

Efecto del pH en la condición de cristalización sobre el daño por radiación en cristales de proteína

Francisco Murphy Pérez

4 de octubre de 2020

Colofón

Este documento se compuso con la ayuda de: KOMA-Script (<https://sourceforge.net/projects/koma-script/>) y L^AT_EX (<https://www.latex-project.org/>), usando la plantilla denominada kaobook (<https://github.com/fmarotta/kaobook/>).

A Lucía, por supuesto.

Resumen

Índice general

Índice general	vi
Acrónimos	ix
1 Introducción	1
1.1 Cristalográfia de rayos X	1
1.2 Daño por radiación	2
Origen	3
Clasificación	3
Consecuencias	3
Métricas	4
1.3 Crioprotección	5
1.4 Sincrotrones	5
1.5 Radioprotectores	6
2 Antecedentes	7
2.1 pH	7
3 Objetivo	9
3.1 Objetivos particulares	9
4 Materiales y métodos	10
Extracción de datos	10
Limpieza de datos	10
4.1 Colecta y análisis de datos	12
A Proteínas más representadas	13
B Análisis visual	14
Bibliografía	15

Índice de figuras

1.1 Patrón de difracción	2
1.2 Diferencia de densidad electrónica	4
2.1 Movimiento de cargas en la cadena polipeptídica	7
B.1 Análisis visual	14

Índice de tablas

1.1 Número de estructuras por método experimental	1
4.1 Proteínas que cristalizan en un intervalo amplio de pH	12
A.1 Las 50 proteínas más representadas	13

Acrónimos

C

CCD Charge-coupled device. 1

CRX Cristalografía de rayos X. 1, 2, 4, 5

P

PDB Protein Data Bank. 1

X

XFEL X-ray Free Electron Laser. 5

1

Introducción

1.1. Cristalografía de rayos X

La cristalografía de rayos X (CRX) es el método experimental más común para obtener la estructura tridimensional de una molécula¹ (Tabla 1.1). En general, los modelos estructurales de las macromoléculas determinadas por medio de cualquier método experimental, se depositan en el banco de datos de proteínas (PDB, por sus siglas en inglés) [1]. El repositorio digital del PDB, de libre acceso, se encuentra en el siguiente enlace <https://www.rcsb.org/>.

Método experimental	Estructuras	Porcentaje (%)
Cristalografía de rayos X	147020	88.88
Resonancia magnética nuclear	12937	7.82
Criomicroscopía electrónica	5037	3.04
Suma	164944	99.74

La teoría de la CRX se explica brevemente a continuación. La energía de los rayos X se puede transferir a los electrones de las moléculas que conforman el cristal. Si la transferencia de energía se da de manera elástica, los electrones oscilarán con la misma frecuencia que la onda de rayos X incidente. Esto, según la electrodinámica clásica, resulta en una nueva emisión de rayos X que a su vez pueden interferir entre sí, de forma destructiva o constructiva. Esta interferencia da lugar al concepto físico conocido como difracción. Si la diferencia entre las fases de las ondas de estos nuevos rayos X es exactamente igual a $n2\pi$ radianes, donde n es un número entero, la interferencia será constructiva. Fue William Lawrence Bragg, quien interpretó la difracción observada como una reflexión² de los rayos X por distintos planos dentro del cristal. Dada la estructura repetitiva del cristal, en general³, toda interferencia constructiva será amplificada y se podrá observar como un punto discreto en el patrón de difracción. Por otra parte, simultáneamente se dará la difracción por distintas familias de planos dentro del cristal, dando su forma final al patrón de difracción (Figura 1.1).

El experimento de CRX es relativamente simple y consiste en:

1. Incidir rayos X sobre el cristal de la macromolécula de interés.
2. Obtener el patrón de difracción.
3. Rotar el cristal en cierto eje.
4. Repetir los pasos anteriores n^4 veces.

Cabe resaltar algunos puntos: (i) Los rayos X son difractados dentro del cristal macromolecular; el producto final, es decir, el patrón de difracción, se obtiene gracias a un detector de fotones, conocido como dispositivo de carga acoplada (CCD, por sus siglas en inglés). El patrón de difracción contiene información de la estructura macromolecular, por lo que es

1: En este proyecto nos atañen las macromoléculas, en particular las proteínas.

[1]: Berman y col. (2000), «The protein data bank»

Tabla 1.1: Número de estructuras depositadas en el PDB por método experimental. La suma de estos tres métodos experimentales, representa el 99.74 % del total de estructuras depositadas. Fuente: búsqueda avanzada en el PDB por método experimental <https://www.rcsb.org/search/advanced>. Actualizada al 21 de junio del 2020.

2: Es tan usual este enfoque por su simplicidad que de manera tradicional y errónea los puntos en un patrón de difracción se denominan *reflexiones*. En este texto se utilizará el término *puntos de difracción* o simplemente *puntos*.

3: Existen ciertas condiciones de simetría que producen la *extinción* total de un punto de difracción.

4: En realidad

necesario mantener una copia digital de cada patrón de difracción para su posterior análisis. (ii) El cristal está montado sobre un goniómetro. Normalmente su rotación es perpendicular a la dirección del haz de rayos X. Se especifica tanto el incremento de rotación, denominado $\Delta\varphi$, como el intervalo total de rotación. Si $\Delta\varphi = 1^\circ$, es decir, después de cada exposición el cristal se rota un grado, un enfoque típico, entonces el intervalo de rotación estará dado por $(\varphi_{\text{final}} - \varphi_{\text{inicial}}) + 1$. (iii) En general, n está dado por la simetría del cristal; a mayor simetría, menor n [2].

[2]: Dauter (1999), «Data-collection strategies»

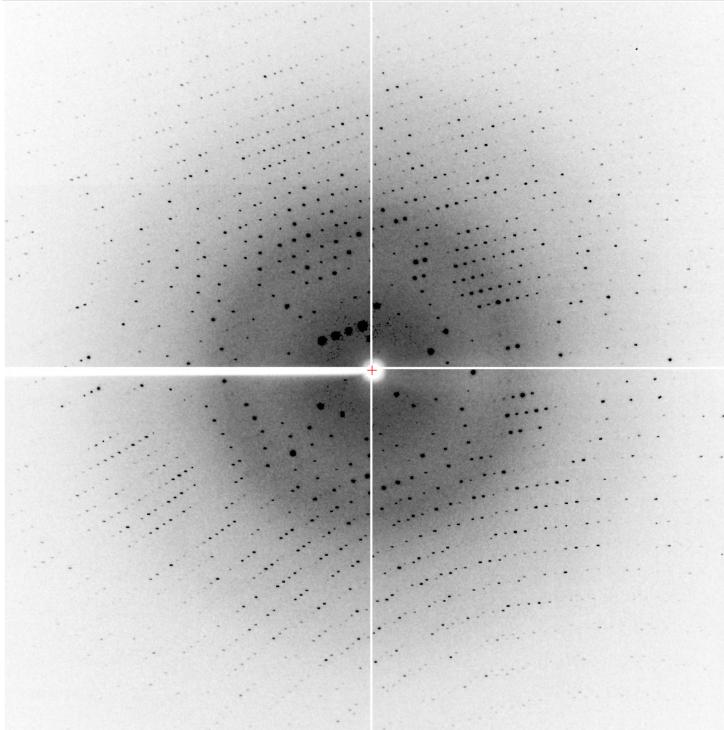


Figura 1.1: Patrón de difracción de la lisozima de la clara de huevo de gallina. Disponible en el siguiente enlace tinyurl.com/ydfw6asv. La resolución es la métrica principal que determina la calidad del modelo estructural de una proteína. Anillos concéntricos en el patrón de difracción, representan lo que se conoce como fajas de resolución. Las fajas de alta/baja resolución corresponden a anillos más externos/internos.

El objetivo del experimento de la CRX, es obtener un «dataset» completo, es decir, suficientes patrones de difracción para los pasos subsecuentes.

1.2. Daño por radiación

Una de las limitantes de la CRX, es el daño que causan los rayos X sobre el cristal macromolecular. Esto se conoce como daño por radiación y provoca que cada cristal macromolecular presente un límite temporal, denominado tiempo de vida útil, bajo el haz de rayos X. El daño por radiación se da porque los rayos X usados tienen una energía relativamente alta⁵. En el caso de un cristal macromolecular, su estabilidad física se da por pocas interacciones no covalentes por unidad de volumen, en comparación con un cristal de sal, por ejemplo. Por esto, su desintegración no requiere de mucha energía⁶. Por otro lado, es evidente que al perderse el orden cristalino, se pierde la amplificación del proceso de difracción y en consecuencia los patrones de difracción cada vez contienen menos información. En otras palabras, la calidad del cristal decae y por lo tanto la calidad de cada patrón de difracción obtenido. El daño por radiación es la principal causa de que sea difícil obtener un «dataset» completo.

5: Más información en la siguiente sección.

6: Incluso su manipulación tiene que darse con extrema precaución.

Origen

El daño por radiación se da porque los electrones de las moléculas que conforman el cristal, absorben la energía de los fotones incidentes. La absorción tiene su causa en uno de dos fenómenos físicos: el efecto fotoeléctrico o la dispersión inelástica. La probabilidad de que se dé el primero es un orden de magnitud mayor que el segundo. El efecto fotoeléctrico consiste en la absorción total de un fotón por un electrón. Como consecuencia el electrón es expulsado de su orbital dejando una vacante electrónica, o hueco positivo (h^+), en la molécula que lo contenía. La energía de los electrones liberados, llamados fotoelectrones, se disipa en la trayectoria que estos hayan tomado; generando miles⁷ de iones, radicales libres y eventos de excitación molecular. Los radicales son especies químicas que poseen uno o más electrones desapareados, por ende su reactividad es muy alta y su tiempo de vida es particularmente corto. Una reacción en cadena de radicales libres es inminente. Si algún radical, o cualquier especie química producto de la radiación, llegase a perturbar la red de contactos cristalinos, se pierde el orden cristalino.

7: El promedio de la primera energía de ionización para los átomos presentes en una proteína es de 12.05 eV, la energía de un fotón con una longitud de onda de 1 Å es de 12 398.4 eV.

Clasificación

El daño por radiación se clasifica de acuerdo con su escala temporal. El daño primario es la ionización dada por el efecto fotoeléctrico. El daño secundario es la subsecuente cascada de radicales libres, dependiente del tiempo y de la temperatura. El daño terciario se define como el daño macroscópico sobre el cristal⁸. Esto implica que una fracción suficiente de macromoléculas dentro del cristal ha sido afectada por el daño primario y secundario [3].

8: Fisuras o cambios en su coloración, por ejemplo.

[3]: Teng y col. (2000), «Primary radiation damage of protein crystals by an intense synchrotron X-ray beam»

Consecuencias

Las consecuencias del daño por radiación se observan en: (i) la disminución de la intensidad de los puntos en el patrón de difracción, sobre todo aquellos en las fajas de alta resolución; (ii) un cambio del volumen de la celda unitaria, que causa la pérdida de la isomorfía cristalina; (iii) el aumento en los parámetros de desplazamiento atómico; y (iv) el empeoramiento de las medidas que indican la calidad global de los datos [3]. El listado anterior se conoce como daño por radiación global. Por otra parte, el daño específico se refiere al daño estructural en la macromolécula cristalizada. Esto conlleva un orden dado: primero ocurre la reducción de átomos metálicos; después se da la ruptura de puentes disulfuro; luego la descarboxilación de aspartatos y glutamatos; y finalmente se pierde el grupo tiometilo de las metioninas [4, 5]. No todos los residuos de aminoácido susceptibles son afectados de la misma manera. Hasta el momento, las razones de esto no han sido esclarecidas. Por este motivo, aunque existen ciertos principios básicos para determinar el daño específico, es difícil predecirlo y saber de antemano el grado en que afectará el modelo estructural obtenido. En el peor escenario, puede ser imposible obtener una estructura macromolecular debido a la inherente susceptibilidad del cristal al daño por radiación o a la pérdida de isomorfía cristalina.

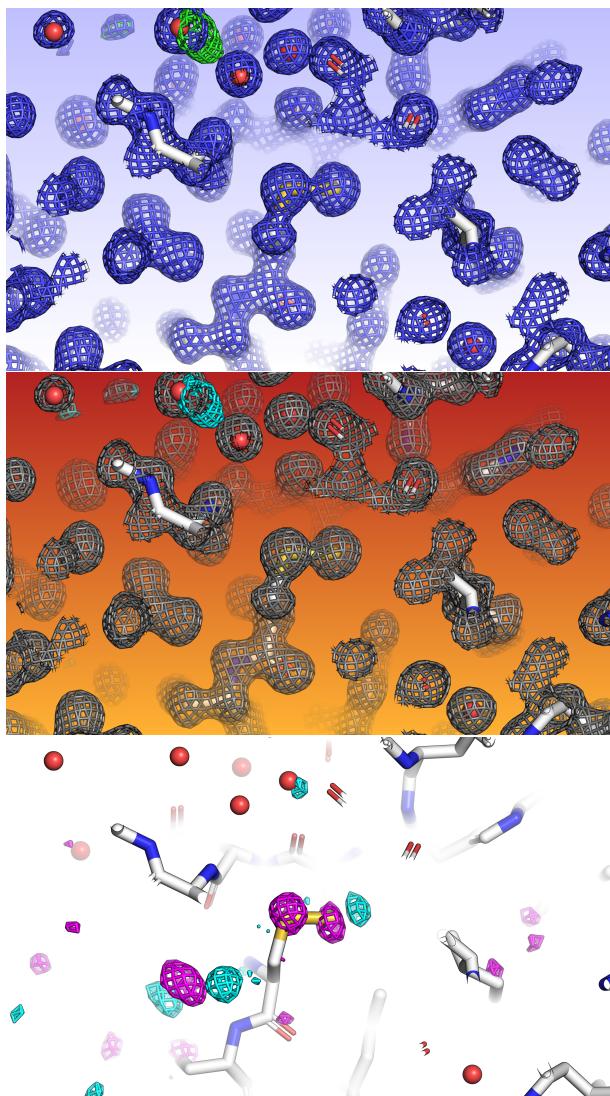
[4]: Weik y col. (2000), «Specific chemical and structural damage to proteins produced by synchrotron radiation»

[5]: Ravelli y col. (2000), «The ‘fingerprint’ that X-rays can leave on structures»

Métricas

La mejor manera de determinar el daño por radiación en un cristal macromolecular, es estimando la dosis de radiación absorbida por el cristal. La dosis depende a su vez de la composición atómica del cristal y de algunos parámetros experimentales referentes al haz de rayos X [6]. La dosis se mide en Gy que, según el Sistema Internacional de Unidades, es equivalente a la absorción de un Joule de energía ionizante por kilogramo de material irradiado. En experimentos de CRX es típico encontrar valores del orden de MGy [7]. En este aspecto, se han propuesto varias métricas para cuantificar el daño por radiación en función de la dosis absorbida; sin embargo, se ha demostrado que el uso de distintas métricas puede dar diferentes resultados [8]. Actualmente no existe una métrica que haya sido acordada por unanimidad en el campo de la CRX.

El daño por radiación específico se puede visualizar al realizar colectas de datos idénticas y tomar la diferencia entre la densidad electrónica del modelo estructural de la colecta n y la colecta inicial.



[6]: Murray y col. (2004), «X-ray absorption by macromolecular crystals: The effects of wavelength and crystal composition on absorbed dose»

[7]: Garman (2010), «Radiation damage in macromolecular crystallography: What is it and why should we care?»

[8]: Allan y col. (2013), «To scavenge or not to scavenge, that is STILL the question»

Figura 1.2: Diferencia de densidad electrónica entre colectas de datos idénticas. Arriba: primer colecta de un cristal de lisozima (dosis absorbida 0.6 MGy). En medio: segunda colecta del mismo cristal (dosis absorbida 3.2 MGy). Los mapas $2F_o - F_c$ y $F_o - F_c$ se encuentran dibujados a 1σ y $\pm 3.5\sigma$, respectivamente. Abajo: La diferencia de densidad electrónica entre colectas $F_{o,2} - F_{o,1}$. Donde la diferencia negativa (magenta) está a -3.5σ y la positiva (cian) a 3.5σ . Nótese como prácticamente no se observa una diferencia entre los mapas $2F_o - F_c$ (malla azul contra malla gris); sin embargo, sí existe una diferencia del modelo estructural inicial (el mismo en las tres imágenes) con respecto a la densidad electrónica de la segunda colecta. En otras palabras, para la segunda colecta, en una fracción de las proteínas cristalizadas este puente disulfuro se ha roto. Imagen realizada con PyMOL [9] y datos de [10].

1.3. Crioprotección

La primer estructura macromolecular determinada fue la de la mioglobina en 1958 por Kendrew y colaboradores [11]. La forma de contender con el daño por radiación en aquella época era utilizando decenas de cristales y promediar los patrones de difracción. La regla de dedo para cambiar el cristal irradiado por uno nuevo, era seguir la intensidad de algunas reflexiones y si esta llegaba a disminuir de 20 a 30 % de su valor inicial, entonces se procedía a reemplazar el cristal.

El primer estudio en el que se valoró el potencial de la crioprotección, para reducir el daño por radiación, surgió por necesidad. Sucedía que ciertos cristales de insulina con átomos pesados sufrían un rápido desgaste por la radiación, en comparación con cristales de insulina sin átomos pesados. Con base en la observación de que el daño secundario es dependiente, en parte, de la temperatura; Low *et al.* compararon, de manera cualitativa, el deterioro de los patrones de difracción colectados a 21, 0 y a -13 °C. Los resultados fueron claros: a menor temperatura, mayor el tiempo de vida útil de los cristales [12].

El problema con la reducción de temperatura en cristales macromoleculares, era la formación de hielo dentro de estos. Fue Haas quien primero usó crioprotectores para prevenir este problema. En el primer caso logró reducir la temperatura hasta -50 °C, remojando cristales de lisozima entrecruzados con glutaraldehído en una mezcla de agua con glicerol [13]. En un estudio posterior con cristales de lactato deshidrogenasa, el proceso de entrecruzamiento destruía los cristales. En cambio, si solo eran remojados por un par de días en una solución con sacarosa, el daño por radiación era diez veces menor [14]. Es hasta 1988 que Hope describe por primera vez lo que conocemos hoy en día como criocrystalografía de rayos X, donde básicamente se añade al cristal, o a la condición de cristalización, un crioprotector y el proceso de difracción del cristal se realiza a -173 °C [15]. Una de las desventajas de esta técnica es encontrar las condiciones de crioprotección adecuadas para cada macromolécula. A pesar de este detalle, la crioprotección fue ganando adeptos de tal forma que para el año 2000 era parte de la rutina de la CRX [16]. Gracias a la crioprotección, en general era suficiente un único cristal macromolecular para obtener un «dataset» completo.

[11]: Kendrew y col. (1958), «A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis»

1.4. Sincrotrones

La principales fuente de rayos X para realizar el experimento de CRX, es la radiación sincrotrón. La historia tecnológica de los sincrotrones se divide en generaciones. La primera generación de sincrotrones eran aquellos pertenecientes al campo de la física de partículas, donde los primeros estudios con respecto a la estructura de proteínas fueron realizados [17]. Para la década de 1980 se construyen los sincrotrones dedicados a la biología estructural, esta es la segunda generación. Para la década de 1990 llega la tercera generación. El primer sincrotrón perteneciente a la cuarta generación empezó a operar en 2016 [18]. Una de las características de un sincrotrón es su brillo espectral, que se define como la distribución del flujo de fotones en el espacio y el rango angular. El flujo se establece como el número de fotones por segundo que atraviesan un área definida por un

[12]: Low y col. (1966), «Studies of insulin crystals at low temperatures: effects on mosaic character and radiation sensitivity»

[13]: Haas (1968), «X-ray studies on lysozyme crystals at -50°C»

[14]: Haas y col. (1970), «Crystallographic studies on lactate dehydrogenase at -75 °C»

[15]: Hope (1988), «Cryocrystallography of biological macromolecules: a generally applicable method»

[16]: Garman (2003), «'Cool' crystals: macromolecular cryocrystallography and radiation damage.»

[17]: Phillips y col. (1976), «Applications of synchrotron radiation to protein crystallography: Preliminary results»

[18]: Owen y col. (2016), «Radiation damage and derivatization in macromolecular crystallography: a structure factor's perspective»

ancho de banda dado [19]. La revolución tecnológica de los sincrotrones se nota en la diferencia del orden de magnitud del brillo espectral [19]. Este aumento en brillo se ha permitido pues permite una gran ventaja: la posibilidad de utilizar cristales de menor tamaño. Esto es porque la principal limitante de la CRX es obtener cristales macromoleculares, en particular cristales de un tamaño adecuado (al menos unos cien micrómetros en sus tres dimensiones⁹.)

Actualmente se está desarrollando la tecnología para cambiar la metodología de la colecta de datos, usando cristales macromoleculares nanométricos y con una fuente de rayos X más poderosa denominada XFEL (del inglés *X-ray Free Electron Laser*). Existen ya varios estudios en los que se ha demostrado la posibilidad de obtener estructuras macromoleculares con esta nueva metodología [20]. Sin embargo, el acceso al tiempo experimental en un XFEL es actualmente muy limitado.

Como se mencionó en la sección anterior, ya para el año 2000, la noción general en el campo de la criocristalográfía era que el daño por radiación era insignificante, un problema del pasado. Precisamente esta noción cambia en ese mismo año, cuando tres estudios independientes muestran el efecto del daño por radiación en la entonces nueva generación de sincrotrones [3-5].

1.5. Radioprotectores

Al ser evidente que el daño por radiación aumentaba con el incremento en brillo, fue necesario buscar estrategias, como la crioprotección, que ayudaran a mitigar el daño por radiación. En el curso de los últimos veinte años, se han investigado varias estrategias pre y posteriores a la difracción con distintos enfoques [21]. Una de las tantas estrategias, es el uso de moléculas pequeñas que interactúan con los radicales libres generados por la radiación. Estas moléculas se denominan radioprotectores. Sin embargo, en la literatura científica existen varias incongruencias con respecto a la efectividad de los radioprotectores y es por esto que la comunidad cristalográfica no ha adoptado al cien por ciento el uso de radioprotectores de manera rutinaria [8, 22].

[19]: Willmott (2019), *An Introduction to Synchrotron Radiation: Techniques and Applications*

9: Existen líneas especiales donde existe la posibilidad de usar cristales con un orden de magnitud menor, las denominadas líneas microfoco.

[20]: Martin-Garcia y col. (2016), «Serial femtosecond crystallography: A revolution in structural biology»

[3]: Teng y col. (2000), «Primary radiation damage of protein crystals by an intense synchrotron X-ray beam»

[4]: Weik y col. (2000), «Specific chemical and structural damage to proteins produced by synchrotron radiation»

[5]: Ravelli y col. (2000), «The ‘fingerprint’ that X-rays can leave on structures»

[21]: Garman y col. (2017), «X-ray radiation damage to biological macromolecules: further insights»

[22]: Nowak y col. (2009), «To scavenge or not to scavenge: that is the question»

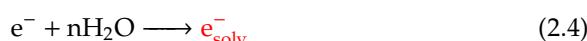
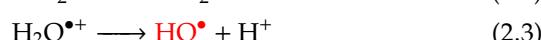
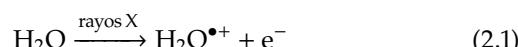
[8]: Allan y col. (2013), «To scavenge or not to scavenge, that is STILL the question»

Antecedentes

2.1. pH

Una estrategia innovadora, investigada en la tesis de maestría del presente autor, fue modificar el pH dentro de los cristales macromoleculares. Esta idea se basa en la idea general de los radioprotectores y se detalla a continuación.

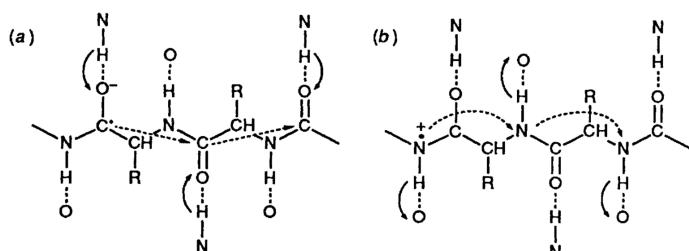
La radiólisis del agua produce las siguientes reacciones [23]:



[23]: Sonntag (2006), *Free-Radical-Induced DNA Damage and Its Repair*

Donde las especies químicas denotadas en rojo, representan los primeros radicales libres presentes en un cristal macromolecular (el radical hidroxilo, el electrón solvatado y el radical hidrógeno).

Con la criocrystalografía de rayos X, se impide la difusión del radical hidroxilo [24]. Sin embargo, el $\text{e}_{\text{solv.}}^-$ y el radical hidrógeno todavía son móviles. El electrón solvatado puede moverse del solvente a la proteína, donde es capaz de viajar a través de la cadena polipeptídica hasta hallar un centro electrofílico, como por ejemplo átomos metálicos o puentes disulfuro (véase la Figura 2.1) [25]. Esta es la razón del origen del orden en el que se presenta el daño por radiación específico, pues la captura de electrones es mucho más específica que la captura de huecos positivos [26]. Por su parte, el radical hidrógeno sigue una reacción de recombinación formando como producto final H_2 , al parecer el responsable directo del daño por radiación global [27].



[24]: Owen y col. (2012), «Outrunning free radicals in room-temperature macromolecular crystallography»

[25]: Symons (1997), «Electron movement through proteins and DNA»

[26]: Close y col. (2019), «Comprehensive model for X-ray-induced damage in protein crystallography»

[27]: Meents y col. (2010), «Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures»

El electrón solvatado y el átomo de hidrógeno se encuentran en un equilibrio ácido-base; por lo que el electrón solvatado se convierte en H^\bullet en una solución ácida.



Figura 2.1: Movimiento de electrones (a) y huecos positivos (b), a través de la cadena polipeptídica. Fuente: [25].

En la tesis de maestría se usó como sistema de estudio la lisozima de clara de huevo de gallina, que presenta cuatro puentes disulfuro. La idea era que el ion hidronio, también conocido como oxidanio, funcionara como radioprotector al interactuar con los electrones solvatados antes que estos interactuaran con los puentes disulfuro de esta proteína. Para esto, se comparó el daño específico sobre los puentes disulfuro en cristales de lisozima a tres niveles de pH: 3.7, 4.7 y 5.7. Los resultados obtenidos fueron opuestos a los esperados: a niveles comparables de dosis de radiación absorbida, el cristal con el pH más ácido presentó mayor daño por radiación que el cristal con el pH más básico.

La variabilidad entre cristales macromoleculares, aún proveniendo de la misma condición de cristalización, puede llevar a conclusiones erróneas [22]. En la tesis de maestría no se pudo concluir si la diferencia observada era estadísticamente significativa, pues no se realizaron las repeticiones necesarias para cada condición experimental.

[22]: Nowak y col. (2009), «To scavenge or not to scavenge: that is the question»

3

Objetivo

El objetivo de este proyecto es determinar el efecto del pH en la condición de cristalización de ciertas proteínas sobre el daño por radiación.

3.1. Objetivos particulares

Para lograr el objetivo principal de este proyecto, se plantean los siguientes objetivos particulares:

(4-10).

Obtener mapas de diferencia de densidad electrónica entre puntos fijos de dosis de radiación absorbida.

1. Realizar un análisis *in silico* para determinar qué proteínas son capaces de cristalizar en un intervalo de pH amplio.
2. Cristalizar las proteínas seleccionadas a diferentes niveles de pH usando cualquiera de los sistemas de amortiguamiento que muestren un intervalo de pH amplio
3. Determinar los parámetros de la colecta de datos que produzcan niveles similares o idénticos de dosis de radiación absorbida en las diferentes proteínas cristalizadas.
4. Realizar colectas de datos continuas y seriales en un sincrotrón.
5. Procesar los patrones de difracción para obtener un modelo inicial de las proteínas.
6. Mapear la diferencia de densidad electrónica entre colectas de datos al modelo inicial de cada proteína y realizar un análisis comparativo, en particular sobre los residuos de aminoácido que son más susceptibles al daño por radiación, para determinar las diferencias en daño por radiación a diferentes niveles de pH.

4

Materiales y métodos

El PDB contiene información acerca de cientos de miles de proteínas cristalizadas en diferentes condiciones de cristalización. Para hallar qué proteínas pueden cristalizar en un amplio intervalo de pH se realizó un análisis *in silico*, como se explica brevemente¹ a continuación.

1: Con detalle en <https://github.com/murpholinox/doctordado>.

Extracción de datos

Se decidió emplear la información cruda del PDB, es decir, extraer la información experimental necesaria directamente del cabezal de los archivos de las estructuras depositadas en el PDB. La principal ventaja es que la extracción de información es de una manera más directa, sin depender de la interfaz de programación de aplicaciones (API, por sus siglas en inglés) del mismo PDB. Para extraer la información de los cabezales se usó el programa gemmi². La información extraída es la siguiente: el contador de la entidad macromolecular, el tipo de entidad, el código de acceso de la base de datos de referencia³, su descripción, el método experimental para crecer los cristales, el pH de la condición de cristalización, los detalles del experimento de la cristalización, la resolución final del modelo estructural, el grupo espacial en el que cristaliza la macromolécula y el identificador de objeto digital de la publicación científica correspondiente. A continuación se muestra un ejemplo de la información extraída:

2: Disponible en el siguiente enlace <https://github.com/project-gemmi/gemmi>.

3: La mayoría de las veces el código usado es aquél de la base de datos UniProt <https://www.uniprot.org/>.

Ejemplo 1

```
6LU7,1,polypeptide(L),P0DTD1,main protease,EVAPORATION,\n6,"2% polyethylene glycol (PEG) 6000, 3% DMSO, 1mM DTT,\n0.1M MES buffer (pH 6.0), protein concentration 5mg/ml,\nVAPOR DIFFUSION, HANGING DROP, temperature 293K",2.16,\n2.16,C 1 2 1,10.1038/s41586-020-2223-y,21728,210031\n6LU7,2,polypeptide(L),P0DTD1,main protease,EVAPORATION,\n6,"2% polyethylene glycol (PEG) 6000, 3% DMSO, 1mM DTT,\n0.1M MES buffer (pH 6.0), protein concentration 5mg/ml,\nVAPOR DIFFUSION, HANGING DROP, temperature 293K",2.16,\n2.16,C 1 2 1,10.1038/s41586-020-2223-y,21728,210031
```

El resultado final es una tabla de datos, con 226 523 observaciones, o filas, y 14 variables, o columnas. El número de observaciones es mayor que el número de estructuras en el PDB, esto es debido a que cada archivo puede tener más de una macromolécula⁴.

4: Como es el caso del ejemplo 1.

Limpieza de datos

Se aplican los siguientes filtros a los datos extraídos para eliminar observaciones que:

1. Carecen de código de acceso.
2. Tienen una resolución peor que 2 Å.
3. Carecen del valor de pH de la condición de cristalización.
4. El número de entidades es mayor o igual a dos.

La lógica de los filtros es la siguiente: (1 y 3) Remover entradas que no tengan las anotaciones correspondientes⁵, en este caso el código de acceso de la proteína y el pH de la condición de cristalización. La última es la anotación más relevante para este proyecto y la primera es la única manera de conocer casi inequívocamente la proteína representada en el archivo. (2) La resolución final del modelo estructural es un indicador de la calidad del cristal obtenido, a mayor resolución menos defectos posee el cristal. Este filtro garantizará que el listado de proteínas obtenidas sean fáciles de cristalizar. Además este filtro servirá para el análisis posterior, al comparar el daño por radiación específico, pues los resultados serán más confiables con una buena resolución. (4) Como se mencionó anteriormente, un archivo puede contener múltiples macromoléculas. Este filtro ayuda a descartar proteínas cristalizadas con otras. En general, las condiciones de cristalización para combinaciones diferentes de macromoléculas serán diferentes, por lo que no tiene caso tener varias observaciones de la misma proteína si presenta diferentes compañeros.

En la tabla Tabla A.1, se muestran las 50 proteínas más representadas en los datos que cumplen los primeros cuatro filtros. A partir de las cuales se aplican los siguientes dos filtros, que ayudan a eliminar las observaciones donde:

5. Su secuencia de aminoácidos sea diferente de la secuencia consenso para cada conjunto de proteínas.
6. No presenten un intervalo de pH amplio en su cristalización.

La lógica de estos dos filtros es la siguiente: (5) en el primer filtro se alegó que el código de acceso de UniProt es la manera de conocer *casi inequívocamente la proteína representada*. En el caso de algunos virus, el código corresponde a un gen que puede codificar para diferentes proteínas. Debido a esto se realizó un alineamiento múltiple, con el programa mafft [28], para eliminar proteínas distintas a la respectiva secuencia consenso con el mismo código de acceso. A continuación se muestra un ejemplo:

Ejemplo 2

```
6W4H,1,polypeptide(L),P0DTD1,SARS-CoV-2 NSP16,...
6W4H,2,polypeptide(L),P0DTD1,SARS-CoV-2 NSP16,...
6W6Y,1,polypeptide(L),P0DTD1,SARS-CoV-2 NSP3,...
6WQD,1,polypeptide(L),P0DTD1,SARS-CoV-2 NSP7,...
6WQD,2,polypeptide(L),P0DTD1,SARS-CoV-2 NSP7,...
7BUY,1,polypeptide(L),7BUY,SARS-CoV-2 virus Main protease,
```

Es claro que el código, P0DTD1, es el mismo; sin embargo, las observaciones corresponden a diferentes proteínas. Este filtro ayuda a mantener proteínas en las que su secuencia no difiera entre sí por más de 15 aminoácidos. Se mantienen entonces proteínas que difieran entre sí por la etiqueta de polihistidinas, pero se excluyen proteínas que contengan el péptido señal y proteínas quimeras, por ejemplo. Y (6) es la condición

⁵: Desafortunadamente la información experimental no siempre se encuentra disponible en los archivos del PDB.

[28]: Katoh y col. (2013), «MAFFT multiple sequence alignment software version 7: Improvements in performance and usability»

experimental que nos interesa en este proyecto, proteínas que cristalicen en un amplio intervalo de pH. Este filtro se aplicó por partes. Primero se realizó un gráfico de caja para cada una de las 50 proteínas más representadas en los datos. Este tipo de gráfica da una representación visual de la distribución de la variable en cuestión. Si la distribución no cubre al menos tres unidades de pH, entonces se descarta dicha proteína, en caso contrario se mantiene. De las proteínas restantes, 25, se realiza un histograma de frecuencias para determinar de manera cuantitativa la frecuencia con la que cada proteína ha sido cristalizada en valores de pH diferentes. Si la frecuencia no es mayor a cinco para la mayoría de las barras en cada histograma, la proteína se descarta, si no se mantiene (véase la Figura B.1). Esto resulta en un listado de 14 proteínas, donde las diez primeras entradas se presentan a continuación (véase la Tabla 4.1).

Número	Código	Intervalo	Nombre
1	P00918	6	Anhidrasa carbónica
2	P00698	5.5	Lisozima
3	P00760	6	Tripsina
4	P02766	5	Transtiretina
5	P42212	6	Proteína verde fluorescente
6	O60885	3.5	Proteína bromodominio 4
7	P19491	4.5	Receptor de glutamato 2
8	O26232	4	Orotidina-5'-fosfato descarboxilasa
9	P00772	4.5	Elastasa
10	P00644	3	Termonucleasa

Tabla 4.1: Proteínas que cristalizan en un intervalo amplio de pH.

4.1. Colecta y análisis de datos

Los cristales obtenidos serán difractados en un sincrotrón, midiendo la dosis de radiación absorbida. Para cada proteína se tendrán que realizar colectas de datos secuenciales de manera repetitiva. Gracias a la presencia de modelos estructurales en el PDB, se puede estimar la dosis absorbida por la proteína antes de realizar el experimento de difracción. Esto se puede realizar gracias al programa raddose [29]. Se resolverán las estructuras por reemplazo molecular y se crearán mapas de diferencia de densidad electrónica para analizar la diferencia en daño por radiación específico a distintos pHs al mismo nivel de dosis de radiación absorbida.

[29]: Zeldin y col. (2013), «RADDSE-3D : time- and space-resolved modelling of dose in macromolecular crystallography»

A

Proteínas más representadas

Número	Código	Cuenta	No.	Código	Cuenta
1	P11838	689	26	P23497	106
2	P00918	619	27	P00800	103
3	P00698	498	28	026232	102
4	P00760	310	29	P22629	97
5	Q6PJP8	296	30	P00489	96
6	Q6B0I6	269	31	P03367	96
7	P02766	258	32	P68400	95
8	095696	257	33	A0A073FPA6	84
9	Q9UIF8	227	34	P46881	79
10	P00644	216	35	P00811	78
11	P00720	215	36	Q16539	77
12	P24941	203	37	P14174	75
13	P42212	182	38	P00431	73
14	P29476	178	39	P01116	73
15	P02185	172	40	P00183	72
16	060885	170	41	P01112	72
17	P18031	152	42	Q00511	70
18	P61823	145	43	Q76353	68
19	P28720	144	44	P00282	64
20	P07900	142	45	P02883	63
21	P61626	139	46	P02945	63
22	P15121	125	47	P06873	63
23	P56817	124	48	P16113	63
24	P0DTD1	123	49	Q04609	63
25	P19491	116	50	P00772	61

Tabla A.1: Las 50 proteínas más representadas en los datos, después de los cuatro primeros filtros. Eliminando estructuras que no contienen: código de acceso ni el valor de pH de la condición experimental. Además descarta aquellas estructuras con una resolución peor que 2 Å y donde el número de entidades es mayor o igual a dos.

B

Análisis visual

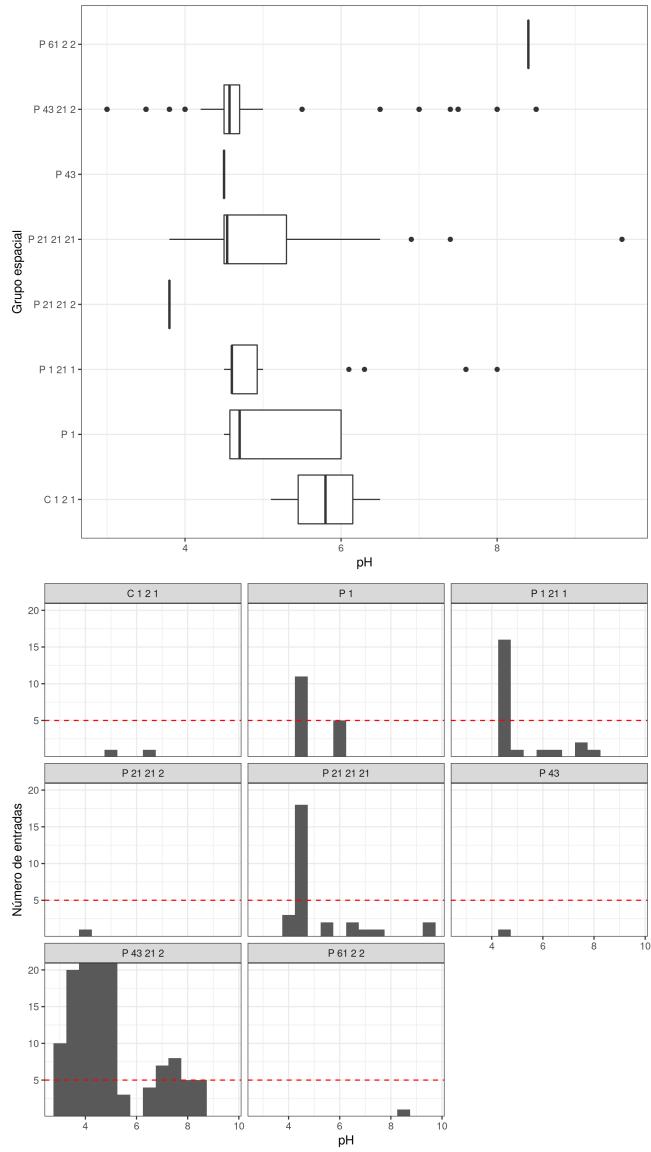


Figura B.1: Ejemplo del análisis visual: histograma de P00698. Nótese, en el gráfico de caja, que esta proteína tiene un amplio intervalo de pH ~ 5 , en particular cuando cristaliza en el grupo espacial P 43 21 2. En este caso, la mayor parte de las estructuras están cristalizadas a un pH cercano a 4.7, de ahí la forma final del gráfico de caja. Su frecuencia de cristalización en este mismo intervalo, según el histograma, es arriba de cinco (denotado por la línea horizontal roja), por lo menos para buena parte del intervalo.

Bibliografía

Aquí se encuentran las referencias citadas en orden de aparición.

- [1] Helen M. Berman y col. «The protein data bank». En: *Nucleic Acids Research* 28.1 (2000), págs. 235-242. doi: [10.1093/nar/28.1.235](https://doi.org/10.1093/nar/28.1.235) (vid. pág. 1).
- [2] Zbigniew Dauter. «Data-collection strategies». En: *Acta Crystallographica Section D: Biological Crystallography* 55.10 (1999), págs. 1703-1717. doi: [10.1107/S0907444999008367](https://doi.org/10.1107/S0907444999008367) (vid. pág. 2).
- [3] Tsu Yi Teng y Keith Moffat. «Primary radiation damage of protein crystals by an intense synchrotron X-ray beam». En: *Journal of Synchrotron Radiation* 7.5 (2000), págs. 313-317. doi: [10.1107/S0909049500008694](https://doi.org/10.1107/S0909049500008694) (vid. págs. 3, 6).
- [4] Martin Weik y col. «Specific chemical and structural damage to proteins produced by synchrotron radiation». En: *Proceedings of the National Academy of Sciences of the United States of America* (2000). doi: [10.1073/pnas.97.2.623](https://doi.org/10.1073/pnas.97.2.623) (vid. págs. 3, 6).
- [5] Raimond B.G. Ravelli y Sean M. McSweeney. «The 'fingerprint' that X-rays can leave on structures». En: *Structure* (2000). doi: [10.1016/S0969-2126\(00\)00109-X](https://doi.org/10.1016/S0969-2126(00)00109-X) (vid. págs. 3, 6).
- [6] James W. Murray, Elspeth F. Garman y Raimond B.G. Ravelli. «X-ray absorption by macromolecular crystals: The effects of wavelength and crystal composition on absorbed dose». En: *Journal of Applied Crystallography* 37.4 (2004), págs. 513-522. doi: [10.1107/S0021889804010660](https://doi.org/10.1107/S0021889804010660) (vid. pág. 4).
- [7] Elspeth F. Garman. «Radiation damage in macromolecular crystallography: What is it and why should we care?». En: *Acta Crystallographica Section D: Biological Crystallography* (2010). doi: [10.1107/S0907444910008656](https://doi.org/10.1107/S0907444910008656) (vid. pág. 4).
- [8] Elizabeth G. Allan y col. «To scavenge or not to scavenge, that is STILL the question». En: *Journal of Synchrotron Radiation* (2013). doi: [10.1107/S0909049512046237](https://doi.org/10.1107/S0909049512046237) (vid. págs. 4, 6).
- [9] LLC Schrödinger. «The {PyMOL} Molecular Graphics System, versión 2.4». Nov. de 2015 (vid. pág. 4).
- [10] Max H. Nanao, George M. Sheldrick y Raimond B. G. Ravelli. «Improving radiation-damage substructures for RIP». En: *Acta Crystallographica Section D Biological Crystallography* 61.9 (sep. de 2005), págs. 1227-1237. doi: [10.1107/S0907444905019360](https://doi.org/10.1107/S0907444905019360) (vid. pág. 4).
- [11] J. C. Kendrew y col. «A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis». En: *Nature* 181.4610 (mar. de 1958), págs. 662-666. doi: [10.1038/181662a0](https://doi.org/10.1038/181662a0) (vid. pág. 5).
- [12] B. W. Low y col. «Studies of insulin crystals at low temperatures: effects on mosaic character and radiation sensitivity». En: *Proceedings of the National Academy of Sciences* 56.6 (dic. de 1966), págs. 1746-1750. doi: [10.1073/pnas.56.6.1746](https://doi.org/10.1073/pnas.56.6.1746) (vid. pág. 5).
- [13] D. J. Haas. «X-ray studies on lysozyme crystals at -50°C». En: *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry* 24.4 (1968), págs. 604-604. doi: [10.1107/s056774086800292x](https://doi.org/10.1107/s056774086800292x) (vid. pág. 5).
- [14] D. J. Haas y M. G. Rossmann. «Crystallographic studies on lactate dehydrogenase at -75 °C». En: *Acta crystallographica. Section B: Structural crystallography and crystal chemistry* 26.7 (1970), págs. 998-1004. doi: [10.1107/S0567740870003485](https://doi.org/10.1107/S0567740870003485) (vid. pág. 5).
- [15] H. Hope. «Cryocrystallography of biological macromolecules: a generally applicable method». En: *Acta Crystallographica Section B* 44.1 (1988), págs. 22-26. doi: [10.1107/S0108768187008632](https://doi.org/10.1107/S0108768187008632) (vid. pág. 5).
- [16] Elspeth Garman. «'Cool' crystals: macromolecular cryocrystallography and radiation damage». En: *Current opinion in structural biology* 13.5 (oct. de 2003), págs. 545-51. doi: [10.1016/j.sbi.2003.09.013](https://doi.org/10.1016/j.sbi.2003.09.013) (vid. pág. 5).

- [17] J. C. Phillips y col. «Applications of synchrotron radiation to protein crystallography: Preliminary results». En: *Proceedings of the National Academy of Sciences of the United States of America* (1976). doi: [10.1073/pnas.73.1.128](https://doi.org/10.1073/pnas.73.1.128) (vid. pág. 5).
- [18] Robin L. Owen y Darren A. Sherrell. «Radiation damage and derivatization in macromolecular crystallography: a structure factor's perspective». En: *Acta Crystallographica Section D Structural Biology* 72.3 (mar. de 2016), págs. 388-394. doi: [10.1107/S2059798315021555](https://doi.org/10.1107/S2059798315021555) (vid. pág. 5).
- [19] Philip Willmott. *An Introduction to Synchrotron Radiation: Techniques and Applications*. 2nd ed. John Wiley & Sons, 2019, pág. 504 (vid. pág. 6).
- [20] Jose M. Martin-Garcia y col. «Serial femtosecond crystallography: A revolution in structural biology». En: *Archives of Biochemistry and Biophysics* 602 (jul. de 2016), págs. 32-47. doi: [10.1016/j.abb.2016.03.036](https://doi.org/10.1016/j.abb.2016.03.036) (vid. pág. 6).
- [21] Elspeth F Garman y Martin Weik. «X-ray radiation damage to biological macromolecules: further insights». En: *Journal of Synchrotron Radiation* 24.1 (ene. de 2017), págs. 1-6. doi: [10.1107/S160057751602018X](https://doi.org/10.1107/S160057751602018X) (vid. pág. 6).
- [22] Elzbieta Nowak y col. «To scavenge or not to scavenge: that is the question». En: *Acta Crystallographica Section D* 65.9 (sep. de 2009), págs. 1004-1006. doi: [10.1107/S0907444909026821](https://doi.org/10.1107/S0907444909026821) (vid. págs. 6, 8).
- [23] Clemens von Sonntag. *Free-Radical-Induced DNA Damage and Its Repair*. Berlin Heidelberg: Springer, 2006 (vid. pág. 7).
- [24] Robin L. Owen y col. «Outrunning free radicals in room-temperature macromolecular crystallography». En: *Acta Crystallographica Section D: Biological Crystallography* 68.7 (2012), págs. 810-818. doi: [10.1107/S0907444912012553](https://doi.org/10.1107/S0907444912012553) (vid. pág. 7).
- [25] Martyn C.R. Symons. «Electron movement through proteins and DNA». En: *Free Radical Biology and Medicine* 22.7 (1997), págs. 1271-1276. doi: [10.1016/S0891-5849\(96\)00548-5](https://doi.org/10.1016/S0891-5849(96)00548-5) (vid. pág. 7).
- [26] David M. Close y William A. Bernhard. «Comprehensive model for X-ray-induced damage in protein crystallography». En: *Journal of Synchrotron Radiation* 26.4 (jul. de 2019), págs. 945-957. doi: [10.1107/S1600577519005083](https://doi.org/10.1107/S1600577519005083) (vid. pág. 7).
- [27] Alke Meents y col. «Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures». En: *Proceedings of the National Academy of Sciences* 107.3 (ene. de 2010), págs. 1094-1099. doi: [10.1073/pnas.0905481107](https://doi.org/10.1073/pnas.0905481107) (vid. pág. 7).
- [28] Kazutaka Katoh y Daron M. Standley. «MAFFT multiple sequence alignment software version 7: Improvements in performance and usability». En: *Molecular Biology and Evolution* 30.4 (2013), págs. 772-780. doi: [10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010) (vid. pág. 11).
- [29] Oliver B. Zeldin, Markus Gerstel y Elspeth F. Garman. «RADDOSE-3D : time- and space-resolved modelling of dose in macromolecular crystallography». En: *Journal of Applied Crystallography* 46.4 (ago. de 2013), págs. 1225-1230. doi: [10.1107/S0021889813011461](https://doi.org/10.1107/S0021889813011461) (vid. pág. 12).