Journal of Computer-Aided Molecular Design, **13:** 21–33, 1999.
KLUWER/ESCOM

Essential dynamics/factor analysis for the interpretation of molecular dynamics trajectories

R. Kaźmierkiewicz, C. Czaplewski, B. Lammek & J. Ciarkowski*
Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

Received 5 November 1997; Accepted 22 June 1998

Key words: essential dynamics, factor analysis, molecular dynamics trajectories

Summary

Subject of this work is the analysis of molecular dynamics (MD) trajectories of neurophysins I (NPI) and II (NPII) and their complexes with the neurophyseal nonapeptide hormones oxytocin (OT) and vasopresssin (VP), respectively, simulated in water. NPs serve in the neurosecretory granules as carrier proteins for the hormones before their release to the blood. The starting data consisted of two pairs of different trajectories for each of the $(NPII/VP)_2$ and $(NPI/OT)_2$ heterotetramers and two more trajectories for the $NPII_2$ and NPI_2 homodimers (six trajectories in total). Using essential dynamics which, to our judgement, is equivalent to factor analysis, we found that only about 10 degrees of freedom per trajectory are necessary and sufficient to describe in full the motions relevant for the function of the protein. This is consistent with these motions to explain about 90% of the total variance of the system. These principal degrees of freedom represent slow anharmonic motional modes, clearly pointing at distinguished mobility of the atoms involved in the protein's functionality.

Introduction

The nonapeptide hormones oxytocin (CYIQNCPLG-NH_2, OT) and vasopressin (CYFQNCPRG-NH_2, VP) are synthesized in the hypothalamus as parts of common precursors with their associated carrier proteins neurophysin I, (NPI), and neurophysin II, (NPII), respectively [1]. The processing of the precursors into hormones complexed 1:1 with NPs occurs in the neurosecretory granules, (NSG), of the posterior pituitary [2], where they remain together until the hormone's secretion into blood. NPI and NPII are small disulfide-rich proteins of 93–95 amino acid residues, 7 disulphide bridges per molecule, very high sequence homology and almost identical hormone-binding and self-association properties [3]. It is thought that NPs, while carrying the hormones in the NSGs, are self-associated into dimers and/or higher oligomers [3, 4].

Recently, we studied the NP/bioligand interactions using molecular dynamics (MD) [5–7]. While trying to find simple and informative ways for managing and possibly reducing enormous amounts of data inherent with the proteins' MD trajectories, we realized that it is possible to reduce any MD trajectory dramatically without loss of significant structural information by applying procedures termed essential dynamics [8]. They enable a reduction of a total configurational space to an 'essential' subspace normally consisting only of a few degrees of freedom, typical of structural motions, and comprising most of the positional fluctuations. The remaining space, whose motional modes have narrow Gaussian distributions and, in fact, represent molecular oscillations, does not contribute to the 'essential modes' [8, 9]. It is a challenge to extract such motions from the molecular trajectories, to identify their functional role, and to reduce the complex protein dynamics to their essential degrees of freedom.

*To whom correspondence should be addressed.

Methods

Theory

The principles of essential MD have been described [8]. After a closer view we have realized that essential MD is actually yet another representation of a well established methodology of Principal Component (or Factor) Analysis (PCA) of very general applicability, widely used not only in physical but also – if not prevailingly – in social and behavioral sciences [10–13]. For convenience, a brief description of PCA is presented below.

Let us express correlation between consecutive conformations **x**, contributing any MD, Monte Carlo or other molecular trajectory, in the form of the covariance matrix **Cov**:

$$\mathbf{Cov} = (\mathbf{x} - \langle\mathbf{x}\rangle)(\mathbf{x} - \langle\mathbf{x}\rangle)^{\mathbf{T}}. \quad (1)$$

According to the linear algebra, any rectangular matrix **D** (e.g. $= \mathbf{x} - \langle\mathbf{x}\rangle$) of internally consistent data can be decomposed into a product of two other matrices, **R** and **C**

$$\mathbf{D} = \mathbf{R} \times \mathbf{C}, \quad (2)$$

i.e.:

$$D_{ij} = \sum_{k=1}^{f} R_{ik} \times C_{ki}. \quad (3)$$

Let the matrices **D**, **R** and **C** have the sizes (r × c), (r × f) and (f × c), respectively. The r rows of data in the matrix **R** are called the row-designees while the f columns of data in this matrix are called the factor scores. Similarly, the f rows of the matrix **C** are called the factor loadings while the c columns of this matrix are called the column-designees [12]. There is an infinite number of ways to accomplish the decomposition subject to Equations (2) and/or (3). However, we are interested in a unique way the two equivalent variants of which are

$$\mathbf{D} = \mathbf{U_1} \times \lambda^{\mathbf{1/2}} \times \mathbf{U_2^T} \quad (4)$$

or

$$\mathbf{D^T} = \mathbf{U_2} \times \lambda^{\mathbf{1/2}} \times \mathbf{U_1^T} \quad (5)$$

and arise from the constraints

$$\lambda = \mathbf{U_2^T}(\mathbf{D^T D})\mathbf{U_2} = \mathbf{U_1^T}(\mathbf{D D^T})\mathbf{U_1}. \quad (6)$$

In Equations (4) to (6) λ is the diagonal eigenvalue matrix of the size f × f determined by the number of the first most significant (i.e. the largest in value) eigenvalues common for both the eigenvalue matrices resulting from the diagonalization (Equation (6)) of either matrix $\mathbf{D^T D}$ and/or $\mathbf{Cov}(= \mathbf{D D^T}$, see Equation (1)) of sizes c × c and r × r, respectively. It appears from the above that f ≤ (c or r), whichever is lower. Hence, the orthonormal eigenvector matrices $\mathbf{U_1}$ and $\mathbf{U_2}$ (corresponding to **T** and $\mathbf{q}^{\mathrm{T}}\lambda^{-1/2}$, respectively in essential dynamics [8]) of the sizes r × f and c × f, respectively, are typically strip-like in shape. The common width of both strips, equal to f, provides an answer as to what size (i.e. a minimum number of significant factors) the dimensionality of **D** may be reduced, without loss of important structural information. Ideally, the retained factors should contain the complete structural information plus a fraction of random error, while the neglected factors should account entirely for the rest of random error. Given c ≥ r, one can write by comparison of Equations (2) and (4):

$$\mathbf{R} = \mathbf{U}_1 \times \lambda^{1/2} = \mathbf{D}\mathbf{U}_2, \quad (7)$$

$$\mathbf{C} = \mathbf{U}_2^{\mathrm{T}} = \lambda^{-1/2}\mathbf{U}_1^{\mathrm{T}}\mathbf{D}. \quad (8)$$

Hence, it appears that a single diagonalization of the r × r matrix is sufficient to yield λ, and **R** [Equations (6) and (7)] and, subsequently, **C** [Equation (8)]. Having λ and **C** ordered in accordance with decreasing λ_i, a sufficient number of structural factors is given by the residual standard deviation (**RSD**) [12]:

$$\mathbf{RSD}_f = \left\{ \left(\sum_{f+1}^{c} \lambda_i \right) \Big/ [\mathbf{r}(\mathbf{c} - \mathbf{f})] \right\}^{1/2} \leq \mathbf{SD} \quad (10)$$

The value of $\mathbf{RSD}_f$ is computed for the systematically increasing number of factors f = 1, 2,..., and judged against the standard deviation (**SD**) of the data in **D** until $\mathbf{RSD}_f$ drops below **SD**, thus yielding the respective f. This may be verified by the reproduction of the original data through the application of Equation (2) to the systematically thickened strip matrices $\mathbf{R}_f$ and $\mathbf{C}_f$, followed by the inspection of the residual matrix Δ_f

$$\Delta_f = \mathbf{D} - \mathbf{R}_f \times \mathbf{C}_f \quad (11)$$

which should eventually contain only random errors thus confirming that f is, indeed, the number of factors necessary and sufficient to reproduce the original data within their **SD** value.

At this point it is worth noting that, according to the above considerations, a single *k*th factor is defined by the *compound set* of three mutually consistent objects: *k*th columns of the eigenvector matrices $U_1^{(k)}$ and $U_2^{(k)}$ ($T^{(k)}$ and $\lambda_k^{-1/2}q^{(k)}$, respectively, in [8]) and their eigenvalue λ_k [12]. The column vector $U_2^{(k)}$ is

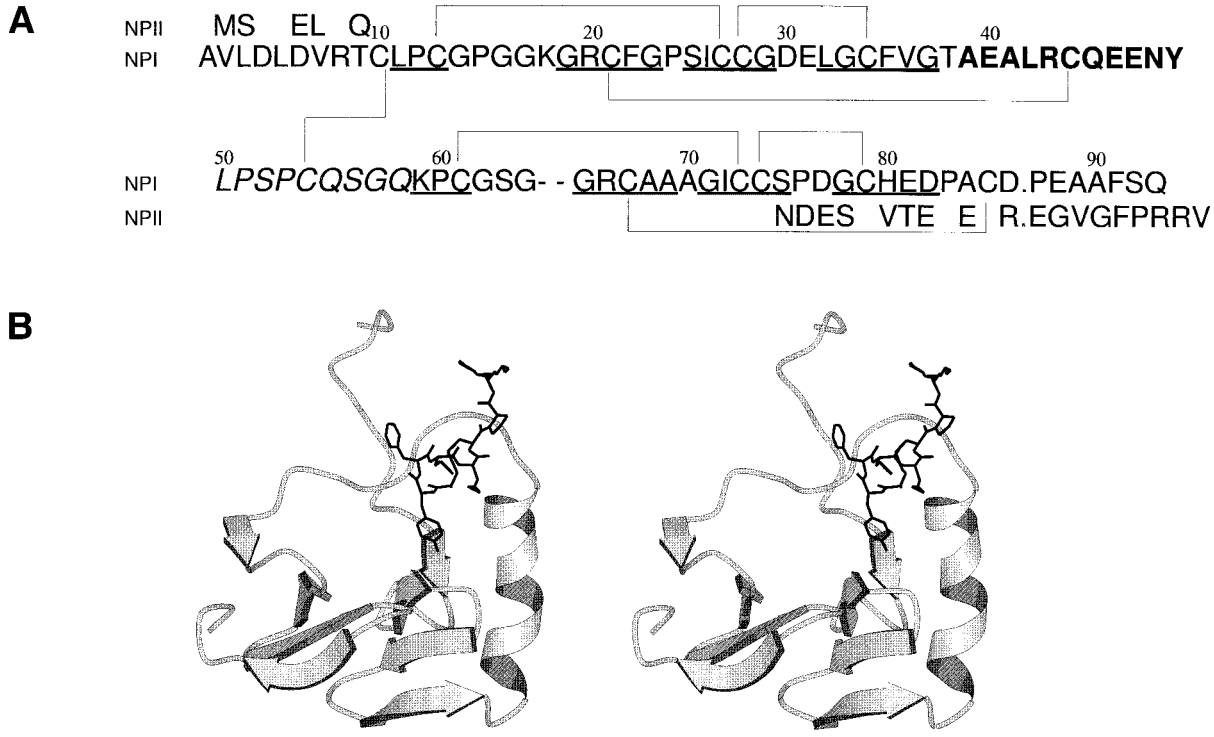

Figure 1. (A) The sequences of bovine NPI and NPII. The β-strands are underlined, the 3_{10} helix is printed in bold and the inter-domain connection in italic. The C-terminal fragment, not included in the simulated trajectories, is separated with a dot. The homological fragments in the amino and carboxyl domains are aligned vertically one under the other. (B) The structure of an NPII/vasopressin heterodimer based on the C^{α}-carbon coordinates (entry 1BN2 in the Brookhaven Protein Data Bank). The NPII molecule is made up of two highly homological domains, composed of similar four-stranded antiparallel β-sheets. In the amino domain the β-sheet is immediately followed by a 3-turn 3_{10} helix, having no match in the carboxyl domain. Both domains are connected by a relatively loose backbone fragment, supported with an *inter*-domain disulphide bridge C10–C54. The remaining 6 disulphide bridges crosslink the *intra*-domain residues: the bridges 13–27, 21–44 and 28–34 within the amino domain and the bridges 61–73, 67–85 and 74–79 within the carboxyl domain. The AVP ligand [8], represented by a stick model, is seen in the binding loop (ENYLPSPC, 47–54) composed of the end of the 3_{10} helix (AEALRCQEENY, 39–49) and the beginning of the inter-domain connection (LPSPCQSGQ, 50–58). The figure was generated using MolScript [Kraulis, P., J. Appl. Crystallography, 24 (1991) 946].

identical to the *k*th factor loading [see Equation (8)], whereas the column vector $U_1^{(k)}$ multiplied by the respective scalar $\lambda_k^{1/2}$ is identical with the *k*th factor score [see Equation (7)]. As will be demonstrated in the following section, a loading is associated with a contribution of a specific factor (mode) to a trajectory while a score is associated with a contribution of the same factor (mode) to the time-averaged distribution of motional freedom over a protein chain. The eigenvalue λ_k is an absolute measure of a fraction of the total variance explained by the *k*th factor [14, 15], i.e. it is a factor that properly scales the afore-mentioned contributions. It is worth to note that the authors of the essential dynamics freely intermingle $T^{(k)}$ and $q^{(k)}$ by calling any of them just an eigenvector (compare the captions to Figures 3 and 8 in [8], respectively). This, while being in line with the *compound factor* definition, sometimes makes their work [8] difficult to follow.

Trajectories

The initial molecular models of the $(NPII/VP)_2$ and $(NPI/OT)_2$ heterotetramers were preprocessed prior to MD in two ways, referred to as Model I and Model II. Only Model I was used for the preprocessing of the $(NPI)_2$ and $(NPII)_2$ homodimers. Both models of data preprocessing are described in detail elsewhere [6,7]. Briefly, Model I consisted of a constrained simulated annealing (CSA, Amber 4.1 implementation [16]), while Model II comprised a series of constrained minimizations framed into a rigorous protocol, gradually releasing the constraints and finally adding water [6]. This resulted in the 6 MD trajectories described in Table 1. The trajectories should in principle reflect both

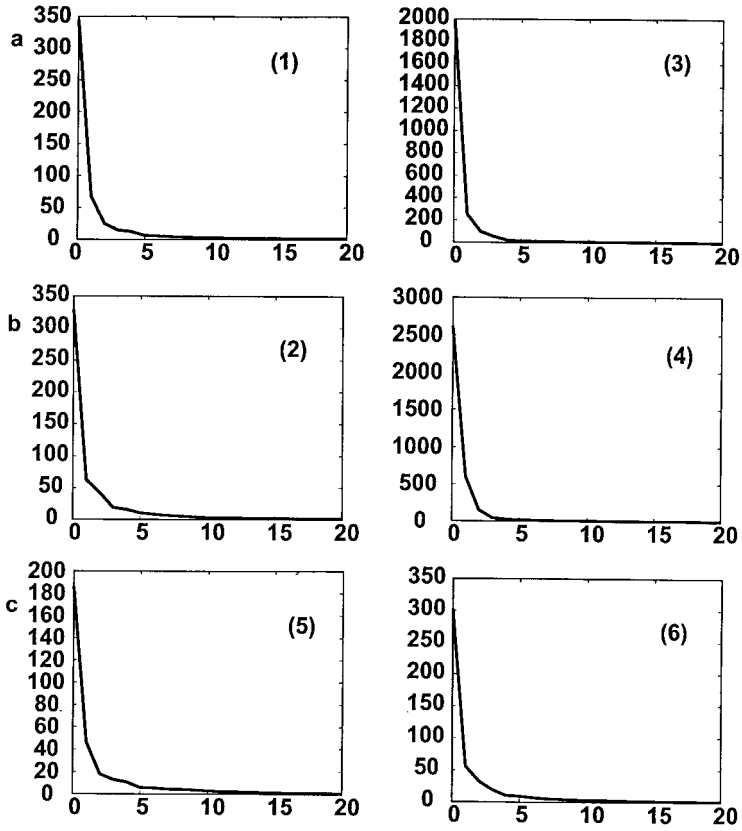

Figure 2. Eigenvalues, in decreasing order of magnitude, obtained from the following trajectories: (a) $(NPII/VP)_2$ Model I (left), Model II (right); (b) $(NPI/OT)_2$ Model I (left), Model II (right); (c) $(NPII)_2$ (left) and $(NPI)_2$ (right), both pre-processed using Model I.

the structural features and diverse preprocessing of the starting $(NPII/VP)_2$ and $(NPI/OT)_2$ structures.

Results and discussion

Six different covariance matrices were diagonalized. The procedures did not take a significant time on a middle class (Sun SPARCstation5 or IBM 42T) UNIX workstation. In Figure 2, the corresponding eigenvalue sets are plotted in the descending order. Since the eigenvalues λ_k provide the measure of the mean square displacements along the anharmonic motional modes associated with the corresponding eigenvectors ($U_1^{(k)}$ and/or $U_2^{(k)}$), it is clear that the configurational space of the protein is not homogeneous in terms of the motion along the eigenvector directions.

The amount of motion associated to a subspace spanned by the first f eigenvectors can be defined as the positional fluctuations included in the corresponding eigenvectors **R** and **C** (see Methods), pertinent to structure and trajectory, respectively.

In each of Figures 3–8 the first 10 and, in addition, the 20th and the 50th motional modes, represented by the corresponding eigenvectors $C^{(k)}$ scaled by $\lambda_k^{1/2}$ (See Methods, Equation (8)), are given. In addition, any panel on these figures contains on its left the sampling distribution function for the displacements along the respective eigenvector overlaid on the corresponding Gaussian function with the same variance and average value. It is clear from these figures that all 'essential' motions that have not yet reached their equilibrium fluctuations belong to the first 10 eigenvectors. Clearly, the non-Gaussian distributions are again found within the first 10 eigenvectors only. As is shown in Appendix B in Ref. [8], Gaussian distribution functions are expected for independent and harmonic motions.

From Figures 3–8 it is seen that there is a great similarity between the corresponding eigenvectors in various systems preprocessed according to the same model. Occasionally, they appear like mirror images, e.g. compare the $C^{(2)}\lambda_2^{1/2}$ eigenvector pair in Figures 7

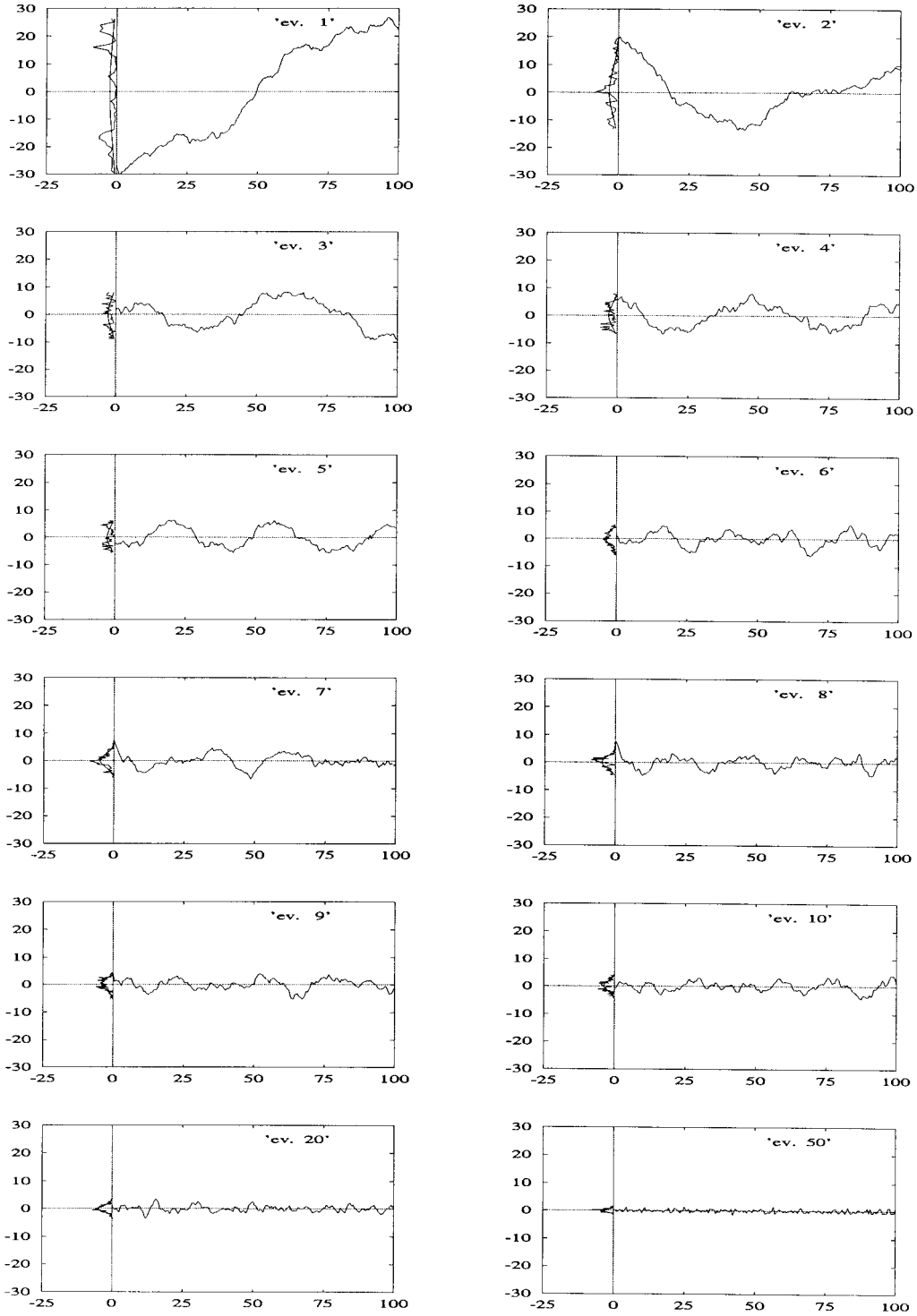

Figure 3. $(NPII/VP)_2$ complex, Model I. The first 10 and, in addition, the 20th and the 50th motional modes are shown, represented by the corresponding eigenvectors $C^{(k)}$ scaled by $\lambda_k^{1/2}$ (see Methods, Equation (8)). Since the dimension of any $C^{(k)}$ is equal to c (= the number of configurations contributing to the trajectory), any eigenvector represents the $\lambda_k^{1/2}$-scaled fluctuations (vertical axis), typical of the $C^{(k)}$-th motional mode in time (horizontal axis, ps). In addition, any panel on its left contains the sampling distribution function for the displacements along the respective eigenvector, overlaid on the corresponding Gaussian function with the same variance and average value.

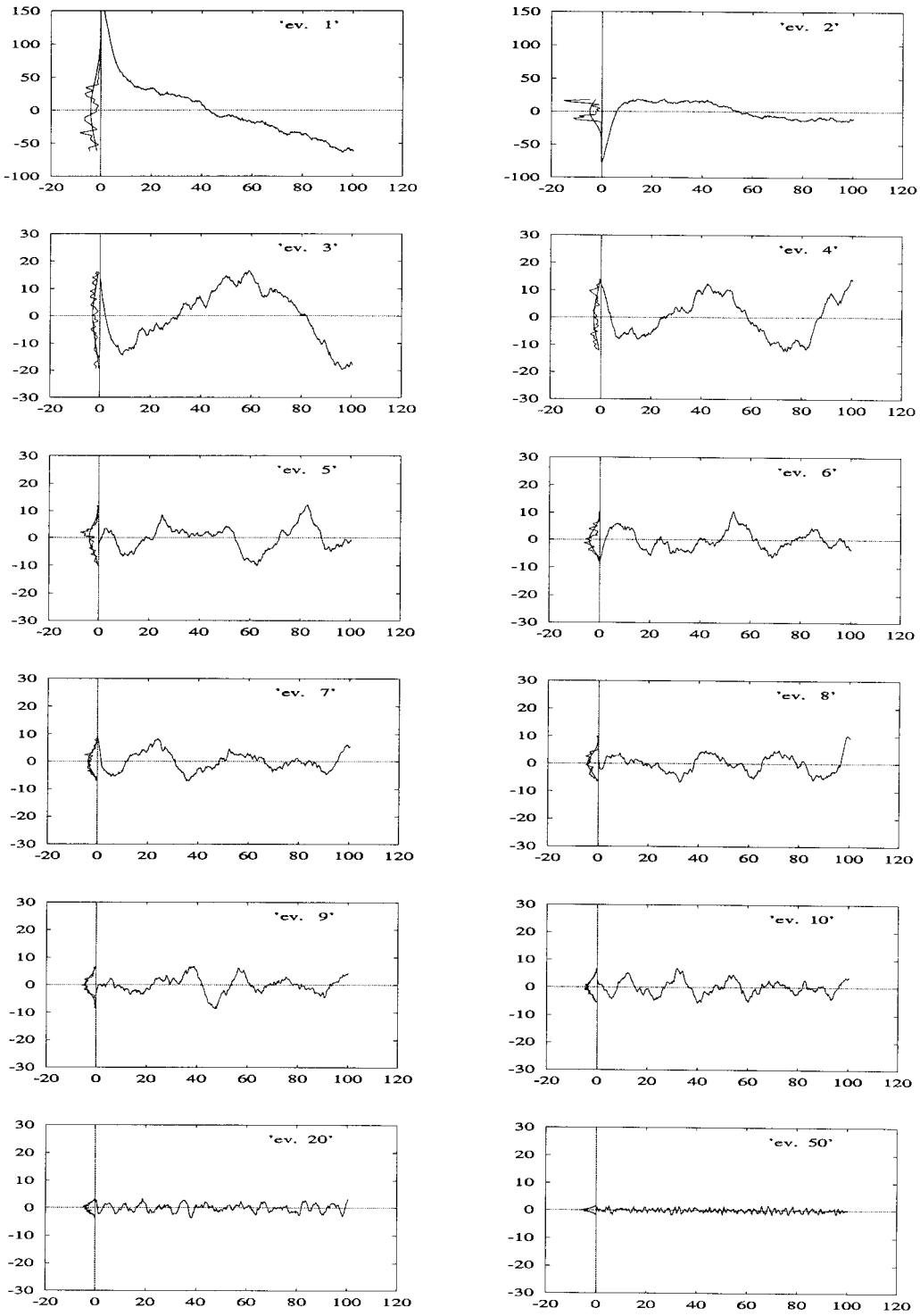

Figure 4. (NPII/VP)$_2$ complex, Model II. See legend to Figure 3.

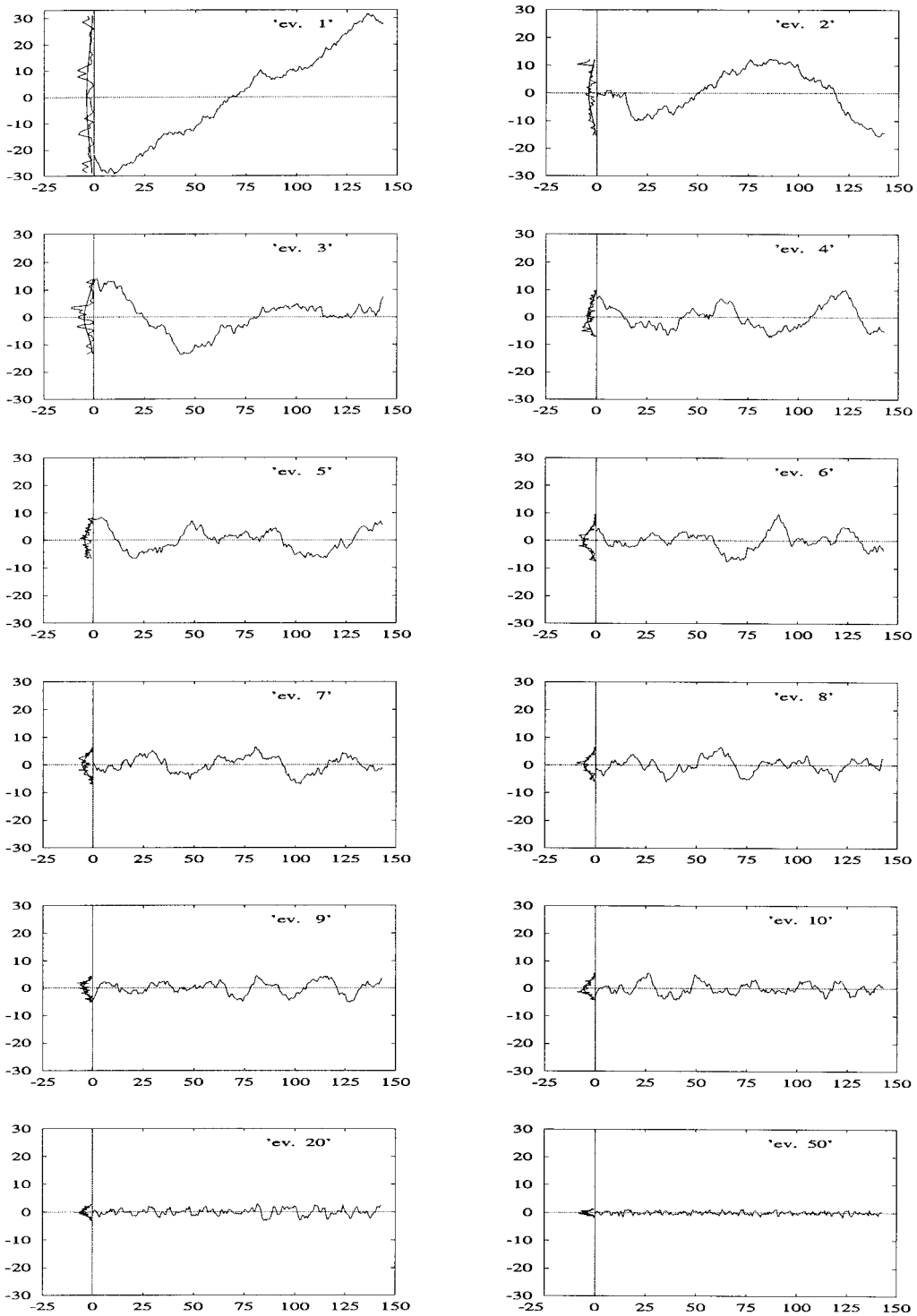

Figure 5. $(NPI/OT)_2$ complex, Model I. See legend to Figure 3.

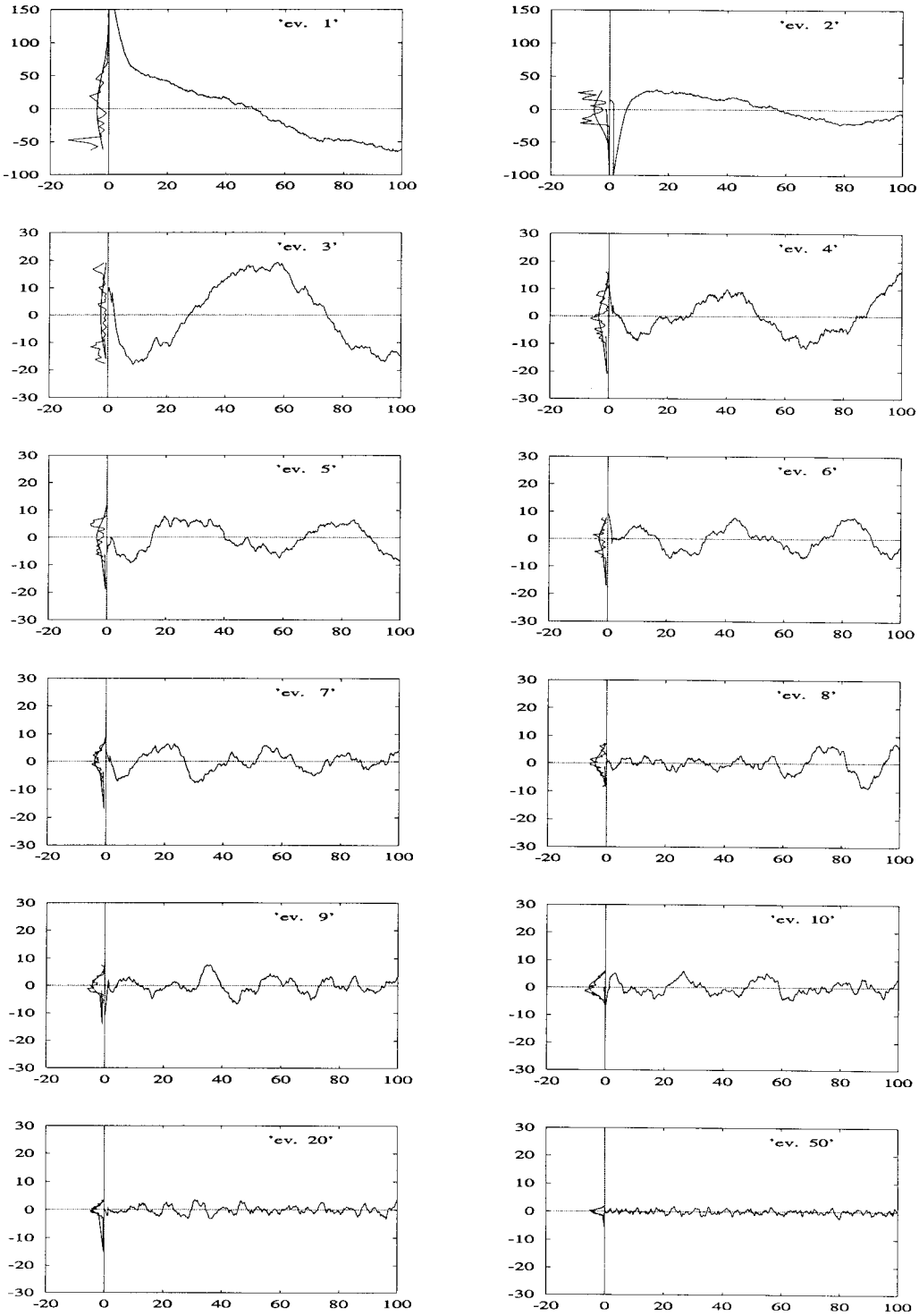

Figure 6. (NPI/OT)$_2$ complex, Model II. See legend to Figure 3.

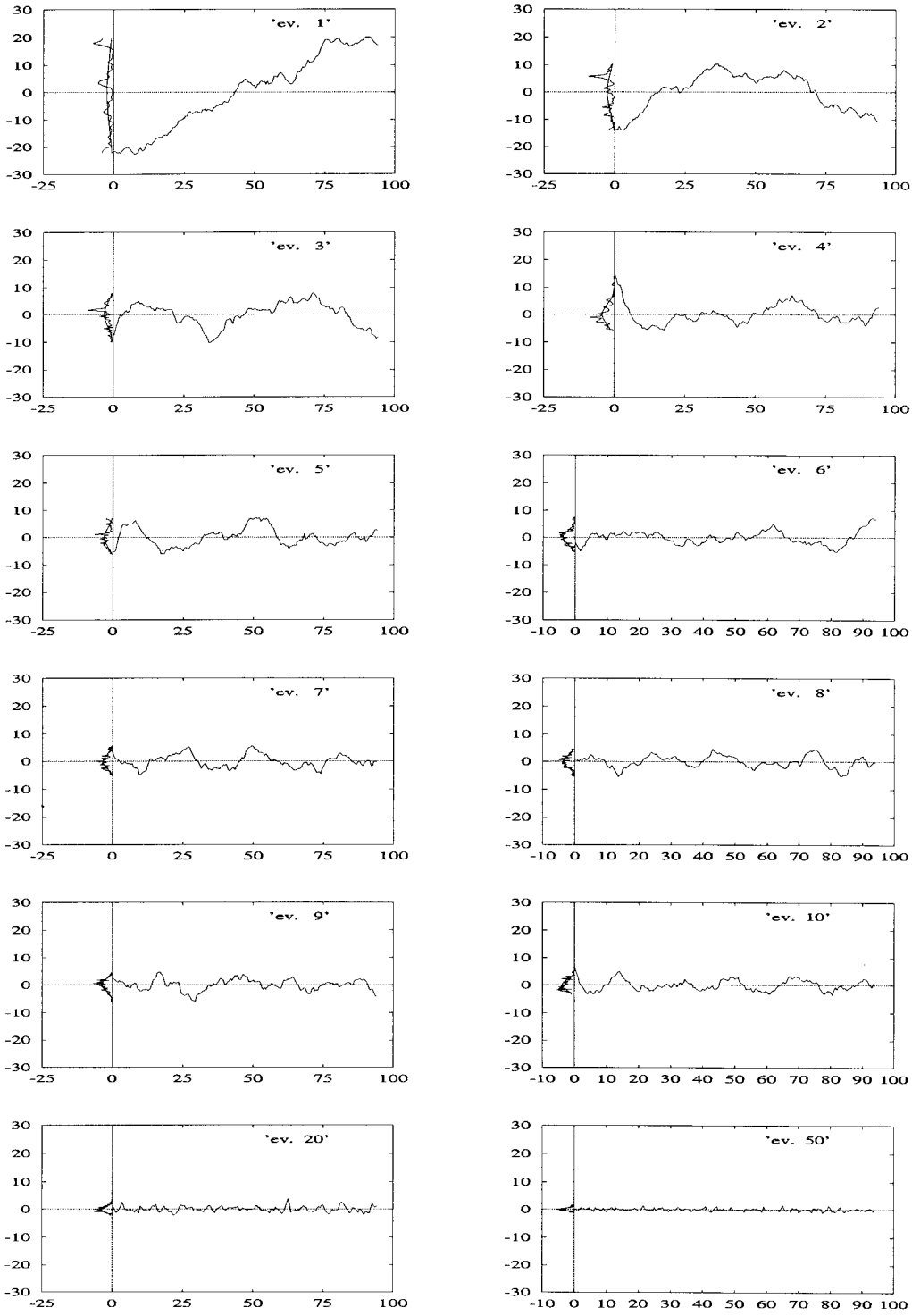

Figure 7. $(NPII)_2$ complex, Model I. See legend to Figure 3.

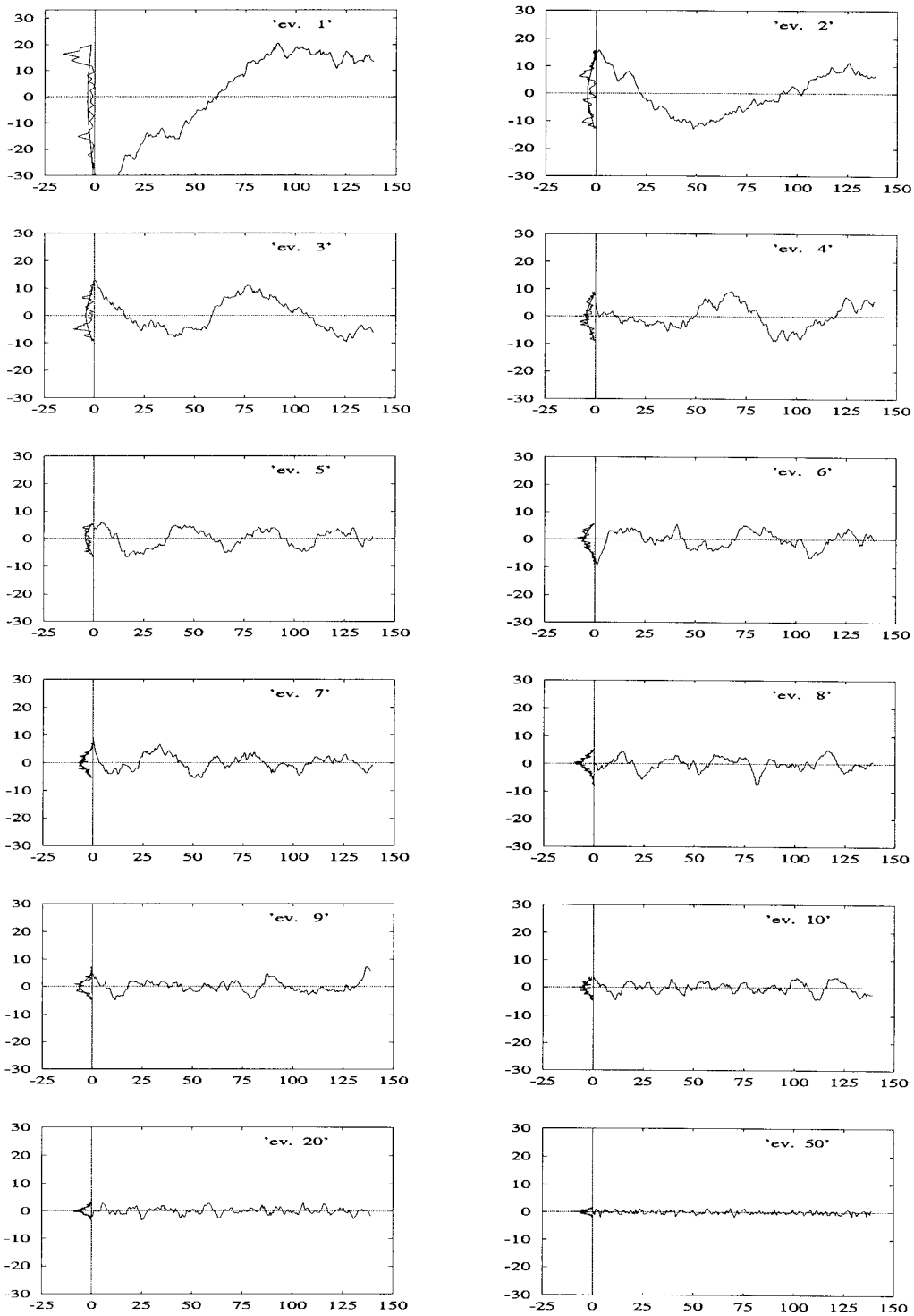

Figure 8. $(NPI)_2$ complex, Model I. See legend to Figure 3.

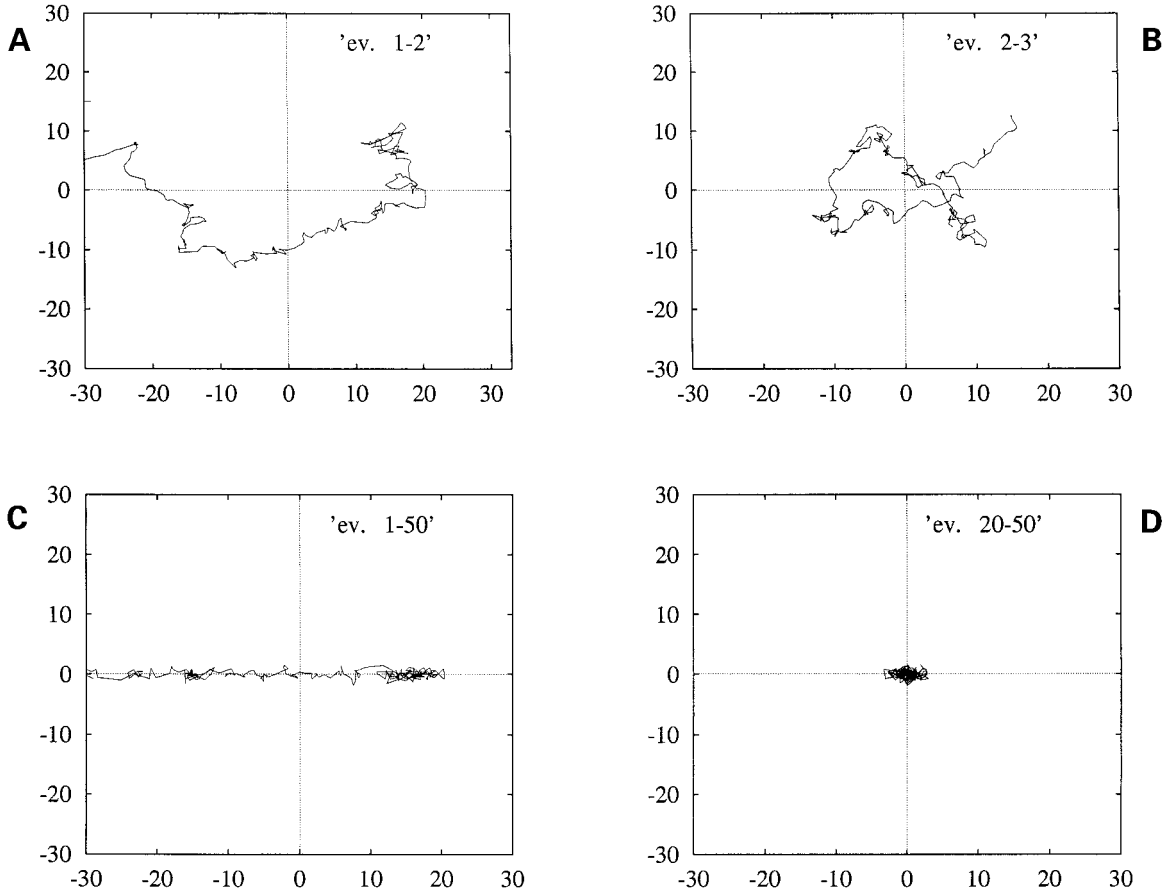

Figure 9. $(NPI)_2$ complex. A two-dimensional development of the trajectory as seen simultaneously along the two selected $\lambda^{1/2}$-scaled **C** components (see Methods, Equation (8)). A: $C^{(1)}$ (horizontal axis) and $C^{(2)}$ (vertical axis) motional modes, upper left; B: $C^{(2)}$ (horizontal axis) and $C^{(3)}$ (vertical axis) motional modes, upper right; C: $C^{(1)}$ (horizontal axis) and $C^{(50)}$ (vertical axis) motional modes, lower left and D: $C^{(20)}$ (horizontal axis) and $C^{(50)}$ (vertical axis) motional modes, lower right. It is clearly seen that while the $C^{(1)}$ and $C^{(2)}$ represent structural anharmonic motional modes, the $C^{(20)}$ and higher motional modes represent harmonic oscillations of lowering amplitudes.

Table 1. The trajectories [6, 7] submitted to the analysis

	Trajectory	Duration (ps)	No. of snapshots, c[a]	No. of Cartesian coordinates (C^{α} atoms only), r[a]
1	$(NPII/VP)_2$, Model I	100	200	570
2	$(NPII/VP)_2$, Model II	100	1000	570
3	$(NPI/OT)_2$, Model I	135	270	570
4	$(NPI/OT)_2$, Model II	100	1000	570
5	$NPII_2$	96	192	516
6	NPI_2	139	278	516

[a] See Methods.

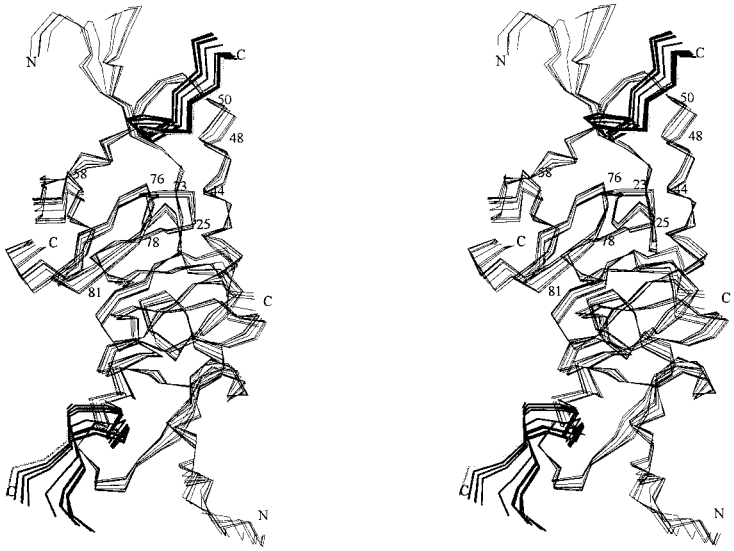

Figure 10. (NPII/VP)$_2$ complex: Superposition of 10 configurations, represented by the time-averaged structures with added motion along the first eigenvector. The ligands are drawn thick.

and 8 for NPII$_2$ and NPI$_2$, respectively. On the other hand , the Model I and Model II trajectories differ considerably for the same object and this seems to be the only consequence arising from different preprocessing of the molecular models prior to the dynamics, see Methods.

Figure 9 is a two-dimensional extension of Figure 7. It shows the development of the (NPI)$_2$ trajectory as seen simultaneously along the two selected $\lambda_k^{1/2}$-scaled $C^{(k)}$ components (see Methods, Equation (8)), with the time axis transparent, i.e. perpendicular to the drawing. Both Panel A, representing the combined motion along the eigenvectors 1 and 2, and Panel B, representing the combined motion along the eigenvectors 2 and 3, indicate that the trajectories 1, 2 and 3 are specific, suggesting the presence of a coupled force field. On the contrary, Panels C and D (representing the combined motions along the eigenvectors 1 versus 50 and 20 versus 50, respectively) indicate that the trajectories 20 and 50 fill the ranges expected for the Gaussian distributions almost completely. This means that in the latter two cases we are dealing with basically independent motions.

As already noted above, there is a great similarity between the corresponding motions along the first few eigenvectors among all four systems uniformly pre-processed, i.e. those referred to as Model I. To investigate this similarity further, we extracted the components from the structure-pertinent $\lambda^{1/2}$-scaled U_1 eigenvectors (i.e. the **R** vectors, see Methods), which would correspond to the sets of time-averaged atomic displacements associated with a specific motional mode.

Figure 10 contains the view of the 10 overlaid structures of the (NPII/VP)$_2$ heterotetramer. They are constructed from the time-averaged structure with added motion along the first eigenvector. Here a diverse motional behaviour both over the time and the sequence, typical of the first factor, is reflected in the varied dispersion of the protein's backbone along the chain. Apart of the N-, the C-termini, the loops in the protein, and the C-termini in the VP ligands, all of which are on the surface and as such structurally mobile, it is the interdomain connection (LPSPC-QSG, 50–58) which looks most mobile in the both monomers. Of other motions, it is noticeable a chain of spatially proximal mobile segments, extending from the N-terminal part of the ligand (CYF, 1–3) via a part of the 3_{10} helix (CQEEN, 44–48), the 2nd loop in the amino (GSP, 23–25) and the 3rd loop in the carboxyl domain (DES, 76–78), to the 4th β-strand in the carboxyl domain (SVTE, 78–81), which is a core of the intermonomeric interface. The other objects (NPI/OT)$_2$, NPII$_2$ and NPI$_2$ (not shown) have demonstrated very similar features. Thus, this string of motions, easily identifiable in the first motional mode (factor), do confirm a working hypothesis on the allosteric communication between the ligand binding site and the intermonomeric interface and its role in the ligand-stimulated dimerization of NPs [3, 4, 17].

Conclusions

We have used the PCA method for the analysis of the molecular dynamics histories of the six systems involving neurophysin and/or neurophysin/bioligand complexes. Our strategy included only the simplest mathematical operations like matrix decomposition (in main step). The method is fast when used for the analysis of only C^{α} traces of the protein chain. One can extract, using the PCA method, the most important motions involved in molecular dynamics simulations. The number of these anharmonical motions is limited and in our simulations was represented exhaustively by about 10 factors (10 eigenvalues with associated eigenvector pairs). The most significant (i.e. the first) factor, as carrying a majority of structural variance, explains most of the structural motion.

Acknowledgements

This work has been supported by the Academic Computer Center in Gdańsk, TASK (CI TASK), and by the Polish Scientific Research Committee (KBN) grant BW/8000-5-0179-7.

References

1. Land, H., Schultz, G., Schmale, H. and Richter, D., Nature, 295 (1982) 299.
2. Gainer, H., Russel, J.T. and Loh, Y.P., Neuroendocrinology, 40 (1985) 171.
3. Breslow, E. and Burman, S., Adv. Enzymol., 63 (1990) 1.
4. Breslow, E.M.G., In Gross, P., Richter, D. and Robertson, G.L. (Eds), Vasopressin, John Libbey Eurotext, 1993, pp. 143–155.
5. Kaźmierkiewicz, R., Czaplewski, C., Lammek, B., Ciarkowski, J. and Lesyng, B., J. Mol. Model, 1 (1995) 135.
6. Kaźmierkiewicz, R., Czaplewski, C., Lammek, B. and Ciarkowski, J., J. Comput.-Aided Mol. Design, 11 (1997) 9.
7. Kaźmierkiewicz, R., Czaplewski, C., Lammek, B. and Ciarkowski, J., Quant. Struct.-Act. Relat., 16 (1997) 193.
8. Amadei, A., Linssen, A.B.M. and Berendsen, H.J.C., Proteins, 17 (1995) 412.
9. Van Aalten, D.M., Findlay, J.B.C., Amadei, A. and Berendsen, H.J.C., Protein Eng., 8 (1995) 1129.
10. Lawley, D.N. and Maxwell, A.E., Factor Analysis as a Statistical Method, 2nd ed., Butterworths, London (1971).
11. Horst, P., Factor Analysis of Data Matrices, Holt, Rinehart and Winston, New York, NY (1965).
12. Malinowski, E. and Hovery, D.G., Factor Analysis in Chemistry, Wiley, New York, NY (1980).
13. Saunders, D.R., Psychometrika, 25 (1960) 199.
14. Murray-Rust, P. and Bland, R., Acta Crystallogr., Sect. C, 34 (1978) 2527.
15. Murray-Rust, P. and Motherwell, S., Acta Crystallogr., Sect B., 34 (1978) 2534.
16. Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, III, T.E., Ferguson, D.M., Seibel, G.L., Singh, U.C., Weiner, P.K. and Kollman, P.A., AMBER 4.1, University of California, San Francisco, CA.
17. Breslow, E., Sardana, V., Deeb, R., Barbar, E. and Peyton, D.H., Biochemistry, 34 (1995) 2137.