Simultaneous Structure Factor and Contrast Transfer Function Parameter Determination in Transmission Electron Microscopy

Olena Sinkevich   Richard A. Tapia   Yin Zhang
Steven J. Ludtke

August 4, 2000

Abstract. We present a new method that allows a fully automated simultaneous determination of the structure factor and the parameters of the Contrast Transfer Function (CTF) and noise function. No previous knowledge of the structure factor or of the CTF parameters is assumed. Our approach is based on the new precise mathematical formulations of the problem as constrained nonlinear least squares that treats the structure factor as a set of undetermined variables, as well as the CTF parameters, and on the interior-point algorithm that satisfies the inequality constraints on the bounds.

Key Words. Electron Microscopy, Structure factor, Contrast Transfer Function, CTF, nonlinear least-squares problem.

1 Introduction

Transmission Electron Microscopy is one of the primary techniques used to determine macromolecular structure (such as for proteins and viruses). Since the microscope is not a perfect optical system, the projected images are not exact representations of the sample particle. This effect is quantified as the contrast transfer function (CTF) of the microscope. In order to allow optimal image correction it is necessary to approximate the parameters of both the CTF and a noise function, as well as the structure of the particle. The density information recovered from this corrected image is thus improved.

Currently, the parameter identification process performed in the laboratories is tedious, time consuming, requires a lot of user interaction and uses ad-hoc procedures.

We present a method for determining the CTF and noise function by estimating their parameters while simultaneously determining the structure factor. We propose two new mathematically sound formulations based on two key ideas: (i) the use of multiple data sets and (ii) the explicit imposition of bounds on the parameters. We treat the problem as a constrained nonlinear least-squares minimization problem where we demand that the corrected structure factor be the same in different micrographs of the same molecule. Combined with a functional form for the CTF and background noise, this provides sufficient information to simultaneously determine the CTF parameters and the corrected structure factor of the molecule.

We construct a numerical algorithm for solving this problem using a new interior-point mathematical methodology, which deals efficiently with the bounds on the variables.

The paper is organized as follows. In section 2, we present background material. We describe the model for the CTF, noise and the total intensity of the signal observed on the micrograph, describe the current "manual" approach to the structure factor and parameter determination problem. In section 3, we present the new mathematical formulations of the problem and the algorithm that allows efficient solution of the problem. In Section 4, we present numerical results that demonstrate the effectiveness of our approach. Finally, we present our conclusions in Section 5.

2 Background

Transmission electron microscopy, along with X-ray crystallography, is one of the primary techniques used to study molecular structure. There are several advantages in using electron cryo-microscopy. In X-ray crystallography phase of the X-ray wave cannot be determined directly, therefore the image in real space is impossible to regenerate because of the absence of phase information necessary for the inverse Fourier transform. It is difficult to crystallize large molecules. Determining the correct crystallization process for a new protein may take months or even years. The biggest advantage of electron microscopy is the collection of the image in real space. In X-ray crystallography the challenge is to recover the lost information, which is not an issue in electron microscopy.

Image formation in electron microscopy is a complex process. The raw images are not faithful representations of the particle [4]. The microscope induces artifacts that need to be corrected.

The basis of image formation is the interaction of the electrons with the object. The theory of image formation states that the wave function at the diffraction plane of an electron lens is the Fourier transform of the object's projected potential function (electron density function); and that the image intensity at the image plane of an electron lens is linearly related to the inverse Fourier transform of the wave function at the diffraction plane convoluted with a Contrast Transfer Function (CTF) of the microscope.

Erickson and Klug (1971) and Hawkes [5] present how instrumental modifications of the signal induced by the electron microscope can be modeled, within certain limitations, in terms of the Contrast Transfer Function (CTF).

CTF correction is the first part of image processing. Each particle image is two-dimensionally Fourier transformed, then the set of images is incoherently averaged (that is, the Fourier intensities are averaged) to reduce the noise. The final average is then rotationally averaged to produce a function of radius only. Each micrograph is characterized by a signal intensity curve which varies with spatial frequency.

2.1 Contrast Transfer Function, Noise, And Total Intensity Model

The Contrast Transfer Function for microscope is traditionally modeled by

$$\mathrm{CTF}(s) = Amp \times e^{-Bs^2}(\sqrt{1 - C_A^2}\ \sin\gamma(s) + C_A \cos\gamma(s)),$$

where s is the spatial frequency, Amp is the overall amplitude, and C_A is the amplitude contrast factor that describes the relative amount of elastic and inelastic scattering. It depends on the specimen and the microscope in use. The term $\sqrt{1 - C_A^2}\ \sin\gamma(s)$ describes the phase contrast and the $C_A \cos\gamma(s)$ term describes the absorption (i.e. amplitude contrast). The term e^{-Bs^2} is an envelope function.

Theoretically, information is transferred in a wide range of spatial frequencies. The increasingly rapid oscillations of CTF make it difficult to exploit all of the high frequency information. In practice, as we go toward higher spatial frequencies, the CTF is damped, limiting the resolution. This effect can be described by the envelope function e^{-Bs^2}. Various sources contributing to the envelope function fall-off include electron microscope factors, specimen movement, photographic emulsion, digital scanner, the specimen itself. Even the skills of the microscope operator (e.g. waiting for too long to take a picture or not getting the defocus right) affect the envelope function fall-off. The known theoretical form for the envelope function is fairly complex. All of the parameters for the full envelope function cannot be determined unambiguously. Empirically, e^{-Bs^2} functional form fits the data within the bound of error in many cases. A simple envelope function is not sufficient for very clear data because it is difficult to fit both the second peak and right-hand part of the signal intensity curve.

Parameter B is a microscope dependent factor. It is weakly dependent on the defocus but in the current model it is considered to be independent. The factor $\gamma(s)$ is the phase shift in reciprocal (Fourier) space for phase contrast microscopes:

$$\gamma(s) = -2\pi\left(\frac{C_s 10^7 \lambda^3 s^4}{4} + 5000 \times \Delta z \lambda s^2\right), \tag{1}$$

where C_s is the spherical abberation of the microscope, $\lambda = \frac{hc}{\sqrt{V^2+2E_0V}}$ is the wavelength of the electrons in the beam, V is the microscope voltage (in kilovolts), E_0V is the rest mass of the electron, h is Plank's constant, and c is the speed of light; Δz is the defocus value.

This dependence of the phase shift on spatial frequency (1) has long been known to give a good agreement with data from the microscope in the absence of astigmatism. In practice, if astigmatism is present, or if the micrograph is corrupted otherwise (such as by specimen drift), the micrograph is undesirable for the purpose of the CTF parameter determination. Therefore, we can assume there is no astigmatism present and we can use the above expression for the phase shift (1).

One can rewrite the CTF as:

$$\begin{aligned}\mathrm{CTF}(s) &= Amp\, e^{-Bs^2}\left(\sqrt{1-C_A^2}\,\sin(\gamma(s)) + C_A\cos(\gamma(s))\right)\\ &= Amp\, e^{-Bs^2}(\sin(\gamma(s)+\theta)) \end{aligned} \tag{2}$$

where $\sin(\theta) = C_A$. The parameter used in our computations is $\theta = \arcsin(C_A)$.

The functional form of the noise intensity is empirical. It includes effects of many different sources and it fits data obtained from several different microscopes well.

$$\text{Noise Intensity}(s) = n_3\, e^{-((\pi/2)n_4 s)^2 - n_1\sqrt{s} - n_2 s}, \tag{3}$$

where n_3 is the squared amplitude of the noise. Since the noise is additive and (empirically) incoherent with regards to the true signal, the total intensity of the signal observed on the micrograph may be expressed by:

$$\text{TI(s)} = \text{CTF}(s)^2\text{F}(s) + \text{Noise Intensity}(s), \tag{4}$$

where Noise Intensity is the squared average of white random noise, and F(s) is the unknown structure factor of the particles (squared Fourier transform of the projected density).

There are four unknown parameters which characterize the CTF: the microscope dependent parameter B, the amplitude contrast C_A, the defocus Δz, and the overall amplitude of the signal Amp. Also there are four unknown parameters describing the noise $n_i, i = 1 \ldots 4$. These model parameters need to be determined, along with the vector of the structure factor values $\text{F}(s)$.

The following requirements bound the parameters. All the parameters are nonnegative, with the exception of the defocus Δz. Parameter B is practically strictly positive; otherwise, the envelope function is equal to one and there is no damping effect. Parameter C_A varies between zero and one, typically $0 \leq C_A \leq 0.2$. There are soft bounds on Δz: -5 $\leq \Delta z \leq$ 5.

The determination of F(s) as a function of spatial frequency is a difficult problem. In case F(s) in unknown, it must be determined by demanding that it is consistent between different defocus settings.

2.2 "Manual" Curve-Fitting Procedure for the CTF Parameter Estimation Problem

Currently, the parameter identification process is done by "manually" fitting the model to the measured data through trial and error (see, for example, [1]). First, one assumes that the structure factor values are all ones. Then, one guesses the parameter values based on the physical characteristics of the CTF and noise functions and applies an unconstrained local optimization method to obtain a fit. The current CTF and two noise parameters fitting procedure goes as follows. For large wave-numbers the intensity is entirely comprised of noise. By plotting the intensity, the coefficients n_3 and n_2 of Noise Intensity $= n_3\, e^{-n_2 s}$ are fitted. To manually fit the data, noise is fitted first from the tail of the curve and the zero points. When one determines the CTF "by hand", the positions of the zeroes of the CTF are measured and fitted to a chart of the CTF characteristics. The four noise parameters can easily be used up to fit the peak but one should avoid using the degrees of freedom for this purpose. If the data is not very clear, the noise parameters can be defined from the tail of the curve. Otherwise, "manual" fitting becomes almost impossible. Currently, the boundaries are enforced by manually restricting the parameter choice to sensible values.

If the fit does not look right or the parameters are out of range, then the guessing and fitting procedures are repeated again until a fit is acceptable. From the fitted parameter values, one computes a set of corresponding structure factor values. Since the structure factor values should be the same for different data sets for the same particle, the fitting procedure requires going back and forth many times between different data sets and parameter values in order to obtain consistent structure factor values.

Other disadvantages of the "manual fitting" approach include the lack of a sound and precise measure for the quality of obtained parameter values and the quality of fit, the excessive sensitivity to the noise intensity parameters, no effective enforcement of parameter bounds, and the inflexibility of accommodating new parameters in the model. Previous attempts (Zhu *et al.* [12]) at automating the parameter estimation process mimic the "manual fitting" approach, hence sharing similar flaws.

The fitting routine does not work well if the envelope function is narrow, or if the defocus and/or other parameters fall out of a smaller subset of the parameter space. It also fails in the presence of distinguished peaks in the high frequency area.

If the structure factor is available from X-ray scattering, it is used in the fitting. Even then, the fit might not be very accurate. Figure 1 illustrates the best "manual" data fit for the Human Fatty Acid Synthase when the structure factor known from X-ray scattering is used and a fit used our automated method that will be described in the next section. In the plot for the automated fit the data is represented by asterisks (*) and the fit is represented by a solid line. We can see that in most cases the fit is accurate. When the X-ray data is absent, the fitting becomes even harder. The approach that we propose in the next chapter will not only allow us to perform the data fit with the structure factor being unknown, but often allows to have a better fit. One of the reasons for a more accurate fit is that there are more degrees of freedom when the structure factor is not fixed (even to a 'good' value).

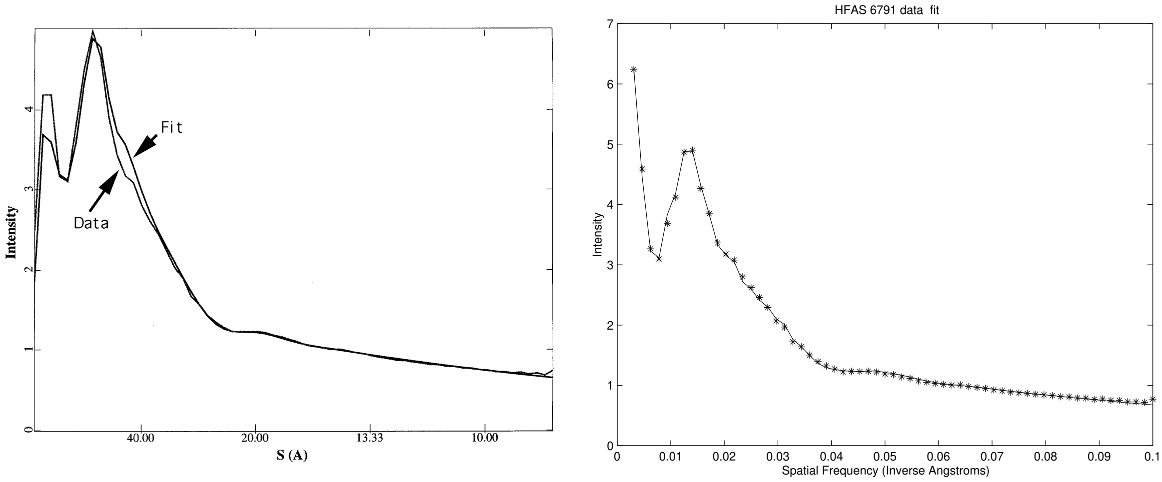

Figure 1: Comparison of the "manual" data fit with the known structure factor (left) and of the fully automatic data fit with the unknown structure factor (right) for human fatty acid synthase.

The attempts of fitting include the "intelligent" (similar to manual) fitting algorithm and a gradient-descent like method.

Zhu, Penczek, Scroder, and Frank [12] proposed an "intelligent" fitting method, which uses zeros of the CTF to determine the defocus and amplitude contrast parameters. The method uses the information from the peaks of the signal intensity curve to define amplitude. The authors do not specify the details of determining the envelope function and noise parameters.

3 Mathematical Formulation of the Problem

We propose two new formulations of the problem. Both are constrained nonlinear least squares that simultaneously determine the structure factor and the Contrast Transfer Function parameters. This new approach requires multiple data sets and finds all the unknown parameters corresponding to all the involved data sets at once.

We consider the functional form of the Contrast Transfer Function and Noise Intensity function as defined in (2) and (3), respectively. Corresponding to each data set there are eight unknown parameters in the CTF and noise function and an unknown vector of structure factor values whose size is that of the data set. If only one data set is used, then the system is underdetermined. This is why it is the current practice, where a single data set is used at a time, to fix the structure factor vector to some value while fitting the other model parameters.

A key observation is that if one uses multiple data sets, then it becomes possible to determine the unknown model parameters, along with the vector of structure factor values which is invariant among different data sets that are generated for the same particle. This approach is feasible because multiple date sets for the same particle are easy to generate in electron microscopy (and are already available for other purposes).

This approach offers considerable advantages and it will be discussed in detail in this chapter. However, we have discussed it in preliminary form at meetings and conferences, e.g. [9], (see also [7]).

We propose to minimize the sum of the squares of the residuals arising from fitting the model to multiple data sets concurrently. The data sets are of the same size, and the minimization is subject to the appropriate bounds on all variables. Thus we estimate the vector of the structure factor values for the sample particles, as well as the parameters of the Contrast Transfer Function and incoherent background (noise) to fit the experimental data.

The CTF and structure factor determination problem can be viewed as a constrained (because of the bounds on the variables) nonlinear least-squares problem. In this problem we fit the total intensity model which is nonlinear in the parameters, as defined in (4), to the total intensity of the signal observed on the micrograph. Usually the data set contains several hundred points.

We chose to formulate the problem as a nonlinear least-squares problem, i.e.

$$\min_{x \in R^n} f(x) = \frac{1}{2} R(x)^T R(x) = \frac{1}{2} \sum_{i=1}^{m} r_i(x)^2 \tag{5}$$

where $m > n$, the *residual function*

$$R : \Re^n \longrightarrow \Re^m$$

is nonlinear in x, and $r_i(x)$ denotes the ith component of $R(x)$. If $R(x)$ is linear, (5) is a linear least-squares problem. If one is attempting to fit the data $(s_i, y_i), i = 1, 2, \ldots, m$, with a model $M(x, s)$ that is nonlinear in x, then the nonlinear least-squares problem consists of choosing x so that the fit is as close as possible in the sense that the sum of squares of the residuals $r_i(x) = M(x, s_i) - y_i$ is minimized. Typically, m is much larger than n.

It is also possible to use the l_1 norm

$$f_1(x) = \sum_{i=1}^{m} |r_i|$$

or the l_∞ norm

$$f_\infty(x) = \max_{1 \le i \le m} |r_i|$$

to evaluate the quality of the fit. The choice of the sum-of-squares measure for data fitting is justified by statistical considerations (see, for example, [2]). The l_∞ norm is not as good in statistical sense as l_2 norm or l_1 norm. Of the three norms we prefer the least-squares formulation also because the objective function is differentiable only when using the l_2 norm and not in the the other two cases.

Note that the objective function is highly nonlinear due to the physics involved in creating the model.

It is possible that the constrained solution is not at the bounds of the feasible region, however, an optimum may be outside the constrained region and removing the constraints may lead to a solution outside of the range of meaningful values.

3.1 Formulations for the Simultaneous Structure Factor and Contrast Transfer Function Parameter Determination

It is convenient to use the following notation for the unknown parameters:

$$x_i = n_i, i = 1, 2, \dots, 4,$$

where n_i are the parameters of the noise intensity function (3), and the rest are the parameters of the CTF (2):

$$x_5 = Amp,\ x_6 = -\Delta z + 5,\ x_7 = \arcsin(C_A),\ x_8 = B.$$

In contrast with the "manual" approach to parameter estimation, parameters Δz, B, and Amp cannot be viewed as unconstrained and the values cannot be simply discarded if they are out of the range of meaningful values. Therefore we introduce a vector of upper bounds b, corresponding to the vector of unknown parameters. The upper bound values are soft. They are determined empirically. The defocus parameter Δz is the only parameter that can have negative values. For convenience, we can shift it to ensure nonnegativity without affecting the formulation. Therefore we consider all the parameters to be nonnegative.

Let us introduce the following notation.

- p is the number of different defocus data sets available for the same particle, usually 5 or 6;
- m is the number of points in each data set, usually several hundred;
- n is the number of unknown CTF and noise parameters for each data set; in the current model $n = 8$;
- $\boldsymbol{x_j}$ is the unknown vector of CTF and noise parameters for jth data set, $\boldsymbol{x_j} \in \Re^n$;
- $\boldsymbol{x}$ is a vector composed of vectors $\boldsymbol{x_j}$, $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_p \end{bmatrix}$, $\boldsymbol{x} \in \Re^{pn}$;
- $\boldsymbol{b_j}$ is a given vector of upper bounds for the CTF and noise parameters $\boldsymbol{x_j}$, $\boldsymbol{b_j} \in \Re^n$; upper bounds are the same for each data set, i.e., $\boldsymbol{b_j} = \boldsymbol{b_i}$ $\forall i, j \in 1, 2, \ldots, p$;
- $\boldsymbol{b}$ is a vector composed of vectors $\boldsymbol{b_j}$, $\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_p \end{bmatrix}$, $\boldsymbol{b} \in \Re^{pn}$;
- y is the unknown vector of structure factors, $y \in \Re^m$;
- $ctf(\boldsymbol{x}_j)$ is the unknown CTF vector for jth data set, $ctf(\boldsymbol{x}_j) \in \Re^m$;
- $noise(\boldsymbol{x}_j)$ is the unknown noise intensity vector for jth data set, $noise(\boldsymbol{x}_j) \in \Re^m$;
- t_j is the vector of total intensity of the signal observed on the micrograph; $t_j \in \Re^m$.

Based on the model of total intensity of the signal observed on the micrograph (4), we can now formulate our problem as

$$\begin{aligned} \min_{\boldsymbol{x},y} \frac{1}{2}\sum_{j=1}^{p} \quad & \|\mathrm{diag}(ctf(\boldsymbol{x}_j))^2 y + noise(\boldsymbol{x}_j) - t_j\|_2^2 \\ & 0 \le \boldsymbol{x} \le \boldsymbol{b},\ \boldsymbol{x} \in \Re^{pn} \\ & y \ge 0,\ y \in \Re^m, \end{aligned} \tag{6}$$

where $noise(\cdot) : \Re^n \to \Re^m$, $ctf(\cdot) : \Re^n \to \Re^m$, $t_j \in \Re^m, j = 1, \ldots, p$.

Let us denote

$$D(\boldsymbol{x}_j) = \mathrm{diag}(ctf(\boldsymbol{x}_j))^2, D(\boldsymbol{x}_j) \in \Re^{m\times m}, j = 1, \ldots, p$$

$$c(\boldsymbol{x}_j) = -noise(\boldsymbol{x}_j) + t_j, c(\boldsymbol{x}_j) \in \Re^m, j = 1, \ldots, p.$$

Omitting the multiplier constant $(\frac{1}{2})$ from the objective function, (6) can then be rewritten as

$$\begin{aligned} \min_{\boldsymbol{x},y} \quad & \textstyle\sum_{j=1}^{p} \|D(\boldsymbol{x}_j)y - c(\boldsymbol{x}_j)\|_2^2 \\ & 0 \le \boldsymbol{x} \le \boldsymbol{b},\ \boldsymbol{x} \in \Re^{pn} \\ & y \ge 0,\ y \in \Re^m \end{aligned} \tag{7}$$

or as

$$\begin{aligned} \min_{\boldsymbol{x},y} \quad & \left\| \begin{bmatrix} D(\boldsymbol{x}_1) \\ \vdots \\ D(\boldsymbol{x}_p) \end{bmatrix} y - \begin{bmatrix} c(\boldsymbol{x}_1) \\ \vdots \\ c(\boldsymbol{x}_p) \end{bmatrix} \right\|_2^2 \\ & 0 \le \boldsymbol{x} \le \boldsymbol{b},\ \boldsymbol{x} \in \Re^{pn} \\ & y \ge 0,\ y \in \Re^m. \end{aligned}$$

We have proven mathematically that the nonnegativity constraint on y can be removed if the upper bound on the noise amplitude parameter is less than the total intensity and if

$$(\sum_{j=1}^{p} D(\boldsymbol{x}_j)^2)_i \equiv (\sum_{j=1}^{p} \mathrm{diag}(ctf(\boldsymbol{x}_j))^4)_i \ne 0 \tag{8}$$

A zero entry on the diagonal of the above matrix can occur for some $i \in 1, 2, \ldots, m$ only if for some s, $ctf(\boldsymbol{x}_j) = 0$ in all p data sets. In practice, this situation is highly unlikely because we use different defocus micrographs and it is improbable that the zeros of the CTF functions will coincide. However, one has to ensure that (8) is not equal to zero. It is possible that the values of (8) are very small, causing numerical problems when dividing by (8). The solution y^* of (7) can be expressed in terms of $\boldsymbol{x}$:

$$y^*(\boldsymbol{x}) = ((\sum_{j=1}^{p} D(\boldsymbol{x}_j)^2))^{-1}(\sum_{j=1}^{p} D(\boldsymbol{x}_j)c(\boldsymbol{x}_j)) \tag{9}$$

or, component-wise,

$$y_i^*(\boldsymbol{x}) = \frac{\sum_{j=1}^{p} d_i(\boldsymbol{x}_j)c_i(\boldsymbol{x}_j)}{\sum_{j=1}^{p} d_i(\boldsymbol{x}_j)^2}, \tag{10}$$

where $d_i(\boldsymbol{x}_j)$ is the ith diagonal element of $D(\boldsymbol{x}_j)$ and $c_i(\boldsymbol{x}_j)$ is the ith element of $c(\boldsymbol{x}_j)$. The non-negativity condition on $y^*(\boldsymbol{x})$ no longer needs to be enforced since $d_i(\boldsymbol{x}_j) \geq 0, c(\boldsymbol{x}_j) \geq 0$.

Using the expression for y^* (10) we propose the second equivalent formulation of the problem. We have proven that if (7) has a local minimizer $(\boldsymbol{x}^*, y^*)$ and

$$(\sum_{j=1}^{p} D(\boldsymbol{x}_j)^2)_i \neq 0 \quad \forall\, i \in 1, 2, \ldots, m$$

then y^* satisfies (9), namely

$$y^* = ((\sum_{j=1}^{p} D(\boldsymbol{x}_j^*)^2))^{-1}(\sum_{j=1}^{p} D(\boldsymbol{x}_j^*)c(\boldsymbol{x}_j^*)) \tag{11}$$

We demonstrated that we can treat the structure factor vector y and the CTF and noise parameters $\boldsymbol{x}$ together, as independent variables, or, y can be treated as a variable dependent on $\boldsymbol{x}$, so that the explicit variable y can be eliminated, leaving fewer variables in the problem formulation. An additional advantage of the second formulation is that no initial guess for the of linear variables y is required. Another observation is that the Hessian approximation matrix is nearly singular when we use the first formulation and the Gauss-Newton method to solve it. A disadvantage of the second formulation is that the function, its Jacobian and Hessian calculations become very cumbersome.

3.2 The Algorithm

We have formulated the problem as nonlinear least squares with simple bound constraints. With these more sophisticated formulations, we need more sophisticated numerical algorithms. An important issue is how to handle the bound constraints. To address this, we will employ interior-point methods, a class of relatively new methods that have enjoyed great success as constrained optimization algorithms in the last decade, in terms of both their theoretical properties and practical performance. The strength of the interior-point methods lies in their ability to treat inequality constraints effectively (see, for example, [3, 10, 11]).

The interior-point methods can be described as follows. Consider a nonlinear programming problem (NLP), which in its most general form is the optimization of an objective function $f(x)$ over some space $x \in X$, subject to constraints on the variable x. The system of optimality conditions for the problem is formulated. The system is then perturbed to avoid certain numerical problems. Then damped Newton's method is applied until optimality conditions are satisfied. It is necessary to damp the Newton's step in order to satisfy the inequality constraints on the variables, which in our case are the upper and lower bound constraints on the CTF and noise parameters and on the structure factor.

The Newton Interior-Point Algorithm

Choose a feasible initial guess.

For $k = 0, 1, 2, \dots$ **do**

STEP 1. Test for convergence: If a convergence criterion is satisfied then exit.

STEP 2. Update the perturbation parameter.

STEP 3. Solve for perturbed Newton step.

STEP 4. Adjust step-length to ensure feasibility with respect to inequality constraints.

STEP 5. Adjust step-length for globalization.

STEP 6. Update unknowns.

EndFor.

The interior-point methods are known for their good local convergence properties and global convergence properties for convex programming problems. We investigate specific primal-dual interior-point methods that best exploit the special structure of the problem at hand. We conduct extensive numerical experiments to find the best combination of the many factors that leads to a reliable and efficient algorithm. The results appear in Section 4.

We proposed a Newton interior-point algorithm that can be easily modified to Gauss-Newton or Levenberg-Marquardt interior-point algorithm. We have also developed mathematical theory (see [8]) that proves the convergence of the algorithm, confirming that it is a good choice in combination with the proposed formulations of the problem.

4 Numerical Experimentation

We have developed an efficient algorithm for the numerical solution of the problem, which can be used as a base for an automated routine for the determination of the structure factor and the CTF and noise function parameters. We construct an algorithm based on interior-point methodology in order to treat bound constraints efficiently. In this section we report our numerical experience. The computational work was done on a SUN Ultrasparc 2 Workstation running SunOS 64 megabytes of memory and with a 200 MHz processor. The programs were written in Matlab and run under Version 5.1.

Recall that the first formulation deals with the CTF and the noise parameters and the structure factor values as independent variables. In the second formulation the structure factor is implicit. Our implementation can be easily modified to use Newton's or Levenberg-Marquardt method instead of the Gauss-Newton method. If Newton's method is desired, the user needs to supply second derivative information, which is avoided if either the Levenberg-Marquardt method or the Gauss-Newton method is used.

The problem is badly scaled due to the difference in magnitude of the variables. We have scaled the variables, however the numerical experimentation showed that scaling B did not effect the computational results significantly.

The code was tested on the standard test set ([6]) as well as test problems created by us.

4.1 User-Supplied Information

The user of the code needs to provide files with the input data: spatial frequency s and corresponding measured total intensity of the signal $TI(s)$ for 5 data sets for each particle. Spatial frequencies s (data points) are the same for a given particle from the same microscope for different defocus data sets. The sharp peak of the signal intensity curve near s=0 represents an artifact of the microscope, which must be ignored in the fitting procedure. Therefore the data corresponding to (0,0) and points before the first peak are excluded from the data fitting process. The range to be ignored must be determined. If the peak is wide, the choice has to be made as to which part of the peak to disregard. For certain data, peaks do represent valuable information. Generally, the fitting starts at the spatial frequency roughly equal to the inverse of the particle size, because for smaller spatial frequencies the useful information is lost.

Some of the noise parameters are very close in value, which contributes to singularity. To fix the linear dependence of the columns containing nearly same values, we can fix the noise parameters (reducing the number of variables), as well as some other parameters that are likely to be similar, such as amplitude contrast parameter C_A.

4.2 Initial Guess

The issue of finding a good initial guess for the variables is an important one. We are concerned with having good initial guess firstly, for the values of CTF and noise parameters, and secondly, for the vector of structure factor values.

It is a challenge to come up with a good initial guess for the vector of structure factor values. There is often no biological insight into such initial guess. The reason is that, if there is an information about structure factor from X-ray scattering, these values are just plugged in to find the values of CTF and noise parameters. If no such information is available then it is hard to suggest a reasonable value One option for a structure factor initial guess is to assume that it is a unit vector, solve the resulting (smaller) nonlinear least-squares problem and use $\frac{\sqrt{TI-NI}}{CTF^2}$ as an initial guess. Another possibility is to use $e^{-const\times s}$, since the structure factor vector mapped against spatial frequency roughly has an exponential shape. We have experimented with several heuristic initial guesses for the vector of structure factor values, such as exponential form and different constant values. It seems that independently of the starting estimate the final answers are consistent.

The use of the second formulation of the problem, where the linear variables, i.e. the structure factors, are implicit, allows one to avoid providing an initial guess for the vector of structure factor values.

4.3 Numerical Results

The results of our numerical experience are summarized in Table 1. The first

Table 1: Computational results for the sample particles.

Type of Particle	n	f	$\|F\|$	Rel	CPU
Human Fatty Acid Synthase	103	0.53338	0.056952	0.005	22.3
Herpes Virus Particles	279	3.7729	0.30941	0.013	1405.1
α-crystallin A	107	0.057627	0.0032962	0.0005	167.91
α-crystallin AB	104	3.5288	0.19409	0.0339	76.860
α-crystallin B	266	2.2884	1.1472	0.008	1261.1

column gives the particle type. The n column gives the dimension of the problem, which equals to the size of the vector of structure factor values (i.e. the size of one data set) plus 40 (i.e. the number of the unknown CTF and noise function parameters for 5 data sets). The third column gives the final objective function value. The forth column states the value of the residual of the optimality system for the problem, that was described in Section 3.2, at the termination of the algorithm. The value indicates how close the algorithm is to the optimal solution. Zero value would indicate an exact minimum was reached, however that is more often than not is precluded by the numerical artifact of the problem The fifth column provides a relative error, defined as the ratio of the objective function value to the size of the vector of unknowns. It gives a better feel for the quality of the fit that the absolute value of the objective function reported in the third column. Last column gives the total computing time in CPU seconds that algorithm took.

The Figures **??**, 2, and 3 illustrate the fit obtained with the use of our formulation and algorithm for human fatty acid synthase, herpes simplex virus (HSV-1), and α-crystallin AB. For α-crystallin AB the plot in the top left corner of each figure shows 5 distinct data sets that need to be fitted for each particle. The other plots show the data, represented by asterisks (*), and the fit, represented by a solid line. We can see that in most cases the fit is accurate. The quality of the fit increases as the points, corresponding to the low frequency part of the data curve are excluded. The current model does not always capture the phenomena responsible for the peaks if the data curve. Therefore, the user can decide how may points to exclude from the fitting process to get the optimal fit.

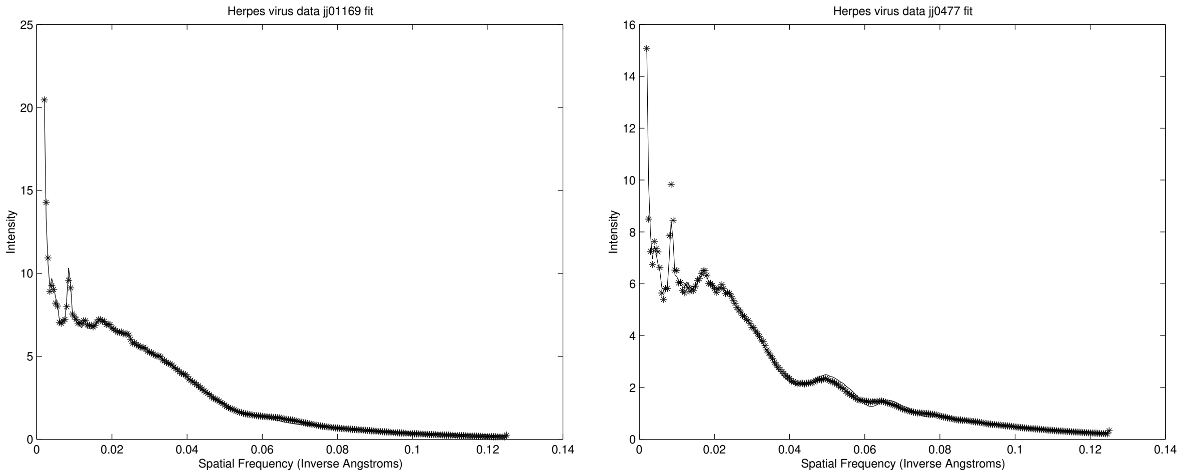

Figure 2: Fully automatic data fit with the unknown structure factor for the herpes simplex virus (HSV-1).

The results in Table 1 were produced using the first formulation of the problem. The second formulation, while a worthwhile idea, is more difficult to implement. Levenberg-Marquardt approach was used in the final calculations since it allows to reduce the numerical errors caused by ill-conditioning.

5 Conclusions

We consider an important problem in electron microscopy. The previous attempts to solve this problem were unsatisfactory. We mathematically formulate the problem in two related but different ways as a nonlinear least-squares problem with simple bounds. Our first formulation is a mixed linear-nonlinear least-squares problem, where the residual function is linear in the variables representing structure factors and nonlinear in the others. Taking advantage of the linearity, we propose the second formulation, that reduces the number of variables, perhaps, at the cost of an increase in the nonlinearity of the problem. The second formulation does avoid the need for an initial guess of the structure factor values.

Our code features a number of advantages over the currently practiced "manual" approach.

It simultaneously fits the CTF and noise parameters and the estimates the structure factor values.

The quality of fit can be precisely judged from the objective function value.

Our fitting procedure can successfully fit the model to the data over a greater range of spatial frequencies than the “manual” procedure.

It is easy to incorporate changes in the model. Currently there are 8 unknown parameters for each data set, as well as a vector of structure factor values, whose size equals that of the data set. As was previously discussed (see Section 2.1), there is interest in making the model more complicated to account for the currently underrepresented phenomena. One of the first such changes, would, perhaps, be the more complicated model of the envelope function, currently modeled as a simple Gaussian. There is no conceptual difficulty in adding as many parameters as desired to the model of the envelope function or to any other part of the model.

The user can take advantage of Gauss-Newton, Levenberg-Marquardt, or Newton's method, when using the algorithm.

The objective value obtained by algorithm is very low in most cases, indicating a close fit of model to the data. In cases, when is was not reduced to a small value, we were attempting to fit the peaks of the intensity curve, that are not described by the current model. Removing points to the left of the peak led to an accurate fit with low objective function values.

We developed an algorithm for this class of problems. It is based on interior-point methodology which deals efficiently with the inequality constraints. We also take advantage of the special structure of our problem. Interior-point approach is combined with the Gauss-Newton method, which is especially successful on zero- and small-residual problems. We would like to point out that our implementation allows for the Levenberg-Marquardt method to be used. Newton's method also can be used, i.e. we can use the second derivative information, if desired.

Our numerical experimentation confirms that our approach is successful and demonstrates a good fit of the model to the data.

The proposed method of the simultaneous structure factor and CTF parameter determination is effective, minimizes user interaction and does not require any previous knowledge about the particle, making it an ideal choice for when neither X-ray crystallographic nor X-ray scattering data is available.

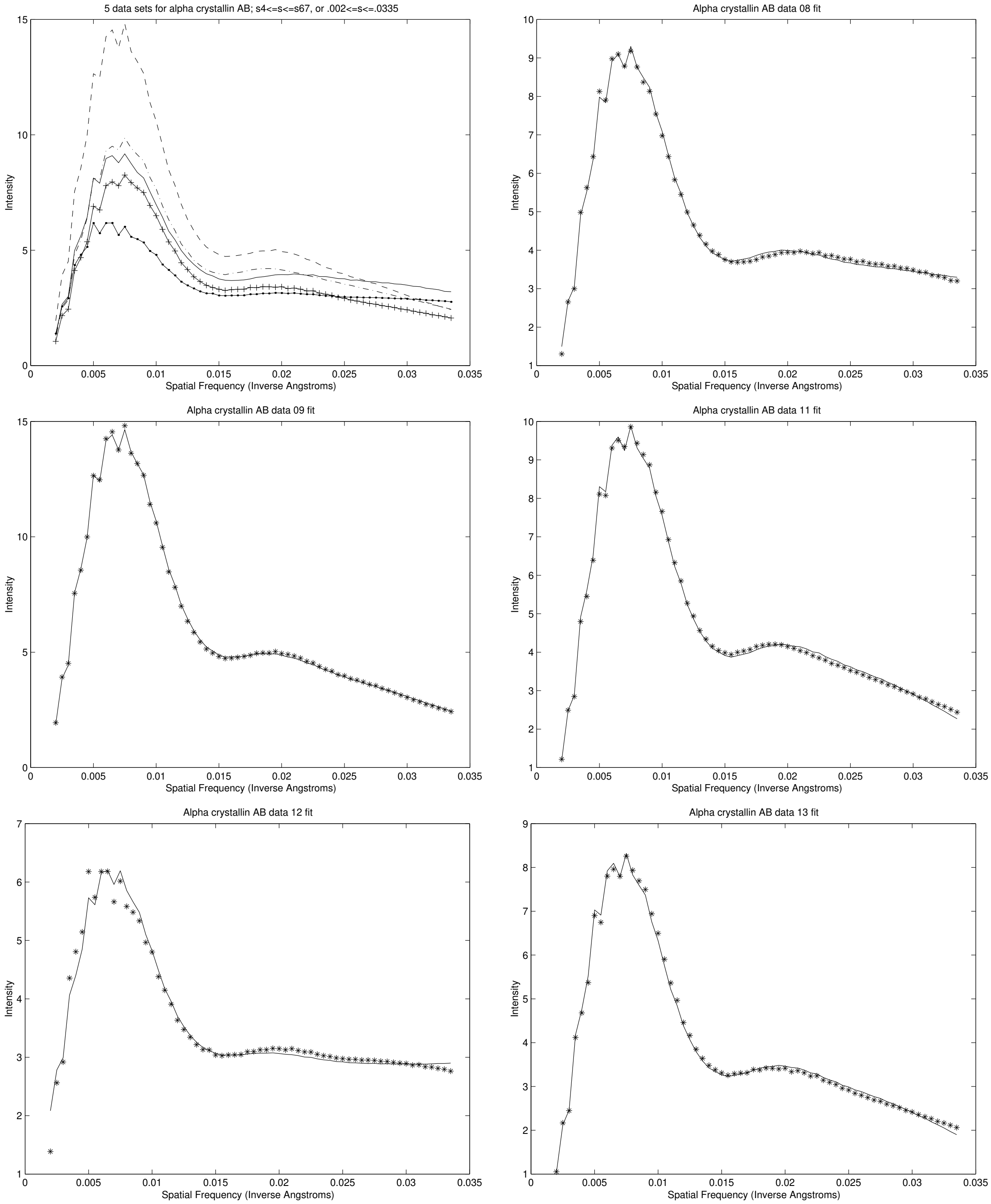

Figure 3: Fully automatic data fit for α-crystallin AB.

References

[1] P.R. Baldwin, S. Ludtke, F. Quiocho, M. Petrash, I. Serysheva, and Wah Chiu. Structure of an isoform of α-crystallin to 30Å resolution by electron cryomicroscopy. *In preparation, Baylor College of Medicine*, 1999.

[2] J.E. Dennis, Jr. and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[3] A. El-Bakry, R.A. Tapia, T. Tsuchiya, and Y. Zhang. On the formulation and theory of the Newton interior point method for nonlinear programming. *Journal of Optimization Theory and Applications*, 89(3):507–541, June 1996.

[4] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, 1996.

[5] P.W. Hawkes. The electron microscope as a structure projector. In J. Frank, editor, *Electron Tomography: Three-Dimensional Imaging with the Transmission Electron Microscope*, chapter 2, pages 17–38. Plenum Press, 1992.

[6] W. Hock and K. Schittkowski. *Test examples for nonlinear programming code*. Springer Verlag, 1981.

[7] S.J. Ludtke, P.R. Baldwin, and W. Chiu. EMAN: Semiautomated software for high resolution single-particle reconstructions. *Journal of Structural Biology*, 128:82–97, 1999.

[8] O. Sinkevich. *Optimixation for Parameter Estimation with Application to Transmission Electron Microscopy*. PhD thesis, Rice University, Houston, Texas, July 2000.

[9] O. Sinkevich, Y. Zhang, R. Tapia, and S. Ludtke. Parameter estimation with application to transmission electron microscopy. SIAM Optimization Conference and Annual Meeting, day honoring John Dennis and Richard Tapia, May 1999.

[10] S. Wright. *Primal-Dual Interior Point Methods.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[11] Y. Ye. *Interior Point Algorithms: Theory and Analysis.* John Wiley & Sons, New York, 1997.

[12] J. Zhu, P.A. Penczek, R. Schroder, and J. Frank. Three-dimensional reconstruction with contrast transfer function correction from energy-filtered cryoelectron micrographs: Procedure and application to the 70s *escherichia coli* ribosome. *Journal of Structural Biology*, 118:197–219, 1997.