

Introdução: Cristalografia de Proteínas

Proteínas executam a maioria das funções específicas dentro de uma célula no nível molecular, incluindo o reconhecimento de substratos, catálise de reações químicas, transporte de moléculas através de membranas, controle de fluxo de informação gênica codificado por ácidos nucleicos e a montagem de complexos mutiprotéicos importantes para a construção e manutenção de estruturas celulares.

A estrutura tridimensional de uma proteína é determinada por sua seqüência de aminoácidos. Há milhões de seqüências de proteínas em bancos de dados públicos (ex: <http://www.ncbi.nlm.nih.gov/>) derivados de projetos de seqüenciamento de genomas que podem ser agrupadas (alinhadas) em famílias de domínios que são relacionados por descendência evolucionária (ex: <http://pfam.wustl.edu/>). Nesta descendência, pequenas modificações em seqüência normalmente não afetam a estrutura tridimensional e logo a função específica é mantida. Portanto, diferenças maiores entre as seqüências dentro de uma família protéica podem refletir diferenças significativas em suas estruturas e conseqüentemente em suas funções. Às vezes, a divergência evolucionária é tão grande que não existe mais semelhança significativa em suas seqüências apesar de manutenção de estruturas e funções comuns. Estimativas do número de famílias de estruturas ou **topologias (“folds”)** protéicas na biosfera variam de entre 1000 e 4000.

A função de uma proteína está diretamente relacionada com sua estrutura. A nossa capacidade de entender esta relação nos seus maiores detalhes é o fator limitante da nossa capacidade de entender processos biológicos e a base molecular da vida. Esta relação é a chave para realizar o potencial de poder manipular o mundo orgânico em benefício à saúde humana, por exemplo o desenho racional de novas drogas e vacinas, o desenho de intervenções moleculares para combater deficiências hereditárias e a produção de animais e plantas transgênicas com propriedades que aumentam rendimento e valor nutricional sem introduzir riscos aos delicados equilíbrios ecológicos.

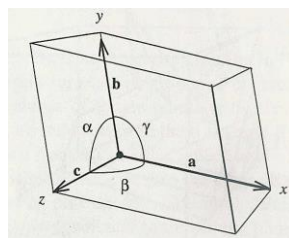
Existem três maneiras de obter informação detalhada (resolução atômica) sobre estruturas protéicas: **cristalografia**, **ressonância magnética nuclear (RMN)** e mais recentemente a técnica de **criomicroscopia eletrônica (Cryo-EM)**. As três técnicas são complementares: Enquanto a cristalografia em geral permite uma maior resolução e determinação de estruturas maiores, RMN fornece, além de informação estrutural, detalhes sobre processos dinâmicos relacionados ao

intercâmbio entre conformações em solução, algumas das quais podem ser transitórias e resistentes à cristalização. Cryo-EM permite estudar proteínas ou complexos macromoleculares que, devido ao seu tamanho e/ou flexibilidade intrínseca não podem ser estudados por cristalografia ou RMN.

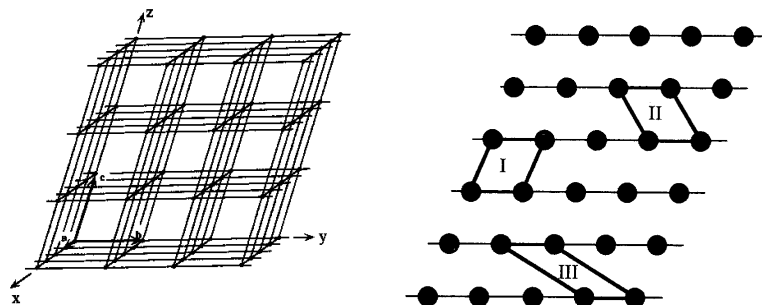
Nas próximas semanas, nós vamos contemplar os vários procedimentos envolvidos na determinação da estrutura tridimensional de uma proteína. Devido às limitações de espaço e de tempo, os detalhes da maioria dos procedimentos serão omitidos. Aos que estão interessados em tratamentos mais profundos sobre o assunto, recomendamos consultar o Material Bibliográfico.

Teoria

Cristais de Proteínas. Em condições específicas, macromoléculas podem formar cristais. Cristalógrafos produzem cristais de proteínas por um processo lento e controlado em condições não desnaturantes. Por definição, cada cristal é um arranjo **periódico** e tridimensional das moléculas, íons ou átomos que o compõe. Contudo, pode-se imaginar um cristal como sendo formado por unidades básicas (paralelepípedos), denominadas de **células unitárias**, que se repetem no espaço por simples translações ao longo de três eixos. Esta definição facilita o estudo cristalográfico já que o conteúdo de uma célula unitária é o mesmo para qualquer outra célula unitária. Além disto, a determinação da configuração de uma célula unitária possibilita determinar a estrutura de todo o cristal. A célula unitária tem uma origem e três eixos a partir dos quais as posições de cada átomo podem ser definidas. A célula unitária de um cristal pode ser descrita por 3 vetores (**a**, **b**, **c**) e 3 ângulos (α , β , γ).

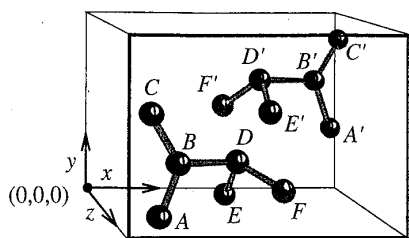


Se desconsiderarmos, por hora, o conteúdo de cada célula unitária, o cristal pode ser visto como uma **rede** de pontos derivados dos vértices das células unitárias.



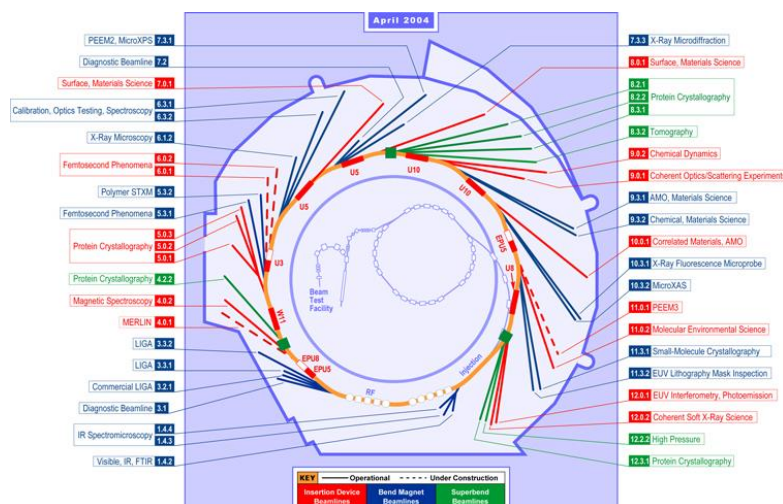
Dentro de cada célula unitária, podem existir um ou mais conjuntos de moléculas equivalentes que são relacionados entre si por **operações de simetria** (combinações de rotações de 60, 90, 120, ou 180° e translações de 1/6, 1/4, 1/3 ou 1/2 do eixo). Estas operações possibilitam subdividir cada célula unitária em regiões menores e equivalentes denominadas de **unidades assimétricas**. Assim, o conteúdo de uma unidade assimétrica estará relacionado com o conteúdo de outra unidade assimétrica por uma determinada operação de simetria. Note que a estrutura de um cristal pode ser determinada com base na estrutura presente em uma única unidade assimétrica e na simetria da célula unitária.

A simetria dentro de uma célula unitária é descrita por seu **grupo espacial**, representado por um símbolo, por exemplo $P2_1$, $P2_12_12_1$ etc. Você pode descrever as unidades assimétricas da figura abaixo? Elas são relacionadas por quais operações de simetria?

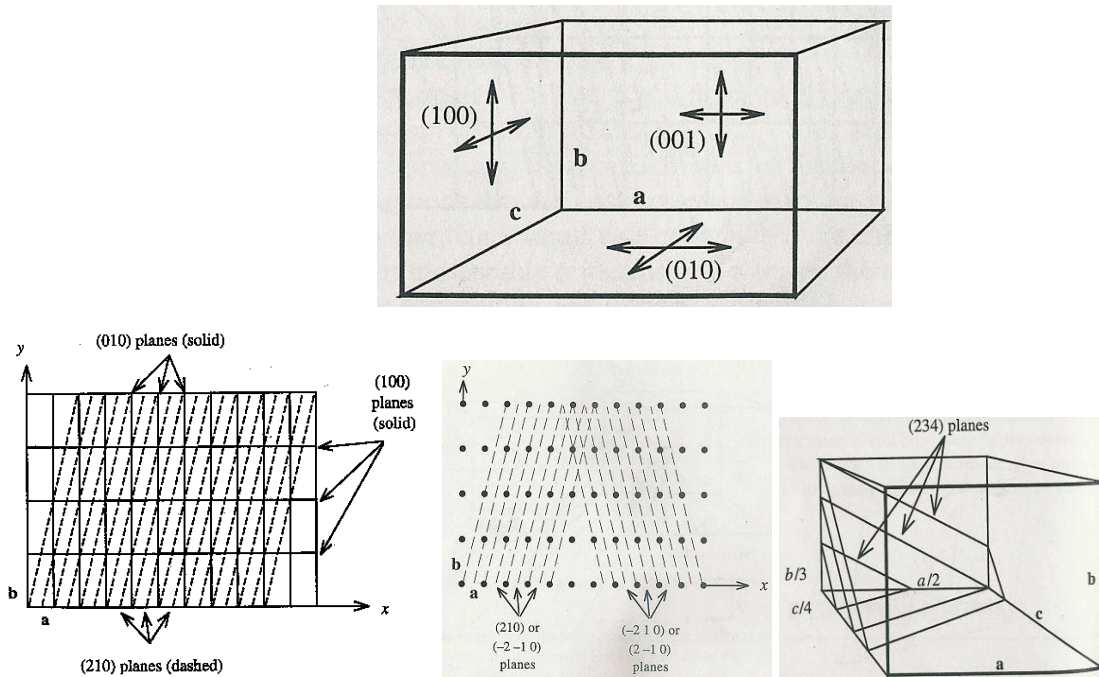


Fontes de Raios X. Raios X são uma forma de radiação eletromagnética com comprimento de onda (λ) entre 0,1 e 100 Å (0,01 e 10 nm). Eles podem ser produzidos de várias maneiras, duas das quais são relevantes para cristalografia: (i) Bombardeamento de um metal (normalmente Cu ou Mb) com elétrons. Um elétron de alta energia colide com um elétron de um orbital de baixa energia do metal resultando na sua ionização. Um elétron de um orbital de alta energia do metal então cai para o orbital vazio, emitindo energia na forma de um fóton de raio X. ii) Emissão de **radiação síncrotron**. Elétrons são acelerados até quase a velocidade de luz em anéis com circunferências de dezenas de metros até quilômetros denominados síncrotrons. A desaceleração dos elétrons nas

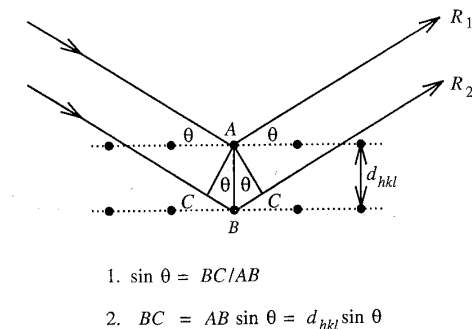
regiões de curvatura do anel resulta na emissão de radiação síncrotron. Essa radiação eletromagnética tem espectro amplo que inclui os raios X. O único síncrotron na América Latina se encontra em Campinas, no Laboratório Nacional de Luz Síncrotron.



Difração de raios X por cristais: Índices de Miller e a Lei de Bragg. Quando um feixe de raios X passa por um cristal, uma parte sofre **difração**. Esta difração pode ser interpretada (simplificado) pela **Lei do Bragg**, que trata os raios de luz como se tivessem sido refletidos por “planos cristalinos” imaginários que cortam os eixos das células unitárias (**a**, **b** e **c**) por números inteiros **h**, **k**, e **l** respectivamente. Os números **h**, **k**, **l** são chamados de **índices de Miller**. Famílias de planos paralelos podem ser descritas por combinações de índices hkl onde as letras h, k e l são números inteiros (vide figuras abaixo). De fato, para qualquer cristal existe um número infinito de índices, cada um que corta os três eixos da célula unitária em um número inteiro de fatias paralelas. Quanto maior o número do índice, as fatias naquela dimensão ficam mais finas. (Normalmente tratamos de índices com módulo menor do que 100.)



Imagine, portanto, um cristal sendo bombardeado por um feixe monocromático de raios-X. Nesta condição, é possível imaginar um conjunto de planos cristalinos imaginários orientados com um ângulo θ em relação aos feixes incidente e refletido, como mostra a figura abaixo. A Lei de Bragg diz que quando a diferença de caminho dos raios que atingem dois ou mais planos ($2BC = 2d \sin \theta$) for igual a um múltiplo inteiro ($n = 1, 2, 3, 4 \dots$) do comprimento de onda da radiação incidente ($n\lambda$), teremos uma condição que resulta em interferência construtiva para a onda difratada (raio-X refletido).

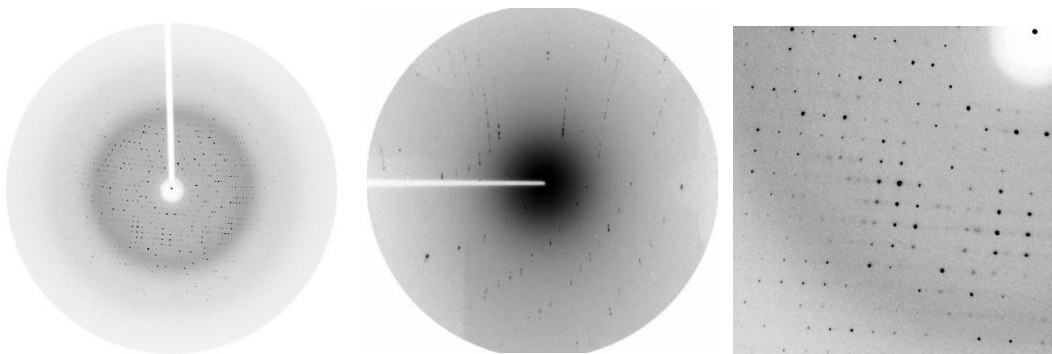


Diferença de caminho entre raios R_1 e R_2 = $2BC = 2AB \sin \theta = 2d \sin \theta$

Condição de Bragg: ocorre interferência construtiva somente quando $n\lambda = 2d \sin \theta$

Cada conjunto de planos hkl tem uma distância (d) entre os planos e um ângulo (θ) em relação à radiação incidente/refletida específicos. Logo, cada conjunto de planos produz uma única **reflexão** somente quando o cristal está orientado de uma maneira que satisfaça à condição de Bragg. Isto gera um padrão de reflexões em três dimensões cujas intensidades podem ser medidas por detectores.

Observação: A Lei de Bragg apenas prevê onde e sob quais condições uma reflexão (ponto de difração) será vista. Contudo, uma teoria mais quantitativa é necessária para descrever em detalhes a intensidade de cada um dos pontos de difração (reflexões). Esta teoria será abordada nas aulas da disciplina.



Espaço real vs espaço recíproco. Notar que as considerações acima implicam que as famílias de planos com distâncias (d) pequenas são representadas por índices (hkl) grandes e possuem ângulos de reflexão (θ) também grandes. Logo, reflexões derivadas de planos com índices grandes se encontram na periferia do padrão de difração. Também, reflexões derivadas de um determinado plano hkl em células unitárias grandes terão ângulos menores que células unitárias pequenas.

Desta relação entre distâncias dentro do cristal e dentro do padrão de difração derivamos a relação entre **espaço real** (do cristal com coordenadas x, y, z) e **espaço recíproco** (do padrão de difração com coordenadas h, k, l). Pode ser demonstrado que cada ponto no espaço recíproco (cada reflexão hkl) se localiza no final de um vetor com comprimento $1/d_{hkl}$, que parte da origem e é perpendicular à família de planos com índices hkl (cujos planos são separados pela distância d_{hkl}). Também, para cada célula unitária com eixos a, b e c, existe uma célula unitária recíproca cujos eixos tem comprimentos proporcionais a $1/a$, $1/b$ e $1/c$. Logo quando um cristal (**rede real**) é rodado no espaço real, podemos associar a **rede recíproca** rodando no espaço recíproco e sendo medida no detector.

Equações de ondas e funções periódicas. Até agora, nossa consideração sobre o padrão de difração ignorou o conteúdo da célula unitária (de fato, as **posições** das reflexões são **independentes** do conteúdo da célula unitária e somente dependem das dimensões da célula unitária e outras variáveis geométricas da coleta dos dados). Para entender como a informação do padrão de difração pode ser utilizada para determinar as estruturas das moléculas dentro da célula unitária (mais precisamente da unidade assimétrica), precisamos considerar a natureza das ondas eletromagnéticas.

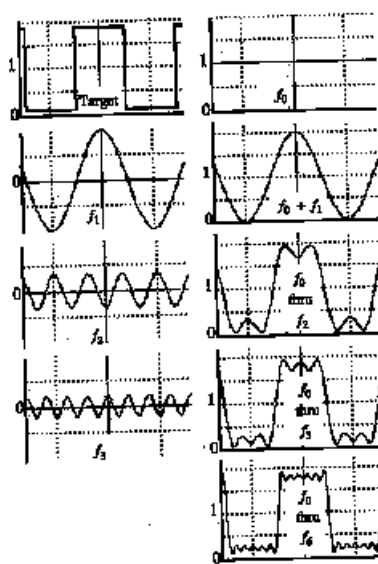
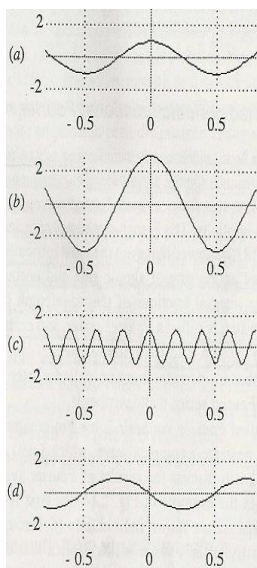
Ondas simples podem ser descritas por funções periódicas simples, por exemplo:

$$f(x) = F \cos 2\pi(hx + \phi) \quad \text{ou}$$

$$f(x) = F \cos 2\pi(hx) + i \sin 2\pi(hx) \quad \text{ou}$$

$$f(x) = F e^{2\pi i(hx)}$$

onde $f(x)$ = amplitude da onda na posição x , F = amplitude máxima da onda, h = frequência, ϕ = fase ou posição da onda em relação à origem. Inspeccionar o diagrama 1 (a), (b), (c) e (d) abaixo e determinar os valores de F , h e ϕ em cada caso.



Cristais são funções periódicas complexas em três dimensões. Imagine-se andando numa linha reta dentro de um cristal. A cada passo você encontrará conjuntos de átomos diferentes em posições diferentes em relação a você. Contudo, depois de andar um pouco você começará a sentir um pouco de *deja vu* ... “... eu já vi aquela histidina e aquela serina antes!” Daria para perceber que um padrão complexo está se repetindo. O que você realmente iria perceber são flutuações nas **densidades** das núvens **eletrônicas**! Caminhando agora numa outra direção constataria que um outro padrão também está se repetindo. Isso ilustra que cada caminhada numa direção diferente produz uma função (padrão) periódica complexa de flutuações de densidade eletrônica. Logo, **um cristal pode ser visto como uma função periódica complexa em três dimensões**. O objetivo da cristalografia é de obter uma função periódica, $\rho(\mathbf{x}, \mathbf{y}, \mathbf{z})$, cujo gráfico representa **o mapa de densidade eletrônica** das moléculas na célula unitária.

Série de Fourier. O matemático francês Jean Baptiste Joseph Fourier demonstrou que qualquer função periódica, independente de sua complexidade, pode ser descrita como uma soma de funções periódicas simples (ex: seno ou co-seno) cujos comprimentos de onda são frações inteiras do comprimento de onda da função complexa original (vide parte 2 da figura acima onde a onda original-alvo (quadrada; “target”) pode ser aproximada pela soma de ondas periódicas simples $f_0 - f_6$).

Em uma dimensão: $f(x) = \sum F_h e^{2\pi i(hx)}$

Em três dimensões: $f(x, y, z) = \sum \sum \sum F_{hkl} e^{2\pi i(hx+ky+lz)}$

Transformada de Fourier. Fourier também demonstrou que para qualquer função $f(x)$ é possível estabelecer uma outra função $F(h)$, tal que:

$$F(h) = \int f(x) e^{2\pi i(hx)} dx$$

A função $F(h)$ é conhecida como a Transformada de Fourier (FT) da função $f(x)$.

Uma das propriedades principais desta função está relacionada à reciprocidade. Se, por exemplo, $f(x)$ é uma função de tempo, $F(h)$ é uma função de frequência e se $f(x)$ for uma função de distância (d), $F(h)$ será uma função de distância recíproca (onde o variável h tem unidades de 1/d).

Desta maneira a FT é uma ferramenta extremamente útil para descrever a relação entre o espaço real e espaço recíproco e logo entre um objeto e seu padrão de difração.

Outra propriedade importante da Transformada de Fourier é que ela é reversível; ou seja:

$$\text{Se } F(h) = \int f(x) e^{2\pi i(hx)} dx, \text{ então } f(x) = \int F(h) e^{-2\pi i(hx)} dh$$

Em outras palavras: se $F(h)$ é a FT de $f(x)$, então $f(x)$ é a FT inversa (FT^{-1}) de $F(h)$.

Em três dimensões:

$$F(h,k,l) = \iiint f(x,y,z) e^{2\pi i(hx+ky+lz)} dx dy dz \quad \text{e} \quad f(x,y,z) = \iiint F(h,k,l) e^{-2\pi i(hx+ky+lz)} dh dk dl$$

Fatores de Estrutura. Cada reflexão hkl no padrão de difração pode ser considerada uma onda complexa que foi formada com contribuições de:

(a) todos os átomos:

$$F_{hkl} = \sum f_j e^{2\pi i(hx_j+ky_j+lz_j)}$$

onde o f_j é o fator de espalhamento do átomo j e $2\pi(hx_j+ky_j+lz_j)$ é a contribuição do átomo j para a fase da reflexão hkl.

ou

(b) todos os elementos de volume da célula unitária:

$$F_{hkl} = \iiint \rho(x,y,z) e^{2\pi i(hx+ky+lz)} dx dy dz$$

Esta onda complexa pode ser descrita como uma série de Fourier cuja soma é chamada o **fator de estrutura** F_{hkl} . A magnitude de F_{hkl} é proporcional à raiz quadrada da intensidade da reflexão hkl $(I_{hkl})^{1/2}$.

A equação anterior mostra que F_{hkl} é a Transformada de Fourier de $\rho(x,y,z)$. Logo, $\rho(x,y,z)$ é a Transformada de Fourier inversa de F_{hkl} :

$$\rho(x,y,z) = (1/V) \sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i(hx+ky+lz)}$$

onde V é o volume da célula unitária e

F_{hkl} é uma série de Fourier cujo resultado final pode ser representado por um número complexo de amplitude ($|F_{hkl}|$) e fase (ϕ_{hkl}).

Pode ser demonstrado que

$$\rho(x,y,z) = (1/V) \sum_h \sum_k \sum_l |F_{hkl}| e^{-2\pi i(hx+ky+lz + \phi_{hkl})}$$

Esta equação mostra como podemos obter a densidade eletrônica em qualquer posição (x,y,z) de uma célula unitária de um cristal a partir do padrão de difração:

As intensidades das reflexões fornecem $|F_{hkl}|$, e os índices fornecem h, k, l.

Mas, como podemos medir as fases de cada uma das milhares de reflexões?

O problema das fases. O problema é que durante o experimento de coletar as imagens do padrão de difração, as informações das fases são perdidas. **A resolução deste problema das fases é um dos passos mais difíceis na determinação de uma estrutura cristalina.**

Existem 3 maneiras principais que podem ser utilizadas para estimar as fases das reflexões. Aquela que você utiliza depende de alguns fatores:

i) *Substituição molecular.* Se dispomos de um modelo estrutural que é suficientemente similar à estrutura da proteína experimental (cristalizada), pode-se utilizar este modelo para calcular $\rho(x,y,z)$ que, por sua vez pode ser utilizada para (através de sua FT) calcular o padrão de difração do modelo (amplitudes dos fatores de estrutura $|F_{hkl}|_{calc}$, incluindo suas fases ϕ_{calc}). Estas fases calculadas podem ser então utilizadas, em combinação com as intensidades observadas experimentalmente ($|F_{hkl}|_{obs}$), para calcular $\rho(x,y,z)$ na célula unitária da proteína que foi cristalizada:

$$\rho(\mathbf{x},\mathbf{y},\mathbf{z}) = (1/V) \sum_h \sum_k \sum_l |\mathbf{F}_{hkl}|_{\text{obs}} e^{-2\pi i(hx+ky+lz + \phi_{\text{calc}})}$$

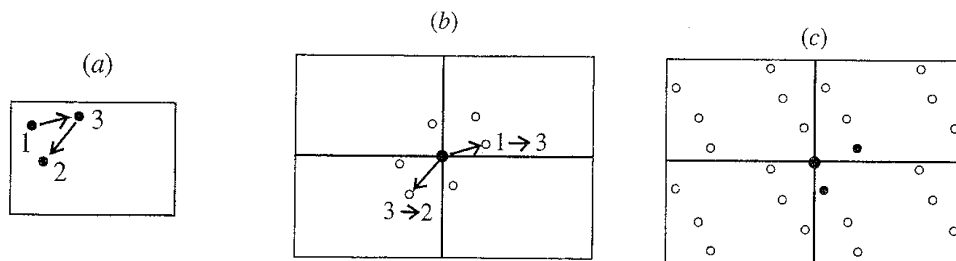
A qualidade das fases calculadas depende da semelhança estrutural entre o modelo e a estrutura sendo estudada. Também existe a questão da **orientação e posição** das moléculas protéicas dentro da célula unitária. Na verdade, a **orientação e posição** das moléculas protéicas do modelo dentro da célula unitária é fundamental para os cálculos dos mapas de densidade eletrônica. A orientação do modelo é definida por três ângulos de rotação e sua posição final será definida pelas translações ao longo dos três eixos da célula unitária. Existem milhões de milhões de possíveis combinações (pelo menos $100^6 = 10^{12}$) destas seis variáveis, o que impossibilita testar todas as possibilidades. O que é feito é: primeiro, determinar a orientação (somente três variáveis <**ângulos eulerianos** α, β, γ >) e depois a posição (mais três variáveis <**translações** ao longo dos três eixos da célula unitária>). Consulte o material bibliográfico para informações mais detalhadas sobre como procurar separadamente as funções de rotação e translação.

Brevemente:

Orientação: Existe uma função chamada **Função de Patterson** que é uma série de Fourier **sem fases** e com termos com amplitudes que são quadrados de fatores de estrutura:

$$\mathbf{P}(\mathbf{u},\mathbf{v},\mathbf{w}) = (1/V) \sum_h \sum_k \sum_l |\mathbf{F}_{hkl}|_{\text{obs}}^2 e^{-2\pi i(hu+kv+lw)}$$

Devido à ausência das fases, uma Função de Patterson pode ser calculada para qualquer conjunto de dados de difração. O ponto importante é que pode ser demonstrado que o gráfico tridimensional $\mathbf{P}(\mathbf{u},\mathbf{v},\mathbf{w})$ da Função de Patterson (chamado de **mapa de Patterson**) tem picos correspondendo aos **vetores entre todos os pares de átomos** na célula unitária. Pensando um pouco sobre isso, temos que para uma célula unitária com N átomos, existem $N(N-1)$ ou quase N^2 vetores quando N for grande. Logo um mapa de Patterson é muito mais complexo do que um mapa de densidade eletrônica. Por exemplo, para uma célula unitária contendo 3 átomos (a), existem seis vetores (b) que produzem o mapa de Patterson mostrado em (c). Pode também ser percebido que **o mapa de Patterson é independente da posição da molécula dentro da célula unitária mas sofrerá rotações se a molécula é rodada**.



Estas considerações apontam a uma maneira em que podemos encontrar a melhor orientação do nosso modelo: calcular mapas de Patterson para várias orientações do nosso modelo (definidas pelos ângulos eulerianos) e compará-los com o mapa de Patterson calculado dos dados experimentais. A melhor sobreposição entre os mapas possibilitará determinar a correta orientação do modelo dentro da célula unitária

Posição: Utilizaremos o modelo na sua melhor orientação para agora posicioná-lo dentro da célula unitária. Em cada posição do modelo podemos calcular as amplitudes dos fatores de estrutura ($|F_{\text{calc}}|$) e compará-las com as amplitudes atualmente observados no experimento de difração ($|F_{\text{obs}}|$). Escolhe-se, dentre todas as opções, aquela em que a soma de todas as diferenças entre $|F_{\text{obs}}|$ e $|F_{\text{calc}}|$ é minimizada:

$$\frac{\sum ||F_{\text{hkl}}|_{\text{obs}} - |F_{\text{hkl}}|_{\text{calc}}|}{\sum |F_{\text{hkl}}|_{\text{obs}}}$$

Vários fatores podem contribuir para o sucesso desta estratégia de calcular as fases; os mais importantes sendo: a) o nível de identidade da sequência primária - normalmente precisa de identidades >30%, b) a ausência de inserções ou deleções em uma estrutura em relação a outra e c) o número de moléculas na unidade assimétrica.

ii e iii) *Métodos dependentes de átomos pesados.* Se não existe um modelo de boa qualidade da nossa estrutura para estimar as fases por substituição molecular, precisamos recorrer a outros métodos para o cálculo das fases dos fatores de estrutura. Estes métodos envolvem a adição de sais de metais pesados ao cristal (ie Pb^{2+} , Hg^{2+} , Au^+ , I^-) ou a incorporação de selenometionina na proteína durante sua produção heteróloga. Além disso, algumas proteínas naturalmente ligam íons metálicos como cofatores (ex.: Fe^{2+} , Zn^{2+} , Mn^{2+}). A adição de átomos mais pesados do que C, N, O ao cristal modificará a padrão de difração em relação ao padrão nativo. Uma vez que a contribuição deste novo átomo para cada uma das reflexões do padrão de difração não se dá de forma igual

(depende da posição deste átomo dentro da célula unitária), espera-se que algumas reflexões sejam mais afetadas do que outras no novo padrão de difração. Logo, é esperado que a adição de um metal pesado irá perturbar algumas reflexões mais do que outras. É justamente esta diferença nas intensidades das reflexões o que possibilita determinar as fases dos fatores de estrutura.

Substituição isomorfa: Comparando os padrões de difração de um **crystal nativo** (obtido sem metal pesado) e de **crystal derivados** (com o metal pesado), as pequenas diferenças nas intensidades de reflexões específicas podem ser utilizadas para determinar as posições dos átomos pesados (mas não dos átomos da proteína). Isso feito, os fatores de estrutura (amplitudes e fases), de todas as reflexões obtidos do cristal com metal pesado ($|F_{PH}|$ e ϕ_{PH}) podem ser separados em contribuições dos átomos pesados ($|F_H|$ e ϕ_H) e da parte protéica ($|F_P|$ e ϕ_P). As variáveis $|F_H|$ e ϕ_H podem ser obtidas pelas posições conhecidos dos átomos pesados, o $|F_P|$ é obtido do padrão de difração do cristal nativo ($|F_{obs}|$; sem átomos pesados) e a desejada ϕ_P pode então ser estimada por procedimentos descritos no Material Bibliográfico e nas aulas. O sucesso desta estratégia depende da obtenção de cristais que incorporam o metal pesado sem perturbar a estrutura da proteína e as dimensões da célula unitária. Em outras palavras, necessita de **crystal derivados isomorfos**. Por isso, esta metodologia é chamada **Substituição Isomorfa**.

Dispersão ou espalhamento anômalo. Átomos podem absorver raios X e emití-los novamente. Este fenômeno acontece somente em comprimentos de ondas específicos. Neste processo de absorção e emissão, ocorre uma mudança de intensidade e de fase dos raios X difratados. Este processo é chamado **dispersão ou espalhamento anômalo**. Logo, se conjuntos de dados do mesmo cristal são coletados em dois comprimentos de onda, um dos quais correspondente ao comprimento de absorção de raios X pelo metal pesado, as mudanças nos padrões de difração podem ser utilizadas para obter informações sobre as fases (ver Material Bibliográfico discutido nas aulas). Além disso, na ausência de dispersão anômala as intensidades e as fases das reflexões hkl e $-h-k-l$ são iguais. Na presença de dispersão anômala esta identidade desapareça (por razões que também são explicados no Material Bibliográfico). As diferenças entre as reflexões $|F_{hkl}|$ e $|F_{-h-k-l}|$ são muito pequenas (normalmente somente alguns % da intensidade) mas podem ser medidas porque são obtidas com alta precisão do mesmo conjunto de dados do mesmo cristal. Estas diferenças também podem ser utilizadas para obter estimativas iniciais sobre as fases.

Obtenção do mapa de densidade eletrônica e refinamento do modelo molecular.
Independentemente do método que você utilizou para obter a primeira estimativa das fases, seu primeiro mapa de densidade eletrônica será calculado simplesmente a partir da seguinte equação:

$$\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}) = (1/V) \sum \sum \sum |\mathbf{F}_{\mathbf{hkl}}|_{\text{obs}} e^{-2\pi i(\mathbf{hx} + \mathbf{ky} + \mathbf{lz} + \phi_{\text{calc}})}$$

(Obs.: Alguns fatores que expressam pesos de qualidade podem ser adicionados à formula sem alterar o real sentido da mesma.)

A inspeção deste mapa deve então apontar regiões que são obviamente derivadas de proteínas em que a cadeia principal (e se tiver sorte algumas cadeias laterais) pode ser reconhecida. Nestas regiões, podemos criar modelos tridimensionais utilizando nossos conhecimentos da sequência primária da proteína específica e nossos conhecimentos básicos de química de proteínas em geral (ângulos e distâncias de ligações químicas, modelos eletrônicos de átomos específicos, principalmente C, N, O e S). Estes modelos podem ser utilizados para calcular as intensidades e fases das reflexões num **processo iterativo** em que procuramos fases calculadas melhores que diminuam as diferenças entre $|\mathbf{F}_{\mathbf{hkl}}|_{\text{obs}}$ e $|\mathbf{F}_{\mathbf{hkl}}|_{\text{calc}}$.

Cada iteração envolve passos de: a) inspeção e manipulação do modelo para que ela encaixe melhor no mapa de densidade eletrônica, b) a construção de novas regiões da molécula em regiões de densidade eletrônica que eram ausentes no ciclo anterior, c) pequenos ajustes no modelo que procuram ao mesmo tempo aproximar valores geométricos das ligações químicas às valores ideais e minimizar a diferença entre $|\mathbf{F}_{\mathbf{hkl}}|_{\text{obs}}$ e $|\mathbf{F}_{\mathbf{hkl}}|_{\text{calc}}$, d) cálculo de novas fases derivadas do novo modelo e e) cálculo de um novo mapa de densidade eletrônica usando as novas fases.

Notas importantes:

Veja que mapas do tipo

$$\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}) = (1/V) \sum_{\mathbf{h}} \sum_{\mathbf{k}} \sum_{\mathbf{l}} ||\mathbf{F}_{\mathbf{hkl}}|_{\text{obs}} - |\mathbf{F}_{\mathbf{hkl}}|_{\text{calc}}| e^{-2\pi i(\mathbf{hx} + \mathbf{ky} + \mathbf{lz} + \phi_{\text{calc}})}$$

chamadas **mapas de diferença** ou **mapas Fo-Fc** enfatizam erros no modelo, com regiões positivas onde está faltando densidade eletrônica (destacando a falta de um átomo ou átomos no modelo

atual) e regiões negativas evidenciam um excesso de elétrons no seu modelo naquela região. Estes mapas são úteis na correção de pequenos erros na estrutura no final do processo de refinamento.

De maneira similar, mapas do tipo

$$\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}) = (1/V) \sum_h \sum_k \sum_l |2|\mathbf{F}_{hkl}|_{\text{obs}} - |\mathbf{F}_{hkl}|_{\text{calc}}| e^{-2\pi i(hx + ky + lz + \phi_{\text{calc}})}$$

chamadas **mapas 2Fo-Fc** são utilizados para visualizar a qualidade do encaixe do modelo com os dados. Estes mapas devem ser sempre positivos e, escolhendo contornos acima de uma densidade eletrônica específica, devem revelar uma superfície molecular.

Note que ao final do refinamento, quando não mais houver diferença entre os dados experimentais e os dados do modelo, o primeiro mapa tenderá a zero enquanto o mapa segundo tenderá a ser calculado apenas com as amplitudes experimentais

Em cada ciclo de refinamento do modelo, monitoramos quanto o modelo está consistente com os dados experimentais, calculando o fator residual (R)

$$R = \frac{\sum ||\mathbf{F}_{hkl}|_{\text{obs}} - |\mathbf{F}_{hkl}|_{\text{calc}}|}{\sum |\mathbf{F}_{hkl}|_{\text{obs}}}$$

A grande maioria das estruturas nos bancos de dados tem valores de $R < 0.3$ e muitos têm valores < 0.2 .

Material Bibliográfico:

Rhodes, G. (2006) Crystallography Made Crystal Clear. 3rd Edition (Academic Press).

Textos mais avançados:

Rupp, B. (2010) Biomolecular Crystallography (Garland Science)

Drenth, J. (1999) Principles of Protein X-Ray Crystallography (Springer).

Blow, D. (2002) Outline of Crystallography for Biologists (Oxford University Press).

McPherson, A. (2009) Introduction to Macromolecular Crystallography (Wiley-Blackwell)

DETERMINAÇÃO DA ESTRUTURA DE LISOZIMA DE CLARA DE OVO (HEWL) POR CRISTALOGRAFIA

Cristalização. A lisozima de clara de ovo de galinha (20 mg/ml) foi cristalizada pela técnica de difusão de vapor em poços de gota sentada em equilíbrio com uma solução de 30% w/v polietileno glicol (PEG) 4000, 1.0 M NaCl, 50 mM acetato de sódio, pH 4.6

Coleta de dados. A coleta de dados foi feita no difratômetro do IQ-USP (bloco 1 superior) usando um dos cristais de lisozima produzidos na etapa de cristalização. O arquivo **lysozyme_intensities.ref** tem todos os dados coletados para o nosso cristal de lisozima. O arquivo **datacollection_log.txt** sumariza alguns aspectos do conjunto de dados coletados para nosso cristal de lisozima.

Resumo do procedimento de estimação inicial das fases por substituição molecular e subsequentes refinamento e validação do modelo.

Passos 3 – 6 nos Exercícios em baixo referem ao procedimento de cálculo inicial das fases por **substituição molecular** e o cálculo do primeiro mapa de densidade eletrônica usando o programa **Phaser MR** dentro do pacote **Phenix**. As amplitudes e as fases no arquivo **lysozyme_0.mtz** (o *zero* significa que é o nosso primeiro arquivo com fases) permite que construimos um mapa de densidade eletrônica **2Fo-Fc**. Este mapa é representado com linhas de contorno azuis no programa **Coot** (<http://www.ysbl.york.ac.uk/~lohkamp/coot/wincoot-download.html>). Neste mapa podemos superimpor o modelo descrito no **lysozyme_0.pdb** que é um modelo de lisozima sem cadeias laterais corretamente posicionado e orientado pelo procedimento de substituição molecular. Este modelo foi utilizado para calcular as fases (**PHIC**). Coot também mostra um mapa de calculado usando as diferenças **Fo-Fc** para enfatizar regiões que têm densidade eletrônica demais (contornos verdes; ou seja, necessita mais átomos) ou de menos (contornos vermelhos; ou seja, necessita remover átomos).

Passos 7 – 10 nos Exercícios em baixo referem ao procedimento de sucessivos ajustes do modelo para melhor encaixar no mapa de densidade eletrônica, usar o novo modelo para calcular novas fases que são utilizados para gerar um novo mapa de densidade eletrônica que é então usado para de novo ajustar o modelo.

Além de permitir visualização de mapas, o Coot permite que modifiquemos manualmente a conformação do nosso modelo para melhor encaixá-lo na densidade eletrônica. Assim, podemos inspecionar resíduo por resíduo, adicionando as cadeias laterais e ajustando a conformação local da cadeia principal e das cadeias laterais. Ao final deste processo de ajuste salvamos a nova conformação do modelo em um novo arquivo pdb.

Se os dados coletados apresentam resolução melhor do que $\sim 2.3 \text{ \AA}$, seremos capazes evidenciar a densidade eletrônica derivada de moléculas de água na primeira camada de solvatação da proteína. Estas moléculas estão normalmente fazendo pontes de hidrogênio com grupos da proteína e/ou outras moléculas de água. Estas moléculas podem ser adicionadas ao modelo e aparecerão no novo arquivo .pdb.

O próximo passo consiste em ajustar parâmetros geométricos (ângulos e comprimentos de ligações químicas e contatos van der Waals) no modelo para que não fogem dos princípios químicos conhecidos para proteínas derivadas de estruturas de alta resolução de moléculas orgânicas pequenas. Brevemente, queremos efetuar ciclos de mudanças nos parâmetros de posição (x, y, z) e fatores de temperatura de todos os átomos do modelo para melhor aproximar os fatores de estrutura calculados (**F_c** calculados a partir do modelo) aos fatores de estrutura observados (**F_o**). Em adição, parâmetros estereoquímicos do modelo como comprimentos de ligações, ângulos de ligações e de torsão e contatos de van der Waals são permitidos variar somente em torno de valores padrões determinados com base em nosso conhecimento de estruturas químicas. Para incorporar todos estes parâmetros, uma equação do seguinte tipo deve ser minimizada:

$$\begin{aligned} \Phi = & \sum \sum \sum \mathbf{w}_{\mathbf{hkl}} (|\mathbf{F}_{\mathbf{hkl}}|_{\text{obs}} - |\mathbf{F}_{\mathbf{hkl}}|_{\text{calc}})^2 \\ & + \sum_{\text{bonds}} \mathbf{w}_i (\mathbf{d}_{\text{ideal}} - \mathbf{d}_{\text{model}})^2 \\ & + \sum_{\text{angulos}} \mathbf{w}_j (\theta_{\text{ideal}} - \theta_{\text{model}})^2 \\ & + \sum_{\text{planos}} \sum_{\text{atomos}} \mathbf{w}_{a,b} (\mathbf{m}_b \cdot \mathbf{r}_{ab} - \mathbf{d}_b)^2 \\ & + \sum_{\text{noncov}} \mathbf{w}_c (\mathbf{d}_{\text{ideal}} - \mathbf{d}_{\text{model}})^4 \\ & + \sum_{\text{Hbond}} \mathbf{w}_d (\mathbf{d}_{\text{ideal}} - \mathbf{d}_{\text{model}})^2 \\ & + \sum_{\text{torsion}} \mathbf{w}_t (\theta_{\text{ideal}} - \theta_{\text{model}})^2 \end{aligned}$$

onde a primeira linha define quanto o modelo aproxima dos dados experimentais e em que $|\mathbf{F}_{\mathbf{hkl}}|_{\text{calc}}$ é uma função de **x, y, z, B e ocupância** definida para cada átomo no arquivo.pdb. As outras linhas descrevem as “penalidades” dos desvios de orientações (distâncias, ângulos) ideais para ligações químicas. Existem vários programas para efetuar esta minimização. Nós utilizaremos o programa **Phenix.refine** dentro do Phenix.

Os arquivos de entrada para o programa Phenix.refine são os arquivos **.mtz** que tem os fatores de estrutura observadas (**F_o**) para cada reflexão e último arquivo **.pdb** salvo que tem os coordenados dos átomos do modelo ajustado (este arquivo **.pdb** será usado pelo programa para calcular **F_c** e a fase (**PHIC**) para cada reflexão). Os arquivos de saída são dois novos arquivos: um **.mtz** (com **F_o** e **F_c** as fases novas calculadas a partir da última versão da estrutura/modelo) e um **.pdb** (com as novas coordenadas da estrutura). Também é produzido um arquivo **.log** que apresenta, além de muitas outras informações, alguns parâmetros globais que são calculados depois de cada ciclo e que refletem quanto o modelo é consistente com os dados experimentais. Deve-se prestar atenção especial aos parâmetros **R_{fact}** e **R_{free}**. Ambas são medidas da qualidade de correlação entre **F_c** e **F_o** e são calculados como o **R** descrito anteriormente, mas o **R_{free}** é calculado utilizando somente as reflexões do conjunto teste (5%) que não foram empregadas antes no refinamento ou para produzir o mapa de densidade eletrônica.

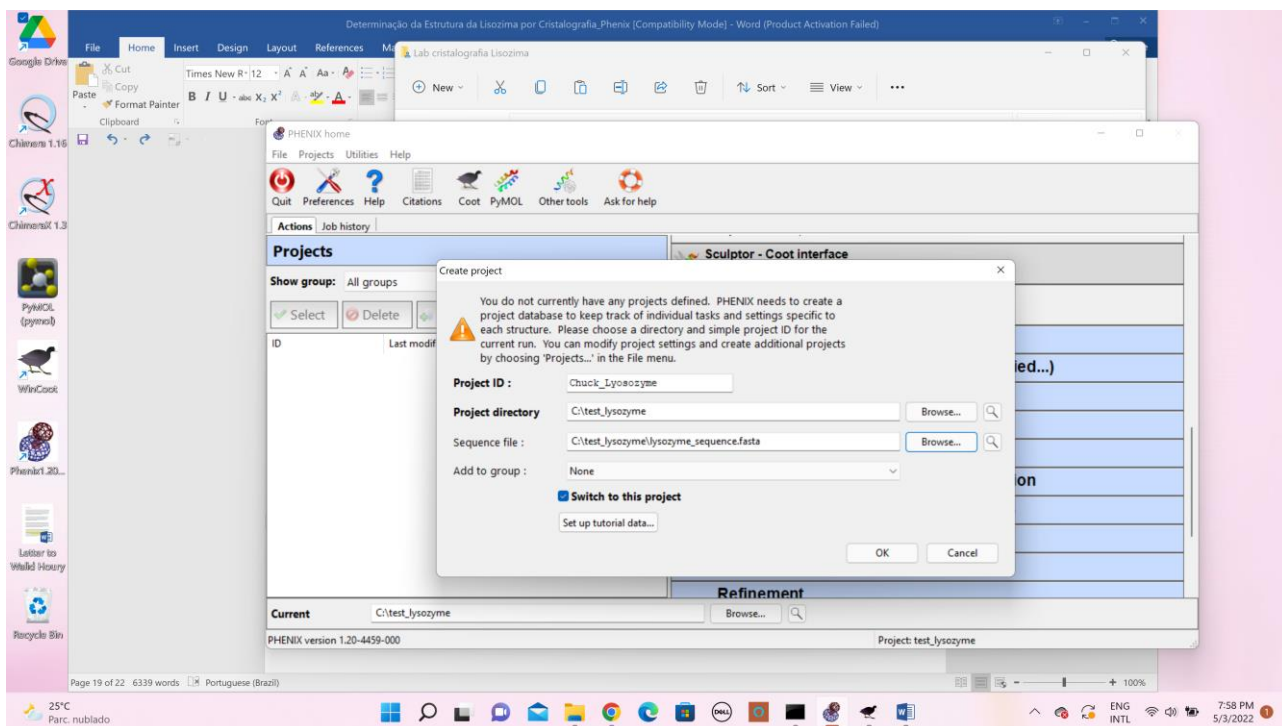
Validação da Estrutura. Depois de vários ciclos/iterações de ajuste manual (refinamento no espaço real) e minimização do Φ (refinamento no espaço recíproco) os valores de **R_{fact}** e **R_{free}** param de diminuir. Neste momento entramos na etapa final de determinação da estrutura em que os últimos desvios de geometria ideal são inspecionados com cuidado. Para isso, prestamos mais atenção nos seguintes fatores (todos podem ser analisados dentro do programa Coot):

- a) ângulos de torção da cadeia principal (ϕ, ψ) fora das regiões permitidas no diagrama de Ramachandran.
- b) desvios de planaridade das ligações peptídicas (ω).
- c) ligações peptídicas na conformação *cis*.
- d) ângulos e ligações com valores muito maior ou muito menor de valores padrões para aquelas ligações.
- e) conformações “raras” das cadeias laterais.
- f) conformações “múltiplas” de cadeias laterais.
- g) contatos próximos demais de van der Waals (“close contacts”).
- h) a presença de moléculas de água sem contato direto ou indireto com a proteína.
- i) “ocupâncias” de átomos menores do 1.0.

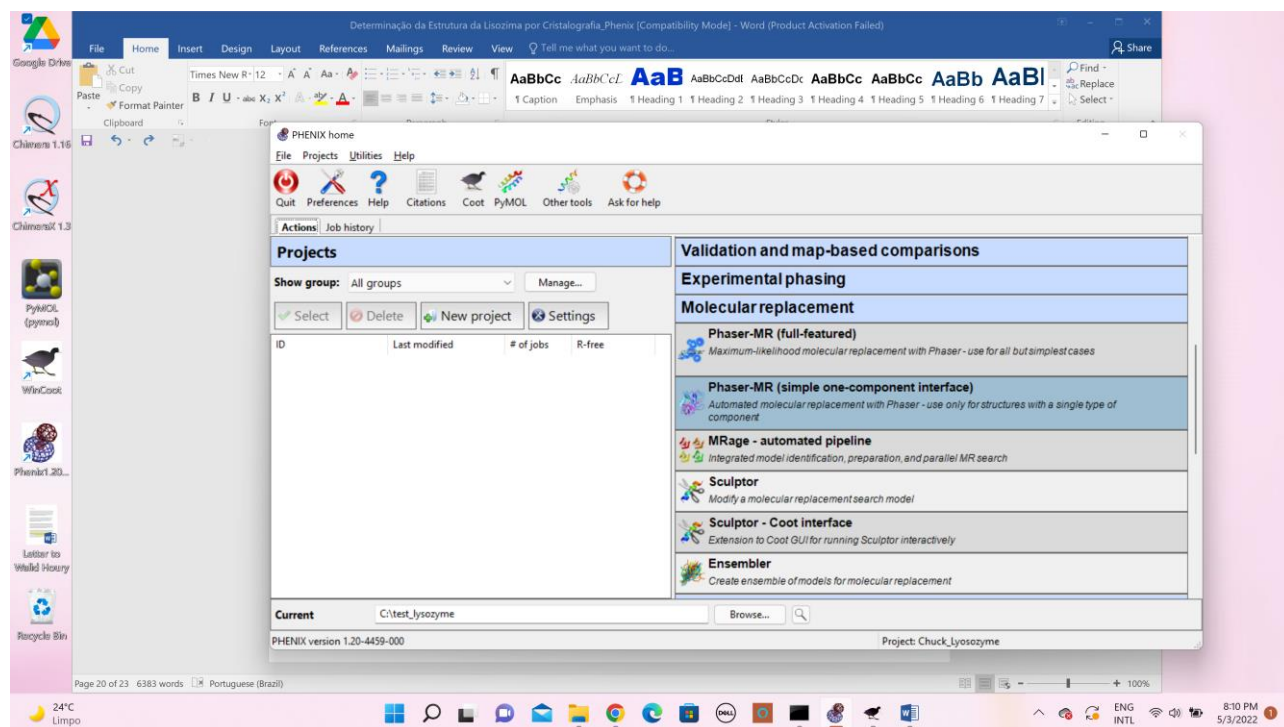
O Protein Data Bank também providencia um serviço online (<https://validate-rcsb-2.wwpdb.org/>) que analisa alguns destes aspectos do seu modelo e produz uma lista de possíveis “problemas” que devem ser checados um a um e corrigidos (se for possível). Esta validação também pode ser feita mais através dentro do pacote Phenix.

EXERCÍCIOS e QUESTÕES

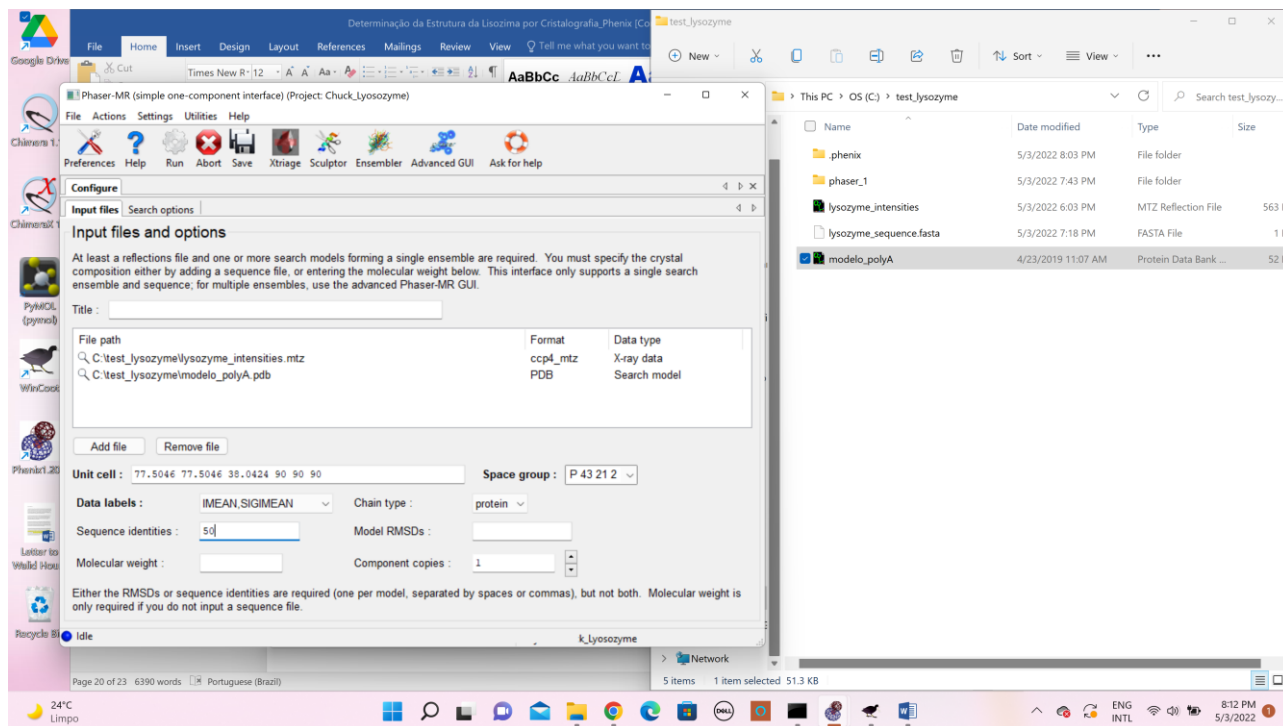
- 1) Já tem várias estruturas da lisozima de gema de ovo de galinha (“hen egg white lysozyme”, HEWL) depositado no Protein Data Base. Faz uma busca e determinar quantos são.
- 2) Vamos empregar uma destas estruturas como um modelo para obter uma estimativa inicial das fases para resolver a estrutura da lisozima cristalizado em nosso laboratório.. Um arquivo pdb contendo as coordenadas dos átomos da cadeia principal desta estrutura (sem as cadeias laterais) foi fornecido com o nome **modelo_polyA.pdb** (pois todos os aminoácidos foram convertidos em alaninas).
- 3) A) O arquivo **datacollection_log.txt** sumariza alguns aspectos do conjunto de dados coletados para nosso cristal de lisozima.
 - i) Qual é o grupo espacial do cristal de lisozima?
 - ii) Quantas reflexões em total e quantas reflexões únicas (unique reflections) foram medidas no experimento de difração de raios X? A partir destes dados podemos calcular quantas vezes, em média, foi medida a intensidade de cada reflexão única. Isso é chamada a “redundância”. Qual é o valor de redundância para o conjunto de dados inteiro e para as reflexões correspondendo à maior resolução?
 - iii) Qual é a maior resolução do conjunto de dados? Qual é o valor de erro sobre ruído (I/σ) para as reflexões correspondendo à maior resolução? Qual é o valor de R_{factor} (ou R_{sym}) para as reflexões correspondendo à maior resolução? Discutir a significância destes valores para as reflexões correspondendo à maior resolução. Você acha que estes dados devem ser mantidos ou descartados? Porque?B) O arquivo **lysozyme_intensities.ref** tem todos os dados coletados para o nosso cristal de lisozima. Abrir o arquivo e descrever seu conteúdo. Para prosseguir temos que converter as intensidades (I) neste arquivo em fatores de estrutura observados (F_o). Também precisamos reformatar o arquivo para o formato **.mtz**. Para isso, usamos o programa **dtrek2mtz** dentro de CCP4i (gerando um novo arquivo com nome **lysozyme_intensities.mtz**).
- 4) Para desenvolver nosso projeto, vamos empregar o pacote de Phenix (quando instalado deve também instalar a ultima versão do Coot). Criar um novo projeto, definindo o nome do projeto, o diretório onde vai guardar os arquivos lidos e gravados e dando endereço do arquivo que contem a sequencia da proteína (**lysozyme_sequence.fasta**).



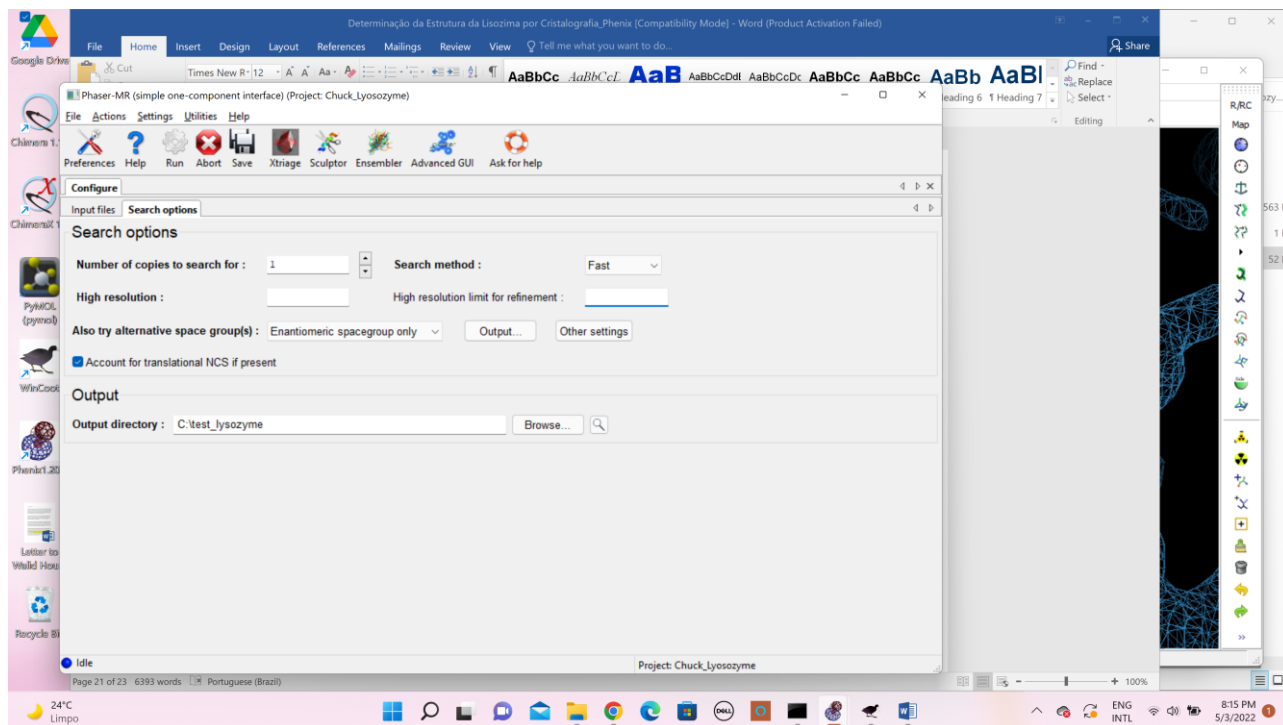
Primeiro executamos o procedimento de Molecular Replacement (substituição molecular). Para isso, empregaremos o programa Phaser for MR dentro de Phenix.



O input do phaser será os arquivos **lysozyme_intensities.mtz** e **modelo_polyA.pdb**.



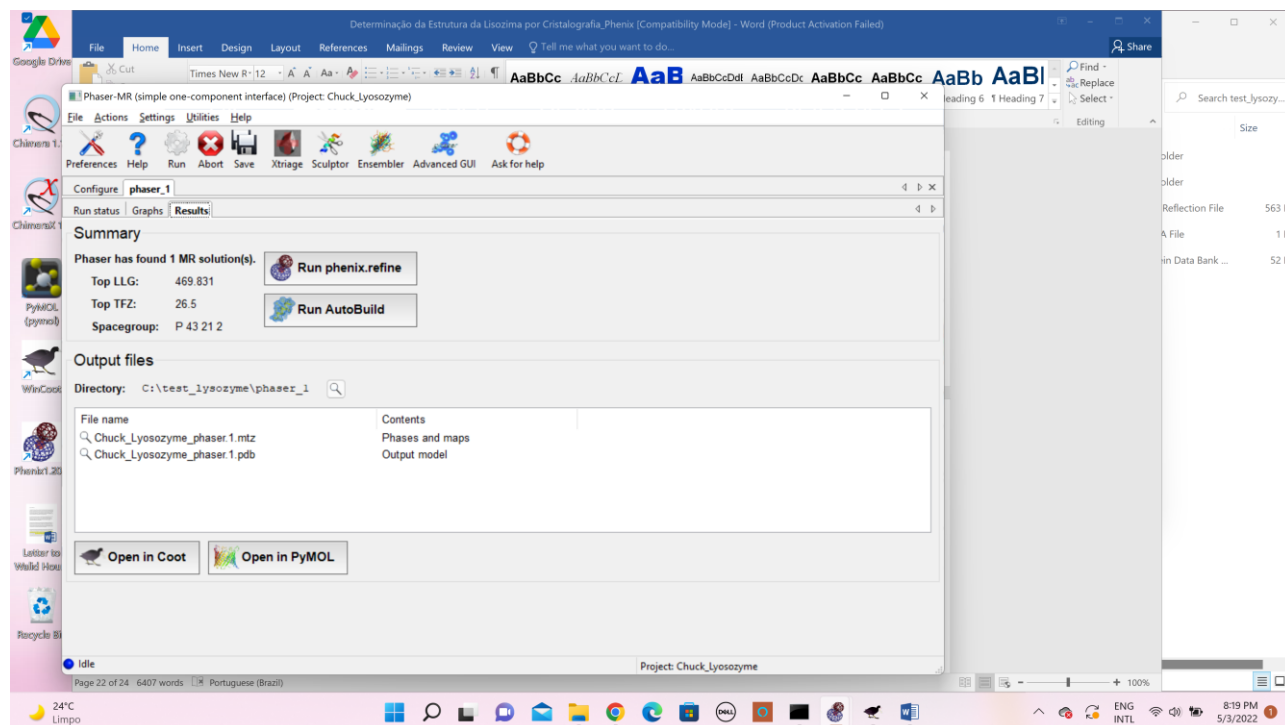
Pode deixar a aba de search options como está



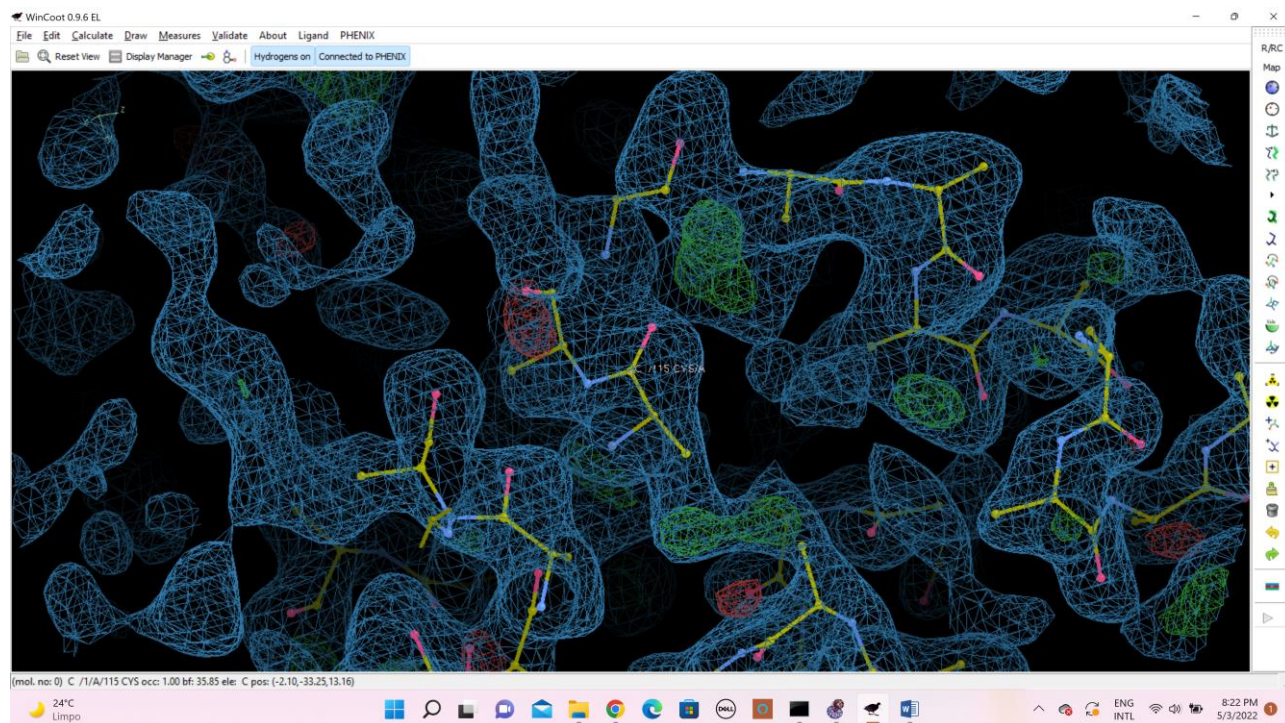
Apertar Run para rodar Phaser.

O output de phaser vai ser dois novos arquivos: um com extensão **.pdb** que tem o nosso modelo agora corretamente orientado (pela função de rotação) e posicionado (pela função de translação) e outro arquivo com extensão **.mtz** que terá os dados originais com mais

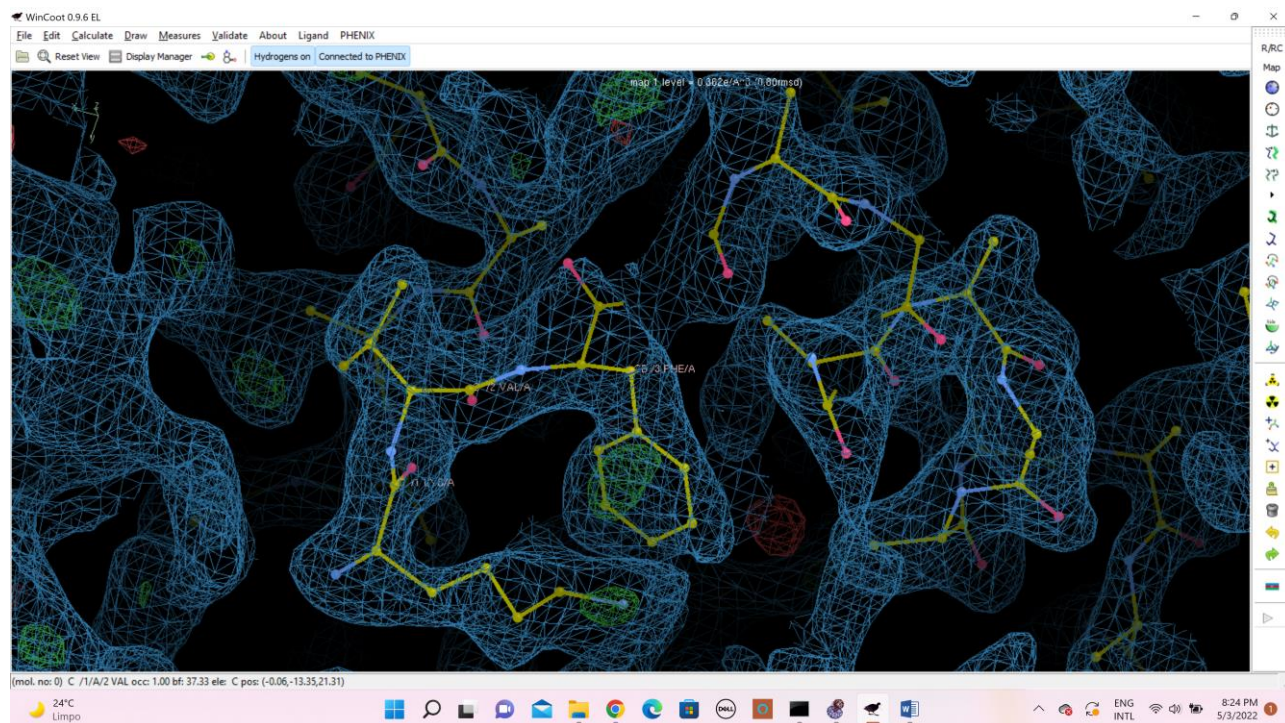
algumas colunas contendo fatores de estrutura calculados a partir do modelo (**Fc**) e fases calculados a partir do modelo (**PHIC**).



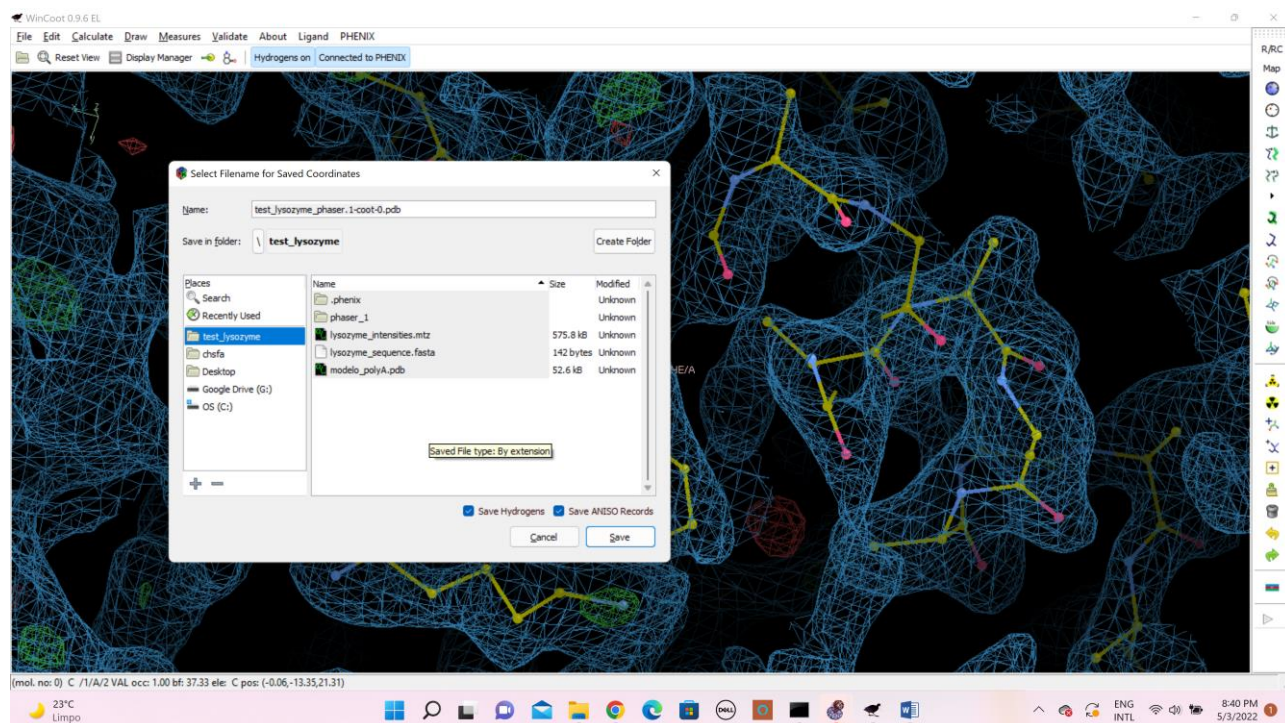
- 5) Abrir o programa Coot e carregar os arquivos **.pdb** e **.mtz** de produzidos pelo Phaser na etapa 4 (ou abrir Coot diretamente dentro de Phenix). Após carregar o arquivo **.mtz**, aparecerá um mapa de densidade eletrônica **2Fo-Fc** representado com linhas de contorno azuis. Após carregar o arquivo **.pdb** aparecerá um modelo de lisozima (sem cadeias laterais) corretamente posicionado e orientado pelo procedimento de Substituição Molecular do Programa Phaser MR. Também é apresentado o mapa calculado usando as diferenças **Fo-Fc** para enfatizar regiões que têm densidade eletrônica demais (contornos verdes) ou de menos (contornos vermelhos).



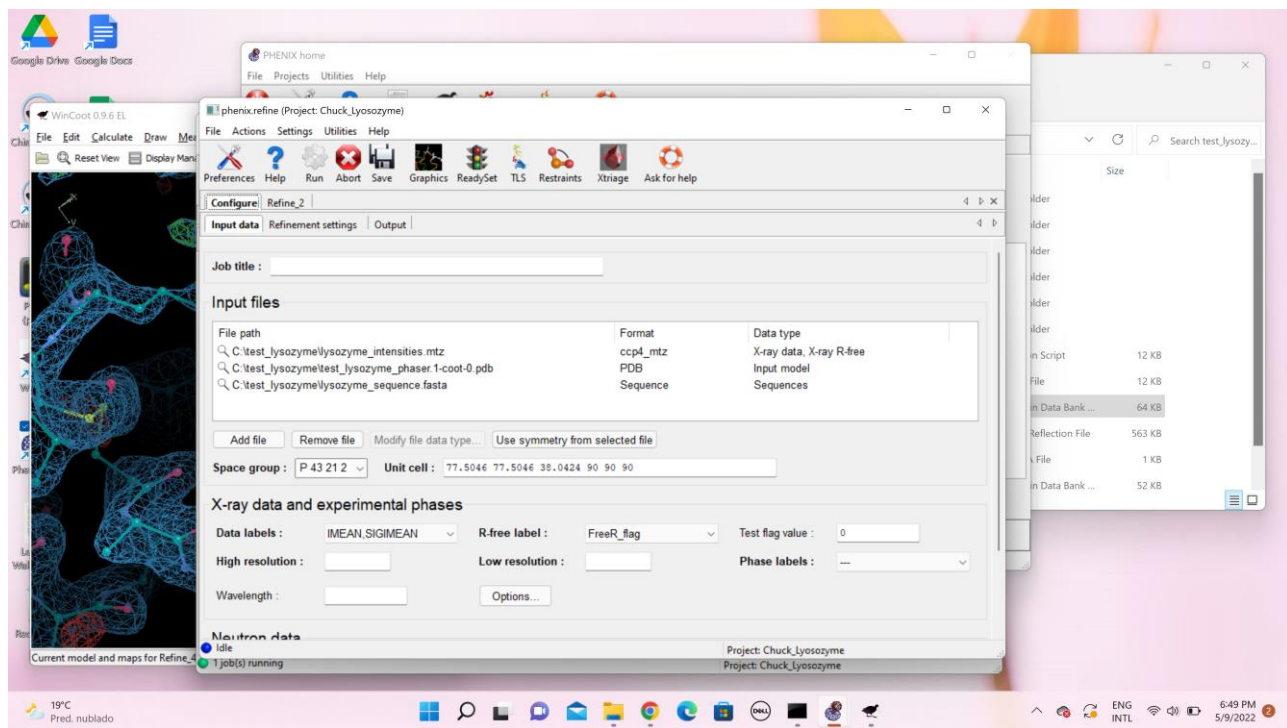
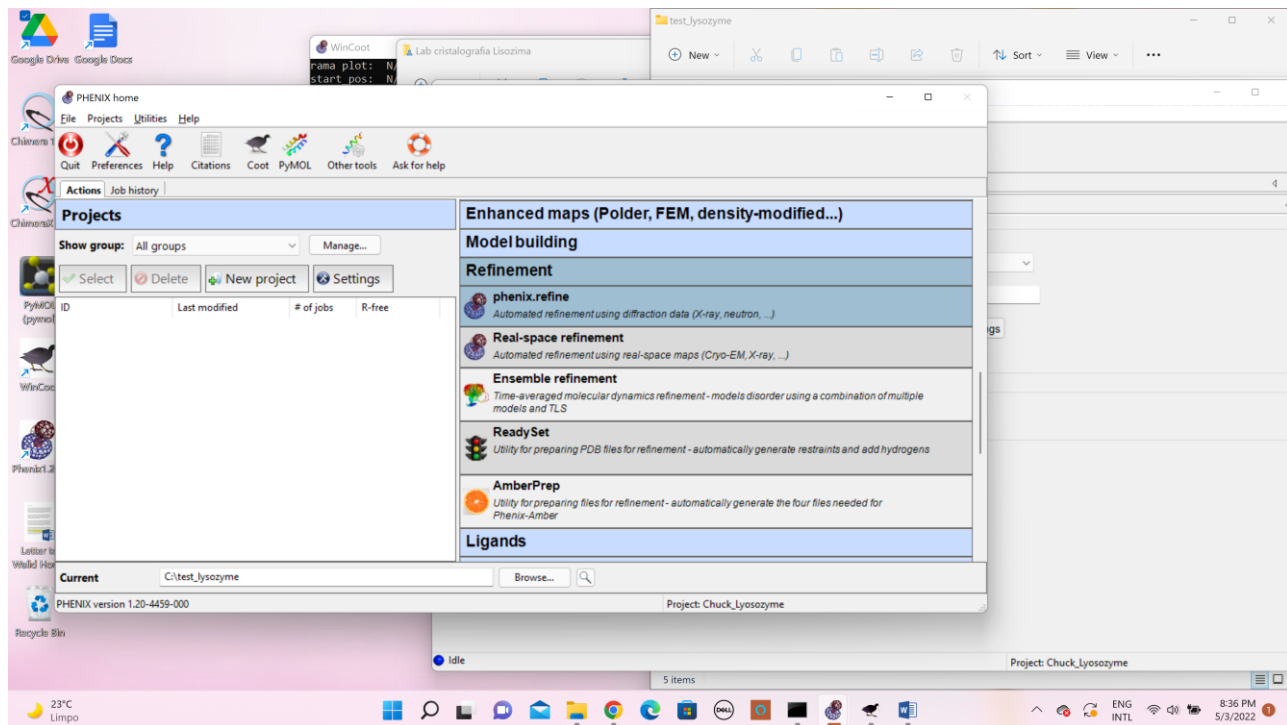
- 6) Além de permitir visualização de mapas, o Coot permite que modifiquemos manualmente a conformação do nosso modelo para melhor encaixá-lo na densidade eletrônica. Assim, podemos inspecionar o modelo, resíduo por resíduo dentro de nossa estrutura, adicionando as cadeias laterais e ajustando a conformação local da cadeia principal e das cadeias laterais.



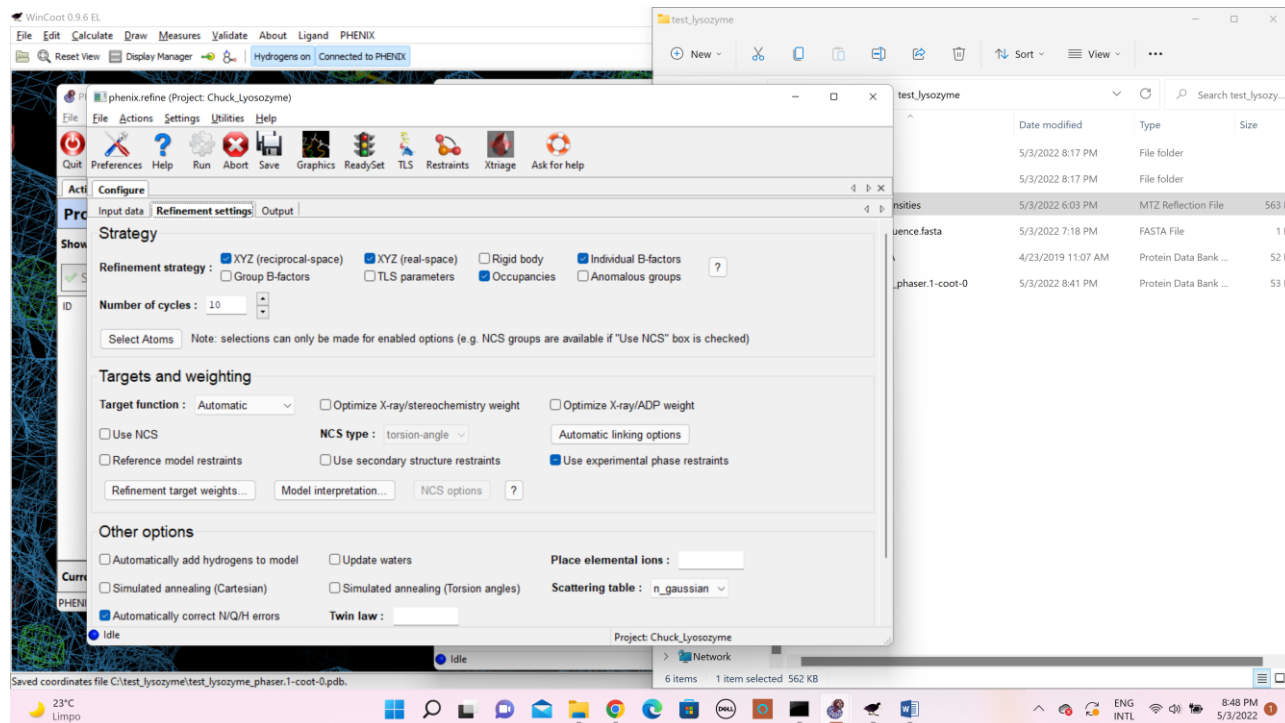
Ao final deste processo de ajuste salvamos (save coordinates) a nova conformação do modelo em um novo arquivo **modelo_coot-0.pdb**.



- 7) a) Rodar o programa **PhenixRefine** dentro da pacote Phenix usando o modelo salvo na etapa 6 como input e o arquivo **.mtz**.

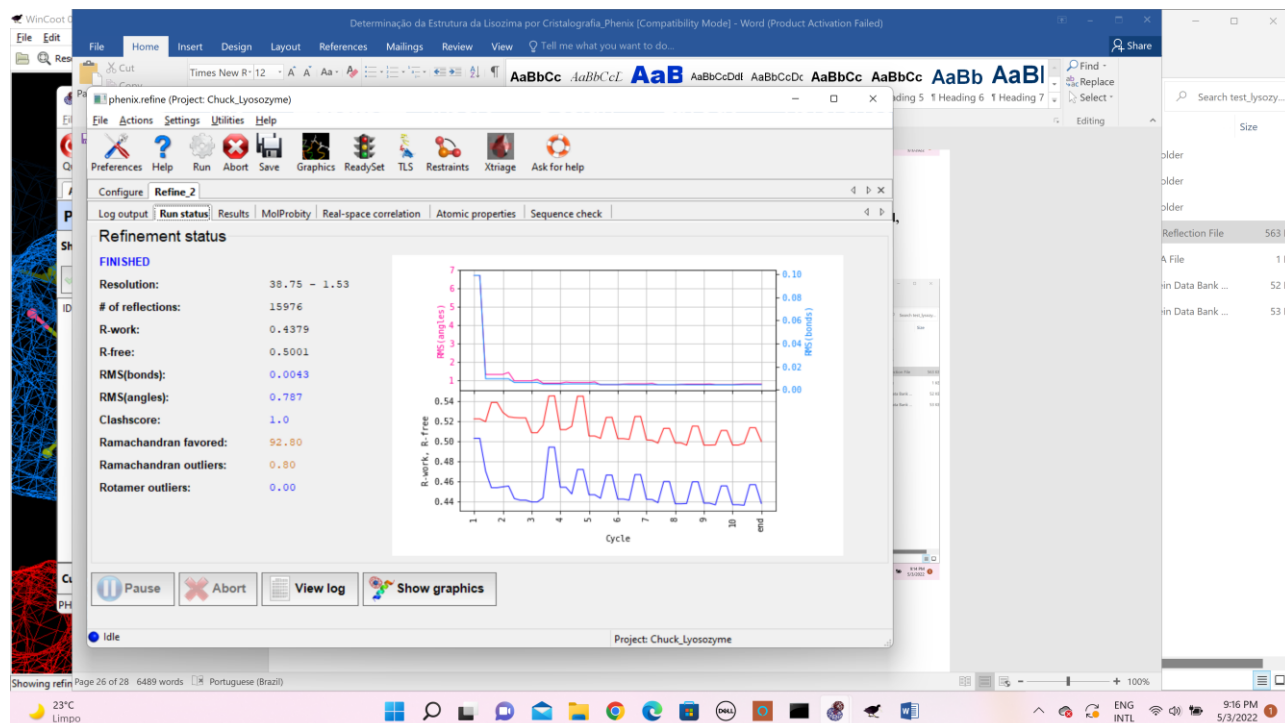


Mudar o número de ciclos de 3 para 10 na aba de refinement settings.



Rodar Phenix.refine (Run).

Isso vai gerar três novos arquivos – um arquivo **.log**, um **.pdb**, e um **.mtz**, este com fases novas calculadas baseadas no novo modelo. Anotar o valor de **R-work**, **R-free**, **RMS-bond**, **RMS-angle**, **Ramachandran favored** e **Ramachandran outliers** no final do ciclo 10.

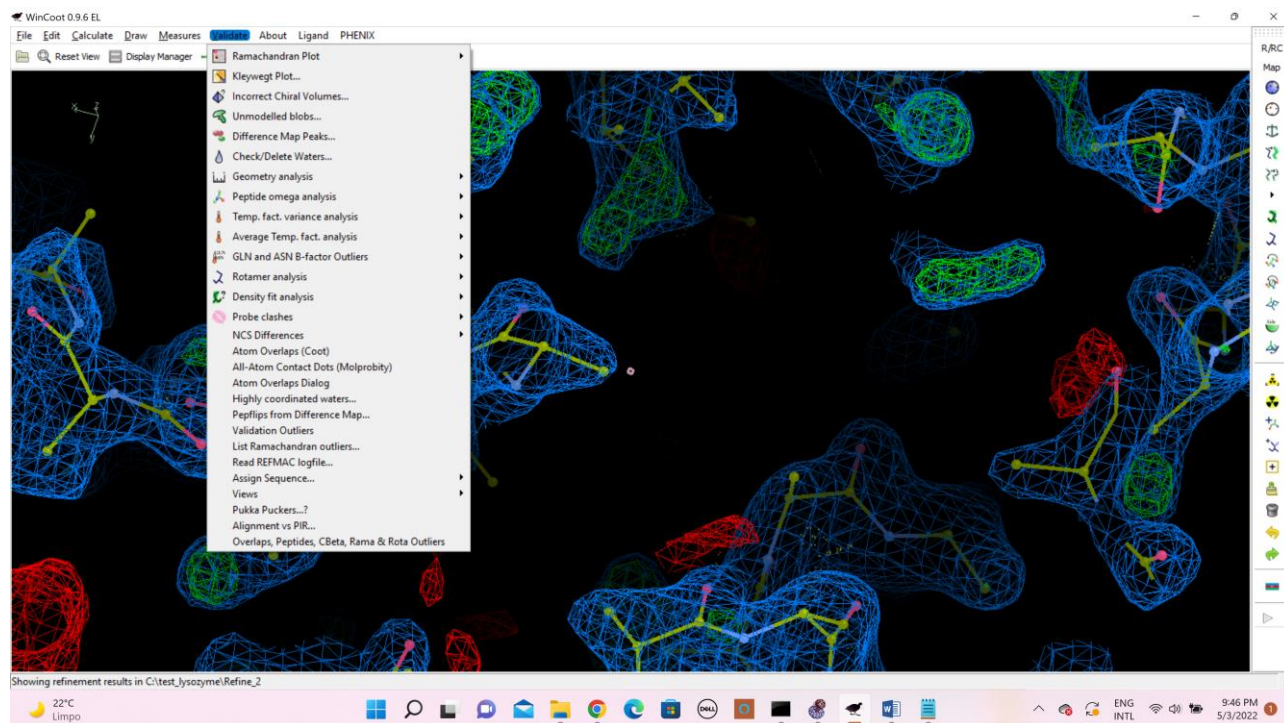


b) Abrir os novos arquivos **.pdb** e **.mtz** no Coot e inspecionar cada resíduo do modelo, introduzindo ajustes quando for necessário.

- REPETE 7a e 7b quantas vezes necessária até você está satisfeito com o encaixe do seu modelo dentro do mapa de densidade eletrônica. Após cada rodada de Phenix-refine, anotar os valores de **R-work**, **R-free**, **RMS-bond**, **RMS-angle**, **Ramachandran favored** e **Ramachandran outliers** no final do ciclo 10. Estes valores devem ser reportados e discutidos no seu relatório. Idealmente, **R-work** e **R-free** devem diminuir paralelamente ao longo do projeto e, no final, não devem diferir por mais do que 0,05.

8) Usando o programa Coot adicionar águas ao modelo.

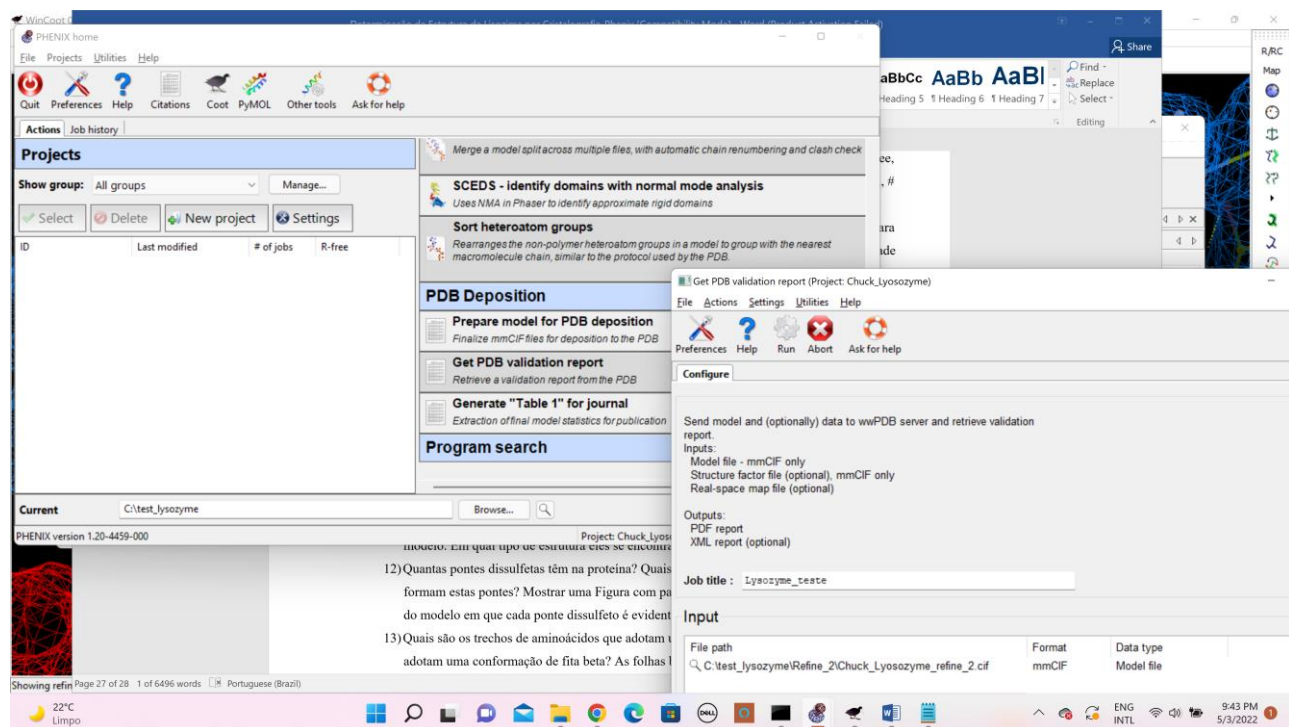
Rodar o programa Phenix-Refine de novo. Repete etapas 7a e 7b mais algumas vezes até você está satisfeito com a correspondência do seu modelo com o mapa de densidade eletrônica. O mapa **Fo-Fc** deve ter poucas regiões com densidade significativa (acima de 3 sigmas) positiva (verde) ou negativa (vermelha). Dentro do coot pode usar as ferramentas dentro da aba “Validate” para checar a qualidade de vários parâmetros geométricos do seu modelo. Isso ajuda em ir direto para os lugares mais “problemáticos”.

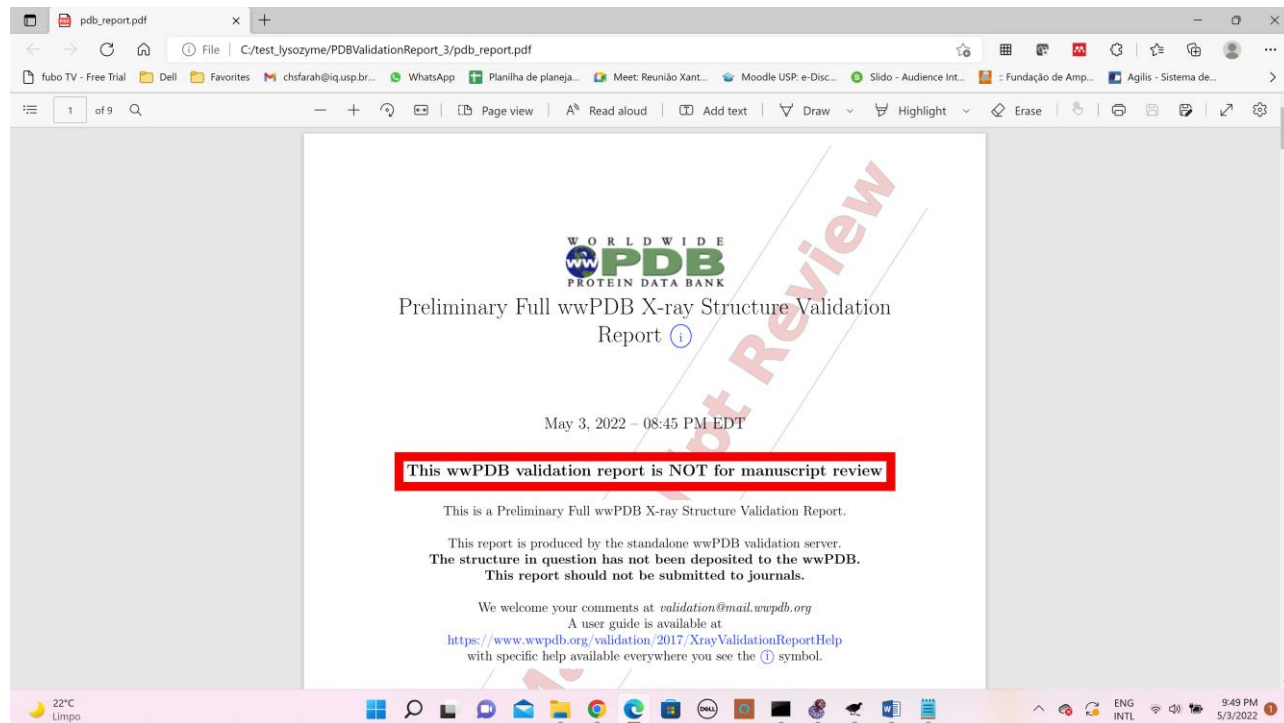


Produzir uma tabela (que mostra a evolução dos valores de **R-work**, **R-free**, **RMS-bond**, **RMS-angle**, **Ramachandran favored** e **Ramachandran outliers** no final de cada rodada de refinamento, indicando em qual rodada foi adicionada as moléculas de água pela primeira vez.

ANALISAR SEU MODELO E MAPA RESPONDENDO ÀS SEGUINTE PERGUNTAS

- 9) a) O arquivo **datacollection_log.txt** sumariza alguns aspectos do conjunto de dados coletados para nosso cristal de lisozima. Produzir uma tabela “**Estatística de Coleção de Dados**” que sumariza:
- grupo espacial, dimensões da célula unitária, resolução, comprimento de onda, # de reflexões, completeza, redundância, I/signal, R_{sym}.
- b) Produzir uma outra tabela “**Estatística de Refinamento**” onde sumariza R-work, R-free, RMS-bond, RMS-angle, Ramachandran favored, Ramachandran outliers do modelo final, # de átomos de proteína, # de átomos de H₂O. Discutir.
- 10) Produzir um relatório de validação (Get PDB Validation Report”) do seu modelo final. Para isso precisa usar o arquivo **.cif** em vez de **.pdb**.





Baseado neste relatório, analisar a qualidade do seu modelo e produz uma lista de possíveis “problemas” ainda não resolvidos. Discutir.

- 11) Produzir o gráfico Ramachandran do seu modelo (usando o programa Coot ou Procheck) e analisá-lo. Há aminoácidos em regiões não permitidas? Quais são? Localizar eles no seu modelo. Em qual tipo de estrutura eles se encontram? Discutir.
- 12) Quantas pontes dissulfetas têm na proteína? Quais são os pares dos aminoácidos que formam estas pontes? Mostrar uma Figura com partes A, B, C etc em que apresenta vistas do modelo em que cada ponte dissulfeto é evidente.
- 13) Quais são os trechos de aminoácidos que adotam uma conformação de alfa hélice? E quais adotam uma conformação de fita beta? As folhas betas são paralelas ou antiparalelas? Produzir uma Figura em Pymol ou Chimera em que as alfa hélices e fitas beta são coloridas diferentemente. Numerar estas estruturas secundárias ($\alpha 1$, $\alpha 2$, $\beta 1$, $\beta 2$, etc).
- 14) Há algumas cadeias laterais cuja conformação não está bem definida? Explicar esta observação. Há algumas cadeias laterais que aparentemente adotam 2 ou mais conformações distintas? Faz uma ou mais figuras mostrando estes casos e discutindo eles.
- 15) É fácil identificar pontes de hidrogênio na estrutura? Produzir uma Figura que aponta com setas, pontes de hidrogênio entre: A) doador e aceitor de grupos da cadeia principal, B) doador da cadeia principal e aceitor da cadeia lateral, C) doador e aceitor de duas cadeias laterais diferentes, D) entre uma molécula de água e a proteína e E) entre duas moléculas de água.

16) Qual é a reação catalizada pela lisozima de ovo de galinha? Quais são os aminoácidos diretamente envolvidas no mecanismo de catálise? Onde se encontra na estrutura? Produzir outras figuras que apresentem esta e outras características interessantes sobre a estrutura de lisozima associada com sua atividade enzimática.