Diseñador de sondas de hibridación

Simón Vergara

27 de noviembre de 2023

Contents:

CAPÍTULO 1

Probe-Designer

1.1 descarga module

Para obtener el ADN que se utilizará como referencia para el diseño de las sondas se va a descargar directamente el archivo genético desde la base de datos 'Nucleotide' de NCBI (National Center for Biotechnology Information) del NIH (National Institute of Health, EEUU) (https://www.ncbi.nlm.nih.gov/nuccore). Esta base de datos contiene millones de registros de secuencias de nucleótidos con mucha información adicional, como la especie, cromosoma, nombre del gen, exones, la fuente de los datos, etc.

Para llevar a cabo la descarga de los archivos genéticos, se implementó un módulo que permite realizar descargas desde esta base de datos a través de la librería Entrez de BioPython y almacenar los archivos en una carpeta. De esta manera se puede automatizar la obtención de las secuencias de referencia.

Por otra parte se implementó la función que retorna una secuencia a partir de un archivo almacenado localmente, además de crear una copia del archivo y almacenarla. Esta función sólo podrá recibir archivos de secuencia anotados como GenBank y GFF3.

El producto de estas funciones implementadas es la secuencia en formato SeqRecord, una estructura de datos propia de la librería BioPython que permite almacenar grandes secuencias con su respectiva notación, como por ejemplo las posiciones de cada uno de los exones, las transcripciones, secciones no codificantes, entre otros. Además contiene información extra de la secuencia tal como la especie, el cromosoma, la fuente de donde salieron los datos, etc.

descarga.**accnum_to_seqrecord**(*accesion_number*)

Función que recibe un accession number de la base de datos "nucleotide" de NCBI https://www.ncbi.nlm.nih.gov/nuccore. Genera el archivo .gbk (GenBank) en la carpeta "files" y retorna la secuencia en formato SeqRecord.

Parámetros
accesion_number – El accesion number del archivo en GenBank, por ejemplo NG_008617.1

Devuelve
La secuencia en formato SeqRecord con todas sus anotaciones.

descarga.**main**()

Ejecutar directamente el archivo "descarga" requiere entregar el parámetro –accessionnumber para descargar una secuencia mediante la función accnum_to_seqrecord.

Parámetros
--accesionnumber – El accesion number del archivo en GenBank, por ejemplo NG_008617.1

descarga.**parse_file_to_seqrecord**(*filepath*)

Función que recibe la ruta de un archivo genético anotado y lo retorna como SeqRecord. Genera una copia del achivo en la carpeta files. Solo se aceptan archivos en formato «GenBank», o sea ".gbk" y ".gb". Estos archivos pueden estar comprimidos en formato ".gz".

Parámetros
filepath – La ruta donde se encuentra el archivo que se quiere utilizar

Devuelve
La secuencia en formato SeqRecord con todas sus anotaciones.

1.2 diseno module

El módulo diseño contiene las funciones que permiten ejecutar el algoritmo diseñador y multiplexador de sondas. El diseño de las sondas de hibridación requiere de cierto procesamiento de datos para poder obtener el resultado adecuado, desde obtener las regiones de interés hasta verificar la validez de las sondas.

Este módulo está desarrollado para ser importado e utilizado por una interfaz de usuario.

diseno.**check_probe**(*seq*, *iscentral*, *minlen*, *maxlen*, *tmmin*, *tmmax*, *gcmin*, *gcmax*, *dgmin_homodim*, *dgmin_hairpin*, *maxhomopol_simple*, *maxhomopol_double*, *maxhomopol_triple*)

Función que recibe una sonda (secuencia de nucleótidos) y las restricciones para verificar que la sonda cumpla los parámetros establecidos. Retorna un valor booleano que determina si cumple las restricciones o no.

Parámetros

- **seq** – La secuencia que se quiere verificar.
- **iscentral** – Booleano que indica si la sonda es central (una mitad en un exón y la otra en el siguiente)
- **minlen** – El largo mínimo de la sonda.
- **maxlen** – El largo máximo de la sonda.
- **tmmin** – Temperatura de melting mínima.
- **tmmax** – Temperatura de melting máxima.
- **gcmin** – Porcentaje de GC mínimo.
- **gcmax** – Porcentaje de GC máximo.
- **dgmin_homodim** – Delta G mínimo permitido para validación de homodimerización.
- **dgmin_hairpin** – Delta G mínimo permitido para validación de hairpin u horquilla.
- **maxhomopol_simple** – Cantidad máxima permitida de homopolímeros simples (AAAAA)
- **maxhomopol_double** – Cantidad máxima permitida de homopolímeros dobles (AGAGAGAG)
- **maxhomopol_triple** – Cantidad máxima permitida de homopolímeros triples (AGCAGCAGC)

Devuelve
Booleano que determina la validez de la secuencia

diseno.**generate_xlsx**(*df*, *name*, *genes*, *minlen=60*, *maxlen=120*, *tmmin=65*, *tmmax=80*, *gcmin=30*, *gcmax=70*, *mindist=0*, *maxdist=50*, *minoverlap=25*, *maxoverlap=50*, *dgmin_homodim=-10000*, *dgmin_hairpin=-10000*, *maxhomopol_simple=6*, *maxhomopol_double=5*, *maxhomopol_triple=4*, *multiplex=True*, *mindg=-13627*, *maxdt=5*, *tiempo_diseno=0*)

Función que genera un reporte excel que contiene todas las sondas generadas y sus características. Además muestra las restricciones iniciales y los grupos de multiplexación.

Parámetros

- **df** – DataFrame de pandas que contiene las sondas y sus parámetros.
- **name** – Nombre que se le quiere dar al reporte.
- **minlen** – El largo mínimo de la sonda.
- **maxlen** – El largo máximo de la sonda.
- **tmmin** – Temperatura de melting mínima.
- **tmmax** – Temperatura de melting máxima.
- **gcmin** – Porcentaje de GC mínimo.
- **gcmin** – Porcentaje de GC máximo.
- **mindist** – Distancia mínima al borde del exón.
- **maxdist** – Distancia máxima al borde del exón.
- **minoverlap** – Sobrelape mínimo entre sondas.
- **maxoverlap** – Sobrelape máximo entre sondas.
- **dgmin_homodim** – Delta G mínimo permitido para validación de homodimerización.
- **dgmin_hairpin** – Delta G mínimo permitido para validación de hairpin u horquilla.
- **maxhomopol_simple** – Cantidad máxima permitida de homopolímeros simples (AAAAA)
- **maxhomopol_double** – Cantidad máxima permitida de homopolímeros dobles (AGAGAGAG)
- **maxhomopol_triple** – Cantidad máxima permitida de homopolímeros triples (AGCAGCAGC)
- **multiplex** – Booleano que indica si el diseño considera multiplexación
- **mindg** – Delta G mínimo para heterodimerización (multiplex)
- **maxdt** – Diferencia máxima de Tm (multiplex)
- **tiempo_diseno** – Tiempo que demoró la ejecución en segundos

Devuelve

Nombre del archivo XLSX generado

diseno.**get_all_genes**(*seqrecord*)

Obtener todos los genes anotados

Parámetros

seqrecord – Subregiones que representan la secuencia codificante o secuencia de exones

Devuelve

Lista con pares de posiciones en una tupla: el fin de un exón y el inicio del siguiente.

diseno.**get_all_transcripts**(*seqrecord*)

Función que retorna todos los transcritos anotados dentro de una secuancia.

Parámetros
seqrecord – Subregiones que representan la secuencia codificante o secuencia de exones

Devuelve
Lista con pares de posiciones en una tupla: el fin de un exón y el inicio del siguiente.

diseno.**get_probes_from_pos**(*empalme*, *record*, *site*, *minlen*, *maxlen*, *tmmin*, *tmmax*, *gcmin*, *gcmax*, *mindist*, *maxdist*, *minoverlap*, *maxoverlap*, *dgmin_homodim*, *dgmin_hairpin*, *maxhomopol_simple*, *maxhomopol_double*, *maxhomopol_triple*)

Esta función genera un par de sondas a partir de la secuencia de referencia y dos posiciones que representan el punto de empalme entre dos exones. El par de sondas generadas pueden ubicarse en el exón donor, acceptor o en ambos (central).

Parámetros

- **empalme** – Tupla con posiciones que representan al punto de empalme o splicing.
- **record** – Secuencia de referencia.
- **site** – Entero que representa si las sondas son donor, acceptor o central.
- **minlen** – El largo mínimo de la sonda.
- **maxlen** – El largo máximo de la sonda.
- **tmmin** – Temperatura de melting mínima.
- **tmmax** – Temperatura de melting máxima.
- **gcmin** – Porcentaje de GC mínimo.
- **gcmin** – Porcentaje de GC máximo.
- **mindist** – Distancia mínima al borde del exón.
- **maxdist** – Distancia máxima al borde del exón.
- **minoverlap** – Sobrelape mínimo entre sondas.
- **maxoverlap** – Sobrelape máximo entre sondas.
- **dgmin_homodim** – Delta G mínimo permitido para validación de homodimerización.
- **dgmin_hairpin** – Delta G mínimo permitido para validación de hairpin u horquilla.
- **maxhomopol_simple** – Cantidad máxima permitida de homopolímeros simples (AAAAA)
- **maxhomopol_double** – Cantidad máxima permitida de homopolímeros dobles (AGAGAGAG)
- **maxhomopol_triple** – Cantidad máxima permitida de homopolímeros triples (AGCAGCAGC)

Devuelve
Lista con dos sondas, cada sonda es una tupla con la secuencia, la posición de inicio, la posición final y la distancia al borde del exón si aplica.

diseno.**get_splicing_pairs**(*locations*)

Función que obtiene las zonas de empalme a partir de una serie de regiones que representan una transcripción. Las zonas de empalme se representan por dos posiciones: el fin de un exón y el inicio del siguiente.

Parámetros
locations – Subregiones que representan la secuencia codificante o secuencia de exones

Devuelve
Lista con pares de posiciones en una tupla: el fin de un exón y el inicio del siguiente.

diseno.**get_splicings**(*seqrecord*, *transcripts*)

Función que retorna los empalmes sin repetición de una secuencia anotada en ciertos transcritos.

Parámetros

- **seqrecord** – Secuencia anotada de referencia
- **transcripts** – Lista de transcritos

Devuelve
Lista con pares de posiciones en una tupla: el fin de un exón y el inicio del siguiente.

diseno.**main**()

diseno.**probe_designer**(*record*, *transcripts*, *progreso_queue*, *minlen=60*, *maxlen=120*, *tmmin=65*, *tmmax=80*, *gcmin=30*, *gcmax=70*, *mindist=0*, *maxdist=200*, *minoverlap=25*, *maxoverlap=50*, *dgmin_homodim=-10000*, *dgmin_hairpin=-10000*, *maxhomopol_simple=6*, *maxhomopol_double=5*, *maxhomopol_triple=4*, *multiplex=True*, *mindg=-13627*, *maxdt=5*)

Función principal diseñadora de sondas. De la secuencia anotada se obtiene una serie de sondas como subsecuencias de ésta, que cumplen varias restricciones ajustadas por los parámetros.

Parámetros

- **record** – Secuencia anotada de referencia para la creación de las sondas.
- **transcripts** – Transcripciones en las cuales se quiere diseñar.
- **minlen** – El largo mínimo de la sonda.
- **maxlen** – El largo máximo de la sonda.
- **tmmin** – Temperatura de melting mínima.
- **tmmax** – Temperatura de melting máxima.
- **gcmin** – Porcentaje de GC mínimo.
- **gcmin** – Porcentaje de GC máximo.
- **mindist** – Distancia mínima al borde del exón.
- **maxdist** – Distancia máxima al borde del exón.
- **minoverlap** – Sobrelape mínimo entre sondas.
- **maxoverlap** – Sobrelape máximo entre sondas.
- **dgmin_homodim** – Delta G mínimo permitido para validación de homodimerización.
- **dgmin_hairpin** – Delta G mínimo permitido para validación de hairpin u horquilla.
- **maxhomopol_simple** – Cantidad máxima permitida de homopolímeros simples (AAAAA)
- **maxhomopol_double** – Cantidad máxima permitida de homopolímeros dobles (AGAGAGAG)
- **maxhomopol_triple** – Cantidad máxima permitida de homopolímeros triples (AGCAGCAGC)
- **multiplex** – Booleano que indica si el diseño considera multiplexación
- **mindg** – Delta G mínimo para heterodimerización (multiplex)

- **maxdt** – Diferencia máxima de Tm (multiplex)

Devuelve
Dataframe con las sondas y sus características

diseno.**segundos_a_hms**(*segundos*)

diseno.**verify_specificity**(*seq*, *iscentral*)

Función que verifica la especificidad o unicidad de una secuencia usando la herramienta de alineamiento BLAST de manera local. Es necesario tener instalado Blast+ y las bases de datos en la carpeta "refseq".

Parámetros

- **seq** – Secuencia de nuecleótidos que se quiere validar.
- **iscentral** – Booleano que indica si se trata de una sonda central. Se utilizan diferentes criterios para chequear la especificidad de una sonda central.

Devuelve
Booleano que indica si la secuencia se considera específica.

1.3 multiplex module

El módulo multiplex contiene las funciones que utiliza el módulo diseñador para poder llevar a cabo la multiplexación de las sondas diseñadas.

multiplex.**are_sequences_compatible**(*seq1*, *seq2*, *mindg*, *maxdt*)

Esta función determina si dos secuencias de ADN son compatibles. Debes personalizar las condiciones de compatibilidad según tus necesidades.

Parámetros

- **seq1** – Primera secuencia de ADN.
- **seq2** – Segunda secuencia de ADN.

Devuelve
True si son compatibles, False en caso contrario.

multiplex.**multiplex_sequences**(*sequences*, *progress_queue*, *mindg=-13627*, *maxdt=5*)

Función principal de multiplexación de sondas. Recibe las secuencias, los parámetros de multiplexación y la cola para informar progreso.

Parámetros

- **sequences** – Lista con todas las diferentes secuencias.
- **progress_queue** – Cola para informar el progreso a la interfaz.
- **mindg** – Mínimo delta G heterodimerización para que dos secuencias sean compatibles.
- **maxdt** – Máxima diferencia de Tm para que dos secuencias sean compatibles.

Devuelve
True si son compatibles, False en caso contrario.

CAPÍTULO 2

Índices y tablas

- genindex
- modindex
- search

Índice de Módulos Python

Índice