Ultramicroscopy 46 (1992) 145–156
North-Holland

ultramicroscopy

A brief look at imaging and contrast transfer

R.H. Wade [1]

Laboratoire de Biologie Structurale, CEA et CNRS URA 1333, Département de Biologie Moléculaire et Structurale, CEN-G, 38041 Grenoble, France

Received at Editorial Office 9 April 1992

The development of the basic notions of electron microscope imaging and of contrast transfer theory are described. The effects of finite source size and chromatic instabilities (defocus fluctuations) can be represented by envelope functions which attenuate the oscillating contrast transfer function and which have significant effects on resolution. Beam-tilt effects are described. It is shown that there is a close relationship between holography and the contrast transfer representation of the imaging of weak-phase objects. This is a personal account which attempts to give an extremely condensed review of the development of the subject with particular emphasis on the last twenty-five years or so, and on matters of interest in biological macromolecular structure investigations.

1. Preamble

In a certain sense one could say that the electron microscope is a Franco-German invention since, when H. Busch showed in 1926 that axially symmetric electrostatic or magnetic fields could focus electron beams, L. de Broglie (1924) had already put forward the hypothesis that matter as well as light might exhibit both wave and corpuscular aspects. It followed that the wavelength λ should be related to the momentum p according to:

$$\lambda = h/p, \tag{1}$$

where h is Planck's constant. These speculations were confirmed by electron diffraction experiments by Davisson and Germer in 1927 and independently in 1928 by G.P. Thomson, son of J.J. Thomson who had discovered the electron in 1897. Construction of the first electron microscope began in the early 1930s. E. Ruska, initially working in collaboration with M. Knoll, built this first instrument which was capable of a magnification of 12000 using two magnetic lenses. For anyone interested in the ins and outs of this period a delightful article by Gabor [1] can be warmly recommended, as can articles by Cosslett [2] and Ruska [3].

Using the relationship between the electron wavelength and momentum we can find an approximate expression for the wavelength as a function of the accelerating voltage V:

$$\lambda = 12V^{-1/2}; \tag{2}$$

this gives $\lambda = 0.037$ Å, for $V = 100$ keV and $\lambda = 0.018$ Å, for $V = 400$ keV. The expression is accurate to within a few percent in this voltage range; for instance, the relativistically corrected value at 400 keV is $\lambda = 0.0164$ Å. These wavelengths are five orders of magnitude smaller than for visible light. The diffraction-limited resolution will be of the order of:

$$d = \lambda/(\sin \alpha) > \lambda, \tag{3}$$

where 2α is the angular aperture of the objective lens. The atomic scattering factors for high-energy electrons are strongly peaked in the forward direction. Consequently, it is possible in theory to resolve distances smaller than the interplanar spacings in crystalline solids without needing to

[1] Present address: Institut de Biologie Structurale, 41 Avenue des Martyrs, 38027 Grenoble Cedex 1, France.

use the full angular aperture of the objective lens. This turns out to be very important since electron lenses have large aberrations. To set the scale of the spacings involved, the interatomic distances in crystalline gold are 2.88 Å, and in the case of organic material the carbon-to-carbon single covalent bond length is 1.54 Å.

Of course the early instruments had resolution limits which were far from approaching atomic resolution. They suffered from many technical problems including electrical and mechanical instabilities and the lack of astigmatism correction. Fundamental contributions to an improved understanding of image formation in the electron microscope were made by Scherzer, who had already shown in 1936 that under the usual operating conditions all rotationally symmetric electron lenses have a convergent effect on the electron beam [4]. Consequently it is not possible to correct third-order aberrations, and in particular spherical aberration. In the presence of spherical aberration the resolution limit (d) is given in terms of geometrical optics by the diameter of the disc of least confusion which is produced because rays passing at larger angles through the objective lens are focused closer to the lens than are paraxial rays:

$$d = \left(C_s \lambda^3\right)^{1/4}. \tag{4}$$

For 100 keV electrons and for a value of the spherical aberration coefficient $C_s = 1$ mm, the resolution predicted by this expression is $d \approx 5$ Å (only a few commercial instruments improve on the above value of C_s). Note that the spacings which can be resolved are two orders of magnitude bigger than the electron wavelength. At the present time the best electron microscopes have a point-to-point resolution below 2 Å. The expression above shows that the resolution depends on the term $C_s^{1/4}$ and consequently beyond a certain limit of electron lens design it becomes practically impossible to improve the microscope performance by reducing the spherical-aberration term. Since the resolution also depends on $\lambda^{3/4}$, the present generation of high-resolution electron microscopes use accelerating voltages of 300 or 400 keV.

It is interesting to note that the work of Scherzer mentioned above led Gabor, searching for a way around the limit imposed by spherical aberration, to publish in 1948 an article entitled "A New Microscopic Principle" [5]. He proposed a two-stage imaging process using coherent monochromatic illumination. The first stage involves making a photographic record of the interference pattern between a strong coherent background wave and the scattered wave from the object. The second step is to replace the developed photographic plate in the position it occupied during recording. An observer, looking upstream through the plate illuminated by the same coherent background wave, will see an image of the object in its original position. Gabor's proposition was to carry out the first stage using electrons and the second stage using light. Unfortunately, it turns out that the reconstruction step also generates an out-of-focus twin image superposed on the in-focus image. The major problem with the method as proposed by Gabor is to remove this twin image. The application of the method to electron microscopy was attempted around 1950 by Haine and Mulvey [6] but this work was unsuccessful, due to major technical problems unsurmountable at that time. A modified version of the method using an inclined reference beam appeared in light optics when a coherent and intense light source, the laser, became available. This time the modified method, *holography*, was a considerable success [7]. At the present time this off-axis holographic technique is being applied in electron imaging with encouraging results by several groups [8–10].

2. Interaction between the electron beam and the sample

The scattering of a high-energy electron beam by a thin sample is strongly peaked in the forward direction because the electron wavelength is small compared to atomic dimensions. Note that this is not the case for X-ray and neutron scattering where typical wavelengths are 1.54 Å for X-rays (Cu Kα radiation) and 1.3 Å for thermal neutrons ($T = 373$ K). The wave transmission function τ at

the exit surface of a thin specimen can be written [11]:

$$\tau(r, z) = \tau_0 \exp\left[-i\pi\lambda \int dz'\, U(r, z')\right], \qquad (5)$$

where the z axis is taken perpendicular to the specimen plane, r lies in the plane, the effective potential is $U(r, z) = 2mV(r, z)/h^2$, $V(r, z)$ is the inner potential of the specimen, $\tau_0 = \tau(r, 0)$ is the incoming wavefunction at the upper surface, $z = 0$, of the specimen. The argument of the exponential term represents the projection of the sample potential along the incident beam direction. It is convenient to set:

$$\phi(r) = \pi\lambda \int dz'\, U(r, z'), \qquad (6)$$

where ϕ, the phase shift of the wave function at the exit surface of the specimen, depends on both the object thickness and the inner potential. When $\phi \ll 1$ the transmission function $\tau = \tau_0 \exp(i\phi)$ can be approximated by:

$$\tau(r, z) = \tau_0[1 + i\phi(r)]. \qquad (7)$$

In this *weak-phase-object approximation* only a small component ϕ of the electron wave is scattered by the specimen. This is the expression usually used to describe the interaction of the electron beam with thin biological samples. The equations above can be inverted allowing $\int dz' U(r, z')$, the projected potential distribution of the sample, to be deduced from $\tau(r, z)$. In the absence of absorption *the phase-object approximation predicts that an image will have no contrast* since the intensity $\tau\tau^*$, see eq. (5), is constant when τ_0 corresponds to an incoming plane wave. Contrast can be produced by interference between the scattered and the unscattered waves, and in the electron microscope the usual way of generating such contrast is to vary the phase of the scattered wave components by simply changing the focus of the objective lens.

The situation described above is highly simplified, and even for thin biological samples there will usually be a weak residual contrast at zero defocus. Many factors contribute to this: scattering outside the objective aperture, inelastic scattering, multiple scattering, residual phase contrast due to C_s, and the essentially complex nature of elastic scattering amplitudes for electrons [12,13]. There are various phenomenological ways of taking account of the existence of an amplitude contrast component, such as including an extra term in the expansion of $\exp(i\phi)$. These all lead to an expression in which the transmission function has both a phase (ϕ) and an amplitude (u) component:

$$\tau(r, z) = \tau_0[1 + i\phi(r) + u(r)].$$

3. Contrast transfer theory

3.1. Parallel monochromatic illumination

The task of contrast transfer theory is to describe quantitatively the relationship between the wavefunction at the exit surface of the specimen and the final image intensity. It turns out that for a weak phase object, subject to conditions such as isoplanicity which are usually satisfied [14,15], the image distortions due to defocus and to spherical aberration can be described quantitatively in terms of spatial frequencies by a simple expression involving a contrast transfer function (CTF). This function, which can be directly assessed from the optical diffraction pattern of the image of a random scattering sample, is particularly well adapted to the many Fourier-based image treatment procedures. For these reasons the image quality of electron micrographs is invariably discussed in terms of the Fourier space representation involving the CTF rather than in terms of the alternative convolution relationship which holds for the image intensity itself.

Contrast transfer theory is based on foundations laid by Scherzer [16] within the framework of the Abbe theory in which image formation is described as a two-stage process. In brief, the theory is developed in the following way:

(i) calculate the Fourier transform of the object wavefunction to obtain the wave amplitude in the back focal plane of the objective lens,

(ii) multiply this by a phase factor describing the wave distortions due to aberrations,
(iii) inverse Fourier transform to obtain the wave amplitude in the image plane,
(iv) calculate the image intensity from the square modulus of this wave amplitude,
(v) calculate the Fourier transform of this intensity.

An excellent introductory article to the basic theory is that of Lenz [17] and on the experimental side reference must be made, of course, to the work of Thon [18]. The theory shows that to a good approximation, *in the case of parallel monochromatic illumination*, the intensity distribution in the image of a *weak phase object* has the Fourier transform $\tilde{I}(\theta)$:

$$\tilde{I}(\theta) = \delta(\theta) + \tilde{\phi}(\theta)\,\sin[(2\pi/\lambda)\,W(\theta)]. \qquad (8)$$

This very useful and simple equation shows that the angular (or spatial frequency) spectrum (diffraction amplitude) of the image intensity is that of the object itself multiplied by an instrument-dependent term. This term, $K(\theta) = \sin[(2\pi/\lambda) \times W(\theta)]$, is called the contrast transfer function. Fourier transforms are indicated by the tilde, $\tilde{\phi}(\theta)$ is the transform of the specimen and $W(\theta)$ is the wave aberration:

$$W(\theta) = -z\theta^2/2 + C_s\theta^4/4, \qquad (9)$$

where z represents the objective lens defocus, C_s the spherical aberration constant and θ is the scattering angle. The wave aberration W describes the distortion of the wavefront in the back focal plane relative to the Gaussian image reference sphere. This wave distortion can be expressed in terms of scattering angle as above or in terms of spatial frequency f, for which the phase distortion due to aberrations is $2\pi W(f) = 2\pi(-z\lambda f^2/2 + C_s\lambda^3 f^4)$, fig. 1. Due to the small value of the electron wavelength, the scattering of high-energy electrons is close to the forward direction. To set the scale, $\theta = \lambda f$ is about 2° for a spatial frequency $f = (1\ \text{Å})^{-1}$.

Over limited spatial-frequency ranges, spherical aberration can be balanced by an appropriate defocus, fig. 1. These regions, in which the phase $2\pi W(f)$ due to the wave aberration is stationary

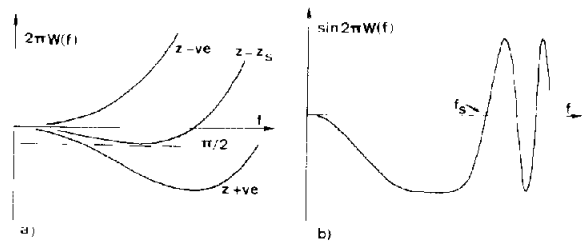


Fig. 1. (a) Dependence of the form of the wave aberration phase $2\pi W(f) = \pi(-z\lambda f^2 + C_s\lambda^3 f^4/2)$ on the defocus z. In this expression positive values of z correspond to an objective lens underfocus. For the Scherzer defocus $z_s = (C_s\lambda)^{1/2}$, the wave aberration has a stationary phase value of $-\pi/2$. (b) At the Scherzer defocus the width of the first oscillation zone of the contrast transfer function, $\sin[2\pi W(f)]$, extends up to $f_s = 1.4(C_s\lambda^3)^{-1/4}$ giving the Scherzer resolution limit $d = 0.7(C_s\lambda^3)^{1/4}$.

($\mathrm{d}W/\mathrm{d}f = 0$), play an important rôle in contrast transfer theory. For example, the defocus $z = (C_s\lambda)^{1/2}$ which sets the stationary phase value to $-\pi/2$ is called the Scherzer defocus. At this defocus object distances down to $d = 0.7(C_s\lambda^3)^{1/4}$ lie within the first peak of the contrast transfer function, fig. 1b, and are consequently transferred to the image with the same contrast. This value of d defines a resolution limit which is essentially identical to that predicted by the disc of least confusion of geometrical optics. We will see below that spatial frequencies in the defocus-dependent stationary phase regions are preferentially transferred when the effect of a finite source size is considered.

Eq. (8) shows that, in the case of weakly scattering samples, the electron-optical aberrations do not destroy the linear relationship between the image intensity and the projected structure of the sample. The spatial frequency content of the image differs from that of the sample only because of the frequency-dependent contrast reversals due to the oscillating term $\sin[(2\pi/\lambda) \times W(\theta)]$ which does not in itself impose a resolution limit.

If the imaging process concerns an object with both phase and amplitude components the

Fourier transform of the object-dependent part of the image intensity will be

$$\tilde{I}(\theta) = \tilde{\phi}(\theta) \sin[(2\pi/\lambda)\, W(\theta)] + \tilde{u}(\theta) \cos[(2\pi/\lambda)\, W(\theta)]. \quad (10)$$

Experimental assessments made from electron micrographs of periodic protein arrays have shown that the amplitude component can be as high as 35% in negative stain [19] and is of the order of 7% for samples preserved in vitreous ice [20].

From here on, scattering angles θ will be replaced either by spatial frequencies $f = \theta/\lambda$ or by the so-called generalised spatial frequency and defocus. The use of the latter gives a considerable simplification of most expressions and has the advantage of giving sets of universal curves. The generalised coordinates can be expressed relative to the Scherzer values:

$$Z = z/z_s \quad \text{and} \quad F = f/f_s, \quad (11)$$

where $z_s = (C_s\lambda)^{1/2}$ is the Scherzer defocus, and the corresponding spatial frequency is $f_s = (C_s\lambda^3)^{-1/4}$.

3.2. Contrast transfer and holography

It is interesting to remark that contrast theory gives a quantitative formulation of holography [21]. All bright-field electron micrographs of weak scatterers are in fact in-line Fresnel holograms produced by interference between the strong unscattered wave and the wave weakly scattered by the sample. The problem of correcting the contrast reversals due to the CTF corresponds to removing the twin image in the holographic reconstruction. A simple way of understanding the imaging process can be seen in the case of a point object illuminated with a parallel beam as shown in fig. 2. An intensity distribution H, similar to a zone plate, is produced by interference between the spherical wave S scattered by the point object O and the reference plane wave P. The reconstruction step involves illuminating the hologram (H) by the same, or by a similar, background wave. It is well known that zone plates produce multiple images and in the special case of a sine-modulated zone plate (the hologram) two

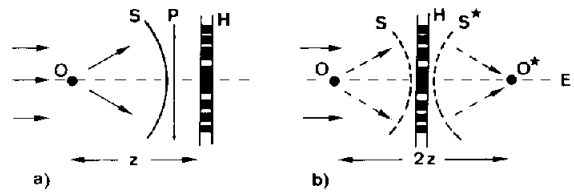


Fig. 2. An image of a weakly scattering object is an in-line Fresnel hologram. A thin sample can be considered as an assembly of point scatterers. (a) For each scatterer a zone-plate-like intensity distribution H is produced by interference between the spherical wave S scattered by the point object O and the reference plane wave P. (b) The reconstruction involves illuminating H with the same background wave. This generates two spherical waves S and S* centred on the twin images O and O*. Both images can be seen by looking through the hologram from E, and if the eye is focussed on one image, the other, which is always exactly superposed, has twice the initial defocus. The Fourier transform of the hologram, not shown, is again a zone-plate-type intensity distribution and corresponds to the contrast transfer function.

images, O and its conjugate O*, are produced on either side of the hologram as viewed from E. These images are separated along the axis by twice the initial defocus, and since they are aligned they are always superposed. The Fourier transform of the hologram corresponds to the contrast transfer function. In this view of the imaging process each atom within a sample will act as an independent scatterer, like the point object O. The final image intensity will then be the sum of the contributions of the individual holograms from all the atoms in the sample.

3.3. Taking account of the angular aperture of illumination

In practice, of course, we never have a perfectly parallel, monochromatic incident electron beam. Electrons are emitted randomly from a part of the filament surface within the gun, and form a cross-over above the anode. A demagnified image of this cross-over is usually formed above the specimen plane using the condenser lenses. If electrons arrive randomly from this "effective source" [22] then the image intensities

due to each individual source element must be summed to obtain the final image intensity. Taking account of the angular distribution of the source intensity gives the source-dependent contrast transfer function K_s:

$$K_s(F) = E_1(Q_0, Z, F)\sin[2\pi W(F)], \qquad (12)$$

where F represents the generalised spatial frequency, and the source-dependent term E_1, usually called the envelope, produces a spatial-frequency-dependent attenuation of the contrast transfer function $K(F)$, previously obtained for parallel illumination. For a Gaussian source intensity distribution,

$$E_1 = \exp\left\{-\left[\pi Q_0 F(F^2 - Z)\right]^2\right\}, \qquad (13)$$

where the source size, in generalised units, is given by Q_0 = angular source size $(C_s/\lambda)^{1/2}$. The effect of the envelope E_1 is shown in the characteristic curves [23], fig. 3. Note that there is a strong preferential transfer of the stationary phase regions which occur at the defocus-dependent frequencies $F = Z^{1/2}$. The physical origin of the envelope term is to be found in the frequency-dependent image displacements due to the finite range of illumination angles, and it is easy to see from the form of the wave aberration surface why the stationary phase regions are favourably transferred. Historically, the effects of the illumination source size were first considered by Frank [24] and by Bonhomme et al. [25]. Only the paper by Frank is directly relevant to the envelope function representation as described here. It is important to note that, in the presence of both spherical aberration and defocus, the envelope function

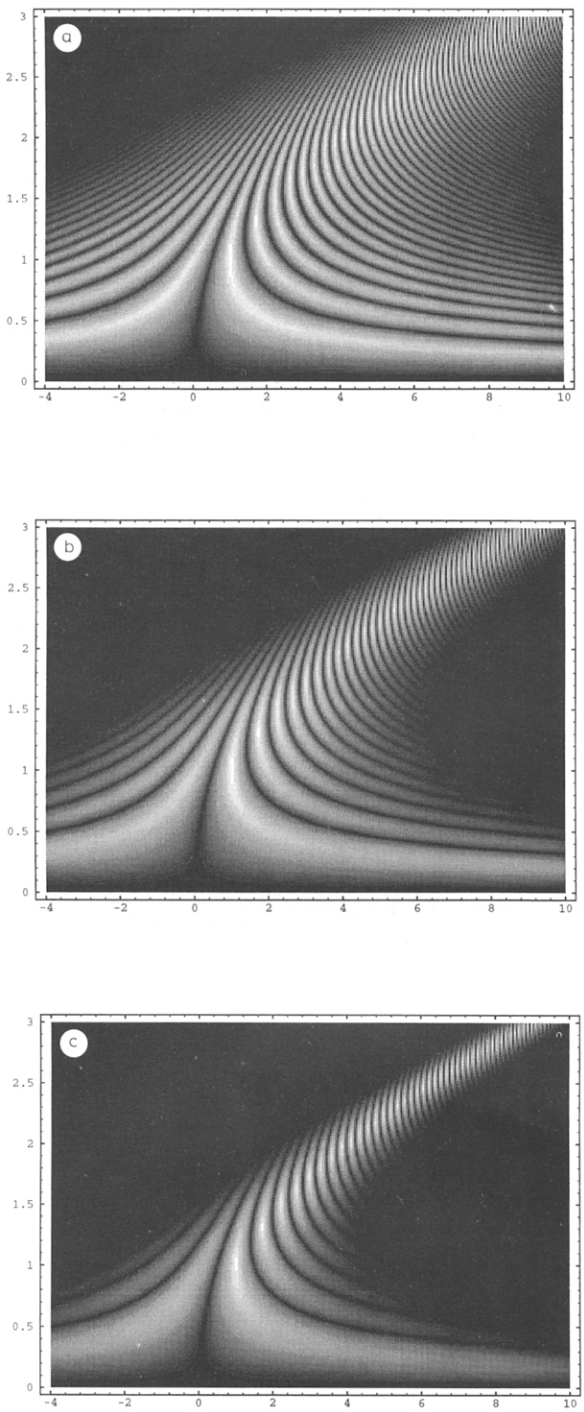
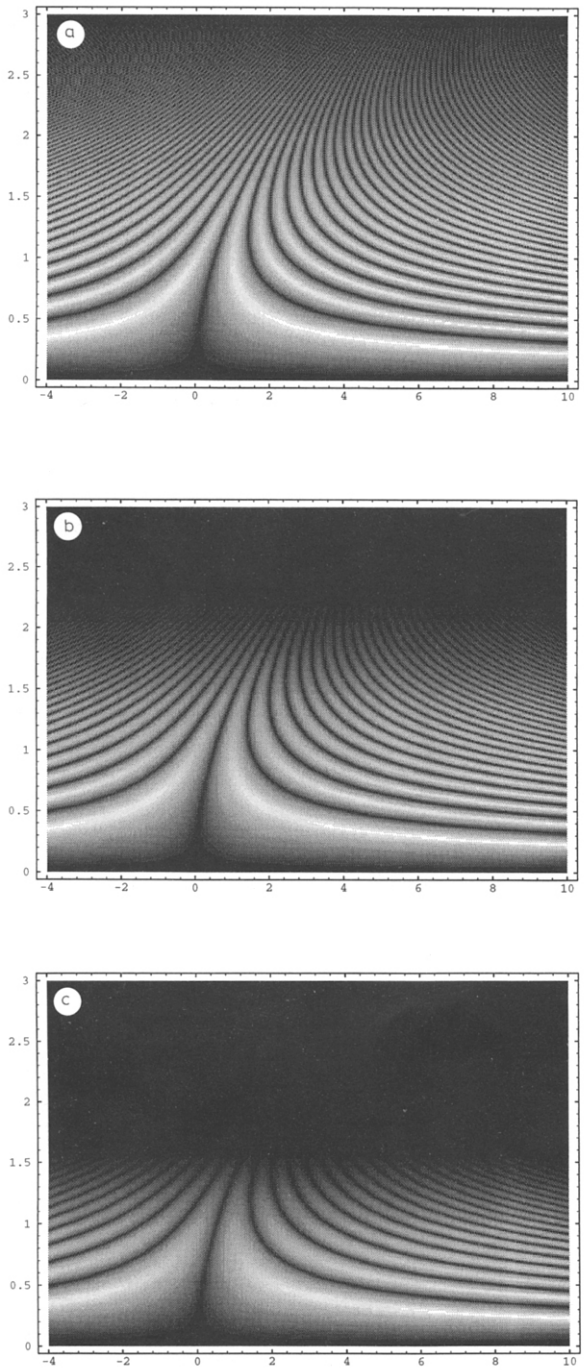


Fig. 3. Representation of the so-called contrast transfer characteristics for the expression $E_1 \sin[2\pi W(F)]$, showing the effect of the envelope E_1 which takes account of the illumination source size. The vertical axis corresponds to generalised spatial frequency $F = f(C_s\lambda^3)^{1/4}$ and the horizontal axis to generalised defocus $Z = z(C_s\lambda)^{-1/2}$ so that a profile along a vertical line gives the CTF for the corresponding Z. The width of the source Q_0, in generalised units, is (a) $Q_0 = 0.05$, (b) $Q_0 = 0.1$, (c) $Q_0 = 0.175$. These values correspond respectively to illumination aperture half angles of 3.5×10^{-4}, 7×10^{-4} and 1.4×10^{-3} rad, for $C_s = 1.4$ mm and $\lambda = 0.037$ Å. For a given defocus the practical resolution will be strongly dependent on the source size. The preferentially transferred zones running to higher resolution correspond to the defocus-dependent stationary phase regions.

gives only an approximate description of the effect of the source size on contrast transfer. When the wave distortion is due to defocus alone, the envelope representation of the effect of source size is mathematically exact [26], so this is an important limiting case in support of the validity of an envelope function representation in the presence of other aberrations.



3.4. Taking account of defocus fluctuations

The focal length of an electromagnetic lens depends on the excitation current and on the electron beam energy. Consequently the combined effects of the energy spread of the beam, the electrical fluctuations of the lens ($\mathrm{d}I/I$), and the instability of the high voltage ($\mathrm{d}V/V$) are to produce a defocus spread Δ. This spread will depend on the chromatic aberration constant (C_c) of the objective lens through a relationship of the type $\Delta = C_c(\mathrm{d}V/V + 2\mathrm{d}I/I)$. Partial chromatic coherence was first dealt with by Hanszen and Trepte [27], and it turns out that in terms of the contrast transfer theory the effects can also be described by an envelope function E_2:

$$E_2 = \exp\left\{-\left(\pi \Delta_G F^2/2\right)^2\right\}, \qquad (14)$$

where Δ_G is the half-width of the defocus spread distribution, $\Delta_G = \Delta/z_s$. It is usually convenient, and physically justified, to represent Δ_G by a Gaussian distribution. The effect of the term E_2, fig. 4, is quite different from that of the source term E_1, and it is this envelope which is ultimately responsible for the electron-optical resolution limit of an electron microscope since it imposes a spatial frequency cut-off depending on the defocus spread Δ_G.

Fig. 4. Characteristic curves as in fig. 3 showing the effect of increasing defocus spread Δ_G as expressed by the envelope term E_2. This is the resolution-limiting term. The values of the term Δ_G are (a) 0.125, (b) 0.25, (c) 0.5. These correspond respectively to defocus fluctuations of 90, 180 and 360 Å for $(C_s\lambda)^{1/2} = 720$ Å, i.e. the same value of C_s and λ as for fig. 3. For $C_c = 1.4$ mm, and ignoring the other instabilities, these values would correspond respectively to electron beam spread half-widths of 0.6, 1.2 and 2.4 eV.

3.5. Combined effect of the source size and defocus fluctuations

Finally we need to know what happens to the envelope functions when the effects of both the source size and the defocus spread are taken into account. It turns out that to a good approximation the modified contrast transfer can be described by the product of the two envelopes indicated above [28]:

$$K_{\text{overall}} = E_1 E_2 \sin[2\pi W(F)]. \qquad (15)$$

Strictly speaking, each of the individual terms is modified by a factor $1/(1+AF^2)$, with $A = (\pi Q_0 \Delta_G)^2$, which can significantly modify $W(F)$ for large values of F [28]. This region is usually attenuated by the combined envelope $E_1 E_2$. There is also an additional envelope term which comes into play when the beam is tilted. Neither of these effects will be considered further here.

4. The effect of beam tilt

In the early 1970s tilted-beam illumination was used to obtain high-resolution images of "amorphous" films of silicon and germanium. The question under examination was whether such vacuum-condensed films were truly amorphous or were agglomerates of randomly oriented microcrystals. The experimental electron microscopy was no doubt inspired by earlier work such as that of Dowell who, in the early 1960s, had been able to obtain 3.2 Å lattice images of tremolite [29]. This was achieved by tilting the electron beam so that the lattice images were obtained with both the direct and the diffracted beams equally inclined relative to the optical axis. With this geometry the two beams have the same aberration-induced phase, and when the envelope terms are taken into account we find that for the critical defocus $z = -C_s f_0^2$ the spatial frequencies on the so-called achromatic circle, radius f_0, are always strongly transferred.

In the case of Si and Ge films Rudee and Howie [30] obtained images showing "lattice fringes" in small regions of the samples. These observations were considered to favour the microcrystalline structural model. This caused considerable controversy in the field and led to a number of investigations of the effects of beam tilt on contrast transfer. Particular mention should be made of the work of McFarlane [31]. Other work on this structural theme rapidly followed, see for example refs. [32–36]. The main importance of this work as far as the use of electron microscopy in structural biology is concerned is the considerable impact on alignment procedures [37,38] and on the awareness of the necessity of phase correction at high resolution [39].

Contrast transfer for a beam tilt of f_0 is described by the function $K(f, f_0)$ [28]:

$$K(f, f_0) = \mathrm{i}\left[t^*(f_0)\, t(f_0+f) - t(f_0)\, t^*(f_0-f)\right], \qquad (16)$$

where $t(f) = \exp[\mathrm{i}2\pi W(f)]$ and $W(f) = -\lambda z f^2/2 + \lambda^3 f^4 C_s/4$. Working through this we find that the term $\sin[2\pi W(f)]$ obtained for parallel axial illumination is now replaced by the product of a phase term and $\sin[W(f, f_0)]$:

$$\sin[W(f, f_0)] = \sin\left\{2\pi\left[\lambda f^2\left(-z + C_s(\lambda f_0)^2\right)/2 + C_s\lambda^3 f^2 f_0^2 \cos\phi + C_s\lambda^3 f^4/4\right]\right\}, \qquad (17)$$

where one of the beam-tilt-dependent terms, $C_s(\lambda f_0)^2$, behaves like an over-focus offset and the other additional term corresponds to a tilt-dependent astigmatism $C_s\lambda^3 f^2 f_0^2 \cos\phi$. The phase term is given by:

$$\exp\left\{\mathrm{i}2\pi\left[\lambda \boldsymbol{f}\cdot\boldsymbol{f}_0\left(-z + C_s(\lambda f_0)^2 + C_s(\lambda f)^2\right)\right]\right\};$$

this includes a term, $\lambda \boldsymbol{f}\cdot\boldsymbol{f}_0(-z + C_s(\lambda f_0)^2)$, corresponding to a defocus and beam-tilt-dependent image shift, and a second term, $\lambda \boldsymbol{f}\cdot\boldsymbol{f}_0 C_s(\lambda f)^2$, which represents a frequency-dependent image shift (axial coma). Naturally, in the axial illumination limit ($f_0 = 0$), $K(f, 0)$ is equivalent to eq. (8).

It is the behaviour of the diffractogram intensity involving the term $\sin^2[W(f, f_0)]$ as a function of tilt angle and direction which is used in the alignment scheme exploited by Zemlin [38].

In the case of the determination of high-resolution protein structures by electron crystallography, Henderson et al. [39] have shown that even for relatively small alignment errors the phase term must be taken into account because of the f^3 dependence of the coma-like term $C_s(\lambda f)^3 f_0 \cos\phi$, ϕ is the angle between f_0, the direction of beam tilt, and a given spatial frequency f.

5. Specimen thickness

The interpretation of electron micrographs and the different three-dimensional reconstruction schemes are based on the linear relationship between the image contrast and the projected potential of the sample. We need to know whether the depth of field is sufficient for this to always be valid. One way of judging this is by reference to contrast characteristic curves like figs. 3 and 4, but let us first of all describe an experiment by Bonhomme and Boerschia [40] in which images were recorded on either side of thickness steps in amorphous carbon films. The positions of the diffactogram maxima show that on the same micrograph there is a defocus difference between the thin and the thicker regions. Taking the thin region as reference, the defocus difference changes sign when the specimen is turned upside down. These results indicate that the in-focus position is half way through the sample thickness and not at the output surface, as would be expected from a direct use of the projected potential as described earlier. An explanation of this result is found if we consider each atom in the sample to scatter independently as for the point scatterer in the holographic scheme, fig. 2. The defocus of each elementary hologram will depend on the position of the atom in the sample. The overall image intensity will be the sum of these independent contributions across the thickness of the sample. To a good approximation we find the same contrast transfer function as previously but with the defocus origin in the middle of the sample thickness and not at the exit surface and with an additional modulation by the thickness-dependent terms shown below:

$$\sin[2\pi W(f)]\left[\sin(\pi\lambda f^2 d/2)\right]/\pi\lambda f^2 d/2,$$

where d represents the sample thickness. Taking the first zero of this thickness-dependent term as an indication of the effect on the contrast transfer we find that there is a cut-off at a resolution of 2 Å for a 200 Å thick sample, whilst at a resolution of 3 Å there is a 0.66 attenuation of the CTF. The effect of specimen thickness on resolution through the sinc function above has also been discussed previously by Zeitler [41].

Another more intuitive way of obtaining an idea of whether a straightforward projection is likely to be a good approximation is to refer to the contrast transfer characteristics. It is easy to see from fig. 4 that, at a relatively strong defocus, $\sin[2\pi W(f)]$ varies very slowly with defocus for low frequencies and much more strongly for higher frequencies. This amounts to having a large depth of field for imaging at resolutions of around 20 Å. For samples a few hundred Å thick this will no longer hold for imaging at higher resolutions.

6. Amorphous carbon and vitreous ice

Ever since optical diffractograms have been used to assess the quality of electron micrographs the standard test objects have been vacuum-condensed carbon films [18]. Such films are amorphous and have been found to give a good approximation to a "white" spatial frequency spectrum. In the case of observations of frozen-hydrated specimens the biological object is observed in a thin layer of vitreous (amorphous) ice. In this case it has been found experimentally that optical diffractograms of the micrographs are no longer much use to reveal the CTF. For some reason the notion of a random scattering, or white, object appears inappropriate for ice even though, like amorphous carbon, it can be supposed to consist of a "random" distribution of scattering centres. Moreover, carbon and oxygen have rather similar atomic scattering factors and the elastic scattering by hydrogen atoms can probably be neglected. Taking an average interatomic separation of 1.5 Å, a sample thickness of 100 to 200 Å will correspond to a stack of some fifty to a hundred atoms. There is unlikely to be a significant varia-

tion in the projected potential from point to point at the exit surface, and consequently very little phase variation. Why then do carbon films, but not ice layers, give defocused images with a strong granularity? A possible explanation is suggested by work in connection with the effect of the substrate roughness on the state of order in thin two-dimensional crystals [42]. It was shown by scanning tunnelling microscopy that vacuum-deposited carbon films can have thickness variations Δ_d of up to 20 Å. A simple estimate based on the experimental value of the inner potential for carbon films $V_0 = 10$ eV shows that this could give rise to phase fluctuations of around $\pi/20$ (where we take phase variations at the exit surface as $\pi\Delta_d V_0/\lambda V$). The difference between carbon and ice could perhaps then be explained in terms of surface smoothness with carbon having at least one rough surface, depending both on the substrate used and on the deposition conditions, and with ice having two atomically smooth surfaces.

7. Correcting for the contrast transfer function

There have been a considerable number of proposals for correcting the contrast transfer function. This was especially true during the early stages in the development of the theory. Most of these methods have fallen into oblivion and it is not opportune or possible to attempt a detailed description of them all. The discussion will be limited to what, as far as I can see, is the first such proposal and then two important practical solutions in use at present will be briefly described. In the framework of the imaging theory presented here, the aim of any correction scheme must be to convert the CTF from $\sin[2\pi W(f)]$ to unity without introducing any additional noise. Naturally this is particularly difficult for the spatial frequencies at or near the zero points of the contrast transfer function.

An early proposal for correction was made in 1951 by Bragg and Rogers [43] in the context of Gabor's holographic method. Although the proposal cannot find any direct application in electron microscopy it is worth consideration for historical reasons, for its elegant simplicity and because it is the precursor of most schemes in that it involves using data from more than one image. Unfortunately for electron microscopists, this method is only valid for an amplitude object and requires a controlled variation of the wave aberrations. This is possible for defocus but not for spherical aberration. Two images are recorded at defocus values of z and $2z$. A holographic reconstruction is made from the first image as shown in fig. 2. The contrast transfer function associated with the reconstruction has the form $\cos^2(\pi\lambda z f^2) = 1 + \cos(2\pi\lambda z f^2)$ [44]. Consequently this CTF can be corrected directly by placing the negative recorded at the defocus $2z$ in register with the reconstruction.

As far as practical solutions are concerned, a two-image method was used in a recent helical reconstruction of the acetylcholine receptor to 17 Å resolution [45] using tubular receptor arrays observed in vitreous ice. Because of the sinusoidal form of the contrast transfer function a single image cannot cover the necessary resolution range with a good signal-to-noise ratio. Consequently micrographs were recorded in pairs, at defocus values of 0.8 and of 2 μm; note how close this is to the two-hologram situation described in the previous paragraph. For these defocus values the first peaks of the contrast transfer function correspond respectively to ~ 25 Å and to ~ 40 Å. The data from both micrographs was combined to give a reasonably equilibrated contrast transfer over the range of spacings from 17 Å to about 100 Å. In addition, the very-low-resolution region along the equator (spacings greater than 100 Å) was corrected using theoretical curves corresponding to a 7% amplitude contrast component [20].

Finally, mention should be made of the treatment of image data in the case of three-dimensional determinations of protein structures to high resolution. This is also a two-image method but it relies on combining data from micrographs and from electron diffraction patterns [46]. The amplitudes of the Fourier components are obtained directly from the intensities of the electron diffraction peaks since these are not influenced by the contrast transfer function. The corresponding phases are determined from the com-

puted Fourier transforms of the micrographs. Amongst other factors account must be taken of the phase reversals due to the oscillating sign of the contrast transfer function and to the phase shifts due to slight electron-optical misalignments of the illumination with respect to the optical axis of the objective lens [39].

8. Conclusion

This is an extremely condensed and personal account of the development of contrast transfer theory over the past twenty-five years or so. It is hoped that those interested in imaging biological specimens will find some useful information such as, for example, the importance of the illumination aperture when imaging at large defocus, fig. 3. An outstanding question, briefly discussed, is the difference between the behaviour, as random scatterers, of vitreous ice and of amorphous carbon. Also, do not forget that bright-field images of weak phase objects are holograms.

Acknowledgements

I would like to thank E. Zeitler for his illuminating comments on this contribution and, in another vein, for many conversations about this and that; mostly that.

References

[1] D. Gabor, in: Proc. 8th Int. Congr. on Electron Microscopy, Canberra, 1974, Vol. 1, Eds. J.V. Sanders and D.J. Goodchild, p. 6.
[2] V.E. Cosslett, in: Advances in Optical and Electron Microscopy, Vol. 10, Eds. R. Barer and V.E. Cosslett (Academic Press, London, 1987) pp. 215–267.
[3] E. Ruska, Rev. Mod. Phys. 59 (1987) 627.
[4] O. Scherzer, Z. Phys. 101 (1936) 593.
[5] D. Gabor, Nature 161 (1948) 777.
[6] M.E. Haine and T. Mulvey, J. Opt. Soc. Am. 42 (1952) 763.
[7] E.N. Leith and J. Upatnieks, J. Opt. Soc. Am. 52 (1962) 1123; 53 (1963) 1377; 54 (1964) 1295.
[8] H. Lichte, in: Advances in Optical and Electron Microscopy, Vol. 12, Eds. T. Mulvey and C.J.R. Sheppard (Academic Press, London, 1991) pp. 25–91.
[9] A. Tonomura, Rev. Mod. Phys. 59 (1987) 639.
[10] G. Matteuci, G.F. Missiroli, E. Nichelatti, A. Migliori, M. Vanzi and G. Pozzi, J. Appl. Phys. 69 (1991) 1835.
[11] B.F. Buxton, in: Imaging Processes and Coherence in Physics, Eds. M. Schlenker, M. Fink, J.P. Goedgebuer, C. Malgrange, J.C. Viénot and R.H. Wade (Springer, Berlin, 1980) pp. 175–184.
[12] J.M. Cowley, Diffraction Physics (North-Holland, Amsterdam, 1975) pp. 75–82.
[13] E. Zeitler and H. Olsen, Phys. Rev. 162 (1967) 1439.
[14] P.W. Hawkes, in: Computer Processing of Electron Microscope Images, Ed. P.W. Hawkes (Springer, Berlin, 1980) pp. 1–33.
[15] F. Lenz, in: Quantitative Electron Microscopy, Eds. G.F. Bahr and E.H. Zeitler (Lab. Inv., Baltimore, 1965) pp. 70–80.
[16] O. Scherzer, J. Appl. Phys. 20 (1948) 20.
[17] F. Lenz, in: Electron Microscopy in Materials Science, Ed. U. Valdrè (Academic Press, London, 1971) pp. 540–569.
[18] F. Thon, in: Electron Microscopy in Materials Science, Ed. U. Valdrè (Academic Press, London, 1971) pp. 571–625.
[19] H.P. Erickson and A. Klug, Phil. Trans. Roy. Soc. B 261 (1971) 105.
[20] C. Toyoshima and P.N.T. Unwin, Ultramicroscopy 25 (1988) 279.
[21] R.H. Wade, in: Computer Processing of Electron microscope Images, Topics in Current Physics, Ed. P.W. Hawkes (Springer, Berlin, 1980) pp. 223–255.
[22] H.H. Hopkins, Proc. Roy. Soc. A 208 (1951) 263; A 217 (1953) 408.
[23] R.H. Wade, Ultramicroscopy 3 (1978) 329.
[24] J. Frank, Optik 38 (1973) 519.
[25] P. Bonhomme, A. Boerschia and N. Bonnet, CR Acad. Sci. (Paris) 277 (1973) B-83.
[26] J.P. Guigay, R.H. Wade and C. Delpla, in: Proc. 25th Anniv. Meeting EMAG, Ed. W.C. Nixon (Institute of Physics, London, 1971) p. 238.
[27] K.J. Hanszen and L. Trepte, Optik 32 (1971) 519.
[28] R.H. Wade and J. Frank, Optik 49 (1977) 81.
[29] W.C.T. Dowell, Optik 20 (1963) 535.
[30] M.L. Rudee and A. Howie, Phil. Mag. 25 (1972) 1001.
[31] S.C. McFarlane, J. Phys. C (Solid State Phys.) 8 (1975) 2819.
[32] S.C. McFarlane and W. Cochran, J. Phys. C (Solid State Phys.) 8 (1975) 1311.
[33] W. Krakow, D.G. Ast, W. Goldfarb and B.M. Seigel, Phil. Mag. 33 (1976) 985.
[34] R.H. Wade, Phys. Status Solidi (a) 37 (1976) 247.
[35] R.H. Wade and K.H. Jenkins, Optik 50 (1978) 1.
[36] D.J. Smith, W.O. Saxton, M.A. O'Keefe, G.J. Wood and W.M. Stobbs, Ultramicroscopy 11 (1983) 263.
[37] F. Zemlin, K. Weiss, P. Schiske, W. Kunath and K.-H. Herrmann, Ultramicroscopy 3 (1978) 49.
[38] F. Zemlin, Ultramicroscopy 4 (1979) 241.
[39] R. Henderson, J.M. Baldwin, K.H. Downing, J. Lepault and F. Zemlin, Ultramicroscopy 19 (1986) 147.

[40] P. Bonhomme and A. Boerschia, J. Phys. D (Appl. Phys.) 16 (1983) 705.
[41] E. Zeitler, in: Advances in Electronics and Electron Physics, Vol. 25, Ed. L. Marton (Academic Press, London, 1968) p. 227.
[42] H.-J. Butt, D.N. Wang, P.K. Hansma and W. Kühlbrandt, Ultramicroscopy 36 (1991) 307.
[43] W.L. Bragg and G.L. Rogers, Nature 167 (1951) 190.
[44] R.H. Wade, Optik 44 (1974) 447.
[45] C. Toyoshima and P.N.T. Unwin, J. Cell Biol. 111 (1990) 2623.
[46] P.N.T. Unwin and R. Henderson, J. Mol. Biol. 94 (1975) 425.