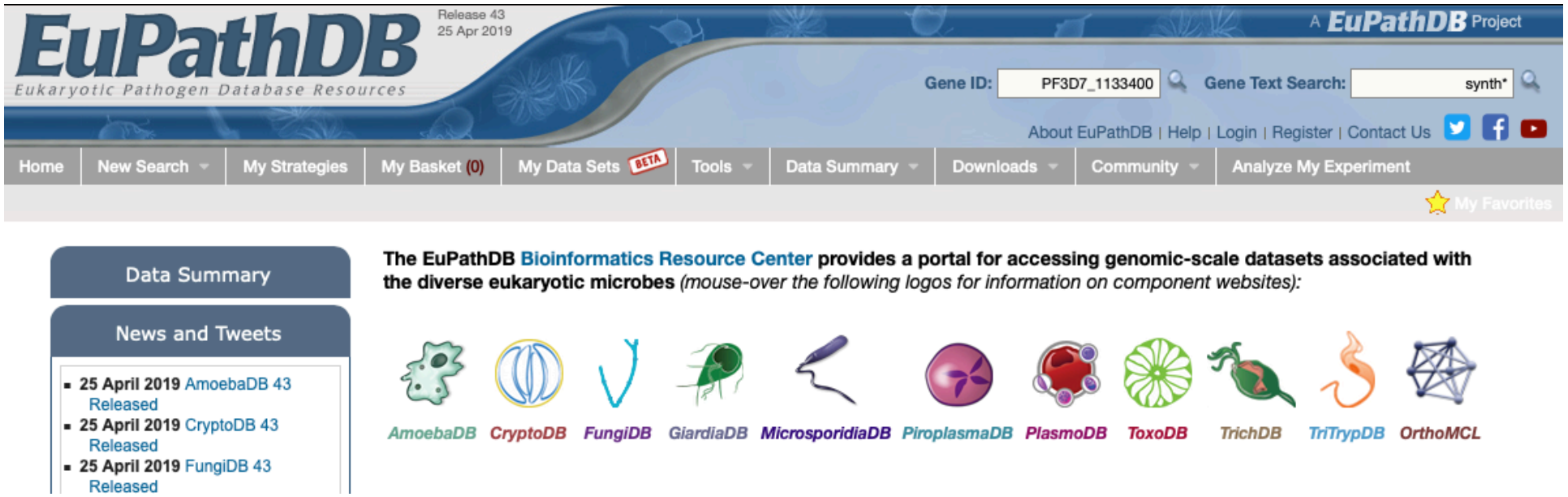


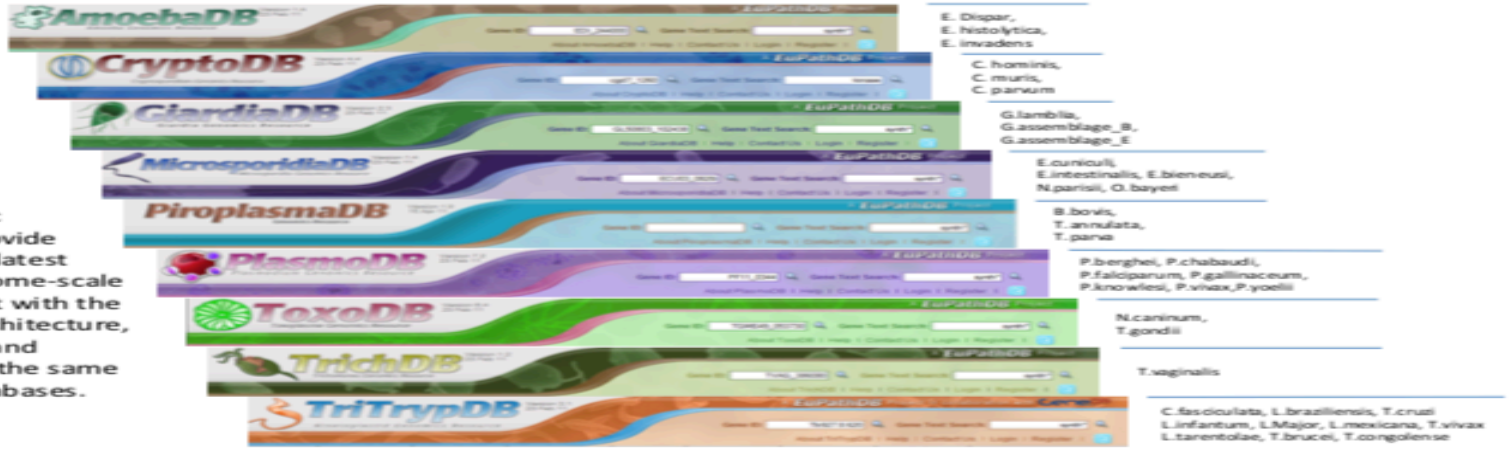
Bases de datos específicas: EupathDB

Especialización en Micología y Parasitología
Facultad de Ciencias Bioquímicas y Farmacéuticas
Universidad Nacional de Rosario
2013

Dra. Victoria Alonso
alonso@ibr-conicet.gov.ar

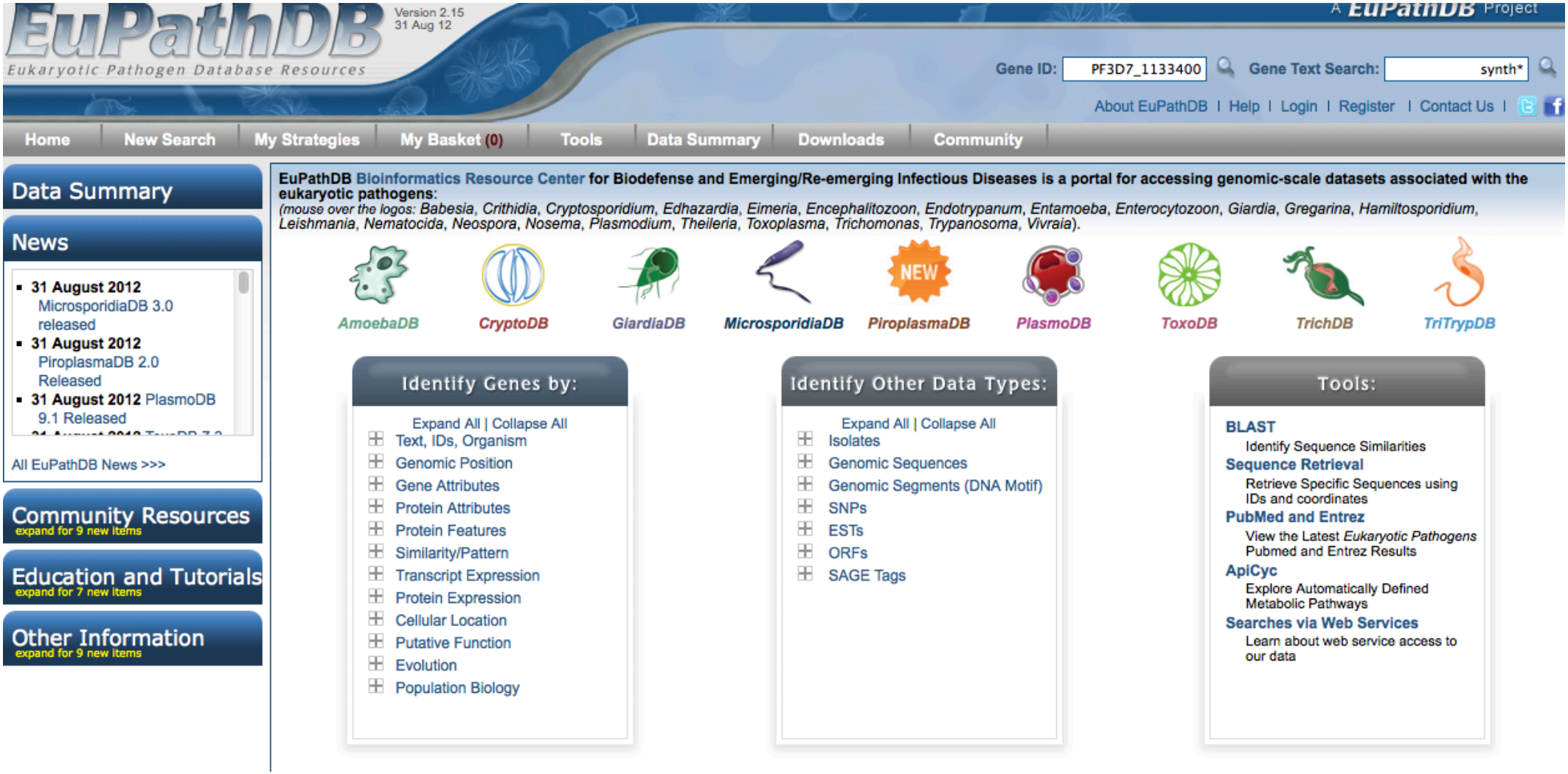
- EuPathDB (**http://EuPathDB.org**) es una base datos integral que engloba los genomas de los organismos parásitos patógenos: *Cryptosporidium, Giardia, Leishmania, Neospora, Plasmodium, Toxoplasma, Trichomonas* o *Trypanosoma*
- Base de datos desarrollada por Bioinformatics Resource Center (BRC), financiada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



→Pulsa en el icono "TrytrypDB"

-Pagina web integral de“TrytrypDB” que incluye los genomas de los kinetoplastidos secuenciados, herramientas para su uso, asi 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 (*).

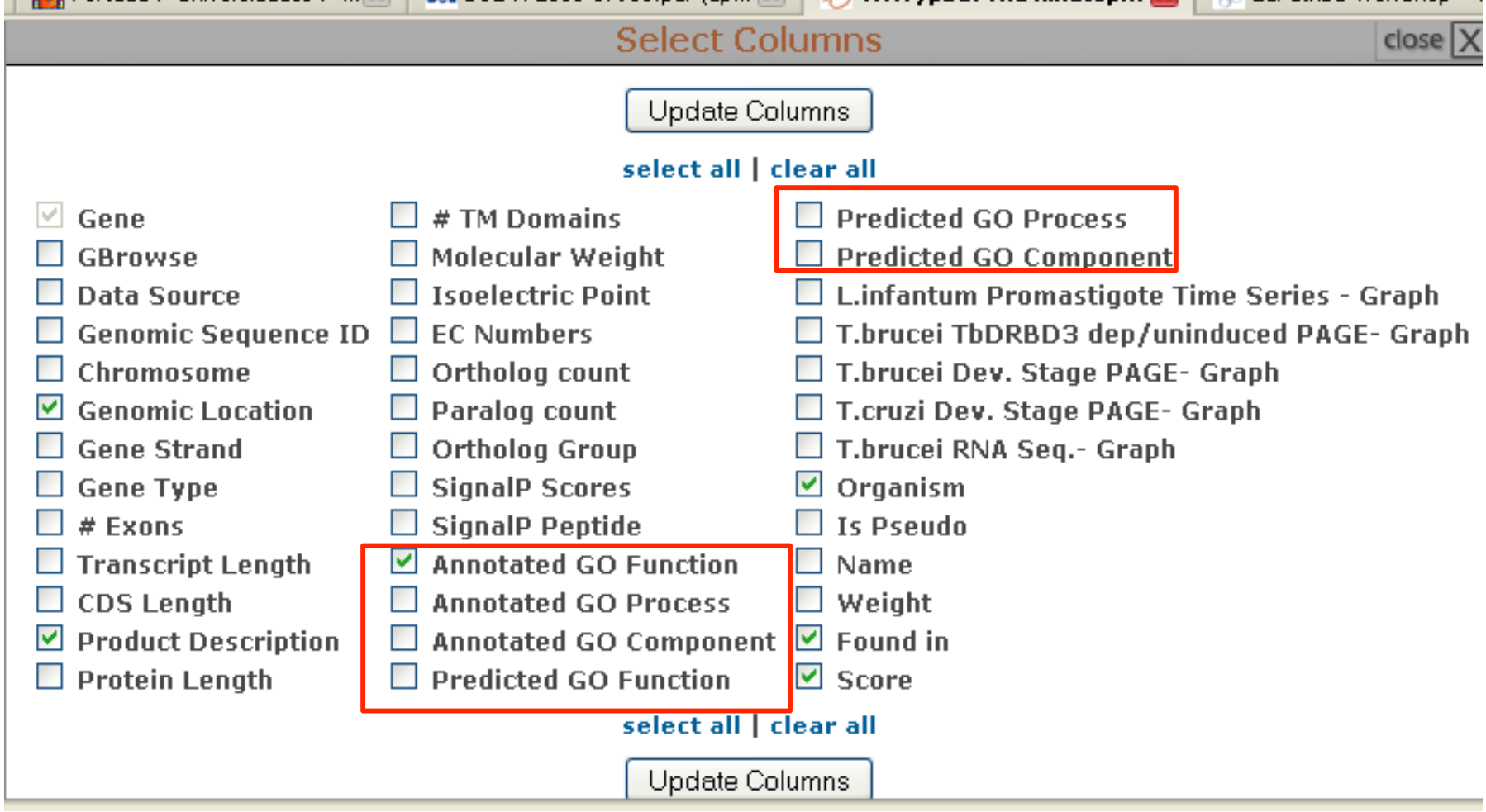
Alertas y noticias especiales

Búsquedas y herramientas

Información y ayuda

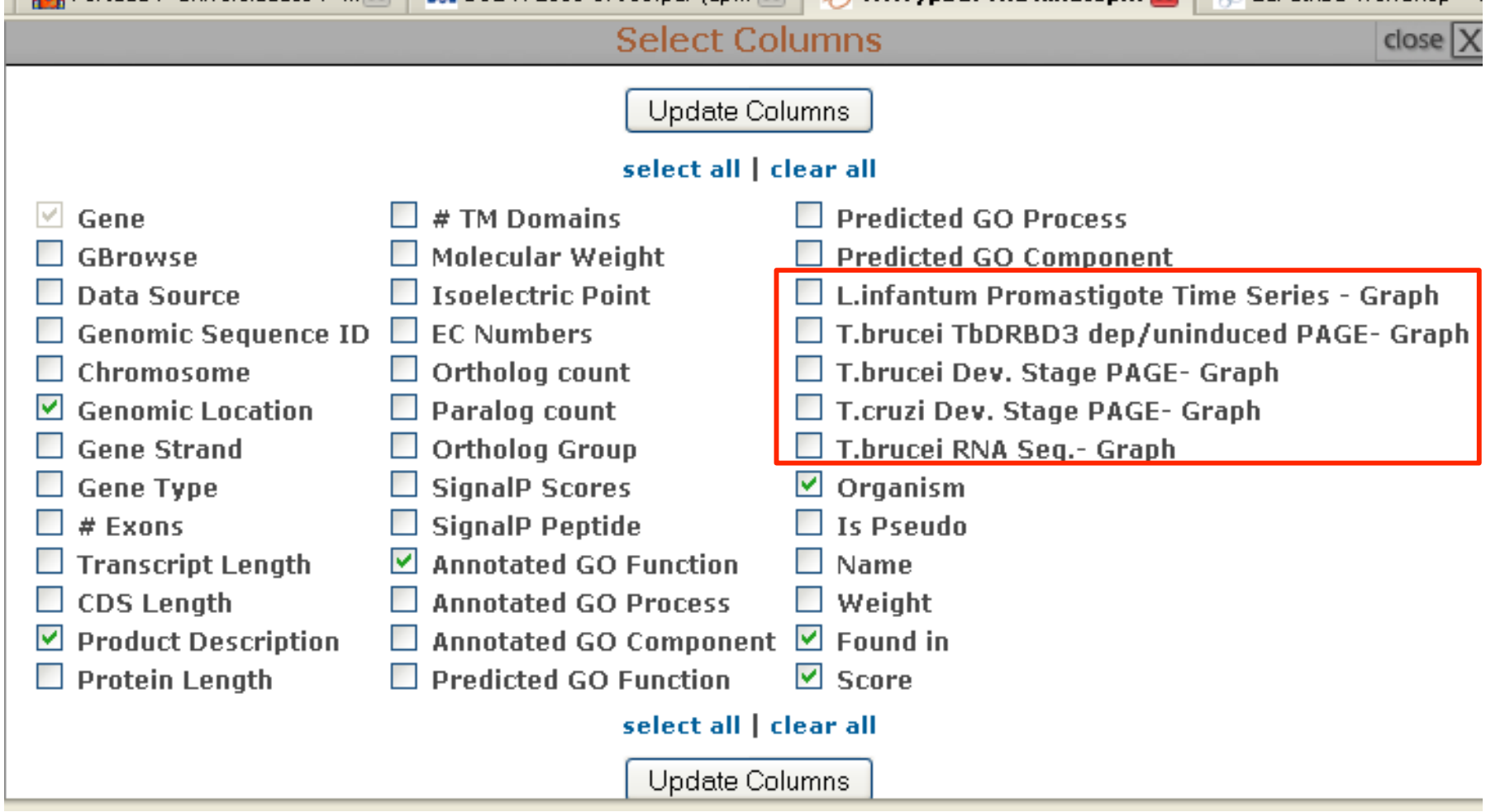
BLAST, PubMed bibliografía, genome browser, más tipos de búsqueda

FUNCIONES BÁSICAS



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 extracelulart.
•*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.



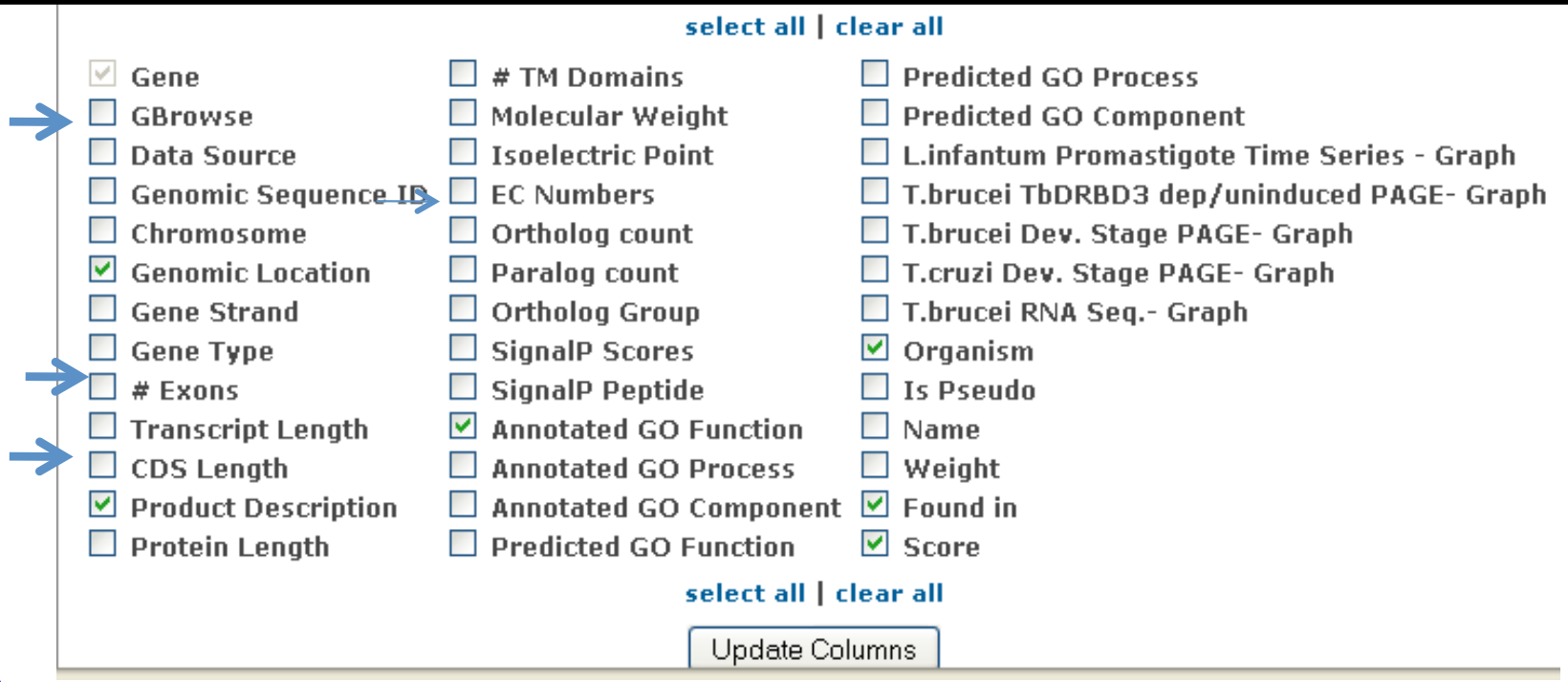
Expresión:

-Transcriptómica

-Microarrays de diferentes tipos de estadíos parasitarios

-RNAseq. Secuenciación de RNA.

- Espectrometría de Masas



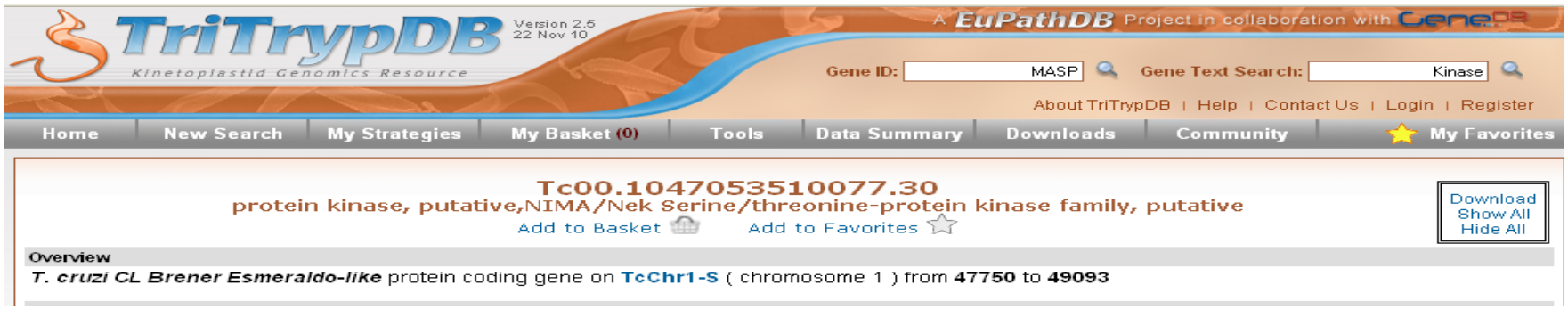
CDS:
Coding sequence. Región de nucleótidos que corresponden a una secuencia de aminoácidos en una proteína y que está incluida entre los codones de inicio y de parada. Las partes no expresadas (5´y 3´UTR e intrones) no se incluyen dentro de un CDS

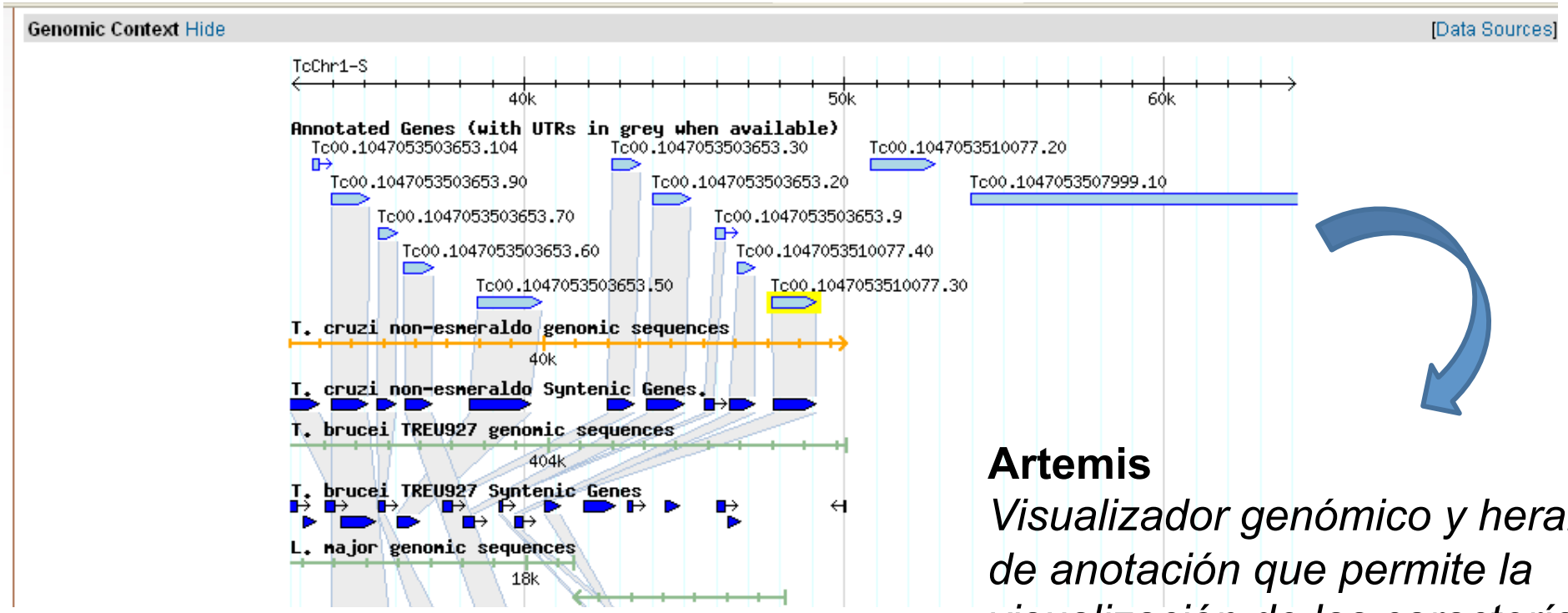
Exones:
Regiones de un gen que no son separadas durante el proceso de splicing y, por tanto, se mantienen en el ARN mensajero maduro

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



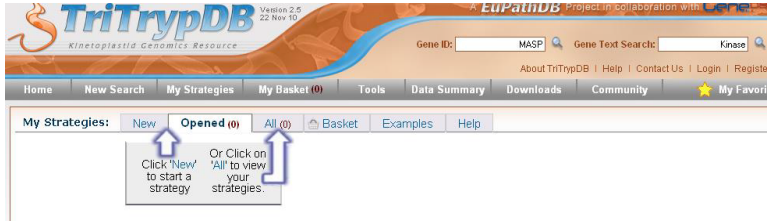


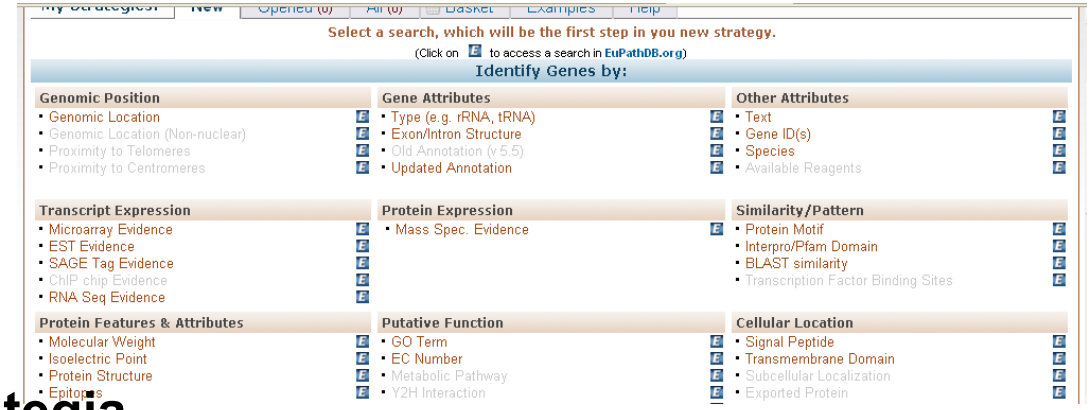
Artemis

Visualizador genómico y heramienta 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

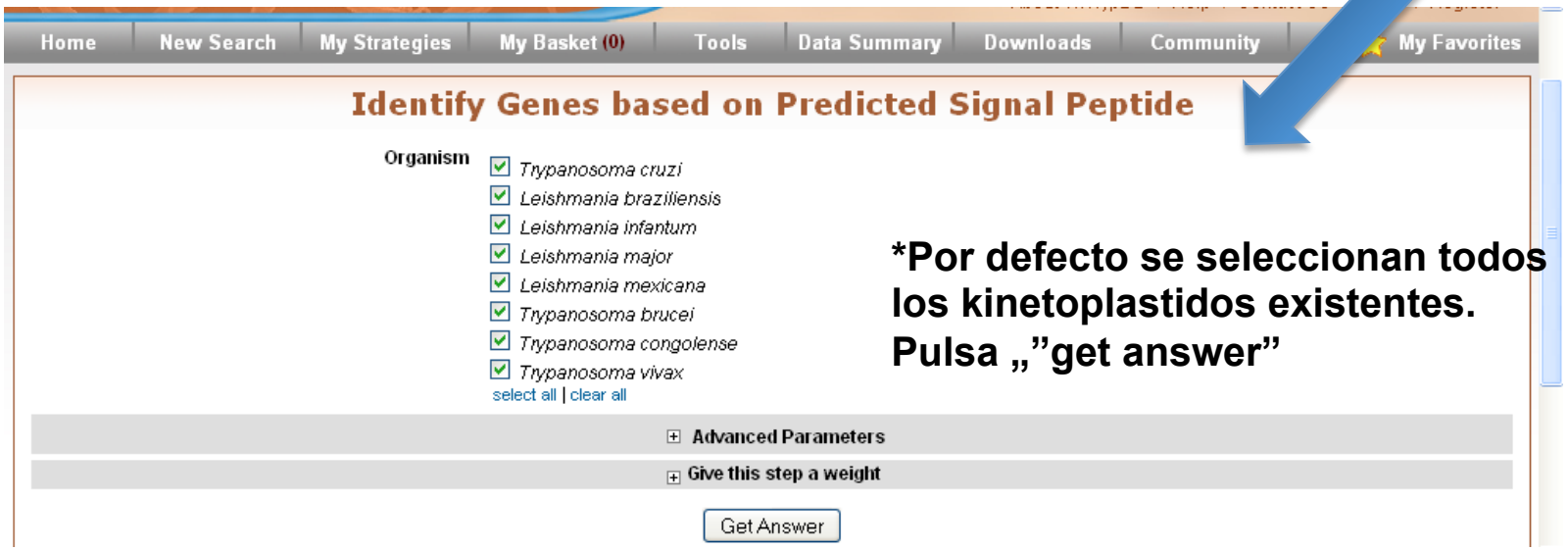
TritrypDB: Estrategias complejas “COMPLEX QUERY”

1.Iniciar una estrategia
Pulsa “My strategies” en el menú principal e inicia una nueva estrategia.



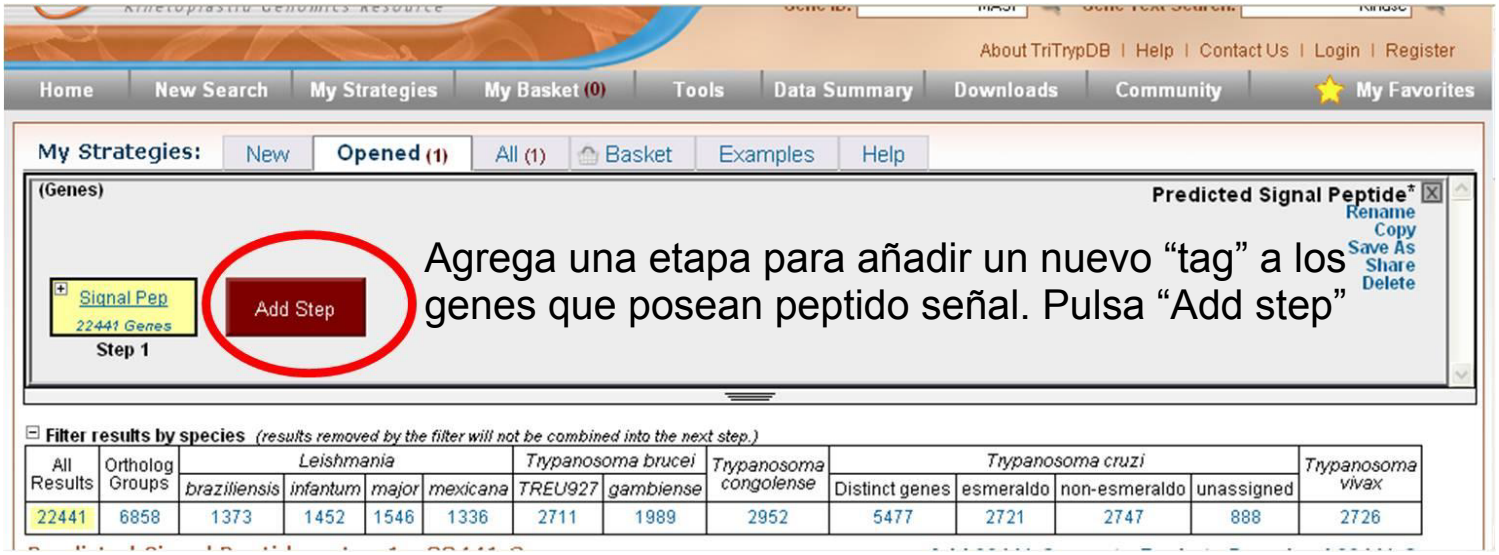


Menú de posibilidades de inicio de estrategia
Todas las opciones contenidas en la base de datos. Ej pulsemos Las opciones para “Cellular location”→ ✍Signal peptide



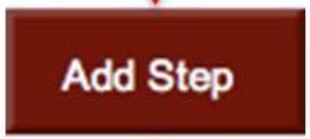
***Por defecto se seleccionan todos los kinetoplastidos existentes. Pulsa „”get answer”**

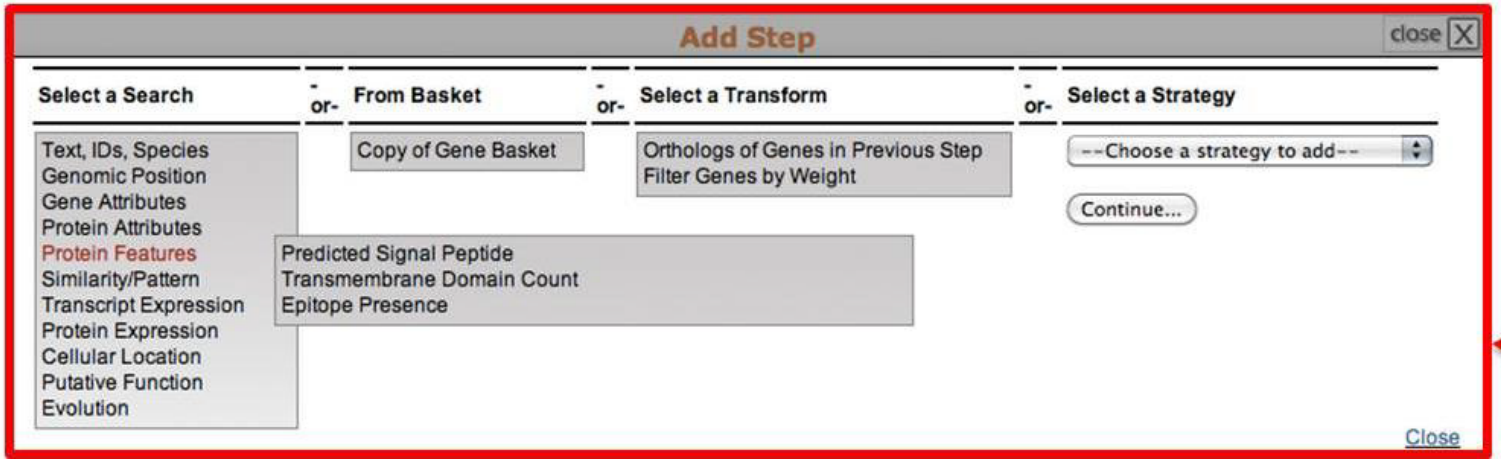
TrytrypDB: Estrategias complejas “COMPLEX QUERY”



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.



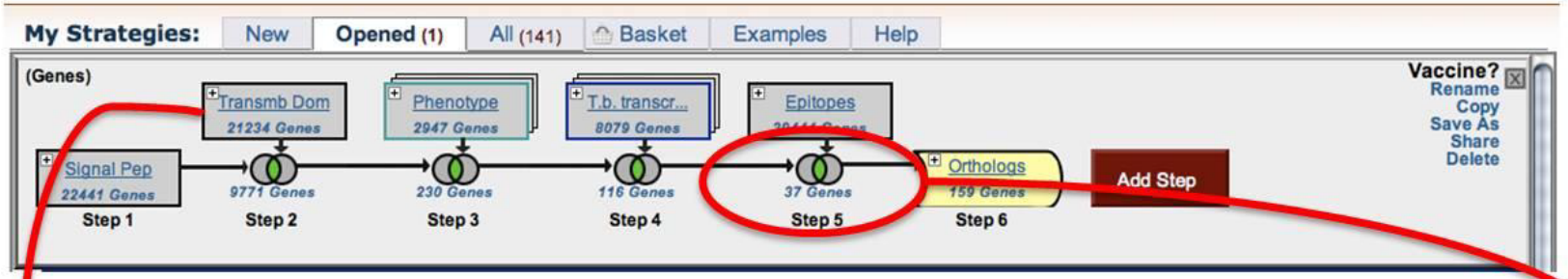


TritrypDB: Estrategias complejas “COMPLEX QUERY”

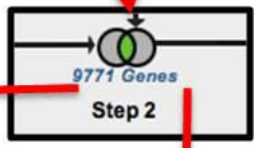
Strategy Builder Guide:

TritrypDB: Estrategias complejas “COMPLEX QUERY”

Strategy Builder Guide (cont.):

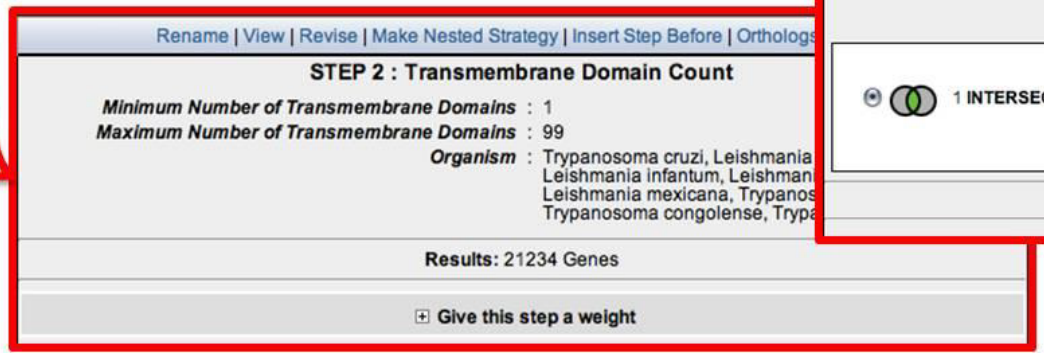


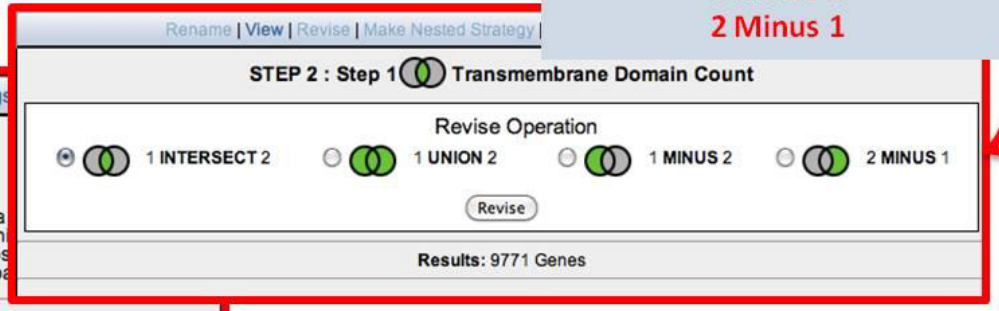
Ej. Al pulsar sobre Dominios Transmembrana se despliegan todos los genes con este tipo de motivo.



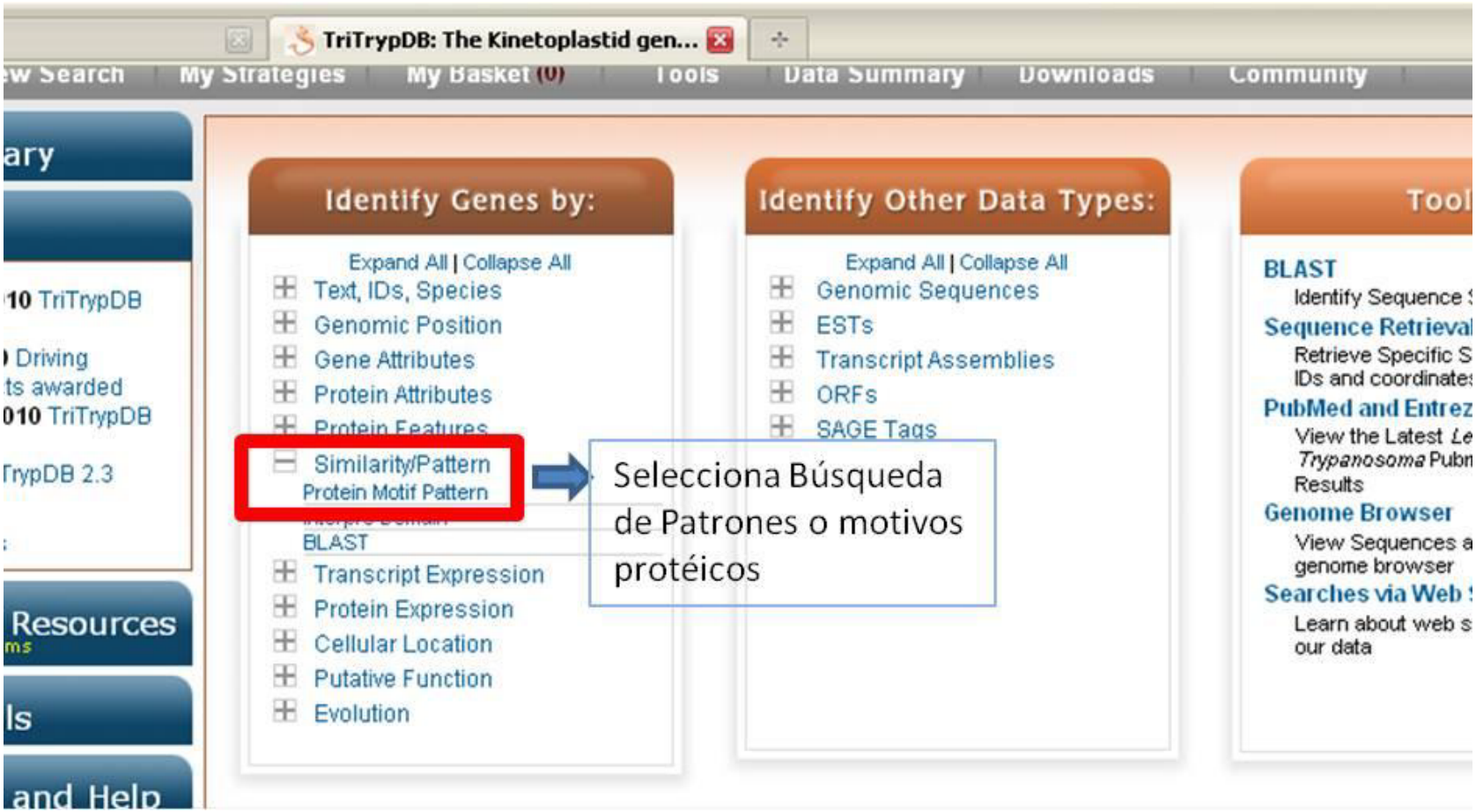
Cada etapa se puede revisar y modificar

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 carácteres que forman un **patrón**, normalmente representativo de otro grupo de carácteres mayor, de tal forma que podemos comparar el patrón con otro conjunto de carácteres para ver las coincidencias



TritrypDB: Busqueda de motivos estructurales (Regular Expressions :RegEx)

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

-**Ambiguedades**: Se presenta/prohibe varios carcacteres

-**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
Ambiguedades

- 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*

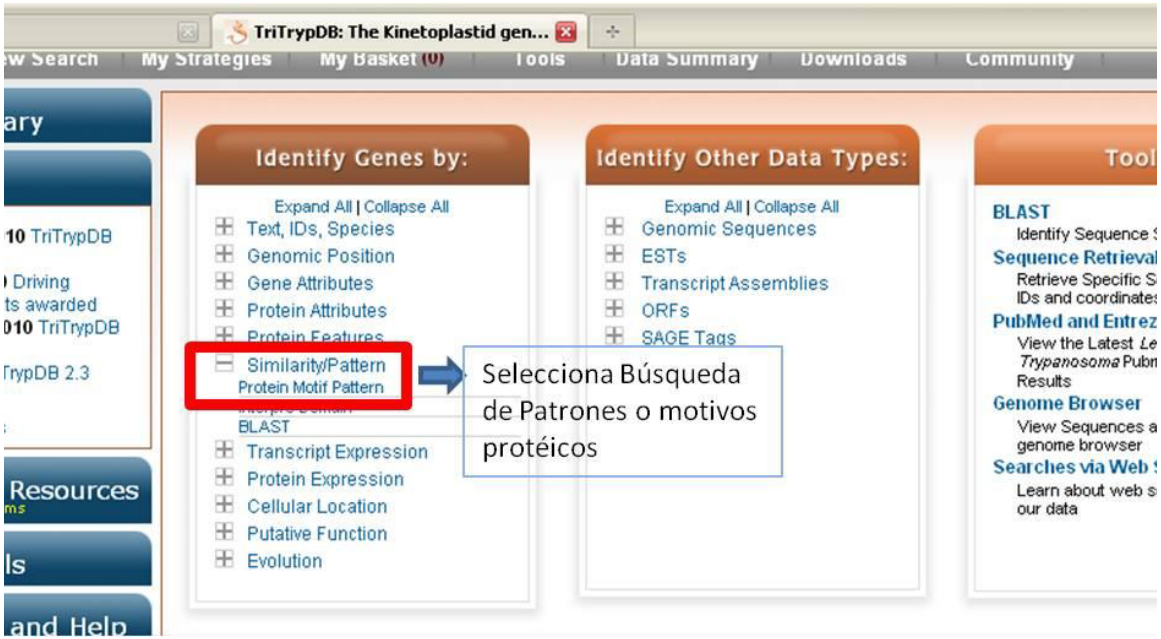
- La repetición de un elemento se indica con éste entre paréntesis: “(“y”)”
  - A(4) indica una “A” repetida 4 veces
  - x(3) indica un caracter cualquiera repetido 3 veces
  - Si el elemento que se repite es uno cualquiera (“x”) puede asignarsele un número variable de repeticiones, incluso el cero
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

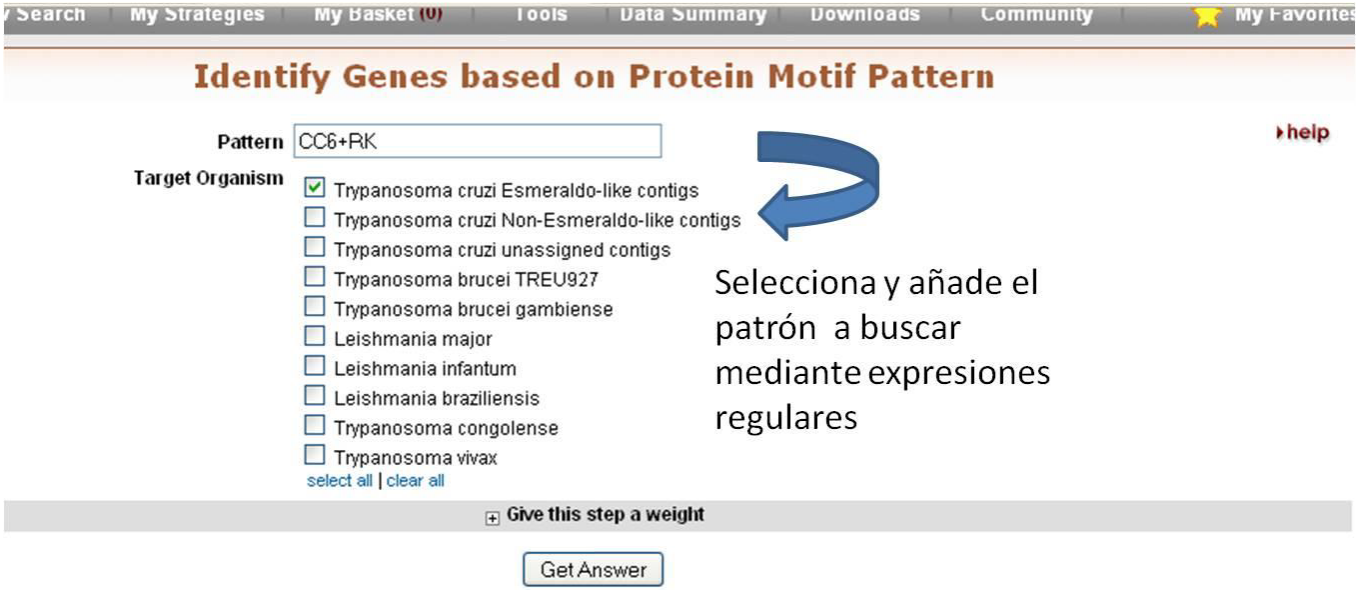
Text: The sequence must start with an alanine, followed by any amino acid, followed by a serine or a threonine, two times, followed by any amino acid or nothing, followed by any amino acid except a valine.

- ALSTDNVANRPMPKPEMF....

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

TritrypDB: Busqueda de motivos estructurales (Regular Expressions :RegEx)





Construye un patron estructural mediante las expresiones regulares del Apendice y selecciona el organismo donde realizar su busqueda.

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].
- "{nº}"→ 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(nº)"→ 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:

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
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
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
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 attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

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 ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa 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)
```