Avance Nº 02

ALUMNO	PROGRAMA	CURSO
MSc. Vicente Enrique Machaca Arceda	Doctorado en Ciencias de la Computación	Tópicos en Computación Gráfica

AVANCE	TEMA	FECHA
02	Predicción y modelado en 3D de estructuras terciarias de proteinas a partir del *contact map*	30-01-2021

1. Introducción

Las proteínas son moléculas complejas que cumplen un rol crítico en nuestro cuerpo, estas cumplen la mayoria de funciones en la células (Anderson and Anderson, 1998). Además, la función de una proteína depende de su estructura (Rangwala and Karypis, 2010) y ultimanmente se ha descubierto que esta función tambien depende de las relación de una proteina con otras (Canzar and Ringeling, 2020). Mas aún, es importante saber, que la estructura de una proteina puede cambiar en el tiempo y su función tambien cambia en el tiempo.

Conocer la estructura de una proteína es de suma importancia para el análisis de su función, generación de medicamentos, etc. (Rangwala and Karypis, 2010). Además, lograr predecir y entender el funcionamiento de estas proteinas y la interacción de redes de proteínas es considerado el nuevo santo crial de la Binformática en estos tiempos (Srihari et al., 2017).

2. Conceptos previos

En esta sección detallaremos algunos conceptos previos del area de Bioinformática/*Proteomics* para comprender el trabajo.

2.1. Estructura de las proteinas

Existen 4 tipos de estructuras de proteínas (Russell and Gordey, 2002):

1. **Estructura primaría:** Secuencia de aminoacidos (ver Figura 1 (a)).
2. **Estructura secundaría:** Pequeños patrones, los mas comunes son las elises α y hojas β (ver Figura 1 (b)).
3. **Estructura terciaría:** Representa la unión de los segmentos de la estructura secundaría (ver Figura 1 (c)). En este caso solo estamos considerando una cadena de aminoacidos (las proteinas a veces son conformadas por varias cadenas de aminoacidos).
4. **Estructura cuaternaría:** Union de varias estructuras terciarias (varias cadenas de aminoacidos) (ver Figura 1 (d)).

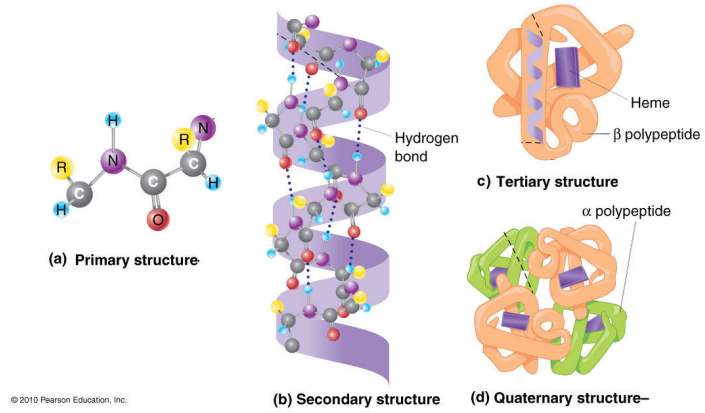

Figura 1: Ejemplo de las 4 estructuras de proteínas. Fuente: (Russell and Gordey, 2002)

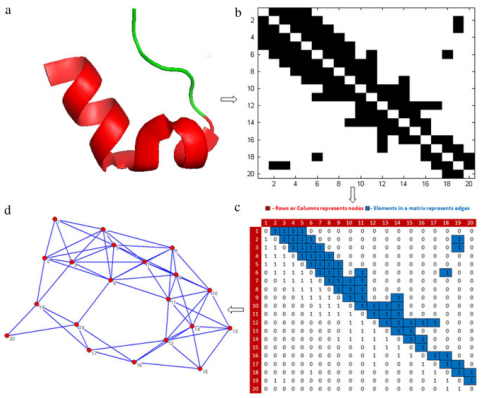

Figura 2: Ejemplo del contact map de una proteína.

2.2. *Contact map*

Representa la distancia de cada posible aminoacido, cuando forman proteínas. El *contact map*, es representado como un gráfico en 2D, y es el elegido por los modelos de machine learning en la predicción de las estructuras de proteínas. En la Figura 2, mostramos como es un *contact map*.

3. Avances

En esta sección, detallaremos el estado anterior del proyecto y los avances realizados estas dos ultimas semanas.

3.1. Estado anterior

Se definio la propuesta del trabajo y se reviso el *paper* propuesto por Adhikari and Cheng (2018), el cúal propone reconstruir la estructura terciaria de una proteína a partir del *contact map*.

3.2. Avance y estado actual

1. Se reviso a detalle el *paper* propuesto por Adhikari and Cheng (2018), el cúal propone la herramienta CONFOLD2.

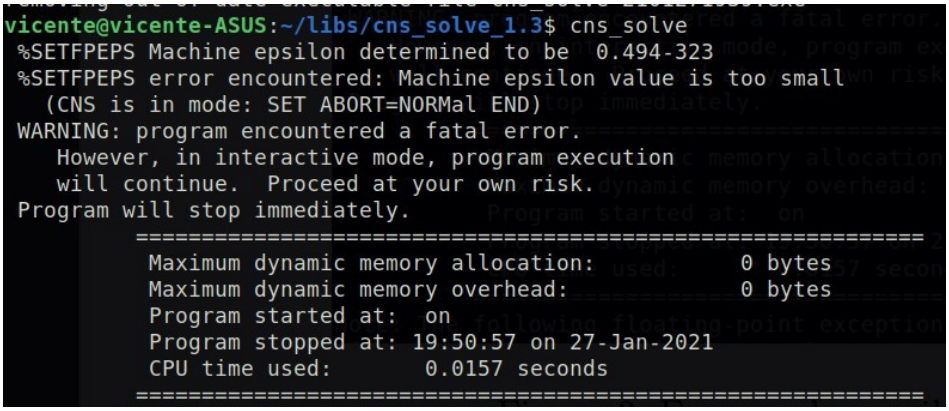

Figura 3: Errores al compilar y ejecutar la herramienta CNS.

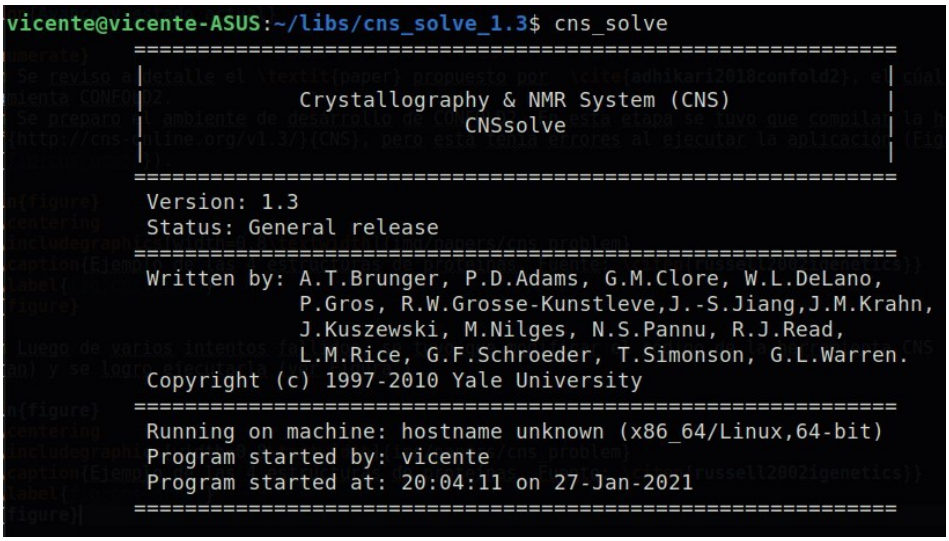

Figura 4: Se logro compilar sin errores la herramienta CNS.

2. Se preparo el ambiente de desarrollo de CONFOLD2. En esta etapa se tuvo que compilar la herramienta CNS, pero esta tenia errores al ejecutar la aplicación (ver Figura 3).

3. Luego de varios intentos fallidos, se tuvo que modificar el código de la herramienta CNS (implementada en Fortran) y se logro ejecutarla (ver Figura 4).

4. Se clono el repositorio de CONFOLD2 desde Github y se probó su código de ejemplo. Este proceso, empieza a generar archivos .pbd que representan posibles estructuras de la proteína, En la Figura 5 mostramos algunas de estas estructuras.

5. El tiempo de procesamiento de CONFOLD2, tomo alrededor de una hora aproximadamente, pero al final, en las ultimas etapas mostro errores. En la Figura 6 y 7 mostramos los errores.

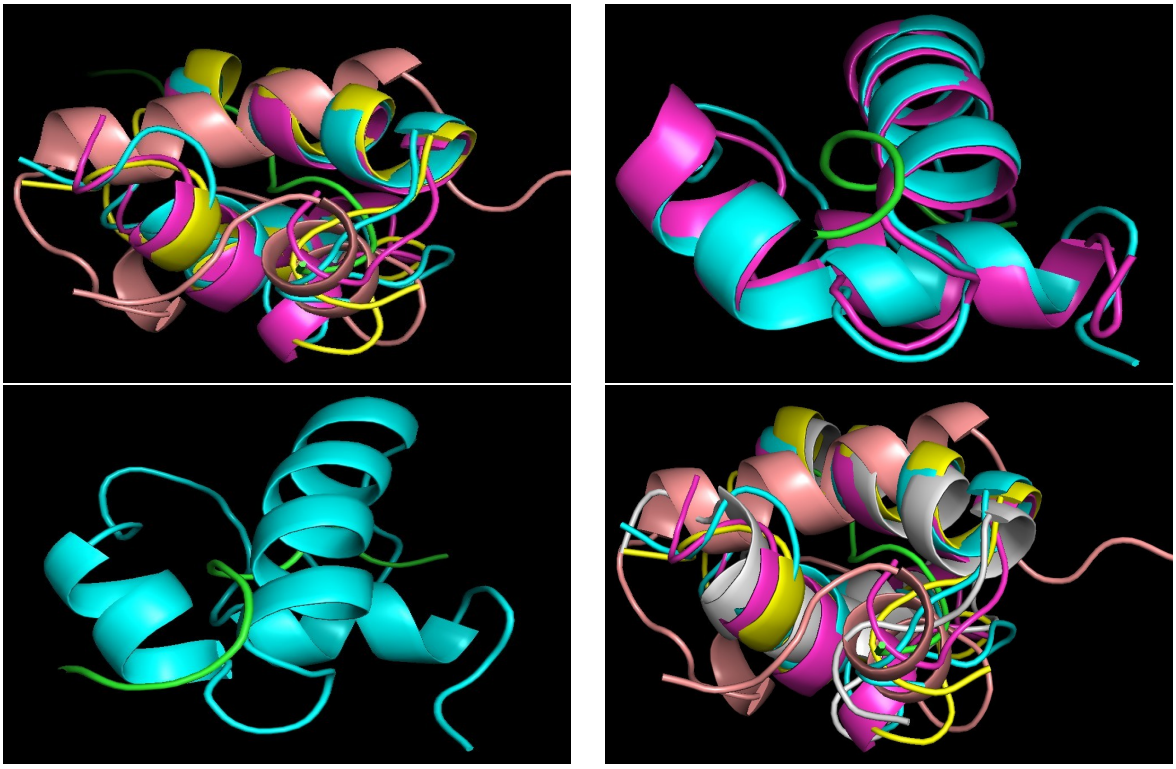

Figura 5: Ejemplo de la construcción de la proteína con CONFOLD2.

```
Forking CONFOLD job with top 3.8L contacts
Log file at -> /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-3.8L/
dando permisos ... hecho
ejecutando job ... hecho

Forking CONFOLD job with top 3.9L contacts
Log file at -> /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-3.9L/
dando permisos ... hecho
ejecutando job ... hecho

Forking CONFOLD job with top 4.0L contacts
Log file at -> /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-4.0L/
dando permisos ... hecho
ejecutando job ... hecho

Wait for all CONFOLD jobs to finish (check individual log files for progress)..
sh: 1: /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/third-party-programs/dssp-2.0.4-linux-amd64: Permission denied
/home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-2.9L/stage1/1guu_13.pdb seems empty! at /home/vicente/projects/B
IOINFORMATICS/CONFOLD2/confold-v2.0/core.pl line 2264.
        main::dssp_result("/home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/o"..., "ss") called at /home/vicente/projects/BIOINFO
RMATICS/CONFOLD2/confold-v2.0/core.pl line 1605
        main::count_ss_match("/home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/o"..., "1guu.ss", "H") called at /home/vicente/pro
jects/BIOINFORMATICS/CONFOLD2/confold-v2.0/core.pl line 2419
        main::assess_dgsa("stage1") called at /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/core.pl line 152
sh: 1: /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/third-party-programs/dssp-2.0.4-linux-amd64: Permission denied
/home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-3.3L/stage1/1guu_13.pdb seems empty! at /home/vicente/projects/B
IOINFORMATICS/CONFOLD2/confold-v2.0/core.pl line 2264.
        main::dssp_result("/home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/o"..., "ss") called at /home/vicente/projects/BIOINFO
RMATICS/CONFOLD2/confold-v2.0/core.pl line 1605
        main::count_ss_match("/home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/o"..., "1guu.ss", "H") called at /home/vicente/pro
jects/BIOINFORMATICS/CONFOLD2/confold-v2.0/core.pl line 2419
        main::assess_dgsa("stage1") called at /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/core.pl line 152
```

Figura 6: Error al finalizar la ejecución de CONFOLD2.

```
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-3.9L/stage2/1guu_model3.pdb not found! at
./confold2-main.pl line 174.
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-3.9L/stage2/1guu_model4.pdb not found! at
./confold2-main.pl line 174.
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-3.9L/stage2/1guu_model5.pdb not found! at
./confold2-main.pl line 174.
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-4.0L/stage2/1guu_model1.pdb not found! at
./confold2-main.pl line 174.
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-4.0L/stage2/1guu_model2.pdb not found! at
./confold2-main.pl line 174.
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-4.0L/stage2/1guu_model3.pdb not found! at
./confold2-main.pl line 174.
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-4.0L/stage2/1guu_model4.pdb not found! at
./confold2-main.pl line 174.
Oops!! Expected model /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-4.0L/stage2/1guu_model5.pdb not found! at
./confold2-main.pl line 174.

Oops!! Only 0 models found in mkdir -p /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models [expected = 200] a
t ./confold2-main.pl line 192.
```

Figura 7: Error al finalizar la ejecución de CONFOLD2.

6. Despues de revisar los errores, estos se debian a que algunos programas no tenian permisos de ejecución. Se solucionó esto y se logró culminar el procesamiento de CONFOLD2. En la Figura 8, mostromas los mensajes retornados por CONFOLD, en este vemos que me indica las 5 mejores estructuras encontradas.

```
Run clustering with the updated list and pick 5 centroids..
Added /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-1.3L-model-2.pdb to top 5 list
Added /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-3.2L-model-1.pdb to top 5 list
Added /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-1.9L-model-2.pdb to top 5 list
Added /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-0.8L-model-5.pdb to top 5 list
Added /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-3.8L-model-3.pdb to top 5 list

Rank the 4 models selected [expected = 5] ..
coping /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-1.3L-model-2.pdb as model1.pdb
coping /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-3.2L-model-1.pdb as model2.pdb
coping /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-1.9L-model-2.pdb as model3.pdb
coping /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-0.8L-model-5.pdb as model4.pdb
coping /home/vicente/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0/output-1guu/top-models/top-3.8L-model-3.pdb as model5.pdb

Finished [./confold2-main.pl]: Thu Feb  4 07:57:41 2021
vicente@vicente-ASUS:~/projects/BIOINFORMATICS/CONFOLD2/confold-v2.0$
```

Figura 8: CONFOLD2 termina de generar las estructuras.

7. Despues de revisar los errores, estos se debian a que algunos programas no tenian permisos de ejecución. Se soluciono esto y se logro culminar el procesamiento de CONFILD2. En la Figura 9, las 4 mejores estructuras encontradas por CONFOLD2.

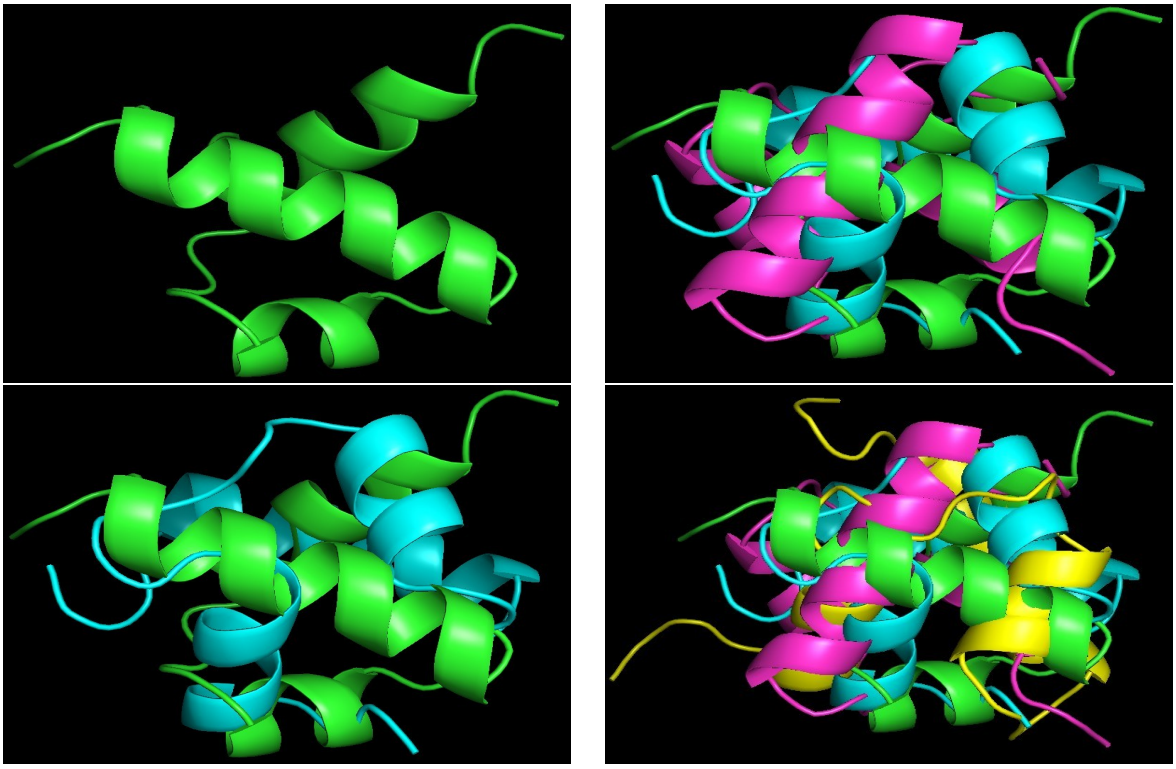

Figura 9: Ejemplo de la construcción de la proteína con CONFOLD2.

Referencias

Adhikari, B. and Cheng, J. (2018). Confold2: improved contact-driven ab initio protein structure modeling. *BMC bioinformatics*, 19(1):22.

Anderson, N. L. and Anderson, N. G. (1998). Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11):1853–1861.

Canzar, S. and Ringeling, F. R. (2020). Protein-protein interaction networks.

Rangwala, H. and Karypis, G. (2010). Introduction to protein structure prediction. *Introduction to Protein Structure Prediction*, 58.

Russell, P. J. and Gordey, K. (2002). *IGenetics*. Number QH430 R87. Benjamin Cummings San Francisco.

Srihari, S., Yong, C. H., and Wong, L. (2017). *Computational prediction of protein complexes from protein interaction networks*. Morgan & Claypool.