Protein Engineering, Design & Selection vol. 18 no. 3 pp. 127–137, 2005
Published online April 8, 2005 doi:10.1093/protein/gzi017

Conservation and specialization in PAS domain dynamics

A.Pandini and L.Bonati[1]

Dipartimento di Scienze dell'Ambiente e del Territorio, Università degli Studi di Milano-Bicocca, Piazza della Scienza 1, 20126 Milan, Italy

[1]To whom correspondence should be addressed.
E-mail: laura.bonati@unimib.it

The PAS (Per-ARNT-Sim) superfamily is presented as a well-suited study case to demonstrate how comparison of functional motions among distant homologous proteins with conserved fold characteristics may give insight into their functional specialization. Based on the importance of structural flexibility of the receptive structures in anticipating the signal-induced conformational changes of these sensory systems, the dynamics of these structures were analysed. Molecular dynamics was proved to be an effective method to obtain a reliable picture of the dynamics of the crystal structures of HERG, phy3, PYP and FixL, provided that an extensive conformational space sampling is performed. Other reliable sources of dynamic information were the ensembles of NMR structures of hPASK, HIF-2α and PYP. Essential dynamics analysis was successfully employed to extract the relevant information from the sampled conformational spaces. Comparison of motion patterns in the essential subspaces, based on the structural alignment, allowed identification of the specialized region in each domain. This appears to be evolved in the superfamily by following a specific trend, that also suggests the presence of a limited number of general solutions adopted by the PAS domains to sense external signals. These findings may give insight into unknown mechanisms of PAS domains and guide further experimental studies.

Keywords: essential motions/molecular dynamics/protein structures

Introduction

The PAS (Per-ARNT-Sim) superfamily of proteins is a widely studied collection of single- and multidomain proteins involved in inducing and regulating some of the basic adaptive mechanisms of the cell (Taylor and Zhulin, 1999; Gu *et al*., 2000; Kewley *et al*., 2004). Their ability to sense changes in the environmental variables (light, redox potentials, ligands) and trigger appropriate responses relies on the efficiency and ductility of a modular unit called the PAS domain. This is an independent domain of ~100 amino acids, suitable for transmitting the signal from the receptive site to other domains or partners through proper conformational changes. Despite the low sequence similarity of PAS domains, a high conservation in the fold and topology of the known structures suggests a strong evolutionary conservation of some functional features. The 'winning' solution developed by this domain seems to include both ductility in single responses and conservation of a stable, reliable and modular fold for signal transmission.

The PAS domains for which structures were first determined were the N-terminal domain of the human potassium channel, HERG (Cabral *et al*., 1998); the LOV domain of the phototropin module of the fern photoreceptor, phy3 (Crosson and Moffat, 2001); the bacterial photoactive yellow protein, PYP (Borgstahl *et al*., 1995; Dux *et al*., 1998; Getzoff *et al*., 2003); and the heme binding domain of the bacterial oxygen-sensor, FixL (Gong *et al*., 1998; Miyatake *et al*., 2000). These domains are characterized by a highly conserved α/β fold: a five-stranded β-sheet hanged on a long helical connector and a bulge of three small helices (Figure 1). In the different proteins the hanging helix is displaced from the sheet in different ways, designing a cavity suitable for arranging, in three of the four cases, different kinds of cofactors. Whereas HERG does not take any cofactor, phy3 non-covalently binds the flavin mononucleotide (FMN) in the interior of the cavity, PYP covalently binds the chromophore 4-hydroxycinnamic acid to Cys69 (lying on α3) and the heme pocket of FixL contains a pentacoordinate heme iron with the vicinal His200 (on the helical connector) as fifth ligand. Additional extra-domain elements that were detected in these structures are an N-terminal bundle of two helices, known as 'helical lariat', in PYP and long helices that lie at the C- or N-terminus in the FixL crystal structures of different species (Gong *et al*., 1998; Miyatake *et al*., 2000).

Subsequently, other PAS structures derived by NMR spectroscopy improved the structural knowledge of the PAS fold: the N-terminal PAS domain of human PAS kinase, hPASK (Amezcua *et al*., 2002), and the C-terminal domain of human hypoxia-inducible factor 2α, HIF-2α (Erbel *et al*., 2003). The solution structure of hPASK PAS A is similar to that of the other known PAS domains, with a distinctive extended loop between a shortened helical connector and the β3 strand. The HIF-2α PAS B adopts the typical α/β fold with a high degree of structural similarity with the PAS domains so far described (Figure 1). Neither of these structures includes any cofactor molecules.

In recent years, significant efforts were also directed to rationalize the mechanisms employed by PAS domains to convert input stimuli into signals that propagate to downstream partners (Gong *et al*., 2000; Amezcua *et al*., 2002; Hao *et al*., 2002; Cusanovich and Meyer, 2003; Erbel *et al*., 2003; Harper *et al*., 2003; Hellingwerf *et al*., 2003).

One of the most intriguing suggestions emerging from these studies is that functionality of the PAS module is intrinsically dynamic and that flexibility of the receptive sites plays a central role in promoting significant pathways of conformational changes, subsequently transmitted to extra-domain units through suitable domain interfaces. However, until now extensive molecular dynamics (MD) simulations were carried out only to characterize the flexibility of PYP in its ground state and with the isomerized chromophore (van Aalten *et al*., 1998a, 2000; Groenhof *et al*., 2002a,b) and only a recent study

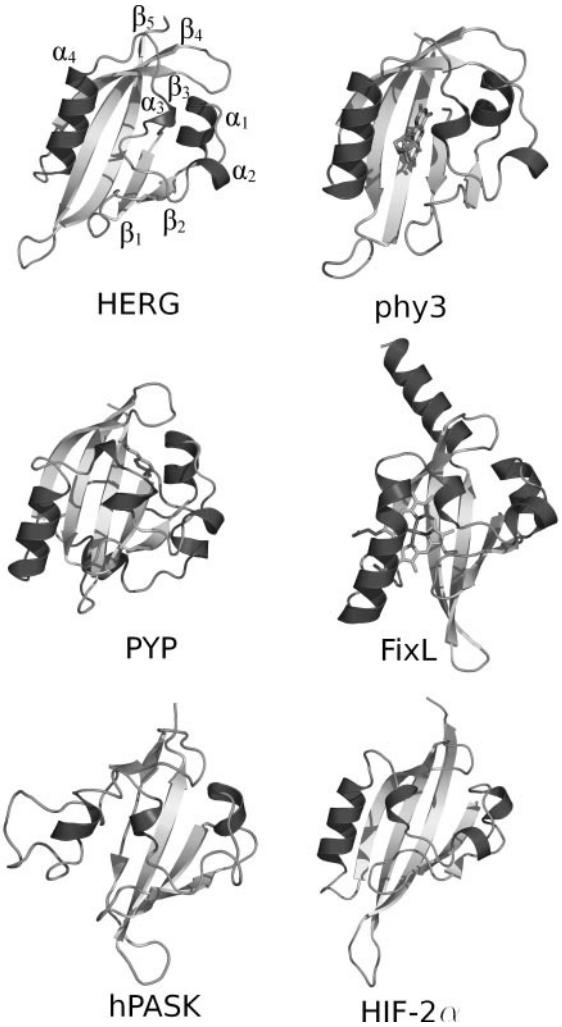

Fig. 1. Cartoon representation of the PAS domains included in the ED analysis, with their cofactors: HERG (PDB i.d.: 1BYW), phy3 (PDB i.d.: 1G28 chain A), PYP (PDB i.d.: 1NWZ), FixL (PDB i.d.: 1DRM), hPASK (PDB i.d.: 1LL8) and HIF-2α (PDB i.d.: 1P97). Secondary structure attribution according to the Kabsch and Sander method (Kabsch and Sander, 1983). The nomenclature adopted for the secondary structure elements is reported on the HERG structure. For clarity, in the comparisons all helices are called 'α' and different types are commented each time; in PYP, the region containing both the π and the 3_{10} helices (residues 62–70) is named α3.

addressed the topic of similarity in flexibilities across the PAS superfamily (Vreede *et al*., 2003).

The aim of this paper is to highlight how the comparison of structural flexibility among proteins belonging to the same superfamily may complement and complete sequence and structure comparisons to shed light on the specialization process that led each system to adapt the general superfamily features to a specific mechanism of action.

The PAS domains are an optimal study case for these purposes thanks to the above-highlighted characteristics of low sequence similarity, high structural conservation and intrinsically dynamic functionality. Moreover, given the rapid growth of experimental information on PAS domain-containing proteins and the large number of these proteins identified in genomes (Cusanovich and Meyer, 2003), many new PAS domains are expected to be described in the future. Therefore, the opportunity to gain insight into unknown functionalities by comparing both structural and dynamic features of the few known domains appears to be a very intriguing task. New crystal structures of cellular PAS domains have already been deposited during the development of this work (Kurokawa *et al*., 2004; Razeto *et al*., 2004).

With the outlined aim, in this work the dynamic properties of the initially available X-ray structures for the receptive states of PAS domains (HERG, phy3, PYP and FixL) were derived by MD simulations. Essential dynamics (ED) analysis on the trajectories obtained was performed to extract the majority of information related to functional motions. On the basis of evidence that collections of NMR structures are also suitable for direct employment in ED analysis (van Aalten *et al*., 1998b), the NMR ensembles of the structures of hPASK and HIF-2α were included in the investigation, along with those of the dark-state PYP (Dux *et al*., 1998), that allowed comparison among different sources of dynamic information. ED analysis was previously employed in the identification of intrinsic structure flexibility related to function, but there are few examples of comparisons of different protein families within a large superfamily (Grottesi and Sansom, 2003; Vreede *et al*., 2003). It is conceivable that the high signal to-noise ratio obtainable with this method for extraction of functional motion might ensure the reliability of comparison even when distant homologous proteins are investigated. The extension of ED for comparative purposes requires the choice of a reference framework to compare the dynamic data. Unlike previous studies (Vreede *et al*., 2003), where only DALI aligned substructures were included in the covariance analysis, here the proposed approach was to identify the motion patterns through independent derivation of the dynamic data for each domain and subsequently to align them. If the flexibility and motion ability of a protein are strictly related to its structure, it is therefore obvious to hypothesize that the structural alignment is a good starting point to compare motion patterns across the superfamily.

The resulting picture suggests a new general representation of the PAS domain as a ductile unit that developed three main solutions to signal reception and transmission, retained through evolution by means of conservation of specific structural and dynamic features.

Materials and methods

Investigation targets

Structures of the PAS domains were collected from the Protein Data Bank (Berman *et al*., 2000; http://www.rcsb.org/pdb/). The X-ray structures of HERG, phy3, PYP and FixL selected for MD simulations were those representing the dark state for the photoreceptors or the unbound state for FixL. When different options were available, deposition with the highest resolution was preferred. The following entries were employed:

1BYW for HERG potassium channel (Cabral *et al*., 1998);
1G28 (chain A) for phy3 phototropin (Crosson and Moffat, 2001);
1NWZ for PYP photoreceptor (Getzoff *et al*., 2003);
1DRM for FixL heme protein (Gong *et al*., 1998).

For the ED analysis on NMR ensembles of structures, the following entries were employed:

1LL8 for PAS kinase, hPASK (Amezcua *et al*., 2002), 20 structures;
1P97 for hypoxia-inducible factor, HIF-2α (Erbel *et al*., 2003), 26 structures;
3PHY for PYP photoreceptor (Dux *et al*., 1998), 26 structures.

Molecular dynamics simulations

Whole domain structures were employed in simulations of HERG, phy3 and PYP, whereas only the PAS fold unit was included for FixL. The excluded C-terminal helix (residues 257–270) shows uncertainty in its position, as recorded in the crystal structure: this flanking element could be located along the PAS domain in direct contact with the β-sheet while its location protruding outside the domain (see Figure 1) should be a partial artifact (Gong *et al*., 1998). Unpublished results on the simulation of the whole domain with the same protocol demonstrate that the exclusion of the C-terminal helix does not affect the locations and relative intensities of highly mobile regions, but enhances the convergence and the consequent reliability of the simulations.

All structures were solvated in SPC water (Berendsen *et al*., 1981) using cubic boxes and simulated with periodic boundary conditions. The dimensions of the box were set to allow at least 0.8 nm between protein and box faces on each side. Solvent was relaxed with 5 ps MD simulation, keeping protein degrees of freedom restrained. After addition of ions to neutralize the systems, a short minimization with steepest descent was performed up to convergence on maximum force lower than 1000 kJ/mol.nm. The resulting systems were employed as starting points for all simulations. These were performed with GROMACS 3.1.4 (Berendsen *et al*., 1995; Lindahl *et al*., 2001), by using the GROMOS96 43a2 version of the GROMOS force field as available in the GROMACS package. For the PYP cofactor (4-hydroxycinnamic acid), topology was derived from the crystal structure coordinates (Getzoff *et al*., 2003) using PRODRG (van Aalten *et al*., 1996), with GROMOS96 force field parameters and then linked to a standard cysteine topology. Quantum mechanically derived charges were employed (Groenhof *et al*., 2002a). The chromophoric environment was carefully checked to ensure that protonation states were correctly set: GLUH topology block instead of the GROMACS default deprotonated state (GLU) was chosen for Glu46.

Simulations were carried out in the NVT ensemble. Four replicas of 10 ns simulation were generated for each PAS domain structure. Each replica started with the same configuration as obtained through the system setup procedure; different random seeds were employed to generate different starting velocities from a Maxwellian distribution at 300 K.

For long-range electrostatic interactions, the particle mesh Ewald summation method (Darden *et al*., 1993) was employed to gain a more accurate description. Van der Waals interactions were described by a 6–12 Lennard–Jones potential with distance cutoff at 0.9 nm; neighbour lists were employed with a list cutoff of 0.9 nm and update frequency every 10 steps.

Protein and solvent were independently coupled with a thermal bath by a Berendsen thermostat at 300 K and a coupling period of 0.1 ps.

The internal degrees of freedom of water molecules were constrained by the SHAKE algorithm (Ryckaert *et al*., 1977) and all bond distances in the proteins were constrained by the LINCS algorithm (Hess *et al*., 1997). Additionally, the interacting site method (Feenstra and Berendsen, 1999) was employed to allow a wider time step of simulation. By introducing these choices, it was possible to increase the integration step up to 4 fs and still obtain fairly stable simulations.

During simulations, configurations and velocities were recorded every 1 ps, collecting 10 000 frames for each replica. All the analyses on trajectory data were performed with GROMACS by taking as reference the starting structures obtained by the system setup procedure. This allowed a consistent picture regardless of different starting resolutions of the crystal structures. For each protein, the equilibrated portions of the four replicas were joined to obtain a combined set representative of different directions of sampling around the starting structures.

Sampling reliability analysis

The efficiency and reliability of simulations were evaluated by different indexes: r.m.s.d. matrices, overlap of replicas on combined trajectories and cosine content of the first eigendirections.

R.m.s.d. values calculated between frames of the trajectories and reported as matrices were proved to be meaningful and synthetic in defining the qualitative extension of sampling and its efficiency.

The overlap between the conformational space spanned by different parts of the simulation was reported as a way to address the convergence of the single part to the overall sampled space (Hess, 2002).

The overlap between two matrices A and B, $s(A, B)$, is defined as

$$s(A,B) = 1 - \frac{d(A,B)}{\sqrt{trA + trB}}$$
$$d(A,B) = \sqrt{tr((A^{1/2} - B^{1/2})^2)} \qquad (1)$$

where tr is the trace of the matrix. When the overlap is 1 the two spanned subspaces are identical, whereas the value of 0 indicates their complete orthogonality. In this work, the overlap between the covariance matrix of each replica and that of the overall combined trajectory were reported for each protein as a function of time. The trend, extension and final value for each replica are clear indexes of the similarity between the sampled section and the complete collection. They can also guide in the definition of the simulation convergence.

The cosine content c_i of the principal component p_i, defined as

$$c_i = \frac{2}{T}\left(\int_0^T \cos(k\pi t) p_i(t) \mathrm{d}t\right)^2 \left(\int_0^T p_i^2(t)\,\mathrm{d}t\right)^{-1} \qquad (2)$$

where T is the time of simulation, is a negative index of the similarity between the dynamics of biological systems and the dynamics of a random diffusion process (Hess, 2000). c_i can take values between 0, no cosine, and 1, a perfect cosine. It was demonstrated that insufficient sampling can lead to behaviours that resemble a functional motion, but describing a random motion. When this happens, different results are obtained every time the simulation is replicated. In analysing this index, it is usually sufficient to evaluate the cosine contribution for the first eigendirection to have a reliable idea of the protein behaviour. Owing to the negative nature of this index, it is well used as a complement and confirmation of the previous two (Hess, 2002).

Essential dynamics analysis

ED analysis is a well-documented application of principal component analysis to MD data (Amadei *et al*., 1993). It is aimed to extract informative directions of motion in a multidimensional space, thus reducing the overall complexity of the simulation and isolating the important motion for the system.

ED application involves the following:

1. Construction of the covariance matrix of the positional fluctuations of atoms.
2. Diagonalization of the covariance matrix by an orthonormal transformation matrix *R*. The columns of the rotational matrix are the eigenvectors defining the principal modes; the relative eigenvalues are the variances of the data set in the new directions.
3. Projection of original data on the eigenvectors to generate the 3*N* principal components.
4. Evaluation of the amount of variance contained in the first eigenvectors allows one to separate the simulation space in two subspaces: the essential subspace and the constrained subspace. This leads to restriction of the description of the system to the more informative essential subspace.

As reported previously (Amadei *et al*., 1993), reduction of the analysis on the C_α atoms can lead to all the relevant information needed to separate the essential subspace and identify the important modes in the protein dynamics. Therefore, for the PAS proteins covariance analysis was performed only on C_α.

To compare essential motions in the PAS fold, only residues belonging to the PAS domain unit were included in the covariance analysis. To identify them, a DALI (Holm and Sander, 1996) structural alignment was derived and taken as reference. Additionally, for PYP the N-terminal helical lariat was included, because it is known to be essential for protein function (Rubinstenn *et al*., 1998; Craven *et al*., 2000).

Two criteria were employed to define the dimensionality of the essential subspace: the fraction of total motion described by the reduced subspace and the distribution of motion along the eigenvectors. The former, computed as the sum of eigenvalues for the included eigenvectors, describes the amount of variance retained by the reduced representation of the system space. The latter is evaluated by projection of motion on the single directions and calculation of the corresponding distributions of motion. Successful applications of ED demonstrated that functional motion often involves presence of different structures sampled by the system along the essential directions and these are visible as distinct peaks in the motion distribution.

Identification and comparison of motion patterns

Patterns of motion were identified on a residue basis for each PAS domain. A smoothed motion value derived from the root mean square fluctuation (r.m.s.f.) on the C_α position in the essential subspace was attributed to each residue. Smoothing was obtained by application of triangular smoothing window of five residues:

$$\text{r.m.s.f.}_{\cdot i}^{(\text{smoothed})} = \text{r.m.s.f.}_{\cdot i-2} + 2 \times \text{r.m.s.f.}_{\cdot i-1} + 3 \times \text{r.m.s.f.}_{\cdot i} + 2 \times \text{r.m.s.f.}_{\cdot i+1} + \text{r.m.s.f.}_{\cdot i+2}$$

This allowed to further increase the signal to noise ratio for all the data set.

Average values of r.m.s.f.$^{(\text{smoothed})}$ were derived for each sequence and those residues with r.m.s.f.$^{(\text{smoothed})}$ over the average were reported as exploiting interesting motion. Groups of contiguous residues with a significant value of motion define the higher mobility regions in a structure and are referred to as 'motion patterns' for the structure within the superfamily.

Comparative analysis of the motion across the PAS domain unit required the choice of a reference framework. For this purpose, a structural alignment was derived for the PAS domains. DALI (Holm and Sander, 1996) pairwise alignments were obtained for each pair of PAS domains, reporting the following *Z*-scores:

	HERG	phy3	PYP	FixL	hPASK	HIF-2α
HERG	–					
phy3	18.3	–				
PYP	10.0	10.1	–			
FixL	10.2	11.1	10.7	–		
hPASK	8.5	8.3	8.1	11.6	–	
HIF-2α	11.4	11.4	9.7	10.7	8.0	–

The phy3 PAS domain was chosen as reference to compose a multiple alignment because it maximizes the average *Z*-score against other PAS domains. As expected, all comparisons score far more than 4 and confirm the reported homology between this set of domains.

Graphs were generated using R (R Development Core Team, 2003). Molecular models images were generated using PyMol (DeLano, 2002).

Results and discussion

Molecular dynamics of the PAS domains

The X-ray receptive structures of the PAS domains of HERG, phy3, PYP and FixL were analysed by MD simulations. MD represents a powerful tool to acquire a picture of the intrinsic flexibility of these domains, provided that extensive sampling of the structure neighbourhood in the conformational space is performed. The computational cost of nanosecond simulations is fairly high, especially when the use of an explicit solvent and detailed long-range electrostatic descriptions are necessary to address the physical behaviour of the system correctly (see Materials and methods). A good compromise to obtain extensive sampling with reduced computational costs was to generate four replicas for the 10 ns simulation of each PAS domain, starting from the same structure but with different sets of atomic velocities. This constitutes the widest available collection of computer simulation data for these structures.

For the four replicas of each protein, average values and standard deviations for r.m.s.d. to the starting structure, radius of gyration and total energy of the system are reported in Table I. All the replicas showed a fairly stable trend with a short equilibration time of ~100–300 ps. For sake of consistency, the following covariance analysis was performed on a productive phase starting at 500 ps and encompassing 9.5 ns for each replica.

Visual inspection of the trajectories and the small changes observed in the radius of gyration highlighted that the domains did not undergo any kind of unfolding process during the simulation process. The overall trend indicates a small contraction, due to the rearrangement of the side chains of the exposed residues. Total energies of the systems are conserved after the short equilibration period, showing that the proteins had reached a condition of ergodicity. Stability is moreover confirmed by the efficiency of the bath coupling and by a constant value of density across the box during each trajectory (data not shown).

Sampling reliability

Recent work reported evidence that incorrect sampling can be misleading in analysing MD results, especially when studies

involve ED, where insufficient sampling can lead to complete misinterpretation of results, masking diffusive motion on random directions within the shape of essential motion (Hess, 2000, 2002). Therefore, a priority in PAS fold analysis is to verify the reliability of the MD analysis, by identifying and measuring the extension of sampling.

A simple but efficient index of sampling is the extension of conformation re-sampling in the phase space. The r.m.s.d. matrices for the combined trajectories of the four PAS structures (Figure 2) show similar ranges of values and a satisfactory sampling of few principal structures. A 10 ns time-scale and four replicas appear to be sufficient to identify a well-sampled main structure for all the domains. Results from cluster analysis on combined trajectories (data not shown) confirmed these findings, indicating that for all four proteins the largest cluster of structures comprises the great majority of the trajectory.

A quantitative confirmation of the sampling convergence is given by the overlap values (Hess, 2002) of increasing portions of the simulation with the overall 38 ns productive phase. These are reported in Figure 3 for the four replicas of each protein. None of the replicas completely overlaps the combined trajectory, despite the fact that the amount of similarity in the sampled information is fairly acceptable for all the proteins, ranging from 60–70% for HERG to 40–50% for PYP. Therefore, none of the PAS structures can be extensively described by a single 10 ns simulation. This suggests that the four replicas technique can guarantee better sampling with an acceptable setting up and computational time. The cosine content (Hess, 2000) for the first principal component derived from the covariance analysis for each simulation (see below) confirmed these

Table I. Average values and standard deviations for some properties in the four simulation replica of PAS domains

	R.m.s.d. (nm)		R_g (nm)		Total energy (kJ/mol)	
	Mean	SD	Mean	SD	Mean	SD
HERG-1	0.253	0.025	1.31	0.01	–260183	377
HERG-2	0.227	0.016	1.32	0.01	–260180	377
HERG-3	0.263	0.030	1.31	0.01	–259681	378
HERG-4	0.259	0.032	1.31	0.01	–259943	379
phy3-1	0.261	0.032	1.33	0.01	–270299	393
phy3-2	0.210	0.020	1.34	0.01	–270878	390
phy3-3	0.265	0.035	1.33	0.01	–270866	393
phy3-4	0.221	0.023	1.34	0.01	–270612	393
PYP-1	0.275	0.033	1.33	0.01	–189258	301
PYP-2	0.253	0.024	1.34	0.01	–189284	303
PYP-3	0.257	0.037	1.33	0.01	–189291	304
PYP-4	0.285	0.029	1.32	0.01	–189259	301
FixL-1	0.250	0.023	1.29	0.01	–189764	298
FixL-2	0.284	0.040	1.31	0.01	–189789	300
FixL-3	0.266	0.037	1.30	0.01	–189775	300
FixL-4	0.264	0.027	1.31	0.01	–189784	301

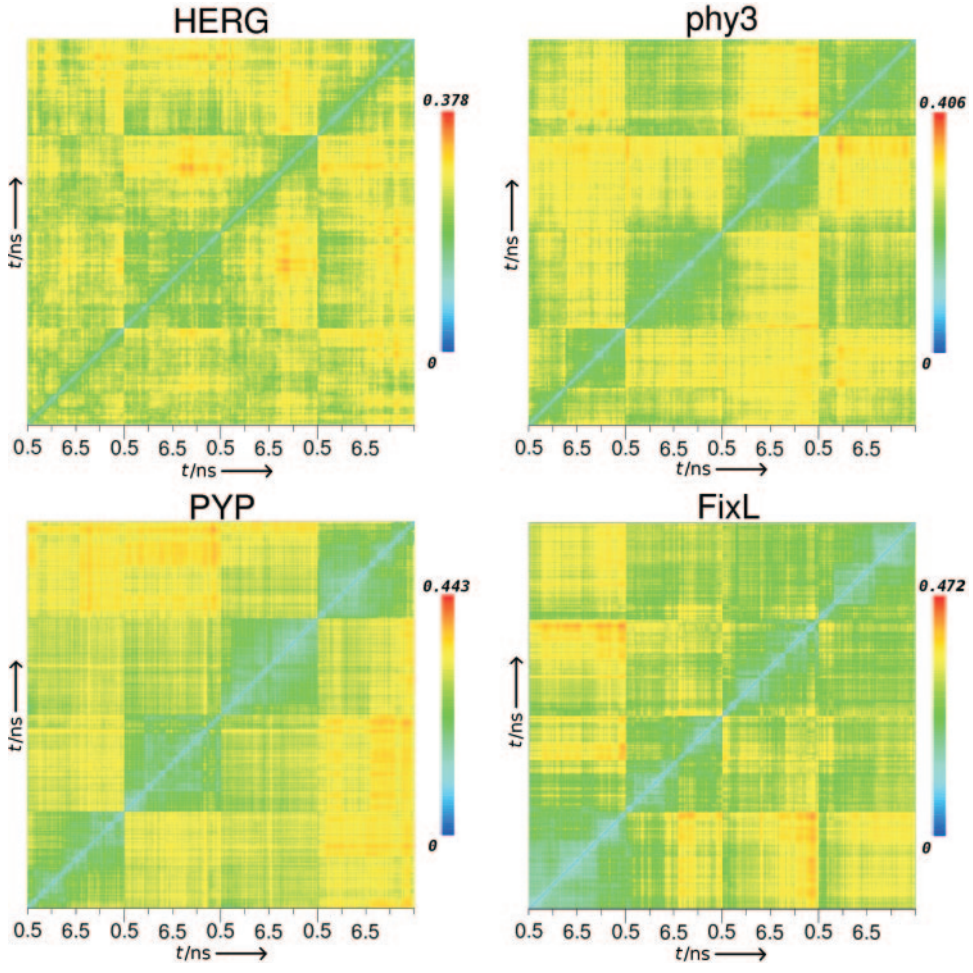

Fig. 2. R.m.s.d. matrices for the combined trajectories of PAS domain simulations, where only the productive frames were considered. Each point represents the r.m.s.d. value after least-squares fitting of the corresponding frames. Values reported are in nm.

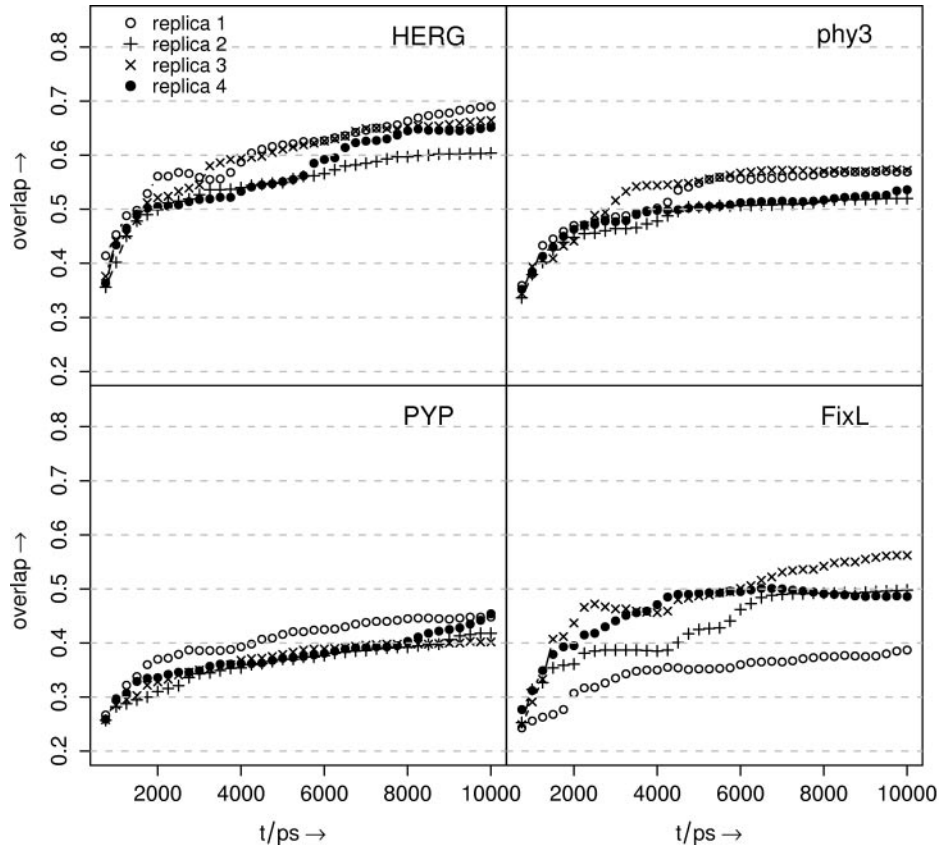

Fig. 3. Overlap graphs for the MD simulations of PAS domain structures. Reported values represent overlap between each replica section and the whole sampled set of structures obtained by combination of the four replicas. Data are presented as relative values with respect to unity, which represents complete overlap.

findings. Data are reported in Table II for the four replicas and the combined trajectory of each protein. In all cases some of the simulations have a high cosine content on the first eigenvectors and are expected partially to describe random diffusion on a flat energy landscape, while the corresponding combined trajectory has almost zero (HERG and phy3) or acceptably low (PYP and FixL) cosine content and can be related to essential motion.

On the basis of the consistency in these indexes, the following covariance analysis was directed to extract an informative and easily representable form for these patterns of motions.

Essential subspace dynamics

Covariance analysis on MD trajectories has been extensively employed to address different questions on the structure and function of a wide range of proteins (van Aalten *et al*., 1995; de Groot *et al*., 1998; Arcangeli *et al*., 2001a,b; Merlino *et al*., 2003). As described in the Materials and methods section, advantages of this type of analysis are the possibility of reducing the analysis to a lower dimensional space, to highlight spatially correlated motions and of extracting the majority of useful information that is located in the low-frequency modes usually involved in the transition between conformations (Amadei *et al*., 1993).

The ED analysis included the MD trajectories for the X-ray receptive structures of HERG, phy3, PYP and FixL, and also the NMR ensembles of structures of PYP, hPASK and HIF-2α. The analysis on the four MD trajectories resulted in a neat separation of the motion across the reoriented space, leading easily to the step of identification of the essential subspace. As indicated by the amount of displacement contained in the first set of eigenvectors (Table II), all the proteins perform more than 80% of the motion along less than 20 eigendirections (4–6% of the original space). The remaining eigenvectors describe a constrained behaviour. The analysis of such an index, and also of the distribution, shape and variance of the motion along each direction (see Materials and methods section) led to defining the extent of the essential space to six principal directions. This extent is compatible with the inclusion of more than 50% of motion in the subspace (Table II), with the employment of all directions with a significant amount

Table II. Distribution of motion in different subspaces and cosine content of the first eigenvector for the four PAS domain simulations

	HERG	phy3	PYP	FixL
No. of eigenvectors	330	312	375	315
Eigenvectors 1–3 (%)	43.6	51.7	56.0	51.9
Eigenvectors 1–6 (%)	58.3	67.3	68.2	63.6
Eigenvectors 1–10 (%)	68.8	75.1	76.2	72.6
Eigenvectors 1–12 (%)	72.5	77.6	78.5	75.4
No. of eigenvectors up to 80% of motion	19	15	14	17
Cosine content of the 1st eigenvector				
Replica 1	0.24	0.72	0.87	0.51
Replica 2	0.08	0.72	0.67	0.59
Replica 3	0.77	0.72	0.87	0.58
Replica 4	0.77	0.02	0.53	0.46
Combined trj.	0.04	0.01	0.40	0.35

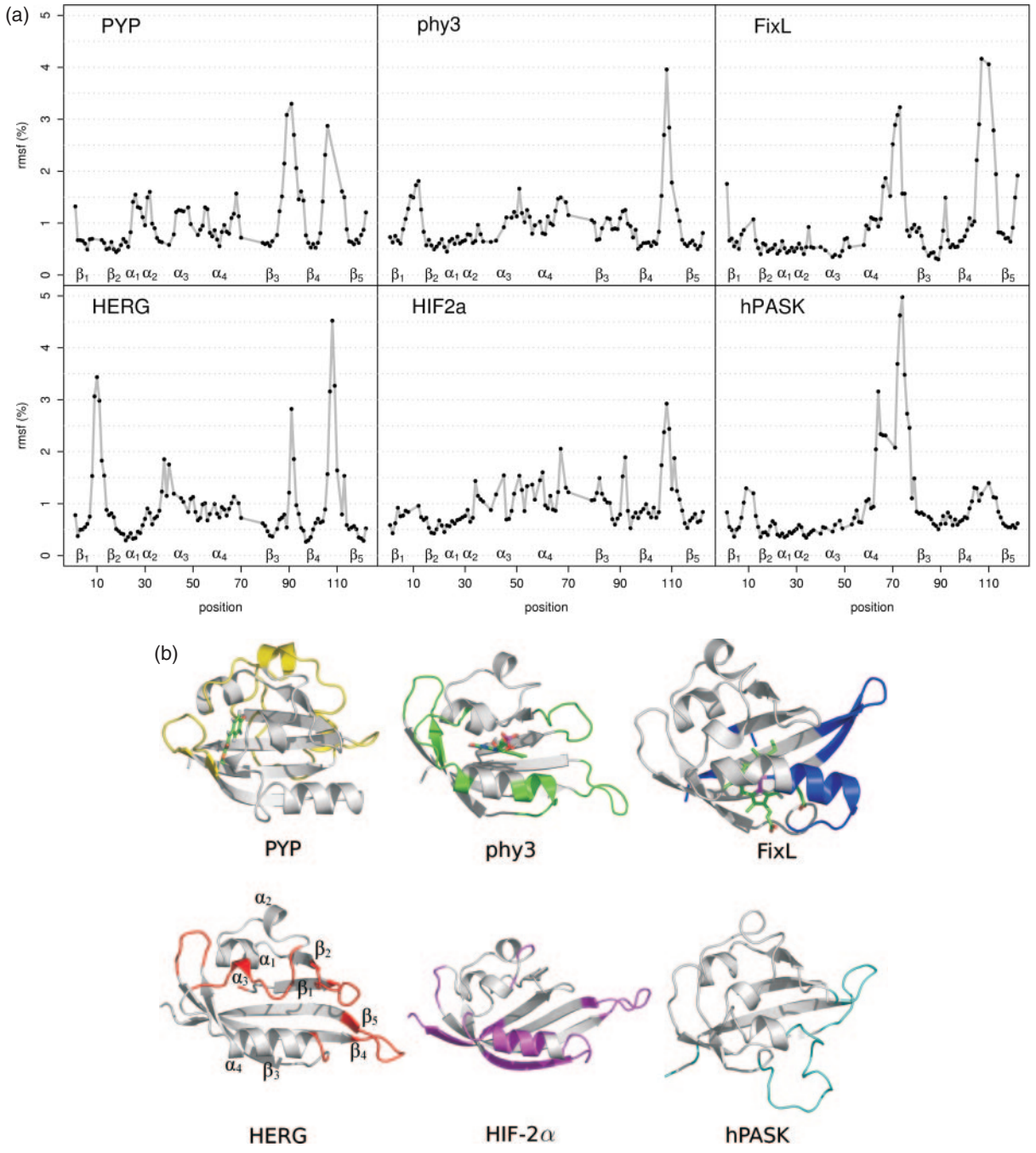

Fig. 4. Mobility in the essential subspace (six dimensions): from MD simulations for the PYP, phy3, FixL and HERG X-ray structures; from ensembles of NMR structures for HIF-2α and hPASK. (**a**) Plot of r.m.s.f. versus residue position (see text). Residue numbers are modified according to the structure-based sequence alignment (Figure 5). Locations of secondary structure elements are qualitatively indicated by labels. (**b**) Locations of motion patterns highlighted by colours in the cartoon representation of the receptive structures. Secondary structure attribution according to the Kabsch and Sander method (Kabsch and Sander, 1983). Secondary structure elements are labeled on the HERG structure.

of variance on displacement and with the inclusion of all anharmonic modes of motion (Amadei *et al*., 1993). For the sake of consistency a six-dimensional essential subspace was selected also in the ED analysis on the NMR ensembles of structures. Also for these structures more than 50% of motion is described by these directions (76.79% for PYP, 82.98 for hPASK, 66.71% for HIF-2α).

Analysis of the r.m.s.f. with respect to the residue position in the essential subspace allowed the definition of the regions with higher mobility within each domain. Plots of r.m.s.f. versus residue position for the six domains are reported in Figure 4a. For comparison purposes, only residues included in the structure-based alignment are considered and r.m.s.f. values are reported as a percentage of the total r.m.s.f. over the selected residues. As described in the Materials and methods section, identification of highly mobile regions was aimed at highlighting groups of residues with significant collective motion compared with the average. To retain only structural fragments with a significant length and to reduce background noise further, a smoothing procedure was applied. Resulting regions, defined as 'motion patterns', are reported on each structure in Figure 4b. Domains containing cofactor molecules

are grouped at the top of Figure 4a and b, whereas domains presenting empty cavities are shown at the bottom.

The ED analysis on the PYP, phy3, FixL and HERG simulations indicates constrained β-sheets for all the domains, with high mobility limited to the inter-strand loops. Beside these, mobility appears to be concentrated in the helical zone, with significant differences in its location. PYP exploits most of its flexibility in the two extra-domain helices at the N-term (behind the β-sheet in Figure 4b). High-mobility regions in the aligned part of the domain (shown in Figure 4a) are located on the two C-terminal inter-strand loops. A certain degree of mobility is also observed in the helical region, where a motion pattern emerges, including the α1–α2 connecting loop and part of α2. The ED analysis of the ensemble of NMR structures of PYP (not shown in Figure 4) indicated mobility in the same regions as evidenced by MD simulations, except for the α1–α2 region. The latter results are consistent with motions detected in solution by ^{15}N relaxation measurements on the dark state (Dux *et al*., 1998); on the other hand, in the same study the authors emphasized that α2 is more mobile than the other helices, as reflected in the incomplete definition of its α-helical conformation in all structures of the NMR ensemble. Moreover, our results from MD simulations are consistent with those obtained in earlier computational studies (van Aalten *et al*., 2000), that indicated the extra-domain and the α1 and α2 helices as the most mobile parts of the ground-state structure of PYP. The agreement between different experimental and computational approaches in detecting locations of PYP mobility, confirmed the reliability of the conformational space sampling obtained by the combined MD trajectories and the informativeness of the ED analysis.

Apart from the inter-strand loops, regions with enhanced dynamics in phy3 include the helical connector and its interconnecting loops. Two patterns of motion were identified at the N- and C-terminal ends of this region (Figure 4b).

ED analysis of the FixL simulation indicated that this domain exploits enhanced dynamics in the region including the C-terminal part of the helical connector and the following α4–β3 loop and in the spatially near β4–β5 loop. Notably, the crystal structure of the FixL PAS domain (Gong *et al*., 1998) shows elevated temperature factors in the α4–β3 loop region (also referred to as the FG loop) consistently with the increased flexibility shown by MD/ED analysis.

HERG regions with the highest mobility are the three inter-strand loops and a portion of the helical zone, where the group of residues from the α2–α3 loop to the α3 element was identified as a motion pattern.

Finally, the ED analysis on the ensembles of NMR structures of HIF-2α and hPASK are reported. For HIF-2α a higher mobility region extends from the α2–α3 loop to the β3 strand, with inclusion of two extended motion patterns encompassing the extremes of the helical connector (see Figure 4b). Other mobile fragments are the two C-terminal inter-strand loops. For hPASK the highest flexibility is located in the long α4–β3 loop and another motion pattern is in the β4–β5 loop. Consistently, the α4–β3 loop was poorly defined in the ensemble of NMR structures and ^{15}N relaxation analysis confirmed its flexibility, in the central part on a fast time-scale and at both ends on a slower time-scale (Amezcua *et al*., 2002).

On the whole, the location of motion patterns obtained by ED analysis is in complete agreement with the experimental evidence for those structures where cofactors act as receptive sites and for which signal reception and transmission mechanisms have been partially elucidated. In fact, it has been shown that in the PYP photocycle (Rubinstenn *et al*., 1998; Craven *et al*., 2000; Anderson *et al*., 2004) conformational changes, restricted to the chromophore and its hydrogen-bond network in the initial stages, subsequently induce long-range changes in the protein tertiary structure, leading to a partially disordered signalling state. The initial changes mostly involve regions that are related to the chromophore rearrangement, where the motion patterns of the PAS core were identified by the ED analysis (the α1, α2 region and the β3–β4 loop). Also, the N-terminal extra-domain helices, where signal is propagated thanks to the close contact with α1, showed high mobility in the ED analysis. In addition, experimental evidence indicates that in LOV domains the light-driven signal transduction is initiated by the formation of a covalent bond between FMN and a cysteine in α3 (Crosson and Moffat, 2001). This induces a general rearrangement in the network of hydrogen bonds, van der Waals and electrostatic interactions between the cofactor and the binding pocket, that mainly involve the α3–α4 clamp (Crosson and Moffat, 2002). The structural changes that follow this first stage have been partially elucidated. Recent NMR studies on the *Avena sativa* LOV (AsLOV) domain suggested signal propagation to a C-terminal extra-domain helix, which undergoes significant destabilization (Harper *et al*., 2003). Accordingly, motion patterns in the helical region of phy3 dark state structure involve the whole α3–α4 clamp (see Figure 4b). Moreover, in FixL, where the main cavity entrance lies between the helical connector and the β3-strand (see the heme location in Figure 4b), the loop connecting these two elements (α4–β3 or FG loop) is involved in the conformational changes that immediately follow ligand binding to the heme cofactor. The same loop is thought to be critical in the PAS-mediated signalling of this protein (Gong *et al*., 2000; Hao *et al*., 2002). The main motion patterns identified by the ED analysis for this domain include this functional loop and the C-terminal end of the helical connector.

Unlike the above systems, neither the nature of the signals nor the reception/transmission mechanisms have been elucidated for the human PAS domains of HERG, HIF-2α and hPASK. These domains had not integrated organic cofactors within their hydrophobic cores. It was hypothesized that the PAS module of hPASK might bind small organic compounds and, on the basis of experimental studies, the long α4–β3 loop was identified as a dynamic portal for ligand binding and also the primary interface for inhibiting the neighbouring kinase domain (Amezcua *et al*., 2002). Location of the main motion pattern in the same loop confirms this first mechanistic evidence.

With the above picture of a set of PAS domains with both a generally elucidated receptive mechanism and a consistent description by ED analysis, it follows that a detailed comparison of essential motion among all the PAS structures could complement sequence and structure comparisons in suggesting hypotheses for the unknown mechanisms.

Conservation and specialization in the essential motion patterns

As outlined above, the PAS domains share a common fold with highly conserved secondary and supersecondary structures; accordingly, their structural alignment may give a useful basis to drive the comparison of essential motions across the

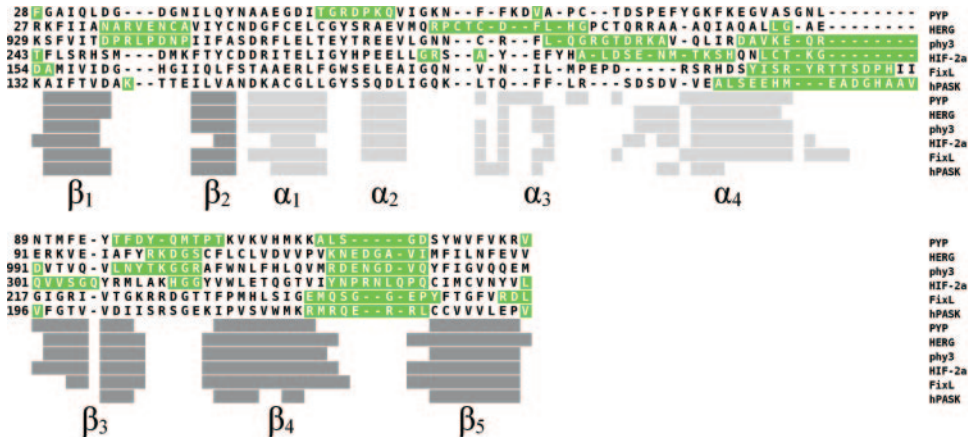

Fig. 5. Patterns of motion on the aligned PAS sequences. Residues belonging to the identified patterns are coloured green. Structural alignment generated by DALI (Holm and Sander, 1986). Secondary structure elements attributed according to the Kabsch and Sander method (Kabsch and Sander, 1983) are reported at the bottom as bars (light grey for helices, dark grey for strands) and labelled according to the adopted nomenclature (see Figure 1).

superfamily. The alignment obtained by DALI (Holm and Sander, 1996) for the target PAS domains (see Materials and methods) is reported in Figure 5. This includes only the common PAS core whereas the extra domain region at the N-terminus of PYP is omitted. The motion patterns in each structure are highlighted on the aligned sequence in the same figure.

With reference to the secondary structures, three main regions can be identified in the alignment: an N-terminal section up to the second small helix, the central region until the end of the loop after the helical connector and the C-terminal three-stranded part. The first and last regions show a high degree of secondary structure similarity, both in type, length and order; differences are limited to hPASK, for which a long turn is observed instead of α2 and the β3 and β4 strands are shorter. The central part presents almost the same secondary structures in all the domains, but with different lengths and with different extension of connecting loops.

Common elements in the dynamic behaviour of the PAS domains are the previously mentioned lack of mobility in the β-strands and high mobility in almost all the inter-strand loops, that are apparent in the aligned motion patterns at the N- and C-termini. The highest differentiation was found in the highly flexible helical zone that constitutes the central part of the alignment; only in PYP does this region extend towards the N-terminus to include α1 and α2.

Both the structural comparisons and, in more detail, the dynamic analysis suggest that this central part of the domain contains the effective specialized segment of the sequence; across the evolution the receptive function was shifted in different part of this region through proper adjustment of loop lengths and related changes of dynamic properties.

Interestingly, specialization in the central section of the sequence appears to be evolved in the different domains by following a specific trend. If patterns are compared in the order proposed in Figure 5, higher mobility regions are located in the α1, α2 region in PYP, concentrate around α3 in HERG, and in the subsequent sequences, phy3 and HIF-2α, they appear at both termini of α4, including part of the helix; finally, mobility shifts to include only the C-terminal end of the helical connector and the following loop in FixL and the same loop in hPASK.

The trend in the location of motion patterns is consistent with cofactor arrangements in the PYP, phy3 and FixL domains. Referring to Figure 4b, the hydroxycinnamic acid in PYP lies in the N-terminal cap, the FMN of phy3 is located in the cleft between this cap and the helical connector, whereas FixL heme protrudes from the binding pocket by a wide entrance offered by α4, β3 and their connecting loop.

The evidence for a similar trend of motion patterns also in the HERG, HIF-2α and hPASK human domains, that lack the cofactors, suggests for the first time that, in a more comprehensive description of the PAS domain receptive functionality, these different dynamic features can be regarded as the three general solutions developed by the domain to sense external signals.

Although the role of binding to small molecules in the regulatory activity of HERG has not been demonstrated, it was reported that some chemicals can specifically bind into the PAS domain of hPASK, stimulating its regulatory pathway (Amezcua *et al*., 2002), and it was proposed that also in HIF-2α ligand binding to the PAS domain could affect its heterodimeric interactions (Erbel *et al*., 2003).

Consequently, it is conceivable that during evolution the PAS unit could have retained dynamic and structural features consistent with the above three solutions, despite the removal of cofactors, and that the role of triggering signal transmission could have been assumed by external ligands that adopted one of these paths in entering the binding cavity.

In this framework, similarities in motion patterns, as they emerge from the proposed alignment, can suggest directions to investigate and understand the unknown mechanisms of the human domains.

The similar location of significant flexibility around the N-terminal cap in PYP and HERG suggests for the latter the involvement of this part of the structure in both signal reception and propagation. With the general lack of knowledge on the HERG mechanism, it can only be hypothesized that the N-terminus would be a suitable location for signal transmission to the disordered 25-residue terminus recognized as functionally important for the K^+ channel deactivation controlled by the HERG PAS domain (Cabral *et al*. 1998).

Following the detected similarities between phy3 and HIF-2α motion patterns, it is expected that the latter could exploit its activity by recepting external signals in the cleft

between the N-terminal cap and the helical connector, where FMN exploits receptor activity in phy3. Although a natural ligand has not been identified for HIF-2α, based on the parallels with other PAS domain signaling, it has been hypothesized that small organic compounds can bind this PAS domain and generate structural changes that are then transmitted to the β-sheet, identified as the interface of heterodimerization (Herbel *et al*., 2003). In our model, ligands are proposed to enter the binding cavity, following the outlined path, in consequence of the intrinsic flexibility of the helical connector and its hinge loops, that dynamically open the gate to the internal pocket.

The observed conservation of motion patterns in hPASK and FixL completes the picture of a conserved set of solutions for signal reception. In this case, the common involvement of the FG loop in both ligand binding and interdomain interactions was partially confirmed by NMR-based studies (Amezcua *et al*., 2002).

The general model for specialized receptive activity proposed here could be extended to newly available PAS domain structures. Annotation and comparison of dynamic properties for a broader range of PAS proteins could help to predict receptive site locations and general features of the unknown mechanisms.

On the other hand, experimental studies could be directed to validate predicted locations of receptive activity in the PAS domain, as in the case of HERG and HIF-2α, and to define better the role of flexibility in partially known mechanisms, as for phy3.

Conclusion

The description obtained by ED analysis for the PAS receptive structure mobility was in complete agreement with the most recent experimental information on the pathways involved in signal reception and propagation of these domains. These findings confirmed the centrality of dynamic information in addressing PAS protein mechanisms and the reliability of the ED approach in extracting the relevant information from the computationally or experimentally sampled conformational spaces.

The structural alignment of the six domains along with the secondary structure location clearly identify two structurally conserved regions (at the N- and C-termini) and a more variable central part. The comparative analysis of motions based on this alignment extends these findings to the dynamic features of the superfamily, where the highest differentiation was found in the flexible helical zone, in the central part of the alignment.

Sequential positions along the alignment, from the N- to the C-termini, are observed in the central region for the patterns of motion of those domains that sense signals by cofactors, by following the order from PYP to phy3 and to FixL. Moreover, the same sequential fashion is maintained in the empty human domains, from HERG to HIF-2α and finally to hPASK. This evidence suggests three general solutions that were developed by the PAS domain for its receptive function and retained by conservation of structural and dynamic features, despite the removal of cofactors.

Once annotation and comparison of dynamic properties are extended to newly available PAS domain structures, it will be possible to employ the model of receptive activity specialization proposed here for gaining insight into receptive site locations and general features of unknown mechanisms.

Acknowledgements

We thank Dr Berk Hess and Dr Alessandra Villa for helpful suggestions on ED with GROMACS and Dr Daan van Aalten for stimulating discussions about the PAS domain dynamics.

References

Amadei,A., Linssen,A.B. and Berendsen,H.J.C. (1993) *Proteins*, **17**, 412–425.
Amezcua,C., Harper,S., Rutter,J. and Gardner,K. (2002) *Structure (Camb.)*, **10**, 1349–1361.
Anderson,S., Srajer,V., Pahl,R., Rajagopal,S., Schotte,F., Anfinrud,P., Wulff,M. and Moffat,K. (2004) *Structure (Camb.)*, **12**, 1039–1045.
Arcangeli,C., Bizzarri,A.R. and Cannistraro,S. (2001a) *Biophys. Chem.*, **90**, 45–56.
Arcangeli,C., Bizzarri,A.R. and Cannistraro,S. (2001b) *Biophys. Chem.*, **92**, 183–199.
Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F. and Hermans,J. (1981) In Pullman,B. (ed), *Intermolecular Forces*. Reidel, Dordrecht, pp. 331–342.
Berendsen,H.J.C., van der Spoel,D. and van Drunen,R. (1995) *Comput. Phys. Commun.*, **91**, 43–56.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Borgstahl,G.E., Williams,D.R. and Getzoff,E.D. (1995) *Biochemistry*, **34**, 6278–6287.
Cabral,J.H.M., Lee,A., Cohen,S.L., Chait,B.T., Li,M. and Mackinnon,R. (1998) *Cell*, **95**, 649–655.
Craven,C.J., Derix,N.M., Hendriks,J., Boelens,R., Hellingwerf,K.J. and Kaptein,R. (2000) *Biochemistry*, **39**, 14392–14399.
Crosson,S. and Moffat,K. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 2995–3000.
Crosson,S. and Moffat,K. (2002) *Plant Cell*, **14**, 1067–1075.
Cusanovich,M.A. and Meyer,T.E. (2003) *Biochemistry*, **42**, 4759–4770.
Darden,T., York,D. and Pedersen,L. (1993) *J. Chem. Phys.*, **98**, 10089–10092.
de Groot,B.L., Hayward,S., van Aalten,D.M.F., Amadei,A. and Berendsen,H.J.C. (1998) *Proteins*, **31**, 116–127.
DeLano,W. (2002). *The PyMOL User's Manual*. DeLano Scientific, San Carlos, CA.
Dux,P. *et al*. (1998) *Biochemistry*, **37**, 12689–12699.
Erbel,P.J., Card,P.B., Karakuzu,O., Bruick,R.K. and Gardner,K.H. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 15504–15509.
Feenstra,K.B.H. and Berendsen,H.J.C. (1999) *J. Comput. Chem.*, **10**, 255–262.
Getzoff,E.D., Gutwin,K.N. and Genick,U.K. (2003) *Nat. Struct. Biol.*, **10**, 663–668.
Gong,W., Hao,B., Mansy,S.S., Gonzalez,G., Gilles-Gonzalez,M.A. and Chan,M.K. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 15177–15182.
Gong,W., Hao,B. and Chan,M.K. (2000) *Biochemistry*, **39**, 3955–3962.
Groenhof,G., Lensink,M.F., Berendsen,H.J., Snijders,J.G. and Mark,A.E. (2002a) *Proteins*, **48**, 202–212.
Groenhof,G., Lensink,M.F., Berendsen,H.J. and Mark,A.E. (2002b) *Proteins*, **48**, 212–219.
Grottesi,A. and Sansom,M.S.P. (2003) *FEBS Lett.*, **535**, 29–33.
Gu,Y.Z., Hogenesch,J.B. and Bradfield,C.A. (2000) *Annu. Rev. Pharmacol. Toxicol.*, **40**, 519–561.
Hao,B., Isaza,C., Arndt,J., Soltis,M. and Chan,M.K. (2002) *Biochemistry*, **41**, 12952–12958.
Harper,S.M., Neil,L.C. and Gardner,K.H. (2003) *Science*, **301**, 1541–1544.
Hellingwerf,K.J., Hendriks,J. and Gensch,T. (2003) *J. Phys. Chem.*, **107**, 1082–1094.
Hess,B. (2000) *Phys. Rev. E*, **62**, 8438–8448.
Hess,B. (2002) *Phys. Rev. E*, **65**, 031910/1–031910/10.
Hess,B., Bekker,H., Berendsen,H.J.C. and Fraaije,J.G.E.M. (1997) *J. Comput. Chem.*, **18**, 1463–1472.
Holm,L. and Sander,C. (1996) *Science*, **273**, 595–603.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kewley,R.J., Whitelaw,M.L. and Chapman-Smith, A. (2004) *Int. J. Biochem. Cell. Biol.*, **36**, 189–204.
Kurokawa,H., Lee,D.S., Watanabe,M., Sagami,I., Mikami,B., Raman,C.S. and Shimizu,T. (2004) *J. Biol. Chem.*, **279**, 20186–20193.
Lindahl,E., Hess,B. and van der Spoel,D. (2001) *J. Mol. Mod.*, **7**, 306–317.
Merlino,A., Vitagliano,L., Ceruso,M.C. and Mazzarella,L. (2003) *Proteins*, **52**, 101–110.
Miyatake,H., Mukai,M., Park,S.Y., Adachi,S., Tamura,K., Nakamura,H., Nakamura,K., Tsuchiya,T., Iizuka,T. and Shiro,Y. (2000) *J. Mol. Biol.*, **301**, 415–431.
Razeto,A. *et al*. (2004) *J. Mol. Biol.*, **336**, 319–329.
R Development Core Team (2003). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rubinstenn,G., Vuister,G.W., Mulder,F.A., Dux,P.E., Boelens,R., Hellingwerf,K.J. and Kaptein,R. (1998) *Nat. Struct. Biol.*, **5**, 568–570.
Ryckaert,J.P., Ciccotti,G. and Berendsen,H.J.C. (1977) *J. Comput. Phys.*, **23**, 327–341.
Taylor,B.L. and Zhulin,I.B. (1999) *Microbiol. Mol. Biol. Rev.*, **63**, 479–506.
van Aalten,D.M.F., Findlay,J.B.C., Amadei,A. and Berendsen,H.J.C. (1995) *Protein Eng.*, **8**, 1129–1135.
van Aalten,D., Bywater,R., Findlay,J., Hendlich,M., Hooft,R. and Vriend,G. (1996) *J. Comput.-Aided Mol. Des.*, **10**, 255–262.
van Aalten,D.M., Hoff,W.D., Findlay,J.B., Crielaard,W. and Hellingwerf,K.J. (1998a) *Protein Eng.*, **11**, 873–879.
van Aalten,D.M., Grotewold,E. and Joshua-Tor,L. (1998b) *Methods*, **14**, 318–328.
van Aalten,D.M., Crielaard,W., Hellingwerf,K.J. and Joshua-Tor,L. (2000) *Protein Sci.*, **9**, 64–72.
Vreede,J., van der Horst,M.A., Hellingwerf,K.J., Crielaard,W. and van Aalten, D.M.F. (2003) *J. Biol. Chem.*, **278**, 18434–18439.

Received August 3, 2004; revised February 12, 2005; accepted March 2, 2005

Edited by David Thirumalai