Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Analysis of experimental time-resolved crystallographic data by singular value decomposition

Sudarshan Rajagopal,[a] Marius Schmidt,[b] Spencer Anderson,[c] Hyotcherl Ihee[d] and Keith Moffat[a,c,e]*

[a]Department of Biochemistry and Molecular Biology, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA, [b]Physik-Department E17, Technische Universität Munchen, Garching, Germany, [c]Center for Advanced Radiation Sources, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA, [d]Department of Chemistry and School of Molecular Science (BK21), Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea, and [e]Institute for Biophysical Dynamics, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA

Correspondence e-mail: moffat@cars.uchicago.edu

Received 22 October 2003
Accepted 22 February 2004

Singular value decomposition (SVD) separates time-dependent crystallographic data into time-independent and time-dependent components. Procedures for the effective application of SVD to time-resolved macromolecular crystallographic data have yet to be explored systematically. Here, the applicability of SVD to experimental crystallographic data is tested by analyzing 30 time-resolved Laue data sets spanning a time range of nanoseconds to milliseconds through the photocycle of the E46Q mutant of photoactive yellow protein. The data contain random and substantial systematic errors, the latter largely arising from crystal-to-crystal variation. The signal-to-noise ratio of weighted difference electron-density maps is significantly improved by the SVD flattening procedure. Application of SVD to these flattened maps spreads the signal across many of the 30 singular vectors, but a rotation of the vectors partitions the large majority of the signal into only five singular vectors. Fitting the time-dependent vectors to a sum of simple exponentials suggests that a chemical kinetic mechanism can describe the time-dependent structural data. Procedures for the effective SVD analysis of experimental time-resolved crystallographic data have been established and emphasize the necessity for minimizing systematic errors by modification of the data-collection protocol.

1. Introduction

Determination of the structures of short-lived intermediates by crystallography is usually achieved by increasing the lifetime of the chemical species by chemical or physical trapping (Moffat & Henderson, 1995) or by improving the time-resolution of the X-ray experiment as in time-resolved Laue crystallography (Moffat, 1989). Fast time-resolution as low as 150 ps (Schotte *et al.*, 2003) allows extremely short-lived species to be visualized. While technically challenging, a number of time-resolved Laue studies have been performed (reviewed in Ren *et al.*, 1999), with the most extensive studies performed on the photocycle of photoactive yellow protein (PYP; Genick *et al.*, 1997; Perman *et al.*, 1998; Ren *et al.*, 2001; Anderson *et al.*, 2004) and the photolysis of the carbon monoxide–myoglobin complex (Šrajer *et al.*, 1996, 2001; Bourgeois *et al.*, 2003; Schotte *et al.*, 2003). It is now possible in a matter of days to collect tens of individual Laue data sets in which each data set has high completeness and redundancy and the sets together span the time range of interest after reaction initiation. Processing such comprehensive data sets requires up to a few weeks (Ren *et al.*, 1999; Bourgeois *et al.*, 2000), resulting in difference electron-density maps across the time range. There remains a major difficulty in the analysis of these maps: the difference electron density associated with

each time point in this range is a superposition of the difference electron densities associated with all structures present in the crystal at that time. For the potential of time-resolved crystallography to be fully realised, separation of the difference electron density associated with a superposition of structures into the difference electron density associated with each structure of a homogeneous time-independent intermediate is essential (Moffat, 1989; Schlichting & Chu, 2000).

Analysis of difference electron density can proceed in a number of ways. Firstly and most simply, the density can be assessed qualitatively (Ren *et al.*, 2001; Šrajer *et al.*, 2001; Schotte *et al.*, 2003). In this strategy, positive and negative difference features are directly associated with the movements of nearby atoms. In the myoglobin example, Šrajer *et al.* (2001) and Schotte *et al.* (2003) associated particular difference electron-density features across many time points with the migration of carbon monoxide after its photolysis from the heme and with the relaxation of the heme and the surrounding protein environment. In PYP, Ren *et al.* (2001) interpreted difference features from nanoseconds to microseconds in the chromophore-binding pocket as changes in the conformation of the 4-hydroxycinnamic acid chromophore. In these studies, the interpretation of the difference electron density and the conclusions were primarily qualitative in nature.

In order to refine the structures of intermediates, more quantitative methodologies must be used. One approach is to collect data at a time point or over a short time range when only one structural species is thought to exist based on prior experiments using, for example, visible absorption spectroscopy. Genick *et al.* (1997) refined a millisecond intermediate of PYP using this approach, as did Perman *et al.* (1998) for a nanosecond intermediate. However, the latter refinement was likely to have been in error owing to the poor signal-to-noise ratio (S/N) of their data set, which was collected at a single 10 ns time delay (Ren *et al.*, 2001). Bourgeois *et al.* (2003) refined a structure against data from a 316 ns time delay after photolysis of carbon monoxide from a myoglobin mutant, but noted that a semi-quantitative analysis of the time-dependent difference electron density showed the presence of multiple intermediates, one in the nanosecond range and another in the microsecond range. If no major density differences are noted between data sets spanning a wider time range, averaging of maps closely spaced in time provides a major improvement in S/N. Anderson *et al.* (2004) took advantage of this in their analysis of 30 data sets obtained during the photocycle of the E46Q mutant of PYP (E46Q PYP) to identify and refine two structures corresponding to the early red-shifted and late blue-shifted spectroscopic intermediates. The assumption in all cases is that only one structure is present in the crystal; if there are in fact multiple structures present, this strategy cannot be used effectively.

To overcome this limitation, it is necessary to include an analysis of the time domain with the refinement of structures, *i.e.* an explicit determination of the chemical kinetic mechanism. Owing to the low S/N present in typical time-resolved crystallographic data, the method of singular value decomposition, SVD (reviewed by Henry & Hofrichter, 1992) is especially applicable (Schmidt *et al.*, 2003). SVD acts as a noise filter in which signal and noise are partitioned into different singular vectors. SVD also provides a reduced representation of the data which greatly simplifies subsequent least-squares fits to structural models. The feasibility of SVD has been demonstrated on simulated time-resolved crystallographic data containing systematic and random errors at various levels (Schmidt *et al.*, 2003). But is SVD applicable to real experimental data? Schmidt *et al.* (2004) have recently applied SVD to Laue data sets of wild-type PYP spanning a more limited microsecond to millisecond time range. They refined two late structures and identified chemical kinetic mechanisms consistent with the crystallographic data. However, it is still unclear what hurdles SVD analysis of experimental time-resolved crystallographic data must overcome over long time ranges and how it should be optimally applied.

Here, we investigate many of the technical issues associated with such an analysis by applying SVD to the E46Q PYP data collected by Anderson *et al.* (2004). Data collected over a wide time range (10 ns to 100 ms), with varying quality in its 30 time points and the possibility of identifying a high number of distinct structures, presents challenges not faced in the previous SVD analysis of experimental data (Schmidt *et al.*, 2004) nor fully assessed in the simulated data (Schmidt *et al.*, 2003). To minimize bias, we present a number of measures that can be used to assess the progress of the SVD procedure, allowing us to interpret time-resolved crystallographic data more objectively. In a companion paper, we show that determination of the chemical kinetic mechanism from these experimental time-resolved crystallographic data is feasible (Rajagopal *et al.*, 2004).

2. Overview of SVD analysis

2.1. Singular value decomposition

For a detailed discussion of this procedure, see Henry & Hofrichter (1992), who review the mathematical aspects of SVD and its application to spectroscopic data, and Schmidt *et al.* (2003), who apply SVD to simulated time-resolved crystallographic data with different S/N levels. SVD decomposes time-dependent data from a data matrix $\mathbf{A}$ into three matrices according to the equation

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{T},$$

where $\mathbf{U}$ is composed of the left singular vectors (lSVs), $\mathbf{S}$ is composed of the singular values and $\mathbf{V}$ is composed of the right singular vectors (rSVs). In an analysis of time-resolved crystallographic data, the lSVs are an orthonormal basis for time-independent difference electron density, the singular values are weighting factors that describe how much the corresponding SVs contribute to the data in a least-squares sense and the rSVs are an orthonormal set that describe the time-dependence of their corresponding lSVs. Thus, an individual lSV is a linear combination of the time-independent difference electron densities associated with each inter-

mediate and an individual rSV is a linear combination of the time-dependent concentrations of each intermediate. SVD has two very useful properties: it acts as a noise filter and presents a reduced representation of the data matrix. The first property allows the data matrix to be reconstructed with a subset of singular vectors (SVs) that contain the majority of signal, discarding those SVs that contain primarily noise. In essence, this property exploits the fact that authentic signal varies smoothly with time, but noise (particularly systematic error) may vary sharply from time point to time point. Furthermore, it allows the derivation of explicit phase information for the difference structure factor. The second property allows a much simpler fit to a chemical kinetic mechanism of the rSVs by identifying the minimum number of structurally distinct components in the data and reducing the dimensionality of the fit considerably (Henry & Hofrichter, 1992). These features are especially well suited to the analysis of crystallographic data (Schmidt *et al.*, 2003).

2.2. Outline of the procedure

SVD analysis of time-resolved crystallographic data can be split into three major stages: data preparation, data evaluation and the determination of the chemical kinetic mechanism. The next three sections of this paper (§§3–5) deal with these procedures in turn. Data preparation includes all steps between experimental data collection and the analysis of the data by SVD. In data collection, first a pulse of polychromatic X-rays illuminates the crystal sample at a set time delay after excitation with a nanosecond laser. One or a few time delays are usually collected from a single crystal; at a later stage, data sets corresponding to the same time delays on different crystals may be merged to yield an average data set for that time point. After data collection and processing and scaling of the data, weighted difference electron-density maps are generated from the difference structure-factor amplitudes $[\Delta F = |\mathbf{F}(t)| - |\mathbf{F}(0)|]$, the errors associated with them ($\sigma_{\Delta F}$) and the phases of the dark-state model (φ). These maps may be subjected to SVD flattening, which increases their S/N by generating $\Delta F_{\rm new}$ and $\varphi_{\rm new}$ (Schmidt *et al.*, 2003). After adjusting individual data sets for crystal-to-crystal variation in the extent of photoinitiation, outliers in the time course may still need to be corrected relative to the first rSV (Schmidt *et al.*, 2003). After these steps, the data matrix is in a form suitable for data evaluation.

In data evaluation, the data matrix is first subjected to SVD, generating a set of lSVs, rSVs and singular values. Those SVs that may contain signal are then identified and subjected to rotation, which tends to partition signal into rSVs with high autocorrelation coefficients (a measure of a function's smoothness) and random noise into rSVs with low autocorrelation. After rotation, a subset of the SVs must be chosen to reconstruct the data matrix $\mathbf{A}' \simeq \mathbf{A}$. The number of significant SVs gives a lower bound on the number of states present in the system (Henry & Hofrichter, 1992). As the lSVs and rSVs are linear combinations of the structures and concentrations of the underlying species, respectively, the rSVs can be fit with a sum of exponentials (Henry & Hofrichter, 1992). If this fit is successful, the number of exponentials corresponds to the number of intermediate states present in the system and the exponents are relaxation times. By combining SVD analysis with information available from other biophysical approaches such as time-resolved spectroscopy, plausible chemical kinetic mechanisms can be fit to the rSVs and time-independent difference electron density corresponding to the underlying chemical species can be calculated (Schmidt *et al.*, 2003).

3. Data preparation

3.1. Experimental data collection

Anderson (2003) and Anderson *et al.* (2004) provide details of the experimental data collection. 54 data sets corresponding to 30 different time points from 10 ns to 100 ms were collected on the E46Q PYP at beamline 14-ID-B of the Advanced Photon Source and beamline ID09B of the European Synchrotron Radiation Facility. A total of 25 different crystals was used. Light and dark data sets were collected from the same crystal, interleaved to minimize the effect of laser and X-ray radiation damage on the difference structure factors. These light ($|F_{\rm L}|$) and dark ($|F_{\rm D}|$) data sets were processed in *LaueView* (Ren & Moffat, 1995*a*,*b*), scaled and redundant data sets were merged in reciprocal space with weighted averaging. Difference structures factors (ΔF) were calculated from ($|F_{\rm L}| - |F_{\rm D}|$) with associated error $\sigma_{\Delta F} = (\sigma_{\rm L}^2 + \sigma_{\rm D}^2)^{1/2}$. Experimental weighted maps were generated using phases

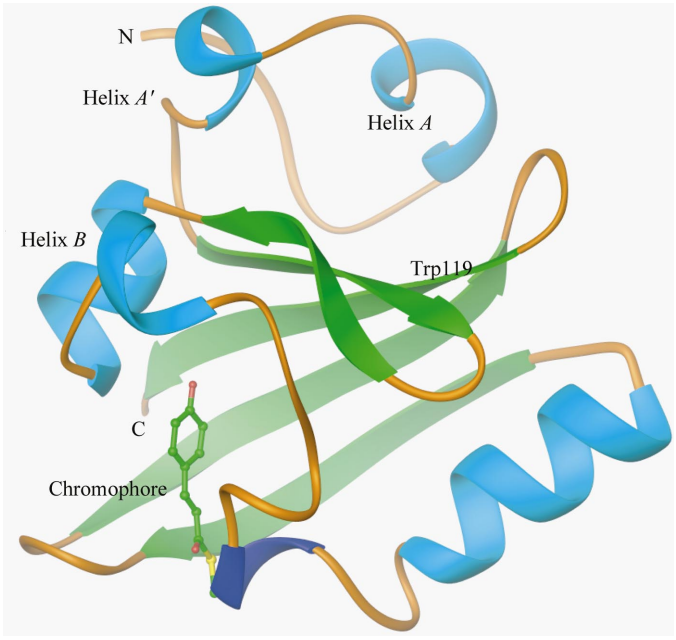

Figure 1
Structure of the E46Q mutant of PYP at room temperature; the 4-hydroxycinnamic acid chromophore is shown in ball-and-stick representation. Labeled are the N- and C-termini and three regions of the protein that are known to contain signal during the photocycle (the chromophore-binding pocket, helix *B* and the N-terminal helices *A* and *A*′) and one region that is largely devoid of signal, centered on Trp119.

from a room-temperature dark-state structure after a round of simulated annealing with the chromophore omitted, in which ΔF values were weighted based on the method of Ursby & Bourgeois (1997) used by Ren *et al.* (2001): $w = 1/[1 + (\sigma^2_{\Delta F}/\langle\sigma^2_{\Delta F}\rangle) + (\Delta F^2/\langle\Delta F^2\rangle)]$.

3.2. SVD flattening of maps

The values of $R_{\rm merge}$ on F ranged from 4.7 to 8.2% for light data sets and from 3.8 to 7.1% for dark data sets, with a typical overall completeness of 75–80% to 1.60 Å. Complete data-reduction statistics on the 54 individual data sets are provided by Anderson (2003). The dark-state structure of PYP is shown in Fig. 1. The strongest difference features during the E46Q PYP photocycle are in the chromophore-binding pocket. Representative difference electron-density maps from selected time delays are shown in Figs. 2(*a*)–2(*d*). Experimental weighted maps from single time delays are of similar quality to simulated maps at the 5 s/3 s noise level generated by Schmidt *et al.* (2003). Analysis of these simulated data sets required SVD flattening to improve S/N in order to determine the chemical kinetic mechanism of the system. Therefore, we subjected all 30 experimental maps to SVD flattening.

The SVD flattening procedure takes advantage of the noise-filtering ability of SVD. As data were collected over a long time range during which a number of intermediates are predicted to exist (Hellingwerf *et al.*, 2003), the total noise content of the data is much higher relative to the signal associated with a single intermediate. Therefore, we separated the 30 data sets into four groups of nine maps; in each group, the signal associated with a single intermediate would not be overwhelmed by the noise in the data. By overlapping each set with its neighbor(s) by one or two time delays, we could compare the same time delay from adjacent groups after flattening to ensure continuity from group to group and the absence of bias introduced by the procedure. These four groups were subjected to SVD of all difference electron-density features within the protein above the $\pm2\sigma$ level, where σ is the r.m.s. value of the difference electron density across the entire map. That is, all features between -2σ and $+2\sigma$ and all features outside of the protein mask were set to zero. After testing a number of different σ values, the 2σ value was chosen to minimize the amount of noise, which is likely to have low σ levels, while preserving features arising from authentic signal. After rotation (see below), significant singular vectors (SVs) were selected to reconstruct the maps (typically four of the nine SVs), which were then back-transformed to provide difference structure factors, $\Delta F_{\rm SVD}$, and phases, $\varphi_{\rm SVD}$. Improved values of $\Delta F_{\rm new}$ and $\varphi_{\rm new}$ were then generated by phase recombination according to Schmidt *et al.* (2003),

$$
\begin{aligned}
|F''_{\rm L}| &= [(|w\Delta F_{\rm old}|\sin\varphi_{\rm SVD} + |F_{\rm D}|\sin\varphi_{\rm D})^2 \\
&\quad + (|w\Delta F_{\rm old}|\cos\varphi_{\rm SVD} + |F_{\rm D}|\cos\varphi_{\rm D})^2]^{1/2}, \\
P &= \left\{\frac{1}{1 + [(\sigma^2_{\rm D} + \sigma^2_{\rm L})/(\langle\sigma^2_{\rm D} + \sigma_{\rm L}L^2\rangle)]}\right\}, \\
|F'_{\rm L}| &= (1 - P)|F''_{\rm L}| + P|F_{\rm L}|, \\
\varphi_{\rm L} &= a\tan\left(\frac{|w\Delta F_{\rm old}|\sin\varphi_{\rm SVD} + |F_{\rm D}|\sin\varphi_{\rm D}}{|w\Delta F_{\rm old}|\cos\varphi_{\rm SVD} + |F_{\rm D}|\cos\varphi_{\rm D}}\right), \\
|\Delta F_{\rm new}| &= [(|F'_{\rm L}|\sin\varphi_{\rm L} - |F_{\rm D}|\sin\varphi_{\rm D})^2 \\
&\quad + (|F'_{\rm L}|\cos\varphi_{\rm L} - |F_{\rm D}|\cos\varphi_{\rm D})^2]^{1/2}, \\
\varphi_{\rm new} &= a\tan\left(\frac{|F'_{\rm L}|\sin\varphi_{\rm L} - |F_{\rm D}|\sin\varphi_{\rm D}}{|F'_{\rm L}|\cos\varphi_{\rm L} - |F_{\rm D}|\cos\varphi_{\rm D}}\right).
\end{aligned}
$$

The mean phase shift $(\varphi_{\rm new} - \varphi_{\rm D})$ after a single round of SVD flattenening at the 2σ level was 35°. Further cycles of flattening did not improve the phase significantly. (It is possible that further phase improvements may be achieved using other density-modification strategies; Cowtan & Main, 1996). Maps

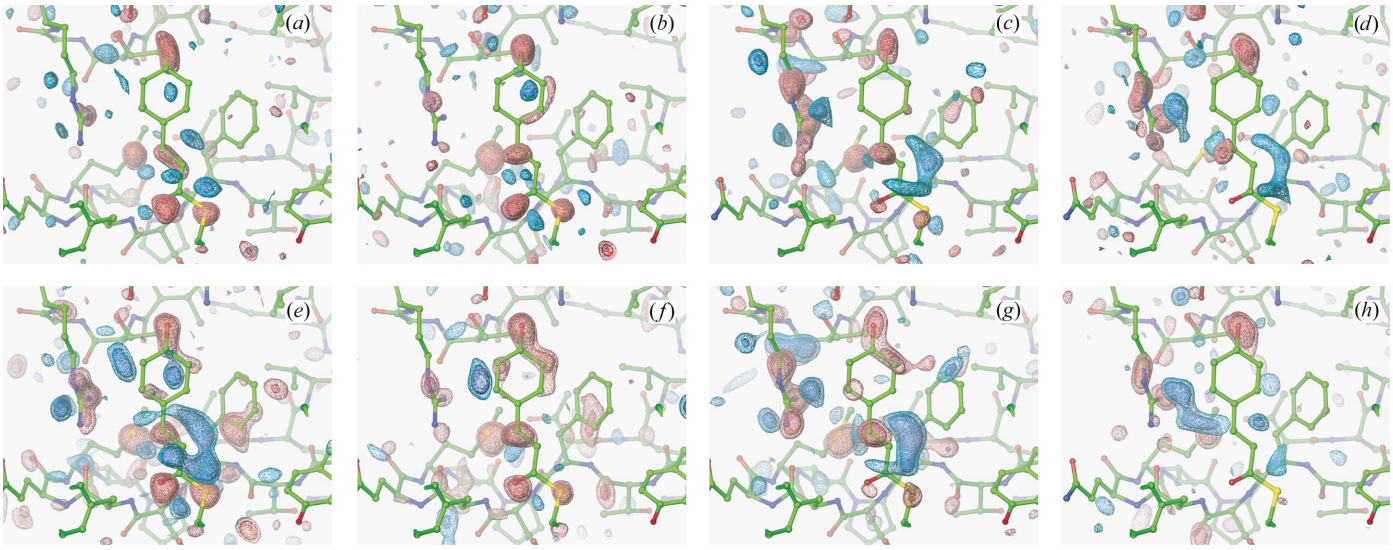

Figure 2
Difference electron-density maps before and after SVD flattening. Experimental weighted maps from (*a*) 10 ns, (*b*) 1 µs, (*c*) 100 µs and (*d*) 7 ms. (*e–h*) Corresponding maps after one round of SVD flattening at the 2σ level (see §3.2). Maps are contoured at -4σ (red), -3σ (pink), $+3\sigma$ (cyan) and $+4\sigma$ (blue).

were then generated from the $\Delta F_{\rm new}$ and $\varphi_{\rm new}$ for construction of the data matrix for subsequent SVD analysis.

3.3. Improvement in signal-to-noise ratio by SVD flattening

The improvement in difference electron-density quality is visually apparent from maps before (Figs. 2*a*–2*d*) and after (Figs. 2*e*–2*h*) SVD flattening. Features present in experimental weighted maps tend to be stronger in the SVD flattened maps and many features not visible in experimental maps are present in the SVD flattened maps at the same contour level. To quantify the improvement in map quality arising from SVD flattening, we calculated the standard deviation of the difference electron-density distributions in spheres of 5 Å radius centered on four regions labeled in Fig. 1: the chromophore-binding pocket, helix *B*, the N-terminal helices *A* and *A*′ and an 'empty' region of protein that lacks major difference features throughout the time course, centered at Trp119. The ratios of the standard deviations (which we refer to as a quality factor, *Q*) of the chromophore-binding pocket, helix *B* and N-terminal electron density distributions to that of the 'empty' region allow us to estimate the S/N of the individual maps. This approach is similar to the real-space free residual used to monitor the progress of density modification (Abrahams, 1997). The improvements in S/N from SVD flattening of all 30 maps is clear in a comparison of the values of *Q* for the chromophore-binding pocket (Fig. 3*a*), helix *B* (Fig. 3*b*) and the N-terminus (Fig. 3*c*) of the SVD flattened maps with those for the experimental weighted maps. In almost all maps, there is a significant improvement in S/N. One clear example is the increased *Q* in flattened maps compared with experimental maps of the N-terminal features in the millisecond time range (Fig. 3*c*). We conclude that SVD flattening significantly increases the S/N in difference electron-density maps.

3.4. Correction of outliers

In the mode of data collection typically used (Ren *et al.*, 1999), a single crystal yields one or at most a few data sets, each at a different time point. However, the absorption properties of the crystal, the quality of its diffraction and the beamline and laser parameters all vary from one data set to another. The individual data sets therefore have different levels of photoinitiation and signal, S/N, crystallographic resolution, redundancy and completeness. In other words, there is significant systematic error from time point to time point across this large E46Q PYP data set. In SVD analysis, such variations may result in signal being interpreted as noise. The rSV 1 of the uncorrected data is shown in Fig. 4(*a*); while there is a significant amount of scatter in the graph, a clear structure is visible in it. We attempted to correct for the systematic error responsible for these deviations in two steps. Firstly, we calculated the level of photoinitiation from the strength of the negative feature in the difference maps associated with the phenolate oxygen of the chromophore, which we know to be an authentic structural signal, by estimating the occupancy of the phenolate oxygen feature using the dark-state model in *STFACT* (McRee, 1999). This feature arises from the displacement of this oxygen that accompanies isomerization of the chromophore and should be present with identical magnitude in all data sets with equal level of photoinitiation where the protein has yet to fully relax back to the dark state. The average percentage of photoinitiation calculated using this method was approximately 20%, partially

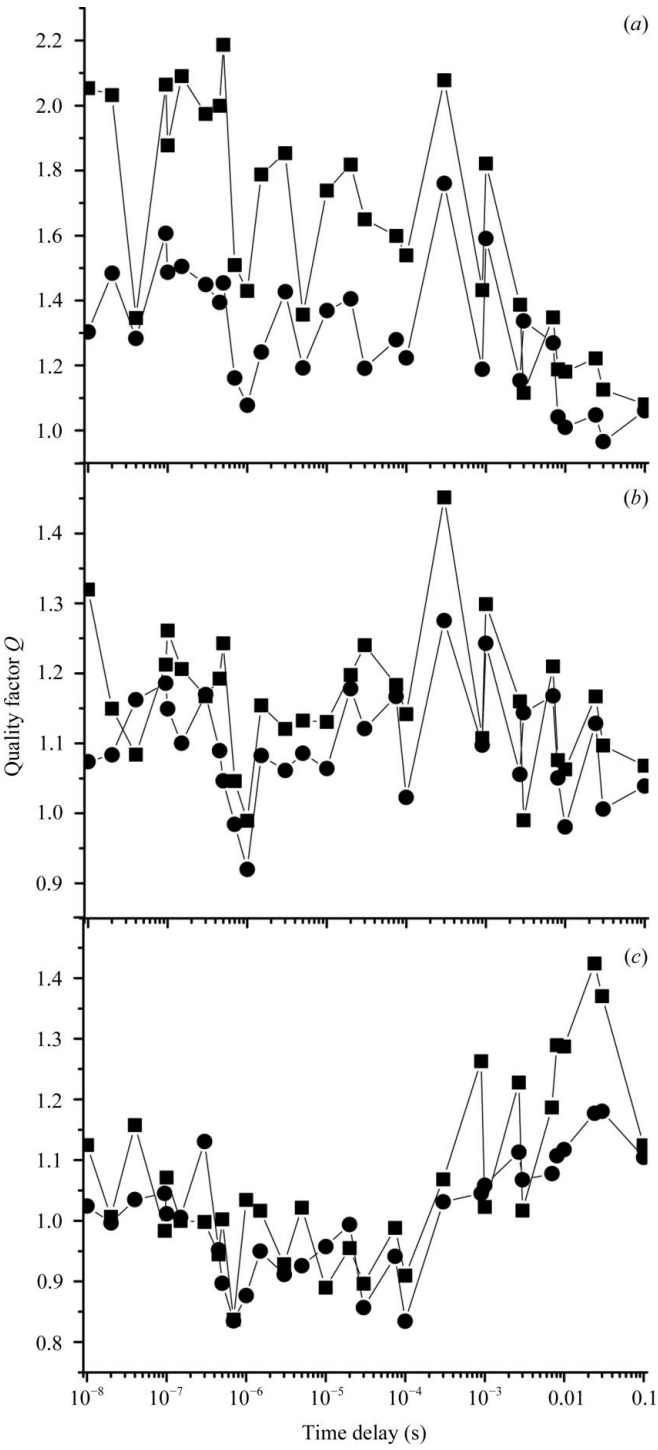

Figure 3
Improvement in S/N by SVD flattening. Quality factors *Q* before (circles) and after (squares) the procedure are shown for (*a*) the chromophore-binding pocket, (*b*) helix *B* and (*c*) the N-terminal helices.

taking into account the significant underestimate arising from the difference Fourier approximation. A plot of rSV 1 after correction for the level of photoinitiation (Fig. 4*b*) shows that it is mostly smoother at early and late times, but makes some time points even larger outliers, *e.g.* the 40 ns time point. This is because individual maps may have a strong or weak phenolate oxygen feature relative to other features in the map, which could introduce large errors into our correction factors. To ensure that the rest of the time course of rSV 1 was not being adversely affected by the presence of large outliers, we conducted a jackknife test. We randomly removed five of the 30 time points nine times and examined rSV 1, with the results shown in Fig. 4(*c*). The time course of rSV 1 does not change when these time points are removed, consistent with rSV 1 containing authentic signal that is present across all time points.

The first SV corresponds to the largest component of signal across all maps and represents an average over all features in a complete data set; thus, it is less susceptible to the errors associated with a correction factor calculated from a single feature such as the chromophore phenolate oxygen. Therefore, we further corrected these data sets based on their deviations from the smoothed first rSV generated by five point adjacent averaging (Fig. 4*d*; Schmidt *et al.*, 2003). Seven (of the total of 30) data sets required corrections of over 30% from the original photoinitiation correction. At later time points in the millisecond time range, the chromophore reisomerizes to its dark-state conformation and a photoinitiation factor

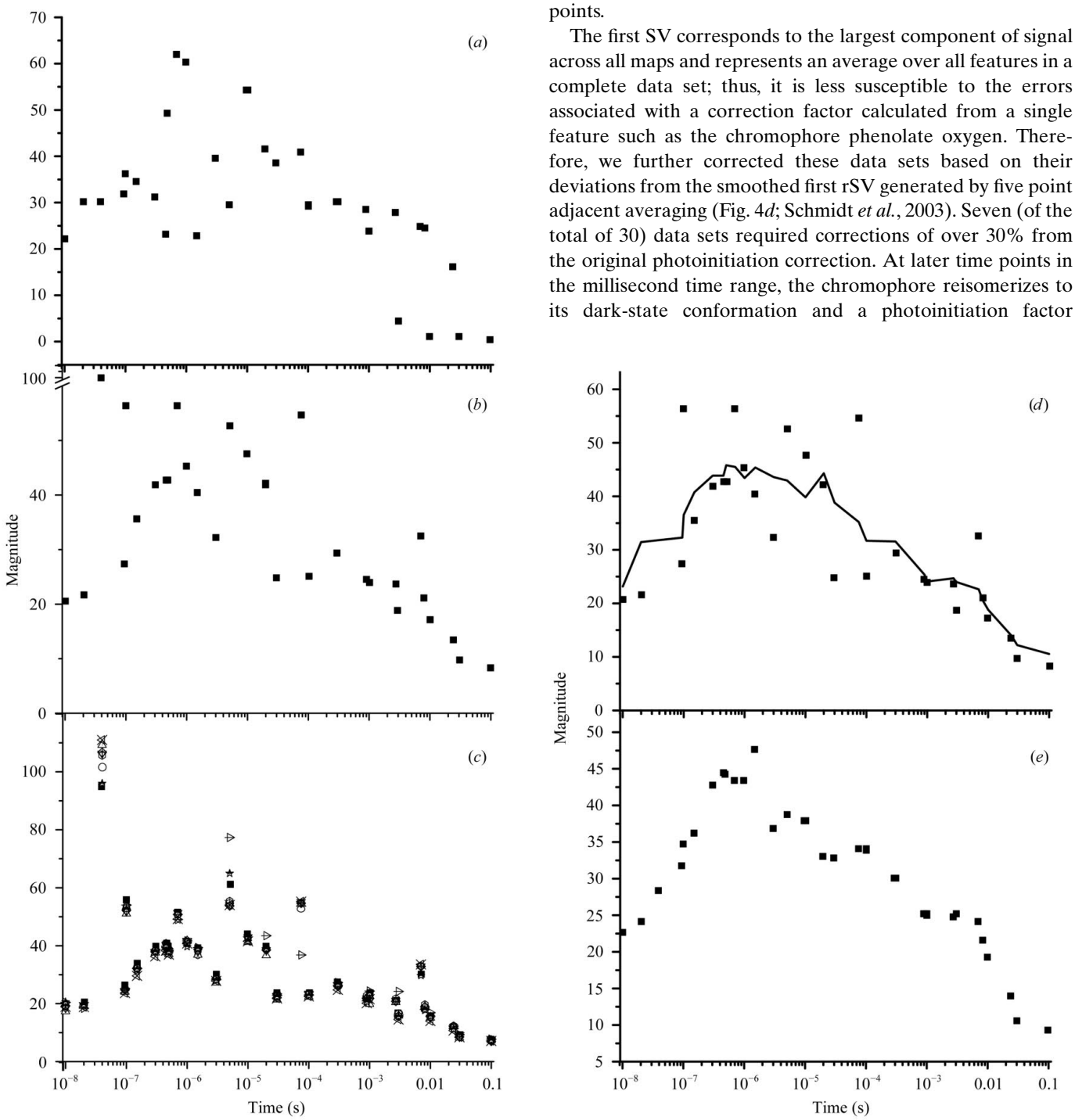

Figure 4
Right singular vector 1 (*a*) before any corrections and (*b*) with a correction based on the strength of the negative feature arising from displacement of the phenolate oxygen. A jackknife test was performed by randomly removing five of the 30 time points nine times and applying SVD, resulting in near-identical values of the rSVs as shown in (*c*). The final correction was based on (*d*) the shape of the smoothed rSV 1, resulting in (*e*) the corrected rSV 1 used for data analysis.

Table 1
Quality factors Q for the chromophore, helix B and N-terminal regions (see §3.3 for the definition of Q).

SVs selected for reconstitution of the data matrix are shown in bold. After removal of outliers by Z value, the mean and standard deviation for the chromophore, helix B and N-terminal Q values are 1.208 ± 0.075, 1.072 ± 0.069 and 1.025 ± 0.054, respectively.

SV	Chromophore	SV	Helix B	SV	N-terminus
1	3.152	**1**	1.521	**2**	1.460
2	1.789	**2**	1.233	**6**	1.325
7	1.423	**6**	1.217	7	1.158
4	1.316	13	1.177	27	1.127
3	1.300	9	1.151	19	1.090
8	1.279	17	1.142	12	1.083
20	1.266	27	1.140	30	1.056
9	1.262	7	1.124	13	1.054
23	1.260	19	1.115	29	1.052
12	1.249	8	1.097	26	1.049
17	1.248	28	1.093	23	1.044
25	1.242	**3**	1.083	25	1.043
5	1.239	**4**	1.078	**3**	1.042
19	1.222	12	1.069	21	1.035
6	1.209	5	1.059	11	1.035
13	1.204	30	1.054	24	1.035
24	1.193	25	1.053	15	1.030
30	1.191	21	1.051	17	1.025
14	1.191	10	1.041	**4**	1.019
28	1.187	23	1.041	**1**	1.017
26	1.177	14	1.038	9	1.006
21	1.169	15	1.034	20	1.000
11	1.156	26	1.025	28	0.994
16	1.140	20	1.021	22	0.988
15	1.134	24	1.018	18	0.985
10	1.129	11	1.016	14	0.961
29	1.119	22	1.006	8	0.961
18	1.116	18	0.980	16	0.961
27	1.100	16	0.976	5	0.958
22	1.088	29	0.952	10	0.901

cannot be applied. To correct these time points, we used a data series spanning 900 µs to 24 ms in which different data sets were collected from the same crystal on the same volume of reciprocal space (Anderson, 2003). As these data sets do not have the large systematic error arising from crystal-to-crystal variation, they could be used to correct temporally adjacent time points based on the shape of the first rSV, the values of which after these corrections are shown in Fig. 4(*e*).

We discuss in §5 how the necessity for the elaborate corrections described above to minimize systematic error in this particular data set may be bypassed.

4. Data evaluation

These 30 fully corrected SVD-flattened maps were then subjected to SVD, resulting in 30 SVs with the singular values and rSV autocorrelation coefficients shown in Fig. 5(*a*). At this stage, we ordered the SVs by the magnitude of their singular value. While the first singular vector has a high singular value and autocorrelation, the other SVs have much lower singular values and autocorrelations. This confirms that a significant amount of noise is present in the data, which mixes signal with noise in a majority of the SVs.

4.1. Rotation

Rotation is required to partition signal from a large number of SVs that contain both signal and noise into a small subset of the SVs that primarily contain signal. For a detailed discussion of the rotation procedure, see Henry & Hofrichter (1992). Rotation generates a new set of rSVs from a linear combination of the old rSVs in order to maximize the autocorrelation of some rSVs at the expense of others. As signal tends to vary smoothly with respect to time while random noise does not, rotation tends to partition signal into rSVs with high autocorrelation and noise into rSVs with low autocorrelation. Rotation, however, results in the loss of the orthonormality property of the SVs and loss of their least-squares properties (Henry & Hofrichter, 1992). The selection of those SVs to be rotated is critical: selection of too few SVs for rotation would result in signal being excluded from the analysis, which could bias the entire process, while selection of too many would introduce noise, thereby decreasing the S/N of the data and generating rSVs made arbitrarily smooth by the addition of noise. We assessed whether individual SVs contained signal

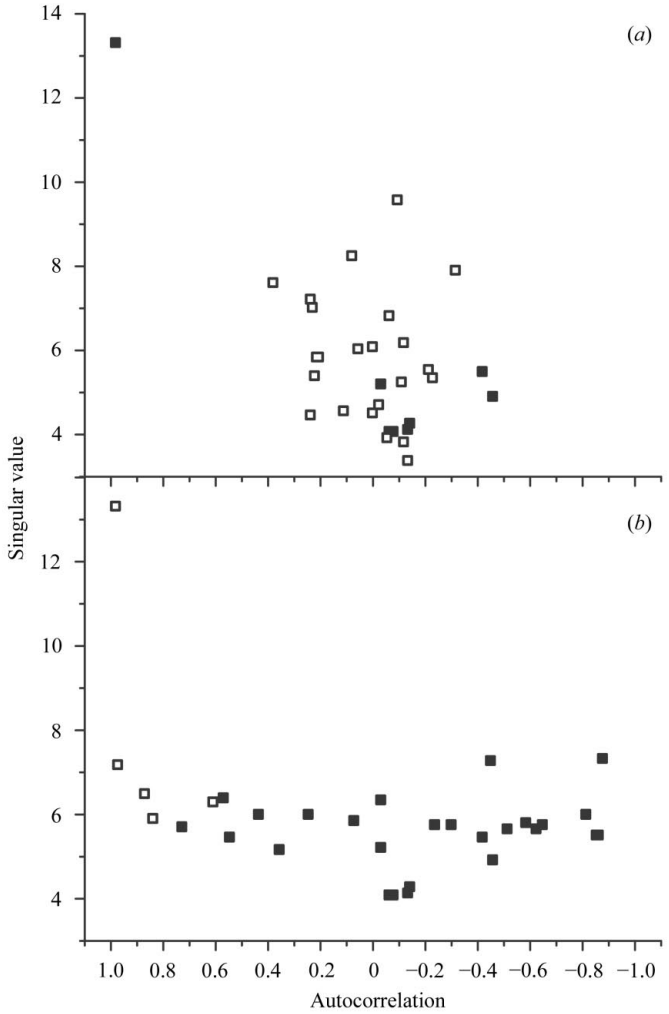

Figure 5
Magnitude of singular value *versus* autocorrelation (*a*) before (open squares: SVs selected for rotation) and (*b*) after (open squares: SVs selected for reconstitution) rotation of selected singular vectors.

using the criteria of Schmidt *et al.* (2003): the magnitude of its singular value, the autocorrelation of its rSV and visual inspection of the lSV. Based on these criteria, we selected SVs 2–13, 15–17, 20–23 and 28–30 for rotation (Fig. 5*a*, open squares), with the results shown in Fig. 5(*b*). At this stage, we reordered all SVs by the magnitudes of their autocorrelations. Most rotated SVs have singular values between 5 and 7 and have autocorrelations spread more or less uniformly from +0.9 to −0.9. While it is clear that those SVs with the lowest autocorrelations do not contain authentic signal and can safely be rejected, autocorrelation alone is not sufficient to assess whether or not a SV contains sufficient signal. Therefore, in order to determine which SVs contain signal and should be retained, more detailed analysis of the lSVs and rSVs is required.

4.2. Selection of singular vectors

Using the criteria of Schmidt *et al.* (2003), we assessed the quality of the rotated SVs. As is clear from Fig. 5(*b*), the first criterion (magnitude of the singular value) is in itself not useful to assess whether a SV contains signal, as the singular values for SVs 2–30 cluster tightly between 5 and 7. The second criterion (magnitude of the rSV autocorrelation) is useful, but outliers can affect the autocorrelation considerably. As the third criterion (visual inspection of the lSV) is useful but open to subjective interpretation, we formulated a statistic that is more objective: spatial clustering in the lSV, *i.e.* positive and negative difference features cluster close to one another in real space. The result of such clustering is that the standard deviation of the difference electron density in that volume is increased relative to those areas that do not contain significant signal. Thus, the quality factor Q previously used to assess the S/N in experimental and SVD-flattened maps is also a suitable measure to assess whether in an lSV certain regions of the protein contain significant amounts of signal.

Quality factors calculated for lSVs 1–30 are shown in Table 1. It is clear from these values that lSVs 1 and 2 have very high S/N, with lSV 1 having a significantly higher chromophore Q values than any of the other lSVs. lSVs 6 and 7 are also above the mean plus two standard deviations in two of the three quality factors. While lSVs 3 and 4 have quite high Q values for the chromophore region, they have much lower Q values in the other two regions. lSV 5 has low Q values in all three regions. On this basis, we tentatively conclude that lSVs 1–4 and 6–7 contain significant signal. Inspection of the

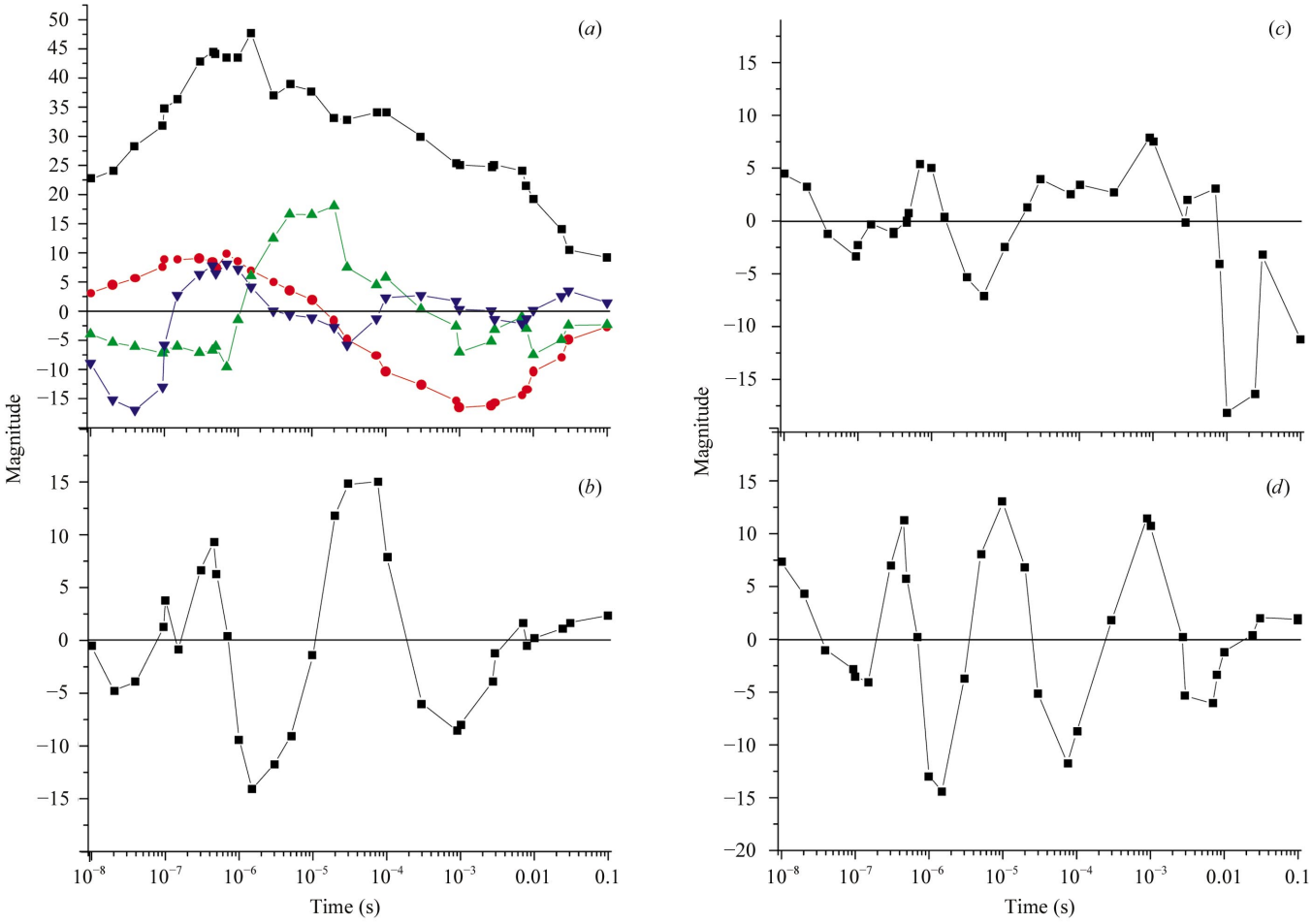

Figure 6
Right singular vectors weighted by the square of their respective singular value, plotted against time. (*a*) rSVs 1–4 (black, red, green and blue, respectively). (*b*) rSV 5, (*c*) rSV 6 and (*d*) rSV 7.

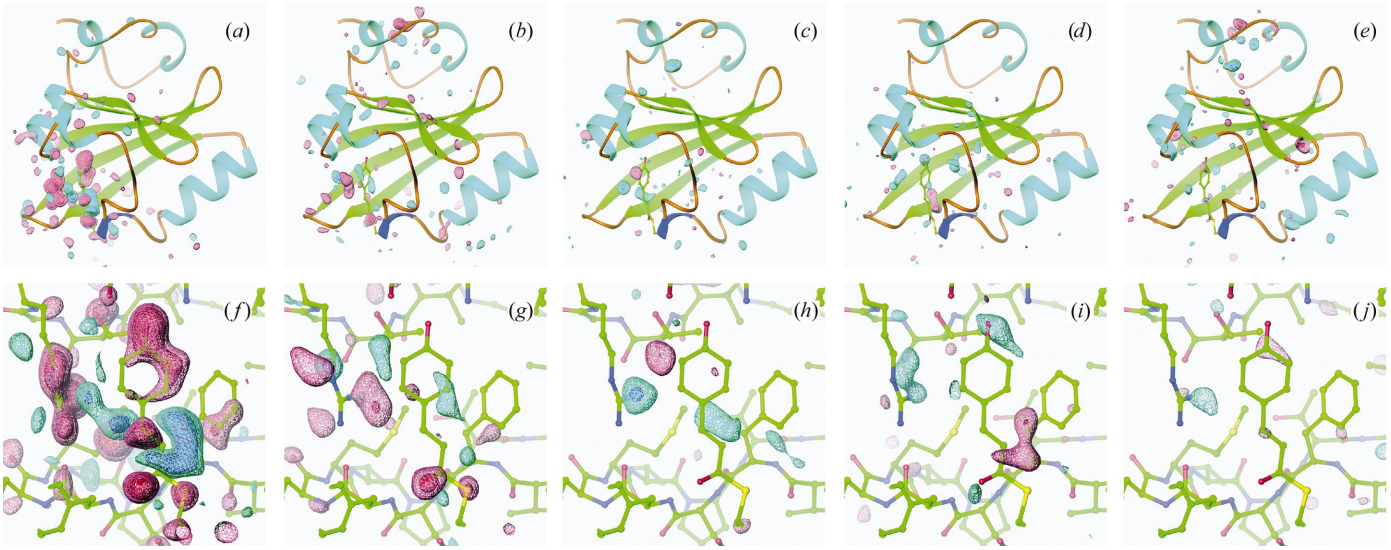

Figure 7
Difference electron density associated with significant left singular vectors (lSVs). Entire protein (*a*)–(*e*) and chromophore-binding pocket only (*f*)–(*j*). (*a*) and (*f*), lSV1; (*b*) and (*g*) lSV2; (*c*) and (*h*) lSV3; (*d*) and (*i*) lSV4; (*e*) and (*j*) lSV 6. Maps are contoured at -6σ (red), -3.5σ (pink), $+3.5\sigma$ (cyan) and $+6\sigma$ (blue).

corresponding rSVs (criterion 2) shows that rSVs 1–4 are smooth and well behaved (Fig. 6*a*). On the other hand, rSVs 5–7 are not smooth (Figs. 6*b* and 6*d*). In particular, rSV 7 oscillates rapidly, which is unlikely to be true of authentic signal. While rSV 5 and 6 also oscillate, rSV 6 is better behaved; its lower autocorrelation may arise from incompletely corrected outliers (Fig. 6*c*).

Because lSV 5 has such low quality factors and because rSV 7 oscillates rapidly, we discard these two SVs and retain SVs 1–4 and 6 for further analysis (Fig. 5*b*, open squares). These five lSVs are shown in Fig. 7. All show features consistent with signal: lSV1 contains strong signal in the chromophore-binding pocket and helix *B*, lSV2 contains strong signal in those two regions and the N-terminus, lSVs 3 and 4 have strong signal in the chromophore-binding pocket and lSV 6 shows strong signal in the N-terminus and weaker signal in helix *B*. Notably, all five lSVs display few features in the rest of the protein, suggesting low noise levels.

While this analysis shows that these SVs have high S/N and contain features consistent with authentic signal, it does not address whether they are sufficient to fully represent the experimental data. Thus, the question of whether they are sufficient to reconstruct the data matrix must be considered.

4.3. Reconstruction of the data matrix

SVs 1–4 and 6 selected above contain approximately 23% of the data from the input matrix (data not shown). Although containing less than a quarter of all data, this subset retains its strongest features. Maps reconstituted with SVs 1–4 and 6 have larger $\Delta\rho_{\rm max}$ and smaller $\Delta\rho_{\rm min}$ than maps reconstituted with all other SVs, SVs 5 and 7–30, in all time delays (except 3 ms) (Fig. 8). This result is consistent with the strongest features in the maps being signal and thus present in the selected subset of SVs. Throughout the time course, $\Delta\rho_{\rm min}$ is of higher magnitude than $\Delta\rho_{\rm max}$, which is again consistent with authentic signal for which negative features are stronger than positive. [Negative features tend to arise from the displacement of atoms from regions with low temperature factors (the dark state); positive features tend to arise from new atomic positions of low occupancy because of the presence of multiple intermediates and perhaps of lower magnitude owing to higher temperature factors in an intermediate state.] While this is strong evidence that these five SVs have captured most of the signal in the data, evaluation of the residual maps is critical in determining whether the data is properly represented by this subset. While residuals for an individual time point may show features consistent with signal, as long as signal does not persist across a number of tempo-

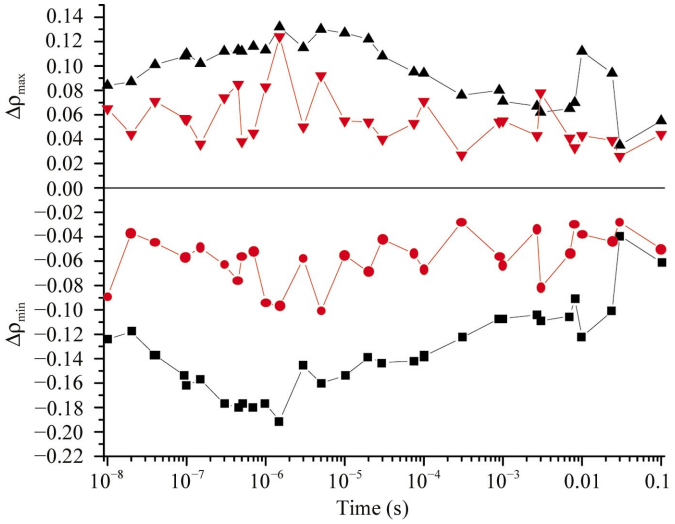

Figure 8
Maximum and minimum values of difference electron density on an absolute scale in maps reconstituted with SVs 1–4 and 6 (black) and SVs 5 and 7–30 (red).

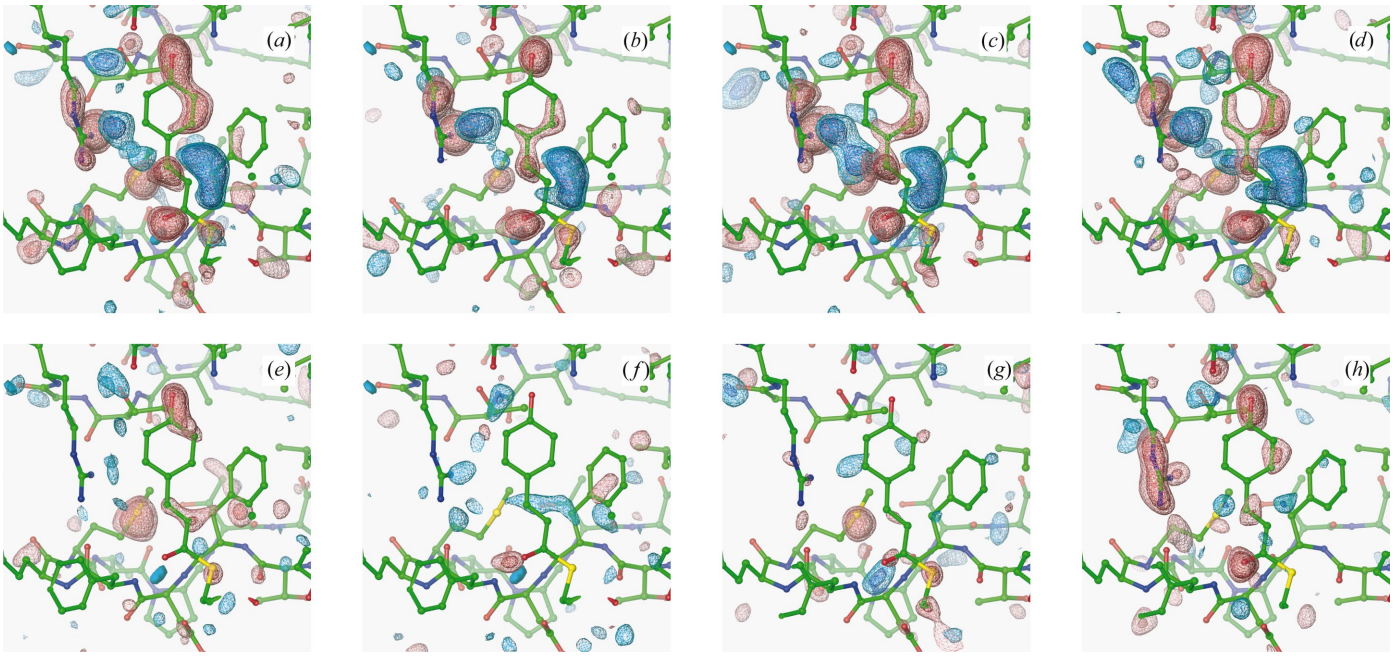

Figure 9
Difference electron-density maps generated from SVs 1–4 and 6 (reconstituted maps) (*a*)–(*d*) or SVs 5 and 7–30 (residual maps) (*e*)–(*h*) for 3, 5, 10 and 20 µs time delays, respectively. Maps are contoured at -4σ (red), -3σ (pink), $+3\sigma$ (cyan) and $+4\sigma$ (blue). The absolute scale of the individual maps can be estimated from Fig. 7.

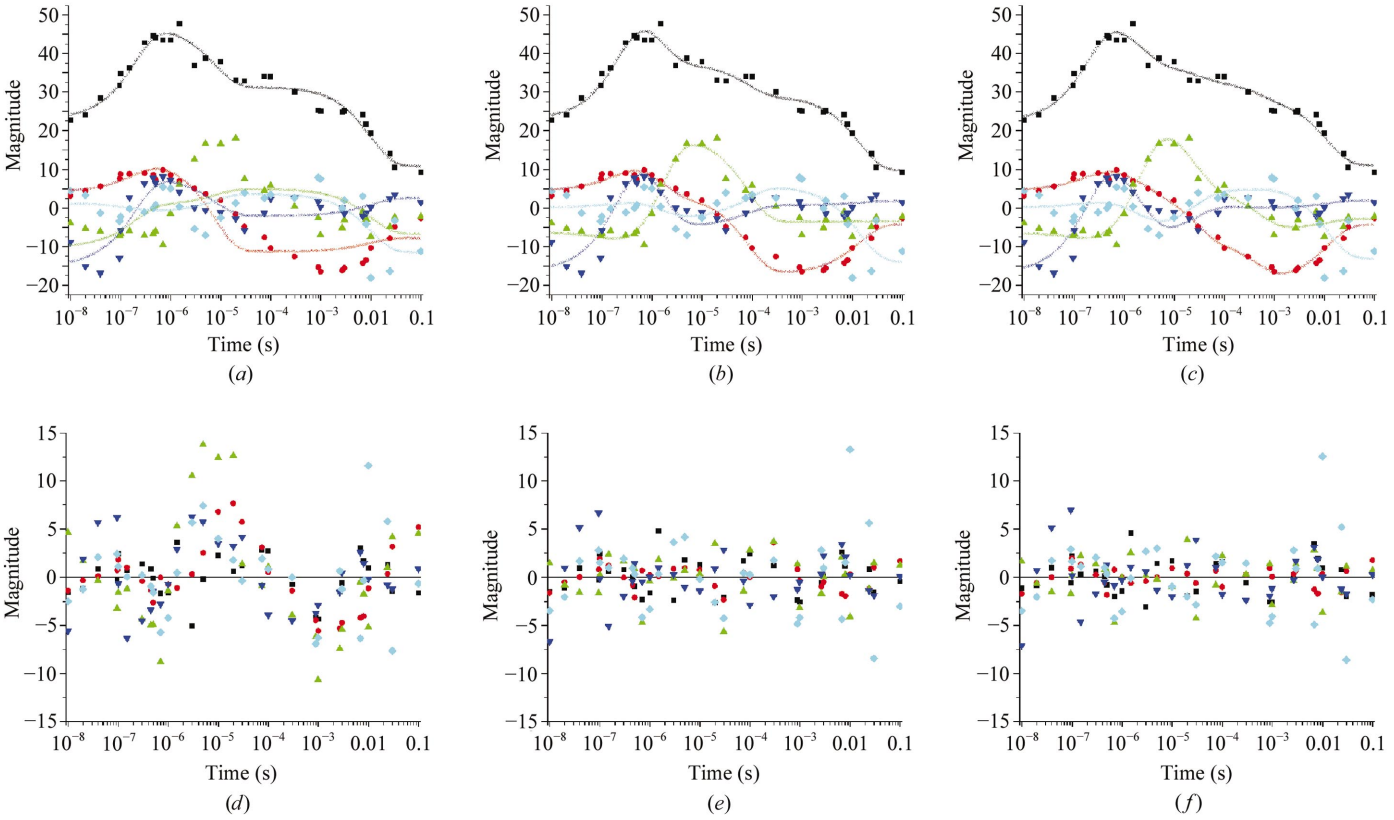

Figure 10
Fits of rSVs 1–4 and 6 (black, red, green, blue and cyan, respectively) weighted by the square of their corresponding singular value by a sum of exponentials. (*a*) Three exponentials [τ_1 = 196 ns; τ_2 = 7.58 µs; τ_3 = 9.9 ms; mean square deviation (MSD) = 17.4]. (*b*) Four exponentials (τ_1 = 233 ns; τ_2 = 1.54 µs; τ_3 = 84 µs; τ_4 = 14 ms; MSD = 6.73). (*c*) Five exponentials (τ_1 = 198 ns; τ_2 = 2.15 µs; τ_3 = 25.9 µs; τ_4 = 485 µs; τ_5 = 11.4 ms; MSD = 6.22). (*d*)–(*f*) Residuals of the fits with three, four and five exponentials, respectively.

rally adjacent time points it is unlikely that significant signal has been rejected by the analysis. Visual inspection of these residual maps over the entire time range does not show persistent signal. For example, residuals for the 3, 5, 10 and 30 µs time delays (Figs. 9*e*–9*h*) are compared with the corresponding maps reconstructed with SVs 1–4 and 6 (Figs. 9*a*–9*d*). Some weak noise peaks persist in the reconstructed maps arising from the retention of small amounts of noise in the SVs. Conversely, some weak signal is present in the individual residual maps, probably owing to random differences in the magnitude of various authentic features between different time delays. Although each individual residual map may contain features that are consistent with signal, as a set they do not display continuity in time nor contain strong positive and negative features, which are required for authentic signal associated with intermediates. We conclude that our selection of the five SVs captures the large majority of the signal and results in a scattered distribution of residuals. It is therefore sufficient to reconstitute the data matrix.

4.4. Fitting of the rSVs

The rSVs are a linear combination of the time course of the concentrations of the underlying chemical species and can therefore be fitted by a sum of relaxation times (Henry & Hofrichter, 1992). If a simple chemical kinetic mechanism holds and first-order reactions are considered, the relaxations should be simple exponentials, but if a complex mechanism holds then the rSVs can also be fitted by functions such as stretched exponentials (Moffat, 2001). To assess the existence of a simple chemical kinetic mechanism, we attempted to fit the rSVs with a sum of simple exponentials. Globally fitting the rSVs with three relaxation times (Fig. 10*a*) results in a poor fit of the rSVs, most notably of rSV 3, as is clear from an inspection of the residuals in the microsecond range (Fig. 10*d*). Fits with four (Figs. 10*b* and 10*e*) and five (Figs. 10*c* and 10*f*) relaxation times are very similar in quality; the addition of a fifth relaxation time leads only to a slightly better fit of rSV 2 in the hundreds of microseconds time range. Examination of the residuals from the four- and five-exponential fits shows that the first four time points of rSV 4 and the last four of rSV 6 fit poorly (Figs. 10*e* and 10*f*). The poor fit to rSV 6 is a consequence of its poor smoothness in the millisecond range (Fig. 6*c*). As for rSV 4, we did not attempt to add another exponential to fit an additional relaxation in the tens of nanoseconds range. This relaxation in the rSV arises from only three time points and is absent in all other rSVs. From the quality of these fits, we conclude that the rSVs are well described by a sum of four simple exponentials; there is no need for stretched exponentials or other more complex functions.

5. Conclusions

We conclude that SVD is a powerful tool in the analysis of experimental time-dependent crystallographic data. However, because of the significant systematic error present in this particular data set, much care had to be applied in its use. Bias can be introduced at a number of stages such as correction for crystal-to-crystal variation, SVD flattening, rotation and the selection of significant SVs. While it is possible to couch many of the subjective criteria used in the procedure in more objective terms, such as the use of spatial clustering instead of visual inspection alone, analysis of this particular data set required much user input and prior information. The major source of the experimental error is crystal-to-crystal variation, although the low S/N of the data also contributes. As discussed above, these errors propagate into the SVD analysis. The SVD analysis thus suggests improvements in data-collection strategy, such as the collection of data with time as the fast variable in which crystal-to-crystal variation is largely masked. Improvements at Laue beamlines have made data collection much faster and allowed the collection of data with higher completeness and redundancy. Preliminary SVD analysis of wild-type PYP data collected in such a manner supports this view; we find that the need for factors to correct for the extent of photoinitiation, a major complication in the present analysis, is completely obviated (H. Ihee *et al.*, personal communication).

This study also lays the groundwork for the determination of a chemical kinetic mechanism for the photocycle of PYP. The fit with four relaxation times is consistent with the presence of a simple chemical kinetic mechanism and the number of relaxations present in the data should correspond to the minimum number of intermediates present (Schmidt *et al.*, 2003). Since the number of significant SVs (five) also gives a lower bound on the number of intermediates, this information provides a powerful constraint in the formulation of candidate mechanisms to be used to fit the data. This then allows for the difference electron density associated with each intermediate to be determined, time-independent intermediate structures to be refined and the chemical kinetic mechanism of E46Q PYP identified. These developments are presented in a companion paper (Rajagopal *et al.*, 2004).

We thank Vukica Šrajer and Jason Key for valuable discussions. HI was supported by a postdoctoral fellowship from the Damon Runyon Cancer Research Foundation. Supported by NIH grants GM36452 and RR07707 to KM.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.
Anderson, S. (2003). PhD thesis, University of Chicago, USA.
Anderson, S., Šrajer, V., Pahl, R., Rajagopal, S., Schotte, F., Anfinrud, P., Wulff, M. & Moffat, K. (2004). Submitted.
Bourgeois, D., Vallone, B., Schotte, F., Arcovito, A., Miele, A. E., Sciara, G., Wulff, M., Anfinrud, P. & Brunori, M. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 8704–8709.
Bourgeois, D., Wagner, U. & Wulff, M. (2000). *Acta Cryst.* D**56**, 973–985.
Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**52**, 43–48.
Genick, U. K., Borgstahl, G. E., Ng, K., Ren, Z., Pradervand, C., Burke, P. M., Šrajer, V., Teng, T. Y., Schildkamp, W., McRee, D. E., Moffat, K. & Getzoff, E. D. (1997). *Science*, **275**, 1471–1475.

Hellingwerf, K. J., Hendriks, J. & Gensch, T. (2003). *J. Phys. Chem. A*, **107**, 1082–1094.
Henry, E. & Hofrichter, J. (1992). *Methods Enzymol.* **210**, 129–192.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Moffat, K. (1989). *Annu. Rev. Biophys. Biophys. Chem.* **18**, 309–323.
Moffat, K. (2001). *Chem. Rev.* **101**, 1569–1581.
Moffat, K. & Henderson, R. (1995). *Curr. Opin. Struct. Biol.* **5**, 656–663.
Perman, B., Šrajer, V., Ren, Z., Teng, T., Pradervand, C., Ursby, T., Bourgeois, D., Schotte, F., Wulff, M., Kort, R., Hellingwerf, K. & Moffat, K. (1998). *Science*, **279**, 1946–1950.
Rajagopal, S., Anderson, S., Šrajer, V., Schmidt, M., Pahl, R. & Moffat, K. (2004). In preparation.
Ren, Z., Bourgeois, D., Helliwell, J., Moffat, K., Šrajer, V. & Stoddard, B. (1999). *J. Synchrotron Rad.* **6**, 891–917.
Ren, Z. & Moffat, K. (1995*a*). *J. Appl. Cryst.* **28**, 482–493.
Ren, Z. & Moffat, K. (1995*b*). *J. Appl. Cryst.* **28**, 461–481.
Ren, Z., Perman, B., Šrajer, V., Teng, T.-Y., Pradervand, C., Bourgeois, D., Schotte, F., Ursby, T., Kort, R., Wulff, M. & Moffat, K. (2001). *Biochemistry*, **40**, 13788–13801.
Schlichting, I. & Chu, K. (2000). *Curr. Opin. Struct. Biol.* **10**, 744–752.
Schmidt, M., Pahl, R., Srajer, V., Anderson, S., Ren, Z., Ihee, H., Rajagopal, S. & Moffat, K. (2004). Submitted.
Schmidt, M., Rajagopal, S., Ren, Z. & Moffat, K. (2003). *Biophys. J.* **84**, 2112–2129.
Schotte, F., Lim, M., Jackson, T. A., Smirnov, A. V., Soman, J., Olson, J. S., Phillips, G. N. Jr, Wulff, M. & Anfinrud, P. A. (2003). *Science*, **300**, 1944–1947.
Šrajer, V., Ren, Z., Teng, T.-Y., Schmidt, M., Ursby, T., Bourgeois, D., Pradervand, C., Schildkamp, W., Wulff, M. & Moffat, K. (2001). *Biochemistry*, **40**, 13802–13815.
Šrajer, V., Teng, T., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M. & Moffat, K. (1996). *Science*, **274**, 1726–1729.
Ursby, T. & Bourgeois, D. (1997). *Acta Cryst.* A**53**, 564–575.