

Actualizaciones y observaciones proyecto Pepti-tools

Deadline envío Paper: Julio-30

Deadline término aplicación: Agosto-30

Desarrollo del paper

Estado actual:

1. Escrito, primera versión

Pendientes:

1. Correcciones revisión de autores
2. Escribir material suplementario
3. Preparar un graphical abstract o imagen general de la aplicación.
4. Preparar cover latter
5. Decidir revista a ser enviado -> Tentativamente, Bioinformatics as App Note

Estado aplicación

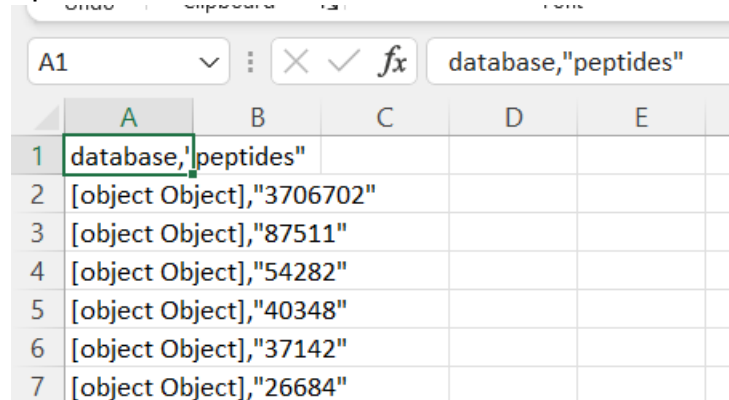
La aplicación se encuentra funcional en una primera versión.

Observaciones

Fueron enviadas por correo previamente, los links son

1. Comentarios Anita:
https://docs.google.com/document/d/1siFzzeFWdmOTiO8hgdPwzv41yaMnT_tOt-2Ud-Wkrxk/edit
2. Comentarios Benjamín:
https://docs.google.com/document/d/1TAnxAJpjnXsTvX_5WMUK_o0to0l-1dKGH5zEABCawSo/edit
3. Comentarios David:
 - a. Actualizar homepage, remover la información que corresponde a About us y a How to cite
 - b. Agregar teams members, Benjamín, Anita, Álvaro y Roberto
 - c. Agregar una sección about us en el menú, esta sección tiene que tener los miembros, la información del equipo.
 - d. Agregar una sección How to cite en el menú, esta sección tiene que tener la información de las citas al paper de peptipedia, este que lanzaremos ahora y todos los trabajos que hagamos en péptidos.
 - e. Database:

- i. Al momento de descargar las tablas en formato csv aparecen de esta forma:



	A	B	C	D	E
1	database,"peptides"				
2	[object Object],"3706702"				
3	[object Object],"87511"				
4	[object Object],"54282"				
5	[object Object],"40348"				
6	[object Object],"37142"				
7	[object Object],"26684"				

- ii. Lo mismo pasa en ambas tablas.
- iii. Deberíamos tener un gráfico de barras en el home page con ambos elementos
- iv. En el home page me falta un cuadro de indicadores. Por ejemplo, cuántas secuencias tenemos, cuántas actividades, cuántas bases de datos y cuándo se actualizaron las cosas.
- v. Al descargar las secuencias, me aparece esta información

```
>1
FNKLKQGSSKRTCACFRKIMPSVHELDERRRGANRWAAGFRKCVSSICRY
>2
RLGTALPALLKTLLAGLNG
>3
YEALVTSILGKLTGLWHNDSVDFMGHICYFRRRPKIRRFKLYHEGKFWCPGWAPFEGRCKYCVVF
>4
VFCTCRGFLCGSGERASGSCTINGVRHTLCCRR
>5
NVILGSAOEFKPSD
```

se necesita que entregue al menos la actividad y el estado, si es predicha o no

- f. Sección Advanced Search
 - i. Falta agregar el cuadro de descripción de cómo emplear la búsqueda avanzada
 - ii. Cuando hago las descargas de las tablas como resultados de las queries, se genera el siguiente CSV, no siendo descargadas todas las secuencias, sólo las de la vista

actual y entregando esos resultados object:

	A	B	C	D	E	F	G	H	I
1	idpeptide,"length","molecular_weight","charge_density","isoelectric_point","charge","Options"								
2	1,"51","5976.99","0.002","11.3267","11.826","[object Object]"								
3	3,"65","7830.16","0.0009","9.8977","7.023","[object Object]"								
4	4,"33","3550.14","0.0013","8.7471","4.691","[object Object]"								
5	6,"67","7748.72","0.0004","9.9897","3.393","[object Object]"								
6	7,"44","4797.32","0.0002","9.4414","1.101","[object Object]"								
7	9,"44","4973.64","0","7.2383","0.129","[object Object]"								
8	11,"40","4359.03","0.0009","8.4001","3.788","[object Object]"								
9	12,"33","3381.1","0.0014","10.6924","4.86","[object Object]"								

- iii. Al agregar el cuadro explicativo, quedará mucho más funcional esta parte.
- g. Sección Fasta converter, falta agregar el cuadro explicativo de como funciona dicho servicio.
- h. Alineamiento
 - i. Falta el cuadro resumen con los inputs
 - ii. No se entiende cuándo es un resultado no significativo.
 - iii. En el cuadro de texto, se debe explicar qué estrategia de Blast emplearon para esto.
 - iv. En la tabla resultados, No es options, es details.
- i. Multi Alignment Sequences
 - i. Se debe permitir descargar el resultado del alineamiento, archivo texto
 - ii. Se debe permitir generar la matriz de distancia
 - iii. Se debe permitir generar el dendrograma
 - iv. Falta el cuadro indicativo de cómo funciona dicha tools
 - v. Se debe poder permitir generar una matriz de distancia en formato heatmap.
 - vi. Limitar a un máximo de 50 secuencias
- j. Pfam Prediction y GO prediction
 - i. Falta el cuadro indicativo con cómo funciona esta cosa
 - ii. Debemos permitir indicar qué cosa estamos utilizando para que no nos hechen la culpa a nosotros si sale mal la cuestión.
 - iii. Se debe indicar cuándo es un resultado no significativo.
- k. Análisis de frecuencias.
 - i. Se debe indicar el cuadro resumen de funcionamiento
 - ii. Cuando son varias secuencias, se hace un gráfico promedio con barras de error y se deben mostrar los gráficos individuales por secuencia ingresada.
- l. Properties Estimation

- i. Incluir el cuadro resumen, en el mismo exponer qué herramienta se utiliza para estimar las propiedades
 - ii. Quitar decimales y dejar todos en orden de 4 decimales.
 - iii. Agregar las unidades de medida de cada respuesta
 - iv. Favor incluir los siguientes índices adicionales
 - 1. Instability Index
 - 2. Aromaticity
 - 3. Aliphatic index
 - 4. Boman Index
 - 5. Hydrophobic Ratio
 - v. Incluir el plot_profile con la imagen que se genera.
 - vi. Incluir el helical_wheel con la imagen que se genera.
- m. Encoding sequences
 - i. Incluir cuadro de texto con la explicación
 - ii. En el mismo cuadro, se debe indicar a qué sección corresponde y la información de interés con respecto a cada codificación.
 - iii. Actualizar la librería/API para trabajar con los nuevos codificadores de propiedades basados en nuestro más reciente artículo.
 - iv. Revisar dado a que me arrojó error cuando intenté generar 1 secuencia.
- n. Clustering
 - i. Actualizar las estrategias de codificación
 - ii. Agregar cuadro resumen descriptivo
 - iii. Corregir indicando cuál es cuál.

Calinski-Harabasz index	Davies-Bouldin Index	Davies-Bouldin Index
9	0.447	0.659

- iv. Agregar cuadro de interpretación de los resultados
 - v. Al descargar, también incluir las secuencias en el archivo CSV que se genera, así se tendría ID, sequence, grupo.
- o. Supervised Learning
 - i. Agregar cuadro descriptivo
 - ii. Agregar cuadro informativo sobre el input y sus características
 - iii. Agregar textos de información sobre los algoritmos
 - iv. Agregar estrategias de interpretación de los resultados.

- v. Permitir exportar el modelo y las estrategias generadas.
- vi. Falta agregar estandarización
- vii. Falta agregar opción de representar en PCA
- p. Otros
 - i. Agregar la sección How to use con el acceso a videos tutoriales de la aplicación
 - ii. Agregar el manual de usuario en formato PDF

Nuevos elementos de interés.

Al sistema actual se le deben agregar las siguientes funcionalidades con el fin de poder hacerlo publicable. Las cuales se nombran y detallan a continuación

1. Clasificación y predicción funcional de nuevas secuencias. Esto es asociado a los modelos de clasificación de funciones que se están actualmente realizando. La idea es tener un formulario similar al que se encuentra en los diferentes tools, contemplando un fasta o subida de archivo y obtener, el resultado es todas las posibles categorías que tenga una secuencia.
2. Usar modelos predictivos. Los modelos que se generan en la parte de supervised learning, los pueden probar de manera inmediata. No obstante, muchas veces esto no es tan factible, debido a que genero nuevas secuencias conforme pase el tiempo. Para ello se necesita que los modelos puedan ser descargados por el usuario. Esta tool debería funcionar así.
 - a. Subir archivo de modelo
 - b. Subir archivo de escalas si corresponde
 - c. Subir archivo de PCA si corresponde
 - d. Subir secuencias nuevas a predecir según respuesta.
 - e. Script aplica escala, PCA según corresponda y luego aplica el modelo
 - f. Se obtienen las respuestas y se muestran en el web, esto es una solución intermedia al hecho de “exportar modelos para usos a posteriori”.
3. Trataremos de agregar el servicio de clustering por comunidades, el cual funcionaría de la siguiente forma
 - a. Usuario selecciona entre evaluación filogenética o numérica
 - b. Si selecciona numérica
 - i. Usuario debe seleccionar tipo de codificación (una de las 8 propiedades)
 - c. El usuario debe ingresar sus secuencias en formato fasta
 - d. El usuario debe seleccionar el umbral de corte

- e. El sistema recibe y aplica el algoritmo con las configuraciones, retornando una matriz de adyacencia para generar un grafo, el csv con la clasificación (sólo los ID de las secuencias) y el rendimiento (tiempos de ejecución y métricas).
 - f. El sistema debe mostrar los resultados y permitir descargarlos.
- 4. Trabajaremos con distancias de secuencias o similitudes para ello, se contemplará los siguientes pasos.
 - a. Alinear todas las secuencias entre sí
 - b. Codificar todas las secuencias con todas las propiedades
 - c. Estimar distancias para todos los casos
 - d. Para cada caso, armar una tabla relación de una secuencia y sus secuencias relacionadas (las primeras 1000)
 - e. Obtener la probabilidad de relación mezclando todos los puntos de vista y armar una tabla de relaciones con estos valores para cada secuencia.
 - f. Finalmente, el sistema al mostrar los péptidos, mostrará las relaciones que tienen con los otros. El interés de esto, está en encontrar relaciones con diferentes actividades, en diferentes niveles del árbol.
- 5. Exploración de Moon light. Este sistema se basa en el análisis de las actividades biológicas de péptidos al mismo tiempo. Por ejemplo, para los antimicrobianos, seleccionar todos aquellos que tienen otra actividad adicional. Esta actividad debe estar al mismo nivel que la secuencia de interés. La idea es que el usuario pueda filtrar secuencias con las propiedades que estime conveniente y seleccionar sólo 1 actividad. En base a esto, el sistema reconoce el nivel y obtendrá el parent, luego debe buscar todas las secuencias que tienen esa actividad y obtener si tiene actividades en el mismo nivel pero con un parent diferente, para poder generar el diagrama de chord. La idea es que se pueda exportar este detalle.
- 6. Incluir servicio de agregar secuencias a la DB. Para ello se debe completar el formulario con la siguiente información.
 - a. Archivo de secuencias
 - b. Actividad biológica (de las que tenemos actualmente)
 - c. Nombre de Fuente
 - d. Una vez recibida las secuencias, el sistema debe generar los siguientes pasos.
 - i. Reclasificar con respecto a los niveles parent de la secuencia
 - ii. Se debe revisar que la secuencia no exista en la DB
 - iii. Si secuencia existe, se debe agregar la propiedad (en caso de que ya tenga una registrada)

- iv. Si secuencias no existe, agregar a la DB
- v. Sistema debe responder con un cuadro resumen que contenga
 - 1. Secuencias totales
 - 2. Secuencias ingresadas correctamente
 - 3. Secuencias rechazadas
 - 4. Número de actividades