

Avance N° 03

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

AVANCE	TEMA	FECHA
03	Predicción y modelado en 3D de estructuras terciarias de proteínas a partir del <i>contact map</i>	20-02-2021

1. Introducción

Las proteínas son moléculas complejas que cumplen un rol crítico en nuestro cuerpo, estas cumplen la mayoría de funciones en la células (Anderson and Anderson, 1998). Además, la función de una proteína depende de su estructura (Rangwala and Karypis, 2010) y ultimamente se ha descubierto que esta función también depende de la relación de una proteína con otras (Canzar and Ringeling, 2020). Mas aún, es importante saber, que la estructura de una proteína puede cambiar en el tiempo y su función también 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 proteínas y la interacción de redes de proteínas es considerado el nuevo santo grial de la Bioinformática en estos tiempos (Srihari et al., 2017).

2. Conceptos previos

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

2.1. Estructura de las proteínas

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

1. **Estructura primaria:** Secuencia de aminoácidos (ver Figura 1 (a)).
2. **Estructura secundaria:** Pequeños patrones, los más comunes son las hélices α y hojas β (ver Figura 1 (b)).
3. **Estructura terciaria:** Representa la unión de los segmentos de la estructura secundaria (ver Figura 1 (c)). En este caso solo estamos considerando una cadena de aminoácidos (las proteínas a veces son conformadas por varias cadenas de aminoácidos).
4. **Estructura cuaternaria:** Unión de varias estructuras terciarias (varias cadenas de aminoácidos) (ver Figura 1 (d)).

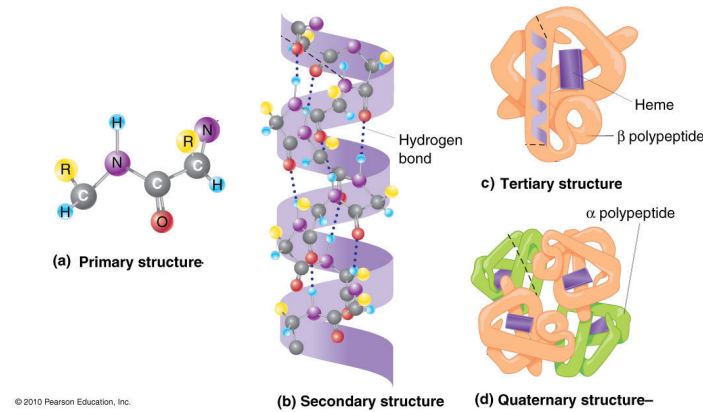


Figura 1: Ejemplo de las 4 estructuras de proteínas. Fuente: (Russell and Gordey, 2002)

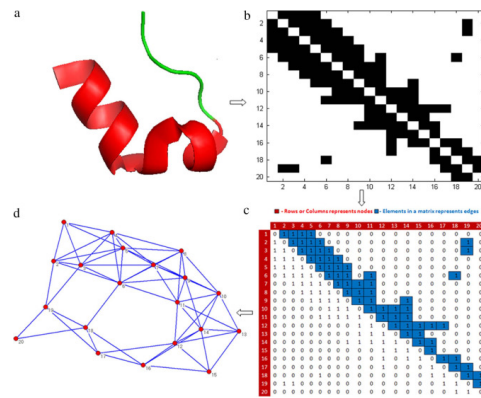


Figura 2: Ejemplo del contact map de una proteína.

2.2. Contact map

Representa la distancia de cada posible aminoácido, 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 cual 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 cual propone la herramienta CONFOLD2.

```
vicente@vicente-ASUS:~/libs/cns_solve_1.3$ cns_solve
%SETFPEPS Machine epsilon determined to be 0.494-323
%SETFPEPS error encountered: Machine epsilon value is too small
(CNS is in mode: SET ABORT=NORMAL END)
WARNING: program encountered a fatal error.
However, in interactive mode, program execution
will continue. Proceed at your own risk.
Program will stop immediately.
Program started at: on
=====
Maximum dynamic memory allocation: used: 0 bytes
Maximum dynamic memory overhead: 0 bytes
Program started at: on
Program stopped at: 19:50:57 on 27-Jan-2021
CPU time used: 0.0157 seconds
=====
```

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

```
vicente@vicente-ASUS:~/libs/cns_solve_1.3$ cns_solve
=====
Crystallography & NMR System (CNS)
CNSsolve
=====
Version: 1.3
Status: General release
=====
Written by: A.T.Brunger, P.D.Adams, G.M.Clore, W.L.DeLano,
P.Gros, R.W.Grosse-Kunstleve, J.-S.Jiang, J.M.Krahn,
J.Kuszewski, M.Nilges, N.S.Pannu, R.J.Read,
L.M.Rice, G.F.Schroeder, T.Simonson, G.L.Warren.
Copyright (c) 1997-2010 Yale University
=====
Running on machine: hostname unknown (x86_64/Linux,64-bit)
Program started by: vicente
Program started at: 20:04:11 on 27-Jan-2021
=====
```

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 pequeñas porciones de la estructura de una proteína, En la Figura 5 mostramos algunas de estas estructuras.
5. El tiempo de procesamiento de CONFOLD2, tomo alrededor de 30 minutos, pero al final en las ultimas etapas mostro errores (al parecer en esta etapa uniría las pequeñas estructuras generadas anteriormente). En la Figura 6 y 7 mostramos los errores, aún se está trabajando en la solución.

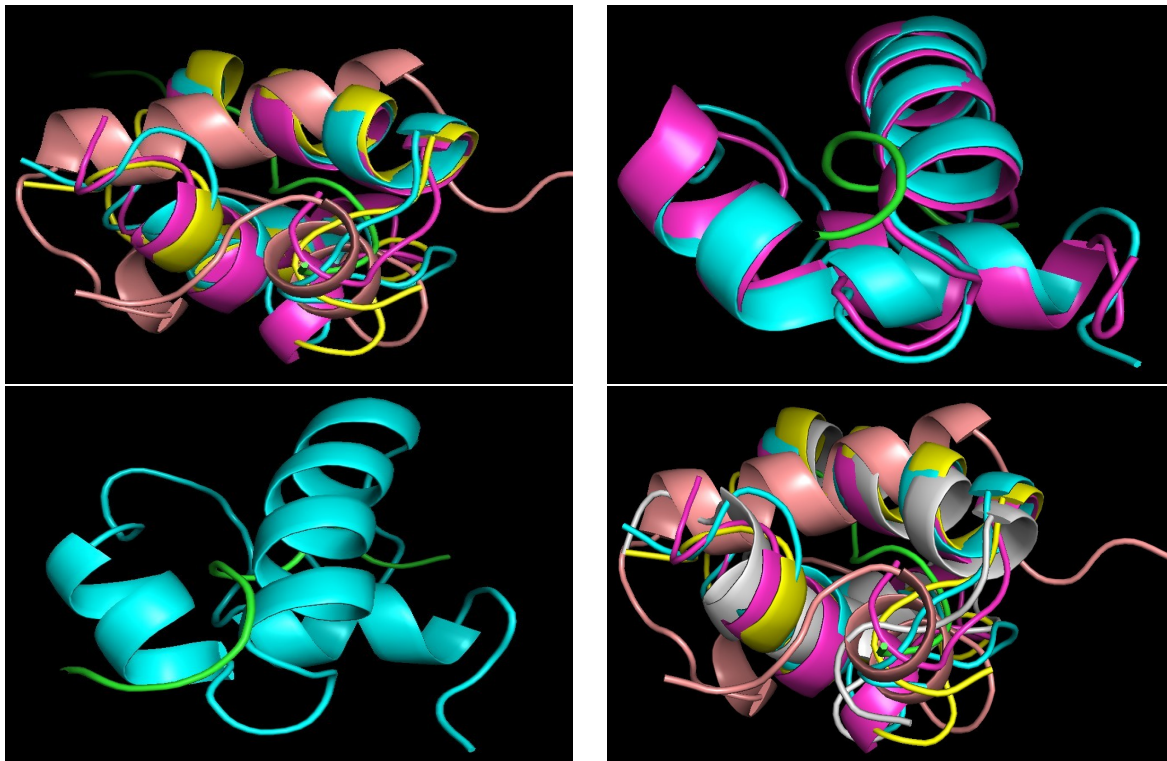


Figura 5: Ejemplo de la construcción de la proteína con CONFOLD2.

```

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.

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.