Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Structural Biology 147 (2004) 211–222
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Analytical trapping: extraction of time-independent structures from time-dependent crystallographic data☆

Sudarshan Rajagopal,[a] Konstantin S. Kostov,[a] and Keith Moffat[a,b,*]

[a] *Department of Biochemistry and Molecular Biology, The University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA*
[b] *Institute for Biophysical Dynamics, The University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA*

Received 18 December 2003, and in revised form 5 March 2004
Available online 12 May 2004

Abstract

All chemical and biological reactions involve atomic motion, embodied in dynamic structural changes. Identifying these changes is the goal of time-resolved crystallography. The "raw" output of a time-resolved macromolecular crystallography experiment is the time-dependent set of difference electron density maps that span the desired time range and display the time-dependent changes in density (and underlying structure) as the reaction progresses. The goal is to interpret such data in terms of a small number of crystallographically refinable, time-independent structures, each associated with a reaction intermediate; to establish the pathways and rate coefficients by which the intermediates interconvert; and thus to establish a chemical kinetic mechanism. We review briefly the various strategies that may be used to achieve this goal and concentrate on two promising advances: singular value decomposition and cluster analysis. The strategies are illustrated by using data on the photocycle of the bacterial blue light photoreceptor, photoactive yellow protein.

Keywords: Time-resolved crystallography; Analytical trapping; Laue diffraction; Singular value decomposition; Cluster analysis; Chemical kinetic mechanism; Photoactive yellow protein

1. Introduction

1.1. The nature of the time-resolved crystallographic experiment

Macromolecular crystallography has enjoyed decades of spectacular success, from the structures of myoglobin and lysozyme to the photosynthetic reaction center, ion channels, icosahedral viruses, and now the ribosome. Nevertheless, mechanism has proven harder to establish than static, time-independent structure per se. This is not altogether surprising since all chemical and biological reactions involve atomic motion, embodied in dynamic structural changes. These structural changes can be very fast and span a wide time range, from femtoseconds to seconds or even longer in certain biological reactions; thus, high time resolution is required, over an extended time range. The key structural changes may be of very limited extent; thus, high spatial resolution is required. The mechanism typically involves a series of intermediates lying between reactants and products; thus, the complexities associated with numerous, short-lived structures and the pathways by which they interconvert must be effectively addressed.

The structures of intermediate states are normally determined by physical or chemical trapping techniques designed both to greatly increase the lifetime of molecules in a desired state and to trap a structurally homogeneous species (Cruickshank et al., 1992; Moffat and Henderson, 1995; Schlichting and Chu, 2000; Stoddard, 1996). Cryocooling is often employed to literally freeze out key atomic motions that accompany reaction, and hence permit cryotrapping of a particular intermediate: physical trapping. Warming of a cryotrapped crystal may then cause a series of structurally

☆ Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2004.04.007.

* Corresponding author. Fax: 1-773-702-0439.
E-mail address: moffat@cars.uchicago.edu (K. Moffat).

1047-8477/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2004.04.007

distinct intermediates to be progressively populated. Alternatively, the chemistry of the system may be perturbed by studying a mutant macromolecule, a variant substrate or an unusual solvent condition, such that a chemical bottleneck develops in the pathway and preceding intermediate(s) accumulate: chemical trapping. Although very widely employed, trapping approaches have limitations. The trapping techniques themselves may perturb the reaction mechanism; the energetic extrapolation necessary to infer the structure of the authentic intermediate from its trapped, very long-lived counterpart may be large particularly if the authentic intermediate is very short-lived; and despite best efforts, the trapped species may still not be spectroscopically or structurally homogeneous (Moffat, 2001).

In contrast, time-resolved crystallography employs no trapping and studies the evolution of the space-average structure in the crystal in real time, as the reaction proceeds. This may be thought odd since the essence of a crystal is time-independent, spatially periodic order. Can time-resolved crystallography even be conducted, when the explicit time dependence of structural populations is sought and spatial periodicity is lost? Can very fast structural transitions be produced in the crystalline state and directly observed? Can the structures of short-lived intermediates and the paths by which they interconvert be identified? That is, can mechanism be directly investigated by structural means? As recent results in the field have indicated, the answer to these questions is "yes," at least for the systems that we and others have investigated so far.

The time-dependent average structure in the crystal varies due to the time-dependent rise and fall of the concentrations of the underlying, time-independent structures (Moffat, 2001). The structures themselves do not vary with time; only their concentrations do. Thus:

$$\rho(r,t) = \sum_i c_i(t)\rho_i(r), \tag{1}$$

where the summation is taken over species i, r denotes the fractional cell coordinate, t denotes time, $c_i(t)$ denotes the fractional concentration of species i at time t, $\rho(r,t)$ denotes the average electron density over all unit cells in the crystal at the fractional cell coordinate r and time t, and $\rho_i(r)$ denotes the electron density at r of species i. An exactly equivalent equation for difference electron density replaces ρ by $\Delta\rho$ throughout. Eq. (1) describes a particularly simple case in which the various species are spatially distributed at random within the crystal and there is no correlation between the conformations of adjacent molecules. That is, the molecules in the crystal behave independently of one another as if they were in dilute solution, and the transitions between intermediate structural states are uncorrelated in time from molecule to molecule. There is no synchronization; the molecules do not move in lock-step. (This cannot be strictly true. Depending on the method of reaction initiation, there is likely to be at least a small concentration gradient across the crystal; and lattice forces, although weak, provide some coupling—but not necessarily correlation of conformations—between adjacent molecules. Nevertheless Eq. (1) turns out to hold experimentally in the systems studied to date.) As in all kinetic experiments, several species i may be present at all times; that is, the crystal exhibits time-dependent structural heterogeneity. The unusual challenge in time-resolved crystallography is to identify and extract the individual values of $\rho_i(r)$ or $\Delta\rho_i(r)$, the time course of each species $c_{\mathrm{i}}(\mathrm{t})$ and the paths by which the species interconvert, from this heterogeneous mixture. As we illustrate, this is achieved by careful measurement of $\rho(t)$ or $\Delta\rho(t)$ over the course of the reaction and by applying certain simple constraints at the data analysis stage.

This research thus directly examines the structural bases of mechanism, at high crystallographic resolution and with temporal resolution from a few nanoseconds (Bourgeois et al., 1996; Perman et al., 1998; Srajer et al., 1996) to ~100 ps (Bourgeois et al., 2003; Schotte et al., 2003). To our knowledge it is the only such experimental approach. Further, the detailed knowledge of intermediate structures and of the paths and rate coefficients by which they interconvert forms a much stronger foundation for computational studies than more accessible and stable reactant, trapped quasi-intermediate, or product structures.

The initial goal of a time-resolved crystallographic experiment is to establish whether any chemical kinetic mechanism holds for a particular biochemical reaction and if so, to identify the distinct, short-lived structural states or intermediates that populate such a mechanism. A further goal is to establish the pathways by which these intermediates interconvert, and the magnitude of the rate coefficients by which the structural transitions occur. A description of the structures, pathways and rates would then constitute a full explication of the mechanism at the experimental level (Moffat, 1989, 2001). In practice, a time-resolved crystallographic experiment seeks to measure the variation with time of all structure factor amplitudes $|\mathbf{F}(hkl,t)|$, spanning the time range from reaction initiation at $t = 0$ until $t = t_{\max}$ when the structural reaction is fully complete. $\Delta\rho(r,t)$ is obtained by Fourier transformation of $\Delta F(hkl,t)\exp(\mathrm{i}\phi_{hkl,0})$, where $\phi_{hkl,0}$ are the known phases for the initial state at $t = 0$ and

$$\Delta F(hkl,t) = |\mathbf{F}(hkl,t)| - |\mathbf{F}(hkl,0)|. \tag{2}$$

All crystallographic experiments deal with a statistically large number of molecules, since a typical crystal contains 10^{12}–10^{14} molecules. Consequently, the time-resolved experiment provides a space average: the variation with time of the average structure of all molecules in the crystal. It also provides a time average over the

duration of the key experimental feature that limits the time resolution. This could be the duration of the laser pulse used to initiate the structural reaction, typically nanoseconds to femtoseconds, or of the X-ray pulse used to monitor the evolution of the diffraction pattern, typically 100 ps at synchrotron X-ray sources, or for the time required for a diffusing substrate to reach an active site or for the photofragmentation reaction of a photolysed caged compound to take place. If the time resolution is substantially shorter than the lifetimes of the structural states being examined (Bourgeois et al., 2003; Schotte et al., 2003), the exact nature of this time average is of less experimental consequence; however, if the time resolution is on the same order as the lifetime of the structural states being examined, interpretation of the resultant difference electron density maps is made significantly more challenging.

Consider the simple, chemical kinetic mechanism involving four states I_0, I_1, I_2, and I_3 illustrated in Eq. (3):

$$I_0 \xrightarrow{k_{01}} I_1 \xrightarrow{k_{12}} I_2 \xrightarrow{k_{23}} I_3. \quad (3)$$

Given the rate coefficients k_{ij} for interconversion of these states, it is straightforward to calculate the fractional concentration of each state as a function of time, as presented in Fig. 1. The fractional concentration of I_0 is high in the earliest time points; measurements in the time range between 100 ns and 1 μs would reveal the (nearly) homogeneous structure of I_0. Likewise, the structure of the product I_3 could be determined from time points in the 50–100 ms range. With the rate coefficients as given, the I_1 state is most populated and 90% homogeneous at the 800 μs time point; the I_2 state, 80% homogeneous at the 2 ms time point. However, at nearly all intervening time points the average structure is heterogeneous and contains a significant population of two or more states. The real experimental problem is the inverse of that presented in this example: given the time-dependent data, obtain the mechanism(s) compatible with it, identify the number and nature of intermediate states and the rate coefficients, and refine the structure of each state. Thus, unscrambling or deconvoluting the time-dependent, heterogeneous mixture of structures at the data analysis stage is the essential challenge. The process of rising to this challenge may be denoted "analytical trapping," in contrast with the more widely used "chemical trapping" and "physical trapping" of candidate intermediate structures (Moffat, 2001; Moffat and Henderson, 1995; Schlichting and Chu, 2000; Stoddard, 1996).

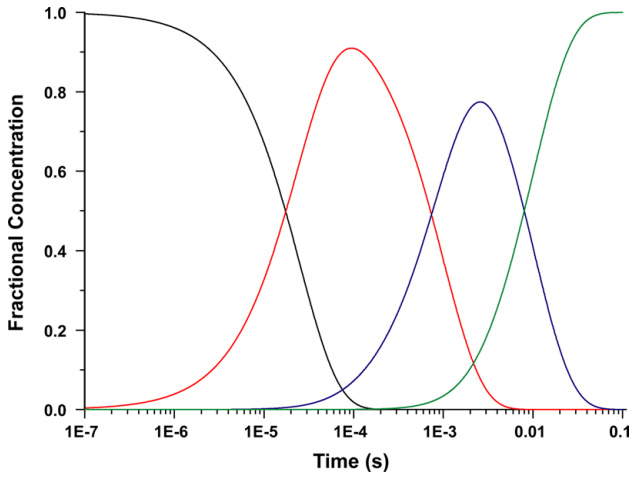

Fig. 1. Example of a simple irreversible, sequential chemical kinetic mechanism containing four intermediates. The time course of the population of each intermediate is plotted: I_0 (black), I_1 (red), I_2 (blue), and I_3 (green), using rate constants $k_{01} = 40\,000\,s^{-1}$, $k_{12} = 1000\,s^{-1}$, and $k_{23} = 100\,s^{-1}$ (see text, Eq. (3)).

1.2. The analytical problem

The experimental problem is made harder by weak signal and the presence of significant noise in the data. That is, the signal-to-noise ratio (S/N)[1] is poor. The signal is weak since the level of reaction initiation in the crystal is never 100% and may be as low as 10–20%. That is, only a fraction of the molecules react. Second, the structural changes may be limited in extent: even in those molecules that do react, most atoms do not move as the reaction progresses and those that do move, do not move far. However, those that do move are often clustered in space, e.g., around a chromophore or in an active site, and this proves to be a distinct advantage. High crystallographic resolution is nevertheless required to identify these small structural changes.

The noise contains both random and systematic contributions. Spatially and temporally random contributions arise from, e.g., error in measurement of the time-dependent structure factor amplitudes. Phase errors may produce spatially random errors that have no temporal component and are constant from time point to time point. Systematic contributions arise from the use of the difference Fourier approximation, and from the fact that an entire time series is typically pieced together from data in which each time point (or at most a few time points) is acquired on a different crystal. Crystal-to-crystal, or experiment-to-experiment, variation in the extent of reaction initiation is a significant source of systematic error (Rajagopal et al., 2004). Radiation damage, to which crystals at room temperature are very sensitive, also can be a source of systematic errors in many time-resolved experiments. Its effects are minimized by acquiring $|\mathbf{F}(hkl,t)|$ and $|\mathbf{F}(hkl,0)|$ on the same crystal at nearly the same time. Since both are affected by radiation damage to (very nearly) the same

[1] *Abbreviations used:* S/N, signal-to-noise ratio; SVs, singular vectors; lSVs, left singular vectors; rSVs, right singular vectors.

extent, the effects of damage on the resultant difference electron density map are minimized.

As experience is gained, the noise can be minimized by careful experimental design. For example, highly redundant measurements of $|\mathbf{F}(hkl,t)|$ are made containing 10–50 observations, from which precise values of both its mean and its standard deviation can be obtained (Anderson, 2003). Knowledge of the standard deviation permits appropriate weighting schemes to be applied which minimize error (Ursby and Bourgeois, 1997). Experiment-to-experiment errors can be minimized by collecting data in the four-dimensional data space *hkl*, *t* with time as the fast variable and hence acquiring only a subset of $|\mathbf{F}(hkl)|$ values on a single crystal.

2. Analytical trapping strategies

The final "raw" output of a time-resolved experiment is the time-dependent set of difference electron density (or electron density) maps spanning the desired time range. Such data can be examined and interpreted in a number of ways: (1) they can be assessed qualitatively or semi-quantitatively without the refinement of intermediate structures; (2) intermediate structures may be refined from a data set from a single time point or from multiple data sets closely spaced in time averaged together; or (3) a complete analysis of the time-resolved data can be performed, which includes an explicit analysis of the time domain along with the refinement of intermediate structures. Until recently, the first two strategies have been applied to most time-resolved crystallographic data, but with improved data quality and speed of data collection, it is now possible to collect sufficient high quality data to apply the third. After briefly reviewing studies that employed the first two approaches, we concentrate on two novel strategies, singular value decomposition and cluster analysis, which allow a determination of a protein's chemical kinetic mechanism by the third approach.

2.1. Qualitative and semi-quantitative approaches

In both of these approaches, positive and negative difference electron density features are directly associated with the movements of nearby atoms. In semi-quantitative approaches, further attempts to quantify these features are made and associated with different intermediate states. Srajer et al. (2001), Schotte et al. (2003), and Bourgeois et al. (2003) used this approach to interpret the results of their experiments on studies on the photolysis of the carbonmonoxy-myoglobin complex. Srajer et al. (2001) associated particular difference electron density features with the migration of carbon monoxide after its photolysis and with the relaxation of the heme and the surrounding protein environment from 1 ns to 1.9 ms. This approach allowed a quantification of the amount of CO bound at different xenon binding sites in the protein. Schotte et al. (2003) achieved extremely short time resolution around 100 ps in their experiment on the L29F myoglobin mutant and associated difference features in six time points over the time range from 100 ps to 3.16 μs to migration of CO around the heme prosthetic group. Bourgeois et al. (2003) used electron difference density map integration to identify motions in the distal E-helix and the CD-turn from 100 to 300 ns that lagged significantly behind the prompt local rearrangements around the heme; this analysis showed that there were at least two intermediates present in their data. In their analysis of 11 time points during the photocycle of the blue light photoreceptor, wild type photoactive yellow protein (WT PYP), spanning the time range from 1 ns to 1 min, Ren et al. (2001) associated difference features in the chromophore binding pocket with specific changes in the conformation of the 4-hydroxycinnamic acid chromophore. In all of these studies, interpretation of the difference electron density was semi-quantitative, thus limiting the conclusions that could be drawn from the data.

2.2. Single time point or adjacent time point averaging

In single time point or adjacent time point averaging, data are collected at a single time point or over a short time range when only one structural species is predicted to exist, based on other evidence such as visible absorption spectroscopy. 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. Using this approach, Genick et al. (1997) refined a millisecond intermediate of PYP from a decay of a photostationary state. Perman et al. (1998) attempted to refine a short-lived PYP intermediate using a similar strategy, but their results were compromised by the low S/N of the data set collected at a single 10 ns time delay (Ren et al., 2001). Two more recent studies have used this approach to refine short-lived intermediates with greater success. Bourgeois et al. (2003) refined a structure of a myoglobin mutant intermediate against data from a 316 ns time delay. Comparison of five different time points allowed them to follow CO through the protein, but also made it clear that more than one conformation may have been present in the crystal at any single time point. Anderson (2003; Anderson et al. 2004) took advantage of an averaging strategy in the analysis of 30 time points obtained during the photocycle of the E46Q mutant of PYP. By comparing differences between adjacent time-points, they identified and refined two structures corresponding to the early (tens of nanoseconds, red-shifted) and late (microseconds to milliseconds, blue-shifted) spectroscopic intermediates in the photocycle. The

assumption in all of these cases is that only one structure is present in the crystal during a given time range; if there are in fact multiple structures present, this strategy cannot be used effectively.

2.3. Singular value decomposition

One approach that allows both an explicit analysis of the time domain along with the refinement of structures is singular value decomposition (SVD), a technique that has been widely used in the analysis of spectroscopic data (Henry and Hofrichter, 1992). Singular value decomposition acts as a noise filter in which signal and noise are partitioned into different singular vectors, making it well-suited to typical time-resolved crystallographic data with a lower S/N compared to typical spectroscopic data (Schmidt et al., 2003). SVD also provides a reduced representation of the data, which greatly simplifies subsequent least-squares fits to chemical kinetic mechanisms. 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) and recently on real experimental data (Rajagopal et al., 2004; Schmidt et al., 2004).

2.3.1. Mathematical basis

For a detailed discussion, refer to Henry and Hofrichter (1992), who review the mathematical aspects of SVD and its application to spectroscopic data, and to Schmidt et al. (2003), who apply SVD to simulated time-resolved crystallographic data with different S/N levels. SVD is a general method that can be applied to any $m \times n$ matrix. In our case, it decomposes time-dependent data from a data matrix $\boldsymbol{A}$. Each column of the data matrix $\boldsymbol{A}$ contains m values of $\Delta\rho(r)$ spanning the crystallographic asymmetric unit, where m is typically 10 000–100 000; each of the n columns describes a time point t, where n is typically 15–30 (Schmidt et al., 2003). SVD decomposes $\boldsymbol{A}$ into three matrices according to the equation:

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

where $\boldsymbol{U}$ is an $m \times n$ matrix composed of the left singular vectors (lSVs); $\boldsymbol{S}$ is a diagonal $n \times n$ matrix composed of the singular values; and $\boldsymbol{V}$ is an $n \times n$ matrix 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 describes 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 intermediate and an individual rSV is a linear combination of the time-dependent concentrations of each intermediate. Two very useful properties arise from the decomposition. By allowing the data matrix $\boldsymbol{A}' = \boldsymbol{U}'\boldsymbol{S}'\boldsymbol{V}'^{\mathrm{T}}$ to be reconstructed with a subset (denoted by primes) of singular vectors (SVs) that contain the majority of signal while discarding those SVs that contain primarily noise, it acts as a noise filter. In crystallographic data, this also allows for phase improvement ("SVD-flattening") in a manner analogous to solvent flattening (Schmidt et al., 2003).

The second property is the reduced representation of the data matrix, which allows a much simpler fit to a chemical kinetic mechanism of the rSVs by reducing its dimensionality (Henry and Hofrichter, 1992). A candidate chemical kinetic mechanism with intermediate concentrations, $\boldsymbol{C}$, is fit to the rSVs by a set of linear parameters, $\boldsymbol{P}$:

$$\boldsymbol{V}' = \boldsymbol{C}\boldsymbol{P}.$$

This fit then allows a determination of the difference electron density corresponding to each intermediate, $\boldsymbol{F}$, by a straightforward transformation. Now

$$\boldsymbol{A}' = \boldsymbol{U}'\boldsymbol{S}'\boldsymbol{V}'^{\mathrm{T}},$$

$$\boldsymbol{U}'\boldsymbol{S}'(\boldsymbol{C}\boldsymbol{P})^{\mathrm{T}} = \boldsymbol{U}'\boldsymbol{S}'\boldsymbol{P}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}},$$

and since $\boldsymbol{A}' = \boldsymbol{F}\boldsymbol{C}^{\mathrm{T}}$ (Henry and Hofrichter, 1992),

$$\boldsymbol{F} = \boldsymbol{U}'\boldsymbol{S}'\boldsymbol{P}^{\mathrm{T}}.$$

The final steps in the determination of mechanism are the refinement of intermediate structures from this difference density $\boldsymbol{F}$, for example by difference refinement (Terwilliger and Berendzen, 1995), and a comparison of the calculated difference densities from the structures with several candidate mechanisms to the experimental data by a procedure called "posterior analysis" (Schmidt et al., 2003).

2.3.2. Application to simulated data

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. We will apply this procedure to one of the simulated data sets (2σ /1σ noise level) from Schmidt et al. (2003), using the mechanism from Fig. 1. Data preparation includes all steps between experimental data collection and the analysis of the data by SVD, such as the weighting of maps (Ursby and Bourgeois, 1997) and SVD-flattening. 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 may be subjected to rotation, a procedure 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. Then a subset of the

SVs must be chosen to reconstruct the data matrix $A' \sim A$. The results of applying this to our simulated data are shown in Fig. 2, where singular values are plotted against the autocorrelation of the rSVs. Based on this plot, we conclude that rotation is unnecessary and after visual inspection of the SVs, we select SVs 1–3 as being significant. The number of significant SVs gives a lower bound on the number of states present in the system (Henry and Hofrichter, 1992). Maps reconstructed with SVs 1–3 retain the majority of signal in the data while excluding the majority of noise (Fig. 3). As the lSVs and rSVs are linear combinations of the structures and concentrations of the underlying species, respectively, a fit of the rSVs with a sum of exponentials can be attempted (Henry and Hofrichter, 1992). If this fit is successful, a chemical kinetic mechanism may hold. The number of exponentials then corresponds to the minimum number of intermediate states present in the system and the exponents are relaxation times. These rSVs along with their corresponding lSVs are shown in Fig. 4. Both the lSVs and rSVs for SVs 1–3 have features consistent with signal: the lSVs have considerable amounts of spatially clustered features near atoms in the chromophore binding pocket (Figs. 4A–C) while the rSVs are smooth, well-behaved and can be fit by a sum of exponentials (Figs. 4E–G) (Schmidt et al., 2003). This is not true of for example SV 9, whose lSV has difference features randomly distributed through the protein (Fig. 4D) and whose rSV fluctuates markedly about 0 (Fig. 4H). 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).

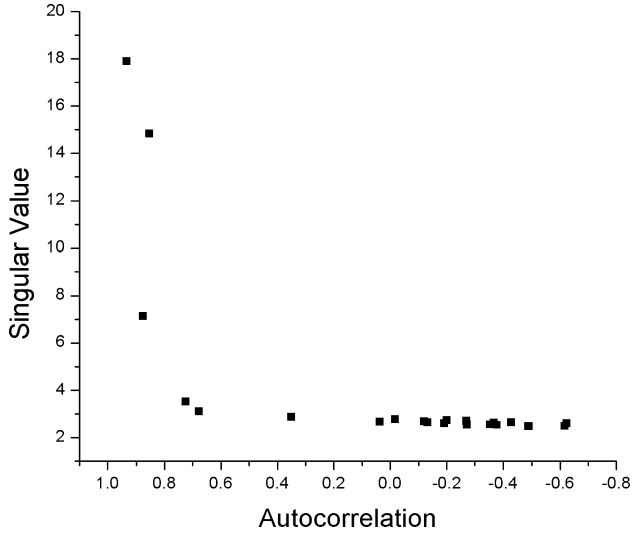

Fig. 2. Magnitude of singular values versus autocorrelations of corresponding right singular vectors after SVD of $2\sigma/1\sigma$ simulated data (Schmidt et al., 2003). Singular vectors that contain authentic signal should both have a high magnitude of singular value and autocorrelation, as such signal should be stronger than small random noise components and should vary smoothly.

2.3.3. Application to real data

Only recently has SVD been applied to real data. Here, we highlight two recent studies that have applied SVD to data spanning the photocycles of WT PYP (Schmidt et al., 2004) and its E46Q mutant (Rajagopal et al., 2004).

Schmidt et al. (2004) applied SVD to the analysis of 15 Laue data sets from 5 μs to 100 ms of the WT PYP photocycle after its initiation with a nanosecond laser pulse. After identification of four significant SVs, the rSVs were first fit with a sum of three exponentials and then with a number of candidate mechanisms. These fits allowed refinement of three time-independent intermediate structures, the first of which corresponded to a red-shifted chemical species and the latter two to blue-shifted chemical species. With these three structures in hand, a plausible chemical kinetic mechanism was identified. This work was the first to demonstrate that a nearly complete description of a chemical kinetic mechanism, in this case of the photocycle of PYP, could be extracted from experimental time-resolved crystallographic data. The strengths of the SVD analysis are clear: identification of a plausible mechanism and refinement of intermediate structures. Some of the weaknesses are also evident: an inability to differentiate similar mechanisms from one another and difficulty in refining the structures of intermediates when multiple chemical species may be present with similar time courses. These weaknesses are due primarily to the relatively high level of random and systematic noise in these data, and not to inherent limitations in the SVD procedure.

The limits of SVD's applicability to real data were further tested by the work of Rajagopal et al. (2004), who applied SVD to a very large set of data consisting of 30 time points from 10 ns to 100 ms during the photocycle of the E46Q mutant of PYP (Anderson, 2003; Anderson et al., 2004). Challenges not present in the SVD analysis of Schmidt et al. (2004) included data collected over a wider time range, a larger number of data sets (30 compared to 15), and with it, high levels of random and systematic errors relative to the signal present in the data. To deal with these problems, Rajagopal et al. (2004) provided a number of methods to limit user bias in SVD analysis of the data, such as spatial clustering of pixels in the lSVs (Fig. 3 of Rajagopal et al., 2004). By a careful application of SVD, the data could be represented by five singular values whose right singular vectors could be well fit by four relaxation times. These constraints help in the determination of a plausible chemical kinetic mechanism of the E46Q PYP photocycle (Rajagopal et al., in preparation).

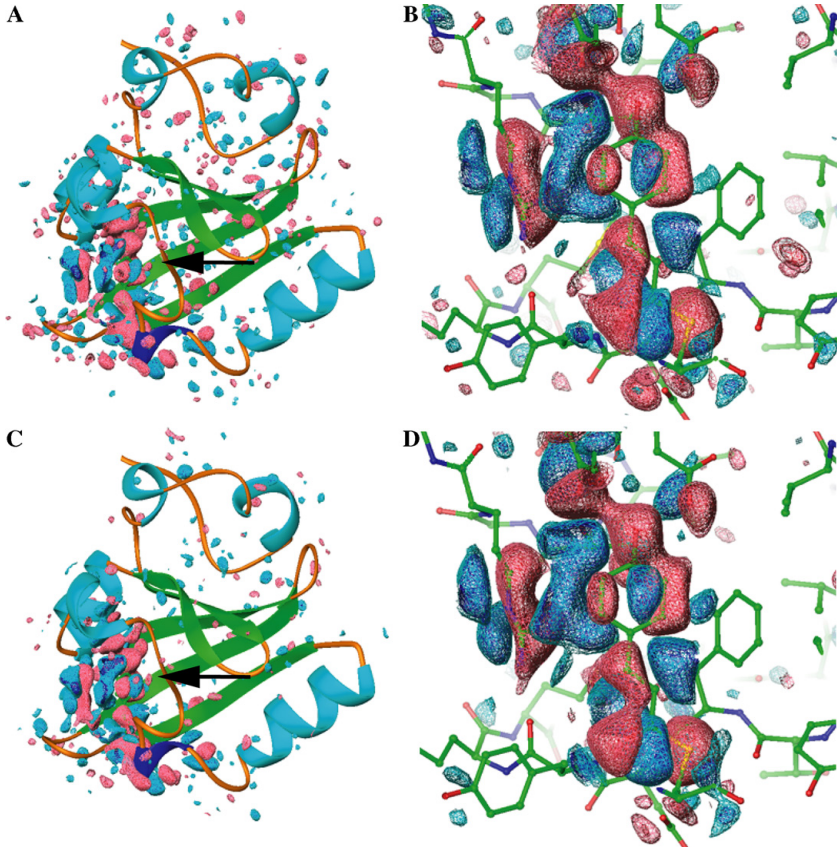

Fig. 3. Maps reconstituted with all singular vectors (A and B) and with three significant singular vectors (C and D). Maps reconstituted with the three significant SVs retain significant signal in the chromophore binding pocket (B and D; arrows in A and C), with much lower levels of noise in the surrounding protein (A and C). Maps are contoured at -4σ (red), -3σ (red), $+3\sigma$ (cyan), and $+4\sigma$ (blue), where σ is the root-mean-square value of the difference electron density across the asymmetric unit.

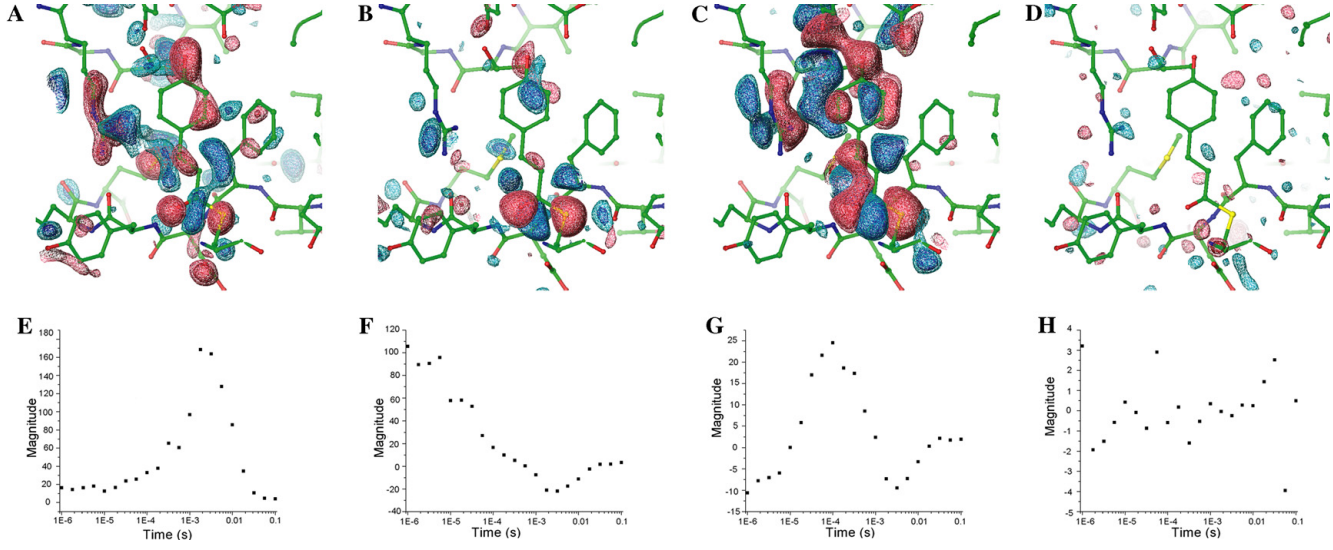

Fig. 4. Left and right singular vectors after SVD of $2\sigma/1\sigma$ simulated data (Schmidt et al., 2003). Left singular vectors for SVs 1–3 (A–C) and 9 (D). Right singular vectors for SVs 1–3 (E–G) and 9 (H). Maps are contoured at -4σ (red), -3σ (red), $+3\sigma$ (cyan), and $+4\sigma$ (blue).

2.4. Cluster analysis

A quite different approach to identifying the time-independent structural intermediates that are present in the time-dependent difference electron density maps is to group or cluster together the pixels (r) in such maps so that the difference density in each group of pixels corresponds to one structural intermediate. If the difference

electron density at a given pixel is sensitive to the presence of a particular intermediate then its time evolution at this pixel will exhibit a shape which is similar to the concentration profile of that intermediate, such as those in Fig. 1. The objective is that, by grouping the pixels with respect to the shapes of their time evolution, we can infer the number of intermediates present, the rates of their inter-conversion, and thus refine the structure of each intermediate. The computational grouping of objects based on their similarity as defined by a suitable mathematical measure is referred to in statistics as cluster analysis (Jain et al., 1999). Here, we report some preliminary results of our work using this technique. The emphasis is on exploratory, rather than inferential, data analysis. We review the principles of cluster analysis and clustering algorithms, describe our progress to date, and outline some future directions.

Cluster analysis has been widely used in many areas of science, including analysis of gene microarrays (Sherlock, 2000), protein dynamics simulations (de Groot et al., 2001), and functional magnetic resonance imaging (fMRI) (Goutte et al., 1999). To our knowledge, this is the first attempt to apply this method in crystallography. For cluster analysis to be successful in the context of time-resolved X-ray crystallography it must: (1) be able to handle the large amount of data generated by time-resolved experiments (typically tens of thousands of pixels over up to 50 time points), (2) be robust with respect to the large amounts of noise present in these experiments, and (3) be able to accurately identify pixels at which the difference electron density is sensitive to only one intermediate as well as those sensitive to more than one intermediate.

The application of cluster analysis to a given problem requires several decisions. First, is there any initial transformation of the raw data that can be applied in order to better distinguish the groupings that are sought? Second, what is the appropriate mathematical description of similarity between the data? And third, once a similarity measure is chosen, what kind of clustering algorithm should be used?

To evaluate the potential of cluster analysis to the analysis of time-resolved crystallography experiments we first use the simulated data for PYP at different noise levels introduced above. The data consist of time series of difference electron density maps, generated as described in Schmidt et al. (2003) and subjected to SVD analysis. Inspection of these maps shows that most pixels at which signal is present contain entirely positive or entirely negative values of the difference electron density throughout the time course: sign changes in the signal are very rare. If a pixel is to exhibit both negative and positive signals during the time course, its position must be occupied in the dark state as well as later in the photocycle with even higher occupancy, which is unlikely. (Sign changes due to noise do occur in the time course where signal is low or absent, for example late in the time course.) We therefore simplify the analysis by taking the absolute value of such maps in order to reduce the total possible number of clusters.

2.4.1. Similarity measures

Of primary importance in clustering is the choice of a metric to quantify whether the time profiles of two pixels in such maps are similar to each other. In this regard, it is useful to consider the values of the electron density for a given pixel at different times as a series of coordinates that define a vector. Among the most commonly used similarity measures are the Euclidean distance, the Pearson correlation coefficient, and Kendall's Tau. Each of these measures has advantages and drawbacks. The Euclidean distance reflects the distance between two points in space, which in this case are defined by two electron density vectors. The Euclidean distance is therefore sensitive to both the direction and the magnitude of the vectors. In contrast, the Pearson correlation coefficient (i.e., the dot product of two normalized vectors) captures the similarity in shape, but places no emphasis on the magnitude of the two series of measurements. Thus, we intuitively expect that this measure will be best suited to capture the similarities in the time evolution of the individual pixels. Kendall's Tau is a measure of correlation based on the tendency of the two pixels to vary in the same direction (i.e., to increase or decrease) from one time point to the next. However, Kendall's Tau does not perform so well in the presence of strong noise. The Euclidean distance and Pearson correlation coefficient do not take into account that the data are ordered in time, while Kendall's Tau does take this ordering into account in a very crude way.

2.4.2. Clustering algorithms

Once a similarity measure is chosen a variety of algorithms exist for clustering the data. It is not the purpose of this paper to provide a comprehensive survey of all such methods so only a brief review will be given. One large group of clustering methods used extensively in analysis of gene microarrays (Eisen et al., 1998) is hierarchical, representing a bottom up approach in which single data profiles are joined to form nodes, which are then joined further. The process continues until all individual profiles and nodes have been joined to form a single hierarchical tree. The branch lengths of the tree represent the degree of similarity between the objects. Such hierarchical clustering methods are not appropriate in the analysis of difference electron density maps as we are not interested so much in the relationships between pairs of pixels as in grouping the pixels into broad classes corresponding to one or more of the structural intermediates. Hence, clustering methods that partition the data into reasonably homogeneous groups are more appropriate for our analysis.

One widely used method to partition data is *k*-means clustering (Jain and Dubes, 1988). In *k*-means clustering the number of partitions is chosen in advance. Each partition has a reference vector, which is initialized randomly. Each electron density vector is then partitioned to its most similar reference vector. Next, each reference vector is recalculated as the average of all electron density vectors assigned to it. These steps are repeated until convergence, that is, until all electron density vectors map to the same partition on consecutive iterations. It should be noted that *k*-means clustering is nondeterministic due to the random initialization and therefore different *k*-means runs on the same data can and do produce different outcomes. In practice, however, *k*-means partitioning is reasonably robust.

An important consideration in *k*-means clustering is how to choose the number of partitions. Often guesswork is needed to decide how many significant patterns there may be in the data. This appears to be less of a problem when using *k*-means clustering in the analysis of time-resolved crystallography data since the available experimental evidence usually places restrictions on the number of possible intermediates, for example, from the number of relaxation times present in the rSVs of an SVD analysis, and this identifies the number of partitions necessary in the absence of noise. In the analysis of the simulated data presented here, the number of intermediates is actually known in advance.

2.4.3. Application to simulated data

We first apply *k*-means clustering to the simulated data set from Schmidt et al. (2003) with no noise. A sequential kinetic mechanism is used with three intermediates that have simulated concentration profiles as shown in Fig. 1. We have performed *k*-means clustering using several similarity measures and varying the number of partitions. For illustration, the results of one such *k*-means clustering run using the Pearson correlation similarity measure on simulated data based on the mechanism shown in Fig. 1 (with no noise added) are depicted in Fig. 5. The calculations were done using the program MeV distributed by The Institute for Genomic Research (Saeed et al., 2003) with the number of partitions set at nine. The data are partitioned into clusters which can clearly be associated with the time course followed by the concentration of the each of the three

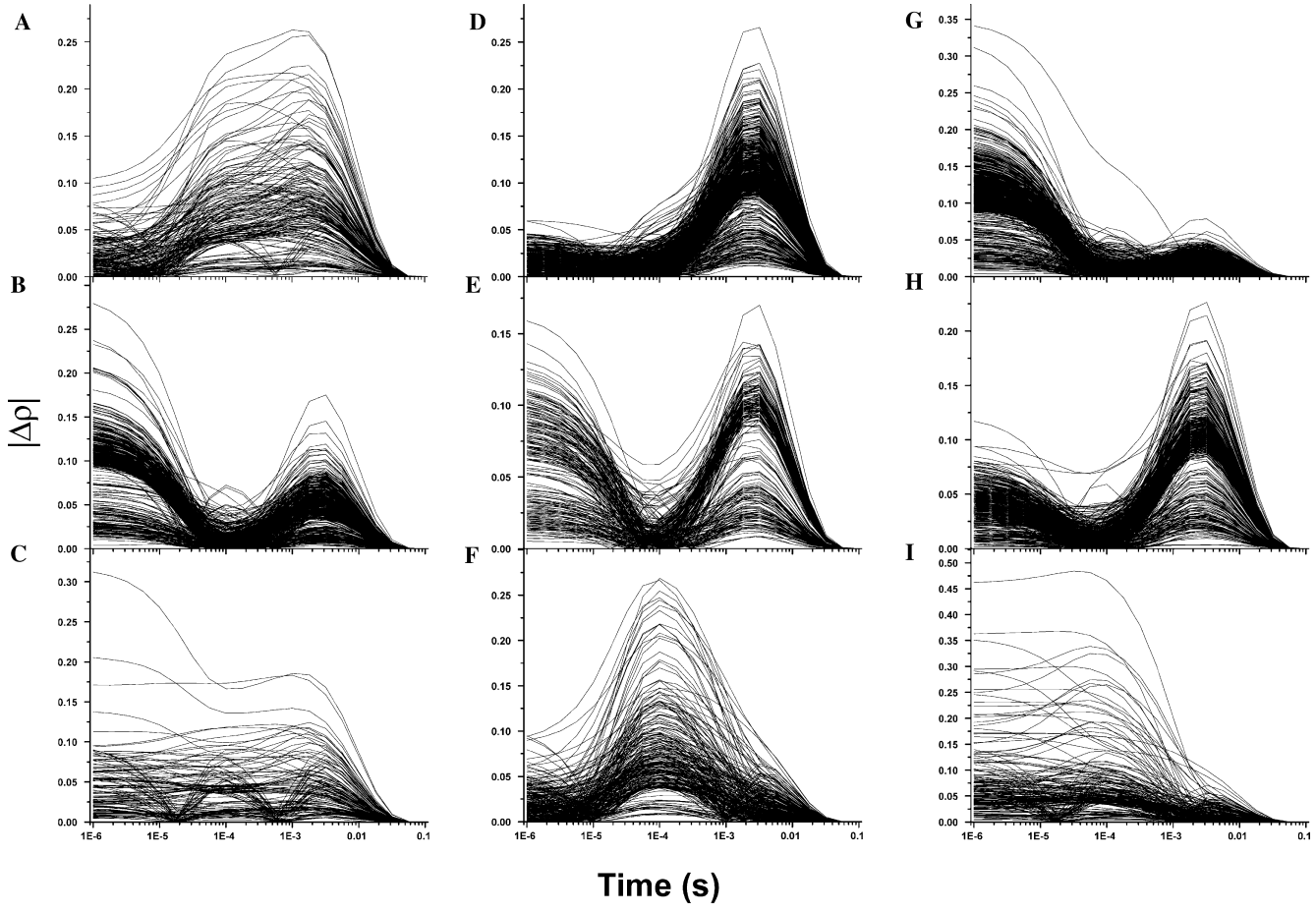

Fig. 5. Time courses of electron density at pixels in nine clusters (A–I) obtained by the *k*-means algorithm applied to simulated data with no added noise generated from the mechanism shown in Fig. 1 (Schmidt et al., 2003).

intermediates, I_0, I_1, and I_2, as well as with combinations of two or all three of the intermediates. Plotting the electron density features associated with each of these clusters (Fig. 6) shows that they are spatially contiguous as would be chemically expected. (Note that the intermediate I_3 is not visualized since it is identical here to the ground state and vanishes in difference electron density maps.)

These preliminary results demonstrate that cluster analysis holds promise in distinguishing the time-independent intermediates present in difference electron-density maps. The next step is to identify the contributions of each intermediate to those pixels that are sensitive to more than one intermediate (e.g., the cluster in Fig. 5A contains I_1 and I_2; that in Fig. 5E contains I_0 and I_2). In addition, to access the applicability of cluster analysis to real data we are currently analyzing the simulated data set in the presence of different levels of noise.

We draw several conclusions that determine our future research directions. First, none of the similarity measures currently in use in the literature appear to be fully satisfactory for the clustering of time-resolved data. For example, the existing measures do not take into account the temporal ordering, which is one of the critical features of our data. Moreover, while metrics such as Pearson correlation form clusters on the basis of shape similarity, when the difference density at a pixel depends on the presence of more than one intermediate, distinct patterns of shapes result that are similar to the eye but are mistakenly grouped into different clusters.

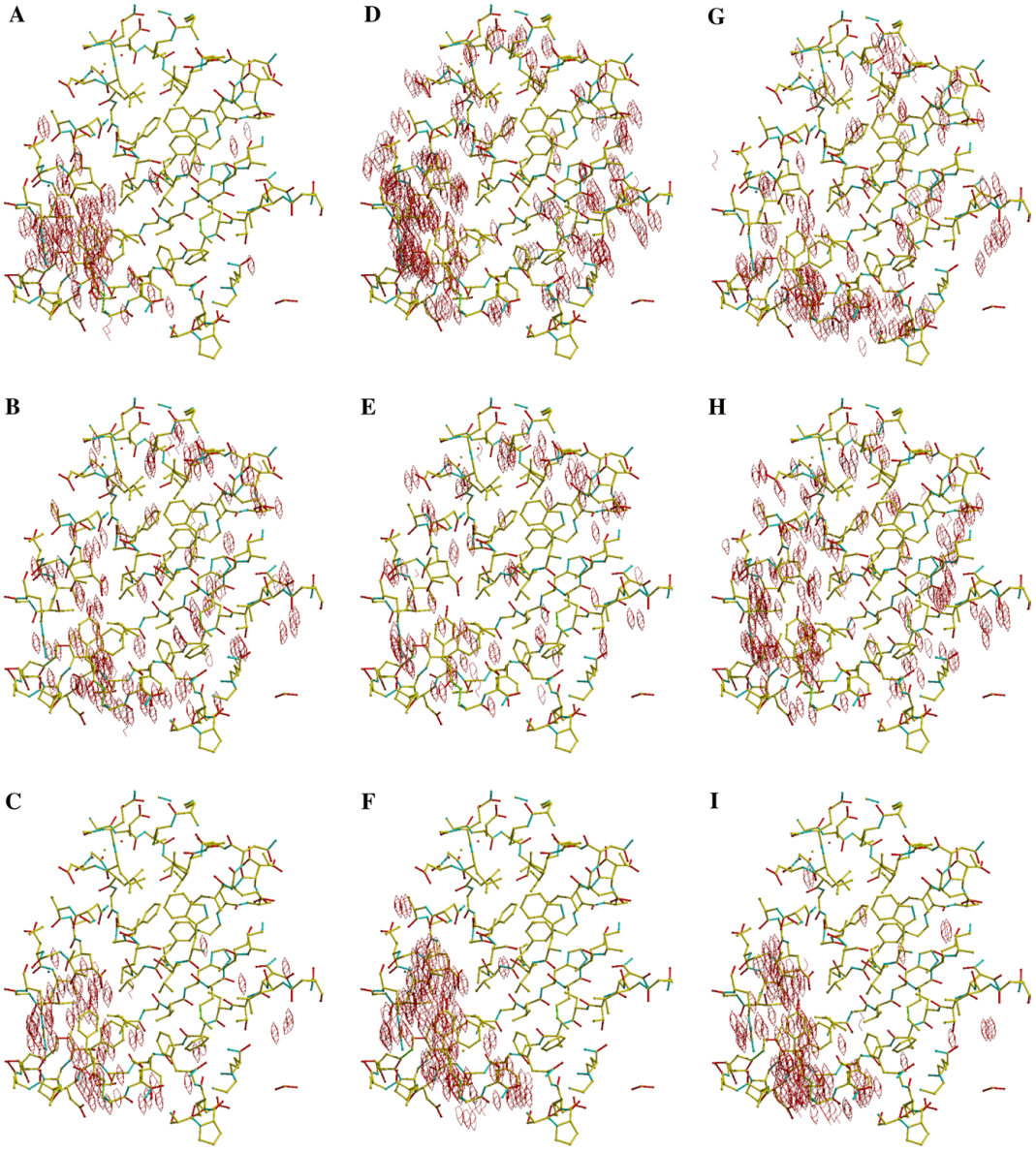

Fig. 6. Spatial distributions of the pixels associated with the nine clusters (A–I) whose time courses are shown in Fig. 5. The chromophore binding pocket of PYP is in the lower left.

For example, the clusters in Figs. 5B, E, and H all contain contributions from I_0 and I_2, but their relative magnitudes differ. Thus, a more sophisticated similarity measure is necessary in order to correctly group such shapes together. One measure that we are evaluating is to use as a basis of similarity the difference in the slopes between consecutive time points in the data.

Second, even when the exact number of intermediates is known it is not always obvious how many partitions to specify in order to display the underlying structure in the data. Due to the imperfect similarity measures and complexity of the data if the pre-defined cluster number is set too low, pixels that should belong to the same cluster are frequently grouped into two or more separate clusters at the expense of pixels that should belong in a separate cluster. Thus, it appears that the number of partitions specified in *k*-means clustering should be overdetermined by trial and error, something that would be difficult to do when the number of intermediates is not known in advance. It would be desirable to go beyond simple *k*-means clustering and use a method in which the number of partitions is not specified in advance. A novel clustering technique proposed by Heyer et al. (1999) requires no such pre-definition of the number of clusters. In this quality clustering algorithm, the diameter of a cluster is defined as the lowest pairwise correlation between any of the electron density vectors that lie within that cluster. A candidate cluster is formed by taking the first electron density vector and adding to it the electron density vector that minimizes the increase in cluster diameter. This process is continued until no further electron density vectors can be added to the cluster without exceeding a pre-defined diameter threshold. A second candidate cluster is formed by starting with the second electron density vector and repeating the process. All electron density vectors are available to the second cluster. This procedure is repeated for each electron density vector so that in the end there are as many clusters as there are electron density vectors. Then, the largest of these clusters is selected and retained and the electron density vectors that it contains are removed from consideration. The remaining electron density vectors are subjected to the same steps and another cluster is identified. The process is repeated until a suitable termination criterion is met, for example that each cluster must be at least of a certain size. This technique has several useful features: it is deterministic; it avoids the problem of specifying the number of clusters in advance; and it allows for the existence of electron density time profiles that do not belong to any cluster. However, instead of setting the number of clusters an arbitrary cluster diameter must be specified. We are currently evaluating the suitability of this technique to the clustering of time-resolved data.

3. Conclusions

As larger and more complex systems are studied by time-resolved crystallography (Baxter et al., 2004; Helliwell et al., 1998), more sophisticated analytical trapping strategies will be important in data interpretation. The above discussion highlights the major limit in the application of SVD and cluster analysis to time-resolved crystallographic data as the quality of the data itself. Most difficulties in the analysis arise from high levels of systematic (rather than random) noise relative to signal present in the data. This noise, however, can be minimized by changes to the data collection strategy (Ren et al., 1999) that have now been applied to a number of time-resolved systems. With data of somewhat higher quality, the application of SVD and cluster analysis to time-resolved data will be simplified considerably (S. Rajagopal, K. Moffat et al., unpublished), perhaps making a determination of chemical kinetic mechanism from such data as routine as the refinement of crystal structures from static monochromatic data.

Acknowledgments

We thank Marius Schmidt for generation of mock data used in this paper and Dan Nicolae and Jason Key for valuable discussions. Supported by NIH Grant GM 36452 to K.M.

References

Anderson, S., 2003. Structural changes in the E46Q mutant of PYP. PhD thesis, University of Chicago, USA.

Anderson, S., Srajer, V., Pahl, R., Rajagopal, S., Schotte, F., Anfinrud, P., Wulff, M., Moffat, K., 2004. Chromophore conformation and the evolution of tertiary structural changes in photoactive yellow protein, Structure, in press.

Baxter, R.H., Pomarenko, N., Srajer, V., Pahl, R., Moffat, K., Norris, J.R., 2004. Time-resolved crystallographic studies of light-induced structural changes in the photosynthetic reaction center. Proc. Natl. Acad. Sci. USA 101, 5982–5987.

Bourgeois, D., Ursby, T., Wulff, M., Pradervand, C., Legrand, A., Schildkamp, W., Laboure, S., Srajer, V., Teng, T.Y., Roth, M., Moffat, K., 1996. Feasibility and realization of single-pulse laue diffraction on macromolecular crystals at ESRF. J. Synchrotron Radiat. 3, 65–74.

Bourgeois, D., Vallone, B., Schotte, F., Arcovito, A., Miele, A.E., Sciara, G., Wulff, M., Anfinrud, P., Brunori, M., 2003. Complex landscape of protein structural dynamics unveiled by nanosecond Laue crystallography. Proc. Natl. Acad. Sci. USA 100, 8704–8709.

Cruickshank, D.W.J., Helliwell, J.R., Johnson, L.N., Royal Society (Great Britain), 1992. Time-resolved Macromolecular Crystallography. Oxford University Press, Oxford, 334 p. ([332], leaves of plates).

de Groot, B.L., Daura, X., Mark, A.E., Grubmuller, H., 2001. Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. J. Mol. Biol. 309, 299–313.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863–14868.

Genick, U.K., Borgstahl, G.E., Ng, K., Ren, Z., Pradervand, C., Burke, P.M., Srajer, V., Teng, T.Y., Schildkamp, W., McRee, D.E., Moffat, K., Getzoff, E.D., 1997. Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. Science 275, 1471–1475.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F., Hansen, L.K., 1999. On clustering fMRI time series. Neuroimage 9, 298–310.

Helliwell, J.R., Nieh, Y.P., Raftery, J., Cassetta, A., Habash, J., Carr, P.D., Ursby, T., Wulff, M., Thompson, A.W., Niemann, A.C., Hadener, A., 1998. Time-resolved structures of hydroxymethylbilane synthase (Lys59Gln mutant) as it is loaded with substrate in the crystal determined by Laue diffraction. Faraday Trans. 94, 2615–2622.

Henry, E., Hofrichter, J., 1992. Singular value decomposition: application to analysis of experimental data. Methods Enzymol. 210, 129–192.

Heyer, L.J., Kruglyak, S., Yooseph, S., 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9, 1106–1115.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, New Jersey.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comp. Surv. 31, 264–323.

Moffat, K., 1989. Time-resolved macromolecular crystallography. Annu. Rev. Biophys. Biophys. Chem. 18, 309–323.

Moffat, K., 2001. Time-resolved biochemical crystallography: a mechanistic perspective. Chem. Rev. 101, 1569–1581.

Moffat, K., Henderson, R., 1995. Freeze trapping of reaction intermediates. Curr. Opin. Struct. Biol. 5, 656–663.

Perman, B., Srajer, V., Ren, Z., Teng, T., Pradervand, C., Ursby, T., Bourgeois, D., Schotte, F., Wulff, M., Kort, R., Hellingwerf, K., Moffat, K., 1998. Energy transduction on the nanosecond time scale: early structural events in a xanthopsin photocycle. Science 279, 1946–1950.

Rajagopal, S., Anderson, S., Schmidt, M., Moffat, K., 2004. Analysis of experimental time-resolved crystallographic data by singular value decomposition. Acta Cryst. D 60, 860–871.

Rajagopal, S., Anderson, S., Srajer, V., Schmidt, M., Pahl, R., Moffat, K., in preparation. A Structural Pathway for Signaling in the E46Q mutant of Photoactive Yellow Protein.

Ren, Z., Bourgeois, D., Helliwell, J., Moffat, K., Srajer, V., Stoddard, B., 1999. Laue crystallography: coming of age. J. Synchrotron Radiat. 6, 891–917.

Ren, Z., Perman, B., Srajer, V., Teng, T.Y., Pradervand, C., Bourgeois, D., Schotte, F., Ursby, T., Kort, R., Wulff, M., Moffat, K., 2001. A molecular movie at 1.8 Å resolution displays the photocycle of photoactive yellow protein, a eubacterial blue-light receptor, from nanoseconds to seconds. Biochemistry 40, 13788–13801.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., Quackenbush, J., 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34, 374–378.

Schlichting, I., Chu, K., 2000. Trapping intermediates in the crystal: ligand binding to myoglobin. Curr. Opin. Struct. Biol. 10, 744–752.

Schmidt, M., Pahl, R., Srajer, V., Anderson, S., Ren, Z., Ihee, H., Rajagopal, S., Moffat, K., 2004. Protein kinetics: Structures of intermediates and reaction mechanism from time-resolved X-ray data. Proc. Natl. Acad. Sci. USA 101, 4799–4804.

Schmidt, M., Rajagopal, S., Ren, Z., Moffat, K., 2003. Application of singular value decomposition to the analysis of time-resolved macromolecular X-ray data. Biophys. J. 84, 2112–2129.

Schotte, F., Lim, M., Jackson, T.A., Smirnov, A.V., Soman, J., Olson, J.S., Phillips Jr., G.N., Wulff, M., Anfinrud, P.A., 2003. Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. Science 300, 1944–1947.

Sherlock, G., 2000. Analysis of large-scale gene expression data. Curr. Opin. Immunol. 12, 201–205.

Srajer, V., Ren, Z., Teng, T.Y., Schmidt, M., Ursby, T., Bourgeois, D., Pradervand, C., Schildkamp, W., Wulff, M., Moffat, K., 2001. Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved Laue X-ray diffraction. Biochemistry 40, 13802–13815.

Srajer, V., Teng, T., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M., Moffat, K., 1996. Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. Science 274, 1726–1729.

Stoddard, B.L., 1996. Caught in a chemical trap. Nat. Struct. Biol. 3, 907–909.

Terwilliger, T.C., Berendzen, J., 1995. Difference refinement—obtaining differences between two related structures. Acta Cryst. D 51, 609–618.

Ursby, T., Bourgeois, D., 1997. Improved estimation of structure-factor difference amplitudes from poorly accurate data. Acta Cryst. A 53, 564–575.