

Data Summary

News and Tweets

- 25 April 2019 AmoebaDB 43 Released
- 25 April 2019 CryptoDB 43 Released
- 25 April 2019 FungiDB 43 Released

The EuPathDB **Bioinformatics Resource Center** provides a portal for accessing genomic-scale datasets associated with the diverse eukaryotic microbes (mouse-over the following logos for information on component websites):



Bases de datos específicas: EupathDB

Especialización en Micología y Parasitología

Facultad de Ciencias Bioquímicas y Farmacéuticas

Universidad Nacional de Rosario

2013

Dra. Victoria Alonso

alonso@ibr-conicet.gov.ar

EuPathDB



- EuPathDB (<http://EuPathDB.org>) es una base de datos integral que engloba los genomas de los organismos parásitos patógenos: *Cryptosporidium*, *Giardia*, *Leishmania*, *Neospora*, *Plasmodium*, *Toxoplasma*, *Trichomonas* o *Trypanosoma*
- Base de datos desarrollada por Bioinformatics Resource Center (BRC), financiada por el National Institutes of Health (NIH).

- Quick access to ID and text search options, login, contact, twitter, etc.

- Main Header Tab Bar: mouse-over 'New Search' to initiate searches; click 'My Strategies' to enter your workspace

- Portal to EuPathDB databases by clicking on icons

- Initiate searches from center panels. Over 100 search types available.

- Identify Genes by: look for Genes based on a variety of datasets, including whole genome sequence, coding vs non-coding genes, transcript evidence (microarray, EST), exon count, etc.

- Identify Other Data Types: Look for ESTs, SNPs or DNA motifs;

- Tools: Access tools like Blast and PubMed from any EuPathDB home page

Taxon specific databases provide access to the latest available genome-scale datasets. Built with the same web-architecture, search types and functions are the same across all databases.

(<http://www.genedb.org/>)

EuPathDB Version 2.15
Eukaryotic Pathogen Database Resources 31 Aug 12

A EuPathDB Project

Gene ID: Gene Text Search:

About EuPathDB | Help | Login | Register | Contact Us |  

Home | New Search | My Strategies | My Basket (0) | Tools | Data Summary | Downloads | Community

Data Summary

News

- 31 August 2012
MicrosporidiaDB 3.0 released
- 31 August 2012
PiroplasmaDB 2.0 Released
- 31 August 2012
PlasmoDB 9.1 Released

All EuPathDB News >>>

Community Resources

expand for 9 new items

Education and Tutorials

expand for 7 new items

Other Information

expand for 9 new items

EuPathDB Bioinformatics Resource Center for Biodefense and Emerging/Re-emerging Infectious Diseases is a portal for accessing genomic-scale datasets associated with the eukaryotic pathogens:
(mouse over the logos: Babesia, Crithidia, Cryptosporidium, Edhazardia, Eimeria, Encephalitozoon, Endotrypanum, Entamoeba, Enterocytozoon, Giardia, Gregarina, Hamiltosporidium, Leishmania, Nematocida, Neospora, Nosema, Plasmodium, Theileria, Toxoplasma, Trichomonas, Trypanosoma, Vivraia).



AmoebaDB



CryptoDB



GiardiaDB



MicrosporidiaDB



NEW



PiroplasmaDB



PlasmoDB



ToxoDB



TrichDB



TriTrypDB

Identify Genes by:

- Expand All | Collapse All
- Text, IDs, Organism
- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- Evolution
- Population Biology

Identify Other Data Types:

- Expand All | Collapse All
- Isolates
- Genomic Sequences
- Genomic Segments (DNA Motif)
- SNPs
- ESTs
- ORFs
- SAGE Tags

Tools:

- BLAST**
Identify Sequence Similarities
- Sequence Retrieval**
Retrieve Specific Sequences using IDs and coordinates
- PubMed and Entrez**
View the Latest *Eukaryotic Pathogens* Pubmed and Entrez Results
- ApiCyc**
Explore Automatically Defined Metabolic Pathways
- Searches via Web Services**
Learn about web service access to our data

→Pulsa en el icono “TrytrypDB”

Data Summary

News

18 November 2010 TriTrypDB 2.5 Released
20 October 2010 Driving Biological Projects awarded
21 September 2010 TriTrypDB 2.4 Released
15 July 2010 TriTrypDB 2.3 released

[All TriTrypDB News](#)

Community Resources
expand for 7 new items

Web Tutorials

Information and Help

Identify Genes by:

[Expand All](#) | [Collapse All](#)

- Text, IDs, Species
- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- Evolution

Identify Other Data Types:

[Expand All](#) | [Collapse All](#)

- Genomic Sequences
- ESTs
- Transcript Assemblies
- ORFs
- SAGE Tags

Tools:

BLAST

Identify Sequence Similarities

Sequence Retrieval

Retrieve Specific Sequences using IDs and coordinates

PubMed and Entrez

View the Latest *Leishmania*, *Trypanosoma* Pubmed and Entrez Results

Genome Browser

View Sequences and Features in the genome browser

Searches via Web Services

Learn about web service access to our data



-Pagina web integral de "TrytrypDB" que incluye los genomas de los kinetoplastidos secuenciados, herramientas para su uso, así como distintos experimentos desarrollados en paralelo

Logotipo presente en todas las páginas Web, incluyendo las ventanas de búsqueda rápida y una barra de herramientas (gris

Búsqueda rápida mediante el identificador del gen o nombre genérico o específico de gen/genes. Soporta plazo para búsquedas de frases exactas y comodines (*).

Version 2.1
12 Mar 10

Gene ID: Gene Text Search:

About TriTrypDB | Help | Contact Us | Omar Harb's Profile | Logout

Home | New Search | My Strategies | My Basket (1) | Tools | Data Summary | Downloads | Community

Data Summary

News

- 15 March 2010 TriTrypDB 2.1 released
- 22 December 2009 TriTrypDB 2.0 released
- 14 November 2009 TriTrypDB 1.3 released
- 13 November 2009 Working with Parasite Database Resources Workshop (April 17-21, 2010)

All TriTrypDB News

Community Resources
expand for 1 new items

Web Tutorials

Information and Help
expand for 6 new items

Identify Genes by:
Expand All | Collapse All

- Text, IDs, Species
- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- Evolution

Identify Other Data Types:
Expand All | Collapse All

- Genomic Sequences
- ESTs
- Transcript Assemblies
- ORFs

Tools:

BLAST
Identify Sequence Similarities

Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates

PubMed and Entrez
View the Latest *Leishmania*, *Trypanosoma* Pubmed and Entrez Results

Genome Browser
View Sequences and Features in the GBrowse genome browser (GMOD)

Searches via Web Services
Learn about web service access to our data

TriTrypDB 2.1 March 12, 2010
©2010 The EuPathDB Project Team

EuPathDB

Please Contact Us with any questions or comments

Alertas y noticias especiales

Búsquedas y herramientas

Búsquedas de genes, búsquedas de distintos tipos de secuencias génicas, como ESTs, ORFs, secuencias genómicas, etc.

Información y ayuda

BLAST, PubMed bibliografía, genome browser, más tipos de búsqueda

FUNCIONES BÁSICAS

TriTrypDB Kinetoplastid Genomics Resource

PathDB Project in collaboration with GeneDB

Gene Text Search:

Home | New Search | My Strategies | My Basket (0) | Tools | Data Summary | Downloads | Community | My Favorites

My Strategies: New | **Opened (1)** | All (8) | Basket | Examples | Help

(Genes) **Text (product name, notes, etc.)(2)***

Text 3417 Genes Step 1 Add Step

Grupos ortólogos: **kinasas**

Listado de genes kinasas

Búsqueda de características individuales

Filter results by species (results removed by the filter will not be combined into the next step.)

All Results	Ortholog Groups	Leishmania				Trypanosoma brucei		Trypanosoma congolense	Trypanosoma emmeraldoense				Trypanosoma vivax
		braziliensis	infantum	major	mexicana	TREU927	gambiense		esmeraldo	non-esmeraldo	unassigned		
3417	642	325	337	356	322	329	303	330	654	392	49	13	291

Text (product name, notes, etc.)(2) - step 1 - 3417 Genes

Add 3417 Genes to Basket | Download 3417 Genes

First 1 2 3 4 5 Next Last Advanced Paging

Select Columns Reset Columns

Gene	Organism	Genomic Location	Product Description	Found in	Score
Tc00.1047053510077.30	<i>T. cruzi</i> CL Brener Esmeraldo-like	TcChr1-S: 47,750 - 49,093 (+)	protein kinase, putative, NIMA/Nek Serine/threonine-protein kinase family, putative	InterPro, GoTerms, Notes, Product	200
Tc00.1047053504181.40	<i>T. cruzi</i> CL Brener Esmeraldo-like	TcChr10-S: 224,812 - 225,717 (+)	cell division protein kinase 2, cdc2-related protein kinase 1	InterPro, Product, GoTerms, Notes	200
Tc00.1047053511133.20	<i>T. cruzi</i> CL Brener	TcChr10-S: 298,185 - 299,432	thymidine kinase, putative	InterPro, GoTerms	200

Select Columns close X

Update Columns

select all | clear all

<input checked="" type="checkbox"/> Gene	<input type="checkbox"/> # TM Domains	<input type="checkbox"/> Predicted GO Process
<input type="checkbox"/> GBrowse	<input type="checkbox"/> Molecular Weight	<input type="checkbox"/> Predicted GO Component
<input type="checkbox"/> Data Source	<input type="checkbox"/> Isoelectric Point	<input type="checkbox"/> L.infantum Promastigote Time Series - Graph
<input type="checkbox"/> Genomic Sequence ID	<input type="checkbox"/> EC Numbers	<input type="checkbox"/> T.brucei TbDRBD3 dep/uninduced PAGE- Graph
<input type="checkbox"/> Chromosome	<input type="checkbox"/> Ortholog count	<input type="checkbox"/> T.brucei Dev. Stage PAGE- Graph
<input checked="" type="checkbox"/> Genomic Location	<input type="checkbox"/> Paralog count	<input type="checkbox"/> T.cruzi Dev. Stage PAGE- Graph
<input type="checkbox"/> Gene Strand	<input type="checkbox"/> Ortholog Group	<input type="checkbox"/> T.brucei RNA Seq.- Graph
<input type="checkbox"/> Gene Type	<input type="checkbox"/> SignalP Scores	<input checked="" type="checkbox"/> Organism
<input type="checkbox"/> # Exons	<input type="checkbox"/> SignalP Peptide	<input type="checkbox"/> Is Pseudo
<input type="checkbox"/> Transcript Length	<input checked="" type="checkbox"/> Annotated GO Function	<input type="checkbox"/> Name
<input type="checkbox"/> CDS Length	<input type="checkbox"/> Annotated GO Process	<input type="checkbox"/> Weight
<input checked="" type="checkbox"/> Product Description	<input type="checkbox"/> Annotated GO Component	<input checked="" type="checkbox"/> Found in
<input type="checkbox"/> Protein Length	<input type="checkbox"/> Predicted GO Function	<input checked="" type="checkbox"/> Score

select all | clear all

Update Columns

The Gene Ontology (GO) es un proyecto colaborativo para describir productos génicos en diferentes bases de datos.

- *cellular component*. Localización celular o en su ambiente extracelular.
- *molecular function*, Las funciones elementales de un producto génico a nivel molecular, ej. Ligando o catálisis.
- *biological process*, Grupos de eventos moleculares con un principio o final definido, pertenecientes al funcionamiento de unidades de vida: células, tejidos, organismos, órganos y organismos.

Select Columns close X

Update Columns

select all | clear all

<input checked="" type="checkbox"/> Gene	<input type="checkbox"/> # TM Domains	<input type="checkbox"/> Predicted GO Process
<input type="checkbox"/> GBrowse	<input type="checkbox"/> Molecular Weight	<input type="checkbox"/> Predicted GO Component
<input type="checkbox"/> Data Source	<input type="checkbox"/> Isoelectric Point	<input type="checkbox"/> L.infantum Promastigote Time Series - Graph
<input type="checkbox"/> Genomic Sequence ID	<input type="checkbox"/> EC Numbers	<input type="checkbox"/> T.brucei TbDRBD3 dep/uninduced PAGE- Graph
<input type="checkbox"/> Chromosome	<input type="checkbox"/> Ortholog count	<input type="checkbox"/> T.brucei Dev. Stage PAGE- Graph
<input checked="" type="checkbox"/> Genomic Location	<input type="checkbox"/> Paralog count	<input type="checkbox"/> T.cruzi Dev. Stage PAGE- Graph
<input type="checkbox"/> Gene Strand	<input type="checkbox"/> Ortholog Group	<input type="checkbox"/> T.brucei RNA Seq.- Graph
<input type="checkbox"/> Gene Type	<input type="checkbox"/> SignalP Scores	<input checked="" type="checkbox"/> Organism
<input type="checkbox"/> # Exons	<input type="checkbox"/> SignalP Peptide	<input type="checkbox"/> Is Pseudo
<input type="checkbox"/> Transcript Length	<input checked="" type="checkbox"/> Annotated GO Function	<input type="checkbox"/> Name
<input type="checkbox"/> CDS Length	<input type="checkbox"/> Annotated GO Process	<input type="checkbox"/> Weight
<input checked="" type="checkbox"/> Product Description	<input type="checkbox"/> Annotated GO Component	<input checked="" type="checkbox"/> Found in
<input type="checkbox"/> Protein Length	<input type="checkbox"/> Predicted GO Function	<input checked="" type="checkbox"/> Score

select all | clear all

Update Columns

Expresión:

-Transcriptómica

-Microarrays de diferentes tipos de estadíos parasitarios

-RNAseq. Secuenciación de RNA.

- Espectrometría de Masas

select all | clear all

<input checked="" type="checkbox"/> Gene	<input type="checkbox"/> # TM Domains	<input type="checkbox"/> Predicted GO Process
<input type="checkbox"/> GBrowse	<input type="checkbox"/> Molecular Weight	<input type="checkbox"/> Predicted GO Component
<input type="checkbox"/> Data Source	<input type="checkbox"/> Isoelectric Point	<input type="checkbox"/> L.infantum Promastigote Time Series - Graph
<input type="checkbox"/> Genomic Sequence ID	<input type="checkbox"/> EC Numbers	<input type="checkbox"/> T.brucei TbDRBD3 dep/uninduced PAGE- Graph
<input type="checkbox"/> Chromosome	<input type="checkbox"/> Ortholog count	<input type="checkbox"/> T.brucei Dev. Stage PAGE- Graph
<input checked="" type="checkbox"/> Genomic Location	<input type="checkbox"/> Paralog count	<input type="checkbox"/> T.cruzi Dev. Stage PAGE- Graph
<input type="checkbox"/> Gene Strand	<input type="checkbox"/> Ortholog Group	<input type="checkbox"/> T.brucei RNA Seq.- Graph
<input type="checkbox"/> Gene Type	<input type="checkbox"/> SignalP Scores	<input checked="" type="checkbox"/> Organism
<input type="checkbox"/> # Exons	<input type="checkbox"/> SignalP Peptide	<input type="checkbox"/> Is Pseudo
<input type="checkbox"/> Transcript Length	<input checked="" type="checkbox"/> Annotated GO Function	<input type="checkbox"/> Name
<input type="checkbox"/> CDS Length	<input type="checkbox"/> Annotated GO Process	<input type="checkbox"/> Weight
<input checked="" type="checkbox"/> Product Description	<input type="checkbox"/> Annotated GO Component	<input checked="" type="checkbox"/> Found in
<input type="checkbox"/> Protein Length	<input type="checkbox"/> Predicted GO Function	<input checked="" type="checkbox"/> Score

select all | clear all

Update Columns

CDS:

Coding sequence. Región de nucleótidos que corresponden a una secuencia de aminoácidos en una proteína y que está incluida entre los codones de inicio y de parada. Las partes no expresadas (5' y 3' UTR e intrones) no se incluyen dentro de un CDS

Exones:

Regiones de un [gen](#) que no son separadas durante el proceso de [splicing](#) y, por tanto, se mantienen en el [ARN mensajero](#) maduro

GBrowse :

Navegador interactivo desarrollado por el proyecto Generic Model Organism Database (GMOD) (www.gmod.org) que se pueden personalizar para mostrar determinadas características cromosómicas, así como anotaciones del usuario.

EC numbers:

Enzyme Commission numbers. Constituyen una esquema de clasificación numérica para enzimas basada en las reacciones químicas que catalizan. Los números EC reflejan las reacciones catalizadas por las enzimas

GEN: CARACTERÍSTICAS BÁSICAS

TriTrypDB Version 2.5
Kinetoplastid Genomics Resource
A **EuPathDB** Project in collaboration with **CeneDB**

Gene ID: Gene Text Search:

Home | New Search | My Strategies | My Basket (0) | Tools | Data Summary | Downloads | Community | My Favorites

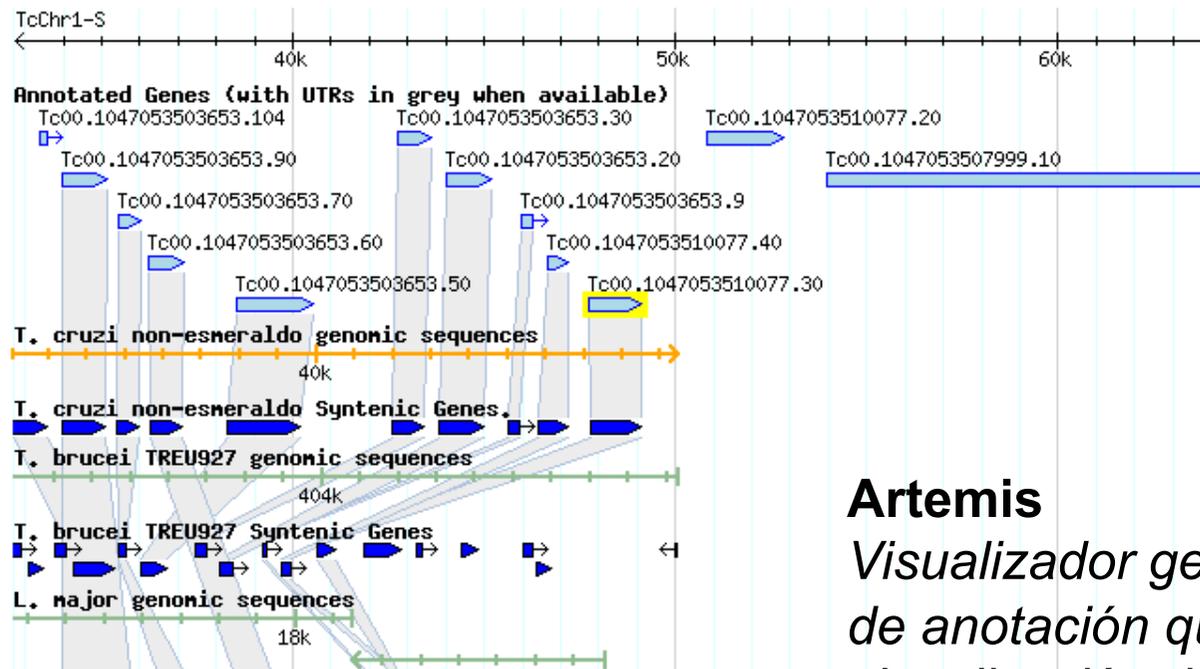
Tc00.1047053510077.30
protein kinase, putative,NIMA/Nek Serine/threonine-protein kinase family, putative
Add to Basket Add to Favorites

Download Show All Hide All

Overview
T. cruzi CL Brener Esmeraldo-like protein coding gene on **TcChr1-S** (chromosome 1) from **47750** to **49093**

Genomic Context [Hide](#)

[\[Data Sources\]](#)

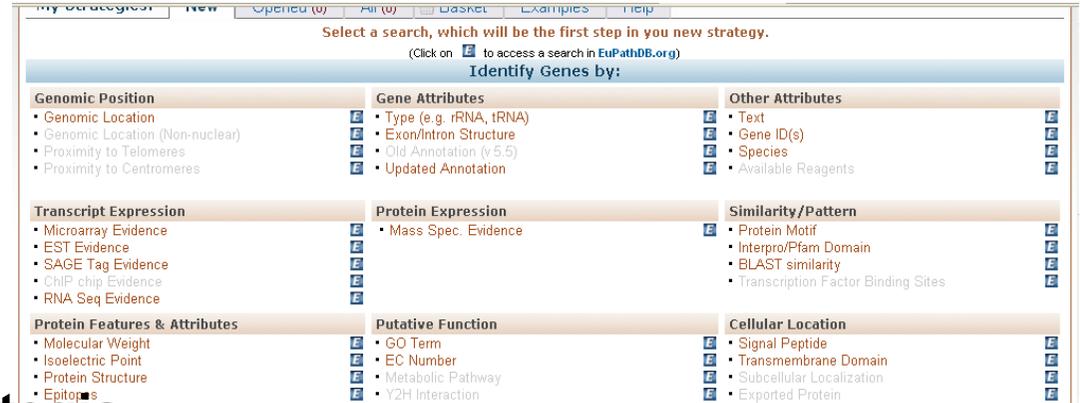


Artemis

Visualizador genómico y herramienta de anotación que permite la visualización de las características de secuencias y los resultados de los análisis de las secuencias, en los seis marcos de lectura

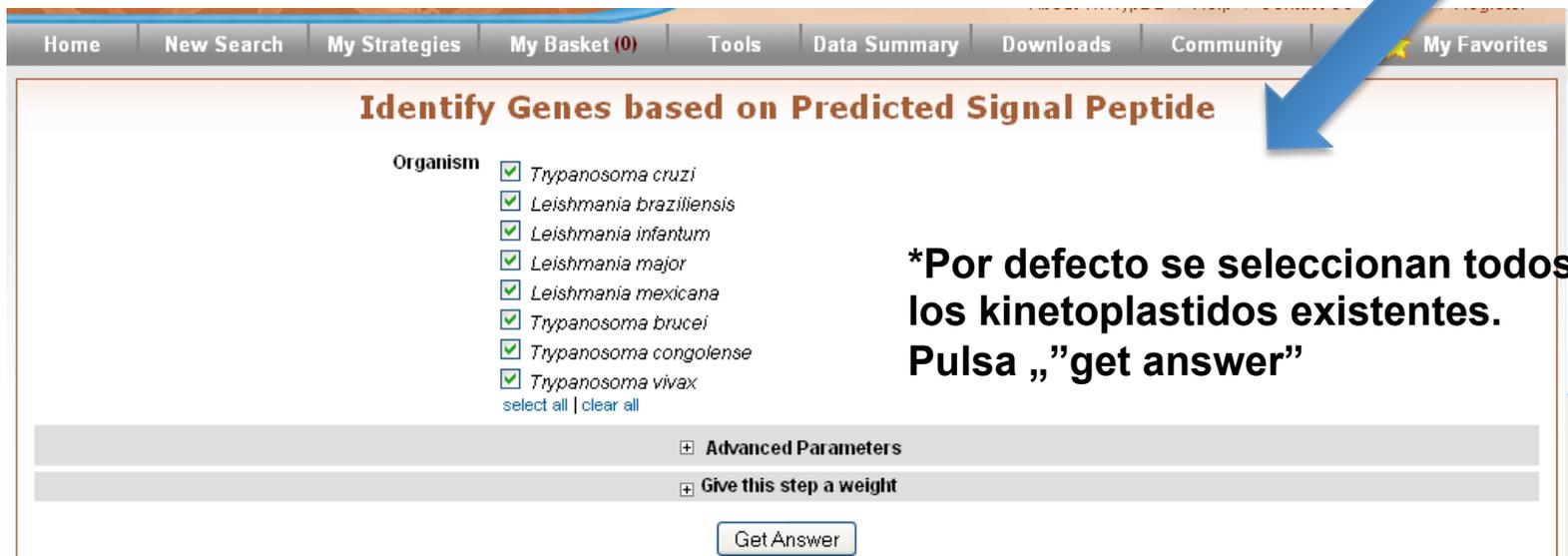
1. Iniciar una estrategia

Pulsa "My strategies" en el menú principal e inicia una nueva estrategia.



Menú de posibilidades de inicio de estrategia

Todas las opciones contenidas en la base de datos. Ej pulsemos Las opciones para "Cellular location" →  Signal peptide



*Por defecto se seleccionan todos los kinetoplastos existentes. Pulsa „"get answer"

TrytrypDB: Estrategias complejas “COMPLEX QUERY”

My Strategies: [New](#) **Opened (1)** [All \(1\)](#) [Basket](#) [Examples](#) [Help](#)

(Genes) Predicted Signal Peptide* [Rename](#) [Copy](#) [Save As](#) [Share](#) [Delete](#)

[+ Signal Pep](#) **Add Step** 22441 Genes
Step 1

Filter results by species (results removed by the filter will not be combined into the next step.)

All Results	Ortholog Groups	Leishmania				Trypanosoma brucei		Trypanosoma congolense	Trypanosoma cruzi			Trypanosoma vivax	
		braziliensis	infantum	major	mexicana	TREU927	gambiense		Distinct genes	esmeraldo	non-esmeraldo		unassigned
22441	6858	1373	1452	1546	1336	2711	1989	2952	5477	2721	2747	888	2726

Add Step:

Construir una estrategia adicionando etapas. Añadiendo etapas se descarga una ventana con múltiples tipos de búsquedas. Una estrategia de búsqueda compleja consiste en la adición de múltiples pasos, añadiendo nuevas características a la búsqueda.

Add Step

Add Step close X

Select a Search or From Basket or Select a Transform or Select a Strategy

Text, IDs, Species
Genomic Position
Gene Attributes
Protein Attributes
Protein Features
Similarity/Pattern
Transcript Expression
Protein Expression
Cellular Location
Putative Function
Evolution

Copy of Gene Basket

Predicted Signal Peptide
Transmembrane Domain Count
Epitope Presence

Orthologs of Genes in Previous Step
Filter Genes by Weight

--Choose a strategy to add--

Continue...

Close

Strategy Builder Guide:

The screenshot shows the TritrypDB Strategy Builder interface. At the top, there are navigation tabs: "My Strategies:", "New", "Opened (1)", "All (141)", "Basket", "Examples", and "Help". Below the tabs is a workflow diagram with six steps:

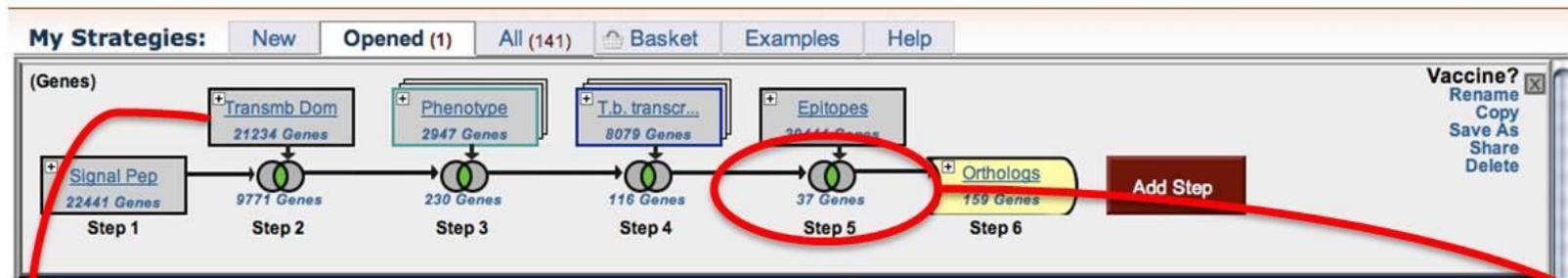
- Step 1: Signal Pep (22441 Genes)
- Step 2: Transmb Dom (21234 Genes) (9771 Genes)
- Step 3: Phenotype (2947 Genes) (230 Genes)
- Step 4: T.b. transcr... (8079 Genes) (116 Genes)
- Step 5: Epitopes (39444 Genes) (37 Genes)
- Step 6: Orthologs (159 Genes)

Annotations in Spanish point to various features:

- "Todas tus previas estrategias" points to the "All (141)" tab.
- "Estrategias abiertas" points to the "Opened (1)" tab.
- "Correr una nueva estrategia" points to the "New" tab.
- "Ver estrategias ejemplo" points to the "Examples" tab.
- "Ayuda" points to the "Help" tab.
- "Renombrar, copiar, salvar o eliminar tu estrategia." points to the context menu for the "Orthologs" step, which includes options: "Rename", "Copy", "Save As", "Share", and "Delete".

Other interface elements include an "Add Step" button and a "Vaccine?" window.

Strategy Builder Guide (cont.):

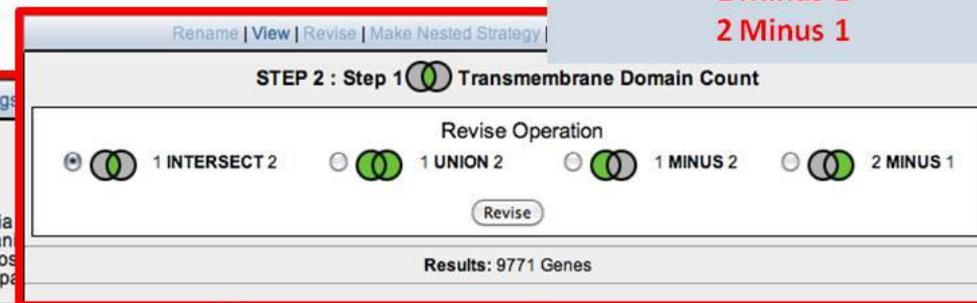
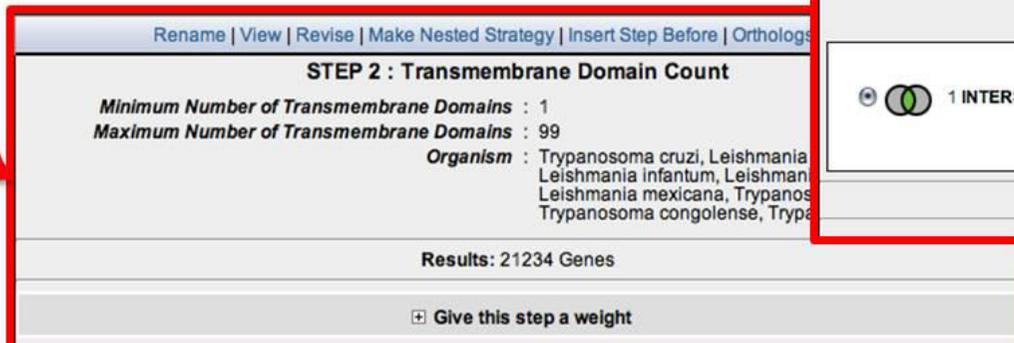


Ej. Al pulsar sobre Dominios Transmembrana se despliegan todos los genes con este tipo de motivo.

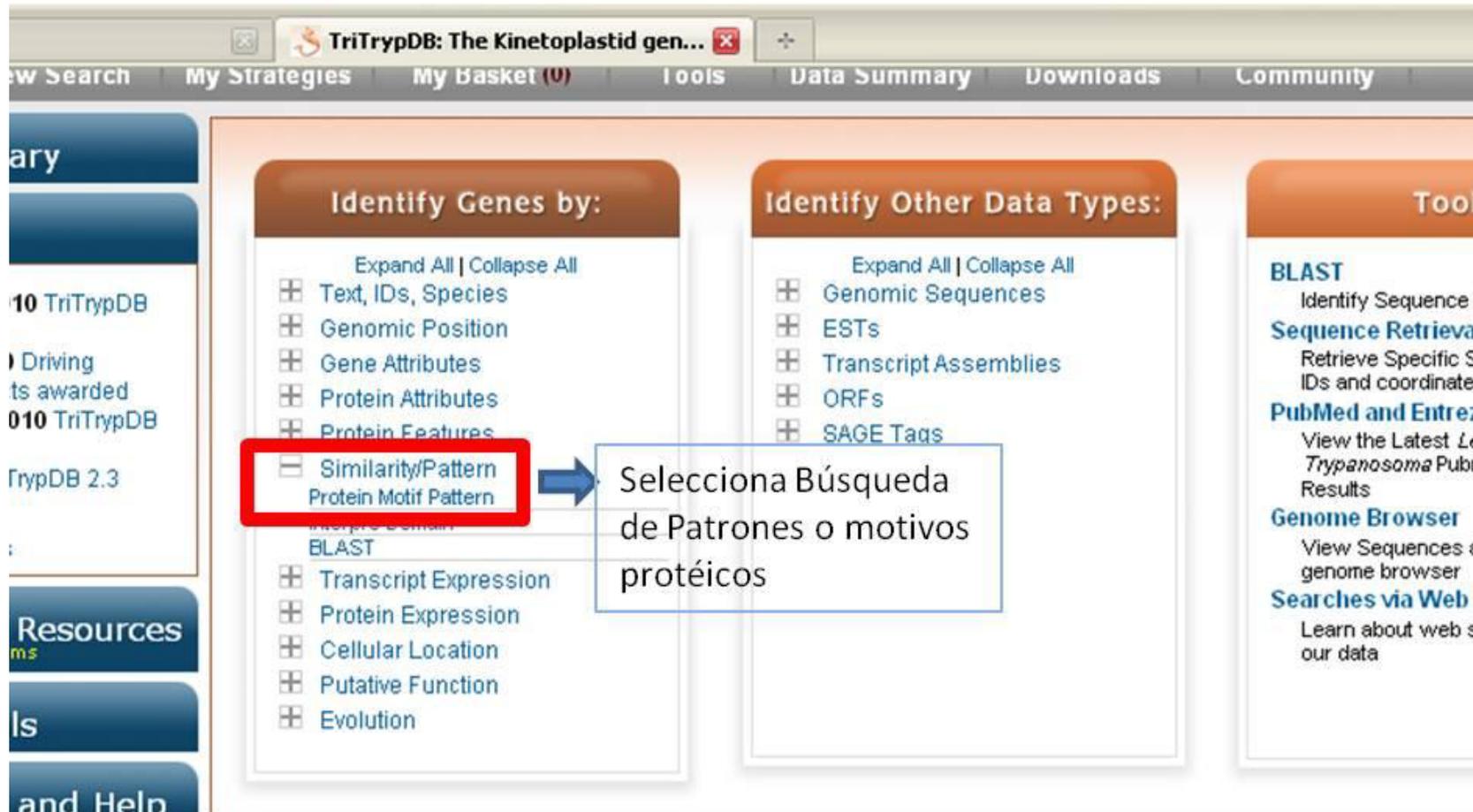


Cada etapa se puede revisar y modificar

Cada paso tiene varias opciones de trabajo.
Intersect 1-2
Unión 1-2
1 Minus 2
2 Minus 1



Las expresiones regulares son una serie de caracteres que forman un **patrón**, normalmente representativo de otro grupo de caracteres mayor, de tal forma que podemos comparar el patrón con otro conjunto de caracteres para ver las coincidencias



MOTIVOS ESTRUCTURALES:

- Es una forma de caracterizar un conjunto de secuencias del alfabeto formado por el ADN o las proteínas.
- Los motivos representan zonas conservadas entre las secuencias que suelen asociarse a características funcionales del grupo de secuencias
- Las expresiones regulares:
 - **Comodines:** Apto para cualquier carácter
 - **Ambigüedades:** Se presenta/prohíbe varios caracteres
 - **Factores de repetición:** número de veces que se presenta [o puede presentarse] un carácter

Sintaxis de expresiones regulares

Caracteres comodín

- Si en una posición dada puede aparecer cualquier carácter se indica con el signo “comodín”
- Aunque en informática éste es a menudo un “*” aquí se utilizará una “x”

G	A	T	T	A	C	A
G	A	C	T	A	C	T
T	A	A	T	A	C	T
A	A	T	T	A	C	C
	A	x	T	A	C	

Patrón: A-x-T-A-C

Sintaxis de expresiones regulares

Ambigüedades

- Si en una posición dada puede aparecer varios caracteres distintos podemos indicarlo de dos formas
 - Aquellos que pueden aparecer: entre “[“ y “]”
 - Aquellos que no se encuentran en la posición: entre “{“ y “}”
- Una misma secuencia se puede indicar de maneras distintas. *P.ej: [ATC] equivale a {G}*

G	A	T	T	A	C	A
G	A	C	T	T	C	T
T	T	A	T	C	C	T
A	T	T	T	A	C	C
	[AT]	x	T	{G}	C	

Patrón: [AT]-x-T-{G}-C={CG}-x-T-[ATC]-C= ...

Sintaxis de expresiones regulares

Elementos repetidos

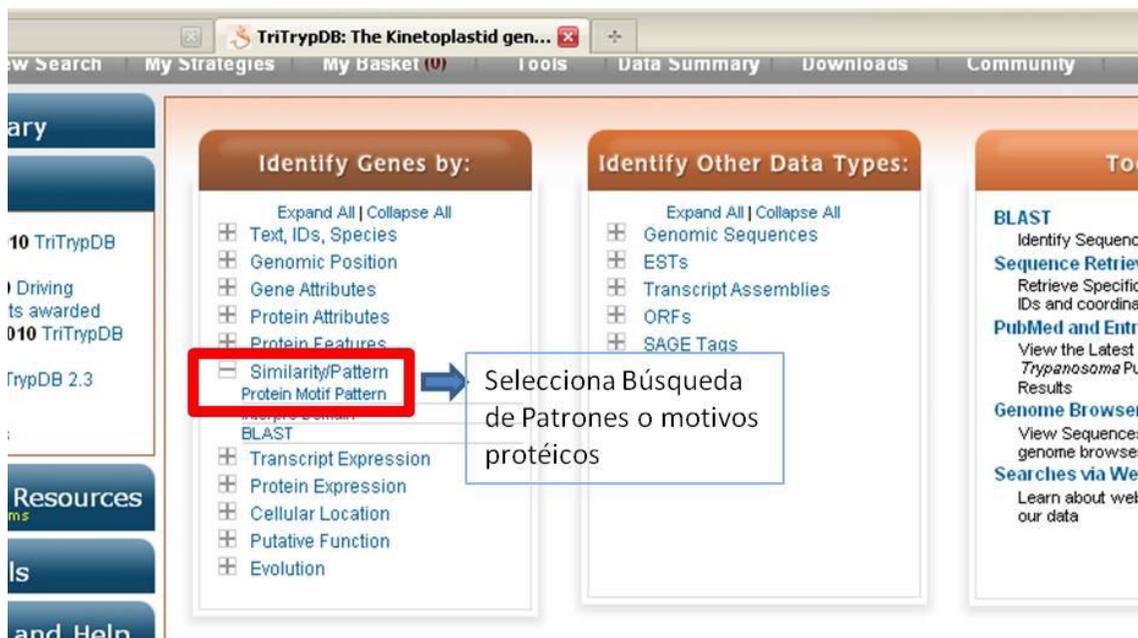
- La repetición de un elemento se indica con éste entre paréntesis: “ (“y”) ”
 - A(4) indica una “A” repetida 4 veces
 - x(3) indica un caracter cualquiera repetido 3 veces
 - Si el elemento que se repite es uno cualquiera (“x”) puede asignarsele un número variable de repeticiones, incluso el cero
 - x(2-4): “x-x”, “x-x-x”, “x-x-x-x”
 - x(0-2): “”, “x”, “x-x”

Text: The sequence must start with an alanine, followed by any amino acid, followed by a serine or a threonine, two times, followed by any amino acid or nothing, followed by any amino acid except a valine.

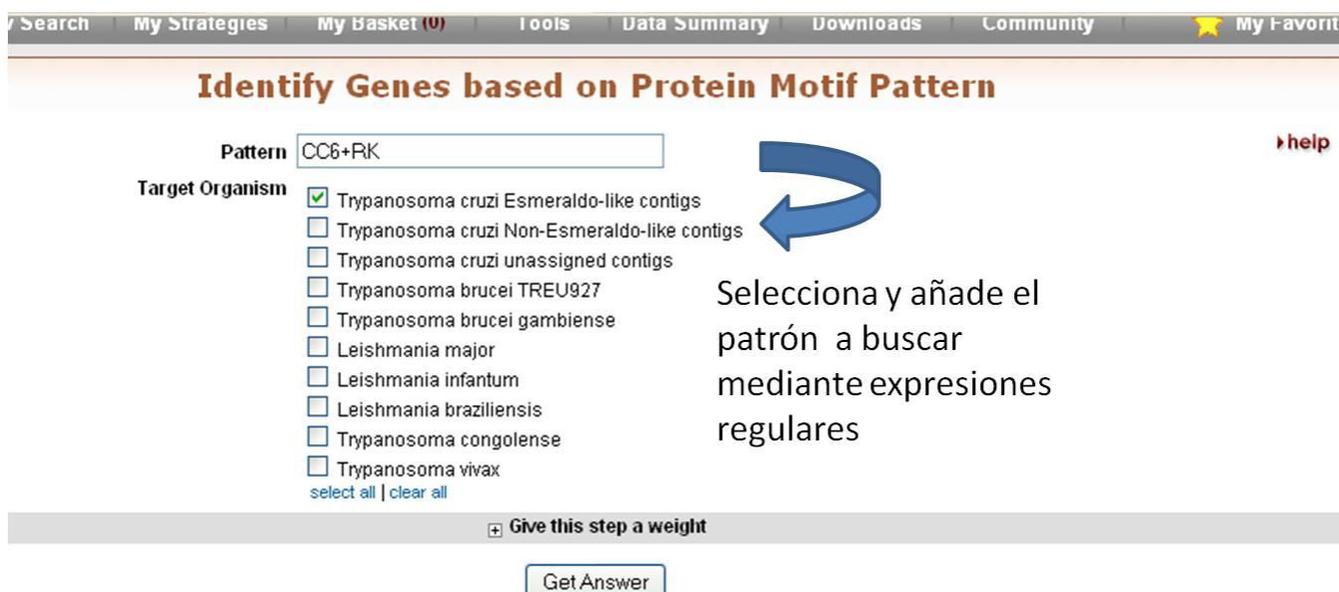
- ALSTDNVANRPMKPEMF....

$^A.[ST]\{2\}.?[^V]$

TriTrypDB: Búsqueda de motivos estructurales (Regular Expressions :RegEx)



The screenshot shows the TriTrypDB website interface. The main navigation bar includes 'New Search', 'My Strategies', 'My Basket (0)', 'Tools', 'Data Summary', 'Downloads', and 'Community'. The 'Identify Genes by:' section is expanded, showing a list of search criteria. The 'Similarity/Pattern' option is highlighted with a red box, and a blue arrow points to it with the text 'Selecciona Búsqueda de Patrones o motivos protéicos'. Other options include Text, IDs, Species; Genomic Position; Gene Attributes; Protein Attributes; Protein Features; BLAST; Transcript Expression; Protein Expression; Cellular Location; Putative Function; and Evolution. The 'Identify Other Data Types:' section includes Genomic Sequences, ESTs, Transcript Assemblies, ORFs, and SAGE Tags. The 'Tools' section includes BLAST, Sequence Retrieval, PubMed and Entrez, Genome Browser, and Searches via Web!



The screenshot shows the 'Identify Genes based on Protein Motif Pattern' search form. The 'Pattern' field contains 'CC6+RK'. The 'Target Organism' section has a list of organisms with checkboxes: Trypanosoma cruzi Esmeraldo-like contigs, Trypanosoma cruzi Non-Esmeraldo-like contigs, Trypanosoma cruzi unassigned contigs, Trypanosoma brucei TREU927, Trypanosoma brucei gambiense, Leishmania major, Leishmania infantum, Leishmania braziliensis, Trypanosoma congolense, and Trypanosoma vivax. There are 'select all' and 'clear all' links. A blue arrow points to the 'Trypanosoma cruzi Esmeraldo-like contigs' option with the text 'Selecciona y añade el patrón a buscar mediante expresiones regulares'. At the bottom, there is a 'Give this step a weight' checkbox and a 'Get Answer' button.

Construye un patrón estructural mediante las expresiones regulares del Apéndice y selecciona el organismo donde realizar su búsqueda.

Apéndice 1. Expresiones Regulares. RegEx

- “[]” → Aminoácidos (Nts) que aparecen en una posición dada. Ej “[AML]” En esa posición puede haber una ocurrencia de que exista el aminoácido Alanina, Metionina o Leucina
- “{ }” → Se indica los aminoácidos (Nts) que no se encuentran en una posición dada. Ej “{A,T,C}”, significa que en esa posición no existe A,T,C, es decir que sólo se encuentra G (aunque esta también se puede indicar como [G]).
- “{n}” → Indica el número de repeticiones de un aminoácido en concreto. Ej “C{7}” Cisteína repetida un total de 7 veces en una secuencia.
- “^” → Inicio de una secuencia. Ej “^A” La secuencia debe empezar en Alanina
- “[^]” → Ocurrencia de cualquier A.a (Nts) excepto los indicados en los corchetes. Ej “[^A]” Ocurrencia de cualquier aminoácido excepto Alanina.
- “.” → Detrás de un A.a (Nts) cualquier otro. Ej “A.” Detrás de una alanina cualquier otra
- “?” → Detrás de un A.a (Nts) cualquier otro o ninguno. Ej “A.?” Detrás de una alanina cualquier otra o ninguna.
- “\$” → Indica en que aminoácido (Nt) debe terminar la secuencia. Ej “\$L” La secuencia acaba en Leucina.
- “+” → Probabilidad de que ocurra uno o más de los caracteres precedentes. Ej. “C+RK” Podemos encontrar 1, 2, 3...o más cisteínas antes de una Arginina y una Lisina.
- “*” → 0 o más ocurrencias del carácter anterior. Ej. “CR*F” Podemos encontrar 0, 1, 2, 3... O más Argininas antes de una Fenilalanina.
- “?” → 0 o 1 ocurrencia de encontrar el carácter precedente. Ej. “CR?F” 0 o una ocurrencia de que exista una arginina antes de una Fenilalanina.
- “-x(n)” → Separación entre aminoácidos (Nts). Ej “C-x(7)C” Separación de 7 aminoácidos entre dos cisteínas.

AA property	Amino acids	Code
Acidic	DE	0
Alcohol	ST	1
Aliphatic	ILV	2
Aromatic	FHWY	3
Basic	KRH	4
Charged	DEHKR	5
Hydrophobic	AVILMFYW	6
Hydrophilic	KRHDENQ	7
Polar	CDEHKNQRST	8
Small	ACDGNPSTV	9
Tiny	AGS	B
Turnlike	ACDEGHKNQRST	Z
Any	ACDEFGHIKLM NPQRSTVWY	.

Formatos de Secuencias

Plain sequence format

A sequence in plain format may contain only [IUPAC characters](#) and spaces (no numbers!).

Note: A file in plain sequence format may only contain **one** sequence, while most other formats accept several sequences in one file.

An example sequence in plain format is:

```
ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTTCCTCGCTTGGTGGTTTGGTGGACCTCCCAGGCCAGTGCCGGGCCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAAC TTC TTCTGGAAGACCTTCTCCTCCTGCAAA TAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

FASTA format

A sequence file in FASTA format can contain several sequences.

Each sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (">") symbol in the first column.

An example sequence in FASTA format is:

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368
ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTTCCTCGCTTGGTGGTTTGGTGGACCTCCCAGGCCAGTGCCGGGCCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAAC TTC TTCTGGAAGACCTTCTCCTCCTGCAAA TAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

EMBL format

A sequence file in EMBL format can contain several sequences.

One sequence entry starts with an identifier line ("ID"), followed by further annotation lines. The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").

An example sequence in EMBL format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
acaagatgcc attgtcccc ggctcctgc tgctgctgct ctccggggcc acggccaccg      60
ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg      120
caggaataag gaaaagcagc ctctgactt tctctgcttg gtggtttgag tggacctccc      180
aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag      240
gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga      300
agaccttctc ctctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca      360
gacctgaa
//
```

GenBank format

A sequence file in GenBank format can contain several sequences.

One sequence in GenBank format starts with a line containing the word LOCUS and a number of annotation lines. The start of the sequence is marked by a line containing "ORIGIN" and the end of the sequence is marked by two slashes ("/").

An example sequence in GenBank format is:

```
LOCUS      AB000263                368 bp    mRNA    linear    PRI 05-FEB-1999
DEFINITION Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION  AB000263
ORIGIN
     1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
    61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
   121 caggaataag gaaaagcagc ctctgactt tctctgcttg gtggtttgag tggacctccc
   181 aggccagtgc cgggccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
   241 gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
   301 agaccttctc ctctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
   361 gacctgaa
```

//

IUPAC nucleic acid codes

To represent ambiguity in DNA sequences the following letters can be used (following the rules of the *International Union of Pure and Applied Chemistry* (IUPAC)):

A = adenine
C = cytosine
G = guanine
T = thymine
U = uracil
R = G A (purine)
Y = T C (pyrimidine)
K = G T (keto)
M = A C (amino)
S = G C
W = A T
B = G T C
D = G A T
H = A C T
V = G C A
N = A G C T (any)
