

Analysing Diffraction Patterns - Subtraction of background noise due to Lipidic Cubic Phase

Markus Klinker, University of Edinburgh, United Kingdom

Supervised by:

Kenneth Beyerlein, Thomas White and Henry Chapman

September 6, 2012

Abstract

This project is concerned with the treatment of background noise present in diffraction patterns. The diffraction patterns were recorded using femtosecond X-ray pulses from a free electron laser to study nanoscale protein crystals. The background in question is due to a lipidic cubic phase used as a carrier substance and strongly features ring-like patterns. A Matlab algorithm will be presented to subtract said ring patterns from diffraction data while leaving any other information intact.

Contents

1. Introduction	3
2. Background and Motivation	3
3. Description of Algorithm	5
3.1. Step 1: Locate Rings	6
3.2. Step 2: Subtract Rings	7
3.3. Step 3: Recover "true" intensity of pixels	7
3.4. Step 4: Further Refinement	8
4. Discussion and Possible Improvments	9
A. Example of how to use Algorithm	10
B. Description of used function	12
C. Hirachy of function	13

1. Introduction

Over the past one hundred years X-Ray crystallography has developed into an indispensable tool for the research of crystal structures. Since the invention of this technique the power of available X-Ray sources has experienced an exponential growth over the course of several decades, culminating in the construction of large scale free electron lasers, enabling us to study increasingly complex and small structures.

One particularly important field of research in X-Ray crystallography is the study of proteins. Proteins are a cornerstone of virtually all life on earth and improving our understanding of their structure may give us insights into a wide variety of biological processes. The first successfully solved protein structure was that of sperm whale myoglobin for which Max Perutz and Sir John Kendrew were awarded the Nobel Prize in 1962[1]. Since then the structures of almost 70000 proteins have been solved using X-Ray crystallography[2].

The object of interest to this project however are proteins which, as yet, have proven elusive to the methods of standard X-Ray crystallography. In particular these include membrane proteins, for which fewer than 300 structures have been solved [3]. These proteins are notoriously difficult to grow into large crystals making them unsuitable for X-Ray crystallography in most X-Ray sources. Over the last few years novel techniques have been developed that have the potential to overcome the limitations researchers have been subjected in the study of submicron protein crystals. These new techniques and the technical challenges they confront us with will be outlined in section 2.

Due to the technical challenges it is imperative to extract as much information as possible from the recorded data. This project will present an algorithm written in Matlab that can be used to reduce inhomogeneous background noise. In particular ring-like structures caused by lipidic cubic phase (LCP) which for some proteins is a suitable material to grow crystals in. Section 3 describes the algorithm developed for this task. Section 4 discussion of the results produced by said algorithm and discusses some ideas how the developed algorithm could be further tested and improved. The appendix contains a detailed description of the code and how to use it.

2. Background and Motivation

As hinted at in the introduction, the reason for the difficulties, we find ourselves presented with, is the submicron size of membrane protein crystals. Up until a few years ago these were near impossible to study, because the dose of radiation necessary to collect diffraction patterns from such small crystals inevitably leads to a Coulomb Explosion destroying the proteins [5]. The development of powerful free electron lasers (such as FLASH at DESY, LCLS at SLAC and in the future XFEL at DESY) has however overcome this problem in a somewhat spectacular way. The capability of these light sources to produce extremely brilliant femtosecond pulses allows us to deliver enough photons to record a diffraction pattern in a time so short as to outrun the processes driving the atoms apart.

It should come as no surprise, that when working with femtosecond pulses to study sub micrometer objects one is presented with tremendous experimental challenges. The main difficulty is to deliver the nanoscale protein crystals to the FEL beam in controlled fashion. Considerable effort has been put into this and has produced one especially promising technique. The protein crystals are carried by a liquid jet (with a diameter of a few micrometers) that continually flows through the X-Ray beam focus (producing these ultra fine jets is a challenge of its own) [7]. Although this methods succeeds in delivering the protein crystals to the beam it does also present us with some problems:

- Most Laser pulses will hit the beam without any protein crstally present (efforts are being made to synchronize the beam pulses with the arrival of protein crystals, these however do not work as of yet)
- The crystals are randomly orientated (gathering information about protein structures from randomly orientated crystals is a not at all trivial problem, the discussion of which would be tangential the objective of this project, suffice it to say the more data gathered the better)
- In the examples considered in this report the protein crystals were grown in an LCP. LCPs have proven a suitable host material to grow membrane proteins in, as it mirrors the structures they naturally occur in and can "store" large numbers of proteins [4]. The drawback of this is that during exposure to the X-Ray pulse the LCP produces a diffraction pattern of it's own, thereby adulterating the information about diffraction on the protein crystals. This LCP background, due to the high symmetry of LCPs, has some very distinct features in the shape of rings and ring segments (LCP-rings)

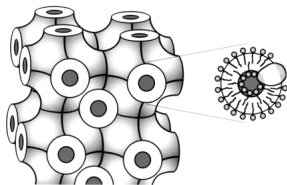


Figure 1: Left: Schematic model of cubic phase. Right: enlarged region with inserted membrane protein [6]

Points one and two illustrate that we should aim to gather as much information as possible from any one diffraction pattern with a protein crystal present. To do so an algorithm was developed in this project to subtract the background described in point three to increase the information we can extract from each image.

3. Description of Algorithm

Figure 2 displays a typical diffraction pattern recorded of a protein crystal in a LCP. Clearly visible are the diffuse LCP rings spanning an entire circle and a few more localized rings resulting from lamellar crystalline phase (these are a result of fast dehydration of the LCP). The aim should be: to recover as many peaks (hits) "hidden" behind the LCP-rings and find their intensities as they would be without a LCP present.

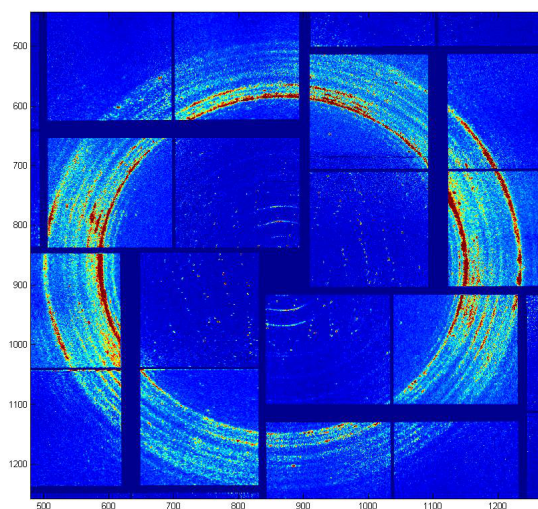


Figure 2: Diffraction recorded exposing nano protein crystal in LCP to FEL pulse. One can clearly see: a) Diffraction Spots (Protein Crystal), b) Large rings (LCP) and c) Small rings (Lamellar crystalline phase)

To achieve this the algorithm processes a diffraction pattern in four steps.

1. Locate the rings
2. Subtract the rings
3. Find Pixels and treat them as to recover their "true" intensities
4. further refinement

3.1. Step 1: Locate Rings

To locate the rings, the diffraction pattern is split up into a number of wedges. In each wedge the radial average intensity is computed. Next, peaks in intensity are identified in each wedge by searching for values over a specified threshold intensity, to learn about where rings are potentially located. Figure 3 illustrates this process schematically.

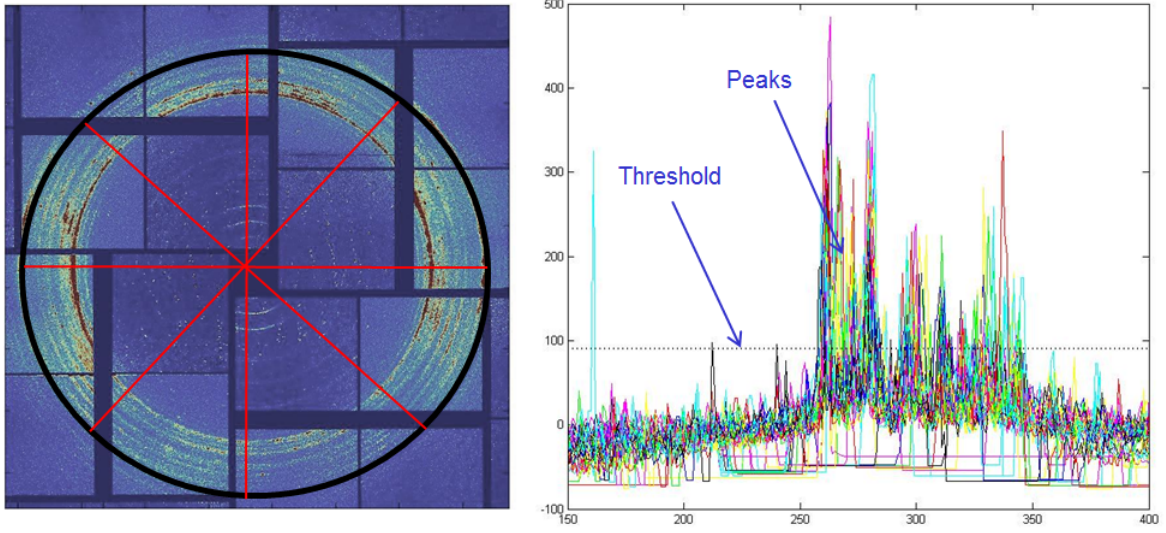


Figure 3: Left: Schematic representation of diffraction pattern split up into wedges. Right: Plot showing intensity versus radius of about 20 wedges. The dotted line represents the threshold used to identify peaks. Between radii 250 and 350 we expect to find some ring segments.

At this stage we have not learned anything about possible ring structures present. To do so, the algorithm looks for peaks in intensity occurring at similar radii in two adjacent wedges. Applying this to all wedges returns all ring segments spanning two wedges. To find out if the previously found ring segments are part of longer rings, this process is repeated finding adjacent sets of ring segments at similar radii. This process is repeated until potentially even complete rings are found. From this information about rings present in the diffraction pattern, a pixel mask is computed identifying all pixels lying within a ring structure. Figure 4 shows a schematic representation of this process, a section of the diffraction pattern with the rings superimposed on it and the pixel mask generated from this information.

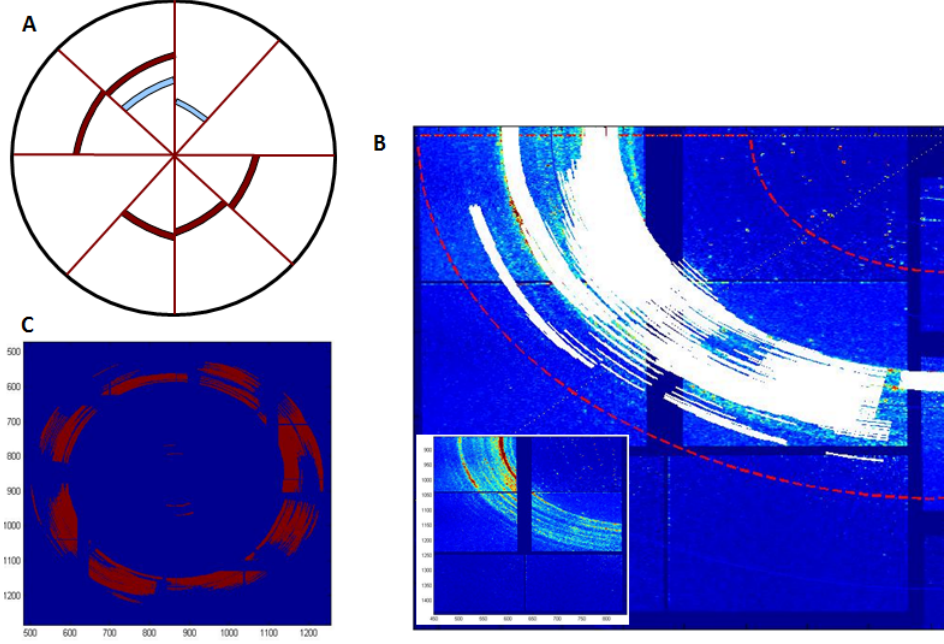


Figure 4: (A) Red and blue arches represent peaks found in the respective wedges. Red: part of a ring segment. Blue: Not part of a ring segment. (B) Section of diffraction pattern on which all rings found in that section are superimposed. (C) Pixel mask found by identifying pixel that are being intersected by ring segments.

3.2. Step 2: Subtract Rings

Having identified the rings, the appropriate values associated with the ring structure need to be found and subtracted. As a first approximation a rectangular box is created around each pixel, tangential to the ring the pixel lies on. The value to be subtracted from a particular pixel is then computed as the median of all pixels lying in the intersection of the rectangular box associated with the pixel in question and the ring mask. We obtain what will be referred to as the reduction mask (see Figure 5).

3.3. Step 3: Recover "true" intensity of pixels

After subtracting the LCP rings as described before any peak we may find in the background reduced regions are likely to underestimate the intensity of that peak as we would find it without the LCP. This is a result of also considering the peaks themselves when computing the median values, thereby shifting the median up. Thus one would like to locate the peaks and compute the reduction mask without taking the peaks into account. For the purpose of this project a peak is considered a strong and local fluctuation in intensity. Hence, a peak was identified as a region of high intensity around which the intensity sharply drops off in every direction (see Figure 6).

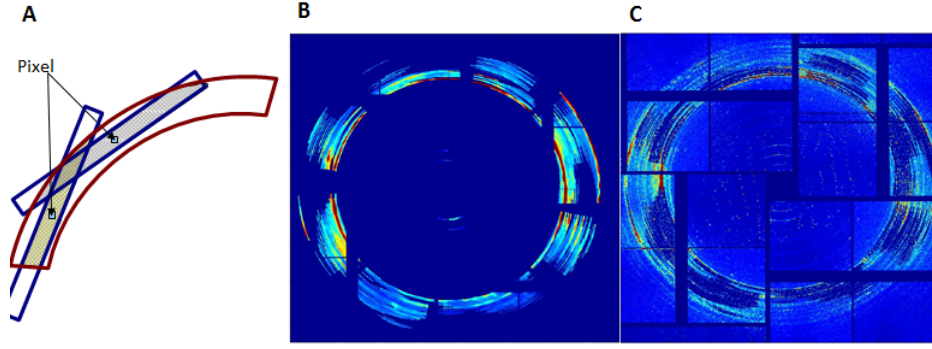


Figure 5: (A) Schematic representation of how intensity value to be subtracted from a pixel is found. The shaded area are the ones considered to find the medians. (B) Mask of values to be subtracted from original diffraction pattern. (C) Resulting pattern after subtracting aforementioned mask.

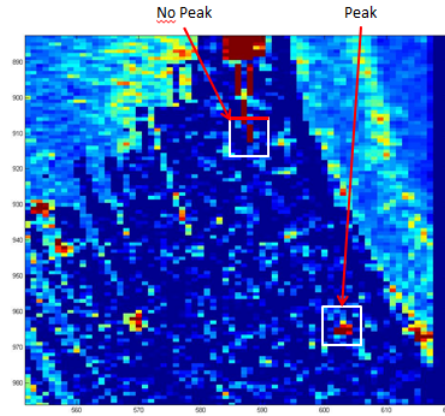


Figure 6: Section of diffraction pattern after subtracting rings, illustrating how peaks are identified as such. The intensity around the peak falls off in all directions. The intensity around the "non-peak" doesn't fall off sufficiently in the top direction.

As a final adjustment in treating peaks only pixels are considered peaks if they have at least one neighbor that was also identified as a peak; this was implemented to avoid individual hot pixels being identified as peaks.

3.4. Step 4: Further Refinement

As a final step two methods are in place further refine the processed data.

- Running the algorithm iteratively: Repeating steps one to three on a data set which has already been processed may improve the results, especially when the

parameters are altered as a function of the number of iterations passed. In particular edges of ring structure not found during the first iteration are likely to be picked up by further iterations (Figure 7).

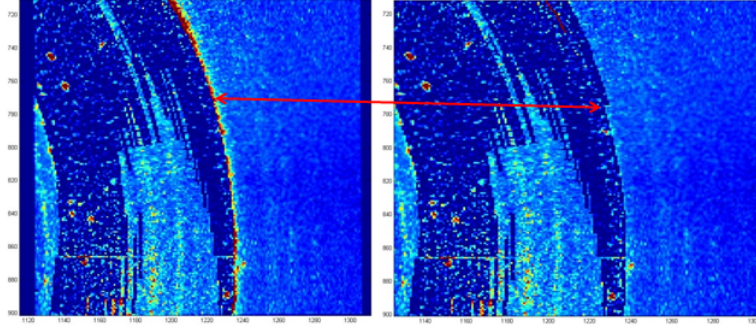


Figure 7: Section of diffraction pattern before and after a second iteration. We can see the leftover ring edge present after one iteration has disappeared after the second.

- Specifying different set of parameters for different regions can be a good idea in particular when one considers data, like in the project, where background structure with differing features are considered. Adjusting the parameters can greatly improve the results in those cases.

4. Discussion and Possible Improvements

In Figure 8 we see two before and after images of what the algorithm has achieved. It is apparent with the naked eye, that regions previously entirely obstructed by ring structures are relatively free of background noise and also exhibit structures some of which are clearly diffraction spots (as they are lining up with other diffraction spots in otherwise "clean" regions) and others which are more difficult to classify (in particular in regions of relatively high residual background with several peak-like structures in close proximity of each other). From the presented images we can also identify regions near detector edges as a trouble spot. Here the ring finding procedure does not work as well regions located more centrally on a detector.

In the following I will present some suggestions as to what can be undertaken to further improve the results produced by this algorithm.

- To improve result near the edges of individual detectors two methods are conceivable. For one, one could implement a variable "wedge density". Drastically increasing the number of wedges has proven to give good results, however only at the expense of very long processing times. Thus decreasing the width of the wedges near trouble spots is likely to improve the results while not slowing down the algorithm drastically. This could be particularly useful if combined with an iterative approach. Another way to approach this issue, could be to compute a mask

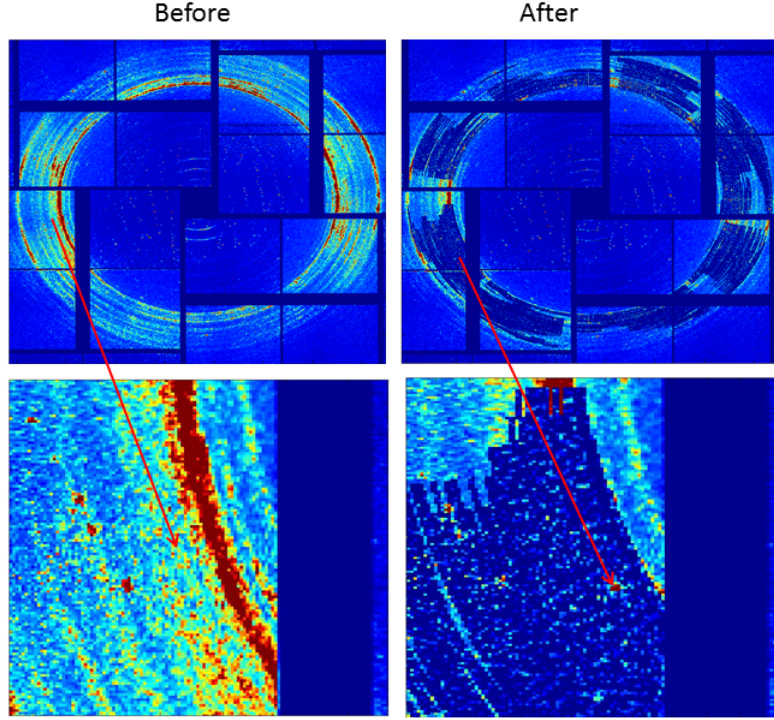


Figure 8: Top row: Entire diffraction pattern before and after being processed. Bottom row: zoomed in region displaying what quite probably is a peak, invisible before the rings were subtracted

identifying "empty" regions in between detectors and replace any values (obviously zeros) by the radially closest values outside this mask.

- To test parameters and get an idea of the success rate of this algorithm in finding peaks it one could plant "fake peaks" in the form of locally raised intensity patterns, to see if the algorithm identifies them correctly.
- In regions where peaks are found one could compute the how strongly the intensity fluctuates in the vicinity of the peak. Strong fluctuations in the immediate proximity could be an indication that what the algorithm found is a result of fluctuations in the intensity of the LCP rings rather than an actual peak.

A. Example of how to use Algorithm

In the following I will briefly outline how to use this algorithm to produce images like the ones presented in the report. The hdf5 files to be considered need to be in the same folder as the Matlab functions. First a set of parameters needs to be established. This is done by defining a $n \times 10$ array. The columns correspond to the individual parameters

described in Table 1; the rows to different sets of parameters should one desire to use different parameters in different parts of the diffraction patterns. The array has the form

```
segments=[intervals, theta_start, theta_fin, theta_res, r_start,
          r_fin, r_res, threshold, ring_guess,tolerance, min_length;...
          intervals, theta_start,..., min_length;...
          intervals,...,min_length];
```

The images above were computed using the following parameter matrix (combining the last four rows into one choosing the lowest threshold only marginally reduces the quality of the output):

```
segments=[150,0,2,300,30,150,800,40,80,4,5;...
          150,1.75,2.25,300,150,400,800,90,80,9,3;...
          150,0.25,0.75,300,150,400,800,145,80,8,5;...
          150,0.75,1.25,300,150,400,800,100,80,9,5;...
          150,1.25,1.75,300,150,400,800,100,80,8,4];
```

Having specified the parameters the function `ring_red_alg` is used to background subtract one set of data saved as two dimensional Matlab array (See Appendix B, `read_in_assem`):

```
[red_data,red_mask]=ring_red_alg(segments, iterations)
```

`red_data` and `red_mask` are two dimensional arrays containing the background subtracted data and the reduction mask respectively. The function `ring_red_alg` should be used to establish a useful set of parameters for a given set of files with similar features. Once this is done, the function `LCP_Reduce` should be used to process a larger set of files. To do so the names of the files have to be written into a text file called `hdf5_files.txt` located in the same folder as the Matlab functions (see Figure 9).

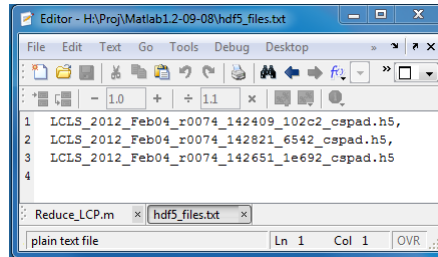


Figure 9: Text file specifying files to be processed by `LCP_reduce`

`LCP_reduce(segments,iterations)` creates two new data sets in the hdf5 files: `data/reduction_mask` and `data/ring_reduced_data`. `red_mask` and `red_data` are written into these respectively.

B. Description of used function

- **box_create**: Creates a rectangular box around a pixel tangential to a circle passing through that pixel. Outputs all values in the intersection of the box and the ring mask.
- **fuzzy_intersect**: Finds common elements in 1D-arrays within some tolerance. If two numbers from ordered arrays `arr_1` and `arr_2` are within a tolerance interval of each other they are considered to be in the intersection of `arr_1` and `arr_2`.
- **Peak_Matrix**: Produces a one dimensional cell. The `nth` entry corresponds to an array specifying the information about peaks found in radial average data in `nth` wedge.
- **Peak_Protect**: Looks for local intensity fluctuations in ring subtracted data and identifies these as peaks. The locations of the peaks are written into an array used to protect said peaks.
- **peakfind**: Finds peaks in intensity in radial average of a wedge.
- **peakwidth**: Determines the width of a peak found by `peakfind`.
- **Rad_Avg2**: Computes radial average of a wedge of data, allowing to specify angular and radial width of the wedge.
- **read_in_assem**: Reads assembled dataset from hdf5 file and outputs it as two dimensional Matlab array.
- **ring_mask3**: Finds coordinates of all pixels being intersected by any rings found by `Ringfind`.
- **Ringfind**: Finds rings an assembled data set by comparing previously computed radial averages of adjacent wedges
- **Ringreduce3**: Combines information found by `Peak_Protect` and `ring_mask3` to find reduction mask using `box_create`.
- **write_to_hdf5**: Creates dataset in specified locations writes desired data to it.

C. Hirachy of function

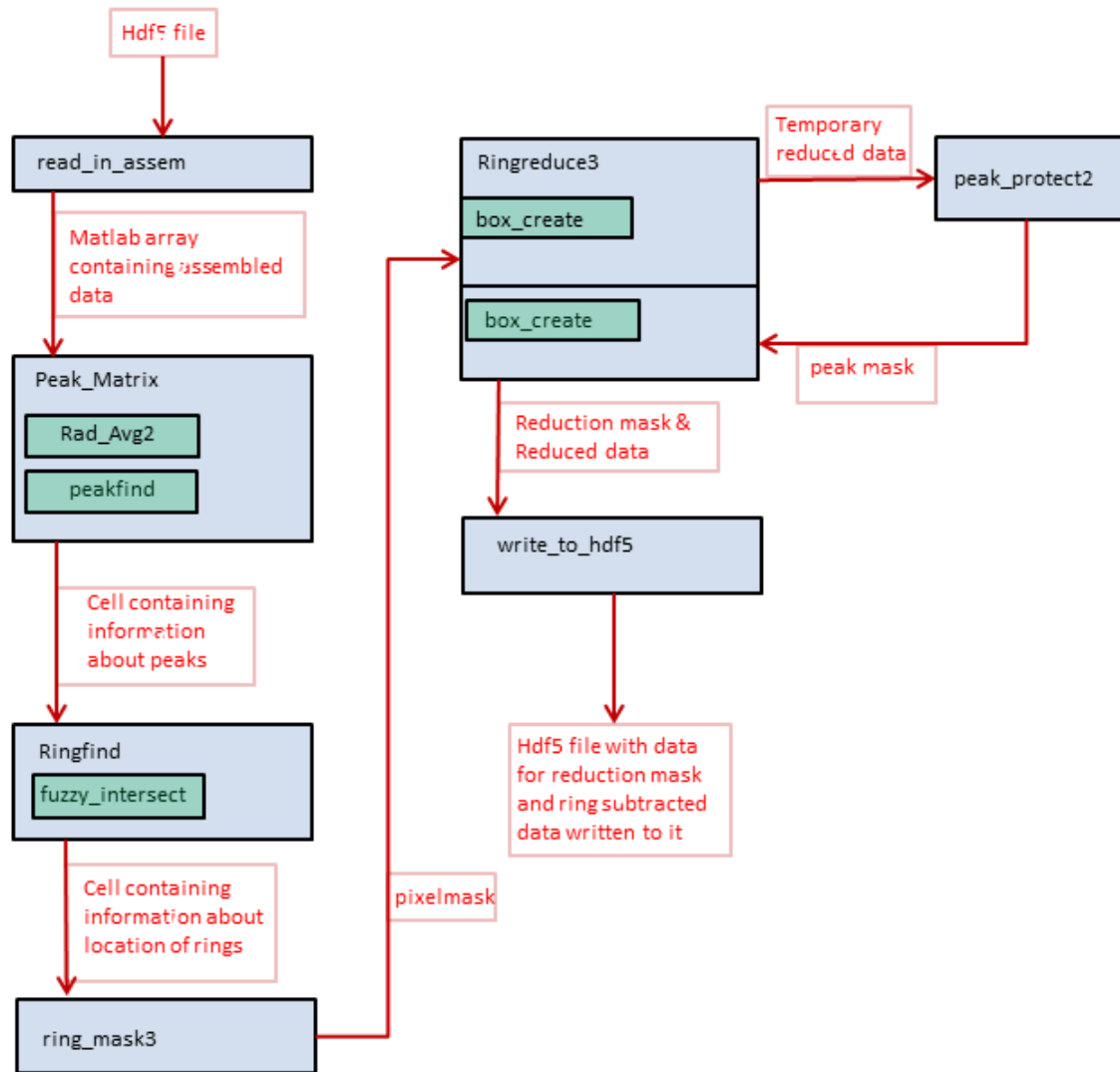


Figure 10: Schematic representation of the steps Reduce_LCP goes through processing an hdf5 file. Red boxes represent the input passed on to the next function, blue boxes major functions and green boxes minor functions utilised by the more complex functions

Table 1: Parameters

parameter	description	suggested values
intervals	number of wedges used in region specified by theta_start, theta_fin	50 to 200 as a first guess
theta_start	starting angle in multiples of π of region to be considered (0 degrees is vertically down and increases in clockwise direction)	
theta_fin	finishing angle in multiple of π of region to be considered	
theta_res	angular resolution; number of radial lines used to calculate radial averages	> 2 times intervals
r_start	starting radius with respect to r_res	choose so that origin is clearly excluded to avoid needless calculations
r_fin	starting radius with respect to r_res	
r_res	number of concentric circles used to compute radial averages	500 to 1500 as first guess
threshold	intensity threshold to identify peaks within wedges	depends on diffraction pattern
ring_guess	can be used to set a maximum number of rings expected in a region	> 50
tolerance	tolerance (with respect to r_res) within which peaks in two adjacent wedges are considered part of the same ring	typically 3 to 10
min_length	minimum length (in multiple of wedges) for a ring segment to be taken into account	> 2

References

- [1] Kendrew J. C. et al. (1958-03-08). "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". *Nature* 181 (4610): 6626.
- [2] PDB - Protein Data Bank. web: <http://www.rcsb.org/pdb/statistics/holdings.do>
- [3] Chapman H.N. et al. (2011-03-02). "Femtosecond X-ray protein nanocrystallography". *Nature* 470 (09750): 73-77.
- [4] Ehteshami M. L. and Rosenbusch J. P. (1996-12-10). "Lipidic cubic phases: A novel concept for the crystallisation of membrane proteins". *PNAS* 93: 14532-14535.
- [5] Neutze R. et al. (2004-04-12). "Potential impact of an X-ray free electron laser on structural biology". *Radiation Physics and Chemistry* 71 (2004): 905916.
- [6] Image taken from Ehteshami M. L. and Rosenbusch J. P. (1996-12-10). "Lipidic cubic phases: A novel concept for the crystallization of membrane proteins". *PNAS* 93: 14532-14535.
- [7] Krian A. et al. (2010-15-03). "Femtosecond protein nanocrystallography data". *Optics Express* Vol. 18, No. 6 (5713). analysis methods