Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Macromolecular structure solution by charge flipping

Christian Dumas[a]* and Arie van der Lee[b]

[a]Centre de Biochimie Structurale, CNRS, UMR5048, INSERM U554, Université Montpellier-1, F-34090 Montpellier, France, and [b]Institut Européen des Membranes, CNRS, UMR5635, Université Montpellier-2, Place Eugène Bataillon, F-34095 Montpellier CEDEX 5, France

Correspondence e-mail: christian.dumas@cbs.cnrs.fr

Received 16 April 2008
Accepted 9 June 2008

The recently discovered charge-flipping phasing algorithm has received growing interest in small-molecule crystallography and powder diffraction. This computational methodology differs from classical direct methods as it does not require *a priori* knowledge of either space-group symmetry or chemical composition and does not rely on probabilistic phase relations. Here, it is shown that the charge-flipping algorithm is capable of solving large macromolecular structures with up to ∼6000 atoms in the asymmetric unit using suitable normalized intensity data at atomic resolution (∼1.0 Å). Moreover, it is demonstrated that this algorithm also provides an efficient tool for the experimental phasing of highly complex heavy-atom or anomalous scattering substructures at medium to low resolution (∼2–6 Å) that are frequently difficult to determine using Patterson techniques or direct methods. With the present extension to macromolecular crystallography, charge flipping has proved to be a very well performing and general phase-recovery algorithm in all fields of kinematical diffraction.

1. Introduction

Direct methods have been the method of choice for structure solution from diffraction data since the start of their development more than 50 y ago, initially only for small-molecule structures but later also for macromolecules, solving not only complete atomic resolution structures but also heavy-atom or anomalous scatterer substructures. In this respect, an important development was the advent of so-called dual-space methods (Weeks & Miller, 1999; Schneider & Sheldrick, 2002) in which direct methods in reciprocal space were alternated with density modification in real space, thus considerably increasing the complexity of solvable protein structures. Recently, a real-space double-cycling method based on Patterson deconvolution techniques and density modification has been introduced, pushing the limit of protein structures solvable by *ab initio* methods even further (Burla *et al.*, 2006). In 2004 a new completely different dual-space method was developed by Oszlányi & Sütő (2004, 2005) which could be used to solve atomic resolution small-molecule structures (Wu *et al.*, 2004). The method, called charge flipping (CF), has its roots in the original Gerberg–Saxton–Fienup iterative algorithm (Fienup, 1982; Gerchberg & Saxton, 1972) for the phase retrieval of nonperiodic objects in optics and can be considered as the crystallographic translation of the mathematical successive over-relaxation procedure used to iteratively solve a linear or nonlinear system of equations. This iterative scheme can be written as

$$\rho_n(\mathbf{r}) = P_2\mathcal{F}P_1\mathcal{F}^{-1}\rho_{n-1}(\mathbf{r}), \quad (1)$$

where $\rho_n(\mathbf{r})$ is the electron density at the nth iteration and $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the forward and inverse Fourier transform operators, respectively. P_1 and P_2 are operators acting in reciprocal and real space, respectively, and are given by

$$P_1G = |G_{\rm obs}|G/|G| \text{ or } P_1G = G\exp(i\Delta\varphi) \quad (2)$$

and

$$P_2\rho(\mathbf{r}) = \text{sign}[\rho(\mathbf{r}) - k_{\rm ed}\sigma]\rho(\mathbf{r}), \quad (3)$$

where G is a complex structure factor and $|G_{\rm obs}|$ is the observed structure-factor amplitude. $\Delta\varphi = \pi/2$ if $|G_{\rm obs}|$ belongs to the w percent weakest amplitudes (Oszlányi & Sütő, 2005) and $P_1G = G\exp(i\Delta\varphi)$ is applied; otherwise, $P_1G = |G_{\rm obs}|G/|G|$. Usually $|G_{\rm obs}| = |F_{\rm obs}|$, where $|F_{\rm obs}|$ is proportional to the square root of the measured intensity. $k_{\rm ed}\sigma$ is the basic CF parameter under which the sign of the charged density is flipped. σ is the standard deviation of the electron density, $\sigma^2 = (1/V^2)\sum_{s\neq 0}|G_{\rm obs}(\mathbf{s})|^2$, where V is the volume of the unit cell. The relaxation parameter $k_{\rm ed}$ is positive and is slightly larger than 1 for proteins, but can be smaller than 1 for small-molecule structures. This iterative scheme thus involves an efficient phase-space exploration arising from a fine balance of perturbations in direct and reciprocal space and experimental moduli constraints. The crystallographic structures are solved in the $P1$ space group using a list of reflections expanded from a unique list of reflections averaged in their correct Laue class. Therefore, symmetry constraints are only imposed weakly during the iterative phasing process. The latter strategy has also occasionally been followed when solving structures using direct methods (Sheldrick & Gould, 1995; Burla *et al.*, 2000; Caliandro *et al.*, 2007).

The unique features of charge flipping compared with direct methods are that neither compositional information nor explicit positivity or atomicity constraints nor crystal symmetry are used, *i.e.* it is a truly *ab initio* phasing method, more so than direct methods. In addition, the parameterization of the algorithm is very simple compared with those of direct method-based dual-space strategies.

It was soon realised that apart from being a powerful small-molecule structure-solution method, charge flipping is also a unique tool to solve structures (Gies, 2007) for which direct methods are not available or less tractable, such as incommensurately modulated structures (Palatinus, 2004) or even quasicrystals (Katrych *et al.*, 2007) and also structures from powder diffraction data (Baerlocher, Gramm *et al.*, 2007; Baerlocher, McCusker *et al.*, 2007). In a recent review (Oszlányi & Sütő, 2008), it was indicated that charge flipping can be used to complete atomic resolution small-protein structures starting from a known relatively simple heavy-atom substructure representing 0.5–1% of the total density. We not only confirm this finding, but we also show that charge flipping can be used to solve *ab initio* the complete structures of relatively complex protein crystals of up to ∼6000 atoms in the asymmetric unit with experimental data measured at atomic resolution (∼1.0 Å).

For macromolecular structures that are larger or diffract to lower resolution, the phase problem is usually solved using experimental phasing techniques such as the selenomethionine-derivatization method (Hendrickson *et al.*, 1990); over two-thirds of novel protein structures are currently determined using single- or multi-wavelength anomalous dispersion (SAD or MAD) phasing strategies (Hendrickson, 1991). Solving novel crystal structures of macromolecules at medium to high resolution involves the localization and use of heavy atoms that serve as markers for the initial phasing through their anomalous or isomorphous scattering signal (Dauter, 2002). We demonstrate here that simple to large heavy-atom or anomalous substructures (8–40 selenium sites in the asymmetric unit) can be solved by CF with the same efficiency but faster and with less effort than the use of Patterson techniques (Burla *et al.*, 2007), direct methods-based dual-space algorithms (Weeks & Miller, 1999; Schneider & Sheldrick, 2002; Xu *et al.*, 2005) or hybrid methods (Grosse-Kunstleve & Adams, 2003). In the case of a complex anomalous substructure, here up to 120 selenium sites in the asymmetric unit, the charge-flipping method may be significantly more efficient than traditional methods. This may noticeably increase the complexity at which protein structures can be resolved in practical protein crystallography using SAD or MAD phasing strategies.

2. Computational details

A flowchart of the steps involved in the charge-flipping procedure is shown in Fig. 1. The first steps of the charge-flipping procedure are the expansion of the amplitude coefficients, averaged according to their correct Laue symmetry, to space group $P1$ with $|F_{000}|$ set to zero and the assignment of random phases to all amplitudes, thus respecting Friedel's law. It is noted that this use of *a priori* known symmetry information is noncritical for the charge-flipping process and that it primarily serves to slightly improve the quality of the resulting electron-density map. The charge-flipping algorithm is also insensitive to the absence or presence of reflections which are extinct according to the space-group symmetry. The correct space-group symmetry can be established after the structure-solution step, instead of before as is traditionally performed. The initial map with zero average density obtained by the Fourier transform is modified by flipping the sign of any voxel density that falls below the positive $k_{\rm ed}\sigma$ threshold. The Fourier coefficients of this perturbed map are computed and the set of phases are taken together with the experimental amplitudes (except for the w percent weakest amplitudes) to generate a new set of structure factors. This cyclical process is repeated until convergence is detected or a maximum tolerated number of cycles is reached. The critical input parameter $k_{\rm ed}$ controls the flipping threshold of the electron-density map. Its optimal value was estimated from a series of initial screening trials in the range 1.0–1.5 with increments of 0.05. These calibration tests were performed with the 2h8t and 2anv data sets used for *ab initio* phasing and with the P4 protein from bacteriophage φ13 data set used for substructure deter-

mination. The default $P1$ map grid sampling was adjusted to $d_{min}/2$ and optimized for prime-number factorization in the *FFT* algorithm. The second fundamental parameter controlling the algorithm in reciprocal space is the weak threshold w that defines the fraction of reflections considered to be weak (Oszlányi & Sütő, 2005). The phases of these weak reflections were shifted by $\pi/2$ and calculated moduli kept rather than reassigned to observed moduli. This parameter was optimized in the same way as k_{ed} in the range 0.0–0.75.

During each cycle of charge flipping, two parameters were monitored: a classical residual R-factor value, where the observed amplitude moduli are compared with the calculated moduli, and a relative skewness coefficient of the density map, which corresponds to a scaled third moment of the density distribution. The procedure was repeated until convergence or a prescribed number of cycles had been reached, typically 50 000–100 000. The suitable convergence criterion is defined by the sharp increase of the relative skewness of the electron-density distribution. This feature is accompanied by a modest drop in R and is followed by a plateau region. An additional 200–500 cycles were added to ensure solution stabilization. Finally, 50–100 cycles of LDE (low-density elimination) density modification (Refaat & Woolfson, 1993) were applied to the CF map in order to incorporate all experimental moduli constraints and to refine the phases.

The correct origin of the randomly shifted density in the $P1$ unit cell was restored with respect to the symmetry elements of the appropriate space group before averaging over symmetry-equivalent grid points. A reference map from a previous successful trial can be used to fix a unique origin and thus allow map averaging between different trials. A secondary parameter, Φ_{sym}, based on the phase equivalence of symmetry-related reflections was used to assess the correctness of a solution. Φ_{sym} is defined as the normalized averaged weighted square phase error between symmetry-equivalent structure-factor phases, where the normalization is performed with respect to the phase errors of a completely random structure (Palatinus & van der Lee, 2008). A slightly different approach has been proposed by Burla *et al.* (2000). It is noted that Φ_{sym} will only be useful if the space-group symmetry is correct; a high Φ_{sym} may indicate either a wrong solution, a wrong space group or both. On the other hand, a high Φ_{sym} and a high relative skewness combined with an interpretable map in $P1$ may indicate that the space-group symmetry is wrong but the structure solution is correct. An analysis of the electron-density map may then reveal the correct space-group symmetry and Φ_{sym} based on the new symmetry will in general be low. It is noted that in macromolecular crystallography determination of the space group is usually not a problem, so that a high Φ_{sym} indicates a wrong solution and a low Φ_{sym} a correct solution. The quality of CF solutions was assessed by calculating, using the program *OVERLAPMAP* (Collaborative Computational Project, Number 4, 1994), the correlation coefficient between the CF density map and the reference density map computed from the refined model. In some cases, automated tracing of the charge-flipping density map and model building was also performed with the program *ARP*/*wARP* (Perrakis *et al.*, 1999).

All CF calculations were performed using the *SUPERFLIP* program (Palatinus & Chapuis, 2007) specially optimized for protein crystallography requirements. Each trial was executed with a single processor on an AMD Opteron (2.2 GHz) cluster operating up to 20 independent trials in parallel. The modifications include loop reorganization and Fortran code optimization to speed up calculations for large density and structure-factor arrays and the implementation of a standardized *X-PLOR* format for density maps. The Fourier transform calculations were performed with the FFTW library (http://www.fftw.org) and represent about 30–50% of the total run time depending on map array size and number of reflections. The new version of *SUPERFLIP* incorporating these specific modifications for macromolecular crystallography with tutorials, scripts and test data will be made available for download at http://superspace.epfl.ch/superflip and http://abcis.cbs.cnrs.fr/crystal.

For *ab initio* phasing tests, we selected six X-ray diffraction data sets collected at atomic resolution ($d_{min} \simeq 1.0$ Å) from

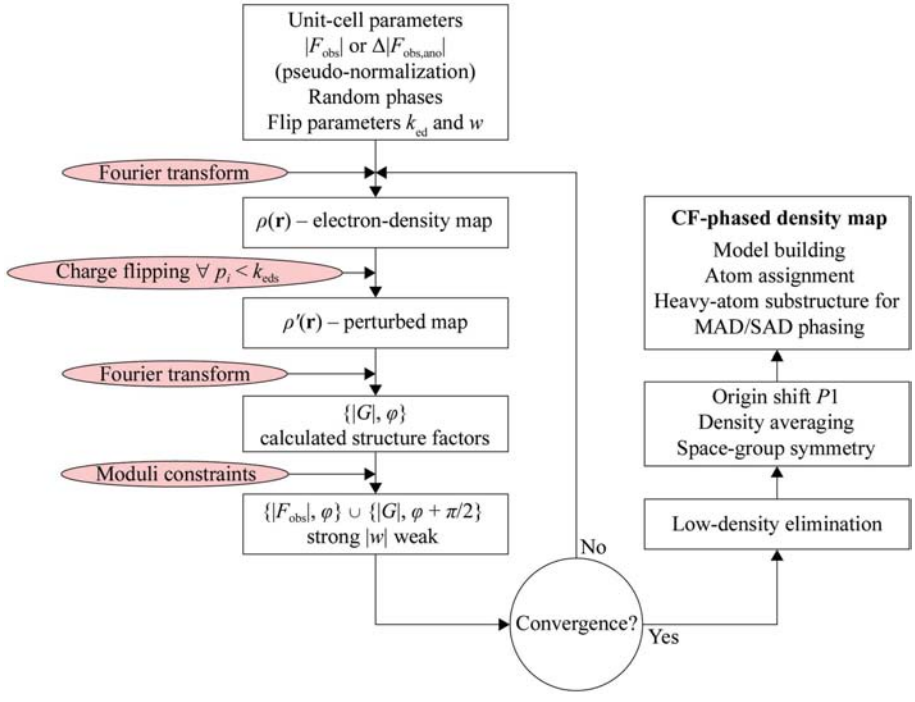


Figure 1
Flow diagram showing the application of the charge-flipping algorithm and the different steps involved in macromolecular crystallography: *ab initio* phasing at atomic resolution and heavy-atom substructure determination.

macromolecular crystals with ∼400–6000 non-H atoms in the asymmetric unit. Both refined coordinates and experimental data are available from the Protein Data Bank for dimeric apamin peptide (PDB code 2h8t; Le-Nguyen *et al.*, 2007), Zn-superoxide dismutase (1mfm; Ferraroni *et al.*, 1999), DNA tetraplex (352d; Phillips *et al.*, 1997), bacteriophage P22 lysozyme (2anv; Mooers & Matthews, 2006), carbon monoxide dehydrogenase (1su8; Dobbek *et al.*, 2004) and alcohol dehydrogenase (2jhf; Meijers *et al.*, 2007). The experimental structure-factor magnitudes were locally normalized in resolution shells using the *ECALC* program (Collaborative Computational Project, Number 4, 1994). The common global normalization *via* the Wilson plot and chemical composition available in the *SUPERFLIP* program (Palatinus & Chapuis, 2007) was also tested. Some data sets were truncated to different resolutions up to 1.4 Å in order to test the effectiveness of CF toward the Sheldrick limit (Sheldrick, 1990; Morris & Bricogne, 2003). In some cases, the experimental diffraction data were extended to higher resolution with simulated data (1su8, from 1.1 to 0.95 Å) or to improve the data completeness (352d, 2jhf). These amplitudes were calculated from the atomic coordinates using the *SFALL* program (Collaborative Computational Project, Number 4, 1994) and appropriate noise (25% of the average amplitude) was added to provide a realistic data set.

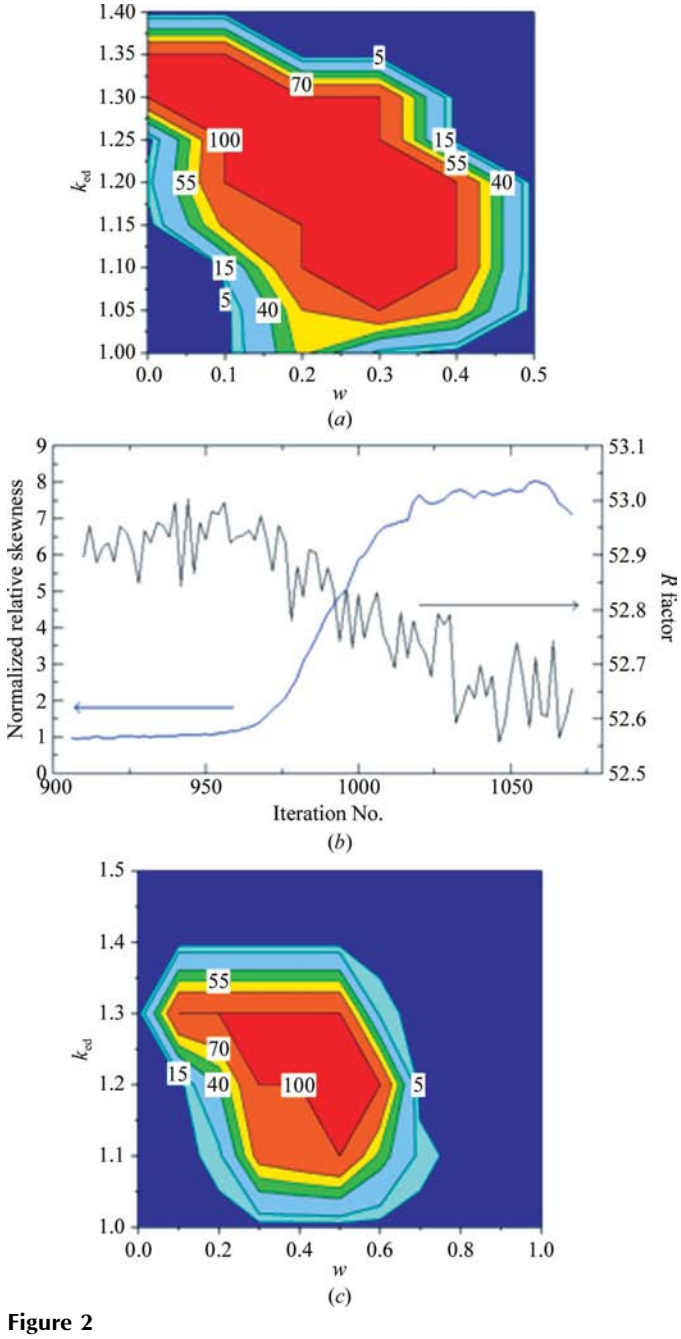


Figure 2
Success rate (%) of *ab initio* atomic structure and substructure solution by charge flipping as a function of the scalar parameters k_{ed} and w. Ten trials were used for each sampling point, with increments of 0.05. (*a*) Diagram corresponding to data for P22 bacteriophage lysozyme using E_{pseudo} normalized structure factors at 1.04 Å resolution. (*b*) A representative successful run of the charge-flipping algorithm for P22 bacteriophage lysozyme using E_{pseudo} normalized structure factors at 1.04 Å resolution. The agreement R factor and the normalized skewness of the density histogram are drawn as a function of dual-space iterations and indicate clear convergence. The CF parameters were k_{ed} = 1.3 and w = 0.05. (*c*) Diagram corresponding to anomalous difference data at 4 Å resolution for P4 packaging enzyme from bacteriophage φ13.

For substructure phasing, the charge-flipping procedure was applied to a set of MAD and SAD diffraction data from the *Autostruct* web site (http://www.ccp4.ac.uk/autostruct/testdata) corresponding to PDB codes 1dw9 (Walsh *et al.*, 2000), 1c8u (Li *et al.*, 2000), 1jc4 (McCarthy *et al.*, 2001) and 1fj2 (Devedjiev *et al.*, 2000). The highly complex substructure of the P4 protein from bacteriophage φ13 (Meier *et al.*, 2005) was also used as a test of CF phasing. In these five data sets, the numbers of heavy-atom sites expected in the asymmetric unit were 8, 20–22, 24, 40 and 120, respectively. The difference amplitude from the anomalous signal of SAD data, the peak and inflection-point in MAD data as well as the dispersive difference values at various wavelengths (inflection point, peak, high-energy remote) were also processed with *ECALC* to obtain pseudo-normalized amplitudes. The maximum Bragg spacing was set to 15 Å and the minimum was varied from 6 to 2 Å in order to investigate the effect of truncating the data at different high-resolution limits. For SeMet and halide-ion site analysis, the program *PEAKMAX* (Collaborative Computational Project, Number 4, 1994) was used to extract the putative heavy-atom site coordinates in the asymmetric unit of the CF map and to sort them according to their relative peak height. A sharp drop in the peak height after the last true site is the best criterion for a significant number of sites. In the case of the P4 substructure, the refined protein model was not available. The Se-site coordinates extracted from *SUPERFLIP* and *SnB* were thus compared with the corresponding site coordinates refined using the program *SHARP* (de La Fortelle & Bricogne, 1997).

3. Results

3.1. Data normalization

The key to success in applying CF to *ab initio* protein structure phasing is the replacement of the input structure-factor moduli $|F_{obs}|$ with appropriately normalized structure

factors $|E_{\rm norm}|$. Recently, the use of standard normalized E values for CF has been suggested (Wu *et al.*, 2004; Coelho, 2007),

$$|E_{\rm norm}(\mathbf{s})| = |F_{\rm obs}(\mathbf{s})/[\langle |F_{\rm obs}(\mathbf{s})|^2\rangle_{\mathbf{s}}]^{1/2}, \quad (4)$$

where the angle brackets indicate the expectation value, which is usually calculated from the approximate chemical composition and the Debye–Waller and scale parameters determined from a Wilson plot (Hauptmann & Karle, 1953). Here, we also used a simple and optimal approach for handling diffraction data from macromolecular crystals (Blessing *et al.*, 1998) where $\langle |F(\mathbf{s})|^2\rangle_{\mathbf{s}}$ is the spherically averaged structure-factor intensity in resolution shells so that $\langle |E_{\rm pseudo}(\mathbf{s})|^2\rangle_{\mathbf{s}} = 1$. This pseudo-normalization has the advantage that it does not depend of the cell size and chemical information, keeping the CF algorithm truly *ab initio*. The difference in success rate between using $|F_{\rm obs}|$ on the one hand and $|E_{\rm pseudo}|$ or $|E_{\rm norm}|$ on the other is remarkable, but is easily understood since the sharpened peaks in the electron-density map computed from normalized structure factors are concentrated into a smaller number of voxels, leaving large plateaus of low density subjected to charge-flipping perturbations. This pseudo-normalization procedure was selected as the default option, but in some cases both $|E_{\rm pseudo}|$ and $|E_{\rm norm}|$ were tried: no significant differences in success rate or map quality were observed except that a faster convergence was frequently obtained using $|E_{\rm pseudo}|$.

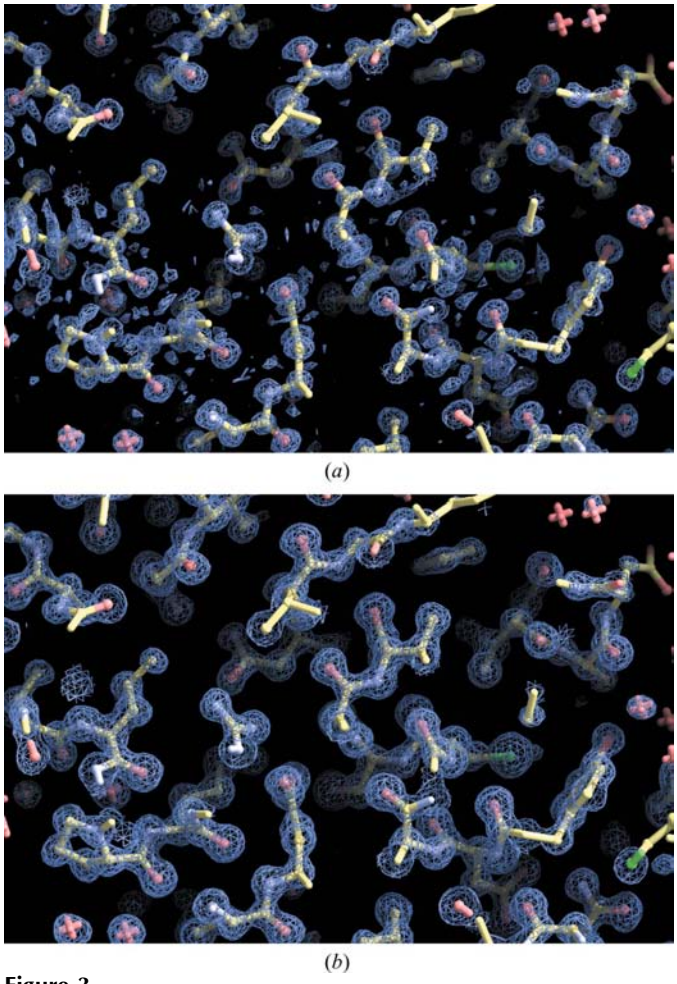


Figure 3
Charge-flipping electron-density map at 1.04 Å resolution of lysozyme from P22 bacteriophage. In (*a*), the map was calculated with E values and phases obtained from the CF algorithm; the blue contour corresponds to the critical CF threshold parameter 1.3σ. In (*b*), the electron-density map was computed with experimental $F_{\rm obs}$ and phases derived from the CF algorithm contoured at 1.3σ. The refined model is superimposed in ball-and-stick representation. This figure was prepared using *Coot* (Emsley & Cowtan, 2004).

3.2. High-resolution *ab initio* phasing of complete structures

In order to provide an evaluation of the power of CF, we selected a medium-sized protein, lysozyme from bacteriophage P22, as a representative model to test the CF algorithm. This crystal structure contains 2385 non-H protein atoms in the asymmetric unit and was one of the largest structures to be solved by direct methods (Mooers & Matthews, 2006). Various combinations of $k_{\rm ed}$ and w were tested in order to delineate the optimal values leading to fast convergence and a high success rate (*i.e.* the percentage of trials that converged to solution; Fig. 2*a*). Under optimal conditions ($k_{\rm ed} = 1.3$, $w = 0.1$), a representative single trial process took approximately 15 min of CPU time, corresponding to 600 CF iterations followed by 50 cycles of polishing by LDE density modification. In comparison, up to 70 000 iterations were required to find a solution with $k_{\rm ed} = 1.1$ and $w = 0.1$, with the success rate being below 10%. The ($k_{\rm ed}$, w) interval with 100% success rate is nevertheless rather large, with w negatively correlated to $k_{\rm ed}$, *i.e.* in order to obtain a 100% success rate for lower w, $k_{\rm ed}$ should be made larger. The results provided useful insights into the behaviour of this algorithm applied to a protein crystal structure with ∼2000 non-H atoms and 50% solvent content. It was found that in the CF process for protein and DNA diffraction data, the skewness of the density histogram, which is related to its third central moment, is an excellent indicator of the phase correctness (Podjarny & Yonath, 1977; Lunin, 1993) and is better than the usual differences between $|G_{\rm obs}|$ and $|G_{\rm calc}|$. These two parameters are calculated at the end of each cycle and a threshold value is defined to detect the convergence of the iterative process.

Usually, three stages in the CF process could be identified: a first initialization stage of about 20 cycles was followed by phase-space exploration before the phase set was attracted to the correct one as indicated by a sharp increase in the relative skewness of the electron-density map and a moderate decrease in the agreement factor R and a final stabilizing plateau (Fig. 2*b*). The quality of the unique solutions was assessed by calculating the correlation coefficient between the averaged density map resulting from the CF process and the reference map, yielding an average value of 0.89. The CF-produced map (Fig. 3) was suitable for automatic building of 288 of the 292 residues of the two molecules using the *ARP*/*wARP* program (Perrakis *et al.*, 1999).

Several other protein or DNA crystal structures were solved by CF from experimental data at atomic resolution (Table 1). Based on the above calibrations, the choice of the critical

Table 1
Results of *ab initio* structure determination at atomic resolution for data from various biological macromolecular crystals.

SG is the crystal space group, d_{min} is the experimental diffraction limit, CC is the correlation coefficient with the reference map computed from the refined model, $\langle\Delta\Phi\rangle$ is the mean phase error for all reflections and Φ_{sym} is the symmetry-agreement score between equivalent structure-factor phases (the normalized averaged weighted square phase error). The values of the k_{ed} and w parameters were 1.30 and 0.1, respectively. $N_{converg}$ is the number of iteration cycles (thousand units) needed to achieve convergence and Success is the average success rate for ten independent trials (up to 30 trials for unsuccessful tests).

Name	PDB code	SG	No. of atoms in ASU (protein/H_2O/heavy atoms)	Completeness (%)	d_{min} (Å)	$N_{converg}$ (1000)/ Success (%)	Φ_{sym}	$\langle\Delta\Phi\rangle$ (°)/ CC
Apamin (Le-Nguyen *et al.*, 2007)	2h8t	$P2_1$	385/70/—	95.5	0.95	2–23/65	2.2	24.2/0.90
					1.1†	—/0	96	—/—
Superoxide dismutase (Ferraroni *et al.*, 1999)	1mfm	$P2_12_12_1$	1152/283/1 Zn, 1 Cu, 9 Cd, 2 Cl	99.2	1.02	1–3/100	3.0	26.1/0.89
				99.4	1.2†	2–5/100	6.4	40.6/0.78
				99.4	1.3†	5–9/100	17	49.2/0.68
				99.4	1.4†	10–12/100	37	56.1/0.55
DNA tetraplex (Phillips *et al.*, 1997)	352d	$P1$	1907/560/9 Ca	89.4	0.95	—/0	—	—/—
				91.0‡	0.95	1.3–13.0/100		33.1/0.82
				94.0‡	0.95	2.0–7.5/100		29.9/0.84
				99.9‡	0.95	0.9–3.1/100		24.4/0.90
				99.9‡	1.1†	—/0		—/—
Lysozyme (Mooers & Matthews, 2006)	2anv	$C2$	2385/517/6 Sm, 8 I	99.4	1.04	0.45–1.0/100	5.3	21.5/0.90
					1.1†	1.0–2.0/100	6.5	22.0/0.90
					1.2†	6–60/80	12	38.0/0.78
					1.3†	—/0	94	—/—
Carbon monoxide dehydrogenase (Dobbek *et al.*, 2004)	1su8	$C2$	4653/1080/6 Ni, 6 Fe	99.8	0.95‡	1.1–2.6/100	1.7	15.3/0.93
					1.1	—/0	95	—/—
Alcohol dehydrogenase (Meijers *et al.*, 2007)	2jhf	$P1$	5866/1241/4 Cd	80.0	1.0	—/ 0	—	—/—
				91.0‡	1.0	1.9–6.0/100		27.3/0.76
				99.0‡	1.0	0.8–1.4/100		15.2/0.92
				99.0‡	1.1†	1.9–6.0/100		19.1/0.90
				99.0‡	1.2†	2.7–13/100		29.1/0.86
				99.0‡	1.3†	—/0		—/—

† Experimental data were truncated in the high-resolution shells. ‡ Experimental data were extended with F_{calc} coefficients calculated from the refined model and 25% simulated noise.

parameters was fixed to the optimal k_{ed} = 1.3 and w = 0.1, typically giving convergence in 500–20 000 cycles with success rates of almost 100%. In all cases, a unique contrasted solution was obtained with a mean absolute phase error for the measured reflections of 15–30° and a correlation coefficient with the reference map of 0.8–0.9. The most remarkable cases were those of carbon monoxide dehydrogenase (PDB code 1su8) containing 4600 non-H protein atoms and alcohol dehydrogenase (2jhf) with 5870 non-H protein atoms, both of which were difficult structures to solve using classical *ab initio* direct methods. Here, they were solved easily at 1.0 Å resolution using default parameters with a 100% success rate and typically 1000–3000 iterations; a single trial required about 5, 8, 10, 12, 15 and 50 min CPU time for 356d, 2jhf, 2h8t, 1mfm, 2anv and 1su8, respectively. This time is significantly shorter compared with the dual-space direct methods that are currently used for *ab initio* phasing.

The resolution and completion-dependence of the CF method have also been analyzed here. A completeness of at least 91–95% is required together with a resolution of around 1.0 Å, below the Sheldrick limit of 1.2 Å. The quality of the CF map is obviously better with higher resolution and more complete data sets. The phasing of the 2jhf structure is illustrative in this respect: at 1.0 Å resolution and with 80% complete data the CF algorithm was not capable of solving the structure, but with 91% complete data the structure, with 5870 non-H atoms in the asymmetric unit (and 1241 water molecules), was solved easily. *ARP*/*wARP* was used to test the quality of the CF map and phases obtained at 1.2, 1.3 and 1.4 Å resolution for the 2jhf and 1mfm proteins; the correlation coefficient and mean phase error were significantly poorer than for the CF map at 1.0 Å resolution (Table 1). In both cases, an almost complete model could easily be built automatically from the primary sequence: 743 amino acids of 748 for 2jhf, and 134 and 120 of 153 for 1mfm at 1.2 and 1.3 Å, respectively. In other regions, the density was improved sufficiently to allow manual model building with *Coot* (Emsley & Cowtan, 2004). The experimental electron-density maps from the charge-flipping method at 1.2 and 1.3 Å resolution were thus suitable for automated model building. The structure of 1mfm also appeared to be solved at 1.4 Å resolution as indicated by the significant increase in the relative skewness score and a moderately low value of Φ_{sym}, but the density map could not be interpreted by *ARP*/*wARP*. In all the tests except peptide apamin (Table 1), the presence of heavy atoms representing 0.5–4% of the total scattering density is actually a necessary condition for the success of the CF algorithm in this atomic resolution range. For instance, we were not capable of solving *ab initio* the structures containing only light atoms and heavy atoms lighter than calcium with PDB codes 2pbv (Declercq *et al.*, 1999), 1mnz (Nowak, Panjikar & Tucker, unpublished results), 1fy4 (Rypniewski *et al.*, 2001) and 1q6z (Bera, Anderson & Hasson, unpublished results), although the almost complete data extended to 0.8–1.0 Å resolution and

Table 2
Heavy-atom substructure determination using the charge-flipping method.

Various amplitude coefficients were used with and without $E_{\rm pseudo}$ normalization: $|\Delta F_{\rm infl\text{-}remot}|$ and $|\Delta F_{\rm ano}{}^{\rm peak}|$ for MAD data sets and $|\Delta F_{\rm ano}|$ for SAD data sets, without selection criteria on $\sigma(|\Delta F|)$ or $|\Delta F|/\sigma(|\Delta F|)$. The values of the $k_{\rm ed}$ and w parameters were 1.30 and 0.35, respectively; the success rate is 100% for ten independent trials. The number of sites found in the *P*1 CF map reduced to the asymmetric unit (ASU) is compared with the number of sites found using *SnB* and *SHELXD* (S), *HySS* (H) or *IL MILIONE* (M). $N_{\rm converg}$ is the number of iteration cycles needed for convergence. The contrast of the peak heights is given in σ units for the highest (H) and lowest (L) significant peaks and the first background peak (B). The average distance of these sites from final deposited PDB structure is given, except for P4-φ13, where the average distance is from the sites refined with *SHARP* at 2.5 Å resolution.

Name/PDB code	SG	$d_{\rm min}$ (Å)	Input coefficients	Final sites in ASU	Sites, S/H/M	Sites, CF	$N_{\rm converg}$ (cycles)	Contrast (H/L/B)	Distance (Å)
Cyanase/1dw9 (Walsh *et al.*, 2000)	*P*1	2.4	$E(\lvert\Delta F_{\rm ano}{}^{\rm peak}\rvert)$	40 Se	38–40 (S), 40 (M)	39	290–350	69.0/10.9/9.0	0.35
		2.4	$\lvert\Delta F_{\rm infl\text{-}remot}\rvert$	—	—	40	340–510	48.0/16.5/8.1	0.33
		2.4	$\lvert\Delta F_{\rm ano}{}^{\rm peak}\rvert$	—	—	39	350–570	69.1/13.0/9.1	0.37
		3.5†	$\lvert\Delta F_{\rm ano}{}^{\rm peak}\rvert$	—	—	38	680–1390	60.8/9.8/8.0	0.42
Thioestase II/1c8u (Li *et al.*, 2000)	$C222_1$	2.5	$E(\lvert\Delta F_{\rm ano}{}^{\rm peak}\rvert)$	8 Se	8 (S), 8 (M), 8 (H)	8	320–590	66.1/48.9/8.7	0.25
		2.5	$\lvert\Delta F_{\rm ano}{}^{\rm peak}\rvert$	—	—	8	180–410	62.1/49.0/8.2	0.24
		2.5	$\lvert\Delta F_{\rm infl\text{-}remot}\rvert$	—	—	8	250–670	57.9/44.1/9.1	0.27
		6.0	$E(\lvert\Delta F_{\rm ano}{}^{\rm peak}\rvert)$	—	—	8	320–1100	18.0/10.3/5.9	0.33
Epimerase/1jc4 (McCarthy *et al.*, 2001)	$P2_1$	2.1	$E(\lvert\Delta F_{\rm ano}\rvert)$	24 Se	24 (S)	24	420–1790	49.3/18.0/12.5	0.16
		2.1	$\lvert\Delta F_{\rm ano}\rvert$	—	—	24	1560–5400	36.2/19.5/14.8	0.21
		3.0†	$\lvert\Delta F_{\rm ano}\rvert$	—	—	24	3410–9160	33.7/15.5/11.1	0.30
P4-φ13 (Meier *et al.*, 2005)	*C*2	2.5	$E(\lvert\Delta F_{\rm ano}\rvert)$	120 Se	96 (S), 114 (M)	120	1260–3680	69.2/13.8/11.9	0.26
		4.0†	$\lvert\Delta F_{\rm ano}\rvert$	—	—	120	1940–4010	67.7/13.9/11.6	0.49
		5.0†	$\lvert\Delta F_{\rm ano}\rvert$	—	—	120	1690–9890	27.8/7.9/6.7	0.73
		6.0†	$E(\lvert\Delta F_{\rm ano}\rvert)$	—	—	116	3100–11230	19.7/6.0/5.6	1.25
		6.0†	$\lvert\Delta F_{\rm ano}\rvert$	—	—	—	—	—	—
Thioesterase/1fj2 (Devedjiev *et al.*, 2000; Dauter *et al.*, 2002)	$P2_1$	1.8	$E(\lvert\Delta F_{\rm ano}\rvert)$	20–22 Br	21–22 (S, H), 22 (M)	22	420–2680	55.3/7.8/7.4	0.24
		1.8	$\lvert\Delta F_{\rm ano}\rvert$	—	—	22	570–1560	50.4/7.6/7.3	0.26
		3.0†	$\lvert\Delta F_{\rm ano}\rvert$	—	—	19	940–3980	34.7/7.9/7.5	0.31

† Experimental data were truncated in the high-resolution shells.

represented 873, 3073, 3747 and 3942 non-H atoms in the asymmetric unit, respectively.

3.3. Medium- and low-resolution phasing of substructures

The second important and attractive extension of this CF methodology for *ab initio* substructure phasing is the use of anomalous or dispersive differences as input coefficients for the iterative CF process. Heavy-atom substructures are usually determined by Patterson or direct methods using isomorphous, dispersive or anomalous difference data or a combination of these. While traditional Patterson methods are limited to about a dozen atoms, the more powerful dual-space direct methods implemented in *SnB*/*BnP* (Smith *et al.*, 1998) and *SHELXD*/*E* (Sheldrick, 2008), the *HySS* hybrid procedure (Grosse-Kunstleve & Adams, 2003) and *IL MILIONE* using an improved Patterson technique (Burla *et al.*, 2007) are successful for complex substructures of up to a few hundred atoms, provided that anomalous difference data have been measured to ∼3.5–4 Å resolution. They often require many trials and the adjustment of numerous parameters, are CPU-time intensive and may give partial or, in some difficult cases, false-positive solutions (Von Delft *et al.*, 2001; Meier *et al.*, 2005). As reported in Table 2, charge-flipping phasing was applied to the data for a series of moderate-size substructures containing 8–40 heavy atoms in the asymmetric unit (cyanase, thioesterase II, epimerase and human thioesterase) and one of the most complex substructures published to date, the packaging enzyme P4 from bacteriophage φ13 (P4-φ13; Meier *et al.*, 2005), which contains as many as 120 Se atoms. The diffraction data were processed according to the CF protocol described in the previous section and the Bijvoet differences from single-wavelength data were used as amplitude coefficients. Various combinations of $k_{\rm ed}$ and w were also explored in order to delineate optimal values (Fig. 2*c*) in this case of the application of CF to substructure determination. It is noted that this ($k_{\rm ed}$, w) diagram presents qualitatively the same characteristics as the ($k_{\rm ed}$, w) diagram for *ab initio* structure determination (Fig. 2*a*).

Table 2 shows that a unique solution for an almost complete list of anomalous scatterers was obtained with a high success rate using both normalized and non-normalized amplitude differences. In most cases, a clear gap is observed between the significant sites and background peaks in terms of peak height. This is particularly true for the P4-φ13 protein crystal, where 480 peaks were clearly identified in the *P*1 working cell (Fig. 4*a*), which represents 120 Se atoms in the asymmetric unit (five in each subunit of the four P4 hexamers). Under the same conditions, the improved Patterson technique was able to localize 114 sites (Burla *et al.*, 2007), while *SnB* and *SHELXD* were able to localize 96 sites in the asymmetric unit, but both were initially trapped in false local minima (Meier *et al.*, 2005). The average distance of the coordinates of the 96 peaks from the correct solution of *SnB* from the corresponding peaks obtained from 2.5 Å resolution CF maps was 0.9 Å. However, in their study Meier and coworkers completed the substructure to 120 sites using anomalous residual Fourier maps calculated using *SHARP*. These 24 additional peaks could be perfectly superimposed with the corresponding sites from the CF-site constellation. Another

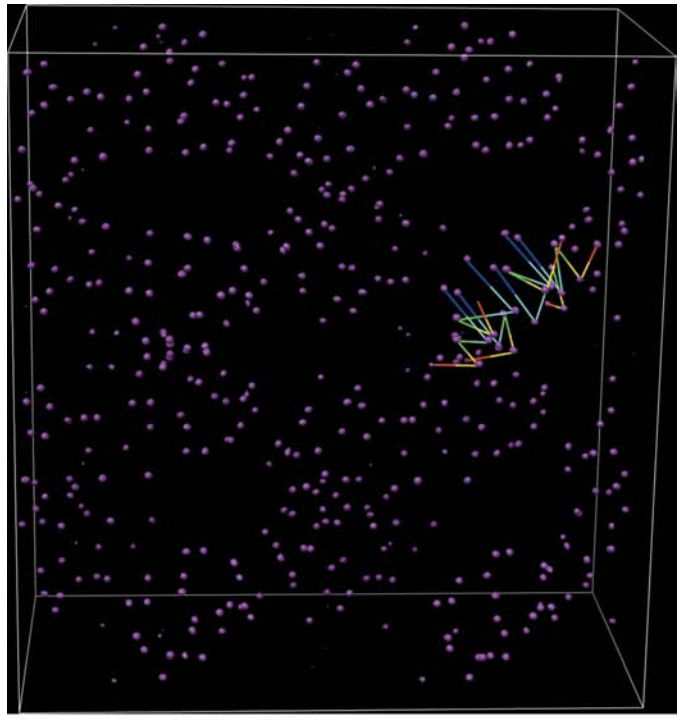
criterion for the correctness of the substructure obtained by *SUPERFLIP* is the presence of clear sixfold noncrystallographic symmetry between the sites (Fig. 4*b*). The 120 Se sites were efficiently refined using the program *SHARP*. The average distance between the 120 selenium positions extracted from the CF map compared with the *SHARP* refined coordinates is 0.26 Å (0.95 Å for the 96 sites obtained with *SnB*) and the refined occupancies highly correlated with the initial peak heights. Interestingly, it was not strictly necessary to use the highest resolution data (2.5 Å), but with data truncated to 6.0 Å the complete substructure was also easily found with an only slightly degraded coordinate precision. The typical CPU times required for a single successful trial at 2.5 and 5 Å are 50 and 3 min, respectively, which are significantly less than the time required by *SnB*, *SHELXD* (300–500 min at 4 Å) or *IL MILIONE* (1100 min at 2.5 Å for five successful trials of 60).

Typically, the coordinate precision of the scatterers in the CF maps is high, within 0.2–0.3 Å of the refined substructures, and the peak heights are in the appropriate order relative to the occupancy and *B* factor of the corresponding atomic sites. Excellent results were obtained from consideration of anomalous differences data alone in a broad range of resolution, here up to 6 Å (P4 and 1c8u), as Se atoms in SeMet-labelled proteins are seldom closer than 4–6 Å to each other.

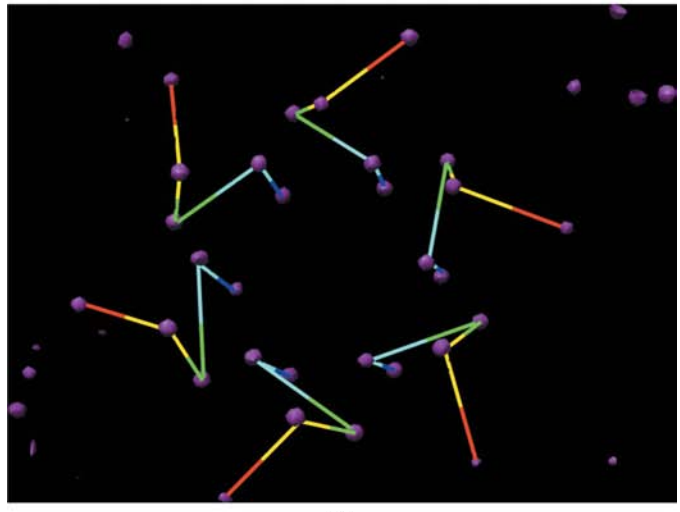
This robust, fast and simple CF technique for *ab initio* substructure determination presents several advantages compared with other conventional methods: (i) *E* normalization is not necessary but is found to slightly increase the convergence rate and improve the success rate at low resolution, (ii) the number of expected sites is not required and most heavy-atom sites are obtained in a single pass, which is in particular illustrated by the substructure determination of human acyl thioesterase, where the number of partially occupied Br sites is *a priori* unknown, and (iii) the CF procedure is fast and particularly insensitive to pseudo-symmetry pitfalls.

4. Discussion

The results presented here demonstrate the feasibility of CF for the solution of structures of biological macromolecules, either proteins or DNA molecules, using experimental data at atomic resolution. This purely computational approach provides a rapid and robust method for phasing complicated structures of up to ∼6000 atoms in the asymmetric unit (2jhf; 8 min CPU time), requiring pseudo-normalization of input structure factors, an almost complete data set and a resolution of at least 1.0 Å. However, Sheldrick's 1.2 Å resolution limit for the applicability of classic direct methods is also reached here in some test data sets (1mfm, 2anv). The results show that this technique is capable of finding unique solutions independently of the starting phases. It only requires very simple parameters, the threshold values $k_{\rm ed}$ in direct space and w in reciprocal space, and is highly parallelizable, starting with different random number seeds on all available CPUs in a cluster. $k_{\rm ed}$ is easily estimated: in all considered cases it was close to 1.3 times the standard deviation of the electron-density map. w is less easy to determine, but it appears that there is a close connection to the quality of the data set, expressed for example by $\langle |G|/\sigma(|G|)\rangle$. A value of 0.05–0.1 was optimal for the high-quality atomic resolution data set tested above, whereas a value of 0.3–0.4 was appropriate for anomalous difference measurements with a low signal-to-noise ratio. Thus, the more time-consuming step of tuning ($k_{\rm ed}$, w) is not required for each test structure. Most importantly, there



(*a*)



(*b*)

Figure 4
Overall view of the *P*1 charge-flipping electron-density map of the selenium substructure of the packaging enzyme P4 of bacteriophage φ13. In (*a*), 480 peaks in the unit cell corresponding to Se atoms from selenomethionine-labelled P4 protein are clearly delineated at a 10σ contour level; 30 Se sites in one hexameric oligomer are connected by coloured lines. In (*b*), an enlarged view of the sixfold-symmetric group of Se density peaks corresponding to one oligomer. The figure was prepared with *CHIMERA* (http://www.cgl.ucsf.edu/chimera).

are clear indications that this valuable tool has not yet reached its limits. Incorporating further constraints such as histogram matching (Baerlocher, Gramm *et al.*, 2007; Baerlocher, McCusker *et al.*, 2007) and molecular-envelope support could further improve the efficiency of the CF method. It is noted that very recently Zhou & Harris (2008) have proposed a charge-flipping variant called 'residue-based charge flipping' (RBCF) which does not use any parameter at all, neither k_{ed} nor w, and which is claimed to give faster convergence than the charge-flipping algorithm as developed by Oszlányi and Sütő. However, the authors only applied RBCF to three small-molecule structures and it remains to be tested whether the method is also successful for macromolecular structures.

The most important limitation of CF for *ab initio* phasing of macromolecular-sized molecules appears to be the resolution, which in the present tests was mostly in the range 0.95–1.2 Å, thus in most cases remaining somewhat above Sheldrick's 1.2 Å rule (Sheldrick, 1990), which states that below this resolution *ab initio* methods cannot be used for phasing macromolecules. It is known that the presence of artificially introduced anomalous scatterers or prosthetic heavy-atom clusters may relax this condition somewhat (Caliandro *et al.*, 2008). Mathematical phase-extension procedures based on model-free methods such as Patterson-MEM calculations (Palatinus *et al.*, 2007), electron-density peak deconvolution (Altomare *et al.*, 2008) or other artificial methods combined with density-modification cycles (Caliandro *et al.*, 2005; Jiaxing *et al.*, 2005) may be used to increase the experimental resolution. This may lead to a significant improvement of the experimental lower resolution limit at which protein-sized structures can be solved *ab initio*: Caliandro *et al.* (2008) were able to solve structures up to 2.0 Å resolution in this way and the largest heavy-atom-containing protein (7890 non-H atoms in the asymmetric unit) at 1.65 Å resolution. Still another method used in the *ACORN* phasing procedure (Foadi *et al.*, 2000) consists of incorporating idealized α-helices or structural motifs as a starting point.

Another limitation of CF which is certainly more severe than for direct methods is that the data need to be nearly complete. The aforementioned extension methods could, however, also be used in these cases for filling in the holes in the data set. The relative amount of data available for substructure determinations is significantly larger than for atomic complete structures with a comparable number of atoms to be positioned. Consequently, the completeness criteria can be significantly relaxed in this case. However, other limitations could hamper the optimum efficiency of the CF method: weak anomalous or dispersive signals in SAD or MAD data sets or low-resolution diffraction data, typically below 5–6 Å. Advances in data-collection strategies and the upgrading of synchrotron beamlines may in the near future result in more complete data sets that are measured to a higher resolution, so that these relative drawbacks of charge flipping compared with direct methods may disappear. The most important result of our study is that charge flipping is now emerging as an efficient and fast method for the determination of high-quality substructures, especially for complex structures, that can be incorporated into the panoply of tools for automatic phasing procedures.

5. Conclusions

It is clear that charge flipping as a phasing tool for macromolecular structures is still at the beginning of its development and that many extensions of its performance can be expected in the near feature. It has been shown in this paper that even in its infancy charge flipping is already capable of solving *ab initio* complex protein structures at atomic resolution which have only become tractable in recent years by classical direct methods combined with the *Shake-and-Bake* algorithm. Astonishingly, this is accomplished using a much easier procedure than used in direct methods, which makes it easier to fine-tune the algorithmic key parameters. The actual limitations of the method are the requirement for complete diffraction data at atomic resolution around ∼1 Å and the presence of a few heavy atoms. Under these conditions, charge flipping can be considered, in view of its simplicity and efficiency, to be a good alternative to direct methods. The CF algorithm can also position the heavy atoms of the substructure itself using the anomalous or dispersive differences generated from these sites. The above test cases show that simple to complex substructure determinations can be performed automatically with a high efficiency at resolutions typical for macromolecular crystal structures (up to ∼5–6 Å). It is difficult to find more complex substructure cases than that of the packaging enzyme P4 from bacteriophage φ13, so that the full potential of the method remains to be explored at this time. Further work now needs to be undertaken to explore the possibility of solving medium-resolution substructures containing nearly a thousand heavy-atom or anomalous scatterers. This provides the potential to phase very large macromolecular assemblies (10–100 MDa) without using the bootstrapping procedures of low-resolution phasing with cryo-electron microscopy models (Hanein, 2007) or labelling with multi-metal clusters (Abrahams & Ban, 2003). Crystallographic studies at atomic detail of supramolecular assemblies are now a major goal in structural biology (Harrison, 2004), especially low-symmetry or unsymmetrical structures (ribosomes, nuclear pore complex *etc.*). Extension of selenium labelling of large complex structures is now a conceivable strategy and improved protocols for labelling proteins or nucleic acids within large macromolecular assemblies are now available (Jiang *et al.*, 2007; Kivelä *et al.*, 2008).

Lukáš Palatinus (EPFL, Lausanne) is thanked for freely distributing the source code of *SUPERFLIP*. We are grateful to Jonathan Grimes (STRUBI, Oxford) for putting at our disposal the experimental data of the P4-φ13 protein crystal. We thank Gérard Bricogne (Cambridge) for helpful comments concerning this manuscript.

References

Abrahams, J. P. & Ban, N. (2003). *Methods Enzymol.* **374**, 163–188.

Altomare, A., Cuocci, C., Giacovazzo, C., Kamel, G. S., Moliterni, A. & Rizzi, R. (2008). *Acta Cryst.* A**64**, 326–336.

Baerlocher, C., Gramm, F., Massüger, L., McCusker, L. B., He, Z. B., Hovmöller, S. & Zou, X. D. (2007). *Science*, **315**, 1113–1116.

Baerlocher, C., McCusker, L. & Palatinus, L. (2007). *Z. Kristallogr.* **222**, 47–53.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1998). *Direct Methods for Solving Macromolecular Structure*, edited by S. Fortier, pp. 47–71. Dordrecht: Kluwer Academic Publishers.

Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2006). *J. Appl. Cryst.* **39**, 728–734.

Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2007). *J. Appl. Cryst.* **40**, 211–217.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2000). *J. Appl. Cryst.* **33**, 307–311.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2008). *J. Appl. Cryst.* **41**, 548–553.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005). *Acta Cryst.* D**61**, 556–565.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2007). *J. Appl. Cryst.* **40**, 883–890.

Coelho, A. A. (2007). *Acta Cryst.* A**63**, 400–406.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Dauter, Z. (2002). *Curr. Opin. Struct. Biol.* **12**, 674–678.

Dauter, Z., Dauter, M. & Dodson, E. J. (2002). *Acta Cryst.* D**58**, 494–506.

Declercq, J. P., Evrard, C., Lamzin, V. & Parello, J. (1999). *Protein Sci.* **8**, 2194–2204.

Devedjiev, Y., Dauter, Z., Kuznetsov, S., Jones, T. & Derewenda, Z. (2000). *Structure*, **8**, 1137–1146.

Dobbek, H., Svetlitchnyi, V., Liss, J. & Meyer, O. (2004). *J. Am. Chem. Soc.* **126**, 5382–5387.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Ferraroni, M., Rypniewski, W., Wilson, K. S., Viezzoli, M. S., Banci, L., Bertini, I. & Mangani, S. (1999). *J. Mol. Biol.* **288**, 413–426.

Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.

Gerchberg, R. W. & Saxton, W. O. (1972). *Optik*, **35**, 237–246.

Gies, H. (2007). *Science*, **315**, 1087–1088.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* D**59**, 1966–1973.

Hanein, D. (2007). *J. Struct. Biol.* **158**, 135–136.

Harrison, S. C. (2004). *Nature Struct. Biol.* **11**, 12–15.

Hauptmann, H. & Karle, J. (1953). *Solution of the Phase Problem*, ACA Monograph No. 3. Wilmington, USA: Polycrystal Book Service.

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *EMBO J.* **9**, 1665–1672.

Jiang, J., Sheng, J., Carrasco, N. & Huang, Z. (2007). *Nucleic Acids Res.* **35**, 477–485.

Jia-xing, Y., Woolfson, M. M., Wilson, K. S. & Dodson, E. J. (2005). *Acta Cryst.* D**61**, 1465–1475.

Katrych, S., Weber, Th., Kobas, M., Massüger, L., Palatinus, L., Chapuis, G. & Steurer, W. (2007). *J. Alloys Comput.* **428**, 164–172.

Kivelä, H. M., Abrescia, N. G. A., Bamford, J. K. H, Grimes, J. M., Stuart, D. I. & Bamford, D. H. (2008). *J. Struct. Biol.* **161**, 204–210.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Le-Nguyen, D., Chiche, L, Hoh, F., Martin-Eauclaire, M. F., Dumas, C., Nishi, Y., Kobayashi, Y. & Aumelas, A. (2007). *Biopolymers*, **86**, 447–462.

Li, J., Derewenda, U., Dauter, Z., Smith, S. & Derewenda, Z. S. (2000). *Nature Struct. Biol.* **7**, 555–559.

Lunin, V. Yu. (1993). *Acta Cryst.* D**49**, 90–99.

McCarthy, A. A., Baker, H. M., Shewry, S. C., Patchett, M. L. & Baker, E. N. (2001). *Structure*, **9**, 637–646.

Meier, C., Mancini, E. J., Bamford, D. H., Walsh, M. A., Stuart, D. I. & Grimes, J. M. (2005). *Acta Cryst.* D**61**, 1238–1244.

Meijers, R., Hans-Werner, A., Dauter, Z., Wilson, K. S., Lamzin, V. S. & Cedergren-Zeppezauer, E. S. (2007). *Biochemistry*, **46**, 5446–5454.

Mooers, B. H. M. & Matthews, B. W. (2006). *Acta Cryst.* D**62**, 165–176.

Morris, R. J. & Bricogne, G. (2003). *Acta Cryst.* D**59**, 615–617.

Oszlányi, G. & Sütő, A. (2004). *Acta Cryst.* A**60**, 134–141.

Oszlányi, G. & Sütő, A. (2005). *Acta Cryst.* A**61**, 147–152.

Oszlányi, G. & Sütő, A. (2008). *Acta Cryst.* A**64**, 123–134.

Palatinus, L. (2004). *Acta Cryst.* A**60**, 604–610.

Palatinus, L. & Chapuis, G. (2007). *J. Appl. Cryst.* **40**, 786–790.

Palatinus, L., Steurer, W. & Chapuis, G. (2007). *J. Appl. Cryst.* **40**, 456–462.

Palatinus, L. & van der Lee, A. (2008). In preparation.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Phillips, K., Dauter, Z., Murchie, A. I. H., Lilley, D. M. J. & Luisi, B. (1997). *J. Mol. Biol.* **273**, 171–182.

Podjarny, A. D. & Yonath, A. (1977). *Acta Cryst.* A**33**, 655–661.

Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.

Rypniewski, W., Østergaard, P., Nørregaard-Madsen, M., Dauter, M. & Wilson, K. S. (2001). *Acta Cryst.* D**57**, 8–19.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.

Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. (1998). *Acta Cryst.* D**54**, 799–804.

Von Delft, F., Lewendon, A., Dhanaraj, V., Blundell, T., Abell, C. & Smith, A. (2001). *Structure*, **9**, 439–450.

Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (2000). *Structure*, **8**, 505–514.

Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* D**55**, 492–500.

Wu, J. S., Spence, J. C. H., O'Keeffe, M. & Groy, T. L. (2004). *Acta Cryst.* A**60**, 326–330.

Xu, H., Weeks, C. M. & Hauptman, H. A. (2005). *Acta Cryst.* D**61**, 976–981.

Zhou, Z. & Harris, K. D. M. (2008). *J. Phys. Chem. A*, **112**, 4863–4868.