Essential Degrees of Freedom of Proteins

ANDREA AMADEI, ANTONIUS B. M. LINSSEN, BERT L. DE GROOT and HERMAN J. C. BERENDSEN
Department of Biophysical Chemistry and BIOSON Research Institute, the University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

Abstract. Analysis of extended molecular dynamics (MD) simulations of several proteins in aquous solutions reveals that it is possible to separate the configurational space into two subspaces: (1) an 'essential' subspace containing only a few degrees of freedom in which anharmonic motion occurs that comprises most of the positional fluctuations; and (2) the remaining space in which the motion has a narrow Gaussian distribution and which can be considered as 'physically constrained'.

Key words. Molecular dynamics, essential degrees of freedom of proteins, lysozyme, haloalkane dehalogenase, HPr.

1. Introduction

Functional proteins are generally stable mechanical constructs that allow certain types of internal motions to enable their biological function. Functional internal motions may be subtle and involve complex correlations between atomic motions, but their nature is inherent in the structure and interactions within the molecule. It is a challenge to derive such motions from the molecular structure and interactions, to identify their functional role, and to reduce the complex protein dynamics to its essential degrees of freedom.

In 1993 we developed a new approach and methodology to identify an 'essential' subspace in a protein configurational space [1]. In this method we analyse the correlations between atomic positional fluctuations and diagonalize the covariance matrix of atomic displacements. We find that most of the positional fluctuations are concentrated in correlated motions in a subspace of only a few degrees of freedom (not more than 1%), while all other degrees of freedom represent much less important, basically independent, Gaussian fluctuations orthogonal to the essential subspace. In our treatment we do not include atomic masses and we do not require any harmonicity for the potential energy. We showed [1] that, if in the configurational space ideal or approximate linear constraints for the motions are present, we are always able, with this method, to identify them and define the complementary subspace (essential subspace) where all the relevant motions should be concentrated. We also do not require a complete equilibration of the system. It is sufficient to have a trajectory that is extended enough to equilibrate the near-constraint subspace and to produce motions in the essential subspace that are large compared to the near constraints fluctuations. Since the near-constraints coordinates have rather short relaxation times and fluctuation ranges, this is usually accomplished within a reasonable simulation time (trajectories within 1 ns).

In this paper we show, in a more general way, the meaning of diagonalizing the covariance matrix. We show the basic results obtained up to now and we describe the new developments we are working on.

2. Theory

In an earlier publication [1] we described the theory which forms the basis of the essential degrees of freedom method. The standard procedure for analysis is as follows: First a simulation of several hundreds of picoseconds is performed. From the trajectory obtained, the overall translational and rotational motion is eliminated. Next a covariance matrix of the atomic displacements with respect to the average positions is built:

$$\mathbf{C} = \mathrm{cov}(\Delta\mathbf{x}) = \langle \Delta\mathbf{x} \otimes \Delta\mathbf{x} \rangle \tag{1}$$

Where $\otimes$ denotes a tensor product and $\Delta\mathbf{x}$ is given as:

$$\Delta\mathbf{x} = \mathbf{x} - \langle \mathbf{x} \rangle \tag{2}$$

By diagonalizing $\mathbf{C}$ we obtain a set of eigenvectors (we choose to sort the eigenvectors in order of decreasing eigenvalue) of which the first few define a subspace in which all the essential motions of the protein appear to occur. The eigenvalues are mean square displacements along the corresponding eigenvector. We now will describe a more general definition of the meaning of the eigenvectors of the covariance matrix $\mathbf{C}$. We will show that the first n eigenvectors define the best fitted hyperplane through the density of the probability distribution in configurational space. This implies that through the density in this space no set of orthonormal vectors can be defined in such a way that one of them has a positional fluctuation exeeding that of the first eigenvector. This means that the first eigenvalue is the maximum possible positional fluctuation along a single direction which was actually sampled. Alternatively, the total positional fluctuation in the $(N-1)$-dimensional plane (where N is the total number of dimensions of the system) orthogonal to this eigenvector is the minimum possible one. Figure 1 illustrates this for a two-dimensional case.

In general, in an N-dimensional space, the subspace orthogonal to the first n eigenvectors has the minimum possible positional fluctuation in respect of the positional fluctuations of any other $N-n$-dimensional subspace. We define the positional fluctuation in an $N-n$-dimensional subspace as:

$$\sigma^2_{N-n} = \sum_{\zeta=n+1}^{N} \langle [(\mathbf{x} - \langle \mathbf{x} \rangle) \cdot \boldsymbol{\nu}_\zeta]^2 \rangle \tag{3}$$

where $\{\boldsymbol{\nu}_\zeta\}$ is any set of N orthornormal vectors.

If we can prove that the vector along which the smallest positional fluctuation is observed, is the last eigenvector, then this can recursively be applied to the complementary $N-1$ dimensional subspace, and so on. The positional fluctuation of $\boldsymbol{\nu}_N$ is:

$$\sigma^2_1 = \langle [(\mathbf{x} - \langle \mathbf{x} \rangle) \cdot \boldsymbol{\nu}_N]^2 \rangle \tag{4}$$

$\boldsymbol{\nu}_N$ can always be written as a linear combination of eigenvectors $\boldsymbol{\eta}_i$:

$$\boldsymbol{\nu}_N = \sum_{i=1}^{N} \alpha_{iN} \boldsymbol{\eta}_i \tag{5}$$

Now $\sigma^2_{(N-1)}$ can be written as:

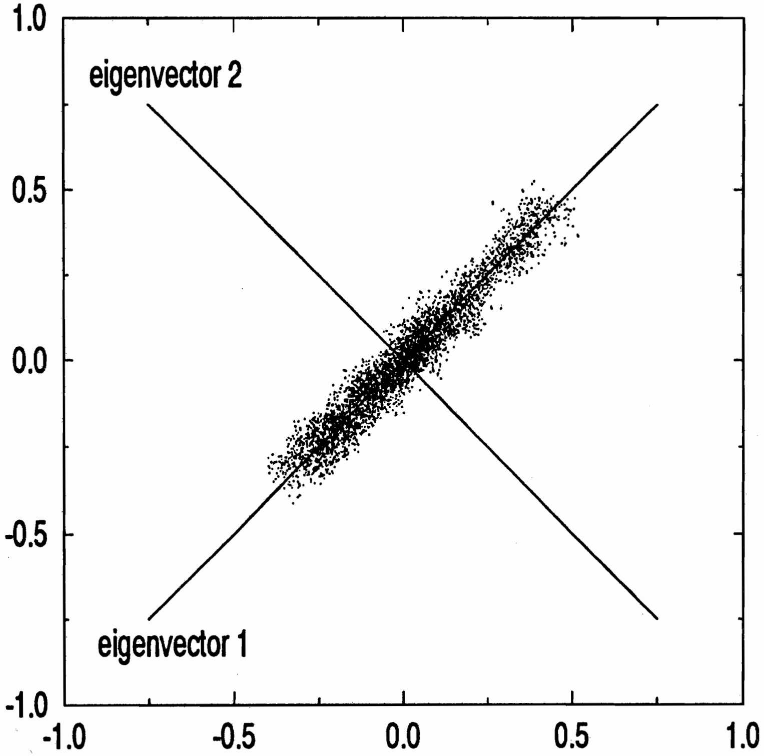

Fig. 1. A two dimensional example of essential dynamics analysis.

$$\sigma_1^2 = \left\langle \left[\sum_{i=1}^{N} (\mathbf{x} - \langle \mathbf{x} \rangle) \cdot \alpha_{iN} \boldsymbol{\eta}_i \right]^2 \right\rangle \tag{6}$$

or

$$\sigma_1^2 = \left\langle \left[\sum_{l=1}^{N} \sum_{l'=1}^{N} [\alpha_{lN} \alpha_{l'N} \{(\mathbf{x} - \langle \mathbf{x} \rangle) \cdot \boldsymbol{\eta}_l\} \{(\mathbf{x} - \langle \mathbf{x} \rangle) \cdot \boldsymbol{\eta}_{l'}\} \right] \right\rangle \tag{7}$$

Since

$$\langle (\Delta \mathbf{x} \cdot \boldsymbol{\eta}_l)(\Delta \mathbf{x} \cdot \boldsymbol{\eta}_{l'}) \rangle = 0 \quad \forall l \neq l' \tag{8}$$

and

$$\langle (\Delta \mathbf{x} \cdot \boldsymbol{\eta}_i)^2 \rangle = \lambda_i \quad (\lambda_i \text{ is the eigenvalue corresponding to } \eta_i) \tag{9}$$

this can be written as:

$$\sigma_1^2 = \sum_{i=1}^{N} \alpha_{iN}^2 \lambda_i \tag{10}$$

We can now write the following inequality:

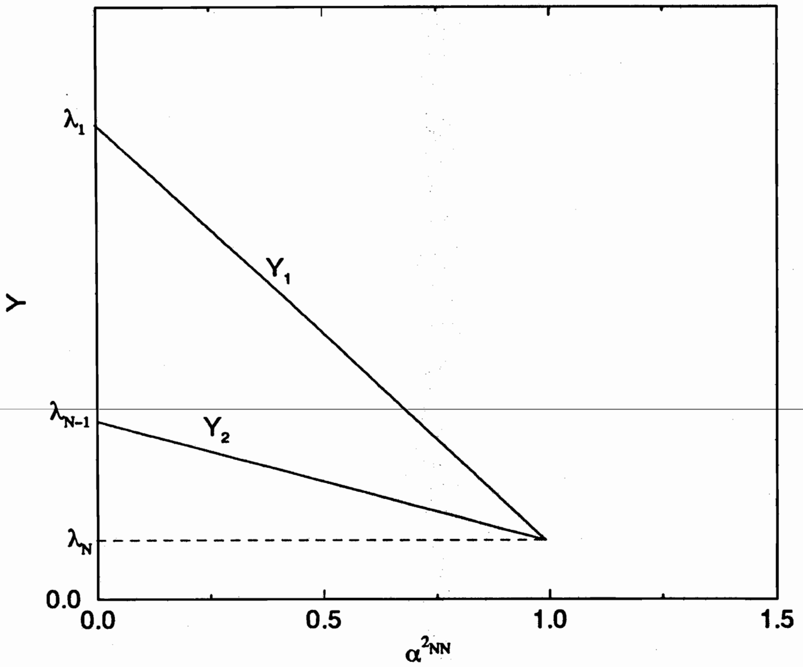

Fig. 2. $\sigma^2_{(N-1)}$ lies between y_1 and y_2.

$$\alpha^2_{NN}\lambda_N + \lambda_1 \sum_{i=1}^{(N-1)} \alpha^2_{iN} \geqslant \sigma^2_1 \geqslant \alpha^2_{NN}\lambda_N + \lambda_{(N-1)} \sum_{i=1}^{(N-1)} \alpha^2_{iN} \tag{11}$$

And since

$$\sum_{i=1}^{N} \alpha^2_{iN} = 1 \tag{12}$$

and

$$\sum_{i=1}^{(N-1)} \alpha^2_{iN} = 1 - \alpha^2_{NN} \tag{13}$$

this can be rewritten as

$$\sigma^2_{NN}\lambda_N + \lambda_1(1 - \alpha^2_{NN}) \geqslant \sigma^2_1 \geqslant \alpha^2_{NN}\lambda_N + \lambda_{(N-1)}(1 - \alpha^2_{NN}) \tag{14}$$

Since we want to find a minimum for σ^2_1 and its value lies between two boundaries, we must determine the minimal values of these boundaries. We can plot both boundaries as a function of α^2_{NN} (Figure 2) using the following properties:

$$\lambda_i \geqslant 0 \tag{15}$$

with $i = 1, 2, \ldots, N$

$$\lambda_{(N-1)} > \lambda_N \tag{16}$$

$$0 \leqslant \alpha_{NN}^2 \leqslant 1 \tag{17}$$

The upper boundary (as a function of α_{NN}^2) looks as follows:

$$y_1(\alpha_{NN}^2) = \alpha_{NN}^2(\lambda_N - \lambda_1) + \lambda_1 \tag{18}$$

and the lower boundary:

$$y_2(\alpha_{NN}^2) = \alpha_{NN}^2(\lambda_N - \lambda_{(N-1)}) + \lambda_{(N-1)} \tag{19}$$

As can be seen from Figure 2 σ_1^2 has a minimum for $\alpha_{NN}^2 = 1$, which means that

$$\boldsymbol{\nu}_N = \boldsymbol{\eta}_N \tag{20}$$

and

$$\sigma_1^2 = \lambda_N \tag{21}$$

This means that the vector for which a minimum mean square positional fluctuation is found, is the last eigenvector. This minimum value is exactly the eigenvalue corresponding to that eigenvector.

3. Results and Discussion

Up to now every protein to which our method was applied showed a very low dimensional essential subspace (which was always defined by less then ten eigenvectors). The kind of structural transitions involved in the essential motions are always connected to possible biological behaviour of the proteins. In lysozyme [1] we found a motion that opened the active site cavity of the enzyme described by the first eigenvector with all the other 'essential' ones involving smaller motions also in the active site region. In thermolysin (manuscript in preparation, in collaboration with D. van Aalten) we found two essential eigenvectors producing clear hinge bending motions between the two domains that enclosed the active site. This motion was already postulated on the basis of crystallographic data. Analysis of haloalkane dehalogenase (manuscript in preparation) revealed a possible entrance or exit of substrate or products. It is interesting to note that the motion along the first eigenvector also involved a tunnel that, on the basis of crystallographic data [2] was supposed to be the entrance as well as the exit. Both tunnels opened and closed in an anticorrelated way. Also the essential eigenvectors of HPr (manuscript in preparation, in collaboration with N. van Nuland, R. Scheek and G. Robillard) of which several structures based on NMR data were available [3] showed a hinge bending (opening and closing) motion between two α-helices enclosing the biological active region of the protein. In Figure 3 two superimposed structures from two (most distant) positions along the first eigenvector of HPr are shown. The hinge bending motion involving both helices can clearly be seen. Figure 4 shows the C_α-eigenvalue curves from three proteins, lysozyme, dehalogenase and HPr. It is evident that after the first few eigenvalues the remaining ones represent negligible motions for all three proteins. We found [1] that the essential eigenvectors derived from the all atoms covariance matrix produced backbone

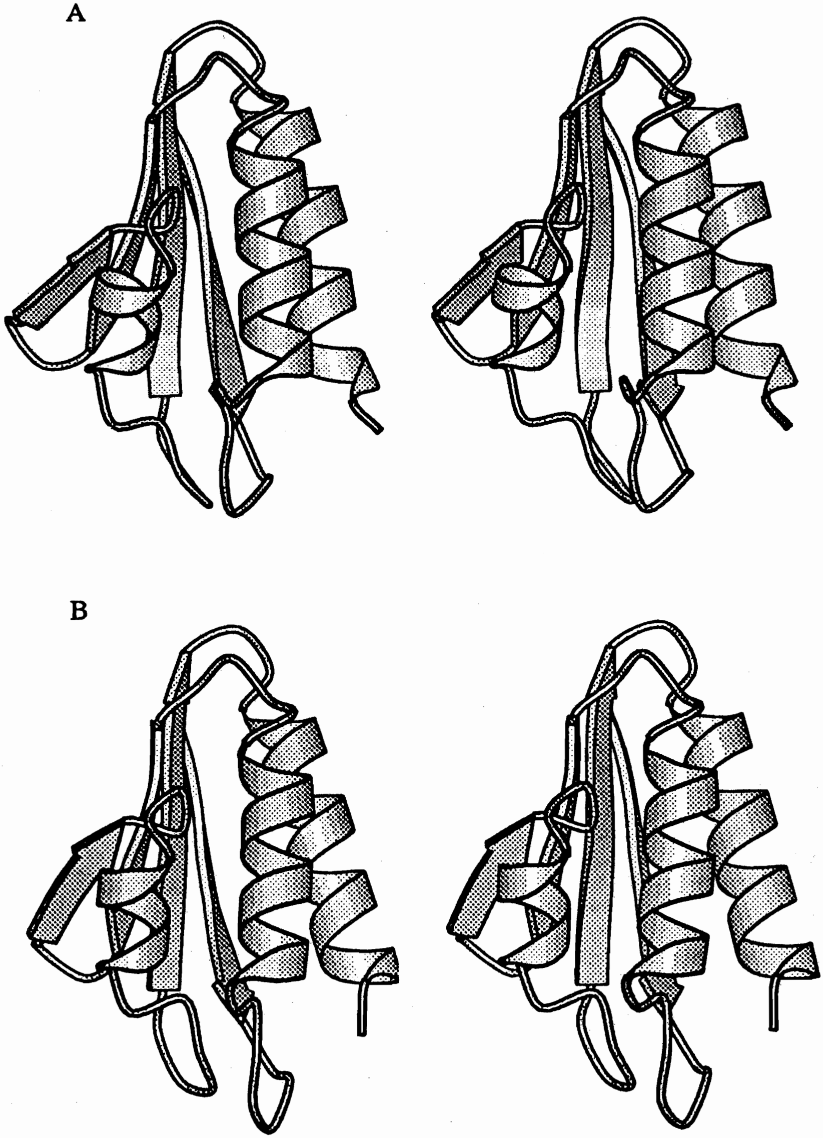

Fig. 3. Stereoview of the closed structure (A) and the open structure (B) of HPr derived from the displacement along the first eigenvector.

motions that were identical to the motions obtained from the essential C_α-eigenvectors.

In Figure 5 we show the projectons of three HPr trajectories (starting from three different NMR-structures) on the three planes defined by the first three C_α-eigenvectors, and one plane defined by two near-constraints eigenvectors (vectors 20 and 50). In the three 'essential planes' the density is far from being equilibrated but the possibility for large displacements is already evident. In fact, the

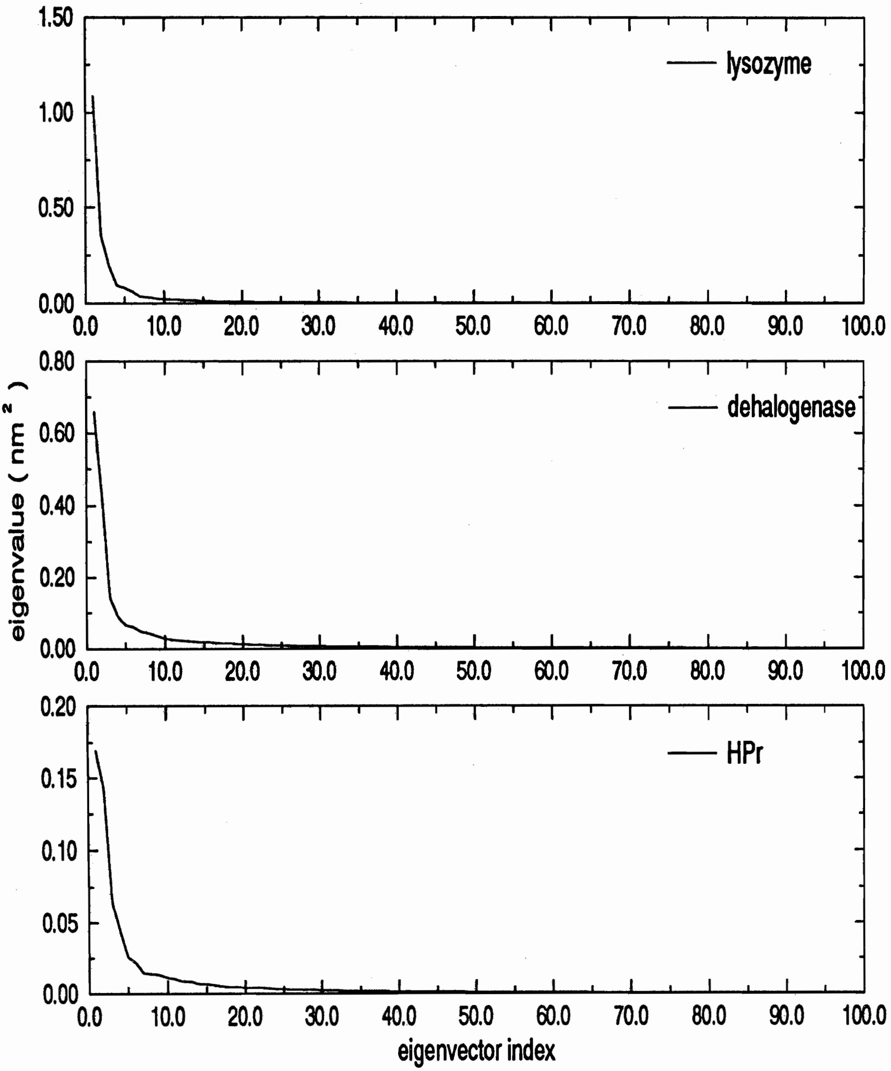

Fig. 4. Eigenvalues of lysozyme, dehalogenase and HPr of C_α-analysis (see text).

trajectories on these essential planes span very different regions (although always in contact). On the contrary, in the near constraints plane the three trajectories fully overlap with the same very narrow region, indicating a full equilibration and a stable near-constraint behaviour.

4. Conclusions

The *essential degrees of feedom* method has, up till now, proved to be useful for detecting protein dynamical behaviour that is related to biological functions. This

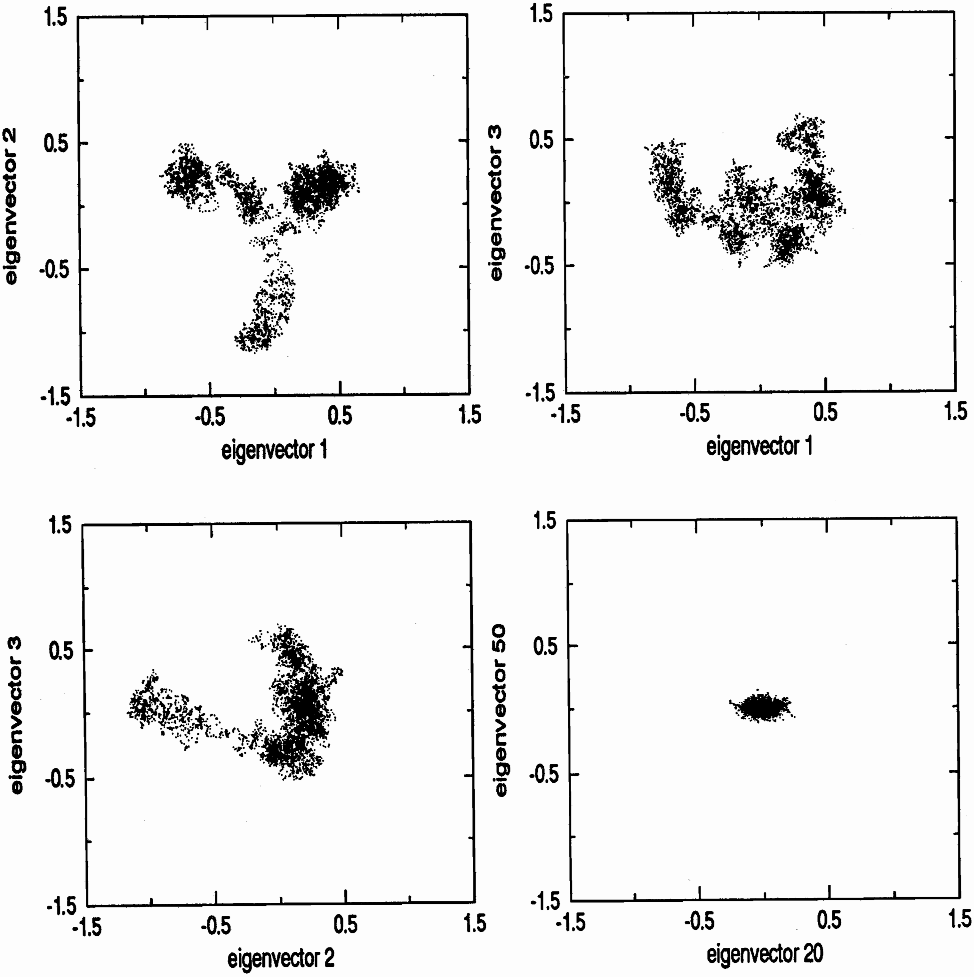

Fig. 5. Three trajectories of HPr projected on four different planes (see text).

of different (but related) proteins. At this moment our main goal is the investigation of the essential subspace of HPr. The geometrical definition of such subspace gives us the capability to sample this subspace in a very efficient way. This can be accomplished by either moving the system along one of the essential eigenvectors or using more complex pathways that are not necessarily linear. If a sufficient sampling can be obtained this should produce all the configurations that are involved in the biological activity. This gives the possibility to predict several properties of the protein and also can give insight into the consequences of mutations.

Acknowledgements

We would like to thank Dr. A. Mark from ETH in Zürich who gave us a 1 ns trajectory of lysozyme, N. van Nuland who provided us with several Molecular Dynamics trajectories of HPr and Dr. K. Verschueren who was so kind to supply us with crystal structure coordinates of dehalogenase with and without substrate. The work concerning HPr was performed in close collaboration with Dr. R. Scheek and Prof. G. Robillard. Dr. A. Amadei was financially supported by 'Istituto Pasteur Fondazione Cenci Bolognetti'. Visualization of HPr was performed with MOLSCRIPT [4].

References

1. Andrea Amadei, Antonius B. M. Linesen, and Herman J. C. Berendsen: *PROTEINS: Structure, Function, and Genetics* **17**, 412–425 (1993).
2. Sybille M. Franken, Henriette J. Rozeboom, Kor H. Kalk, and Bauke W. Dijkstra: *EMBO J.* **10**, 1297–1302 (1991).
3. Nico A. J. van Nuland, Ilona W. Hangyi, René C. van Schaik, Herman J. C. Berendsen, Wilfred F. van Gunsteren, Ruud M. Scheek, and George T. Robillard: *J. Mol. Biol.* **237**, 544–559 (1994).
4. P. J. Kraulis: *J. Appl. Crystalogr.* **24**, 946–950 (1991).