jose Luís Ruiz per prendre el relleu com a cotutor i la seva ajuda en caires més informàtics.

Posteriorment em cal agrair al Dr. Santi Garcia-Vallvé per depositar la seva confiança en mi com al meu tutor de treball de fi de grau i estar present en els moments necessaris.

En darrer lloc, agrair a totes les persones que sempre han estat presents i m'han motivat a seguir sempre endavant, de manera curricular i extra curricular.

# Índex

1. Agraïments.....	4
2. Resum .....	6
3. Paraules Claus.....	7
4. Abreviatures.....	7
5. Definicions:.....	7
6. Introducció.....	9
6.1 Ortologia i malalties genètiques .....	9
6.2 Motivació del treball.....	12
7. Hipòtesi de treball i objectius .....	15
8. Metodologia.....	16
8.1 Selecció dels organismes model.....	16
8.2 Selecció de les malalties genètiques.....	16
8.3 Bases de dades utilitzades .....	18
8.4 Selecció de gens ortòlegs.....	19
8.4.1 Mètriques per escollir els gens ortòlegs .....	20
8.5 Creació del codi Python.....	21
8.5.1 Dependències.....	21
8.5.2 Funcionament del programa .....	22
9. Resultats i discussió.....	25
10. Conclusions .....	31
11. Bibliografia .....	33
12. Annexos .....	37
12.1 Taules SQL per a cada alineament.....	37

## 2. Resum

Aquesta tesi explora l'ús de les relacions d'ortologia entre els gens humans i els organismes model per identificar els millors candidats per estudiar malalties genètiques. Utilitzant eines i mètriques bioinformàtiques com InParanoid DIAMOND, l'estudi alinea proteomes i avalua similituds de seqüència, mitjançant un programa escrit en Python. Aquesta aplicació descarrega d'UniProt els proteomes dels organismes model i les seqüències de les proteïnes de cada malaltia seleccionada. Els organismes model han estat cinc, escollits en base als més comuns per realitzar estudis al laboratori. Les malalties genètiques humanes han estat sis, tres escollides per demostrar el correcte funcionament del programa i tres per proposar solucions en aquelles que no existeix cap estudi actual en organismes model. Un cop es tenen totes les seqüències, es realitza un alineament mitjançant InParanoid DIAMOND. Posteriorment, el programa avalua les mètriques d'aquest mètode d'alineament gen a gen i escull de tots els organismes model quin és el més proper a l'humà. El programa creat demana a l'usuari introduir els organismes model i els noms de les malalties, i extreu, per a cada malaltia, les mètriques gen a gen i la comparativa en global. Aquests resultats han estat analitzats i discutits per comprovar que les sortides són les esperades i que el programa funciona. Finalment, per veure la eficàcia del programa, es compararan els millors organismes model que proposa el programa creat amb altres alternatives com per exemple Ensembl Compara. Posteriorment es realitzen propostes de millora i objectius a llarg termini per la aplicació.

### 3. Paraules Claus

Ortologia, Gen/s Ortòlegs, Malalties Genètiques, Bioinformàtica, Proteïnes, UniProt, Python.

### 4. Abreviatures

- BBS: Bardet-Biedl syndrome
- BLAST: Basic Local Alignment Search Tool
- CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
- FHS: Floating Harbor Syndrome
- HBL: Hipobetalipoproteinèmia
- HD: *Huntington's Disease* (Malaltia de Huntington)
- LUCA: Last Universal Common Ancestor
- PV: Policitemia Vera

### 5. Definicions:

- Aneuploidia: Condició present en organismes on el cariotip s'ha vist modificat en la quantitat de cromosomes presents respecte al característic per l'espècie.
- Cariotip: Conjunt de parells de cromosomes d'una cèl·lula, de forma, mida i nombre característics de cada espècie.
- Ciliopatia: Malaltia que afecta el cili primari dels microtúbuls, encarregat de senyals sensorials molt importants, com per exemple la vista.
- *Clustering*: Agrupament de dades amb aspectes en comú, en aquest cas, que comparteixen de la mateixa llavor.
- E-Value: És un paràmetre estadístic que descriu el nombre de vegades que es pot esperar veure per casualitat el resultat desitjat quan es realitza un alineament. En aquest cas, disminueix quan l'alineament és més exacte.
- FASTA: Format de dades bioinformàtiques basat en text que s'utilitza per emmagatzemar seqüències de nucleòtids o aminoàcids
- *Hashing*: Mètode informàtic per compactar informació semblant en caràcters de longitud fixa per reduir el cost de comparació.

- InParanoid: Algorisme emprat per trobar *best reciprocal hits*, encarregat de determinar la ortologia entre dos gens. Genera un *BitScore* a partir de la qualitat dels alineaments obtinguts ja sigui mitjançant BLAST o DIAMOND.
- InParanoid DIAMOND: Algorisme de funcionalitat similar a BLAST, encarregat de realitzar alineaments de seqüències gèniques per al seu posterior processat.
- InParanoidDB: Base de dades encarregada de desar els gens ortòlegs en parelles o per dominis trobats mitjançant InParanoid. Hi ha dos versions rellevants, InParanoidDB8, que empra BLAST per als alineaments, i InParanoidDB9 que fa servir DIAMOND. En major part es farà referència a la base de dades més actual.
- *Pipeline*: estructuració dels passos que segueix Anaxomics S.L. per realitzar els seus estudis *in silico* per poder oferir els resultats als clients. Més informació a <http://anaxomics.com>
- Paral·lelització: Mètode de programació que permet distribuir el codi a executar de manera simultània a diferents medis d'execució (diversos processadors, o processadors amb més d'un nucli) per reduir el temps d'execució.
- Taula SQL (*SQL Table*): Mètode d'emmagatzemar la informació en bases de dades relacionals de manera estructurada. Permet fer posteriors consultes, anomenades *quèries*, sobre aquestes relacions.

## 6. Introducció

### 6.1 Ortologia i malalties genètiques

Les malalties genètiques han estat presents a la espècie humana des de anys enrere. Això es coneix degut a múltiples jaciments arqueològics on s'han caracteritzat rastres biològics pertanyents a *Homo sapiens*. Aquestes troballes han pogut ser estudiades amb eines computacionals especialitzades en detectar aneuploïdies (Anastasiadou et al. 2024). Els resultats d'aquestes investigacions reporten la presència de malalties com la síndrome de Klinefelter que daten sobre la edat de coure, aproximadament al 4500 A.C. a la península ibèrica. Aquesta síndrome, causada per una trisomia dels cromosomes sexuals, és a dir, un cariotip 47 XXY, provoca infertilitat i certa predisposició a altres tipus de malalties com diabetis de tipus II (Lanfranco 2004).

Aquestes malalties, fins fa relativament poc, eren una gran causa de mortalitat degut al desconeixement del què les causava. Fins al descobriment del ADN i tot el que això implica, qualsevol aspecte relacionat amb la genètica passava completament desapercbut. Gregor Mendel va ser el pioner en introduir conceptes de herència genètica a finals del segle XIX. No va ser fins més tard, gràcies a la caracterització de la doble hèlix de ADN (Watson & Crick 1953), que es va començar a debatre sobre la funció dels gens dins del nostre conjunt genòmic (cariotip) de 46 cromosomes i com aquest passa a la nostra descendència.

Mitjançant processos de transcripció i traducció de gens codificants podem obtenir proteïnes, les quals realitzen múltiples funcions dins de l'organisme dels éssers vius. Una mutació no sinònima (és a dir, un canvi de codó que no resulta en la traducció al mateix aminoàcid) en la seqüència del ADN pot implicar errors en la traducció, per tant proteïnes amb lleugers canvis sobre l'estructura primària. Hi ha diversos tipus de mutacions: inserció, deleció, canvi del marc de lectura o polimorfisme. Segons el tipus de mutació, per molt petit que sigui el canvi en la seqüència d'aminoàcids pot provocar canvis significatius a la estructura secundària o terciària de la proteïna. Això es tradueix en canvis de la conformació en l'espai de la proteïna. Les proteïnes depenen de la seva estructura i plegament dins de les cèl·lules per poder realitzar la seva funció. Per exemple, en el cas dels enzims, no seria possible la unió del substrat al centre actiu d'aquests si no presenten

una estructura tridimensional capaç d'acceptar-lo. Si el substrat no es pot unir a l'enzim, la reacció no es catalitzaria mai, per tant es perdria la funció. (Pierce 2020, 471-502). Aquest mateix exemple és la base de múltiples malalties genètiques causades per mutacions gèniques, com per exemple la malaltia de Huntington. També hi ha mutacions cromosòmiques, que afecten a un fragment del cromosoma, i mutacions genòmiques, com la síndrome de Klinefelter, que afecten al cariotip.

En el cas de la malaltia de Huntington, descoberta inicialment el 1872 pel doctor nord-americà George Huntington, era molt imprevisible ja que no es coneixien els seus mecanismes d'acció ni es podia predir la manera d'heretar-la. Finalment, gràcies als avenços genètics mitjançant estudis amb bacteris i altres organismes que van donar lloc al descobriment dels enzims de restricció, es va trobar al 1983 el marcador genètic de la malaltia de Huntington (Kolata 1983). Els científics referien a aquest marcador com un lloc de restricció tan a prop del gen de la proteïna causant de la malaltia que la seva presència indica on està el gen causant de la malaltia. Ara quedava demostrar que hi havia mutacions en els gens presents a les famílies on hi era present, però això va obrir un debat ètic sobre l'experimentació en humans degut a la impossibilitat de simular la malaltia fora del cas real.

Actualment, gràcies a les tecnologies de nova generació, experimentar amb organismes model s'ha convertit en la alternativa més ràpida i eficaç per poder estudiar aquestes malalties. Per exemple, usant CRISPR, podem modificar la seqüència de bases nitrogenades, introduint material genètic propi de la malaltia humana, i així induir nosaltres mateixos la malaltia amb aquesta mutació dirigida, evitant mètodes de mutació menys específics, com l'ús de radiacions o compostos químics. Un cop localitzada la malaltia dins del genoma, el que realment és costós es determinar si les variacions gèniques esmentades anteriorment son les causants de la malaltia en si. Per aquest cas, s'utilitzen diversos mètodes per no haver d'experimentar en humans o amb cultius, ja que el cost tant econòmic com temporal seria molt elevat degut al propi cicle de vida de la espècie (Yamamoto et al. 2024). La opció més viable és emprar **organismes model**.

Moltes de les característiques genètiques son compartides entre organismes degut a la evolució, on, a través d'events d'especiació i diferenciació s'han generat tots els organismes a partir del *Last Universal Common Ancestor* (LUCA) (Weiss et al. 2018). Aquest organisme, com el seu nom insinua, és el darrer ancestre comú entre els

organismes dels tres dominis. Va ser la primera cèl·lula amb la capacitat de portar a terme funcions metabòliques de síntesi de proteïnes a partir del seu genoma, que conté la informació genètica que ha passat a la seva descendència gràcies a la evolució. (Forterre 2024). Per aquest motiu, organismes com la mosca de la fruita o el cuc pla (*Drosophila melanogaster* | *Caenorhabditis elegans*) presenten certa similitud als seus gens respecte del genoma humà. Si ens apropem filogenèticament als humans, trobem organismes més complexos, com és el cas de la rata o el ratolí (*Rattus norvegicus* | *Mus musculus*) que encara tenen més similituds que els esmentats anteriorment. Això implica poder estudiar i conduir experiments genètics en aquests organismes model sense haver d'usar éssers humans. Aquests gens que comparteixen similituds d'estructura i funció a través de les diferents espècies s'anomenen gens ortòlegs.

Els gens ortòlegs són un tipus de gens homòlegs que estan relacionats entre si per events d'especiació, és a dir, de manera vertical en la formació de noves espècies en el curs de la evolució. Cal no confondre'ls amb els gens paràlegs, que son aquells que es relacionen per events de duplicació, és a dir, de manera horitzontal s'obtenen dos gens que no tenen perquè compartir funció o especificitat (Koonin 2005) (Altenhoff et al. 2019). Degut a la importància en la similitud d'estructura i funció en els gens ortòlegs, s'han establert nombroses bases de dades específicament per desar aquestes relacions d'espècie a espècie per facilitar els estudis. Per tant, es pot concloure que l'ortologia estableix les bases per experimentar en organismes model les variacions en la funció molecular darrere de mutacions genètiques humanes.

## 6.2 Motivació del treball

El propòsit d'aquest treball és continuar desenvolupant de manera personal el projecte font de les pràctiques externes realitzades l'estiu del 2023, on vaig realitzar un estudi comparatiu d'estratègies per obtenir gens ortòlegs. Aquesta comparativa serviria posteriorment per generar una base de dades amb els gens ortòlegs de la espècie humana envers a els organismes model d'interès de la empresa per futures investigacions. Les tres estratègies proposades pel tutor de pràctiques van ser les següents:

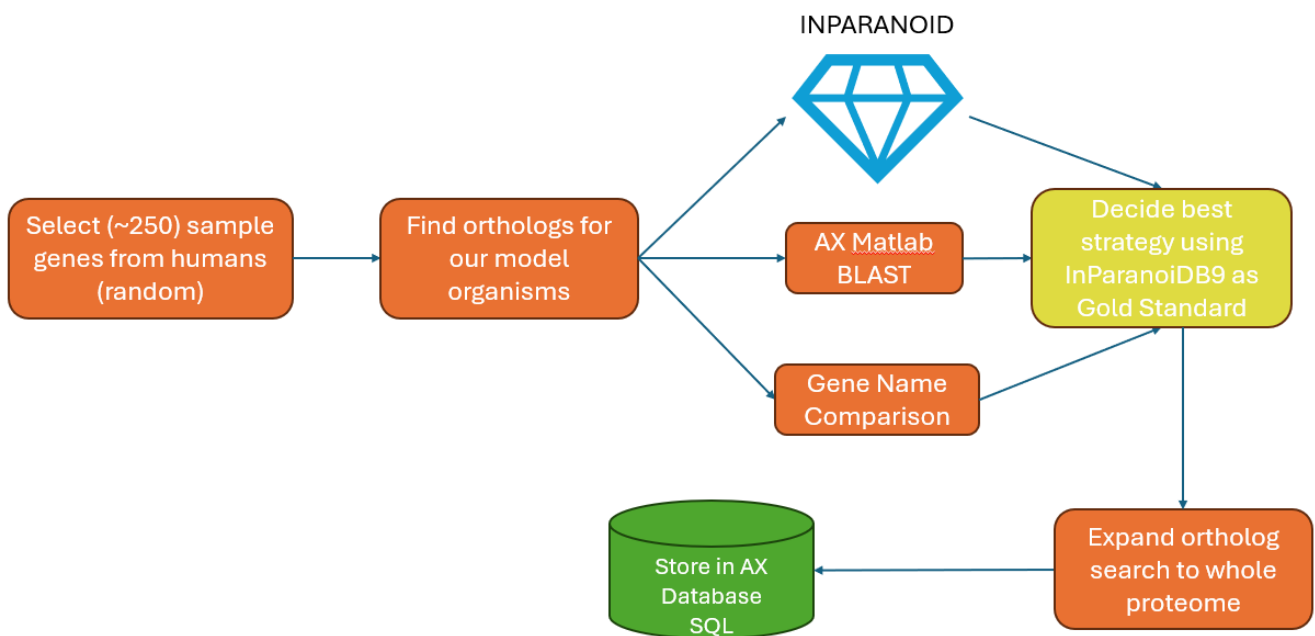


Figura 1: Pipeline de l'objectiu de la estada de pràctiques externes a Anaxomics S.L.

Primerament, realitzar cerques pel mateix nom de gen a bases de dades com UniProt. Aquest mètode és bastant fiable, ja que majoritàriament aquestes bases de dades empen els mateixos noms o molt similars per a gens ortòlegs ([referència UniProt](#)) (UniProt Consortium, 2023). El problema es que no tots els que ho son tenen el mateix nom, degut a conflictes segons el conveni emprat per donar el nom. En segon lloc, realitzar alineaments de forma manual i comparant percentatges de similitud mitjançant BLAST. Aquest segon mètode també es bo, ja que la característica d'aquests gens es que comparteixen un percentatge d'homologia bastant elevat, la qual cosa es pot fer servir per determinar ortologia entre dos seqüències. De totes maneres, realitzar els alineaments és lent, i fent servir els paràmetres predeterminats de BLAST+ (Matriu BLOSUM62, utilitzant un *gap open penalty* de 11 i *gap extension penalty* de 1) no obtenim el màxim

d'ortòlegs possibles. Finalment, i també el mètode més precís dels tres, és fer servir InParanoid DIAMOND (Persson & Sonnhammer 2022).

Aquest algorisme permet utilitzar millor el poder computacional de la màquina on s'està fent servir, de tal manera que és més ràpid realitzar els alineaments. També es pot seleccionar la sensibilitat amb la que es realitzen, per millorar la predicció sacrificant el temps que triga en executar els alineaments, tal com mostra la taula 1.

Taula 1: Comparativa de les configuracions de sensibilitat de InParanoid DIAMOND (Persson & Sonnhammer 2022)

Ortholog prediction method	Number of ortholog pairs	Average Jaccard	Average Unique InParanoid-BLAST	Average Unique InParanoid-DIAMOND	Runtime (hours)	Runtime (CPU hours)
InParanoid-BLAST	12,576,950	-	-	-	166.50	5940.12
InParanoid-DIAMOND sensitive	9,594,028	0.698	22.05%	8.17%	11.42	90.37
InParanoid-DIAMOND more sensitive	9,661,846	0.703	21.41%	8.30%	11.98	108.38
InParanoid-DIAMOND very sensitive	9,946,816	0.743	17.18%	8.50%	10.48	135.82
InParanoid-DIAMOND very sensitive with composition based statistics (option 4)	9,640,672	0.683	22.22%	9.48%	16.13	424.42
InParanoid-DIAMOND ultra sensitive	10,065,934	0.751	16.00%	8.86%	19.92	386.32

Al ser una aproximació més permissiva, permet detectar alineaments com a gens ortòlegs que amb altres mecanismes s'escaparien, amb la suficient confiança de que son veritablement ortòlegs. També permet detectar ortologia de dominis, mitjançant l'algorisme Domainoid, que pot resultar interessant, ja que aquests fluctuen de manera independent a la proteïna sencera durant l'especiació (Persson & Sonnhammer 2023).

Com es pot observar a la figura 2, d'un estudi comparatiu sobre un recull de 234 proteïnes a l'atzar, les quals podien tenir relació d'ortologia o no, el mecanisme que més ha trobat ha estat InParanoid DIAMOND. Per aquest motiu s'utilitzarà al llarg de tot l'estudi. Es pot observar que els mètodes de *Gene Name* i BLAST son prou correctes, i per tant, si no es disposa del temps i/o els requeriments per usar DIAMOND son alternatives prou bones, ja que no donen falsos positius.

ESTADÍSTICAS:	InParanoid	Gene Name	BLAST
TOTAL Ortólogos	135	86	83
%Ortólogos encontrados	57,69%	36,75%	35,47%
%Relativizado a InParanoid	100,00%	63,70%	61,48%
TRUE POSITIVES	-	86	83
TRUE NEGATIVES	-	27126	27126
FALSE POSITIVES	-	0	0
FALSE NEGATIVES	-	49	52

Figura 2: Estudi de la precisió dels mètodes esmentats prèviament per detectar ortòlegs.

InParanoidDB 9 és de les millors bases de dades que registra *Quest for Orthologs* (QfO), des del seu llançament al 2023. QfO és un esforç de tota la comunitat per crear un consorci per establir referents i avaluar nous mètodes en la cerca d'ortòlegs a partir de conjunts de dades preestablerts. Gràcies a aquest consorci, es pot reafirmar la fiabilitat de la base de dades i del seu algorisme DIAMOND, el qual serà emprat per realitzar els alineaments.

Un cop obtingut quin és el millor mètode, va faltar generar la base de dades amb totes les ortologies, però, degut a la finalització de la meua estada, no es va poder formalitzar a un model relacional. Simplement es van extreure els alineaments en format Excel, resultat més que satisfactori per a les pràctiques.

Ara, amb el coneixement de com realitzar els alineaments per obtenir la ortologia entre els gens, reprenc aquest projecte al Treball de Fi de Grau per usar el mateix procediment. El propòsit no és tornar a crear una base de dades, sinó determinar a partir d'aquesta relació de gens ortòlegs si es poden estudiar malalties genètiques humanes en organismes model.

## 7. Hipòtesi de treball i objectius

La hipòtesi d'aquest treball, és, doncs, que **l'estudi de relacions d'ortologia entre els gens de malalties genètiques humanes amb els gens dels organismes model permeten escollir quin seria idealment el millor organisme per estudiar aquestes malalties, prioritzant la fidelitat genètica.**

És de vital interès demostrar aquesta hipòtesi ja que, mitjançant enginyeria genètica, es podria arribar a induir la malaltia en aquests organismes. D'aquesta manera es podria modelitzar i administrar diferents tractaments o desenvolupar possibles cures que posteriorment es puguin traslladar a escala humana.

Ara bé, l'objectiu d'aquest treball és **crear un programa en python que, a partir de l'entrada del nom de la malaltia, generi una taula comparativa entre organismes model.** Aquesta taula contindrà les proteïnes de la malaltia humana, i les ortòlogues a cada organisme model. Segons la quantitat de correspondències i la confiança que retorni l'algorisme DIAMOND, s'escollirà l'organisme model més adequat per estudiar la malaltia per proximitat als gens humans.

Posterior a la creació del programa cal establir el segon objectiu, que serà demostrar que el programa és efectiu i funciona. Per aconseguir-ho s'escolliran tres malalties genètiques actualment estudiades en algun dels organismes model proposats. Aquestes serviran per validar si l'organisme que proposa el programa és el mateix o similar genèticament a el que fan servir els investigadors. Posteriorment, per donar-li ús al programa, es faran servir tres malalties de les quals no hi consta cap estudi actual amb organismes model. El programa farà la proposta de quin seria el millor per modelar aquesta malaltia en futures investigacions.

La primera tasca és obtenir la seqüència de les proteïnes humanes associades als gens de la malaltia de forma programàtica mitjançant UniProt. Un cop obtingudes, la segona tasca serà realitzar els alineaments mitjançant InParanoid DIAMOND. Per a cada correlació, la tercera tasca serà comparar-les i determinar quin serà l'organisme preferit per l'estudi. Finalment, la darrera tasca per validar la hipòtesi serà comparar aquesta estratègia de validació d'ortologia amb altres ja presents, com per exemple Ensembl Compara.

## 8. Metodologia

### 8.1 Selecció dels organismes model

Els organismes model emprats per l'estudi de l'ortologia sobre les malalties genètiques fa referència als utilitzats durant les pràctiques externes realitzades a l'estiu. Aquests organismes, presents a la taula 2, eren els de interès per a l'empresa i per aquest motiu han estat escollits, tot i que podrien haver sigut uns altres.

Taula 2: Noms i identificadors taxonòmics dels organismes model emprats en l'estudi

<b>Nom Comú</b>	<b>Nom Científic</b>	<b>Taxonomy ID</b>
<b>Rata</b>	<i>Mus musculus</i>	10090
<b>Ratolí</b>	<i>Rattus norvegicus</i>	10116
<b>Mosca de la Fruita</b>	<i>Drosophila melanogaster</i>	7227
<b>Peix Zebra</b>	<i>Danio rerio</i>	7955
<b>Cuc pla</b>	<i>Caenorhabditis elegans</i>	6239

El tutor de pràctiques durant la meua estada va suggerir aquests organismes. Els motius en són diversos, els mamífers per estreta relació, i la resta perquè són de curta gestació i/o esperança de vida i faciliten molt els estudis. També, per interès de cara al *pipeline* de treball de la empresa.

### 8.2 Selecció de les malalties genètiques

En aquest cas, s'ha volgut estudiar un primer recull de tres malalties genètiques presents a la taula 3 que actualment s'estan estudiant en organismes model. D'aquesta manera es pot comprovar si l'organisme escollit per la investigació coincideix amb el que la nostra aplicació proposa. Un segon recull de tres malalties permetrà poder recomanar quins serien els millors organismes model a utilitzar per al cas de futures investigacions.

Taula 3: Malalties genètiques humanes amb el seu Disease ID (UniProt) i un experiment amb un organisme estudiat, si escau. Les malalties amb variacions similars de la malaltia presenten Disease ID addicionals. La BBS conté molts, es posa l'enllaç directament.

Nom de la Malaltia	Organisme Estudiad	Disease ID	Alternative Disease ID's
Huntington's Disease (HD)	Ratolí (Alkanli et al. 2023)	<a href="#">DI-01754</a>	<a href="#">DI-01755</a> , <a href="#">DI-01756</a> , <a href="#">DI-01954</a> , <a href="#">DI-00414</a> , <a href="#">DI-01079</a>
Acatlasemia	Ratolí (Sunami et al. 2004)	<a href="#">DI-00016</a>	N/A
Hypobetalipoproteinemia (HBL)	Peix Zebra (Templehof et al. 2021)	<a href="#">DI-01587</a>	<a href="#">DI-03014</a> , <a href="#">DI-00308</a>
Bardet-Biedl syndrome (BBS)	N/A	<a href="#">DI-03107</a>	Link: <a href="https://www.uniprot.org/diseases?query=Bardet-Biedl+syndrome">https://www.uniprot.org/diseases?query=Bardet-Biedl+syndrome</a>
Polycythemia vera (PV)	N/A	<a href="#">DI-02712</a>	<a href="#">DI-01300</a>
Floating Harbor Syndrome (FHS)	N/A	<a href="#">DI-03389</a>	N/A

Algunes de les malalties sense organisme model en estudi és principalment degut a la presència de moltes variants fenotípiques de la mateixa malaltia, com és el cas de la síndrome Bardet-Biedl. La BBS és un trastorn hereditari poc freqüent considerat un model de ciliopatia no mòbil (Melluso et al. 2023), per tant té molts possibles quadres clínics. Això les fa malalties difícils de caracteritzar i poder portar-les a estudi sense fixar-se en una en concret. Per aquest estudi es proposaran quins són els organismes més fidels a la majoria de variants, seleccionant totes les proteïnes presents a UniProt involucrades en algun aspecte de la malaltia segons la base de dades Human Diseases. Es donarà el resultat de quin model es podria utilitzar per futurs estudis.

També existeix el cas dels trastorns genètics rars, com és el cas de la policitèmia vera (PV) o la síndrome Floating Harbor (FHS). En el cas de la PV, aquesta malaltia presenta un increment en la producció de glòbuls vermells. És una malaltia que s'acostuma a manifestar tard i l'enfoc és tractar els símptomes per reduir riscos de trombosi en pacients d'alt risc. Tot i que ha sigut possible induir la mutació *gain-of-function* al gen JAK2 en ratolins (Tefferi et al. 2021), principal causant d'aquesta malaltia, actualment no hi ha altres estudis en curs. Per altres motius personals també ha estat escollida. Per la FHS, es coneix que va lligada a una mutació del gen SRCAP que causa malformacions cranials,

retard en el creixement ossi i altres afectacions en el desenvolupament neuronal del individu (Dobrzynski et al. 2024). No hi ha cap estudi en curs en cap organisme model.

Respecte les malalties genètiques ben caracteritzades, és a dir, que han estat o actualment estan essent estudiades amb organismes model en podem destacar tres: La malaltia de Huntington, mencionada prèviament, és una malaltia neurodegenerativa que presenta disparitat cognitiva i dificultats en la mobilitat. Té un diagnòstic senzill, ja que els malalts presenten repeticions CAG al gen de la huntingtina (Stoker et al. 2022). En segon lloc, l'acatalasèmia, com el seu nom insinua, denota una falta de catalasa a la sang. La catalasa és un enzim antioxidant portador de grup Hemo, present sobretot a glòbuls vermells en humans. De vegades va associada a la presència de diabetis mellitus (Ando et al. 2024). Finalment, la HBL, caracteritzada per baixes concentracions o absència de colesterol de lipoproteïnes de baixa densitat i apolipoproteïna B al plasma sanguini. Això pot causar cirrosi, problemes de fetge i carcinomes hepatocel·lulars (Welty 2020).

### 8.3 Bases de dades utilitzades

- UniProtKB: La UniProt Knowledge Base recull informació funcional sobre proteïnes, tals com el nom de l'organisme, la seqüència d'aminoàcids, la funció... conté dades de dos fonts, Swiss-Prot, on les dades estan curades i TrEMBL, on l'anàlisi ha estat computacional i requereix d'autenticació manual. link: <https://www.uniprot.org/uniprotkb>
- UniProt Proteomes: Apartat de suport d'UniProt, conté el proteoma sencer de les espècies especificades pel Taxonomy ID. Utilitzat per extreure el proteoma sencer dels organismes model per alinear-lo sota DIAMOND. link: <https://www.uniprot.org/proteomes>
- UniProt Human Diseases: Apartat de suport d'UniProt, conté informació sobre les malalties genètiques humanes sota un Disease ID. Aquestes també presenten els gens i l'enllaç a la seva entrada a UniProtKB. Utilitzar per trobar les proteïnes associades als gens de cada malaltia de manera programàtica. link: [https://www.uniprot.org/diseases?query=\\*](https://www.uniprot.org/diseases?query=*)
- EnsemblCompara: Base de dades de genòmica que posa al abast múltiples eines bioinformàtiques, com BLAST, buscador de variants... Dona informació

sobre els gens de diverses espècies, i conté un apartat d'ortologia. link: <https://www.ensembl.org/index.html>

## 8.4 Selecció de gens ortòlegs

Com es pot observar a la taula 1, l'algorisme DIAMOND sacrifica una part de sensibilitat a l'hora de fer l'alineament respecte al *Gold Standard* de BLASTP (Altschul et al. 1997), però redueix el temps d'espera a hores. (Buchfink et al. 2021). La seva opció *--very-sensitive* i *--ultra-sensitive* son el millor equilibri entre temps emprat i precisió, la qual fa perfecte aquest algorisme per realitzar estudis amb grans conjunts de dades. També, degut a la seva imponent presència al *Quest for Orthologs* i un previ estudi comparatiu de bases de dades d'ortologia realitzat durant les pràctiques d'estiu, present als annexos, fa d'aquesta la millor opció per trobar gens ortòlegs.

InParanoid DIAMOND és un programa que requereix com a entrada els fitxers en format FASTA de les proteïnes a alinear. També es pot modificar altres paràmetres d'entrada, com la sensibilitat o com volem els resultats. Nosaltres la deixarem en la sensibilitat *--very-sensitive*, que és la predeterminada, i especificarem les carpetes on el nostre programa haurà descarregat prèviament els proteomes en format FASTA. També especificarem una carpeta de sortida, on, per cada espècie, obtindrem les *SQL Tables*.

DIAMOND aprofita tot el poder computacional del que disposa la màquina mitjançant mètodes de paral·lelització, per tant vol dir que és un programa que gasta molts recursos, per realitzar els alineaments. De totes maneres, s'adapta a cada màquina, per tant modera el consum segons dels recursos computacionals que disposa. Utilitza índexs especialitzats per accelerar les cerques, un per a les seqüències de consulta i un altre per a les seqüències de la base de dades. També permet paral·lelització i estratègies de filtratge en clústers, cosa que permet localitzar coincidències inicials ràpidament a partir de fragments anomenats k-mers. Aquestes coincidències s'estenen utilitzant variants de l'algorisme Smith-Waterman, que alinea les seqüències localment avaluant similituds amb matrius de substitució com BLOSUM. Els resultats es filtren segons criteris com la puntuació de l'alineament, la cobertura i el E-Value, garantint que els millors alineaments indiquin relacions d'ortologia. (Buchfink et al. 2021).

### 8.4.1 Mètriques per escollir els gens ortòlegs

El nostre programa realitza els alineaments mitjançant DIAMOND. Un cop s'han trobat totes les relacions d'ortologia segons els paràmetres de sensibilitat especificats, DIAMOND assigna un valor de *BitScore* a cada alineament semblant al que donen altres mètodes d'alineament (Buchfink et al. 2021) i inversament relacionada amb l'E-value. Posteriorment, s'escull quin serà el parell de gens ortòlegs de llavor. Aquesta llavor és l'alineament més precís que s'ha trobat dins dels dos proteomes escollits, per tant amb un *BitScore* més elevat. Aquesta llavor s'utilitza com a referència per assignar puntuació a la resta d'alineaments. Per tal de determinar que els ortòlegs de llavor ho son veritablement, es calcula la diferència entre el *BitScore* de la llavor i la següent llavor més probable. Aquest paràmetre s'anomena *SeedScore*, i està normalitzat respecte a 1. Valors de *SeedScore* més propers a 1 indica major confiança en què la mètrica de *BitScore* serà correcta per tots els alineaments (Persson & Sonnhammer. 2023). Per poder comparar ortologies de diferents alineaments, s'utilitza la mètrica de *bootstrap*. Aquesta mètrica té un propòsit similar al *SeedScore*, però no només ho fa amb la llavor, i per tant per al nostre cas servirà per determinar quin ortòleg és més proper entre espècies per al mateix gen de la malaltia. Actualment es calcula amb un ajust a una corba de regressió, ja que està fortament correlacionada amb la diferència de *BitScores* (Persson & Sonnhammer. 2023), però el seu significat ve de la necessitat anterior de comparar la quantitat de vegades que, dins d'un mateix clúster de gens, la parella escollida com a ortòloga ha sortit com a més propera (Östlund et al. 2010). Valors més elevats de *bootstrap* indiquen que la parella alineada ha sortit més cops escollida com a llavor respecte al mateix nombre de comparacions, per tant, podem assumir que hi ha major percentatge d'ortologia comparant dos gens d'espècies diferents respecte el mateix gen de cada malaltia.

Un cop acabada l'execució, el codi InParanoid genera uns fitxers de sortida per a cada parella de fitxers FASTA alineats, que contenen una taula SQL amb el *SeedScore*, el *bootstrap* i altres dades d'interès obtingudes als FASTA. Aquestes contenen els alineaments ordenats prèviament. D'aquesta manera obtenim els gens ortòlegs per a cada parell d'espècies del nostre interès, en aquest cas, de cada organisme model respecte a la malaltia en humans.

Per al nostre estudi, el programa dissenyat prendrà com a ortòlegs de referència aquells amb un *SeedScore* superior al 0.99, i el valor de *bootstrap* servirà per desempatar i

escollir, per a la mateixa proteïna de la malaltia en humans, quina ha estat la proteïna més semblant en organismes model.

## 8.5 Creació del codi Python

Com la conclusió de les pràctiques externes realitzades amb Anaxomics és que InParanoid DIAMOND és dels millors mecanismes que existeixen actualment per trobar gens ortòlegs, cal crear un codi en Python capaç d'incorporar l'algorisme DIAMOND en l'aplicació. Aquest programa ha de oferir a l'investigador la capacitat de seleccionar els organismes model dels que disposa i la malaltia a investigar. D'aquesta manera pot extreure les seqüències de manera programàtica mitjançant UniProt. Un cop es tenen les seqüències descarregades en format FASTA, el programa executarà el mètode d'alineament InParanoid DIAMOND entre les seqüències proteïques de la malaltia i el proteoma de cada organisme model escollit. Un cop finalitzen els alineaments, el programa ha de ser capaç de comparar les mètriques que retorna InParanoid DIAMOND en format *SQL Table*. D'aquesta manera, el programa proposarà quin és l'organisme model més adequat per l'estudi després de fer les comparatives proteïna a proteïna.

### 8.5.1 Dependències

El codi està pensat per ser utilitzat en entorn Linux, amb una versió de Python no inferior a 3.10, i amb la dependència de Jupyter Notebook (format que intercala cel·les de documentació i cel·les de codi) instal·lada. Els requisits recomanats per córrer el codi sense problemes són com a mínim 8GB de memòria RAM i un processador de gamma mitja alta, com pot ser Intel Core i7-8700K. Les llibreries emprades per al codi del programa han estat diverses. La llibreria principal ha estat *pandas*, per llegir i treballar amb conjunts de dades a partir dels fitxers *SQL Table*. La llibreria *subprocess* ha permès cridar al algorisme InParanoid DIAMOND mentre s'executa el programa. La llibreria *os* ha servit per gestionar els directoris on es trobaven els fitxers necessaris per l'execució del programa. Finalment, la llibreria *tkinter* per permetre un format més amigable amb interfície gràfica per algunes cel·les com es pot apreciar a la figura 3.

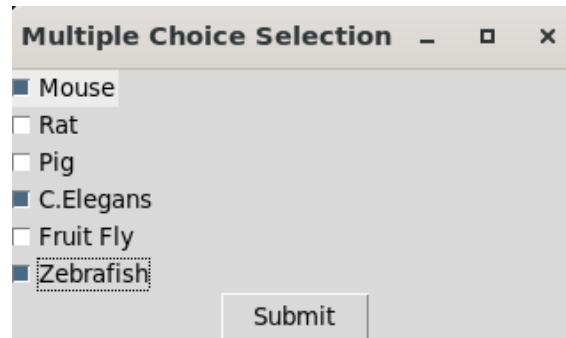


Figura 3: Exemple de interfície gràfica del programa per escollir els organismes model dels quals es vol trobar els gens ortòlegs.

### 8.5.2 Funcionament del programa

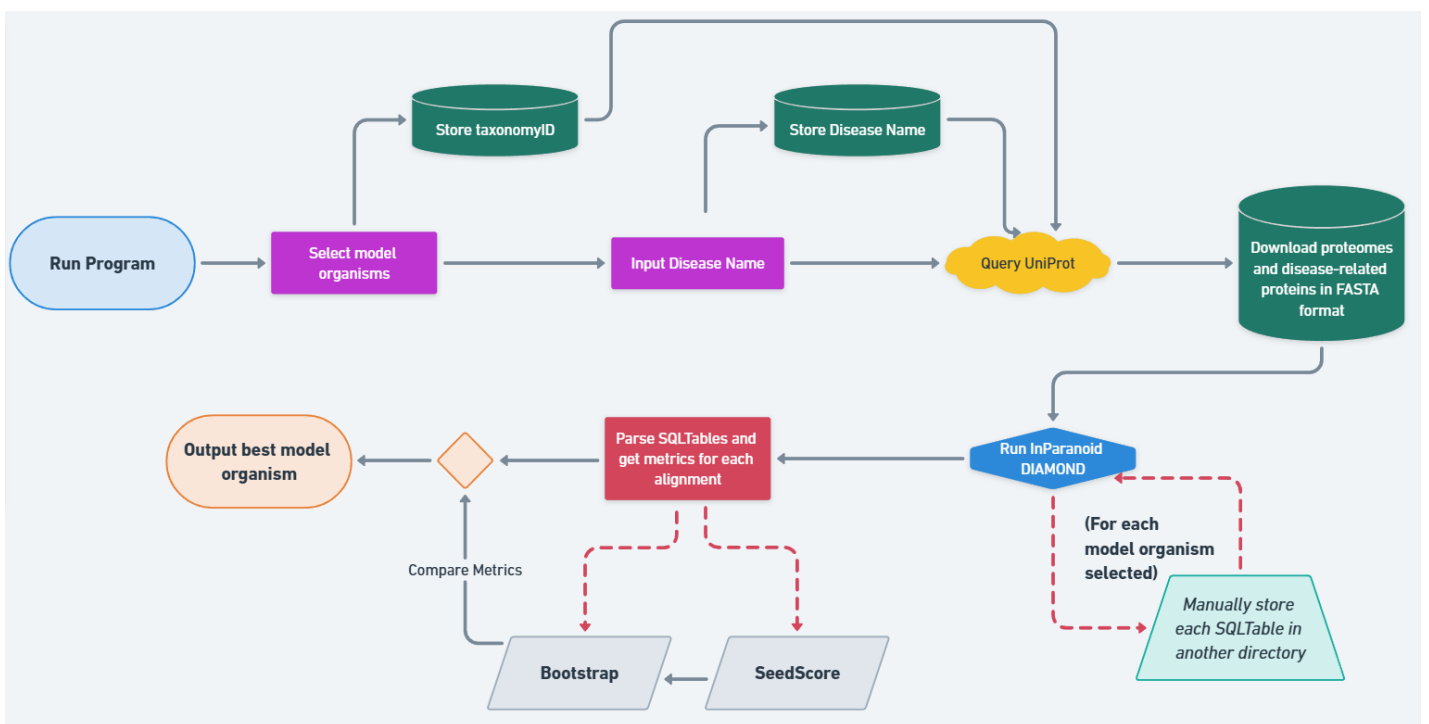


Figura 4: Diagrama que representa l'esquema de treball del codi des de la selecció d'organismes i malaltia fins a obtenir quin és el millor model.

El programa creat amb la intenció d'oferir el millor organisme model a partir del nom d'una malaltia genètica per al seu estudi està distribuït en cel·les de codi executable per parts, formant part del conjunt en un Jupyter Notebook. Per poder-lo executar només cal seguir els passos documentats a sobre de cada cel·la, que segueixen l'esquema plantejat a la figura 4, de manera resumida i esquemàtica, que també serveix per distingir quan s'està accedint a bases de dades, descarregant fitxers o realitzant alineaments.

La primera cel·la permet escollir quin/s del/s organismes proposats a la taula 4 es volen tenir en consideració de cara a l'experiment. Aquest pas és important per als investigadors

ja que potser no es disposa de tots els models a un estabulari, o no se'n han obtingut permisos per tots, així no es perd temps en computar tots els genomes d'aquells organismes. La selecció es fa mitjançant el *Taxonomy ID*.

En segon lloc es realitzarà una cerca en UniProt del nom de la malaltia, a l'apartat de [Human Diseases](#). Aquesta cerca es realitzarà de manera programàtica, mitjançant [UniProt REST API](#), (UniProt Consortium, 2023) per poder facilitar al codi la interpretació dels resultats, ja que la interfície d'UniProt és fàcil d'utilitzar i entendre per als humans però no per a les màquines.

De cada malaltia s'extrauran els gens més significants, per tal de no haver de comparar tot el proteoma humà, i es trobarà la proteïna associada a UniProt per *Homo sapiens*. Totes aquestes seran desades en un format FASTA una darrere l'altra.

En cada organisme model sí que serà necessari obtenir tot el proteoma, a [l'apartat de proteomes de UniProt](#) (UniProt Consortium, 2023). Es farà servir el filtre de proteoma de referència, per obtenir un estàndard vàlid. Es desarà també en format FASTA amb totes les entrades.

Un cop obtinguts tots els proteomes dels organismes model i les seqüències proteiques de la malaltia humana desitjada (només caldrà realitzar-ho el primer cop que s'executa el codi) finalment, es pot realitzar l'alineament mitjançant InParanoid DIAMOND per trobar la relació d'ortologia. Es separarà cada proteoma a una carpeta on seran afegides les proteïnes de la malaltia humanes, i hi haurà tantes carpetes com organismes model s'hagi decidit emprar. Posteriorment es cridarà l'algorisme per a cada carpeta.

Això produeix una nova carpeta per als resultats de l'alineament mitjançant DIAMOND, desats en un fitxer en format *SQL Table*, visible a la taula 4. Aquest fitxer serà l'entrada per la darrera cel·la de codi del nostre programa. Aquesta taula conté els alineaments amb una sèrie de mètriques explicades prèviament. Totes les taules amb els valors de *bootstrap* per cada gen estan presents als annexos.

Taula 4: Exemple de SQL Table dels ortòlegs per la malaltia de la hipobetalipoproteinèmia amb el proteoma del peix zebra amb títols que la fan més entenedora i amigable.

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	2383	Zebrafish	1,000	tr A0A8N7T7N7 A0A8N7T7N7_DANRE	100%	1,000
1	2383	Human	1,000	sp P04114 APOB_HUMAN	100%	1,000
2	416	Zebrafish	1,000	tr A0A8N7UZN6 A0A8N7UZN6_DANRE	99%	0.986
2	416	Human	1,000	sp Q9Y5C1 ANGL3_HUMAN	100%	1,000
3	376	Zebrafish	1,000	tr Q6PBX2 Q6PBX2_DANRE	72%	0.589
3	376	Human	1,000	sp Q9Y6B6 SAR1B_HUMAN	100%	1,000

D'entre les que més valor tenen per l'estudi, està el *BitScore*. Aquest, similar a BLAST, mesura la similitud de la seqüència independentment de la longitud de la seqüència de consulta i la mida de la base de dades i es normalitza en funció de la puntuació d'alineament en brut. Anteriors versions d'InParanoid usaven la mètrica de *bootstrap* (Persson & Sonnhammer. 2022), que servia per determinar la distància de les seqüències proteiques al graf que les interconnecta per determinar la ortologia. Com més gran el *bootstrap*, menor la distància entre aquestes seqüències, i per tant major confiança de que son veritablement ortòlogues. Actualment, InParanoid 9 utilitza el *SeedScore*, ja que van trobar una forta correlació entre la diferència de *BitScore* de les seqüències de cada espècie, i això era més barat de calcular. Està normalitzat, prenent valors de 0 a 1, essent 1 la distància més propera de *BitScores*.

Per a l'estudi, es combinaran les dues darreres mètriques emprades per InParanoid. Primerament, el programa descartarà els alineaments amb un *SeedScore* inferior a 0.99. Posteriorment, es sumaran els valors obtinguts de *bootstrap* per a cada comparació gen a gen. L'organisme model amb major *bootstrap* acumulat serà escollit. La espècie resultant serà proposada com la millor per realitzar estudis genètics per proximitat a les proteïnes de la malaltia humanes.

## 9. Resultats i discussió

Després de l'estudi sobre gens ortòlegs i organismes model, s'ha aconseguit crear una aplicació que fa l'esmentat prèviament. Aquest programa és capaç de rebre dos entrades. Primerament s'escull el *Taxonomy* ID dels organismes model de les espècies d'interès, acotades per les definides a la taula 3. La segona entrada consisteix en el nom de la malaltia genètica humana d'interès. Amb aquestes dades, el programa descarrega d'UniProt el proteoma sencer de cada organisme model seleccionat i les proteïnes específiques de la malaltia en fitxers FASTA individuals. La següent opció que dona la aplicació és realitzar els alineaments utilitzant l'eina InParanoid DIAMOND. Un cop obtingudes les *SQL Tables* de DIAMOND, l'usuari ha de moure a una carpeta aquelles de les quals en vol fer la comparativa. El darrer pas que fa l'aplicació és llegir les mètriques de *SeedScore* i *bootstrap* de cada *SQL Table* dins de la carpeta especificada i fer un filtratge. Només accepta aquells valors de *SeedScore* superiors a 0.99. Un cop filtrat, es fa una suma acumulada per cada espècie del valor de *bootstrap* associat a cada gen. Finalment, l'aplicació mostra per consola per cada malaltia quin ha sigut l'organisme model amb un *bootstrap* més gran entre les altres espècies. Això indica que, genèticament parlant, en conjunt dels gens implicats aquell organisme és millor que la resta.

Un cop executat el programa amb tots els organismes model per cada malaltia, cal comentar les sortides d'aquest. Com ha estat esmentat, el codi retorna uns fitxers en format *SQL Table*, que ens permetran realitzar manualment la comparativa gen a gen en cas d'interès particular. La darrera cel·la de codi s'encarrega de recórrer totes les *SQL Tables* i realitzar el còmput acumulatiu de *bootstrap* per a cada proteïna. Això permet no haver de buscar a cada fitxer i anar realitzant la suma acumulativa per veure qui té més punts. Al final, el programa ens mostra per pantalla per a quina malaltia present a les *SQL Tables* quin és el millor organisme per estudiar-lo, amb el valor de *bootstrap* acumulat (no té cap importància al comparar-ho entre malalties, ja que depèn del nombre de gens involucrats).

Analitzant primer el balanç global dels gens, podem observar la figura 5:

```

Processing disease: Acatalasemia
The best model organism for Acatalasemia is taxonomy ID 10090 with a cumulative bootstrap score of 1008.0.

Processing disease: Huntington
The best model organism for Huntington is taxonomy ID 10116 with a cumulative bootstrap score of 11288.0.

Processing disease: PV
The best model organism for PV is taxonomy ID 10116 with a cumulative bootstrap score of 6971.0.

Processing disease: BBS
The best model organism for BBS is taxonomy ID 10090 with a cumulative bootstrap score of 27869.0.

Processing disease: FHS
The best model organism for FHS is taxonomy ID 10116 with a cumulative bootstrap score of 4118.0.

Processing disease: HBLPPN
The best model organism for HBLPPN is taxonomy ID 10116 with a cumulative bootstrap score of 6574.0.

```

*Figura 5: Registre de consola per on l'aplicació ens retorna els millors organismes model amb el bootstrap acumulat, excloent aquells de SeedScore inferiors a 0.99*

Que podem resumir-ho en la següent taula, continuació de la taula 3 on hem seleccionat les malalties segons si tenien o no organisme model en estudi actualment:

*Taula 5: Comparativa per a les tres primeres malalties dels organismes principalment estudiats envers als organismes proposats pel programa. Proposta per a l'estudi de les tres darreres.*

<b>Nom de la Malaltia</b>	<b>Organisme esperat</b>	<b>Organisme Obtingut</b>
Huntington's Disease (HD)	Ratolí	Rata
Acatalasemia	Ratolí	Ratolí
Hypobetalipoproteinemia (HBL)	Peix Zebra	Rata
Bardet-Biedl syndrome (BBS)	N/A	Ratolí
Polycythemia vera (PV)	N/A	Rata
Floating Harbor Syndrome (FHS)	N/A	Rata

Es pot veure que el programa ha predit correctament l'organisme model per la malaltia de l'acatalasemia. En el cas de la malaltia de Huntington, l'organisme proposat (rata) es prou proper filogenèticament a la seva contrapart estudiada (ratolí), en el cas de la HBL, l'organisme no ha estat encertat.

Cal mencionar, que tots els organismes model han presentat com a mínim algun ortòleg respecte d'alguna de les proteïnes d'alguna malaltia. Per certes malalties alguns

organismes en concret no han trobat ortòlegs, com per exemple el cas de la policitèmia vera i *Caenorhabditis elegans*.

Per poder discutir amb més profunditat cada cas, cal analitzar cada *SQL Table* d'aquestes malalties que han presentat resultats diferents per veure gen a gen quin té el major *bootstrap*, denotant major percentatge d'ortologia.

Malaltia de Huntington: 6 gens en total

Taula 6: Comparativa entre el millor organisme proposat per el programa gen a gen, i comparativa del Bootstrap entre rata i ratolí

<b>Gen humà</b> (Gene Name)	<b>Gen organisme model</b> (UniProt ID   UniProt Entry Name)	<b>Bootstrap</b> (1r vs 2n)
HTT	P51111   HD_RAT	5450 (a 2 del ratolí)
TAF1	A0A8I6A7M0   A0A8I6A7M0_RAT	3498 (a 10 del ratolí)
JPH3	D3ZWH2   D3ZWH2_RAT	1211 (a 14 del ratolí)
XK	Q9QXY7   XK_MOUSE	751 (a 9 de la rata)
TBP	P29037   TBP_MOUSE i Q66HB1   Q66HB1_RAT	513 (empat de <i>bootstrap</i> amb ratolí però avantatge de <i>SeedScore</i> )
PRNP	P13852   PRIO_RAT	387 (a 29 del ratolí)

En el cas de la malaltia de Huntington, a la comparativa gen a gen de la taula 6, hem pogut apreciar que la diferència entre ambdues espècies murines no ha estat tan gran. De sis gens que componen la malaltia, quatre han estat lleugerament més propers els de rata. Un d'ells s'ha apropat més el de ratolí, i en el darrer, hi ha hagut un empat de *bootstrap*, però per *SeedScore* s'ha vist beneficiada la rata. Això ens demostra que aquests organismes són prou propers genèticament com per emprar-los indistintament. El motiu per escollir un o altre pot variar en funció de disponibilitat, costos, temps de vida i durada de l'estudi. També cal tenir en compte que no s'estan emprant subespècies de rata o ratolí en la comparativa proteòmica, s'estan emprant els models d'espècie representatius (*Rattus norvegicus* i *Mus musculus*) presents a UniProt.

## HBL: 3 gens en total

Taula 7: Comparativa entre els organismes model de rata, ratolí i peix zebra, amb diferències de bootstrap i SeedScore. En el cas del tercer gen, cap dels organismes ha demostrat una confiança prou com per dir que son ortòlegs.

Gen humà (Gene Name)	Gen Rata (UniProt Entry Name)	Gen Ratolí (UniProt Entry Name)	Gen Peix Zebra (UniProt Entry Name)
APOB	F1M6Z1_RAT <i>Bootstrap</i> : 5880 <i>SeedScore</i> : 1.000	APOB_MOUSE <i>Bootstrap</i> : 5944 <i>SeedScore</i> : 1.000	A0A8N7T7N7_DANRE <i>Bootstrap</i> : 2383 <i>SeedScore</i> : 1.000
ANGPTL3	F7FHP0_RAT <i>Bootstrap</i> : 694 <i>SeedScore</i> : 1.000	ANGL3_MOUSE <i>Bootstrap</i> : 594 <i>SeedScore</i> : 1.000	A0A8N7UZN6_DANRE <i>Bootstrap</i> : 416 <i>SeedScore</i> : 0.986
SAR1B	SAR1B_RAT <i>Bootstrap</i> : 400 <i>SeedScore</i> : 0.761	SAR1B_MOUSE <i>Bootstrap</i> : 400 <i>SeedScore</i> : 0.518	Q6PBX2_DANRE <i>Bootstrap</i> : 376 <i>SeedScore</i> : 0.589

Finalment, la malaltia de la hipobetalipoproteïnèmia, que la nostra aplicació proposa la rata, tot que ha estat estudiada en peix zebra. La comparativa present a la taula 7 ens mostra que el nostre programa ha detectat que el gen més versemblant al APOB en humans ha estat el de ratolí, el de ANGPTL3 ha estat rata i el gen SAR1B no ha sigut el prou proper per considerar-lo ortòleg veritable. En *bootstrap* acumulat, ha sortit beneficiada la rata, però un altre cop amb diferències poc significatives respecte al ratolí. El peix zebra sí que s'allunya una mica més.

Es possible que els estudis emprin el peix zebra en comptes de la rata per qüestions econòmiques o per altres motius més enllà de les qüestions d'ortologia i semblança en funció, com per exemple el fet que el peix zebra ha estat l'únic organisme model dels més senzills que contenia els 3 gens relacionats amb la malaltia. A la resta d'organismes els hi faltava més d'un gen i per tant han perdut punts al *bootstrap* acumulat.

Per tal de veure si la nostra aplicació ha estat encertada, cal comparar els nostres resultats amb el que hi ha disponible a la base de dades Ensembl Compara. Aquesta base de dades és una eina que ofereix una gran varietat de serveis per a estudis genòmics per moltes espècies. Un d'aquests serveis és l'apartat d'ortologia, on es pot introduir el nom d'un

gen per humans, i per als organismes que hi ha ortòleg et diu el *Target identity percentage*. Aquesta mètrica indica el tant per cent de la seqüència ortòloga que coincideix amb la seqüència humana (Harrison et al. 2024). És interessant realitzar una taula comparativa amb els organismes model predits a la aplicació envers dels millors organismes model (és a dir, amb el major *Target ID percentage*) que proposa Ensembl.

Per aquesta comparativa s’han escollit tres de les malalties que poden ser les més interessants segons el que s’ha vist a la taula 5. La hipobetalipoproteïnèmia, per determinar si el peix zebra té el major percentatge d’ortologia que la rata, l’acatalasèmia, per veure si el ratolí segueix essent la millor opció, i de les malalties sense estudis, la policitemia vera, que no té gaires gens, per veure si la predicció també és correcta.

Taula 8: Comparativa entre l’organisme predit pel nostre programa i l’organisme més estret relacionat al gen humà segons Ensembl Compara

Malaltia	Gene Name	Organisme Estudiat	Global Programa	Gen a Gen Programa	Ensembl Compara
HBL	APOB	Peix Zebra	Rata	Ratolí	Ratolí 70.79%
	ANGPTL3			Rata	Rata 78.46%
	SAR1B			-	Ratolí 98.99%
Acatalasèmia	CAT	Ratolí	Ratolí	Ratolí	Ratolí 89.37%
PV	JAK2	N/A	Rata	Ratolí	Rata 94.61%
	TET2			Rata	Rata 63.85%
	F5			Ratolí	Ratolí 70.96%

Com podem veure a la taula 8, la majoria de vegades que el programa prediu gen a gen el millor organisme model, gràcies a la mètrica de *bootstrap*, ho encerta respecte l’alternativa Ensembl. Això vol dir que podem extrapolar aquesta suma acumulada de *bootstraps* per dir, en general, per a tots els gens, quin organisme és més fidel genèticament als gens de la malaltia humana. En el cas del gen SAR1B, el nostre programa no ha tret cap ortòleg perquè la mètrica de *SeedScore* no era el prou bona per assegurar que els ortòlegs de llavor per aquell alineament eren ortòlegs veritables. En el cas de la JAK2, la base de dades Ensembl donava per vàlid el ratolí amb un percentatge

del 93.73%, com es pot veure a la figura 6. Per tant, no s'allunya tant de la sortida del nostre programa.

Selected orthologues [Hide](#)

Show	All	entries	Show/hide columns					Mus Musculus	
Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence		
Mouse ( <i>Mus musculus</i> )	1-to-1 <a href="#">View Gene Tree</a>	Jak2 ( <a href="#">ENSMUSG00000024789</a> ) <a href="#">Compare Regions</a> (19,29,229,228-29,290,480:1) <a href="#">View Sequence Alignments</a>	93.73 %	93.73 %	100	100.00	Yes		

Figura 6: Exemple de captura de pantalla d'Ensembl Compara per al gen de la JAK2, filtrant per nom a *Mus musculus*. Es pot observar el percentatge d'identitat, així com altres paràmetres i informació sobre l'ortòleg.

També veig interessant per al cas de la HBL, que, tot i que els estudis actuals s'estan realitzant en peix zebra, Ensembl dona com a millor alternativa les espècies murines. Això dona peu a pensar el que s'ha esmentat abans a la comparativa gen a gen. De vegades, no només es busca la major fidelitat genètica, sinó que també entren en joc altres factors externs. Aquests poden ser: facilitat a l'hora de modificar o mantenir aquests organismes, cicles de vida i reproducció, limitacions econòmiques...

Aquest treball de fi de grau aïlla aquestes condicions òptimes on no en falten recursos i s'hi pot dedicar temps i material sense escatimar. Evidentment això al món de la investigació no és així, però l'aplicació desenvolupada no té altres factors en compte que no siguin les relacions d'ortologia.

Una proposta de millora de l'aplicació podria ser precisament això. Afegir uns paràmetres d'entrada, com per exemple aquest tipus de restriccions. D'aquesta manera es podria, potser incorporant models d'intel·ligència artificial (IA), assegurar que cada organisme model predit s'ajusta a les condicions i necessitats de cada experiment, intentant buscar la major relació d'ortologia possible dins de les restriccions per garantir l'èxit de la investigació.

## 10. Conclusions

La hipòtesi inicial ha resultat ser certa. L'estudi de relacions d'ortologia entre els gens de malalties genètiques humanes amb els gens dels organismes model han permès escollir el millor organisme per a futures investigacions sobre cada malaltia. Els gens ortòlegs demostren similituds en quant a seqüència i funció de la proteïna resultant entre humans i organismes model, i aquesta característica ha sigut aprofitada per desenvolupar el programa.

Sabent que els gens ortòlegs presenten homologia de seqüència, el primer objectiu ha estat crear un programa per buscar aquestes relacions. La primera tasca ha consistit en descarregar les proteïnes involucrades a partir de l'API REST d'UniProt. Aquestes proteïnes han estat tant les de la malaltia en humans, com el proteoma de cada organisme model. Posteriorment, la segona tasca tracta de alinear les seqüències de cada proteïna humana amb tot el proteoma dels organismes model. D'aquesta manera, usant InParanoid DIAMOND obtenim aquelles relacions d'ortologia amb la suficient confiança que veritablement son ortòlegs. El nostre programa recull els paràmetres de *bootstrap* per a cada alineament que dona DIAMOND. Aquest paràmetre relaciona la distància que hi ha entre la seqüència de la proteïna de la malaltia i la proteïna ortòloga trobada a l'organisme model. Realitzant aquest pas per a cada malaltia i cada organisme model permet comparar quin de tots és el millor per a estudiar aquelles malalties que no en tenen cap organisme model assignat.

Per tal de demostrar l'efectivitat del programa al nostre segon objectiu, s'ha executat la nostra aplicació sobre malalties que ja estaven estudiades en determinats organismes model. Per saber més en detall què passava a cada cas, s'han fet taules comparatives gen a gen per a casos particulars i també s'ha utilitzat l'alternativa a la nostra aplicació, Ensembl Compara, per acabar de realitzar les comparatives.

Els resultats obtinguts de la demostració del programa han estat prou satisfactoris, tenint en consideració que InParanoid DIAMOND és un mètode d'alineament que sacrifica possibles ortòlegs assegurant la sensibilitat dels que si ho són. El programa tampoc té en consideració efectes externs com per exemple disponibilitat, temps... Com s'ha mencionat prèviament, una millora seria considerar totes aquestes variables com a restriccions. Mitjançant models d'intel·ligència artificial i la sortida del nostre programa es podria

predir el millor organisme model segons les situacions de cada investigació. Això obre les portes a futures modificacions i adaptacions del programa. Aquestes versions millorades podrien oferir un servei únic i personalitzat a cada equip de recerca per als seus estudis amb organismes model. En qüestió de menys d'una hora, es podria saber amb total certesa quins son els gens i en quin organisme model ens podem basar per estudiar aquestes malalties. Amb aquesta informació, s'estalviaria tota la recerca manual prèvia per començar la investigació i es podria dedicar més temps i recursos a altres aspectes que més ho necessitin.

## 11. Bibliografia

- Alkanli SS, Alkanli N, Ay A, Albeniz I. CRISPR/Cas9 Mediated Therapeutic Approach in Huntington's Disease. *Mol Neurobiol*. 2023 Mar;60(3):1486-1498. doi: [10.1007/s12035-022-03150-5](https://doi.org/10.1007/s12035-022-03150-5)
- Altenhoff AM, Glover NM, Dessimoz C. Inferring Orthology and Paralogy. *Methods Mol Biol*. 2019;1910:149-175. doi: [10.1007/978-1-4939-9074-0\\_5](https://doi.org/10.1007/978-1-4939-9074-0_5)
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997 Sep 1;25(17):3389-402. doi: [10.1093/nar/25.17.3389](https://doi.org/10.1093/nar/25.17.3389)
- Anastasiadou, K., Silva, M., Booth, T. *et al*. Detection of chromosomal aneuploidy in ancient genomes. *Commun Biol* 7, 14 (2024).<https://doi.org/10.1038/s42003-023-05642-z>
- Ando M, Fukushima K, Nishizaki K. The discovery of acatalasemia (lack of catalase in the blood) and its significance in human genetics. *Proc Jpn Acad Ser B Phys Biol Sci*. 2024;100(7):353-367. doi: [10.2183/pjab.100.024](https://doi.org/10.2183/pjab.100.024)
- Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 2021 Apr;18(4):366-368. doi: [10.1038/s41592-021-01101-x](https://doi.org/10.1038/s41592-021-01101-x)
- Dobrzynski W, Stawinska-Dudek J, Moryto N, Lipka D, Mikulewicz M. Floating-Harbor Syndrome: A Systematic Literature Review and Case Report. *J Clin Med*. 2024 Jun 12;13(12):3435. doi: [10.3390/jcm13123435](https://doi.org/10.3390/jcm13123435)
- Emma Persson, Erik L.L. Sonnhammer, InParanoid-DIAMOND: faster orthology analysis with the InParanoid algorithm, *Bioinformatics*, Volume 38, Issue 10, May 2022, Pages 2918–2919, <https://doi.org/10.1093/bioinformatics/btac194>
- Emma Persson, Erik L.L. Sonnhammer, InParanoidDB 9: Ortholog Groups for Protein Domains and Full-Length Proteins, *Journal of Molecular Biology*, Volume 435, Issue 14, 2023, 168001, ISSN 0022-2836 <https://doi.org/10.1016/j.jmb.2023.168001>.

- Forterre P. The Last Universal Common Ancestor of Ribosome-Encoding Organisms: Portrait of LUCA. *J Mol Evol.* 2024 Oct;92(5):550-583. doi: [10.1007/s00239-024-10186-9](https://doi.org/10.1007/s00239-024-10186-9)
- Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D196-203. doi: [10.1093/nar/gkp931](https://doi.org/10.1093/nar/gkp931)
- Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics.* 2011 Aug 31;12:357. doi: [10.1186/1471-2105-12-357](https://doi.org/10.1186/1471-2105-12-357)
- Kolata G. Huntington's disease gene located. *Science.* 1983 Nov 25;222(4626):913-5. <https://doi.org/10.1126/science.6314504>
- Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* 2005;39:309-38. doi: [10.1146/annurev.genet.39.073003](https://doi.org/10.1146/annurev.genet.39.073003).
- Lanfranco F, Kamischke A, Zitzmann M, Nieschlag E. Klinefelter's syndrome. *Lancet.* 2004 Jul 17-23;364(9430):273-83. doi: [10.1016/S0140-6736\(04\)16678-6](https://doi.org/10.1016/S0140-6736(04)16678-6)
- Melluso A, Secondulfo F, Capolongo G, Capasso G, Zacchia M. Bardet-Biedl Syndrome: Current Perspectives and Clinical Outlook. *Ther Clin Risk Manag.* 2023 Jan 30;19:115-132. doi: [10.2147/TCRM.S338653](https://doi.org/10.2147/TCRM.S338653)
- Peter W Harrison, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, and many more. Ensembl 2024, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D891–D899, <https://doi.org/10.1093/nar/gkad1049>
- Pierce, Benjamin A. *Genética: Un Enfoque Conceptual*. Panamericana, 2020. ISBN:[978-84-9835-216-0](https://doi.org/10.1016/S0140-6736(04)16678-6)
- Stoker TB, Mason SL, Greenland JC, Holden ST, Santini H, Barker RA. Huntington's disease: diagnosis and management. *Pract Neurol.* 2022 Feb;22(1):32-41. doi: [10.1136/practneurol-2021-003074](https://doi.org/10.1136/practneurol-2021-003074)

- Sunami R, Sugiyama H, Wang DH, Kobayashi M, Maeshima Y, Yamasaki Y, Masuoka N, Ogawa N, Kira S, Makino H. Acatlasemia sensitizes renal tubular epithelial cells to apoptosis and exacerbates renal fibrosis after unilateral ureteral obstruction. *Am J Physiol Renal Physiol*. 2004 Jun;286(6):F1030-8. doi: [10.1152/ajprenal.00266.2003](https://doi.org/10.1152/ajprenal.00266.2003)
- Tefferi, A., Vannucchi, A.M. & Barbui, T. Polycythemia vera: historical oversights, diagnostic details, and therapeutic views. *Leukemia* **35**, 3339–3351 (2021). <https://doi.org/10.1038/s41375-021-01401-3>
- Templehof H, Moshe N, Avraham-Davidi I, Yaniv K. Zebrafish mutants provide insights into Apolipoprotein B functions during embryonic development and pathological conditions. *JCI Insight*. 2021 Jul 8;6(13):e130399. doi: [10.1172/jci.insight.130399](https://doi.org/10.1172/jci.insight.130399)
- UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2023 Jan 6;51(D1):D523-D531. DOI: [10.1093/nar/gkac1052](https://doi.org/10.1093/nar/gkac1052)
- Watson, J., Crick, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **248**, 765, April 25 1953 (issued 1974). <https://doi.org/10.1038/248765a0>
- Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet*. 2018 Aug 16;14(8):e1007518. doi: [10.1371/journal.pgen.1007518](https://doi.org/10.1371/journal.pgen.1007518)
- Welty FK. Hypobetalipoproteinemia and abetalipoproteinemia: liver disease and cardiovascular disease. *Curr Opin Lipidol*. 2020 Apr;31(2):49-55. doi: [10.1097/MOL.0000000000000663](https://doi.org/10.1097/MOL.0000000000000663)
- Yamamoto, S., Kanca, O., Wangler, M.F. *et al*. Integrating non-mammalian model organisms in the diagnosis of rare genetic diseases in humans. *Nat Rev Genet* **25**, 46–60 (2024). <https://doi.org/10.1038/s41576-023-00633-6>

Zaleskas VM, Krause DS, Lazarides K, Patel N, Hu Y, Li S, Van Etten RA. Molecular pathogenesis and therapy of polycythemia induced in mice by JAK2 V617F. PLoS One. 2006 Dec 20;1(1):e18. doi: [10.1371/journal.pone.0000018](https://doi.org/10.1371/journal.pone.0000018)

## 12. Annexos

### 12.1 Taules SQL per a cada alineament

#### *Malaltia de Huntington – Rata*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	5450	Rat	1,000	tr G3V9P7 G3V9P7_RAT	100%	1,000
1	5450	Rat	1,000	tr A6IK16 A6IK16_RAT	100%	1,000
1	5450	Rat	0.919	tr A0A8I6AXC3 A0A8I6AXC3_RAT		
1	5450	Rat	0.909	tr A0A8I5Y907 A0A8I5Y907_RAT		
1	5450	Rat	0.867	sp P51111 HD_RAT		
1	5450	Rat	0.779	tr A0A8I6A113 A0A8I6A113_RAT		
1	5450	Human	1,000	sp P42858 HD_HUMAN	100%	1,000
2	3498	Rat	1,000	tr A0A8I6A7M0 A0A8I6A7M0_RAT	100%	1,000
2	3498	Rat	0.370	tr A6IQC3 A6IQC3_RAT		
2	3498	Rat	0.350	tr A0A8I6A7N8 A0A8I6A7N8_RAT		
2	3498	Human	1,000	sp P21675 TAF1_HUMAN	100%	1,000
3	1211	Rat	1,000	tr D3ZWH2 D3ZWH2_RAT	100%	1,000
3	1211	Rat	1,000	tr A6IZQ1 A6IZQ1_RAT	100%	1,000
3	1211	Human	1,000	sp Q8WXH2 JPH3_HUMAN	100%	1,000
4	742	Rat	1,000	tr Q3B7L3 Q3B7L3_RAT	100%	1,000
4	742	Rat	0.923	sp Q5GH61 XK_RAT		
4	742	Human	1,000	sp P51811 XK_HUMAN	100%	1,000
5	513	Rat	1,000	tr Q66HB1 Q66HB1_RAT	95%	0.897
5	513	Rat	1,000	tr F7FBK7 F7FBK7_RAT	94%	0.897
5	513	Human	1,000	sp P20226 TBP_HUMAN	100%	1,000
6	387	Rat	1,000	tr A6HQE9 A6HQE9_RAT	100%	1,000
6	387	Rat	1,000	sp P13852 PRIO_RAT	100%	1,000
6	387	Human	1,000	sp P04156 PRIO_HUMAN	100%	1,000

#### *Malaltia de Huntington – Peix Zebra*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	4381	Zebrafish	1,000	tr A0A8M2B6V6 A0A8M2B6V6_DANRE	100%	1,000,000
1	4381	Zebrafish	0.988	tr O42269 O42269_DANRE		
1	4381	Zebrafish	0.983	tr A0A8M3B295 A0A8M3B295_DANRE		
1	4381	Zebrafish	0.979	tr A0A8M3AIT6 A0A8M3AIT6_DANRE		
1	4381	Zebrafish	0.974	tr A0A8M3AIU2 A0A8M3AIU2_DANRE		
1	4381	Zebrafish	0.956	tr A0A8M3AU50 A0A8M3AU50_DANRE		
1	4381	Human	1,000	sp P42858 HD_HUMAN	100%	1,000,000
2	2764	Zebrafish	1,000	tr A0A8M9QDY7 A0A8M9QDY7_DANRE	100%	1,000,000
2	2764	Zebrafish	0.995	tr A0A8M9Q3C1 A0A8M9Q3C1_DANRE		
2	2764	Zebrafish	0.964	tr A0A8M9Q8M9 A0A8M9Q8M9_DANRE		
2	2764	Zebrafish	0.920	tr Q1LYC2 Q1LYC2_DANRE		
2	2764	Zebrafish	0.915	tr A0A8M2BCS6 A0A8M2BCS6_DANRE		
2	2764	Human	1,000	sp P21675 TAF1_HUMAN	100%	1,000
3	759	Zebrafish	1,000	tr A0A0R4IT36 A0A0R4IT36_DANRE	99%	1,000
3	759	Human	1,000	sp Q8WXH2 JPH3_HUMAN	100%	1,000
4	569	Zebrafish	1,000	tr Q5GH44 Q5GH44_DANRE	99%	0.984
4	569	Zebrafish	0.495	tr A0A8M9QLM0 A0A8M9QLM0_DANRE		
4	569	Human	1,000	sp P51811 XK_HUMAN	100%	1,000
5	460	Zebrafish	1,000	sp Q7SXL3 TBP_DANRE	99%	0.991
5	460	Human	1,000	sp P20226 TBP_HUMAN	100%	1,000

### Malaltia de Huntington – Ratolí

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	5447	Mouse	1,000	tr G3X9H5 G3X9H5_MOUSE	100%	1,000
1	5447	Mouse	0.989	sp P42859 HD_MOUSE		
1	5447	Human	1,000	sp P42858 HD_HUMAN	100%	1,000
2	3488	Mouse	1,000	sp Q80UV9 TAF1_MOUSE	100%	1,000
2	3488	Mouse	0.977	tr B1Q2W7 B1Q2W7_MOUSE		
2	3488	Mouse	0.488	tr A0AAQ4VMV6 A0AAQ4VMV6_MOUSE		
2	3488	Human	1,000	sp P21675 TAF1_HUMAN	100%	1,000
3	1197	Mouse	1,000	sp Q9E177 JPH3_MOUSE	100%	1,000
3	1197	Mouse	0.993	tr Q8BR43 Q8BR43_MOUSE		
3	1197	Human	1,000	sp Q8WXH2 JPH3_HUMAN	100%	1,000
4	751	Mouse	1,000	tr Q059K7 Q059K7_MOUSE	100%	1,000
4	751	Mouse	1,000	sp Q9QXY7 XK_MOUSE	100%	1,000
4	751	Human	1,000	sp P51811 XK_HUMAN	100%	1,000
5	513	Mouse	1,000	sp P29037 TBP_MOUSE	99%	0.986
5	513	Mouse	1,000	tr Q6R165 Q6R165_MOUSE	99%	0.986
5	513	Human	1,000	sp P20226 TBP_HUMAN	100%	1,000
6	358	Mouse	1,000	tr Q4FJQ7 Q4FJQ7_MOUSE	100%	1,000
6	358	Mouse	1,000	sp P04925 PRIO_MOUSE	100%	1,000
6	358	Mouse	1,000	tr Q3UG89 Q3UG89_MOUSE	100%	1,000
6	358	Mouse	1,000	tr Q3TZI7 Q3TZI7_MOUSE	100%	1,000
6	358	Mouse	0.960	tr Q3UF68 Q3UF68_MOUSE		
6	358	Mouse	0.960	tr Q3UBH0 Q3UBH0_MOUSE		
6	358	Mouse	0.940	tr Q9QYT9 Q9QYT9_MOUSE		
6	358	Human	1,000	sp P04156 PRIO_HUMAN	100%	1,000

### Malaltia de Huntington – C. elegans

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	957	Human	1,000	sp P21675 TAF1_HUMAN	100%	1,000
1	957	<i>C.Elegans</i>	1,000	tr A0A8S4Q965 A0A8S4Q965_CAEEL	100%	1,000
1	957	<i>C.Elegans</i>	0.997	sp G5EGM3 TAF1_CAEEL		
2	294	Human	1,000	sp P20226 TBP_HUMAN	100%	1,000
2	294	<i>C.Elegans</i>	1,000	sp P32085 TBP_CAEEL	100%	1,000
3	271	Human	1,000	sp Q8WXH2 JPH3_HUMAN	100%	1,000
3	271	<i>C.Elegans</i>	1,000	tr Q22667 Q22667_CAEEL	100%	1,000
3	271	<i>C.Elegans</i>	0.995	tr Q966S4 Q966S4_CAEEL		
4	55	Human	1,000	sp P51811 XK_HUMAN	100%	1,000
4	55	<i>C.Elegans</i>	1,000	sp O17386 CED8_CAEEL	100%	1,000

### Malaltia de Huntington – D. melanogaster

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	1612	<i>D.Melanogaster</i>	1,000	tr E1JJ72 E1JJ72_DROME	100%	1,000
1	1612	<i>D.Melanogaster</i>	1,000	sp P51123 TAF1_DROME	100%	1,000
1	1612	<i>D.Melanogaster</i>	1,000	tr A0A0B4LGU9 A0A0B4LGU9_DROME	100%	1,000
1	1612	<i>D.Melanogaster</i>	0.991	tr A0A0B4K602 A0A0B4K602_DROME		
1	1612	<i>D.Melanogaster</i>	0.890	tr Q29R44 Q29R44_DROME		
1	1612	Human	1,000	sp P21675 TAF1_HUMAN	100%	1,000
2	318	<i>D.Melanogaster</i>	1,000	sp P20227 TBP_DROME	100%	0.992
2	318	Human	1,000	sp P20226 TBP_HUMAN	100%	1,000

*Acatalasemia – Rata*

Rank	Bootstrap	Espèce	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	1004	Human	1,000	sp P04040 CATA_HUMAN	100%	1,000
1	1004	Rat	1,000	sp P04762 CATA_RAT	100%	1,000
1	1004	Rat	1,000	tr A6HNR6 A6HNR6_RAT	100%	1,000

*Acatalasemia – Peix Zebra*

Rank	Bootstrap	Espèce	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	875	Human	1,000	sp P04040 CATA_HUMAN	100%	1,000
1	875	ZebraFish	1,000	sp Q9PT92 CATA_DANRE	100%	1,000

*Acatalasemia – Ratolí*

Rank	Bootstrap	Espèce	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	1008	Human	1,000	sp P04040 CATA_HUMAN	100%	1,000
1	1008	Mouse	1,000	tr Q542K4 Q542K4_MOUSE	100%	1,000
1	1008	Mouse	0.989	sp P24270 CATA_MOUSE		-
1	1008	Mouse	0.979	tr Q91X12 Q91X12_MOUSE		-
1	1008	Mouse	0.979	tr Q3TVZ1 Q3TVZ1_MOUSE		-
1	1008	Mouse	0.968	tr Q3UZE7 Q3UZE7_MOUSE		-
1	1008	Mouse	0.947	tr Q8C6E3 Q8C6E3_MOUSE		-
1	1008	Mouse	0.904	tr Q3UF58 Q3UF58_MOUSE		-

*Acatalasemia – C. elegans*

Rank	Bootstrap	Espèce	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	677	Human	1,000	sp P04040 CATA_HUMAN	100%	1,000
1	677	C.Elegans	1,000	sp Q27487 CATA1_CAEEL	100%	1,000
1	677	C.Elegans	0,503	sp O61235 CATA2_CAEEL		-
1	677	C.Elegans	0,411	tr Q8MYL7 Q8MYL7_CAEEL		-
1	677	C.Elegans	0,406	tr U4PE16 U4PE16_CAEEL		-

*Acatalasemia – D. melanogaster*

Rank	Bootstrap	Espèce	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	701	<i>D. Melanogaster</i>	1,000	sp P17336 CATA_DROME	100%	1,000
1	701	<i>D. Melanogaster</i>	0.234	tr Q9VLI6 Q9VLI6_DROME		-
1	701	<i>D. Melanogaster</i>	0.231	tr A0AMY5 A0AMY5_DROME		-
1	701	<i>D. Melanogaster</i>	0.228	tr C0MII7 C0MII7_DROME		-
1	701	<i>D. Melanogaster</i>	0.214	tr C0MII2 C0MII2_DROME		-
1	701	Human	1,000	sp P04040 CATA_HUMAN	100%	1,000

### HBL – Rata

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	5880	Rat	1,000	tr F1M6Z1 F1M6Z1_RAT	100%	1,000
1	5880	Rat	0.980	tr A0A816AJ20 A0A816AJ20_RAT		
1	5880	Rat	0.955	sp Q7TMA5 APOB_RAT		
1	5880	Rat	0.938	tr A0A8L2QJJ3 A0A8L2QJJ3_RAT		
1	5880	Rat	0.455	tr A6HAL1 A6HAL1_RAT		
1	5880	Human	1,000	sp P04114 APOB_HUMAN	100%	1,000
2	694	Rat	1,000	tr F7FHP0 F7FHP0_RAT	100%	1,000
2	694	Rat	1,000	tr A6JRL4 A6JRL4_RAT	100%	1,000
2	694	Human	1,000	sp Q9Y5C1 ANGL3_HUMAN	100%	1,000
3	400	Rat	1,000	tr A6HE76 A6HE76_RAT	98%	0.761
3	400	Rat	1,000	sp Q5HZY2 SAR1B_RAT	98%	0.761
3	400	Human	1,000	sp Q9Y6B6 SAR1B_HUMAN	100%	1,000

### HBL – Peix Zebra

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	2383	Zebrafish	1,000	tr A0A8N7T7N7 A0A8N7T7N7_DANRE	100%	1,000
1	2383	Human	1,000	sp P04114 APOB_HUMAN	100%	1,000
2	416	Zebrafish	1,000	tr A0A8N7UZN6 A0A8N7UZN6_DANRE	99%	0.986
2	416	Human	1,000	sp Q9Y5C1 ANGL3_HUMAN	100%	1,000
3	376	Zebrafish	1,000	tr Q6PBX2 Q6PBX2_DANRE	72%	0.589
3	376	Human	1,000	sp Q9Y6B6 SAR1B_HUMAN	100%	1,000

### HBL – Ratolí

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	5944	Mouse	1,000	sp E9Q414 APOB_MOUSE	100%	1,000
1	5944	Mouse	0.956	tr E9Q1Y3 E9Q1Y3_MOUSE		
1	5944	Human	1,000	sp P04114 APOB_HUMAN	100%	1,000
2	594	Mouse	1,000	sp Q9R182 ANGL3_MOUSE	100%	1,000
2	594	Mouse	1,000	tr Q3UEF5 Q3UEF5_MOUSE	100%	1,000
2	594	Human	1,000	sp Q9Y5C1 ANGL3_HUMAN	100%	1,000
3	400	Mouse	1,000	tr Q0VGU0 Q0VGU0_MOUSE	59%	0.518
3	400	Mouse	1,000	sp Q9CQC9 SAR1B_MOUSE	56%	0.518
3	400	Human	1,000	sp Q9Y6B6 SAR1B_HUMAN	100%	1,000

### HBL – C. elegans

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	272	Human	1,000	sp Q9Y6B6 SAR1B_HUMAN	100%	1,000
1	272	C.Elegans	1,000	sp Q23445 SAR1_CAEEL	100%	1,000

### HBL – D. melanogaster

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	291	D.Melanogaster	1,000	tr Q9VD29 Q9VD29_DROME	100%	1,000
1	291	D.Melanogaster	0.564	tr Q9NFN8 Q9NFN8_DROME		
1	291	D.Melanogaster	0.264	tr Q95SY7 Q95SY7_DROME		
1	291	Human	1,000	sp Q9Y6B6 SAR1B_HUMAN	100%	1,000
2	144	D.Melanogaster	1,000	tr Q95SM8 Q95SM8_DROME	81%	0.806
2	144	D.Melanogaster	0.282	tr Q4V4U2 Q4V4U2_DROME		
2	144	D.Melanogaster	0.146	tr Q9VME7 Q9VME7_DROME		
2	144	D.Melanogaster	0.142	tr A1ZAJ8 A1ZAJ8_DROME		
2	144	D.Melanogaster	0.140	tr A1Z7G6 A1Z7G6_DROME		
2	144	D.Melanogaster	0.129	tr A0A0B4K7T3 A0A0B4K7T3_DROME		
2	144	D.Melanogaster	0.058	tr M9PHH5 M9PHH5_DROME		
2	144	D.Melanogaster	0.058	tr Q9W302 Q9W302_DROME		
2	144	D.Melanogaster	0.054	tr Q9VX04 Q9VX04_DROME		
2	144	D.Melanogaster	0.052	tr Q8MR72 Q8MR72_DROME		
2	144	Human	1,000	sp Q9Y5C1 ANGL3_HUMAN	100%	1,000

*Síndrome Bardet Biedl – Rata*

Rank	Bootstrap	Espèce	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	3920	Human	1,000	sp Q15078 CE290_HUMAN	100%	1,000
1	3920	Rat	1,000	tr A0A0G2K715 A0A0G2K715_RAT	100%	1,000
1	3920	Rat	0.958	tr A0A8I5ZYJ8 A0A8I5ZYJ8_RAT		
1	3920	Rat	0.782	tr A0A0G2K929 A0A0G2K929_RAT		
2	3402	Human	1,000	sp Q9UG01 IF172_HUMAN	100%	1,000
2	3402	Rat	1,000	tr A0A096MJL0 A0A096MJL0_RAT	100%	1,000
2	3402	Rat	0.933	sp Q9JKU3 IF172_RAT		
2	3402	Rat	0.933	tr A0A0G2JVF8 A0A0G2JVF8_RAT		
2	3402	Rat	0.567	tr A0A096MK45 A0A096MK45_RAT		
3	1930	Human	1,000	sp Q2M1P5 KIF7_HUMAN	100%	1,000
3	1930	Rat	1,000	tr D4A9P0 D4A9P0_RAT	100%	1,000
4	1722	Human	1,000	sp Q5HYA8 MKS3_HUMAN	100%	1,000
4	1722	Rat	1,000	sp POC152 MKS3_RAT	100%	1,000
4	1722	Rat	0.986	tr F1M947 F1M947_RAT		
5	1573	Human	1,000	sp Q3SYG4 PTHB1_HUMAN	100%	1,000
5	1573	Rat	1,000	tr A0A1B0GWY0 A0A1B0GWY0_RAT	100%	1,000
5	1573	Rat	0.908	tr A0A1B0GWW7 A0A1B0GWW7_RAT		
5	1573	Rat	0.806	tr F1M285 F1M285_RAT		
5	1573	Rat	0.429	tr A6JP05 A6JP05_RAT		
6	1331	Human	1,000	sp Q8IWZ6 BBS7_HUMAN	100%	1,000
6	1331	Rat	1,000	tr Q66H90 Q66H90_RAT	100%	1,000
6	1331	Rat	0.761	tr A0A8I6G5H8 A0A8I6G5H8_RAT		
6	1331	Rat	0.707	tr A6IHY7 A6IHY7_RAT		
6	1331	Rat	0.120	tr A6IHY6 A6IHY6_RAT		
6	1331	Rat	0.076	tr A0A8I5ZN21 A0A8I5ZN21_RAT		
7	1307	Human	1,000	sp Q9BXC9 BBS2_HUMAN	100%	1,000
7	1307	Rat	1,000	tr A6JY76 A6JY76_RAT	100%	1,000
7	1307	Rat	1,000	sp Q99MH9 BBS2_RAT	100%	1,000
7	1307	Rat	0.185	tr A0A8I6A6C7 A0A8I6A6C7_RAT		
8	1275	Human	1,000	sp Q13049 TRI32_HUMAN	100%	1,000
8	1275	Rat	1,000	tr A0A8I5ZVV9 A0A8I5ZVV9_RAT	100%	1,000
8	1275	Rat	1,000	tr Q66H79 Q66H79_RAT	100%	1,000
9	1145	Human	1,000	sp Q95876 FRITZ_HUMAN	100%	1,000
9	1145	Rat	1,000	tr A0A0G2JXB0 A0A0G2JXB0_RAT	100%	1,000
9	1145	Rat	0.953	tr A0A8I5Y2A3 A0A8I5Y2A3_RAT		
9	1145	Rat	0.894	sp B1WC10 FRITZ_RAT		
9	1145	Rat	0.798	tr A0A8I6ABY1 A0A8I6ABY1_RAT		
9	1145	Rat	0.639	tr A6JQ43 A6JQ43_RAT		
9	1145	Rat	0.458	tr A0A8I5Y4G7 A0A8I5Y4G7_RAT		
10	1077	Human	1,000	sp Q8NFI9 BBS1_HUMAN	100%	1,000
10	1077	Rat	1,000	tr D4A4U2 D4A4U2_RAT	99%	0.998
10	1077	Rat	1,000	tr A6HZ05 A6HZ05_RAT	99%	0.998

*Síndrome Bardet Biedl – Peix Zebra*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	2730	Human	1,000	sp Q9UG01 IF172_HUMAN	100%	1,000
1	2730	ZebraFish	1,000	sp Q5RHH4 IF172_DANRE	100%	1,000
1	2730	ZebraFish	0.986	tr B3DK94 B3DK94_DANRE		
2	2460	Human	1,000	sp O15078 CE290_HUMAN	100%	1,000
2	2460	ZebraFish	1,000	sp P85001 CE290_DANRE	100%	1,000
2	2460	ZebraFish	0.962	tr F8W5U5 F8W5U5_DANRE		
2	2460	ZebraFish	0.959	tr A0A8M3AP13 A0A8M3AP13_DANRE		
2	2460	ZebraFish	0.896	tr A0A8M3AWH7 A0A8M3AWH7_DANRE		
2	2460	ZebraFish	0.892	tr A0A8M3AZL6 A0A8M3AZL6_DANRE		
2	2460	ZebraFish	0.849	tr A0A8M2BA16 A0A8M2BA16_DANRE		
3	1176	Human	1,000	sp Q5HYA8 MKS3_HUMAN	100%	1,000
3	1176	ZebraFish	1,000	tr E9QB24 E9QB24_DANRE	100%	1,000
3	1176	ZebraFish	0.981	tr A0A8M3AJ63 A0A8M3AJ63_DANRE		
4	1123	Human	1,000	sp Q8IWZ6 BBS7_HUMAN	100%	1,000
4	1123	ZebraFish	1,000	tr Q08C18 Q08C18_DANRE	100%	1,000
5	1120	Human	1,000	sp Q3SYG4 PTHB1_HUMAN	100%	1,000
5	1120	ZebraFish	1,000	tr A0A8M3AT49 A0A8M3AT49_DANRE	100%	1,000
5	1120	ZebraFish	0.977	tr A0A8M9PJW4 A0A8M9PJW4_DANRE		
6	1113	Human	1,000	sp Q2M1P5 KIF7_HUMAN	100%	1,000
6	1113	ZebraFish	1,000	tr F1QH25 F1QH25_DANRE	100%	1,000
6	1113	ZebraFish	0.988	sp Q58G59 KIF7_DANRE		
7	1065	Human	1,000	sp Q9BXC9 BBS2_HUMAN	100%	1,000
7	1065	ZebraFish	1,000	sp Q98SP7 BBS2_DANRE	100%	1,000
8	858	Human	1,000	sp Q13049 TRI32_HUMAN	100%	1,000
8	858	ZebraFish	1,000	tr A0A0R4IL45 A0A0R4IL45_DANRE	100%	1,000
8	858	ZebraFish	0.996	tr A8WGA9 A8WGA9_DANRE		
9	800	Human	1,000	sp Q8NFJ9 BBS1_HUMAN	100%	1,000
9	800	ZebraFish	1,000	tr A4QP86 A4QP86_DANRE	100%	1,000
10	702	Human	1,000	sp Q9NXB0 MKS1_HUMAN	100%	1,000
10	702	ZebraFish	1,000	tr A3QK09 A3QK09_DANRE	100%	1,000

*Síndrome Bardet Biedl – Ratolí*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	4024	Human	1,000	sp O15078 CE290_HUMAN	100%	1,000
1	4024	Mouse	1,000	tr E9Q9M0 E9Q9M0_MOUSE	100%	1,000
1	4024	Mouse	0.987	sp Q6A078 CE290_MOUSE		
2	3401	Human	1,000	sp Q9UG01 IF172_HUMAN	100%	1,000
2	3401	Mouse	1,000	sp Q6VH22 IF172_MOUSE	100%	1,000
3	1937	Human	1,000	sp Q2M1P5 KIF7_HUMAN	100%	1,000
3	1937	Mouse	1,000	tr J3QNW9 J3QNW9_MOUSE	100%	1,000
3	1937	Mouse	0.995	sp B7ZNG0 KIF7_MOUSE		
3	1937	Mouse	0.988	tr E9QMU1 E9QMU1_MOUSE		
4	1739	Human	1,000	sp Q5HYA8 MKS3_HUMAN	100%	1,000
4	1739	Mouse	1,000	tr A2AJP5 A2AJP5_MOUSE	100%	1,000
4	1739	Mouse	1,000	sp Q8BR76 MKS3_MOUSE	100%	1,000
5	1576	Human	1,000	sp Q3SYG4 PTHB1_HUMAN	100%	1,000
5	1576	Mouse	1,000	sp Q811G0 PTHB1_MOUSE	100%	1,000
6	1335	Human	1,000	sp Q8IWZ6 BBS7_HUMAN	100%	1,000
6	1335	Mouse	1,000	sp Q8K2G4 BBS7_MOUSE	100%	1,000
6	1335	Mouse	0.967	tr Q3V165 Q3V165_MOUSE		
6	1335	Mouse	0.934	tr B1ATV5 B1ATV5_MOUSE		
6	1335	Mouse	0.890	tr B1ATV6 B1ATV6_MOUSE		
6	1335	Mouse	0.088	tr E9Q0R0 E9Q0R0_MOUSE		
7	1315	Human	1,000	sp Q9BXC9 BBS2_HUMAN	100%	1,000
7	1315	Mouse	1,000	sp Q9CWF6 BBS2_MOUSE	100%	1,000
8	1269	Human	1,000	sp Q13049 TRI32_HUMAN	100%	1,000
8	1269	Mouse	1,000	sp Q8CH72 TRI32_MOUSE	100%	1,000
8	1269	Mouse	1,000	tr Q3TLR3 Q3TLR3_MOUSE	100%	1,000
9	1128	Human	1,000	sp O95876 FRITZ_HUMAN	100%	1,000
9	1128	Mouse	1,000	sp Q8C456 FRITZ_MOUSE	100%	1,000
9	1128	Mouse	1,000	tr B1ATJ7 B1ATJ7_MOUSE	100%	1,000
10	1084	Human	1,000	sp Q8NFJ9 BBS1_HUMAN	100%	1,000
10	1084	Mouse	1,000	sp Q3V3N7 BBS1_MOUSE	100%	1,000

### Síndrome Bardet Biedl – *C. elegans*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	1434	Human	1,000	sp Q9UG01 IF172_HUMAN	100%	1,000
1	1434	<i>C.Elegans</i>	1,000	sp Q22830 OSM1_CAEEL	100%	1,000
2	399	Human	1,000	sp Q81WZ6 BBS7_HUMAN	100%	1,000
2	399	<i>C.Elegans</i>	1,000	sp Q9XW70 BBS7_CAEEL	100%	1,000
3	369	Human	1,000	sp Q9BXC9 BBS2_HUMAN	100%	1,000
3	369	<i>C.Elegans</i>	1,000	sp Q19640 BBS2_CAEEL	100%	1,000
4	348	Human	1,000	sp Q5HYA8 MK33_HUMAN	100%	1,000
4	348	<i>C.Elegans</i>	1,000	tr Q20046 Q20046_CAEEL	100%	1,000
5	324	Human	1,000	sp Q8N3I7 BBS5_HUMAN	100%	1,000
5	324	<i>C.Elegans</i>	1,000	sp Q21626 BBS5_CAEEL	100%	1,000
6	270	Human	1,000	sp Q8TAM2 TTC8_HUMAN	100%	1,000
6	270	<i>C.Elegans</i>	1,000	sp Q23049 TTC8_CAEEL	100%	1,000
7	262	Human	1,000	sp Q8NFJ9 BBS1_HUMAN	100%	1,000
7	262	<i>C.Elegans</i>	1,000	sp Q9NEZ7 BBS1_CAEEL	100%	1,000
8	220	Human	1,000	sp Q96RK4 BBS4_HUMAN	100%	1,000
8	220	<i>C.Elegans</i>	1,000	sp Q5CZ52 BBS4_CAEEL	100%	1,000
9	165	Human	1,000	sp Q9H0F7 ARL6_HUMAN	100%	1,000
9	165	<i>C.Elegans</i>	1,000	sp Q18510 ARL6_CAEEL	81%	0.739
10	134	Human	1,000	sp Q3SYG4 PTHB1_HUMAN	100%	1,000
10	134	<i>C.Elegans</i>	1,000	sp O01514 PTHB1_CAEEL	100%	1,000

### Síndrome Bardet Biedl – *D. melanogaster*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	1600	<i>D.Melanogaster</i>	1,000	sp Q9W040 OSM1_DROME	100%	1,000
1	1600	Human	1,000	sp Q9UG01 IF172_HUMAN	100%	1,000
2	322	<i>D.Melanogaster</i>	1,000	tr Q9I7P3 Q9I7P3_DROME	100%	1,000
2	322	<i>D.Melanogaster</i>	0.977	tr Q4V3X5 Q4V3X5_DROME		
2	322	<i>D.Melanogaster</i>	0.628	tr Q4V3K1 Q4V3K1_DROME		
2	322	<i>D.Melanogaster</i>	0.626	tr Q4V3Q7 Q4V3Q7_DROME		
2	322	<i>D.Melanogaster</i>	0.599	tr Q4V3F8 Q4V3F8_DROME		
2	322	Human	1,000	sp Q8N3I7 BBS5_HUMAN	100%	1,000
3	310	<i>D.Melanogaster</i>	1,000	sp Q9VTW2 INP5E_DROME	100%	1,000
3	310	Human	1,000	sp Q9NRR6 INP5E_HUMAN	100%	1,000
4	299	<i>D.Melanogaster</i>	1,000	tr B7YZM8 B7YZM8_DROME	100%	1,000
4	299	Human	1,000	sp Q3SYG4 PTHB1_HUMAN	100%	1,000
5	275	<i>D.Melanogaster</i>	1,000	tr Q9VS19 Q9VS19_DROME	100%	1,000
5	275	<i>D.Melanogaster</i>	1,000	tr F3YDL9 F3YDL9_DROME	100%	1,000
5	275	Human	1,000	sp Q8NFJ9 BBS1_HUMAN	100%	1,000
6	252	<i>D.Melanogaster</i>	1,000	tr M9PE00 M9PE00_DROME	100%	1,000
6	252	<i>D.Melanogaster</i>	1,000	sp Q9VQ36 FRITZ_DROME	100%	1,000
6	252	<i>D.Melanogaster</i>	1,000	tr M9PBR4 M9PBR4_DROME	100%	1,000
6	252	Human	1,000	sp Q95876 FRITZ_HUMAN	100%	1,000
7	241	<i>D.Melanogaster</i>	1,000	tr A0A0B4KGJ7 A0A0B4KGJ7_DROME	100%	1,000
7	241	<i>D.Melanogaster</i>	1,000	tr Q9VDK3 Q9VDK3_DROME	100%	1,000
7	241	Human	1,000	sp Q5HYA8 MK33_HUMAN	100%	1,000
8	238	<i>D.Melanogaster</i>	1,000	sp A1Z8E9 BBS4_DROME	100%	1,000
8	238	<i>D.Melanogaster</i>	0.898	tr A0A0B4LF63 A0A0B4LF63_DROME		
8	238	Human	1,000	sp Q96RK4 BBS4_HUMAN	100%	1,000
9	213	<i>D.Melanogaster</i>	1,000	tr A0A0B4LFW0 A0A0B4LFW0_DROME	99%	0.968
9	213	<i>D.Melanogaster</i>	0.963	tr A1ZBK9 A1ZBK9_DROME		
9	213	<i>D.Melanogaster</i>	0.873	tr Q4V546 Q4V546_DROME		
9	213	Human	1,000	sp Q9H0F7 ARL6_HUMAN	100%	1,000
10	159	<i>D.Melanogaster</i>	1,000	tr Q9VPP9 Q9VPP9_DROME	100%	1,000
10	159	Human	1,000	sp Q8TAM2 TTC8_HUMAN	100%	1,000

### Policitèmia Vera – Rata

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	2668	Human	1,000	sp P12259 FA5_HUMAN	100%	1,000
1	2668	Rat	1,000	tr A0A0G2K3W2 A0A0G2K3W2_RAT	100%	1,000
1	2668	Rat	0.124	tr Q7TPK2 Q7TPK2_RAT		
1	2668	Rat	0.073	tr A6IDD7 A6IDD7_RAT		
2	2214	Human	1,000	sp O60674 JAK2_HUMAN	100%	1,000
2	2214	Rat	1,000	tr A6I0V5 A6I0V5_RAT	100%	1,000
2	2214	Rat	1,000	sp Q62689 JAK2_RAT	100%	1,000
3	2089	Human	1,000	sp Q6N021 TET2_HUMAN	100%	1,000
3	2089	Rat	1,000	tr D4AC33 D4AC33_RAT	100%	1,000

### *Policitèmia Vera – Peix Zebra*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	1616	Human	1,000	sp O60674 JAK2_HUMAN	100%	1,000
1	1616	ZebraFish	1,000	tr C0RW38 C0RW38_DANRE	96%	1,000
1	1616	ZebraFish	0.460	tr A0A8M6Z6T4 A0A8M6Z6T4_DANRE		
2	1281	Human	1,000	sp P12259 FA5_HUMAN	100%	1,000
2	1281	ZebraFish	1,000	tr B7ZVT7 B7ZVT7_DANRE	99%	1,000
3	979	Human	1,000	sp Q6N021 TET2_HUMAN	100%	1,000
3	979	ZebraFish	1,000	tr E7F958 E7F958_DANRE	99%	1,000

### *Policitèmia Vera – Ratolí*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	2743	Human	1,000	sp P12259 FA5_HUMAN	100%	1,000
1	2743	Mouse	1,000	sp O88783 FA5_MOUSE	100%	1,000
2	2204	Human	1,000	sp O60674 JAK2_HUMAN	100%	1,000
2	2204	Mouse	1,000	sp Q62120 JAK2_MOUSE	100%	1,000
3	1969	Human	1,000	sp Q6N021 TET2_HUMAN	100%	1,000
3	1969	Mouse	1,000	tr D0VLQ5 D0VLQ5_MOUSE	100%	1,000
3	1969	Mouse	1,000	sp Q4JK59 TET2_MOUSE	100%	1,000
3	1969	Mouse	0.996	tr A0A0G2JF55 A0A0G2JF55_MOUSE		

### *Policitèmia Vera – C. elegans*

No hi ha hagut resultats.

### *Policitèmia Vera – D. melanogaster*

No hi ha hagut resultats.

### *Síndrome Floating Harbor – Rata*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	4118	Human	1,000	sp   Q6ZRS2   SRCAP_HUMAN	100%	1,000
1	4118	Rat	1,000	tr   A0A8I5ZYJ1   A0A8I5ZYJ1_RAT	100%	1,000
1	4118	Rat	0.872	tr   M0R750   M0R750_RAT		
1	4118	Rat	0.755	tr   A0A8I6AE82   A0A8I6AE82_RAT		

### *Síndrome Floating Harbor – Peix Zebra*

Rank	Bootstrap	Espècie	Seed-Score Human-Model	Uniprot Gene ID and Name	Correlation	Seed-Score Model-Human
1	1713	Human	1,000	sp   Q6ZRS2   SRCAP_HUMAN	100%	1,000
1	1713	ZebraFish	1,000	tr   A0A8M2B358   A0A8M2B358_DANRE	100%	1,000
1	1713	ZebraFish	0.994	tr   E7FAZ7   E7FAZ7_DANRE		
1	1713	ZebraFish	0.975	tr   A0A8M2B2C4   A0A8M2B2C4_DANRE		

### *Síndrome Floating Harbor – Ratolí*

No hi ha hagut resultats.

### *Síndrome Floating Harbor – C. elegans*

No hi ha hagut resultats.

### *Síndrome Floating Harbor – D. melanogaster*

No hi ha hagut resultats.