# 5

# *Functional Information from Slow Mode Shapes*

**Yves-Henri Sanejouand**

**CONTENTS**



## 5.1 Introduction

The idea that protein functional motions can be well described with a few slow normal modes *only*, probably originates from the seminal study of hen-egg lysozyme hinge-bending motion, by Martin Karplus and coworkers, 30 years ago [1]. Indeed, after the calculation of an adiabatic potential for the angle-bending, found to be approximately parabolic, these authors treated the relative motion of the two structural domains as an angular harmonic oscillator composed of two rigid spheres with moments of inertia corresponding to those of the domains. A vibrational frequency of 4.3 $cm^{-1}$ was obtained, quite close to the lowest-frequency value found afterward, when standard normal mode analysis (NMA) was performed [2,3].

Then, approximate low-frequency (slow) normal modes were obtained in the case of the quite large yeast hexokinase enzyme (nearly 450 amino-acids), using the Raleigh–Ritz method, and compared to the conformational change observed upon inhibitor binding. It was noticed that two of them had strong components along the conformational change [4].

Later on, such a relationship between protein functional motion and slow mode shapes was also observed for proteins whose structural domains (in particular, their limits) cannot be determined easily, like those of citrate synthase [5]. Notably, as one of the more striking examples, it was found that the second lowest-frequency mode of the T-form of hemoglobin is enough for describing two-third of the transition between T- and R-forms [6, 7].

The fact that a protein motion with a high "collective" character, that is, a motion in which many atoms are involved, can be accurately described with a subset of low-frequency modes is not a surprising result because the corresponding (normal) coordinates themselves have such collective character. However, the fact that one, or a few, of them may prove enough for obtaining a fair description of a conformational change was not a priori expected.

For instance, from a physical point of view, the energy function used to compute protein normal modes is an approximate one, and frequency values would be significantly different, if it were possible to compute them at *ab initio* levels. Moreover, low-frequency parts of protein normal mode spectra are usually not characterized by clear gaps. More generally, NMA is based on a small displacements approximation, which amounts to suppose that a protein behaves like a solid does at low temperature, although it is well known that a protein is a somewhat flexible polymer, undergoing many local conformational transitions at room temperature. Furthermore, from a biological point of view, proteins are known to fold and function in a water environment, within a narrow range of pH, temperature, ionic strength, etc., whereas standard NMA is performed *in vacuo*. As a matter of fact, it requires a preliminary energy minimization, which drifts the atoms of the protein up to several Ångstroms away from their positions in the crystallographic structure. As a consequence, the structure studied with standard NMA is a distorted one. Note that, nowadays, this later point can be partly disregarded, thanks to the development of implicit solvent models, like EEF1 [8] or ACE [9, 10], within the frame of the generalized Born approximation. Indeed, some normal mode studies are now being performed with such a kind of description for protein–water interactions [11].

However, recent results have shed some light on this paradox. Notably, it was shown that using a single parameter Hookean potential for taking into account pairwise interactions between neighboring atoms, the so-called elastic network model (ENM) [12–14], yields results in good agreement with those obtained when NMA is performed with standard semi-empirical potentials, as far as low-frequency normal modes are concerned [15–17].

The purpose of the present contribution is to compare protein functional motions and slow mode shapes, as they are obtained with standard NMA or with various, less detailed, approaches, including ENM. Hereafter,

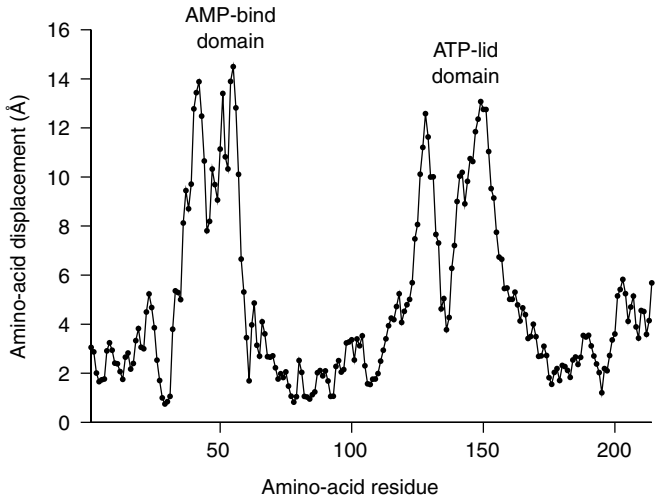


**FIGURE 5.1**
The conformational change of adenylate kinase upon ligand binding.

approximate methods are described and two cases studied previously [12, 18, 19] are considered in more depth, namely Adenylate Kinase (AdK) and dihydrofolate reductase (DHFR).

## 5.2 Conformational Change of AdK Arising from NMA

### 5.2.1 Standard Normal Mode Calculation

Adenylate kinase is a "classic" three-domain enzyme [20]. Upon binding of AdK substrates, ATP and AMP, large-amplitude motions (up to 15 Å; see Figure 5.1) of the two small "AMP-bind" (residues 31 to 72) and "ATP-lid" (residues 119 to 156) structural domains allow for the closure of the active site, as shown in Figure 5.2 in the case of *Escherichia Coli* structures (PDB codes 4AKE and 1ANK).

Standard NMA was done as follows, starting from the "open" form of AdK (Figure 5.2[a]). First, an extensive energy minimization was performed, with the CHARMM package [22], version 27, using extended atoms, the PARAM19 force-field, a distance-dependent dielectric constant, and a 9 Å cutoff for electrostatic interactions. The minimization process was stopped at a gradient root-mean-square (RMS) of $10^{-6}$ kcal/(mole Å), after nearly 20,000 adopted basis Newton–Raphson (ABNR) steps. At this point, the $C_\alpha$-RMS deviation from the crystal structure is significant: 1.9 Å. Next, using the VIBRAN module of CHARMM, **F**, the Hessian, that is, the mass-weighted second derivatives of the potential energy matrix, was diagonalized. Because in this case the matrix is not large (matrix order is $3N = 6093$), the standard

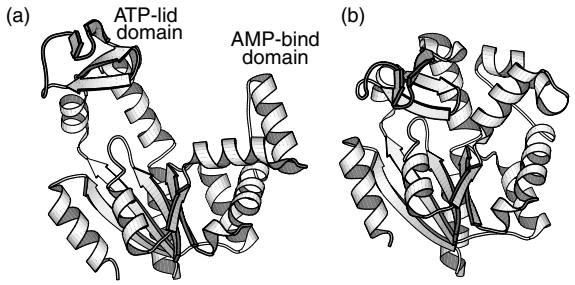


**FIGURE 5.2**
AdK open (a) and closed (b) conformations, drawn with Molscript [21].

DIAGQ routine available in CHARMM was used [23]. Among the six "zero-frequency" values found, corresponding to the overall translations and rotations of the whole protein, the largest one is close to expected numerical limits, namely 0.0035 $cm^{-1}$. This means that the minimization process was efficient enough.

### 5.2.2 Comparison with the Conformational Change

In order to quantify how well a conformational change is described by normal mode $j$, one can calculate $I_j$, the scalar product (overlap) between $\mathbf{\Delta x} = \{\Delta x_1, \ldots, \Delta x_k, \ldots, \Delta x_{3N}\}$, the conformational change observed by crystallographers, and $\mathbf{y_j} = \{y_{1j}, \ldots, y_{kj}, \ldots, y_{3Nj}\}$, the $j$th normal mode of the protein. This is a measure of the similarity between the direction of the conformational change and the one given by mode $j$. It is obtained as follows [5]:

$$I_j = \mathbf{\Delta x} \cdot \mathbf{y_j} = \frac{\sum \Delta x_k y_{kj}}{\sqrt{\sum \Delta x_k^2}} \tag{5.1}$$

where $\Delta x_k = x_k^{\mathrm{o}} - x_k^{\mathrm{c}}$, $x_k^{\mathrm{o}}$ and $x_k^{\mathrm{c}}$ are, respectively, the $k$th atomic coordinate of the protein in the open crystallographic structure and in the closed one. A value of $\pm 1$ for the overlap ($\mathbf{y_j}$ is normalized) means that the direction given by $\mathbf{y_j}$ is identical to $\mathbf{\Delta x}$. From a practical point of view, $\mathbf{\Delta x}$ is calculated after both crystallographic conformations of the protein are superimposed, using standard fitting procedures. Note that $Q_{\mathrm{d}}$, the quality of the motion description, calculated as:

$$Q_{\mathrm{d}} = 100 \sum_{j=1}^{n} I_j^2 \tag{5.2}$$

is equal to 100% when $n = 3N$, that is, when all modes are taken into account, since the $3N$ modes form a complete basis set [24].

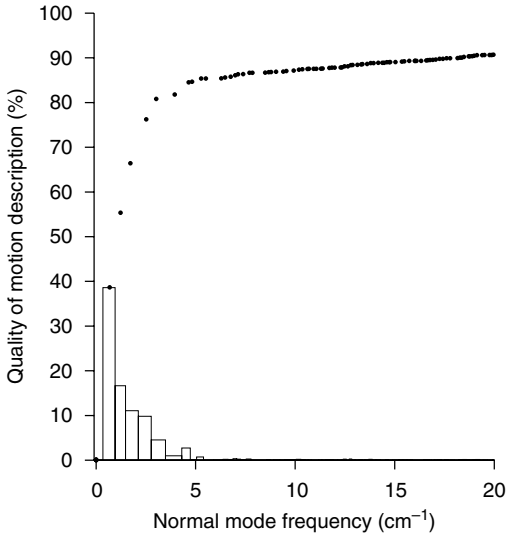


**FIGURE 5.3**
Description of AdK conformational change with standard normal modes.

In Figure 5.3, $Q_d$ is given for the AdK conformational change shown in Figure 5.2, when more and more low-frequency modes of the open form are added to the description (black circles). The contribution of each normal mode is also shown (white boxes). Note that a single normal mode, the one with lowest frequency ($\nu = 0.68$ cm$^{-1}$), is enough for describing nearly 40% of the conformational change, whereas the five with lowest frequency modes allow for the description of more than 80% of this motion. Of course, the six zero-frequency modes do not contribute to the description, because overall rigid body motions are removed when the least-square fit of the closed form with respect to the open form is performed.

### 5.2.3 Effective Number of Modes Required for the Description

In order to determine $n_{\text{eff}}$, the minimum number of modes that are sufficient for accurately describing a conformational change, one can try to evaluate the information contained in the $I_j^2$s, as follows (a related, recently proposed, quantity was coined "mode concentration" [25]):

$$\log(n_{\text{eff}}) = -\sum^{n} I_j'^2 \log(I_j'^2) \tag{5.3}$$

where

$$I_j'^2 = \frac{I_j^2}{\sum^n I_j^2}$$

The above normalization means that the $n$ low-frequency normal modes considered are supposed to yield the best possible description of the conformational change. In the case of the AdK conformational change, when $n = 3N = 6093$, $n_{\text{eff}} = 14.8$, whereas when $n = 90$, that is, when all modes considered in Figure 5.3 are taken into account, $n_{\text{eff}} = 6.9$. The difference comes from the fact that many modes contribute somehow to the description of the 10% of the conformational change that are not described by the 90 modes with the lowest frequency. Note that 6 to 8 modes describe more than a few percentages of the conformational change each (see Figure 5.3), a figure in good agreement with the latter evaluation of $n_{\text{eff}}$.

### 5.2.4 RTB Approximation

Owing to its size, diagonalizing the Hessian can be the technically limiting step. Indeed, though the NMA of the small, 58 amino-acids, BPTI protein was performed as early as 1982 [26], 10 years later the largest protein studied at the atomic level of description was still myoglobin, with 153 amino-acids [27], although most interesting proteins are much larger. Since then, efficient algorithms were designed (e.g., DIMB [28]) or adapted to the case of macromolecular assemblies (e.g., the block Lanczos approach [5]) in order to compute the lowest-frequency normal modes, that is, the most informative ones.

Instead of diagonalizing the Hessian, $\mathbf{F}$, as in standard NMA, the principle of the RTB approximation (RTB stands for rotation–translation of blocks) is to diagonalize $\mathbf{F}_{\text{b}}$, a smaller $6n_{\text{b}} \times 6n_{\text{b}}$ matrix defined as follows [18, 29, 30]:

$$\mathbf{F}_{\text{b}} = \mathbf{P}^{\text{t}}\mathbf{F}\mathbf{P} \tag{5.4}$$

where $\mathbf{P}$ is an orthogonal $3N \times 6n_{\text{b}}$ projection matrix built with the vectors describing the six rigid-body rotations and translations of each of the $n_{\text{b}}$ blocks the protein is split into. For instance, each block can contain a single amino-acid residue. $\mathbf{U}_{\text{p}}$, the $3N \times 6n_{\text{b}}$ matrix with the $6n_{\text{b}}$ approximate lowest-frequency normal modes of the protein, is then obtained as follows:

$$\mathbf{U}_{\text{p}} = \mathbf{P}\mathbf{U}_{\text{b}}$$

where $\mathbf{U}_{\text{b}}$ is the matrix diagonalizing $\mathbf{F}_{\text{b}}$, $\mathbf{U}_{\text{b}}$ being obtained with standard diagonalization techniques. DIAGRTB, the corresponding Fortran program is available on the web (http://ecole.modelisation.free.fr/modes.html). An efficient, more general, implementation, called BNM (standing for Block Normal Modes) [30], where each block can be treated as a flexible body, in the spirit of dynamical models of the MB(O)ND family [31, 32], is also available in CHARMM [22], since version 32. Note that approximate modes thus obtained can then be refined, for instance, using the effective Hamiltonian theory, as originally proposed [29]. However, as far as slow mode shapes are concerned,

approximate modes are usually so close to exact modes [18, 29] that it is not worth the extra computational cost.

As a matter of fact, the RTB approximation allows for quick calculations of the lowest-frequency modes of large systems described at atomic level [18]. Indeed, when two residues are placed in each block, $\mathbf{F}_b$ is a $3N_r \times 3N_r$ matrix, where $N_r$ is the number of residues. So, it has the same size as matrices diagonalized within the frame of methods based on simplified protein representations, when only $C_\alpha$ atoms are taken into account [12, 13, 17]. When six residues are placed in each block, $\mathbf{F}_b$ is a $N_r \times N_r$ matrix, that is, it has the same size as contact matrices diagonalized within the frame of the fastest method allowing for *B*-factors calculation [16].

Of course, the RTB approximation can only be used for calculating modes in which the so-defined blocks behave almost rigidly. Even in that case, calculated frequencies are found to be higher than exact ones, reflecting the fact that atoms belonging to a given block cannot relax so as to lower the energetical cost of the normal mode motion. However, for frequencies lesser than 40 $cm^{-1}$, at least when one amino-acid is put in each block, a linear relationship between approximate and exact frequencies holds, that is,

$$\nu_{rtb} = d_p \cdot \nu_s$$

where $\nu_s$ and $\nu_{rtb}$ are frequencies obtained using, respectively, standard approaches or the RTB approximation. In the case of a set of proteins of various sizes, using CHARMM force-field [22] and an 8.5 Å cutoff for electrostatic interactions, it was found that $d_p$ does not depend upon protein size or fold type ($d_p = 1.7 \pm 0.1$) [18]. This enables us to get fair estimates for exact frequencies, once the approximate ones are known. Note that $d_p$ seems to increase linearly, as a function of the number of amino-acid residues put in each block. Indeed, $d_p$ is nearly equal to 1.7, 2.1, 2.4, and 3.0, when each block contains 1, 2, 3, or 5 residues, respectively. However, in the later case, the linear relationship between $\nu_s$ and $\nu_{rtb}$ only holds for frequencies below 15 to 20 $cm^{-1}$ [18]. Note also that $d_p$ depends little upon the details of the electrostatic potential. In the present study of AdK normal modes, where a 9.0 Å cutoff and a distance-dependant dielectric constant are used, $d_p$ is found equal to 1.8 and 3.2, respectively, when each block contains one or five residues.

In Figure 5.4, $Q_d$, the quality of the motion description (see Equation 5.2), is given for each standard normal mode of AdK when the 100 lowest-frequency approximate modes are taken into account in Equation 5.2 ($n = 100$), as they are calculated with the RTB approximation, with one (black squares) or five (white squares) residues per block (results are also shown when Tirion's modes are used for the description; see Section 5.2.5). With one residue per block, RTB low-frequency modes are able to describe more than 80% of each standard mode of frequency lower than 10 to 15 $cm^{-1}$. Similar results were obtained previously, in the case of the HIV-1 protease [18]. With five residues per block, the quality of the description drops significantly as the frequency

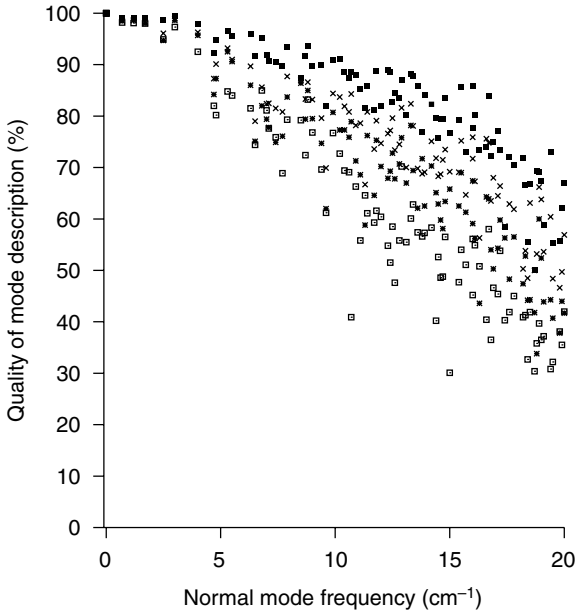


**FIGURE 5.4**
Quality of the description of each AdK normal mode with 100 approximate ones. Approximate low-frequency modes were calculated as follows: standard Hessian and the RTB approximation, with one (black squares) or five amino-acid residues per block (white squares); Tirion's Hessian (stars); Tirion's Hessian and the RTB approximation (crosses).

of the mode increases, except for the five lowest-frequency modes ($\nu = 0.68$, 1.23, 1.72, 2.52, and 3.02 cm$^{-1}$).

In Figure 5.5, $n_{\text{eff}}$, the effective number of modes required for the description (see Equation 5.3), is also given for each standard normal mode of AdK, when the 100 lowest-frequency approximate modes are taken into account in Equation 5.3 ($n = 100$), as they are calculated with the RTB approximation. Only the five lowest-frequency standard normal modes can be accurately described with less than five approximate modes. The sixth one ($\nu = 3.95$ cm$^{-1}$) is well described with $n_{\text{eff}} = 3.5$ modes calculated with the RTB approximation and one residue per block, but $n_{\text{eff}} = 13.8$ when the RTB approximation is used with five residues per block.

### 5.2.5 Tirion's Approach

Within the frame of the approach proposed by Tirion [15], the standard detailed potential energy function is replaced by

$$E_{\text{p}} = \sum_{d_{ij}^0 < R_{\text{c}}} C(d_{ij} - d_{ij}^0)^2 \tag{5.5}$$

where $d_{ij}$ is the distance between atoms $i$ and $j$, $d_{ij}^0$ being the distance between these two atoms in the studied structure. The strength of the potential $C$ is a constant assumed to be the same for all interacting pairs. As such, it has to be set only in order to define energy (and frequency) units. Note that this energy function was designed so that for any chosen configuration the potential energy, $E_p$, is a minimum of the function ($E_p = 0$). Thus, with such an approach *par définition* NMA does not require any prior energy minimization.

Note also that in Equation 5.5 the sum is restricted to atom pairs separated by less than $R_c$, which is an arbitrary cutoff parameter. When, as proposed by Bahar et al., only $C_\alpha$ atoms are taken into account [16], a cutoff of 8 to 13 Å can be used [12, 13]. The corresponding ENM [12–14] ($C_\alpha$-ENM) is enough to study backbone motions, since it proves sufficient for characterizing low-frequency normal modes of proteins. Moreover, it allows for studying proteins of large size on common workstations, using small amounts of CPU time, since, with such a simple model, the matrix to be diagonalized is a $3N_r \times 3N_r$ one. As a matter of fact, with such models, modes of systems as large as the whole ribosome have been calculated on desktop computers [33].

Using this kind of highly simplified potential, as with detailed modes, a few low-frequency normal modes are often found to yield a good description of protein functional motions, especially when the corresponding conformational change has a highly collective character [12, 14, 25]. Thus, results obtained with NMA in the field of low-frequency protein dynamics seem to be of a very good quality even when most atomic details are ignored.

In Figure 5.4 and Figure 5.5, respectively, $Q_d$, the quality of the motion description and $n_{eff}$, the effective number of modes required for an accurate description, are given for each standard normal mode of AdK when the 100 lowest-frequency approximate modes are taken into account in Equations 5.2 and 5.3 ($n = 100$), as they are calculated with Tirion's approach. For the sake of comparison, the structure considered is the open form of AdK studied with standard NMA, that is, the energy-minimized one. Here, $R_c = 5$ Å and, as initially proposed [15], all atoms are included in the model. First, the diagonalization of the corresponding Hessian is performed with standard techniques (stars). Next, the RTB approximation is used, with one residue per block (crosses). Note that this kind of calculation can now be performed through the Web, thanks to the ELNEMO Web site of Karsten Suhre (http://igs-server.cnrs-mrs.fr/elnemo/) [34, 35].

Interestingly, within the frame of Tirion's approach, the RTB approximation seems to improve the quality of the description of most, if not all, standard modes considered (see Figure 5.4 and Figure 5.5). This is not an unexpected result, because RTB adds informations to Tirion's model, about amino-acid sizes for instance, through the projection process (see Equation 5.4). Note that in the method originally proposed by Tirion, topological informations were also included in the model through the use of internal coordinates [15]. On the other hand, Tirion's modes yield better descriptions of standard modes than those obtained with detailed potentials, when the

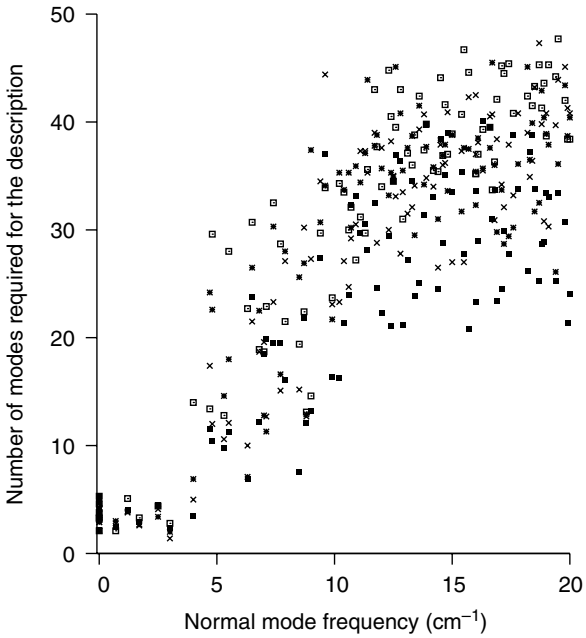


**FIGURE 5.5**
Minimum number of approximate modes required for an accurate description of each AdK normal mode. Approximate low-frequency modes were calculated as follows: standard Hessian and the RTB approximation, with one (black squares) or five amino-acid residues per block (white squares); Tirion's Hessian (stars); Tirion's Hessian and the RTB approximation (crosses).

latter are obtained with the RTB approximation and five residues per block (see Figure 5.4 and Figure 5.5).

However, with all four approximations, the five lowest-frequency standard normal modes of the open form of AdK are all found to be extremely well described ($Q_d$ over 95%; see Figure 5.4), with a small effective number of approximate modes ($n_{eff} = 5$ or less; see Figure 5.5). This means that these modes are very robust and, specifically, that the subspace spanned by the five corresponding coordinates (the normal coordinates) is almost not perturbed when most atomic details are missing in the protein model. A similar conclusion was also reached in a study of large protein normal modes, through a hierarchy of coarse-grained models [36], as well as in a study using low-resolution structural data [19].

Such results support the idea that the few lowest-frequency modes depend mainly upon the shape of the protein, that is, upon the distribution of its masses in space. As expected, such modes can be described well in terms of relative motions of structural domains of the protein. Reciprocally, when the limits of the domains of a given protein are not obvious, they can be used in order to delineate the domains, as proposed by Hinsen [17].

**TABLE 5.1**

Quality of the Description of AdK Conformational Change with Low-Frequency Normal Modes, either Standard or Approximate Ones.

| Structure | Method | Largest overlap (mode rank) | $n_{\text{eff}}$ | $Q_{\text{d}}$(%) |
|---|---|---|---|---|
| Energy-minimized | Standard | −0.62 (#1) | 6.9 | 91 |
| Energy-minimized | Standard+RTB(1) | 0.74 (#1) | 4.4 | 91 |
| Energy-minimized | Standard+RTB(5) | −0.71 (#1) | 4.5 | 91 |
| Energy-minimized | Tirion | −0.75 (#6) | 4.8 | 91 |
| Energy-minimized | Tirion+RTB(1) | −0.77 (#1) | 4.3 | 92 |
| Energy-minimized | $C_\alpha$-ENM | −0.74 (#1) | 4.9 | 96 |
| Crystallographic | Tirion | 0.81 (#9) | 3.6 | 94 |
| Crystallographic | Tirion+RTB(1) | −0.81 (#1) | 3.6 | 95 |
| Crystallographic | $C_\alpha$-ENM | 0.81 (#1) | 3.8 | 97 |

The normal modes are calculated either for the energy-minimized structure obtained within the frame of the standard approach, or for the initial, crystallographic, structure of the open form. When the RTB approximation is used for diagonalizing the Hessian, the number of amino-acid residues put in each block is given between parentheses.

### 5.2.6 Description of the Conformational Change with Approximate Modes

In Table 5.1, a summary of the results obtained is given, when the AdK conformational change is compared to the 100 lowest-frequency normal modes obtained with the various methods described above.

Amazingly, results obtained with standard NMA, that is, with the more detailed protein model, are the less spectacular ones. Indeed, the mode that is most involved in the conformational change [37] has an overlap with the conformational change (see Equation 5.1) of −0.62, whereas it has more significant values when approximate modes are considered (up to 0.81). Moreover, the number of modes required for an accurate description of the conformational change, $n_{\text{eff}}$ (see Equation 5.3), is significantly smaller when approximate methods are used (down to 3.6, instead of 6.9), whereas the quality of the description with the 100 lowest-frequency modes, $Q_{\text{d}}$ (see Equation 5.2), is better (up to 97%, instead of 91%).

One major advantage of the family of methods based on Tirion's approach is that they enable calculating low-frequency modes of the crystallographic structure itself. As mentioned above, during the minimization process required within the frame of standard NMA, the open form of AdK drifts away from the crystal structure by 1.9 Å. As a matter of fact, better results are obtained when normal modes are calculated for the crystallographic structure (see Table 5.1). Note that, in this later case, as far as the description of the conformational change is concerned, the RTB approximation does not improve the modes obtained with Tirion's approach. However, it lowers the rank of the mode found to be the most involved in the conformational change. This is likely to be related to the following artifact, found when Tirion's approach