doi:10.1016/S0022-2836(02)00562-4 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) **320**, 1011–1024

JMB

Simplified Normal Mode Analysis of Conformational Transitions in DNA-dependent Polymerases: the Elastic Network Model

M. Delarue[1]* and Y.-H. Sanejouand[2]

[1]*Unité de Biochimie Structurale*
URA 2185 du CNRS
25 rue du Dr Roux
Institut Pasteur, 75015 Paris
France

[2]*Centre de Recherches Paul*
Pascal, UPR 8641 du CNRS
Avenue Albert Schweitzer
33600 Pessac, France

**Corresponding author*

The Elastic Network Model is used to investigate the open/closed transition in all DNA-dependent polymerases whose structure is known in both forms. For each structure the model accounts well for experimental crystallographic *B*-factors. It is found in all cases that the transition can be well described with just a handful of the normal modes. Usually, only the lowest and/or the second lowest frequency normal modes deduced from the open form give rise to calculated displacement vectors that have a correlation coefficient larger than 0.50 with the observed difference vectors between the two forms. This is true for every structural class of DNA-dependent polymerases where a direct comparison with experimental structural data is available. In cases where only one form has been observed by X-ray crystallography, it is possible to make predictions concerning the possible existence of another form in solution by carefully examining the vector displacements predicted for the lowest frequency normal modes. This simple model, which has the advantage to be computationally inexpensive, could be used to design novel kind of drugs directed against polymerases, namely drugs preventing the open/closed transition from occurring in bacterial or viral DNA-dependent polymerases.

Keywords: polymerases; normal modes; open/closed transition; domain movements; drug design

Introduction

DNA replication, transcription and maintenance are essential functions in all forms of life. All these tasks are performed, at least for the polynucleotide synthesis part, by DNA-dependent polymerases, which have been characterized in all kingdoms of life.[1]

Following the first structure determination of a DNA polymerase,[2] a large body of structural and functional data has been accumulated in the last decade on this otherwise very diverse superfamily of proteins. In particular, an avalanche of new polymerase structures has appeared recently, from which it has been possible to identify a few unifying principles.[3]

One of them is that the number of different folds adopted by the different subfamilies identified through sequence alignments[4] is smaller than previously thought. For instance the catalytic domains of pol A (pol I) and pol B (pol α) families of DNA-dependent DNA polymerases were found by X-ray analysis to adopt the same fold,[5] as predicted earlier by sequence motifs alignment.[6] Similarly, single subunit DNA-dependent RNA polymerases were also predicted to adopt the pol I fold[6] and this prediction has been confirmed later by crystallography.[7] However, the catalytic domain of human DNA polymerase β (pol β) adopts a completely different fold.[8] Multisubunit RNA polymerases also have a different and more complicated fold and architecture.[9]

Even if the folds are different, their overall architecture can all be described with the hand metaphor.[2,3] DNA polymerase structural organization is highly modular and made up of several domains which are described as the palm, fingers and thumb domains. The catalytic site is on the surface of the palm domain.

Abbreviations used: pol, polymerase; NMA, mormal mode analysis; TdT, terminal desoxynucleotidyltransferase.

E-mail address of the corresponding author: delarue@pasteur.fr

At the atomic level, their active sites can all be described with the so-called two-metal-ions mechanism.[10] They all contain a few strictly conserved and crucial aspartate residues arranged in space in a similar manner to coordinate two metal ions that will (i) activate the 3′OH of the primer strand to be elongated and (ii) assist in the departure of the PPi moiety of the incoming dNTP. This is true for the pol α and pol I structural class, the pol β structural class as well as for multisubunit RNA polymerases, even though their catalytic (palm) domains all have different folds.

Finally, at least one member of these different classes has been found in two different forms, the open and closed forms.[11,12] The existence of the two different forms was found when tertiary complexes could be crystallized, some of them displaying catalytic activity in the crystal state.[8,13–16] The closed form allows for an intimate "grip" of the enzyme on its template substrate while the open form is necessary as a relaxed form in order for translocation to occur. For processive enzymes, the protein is expected to cycle between these two forms.

All of these points are also valid for RNA-dependent polymerases, as determined by X-ray crystallography. They all adopt the pol I fold as predicted earlier by sequence analysis.[6,17] However, we will leave aside reverse transcriptases in this article since there are already a number of recent studies devoted to a dynamical transition in reverse transcriptase by molecular dynamics[18,19] or by normal mode analysis.[20] We will only touch upon RNA-dependent RNA polymerases to make predictions, as there exists only one known form of these polymerases up to now and mainly concentrate on DNA-dependent polymerases where there are several examples of open and closed conformations.

The transition between the open and closed forms of polymerases is usually associated with the limiting step of the polymerization reaction, which is described in terms of an induced fit transition.[21] Indeed, this transition is thought to occur only if the incoming dNTP is complementary to the base being copied. This step is therefore directly associated with the fidelity of the copying reaction.[22,23]

Recently, two new structures of polymerases belonging to the so-called pol Y family[24] have been solved. This family is able to bypass lesions on the DNA (hence its name of translesion polymerases) and is involved in DNA repair mechanisms.[25] By virtue of its function, this family replicates DNA in an error-prone fashion. One of the recently solved structures is the yeast pol η polymerase[26] and the other is the archaebacterial analogue of the *Escherichia coli* dinB protein.[27,28] Both were found in the closed conformation, even though they were crystallized in their unliganded form. The general interpretation of this phenomenon is that polymerases with low fidelity do not display the open/closed transition, which is needed only when high fidelity is needed.[22]

An additional line of evidence in favour of this interpretation can be found in the recent structure of murine terminal desoxynucleotidyltransferase (TdT).[29] TdT, whose sequence is closely related to pol μ, a newly discovered human polymerase thought to be involved in hypermutagenesis,[30] turns out to be structurally highly similar to the structure of pol β, which is itself involved in DNA repair. TdT can be considered as an extreme case of an error-prone polymerase since it is a template-independent polymerase, which adds random nucleotides to the N regions of V(D)J junctions of immunoglobulin genes, thereby contributing to the generation of the diversity of response of the immune system. Again, TdT was found to resemble most closely to the closed form of pol β, even in the unliganded state. A detailed analysis of the structure suggests that TdT may be permanently locked in the closed form in solution.[29]

Pol β actually stands as a representative member of an entire family comprising pol X type polymerases (pol β and pol λ), template independent polymerases (pol μ and TdT, poly (A) polymerase, oligo(A) 2′-5′ polymerase) as well as several other nucleotidyltransferases.[31]

It seems clear that the open/closed transition is of utmost functional significance in polymerases. Several questions can be raised: is this transition encoded in the structure itself? Is it an intrinsic property, a natural tendency of this mechanical system? Can it be predicted from the structure alone?

We have sought further clarification of this conformational transition by studying the normal modes calculated from both the open and closed forms, in all cases where they are available. The reason behind using Normal Mode Analysis (NMA) is that we are looking for large amplitude movements that are unlikely to occur during the time scales available to the molecular dynamics method.[32–35] Instead, the lowest frequency modes of NMA should give some clues about the correlated movements likely to occur in these large proteins. Because they are highly collective movements, the lowest frequency modes are the ones with the largest amplitudes, at a given temperature (and energy).

Whenever structural information is available both for the open and the closed forms, it is possible to do a direct comparison between the movements predicted by the theory for the different modes and the difference vectors between the closed and open forms.

Here, we show that the open/closed transition can be mapped to a handful of the lowest frequency normal modes determined directly from the structure of the open form using a simple and computationally inexpensive method borrowed from the field of structural mechanics.[36] The method, which essentially says that all atoms are

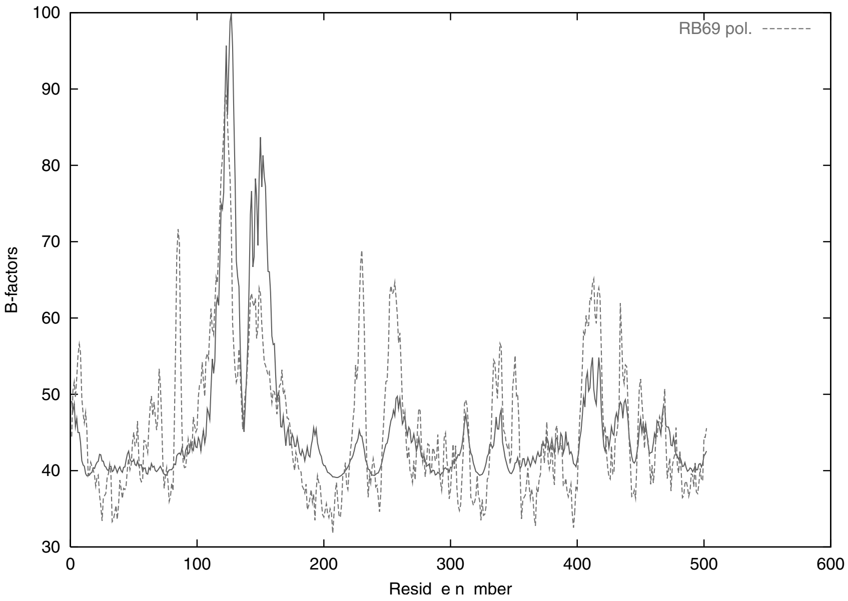

Figure 1. Plot of experimental (dotted line) and calculated (continuous line) temperature factors, using the Elastic Network Model, as a function of residue number, for phage RB69 polymerase (PDB code 1IH7, pol I structural class and pol B, or pol α, sequence class). The numbering of the sequence begins at residue 383.

linked by springs, i.e. experiencing an harmonic potential bringing them back to their equilibrium position, was simplified by Hinsen[37,38] who showed that it was possible to use only C^{α} coordinates instead of all atoms. Bahar and colleagues[39,40] also developed the method along slightly different lines involving the inversion of a Kirchhoff contact matrix.

Lowest frequency modes calculated by this method (The Elastic Network Model) can be shown to be sufficient to explain the open/closed transition for all the cases where both the open and closed forms are known, in the different structural classes of pol I and pol α, pol β and multisubunit RNA polymerases. The latter case, which concerns very large proteins of about 3500 residues,[9,41] was made possible by a recent and even more drastic simplification of the method, which was shown recently to perform well in test cases,[42,43] and which consists in grouping residues in super particles.[44,45] Normal mode analysis of systems of such a large size would simply be impossible with currently available computers, using standard methods, or would require large amounts of CPU-time on supercomputers, using iterative methods like DIMB.[46]

Although this method uses very grossly approximated energy functions (for instance, there is no explicit account of the solvent), it has the advantage of computational speed, allowing different hypotheses to be tested quickly. Also, it does not require energy minimization of the X-ray structure prior to normal mode calculations since it assumes that the X-ray structure is the minimum of the energy function. The method can also be used to make predictions in cases where only one form is known, as in the case of the RNA-dependent RNA polymerase of hepatitis C virus.

Results

Correlation between the calculated and the observed *B*-factors.

To test the validity of the Elastic Network Model, the experimental crystallographic *B*-factors were compared to the mean quadratic displacement values of each of the C^{α} atoms, based on a subset of the 100 lowest frequency normal modes. This is illustrated in Figure 1 for pol α, as a function of residue number, and in Table 1 for all known structures of DNA-dependent polymerases. In most cases, the correlation coefficient between these two quantities is around 50% (see Table 1). This compares well with results previously described using a closely related method[39] and also with a recent survey on 40 high-resolution structures, where the method used here gave a correlation coefficient of 0.52 (±0.15) (Tama, Marques & Sanejouand, unpublished results). In those cases where the correlation coefficient is lower (*Taq* pol I and pol β), i.e. around 30% only, it was verified that this value could be increased if the DNA was included in the model (Table 1). In this case, the DNA was represented with only three atoms per nucleotide, which were chosen so that the sampling was roughly equivalent with the sampling of one C^{α} atom per residue used to represent the polypeptide chain. Only those base-pairs buried by the protein were considered (about four base-pairs).

Projection of the open/closed transition onto the lowest frequency normal modes

For every crystallographic DNA-dependent polymerase known in at least two forms, each of the lowest frequency normal mode displacement vectors attached to each C^{α} position was projected

Table 1. X-ray structures of DNA-dependent polymerases used in this study

Name	Most significant lowest mode(s)	*B*-factors correlation coefficent (w/o DNA)	Struct. class (X-ray)	Seq. class according to Ref. 4	PDB code	Resolution (in Å)	Number of residues	Form
Phage RB69 DNA pol	#1 (0.63); #2 (0.71)	0.713	Pol I	Pol B (pol α)	1ih7	2.20	502	Open
Phage RB69 DNA pol	#1 (0.42); #8 (0.47)	0.530	Pol I	Pol B (pol α)	1ig9	2.60	502	Closed
T. aquaticus DNA pol I	#4 (0.51)	0.29 (0.51)	Pol I	Pol A (pol I)	2ktq	2.30	528	Open
T. aquaticus DNA pol I	#11 (0.51)	0.32 (0.35)	Pol I	Pol A (pol I)	3ktq	2.30	528	Closed
DNA pol β	#1 (0.73)	0.27 (0.39)	Pol β	Pol X (pol β)	1bpx	2.40	326	Open
DNA pol β	#1 (0.46) #2 (0.46)	0.36 (0.33)	Pol β	Pol X (pol β)	1bpy	2.20	326	Closed
TdT	NA	0.58	Pol β	Pol X (pol β)	1jms	2.35	360	Closed
Pol η	NA	0.36	Pol I	Pol Y (pol IV/V)	1jih	2.25	509	Closed
dinB free full length	#3 (0.73)	0.51	Pol I	Pol Y (pol IV/V)	1k1q	2.80	333	Closed free
dinB bound full length	#9 (0.50)	0.37 0.48	Pol I	Pol Y (pol IV/V)	1jx4 1jxl	1.70 2.10	341 341	Closed bound
Yeast RNA pol II	#1 (0.69)	0.77	RNA pol	RNA pol multisubunit	1i6h	3.30	3500	Open
Yeast RNA pol II	#1 (0.64)	0.61	RNA pol	RNA pol multisubunit	1i50	2.80	3500	Closed
Bacterial RNA pol II	#1 (0.51)	NA	RNA pol	RNA pol multisubunit	1hqm	3.30	2687	Open
Bacterial RNA pol II	#1 (0.54)	NA	RNA pol	RNA pol multisubunit	ftp	15.0 (electron microscopy)	2687	Closed
Phage T7 RNA pol	#1 (0.66)	0.56	Pol I	RNA pol monosubunit	1aro	2.80	774	Open free
Phage T7 RNA pol	#1 (0.59)	0.42	Pol I	RNA pol monosubunit	1cez	2.40	774	Binary complex

The correlation coefficient between the predicted and the experimental *B*-factors is given, as well as the normal mode numbers which give a correlation larger than 0.5 for the predicted displacement vectors and the observed difference vectors. Only the lowest 100 lowest frequency modes are considered. The PDB code, highest resolution and number of residues are indicated for each structure, as well as the belonging to one of the sequence classes defined in Ref. 4 and to one of the three possible structural classes found up to now. The structural classes are named after the first representative member that was solved by X-ray crystallography. Alternative designations of the sequence classes are indicated in parenthesis: pol A is also known as pol I, pol B as pol α and pol X as pol β, as in Ref. 6. The polymerases are grouped so that DNA-dependent DNA polymerases are on top and DNA-dependent RNA polymerases are at the bottom of the Table.

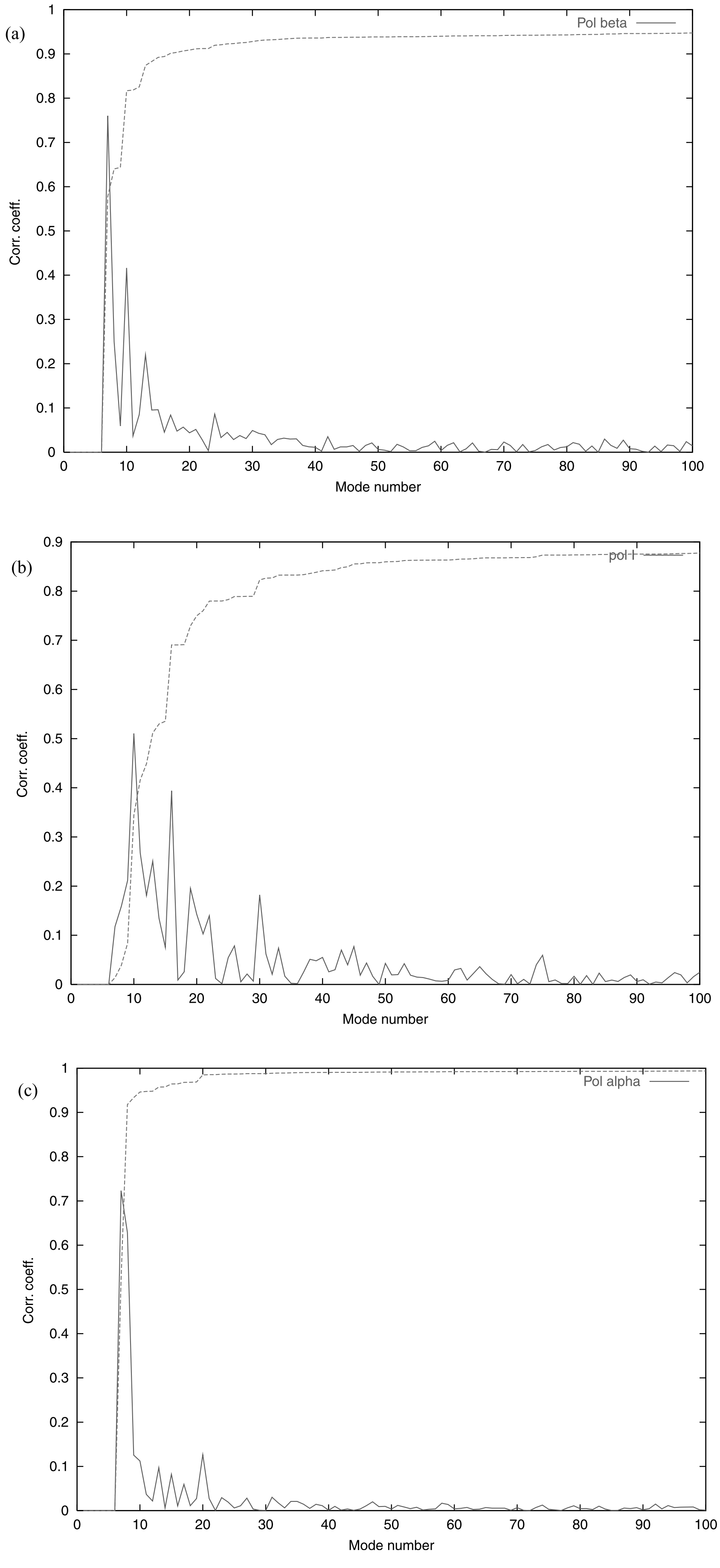

Figure 2. Mean cosine of the projection of the displacement vectors of each of the 100 lowest normal modes onto the open/closed difference vectors for DNA-dependent DNA polymerases. The dotted line is the cumulated square cosine. (a) Pol β (1BPX) (b) pol I (2KTQ); (c) pol α (1IH7). In this representation, modes 1–6 should be ignored: they represent overall translation and rotation movements to superpose the two forms of the protein, through a rigid body movement. All pairs of structures were superimposed with Profit (see the text) prior to calculations.

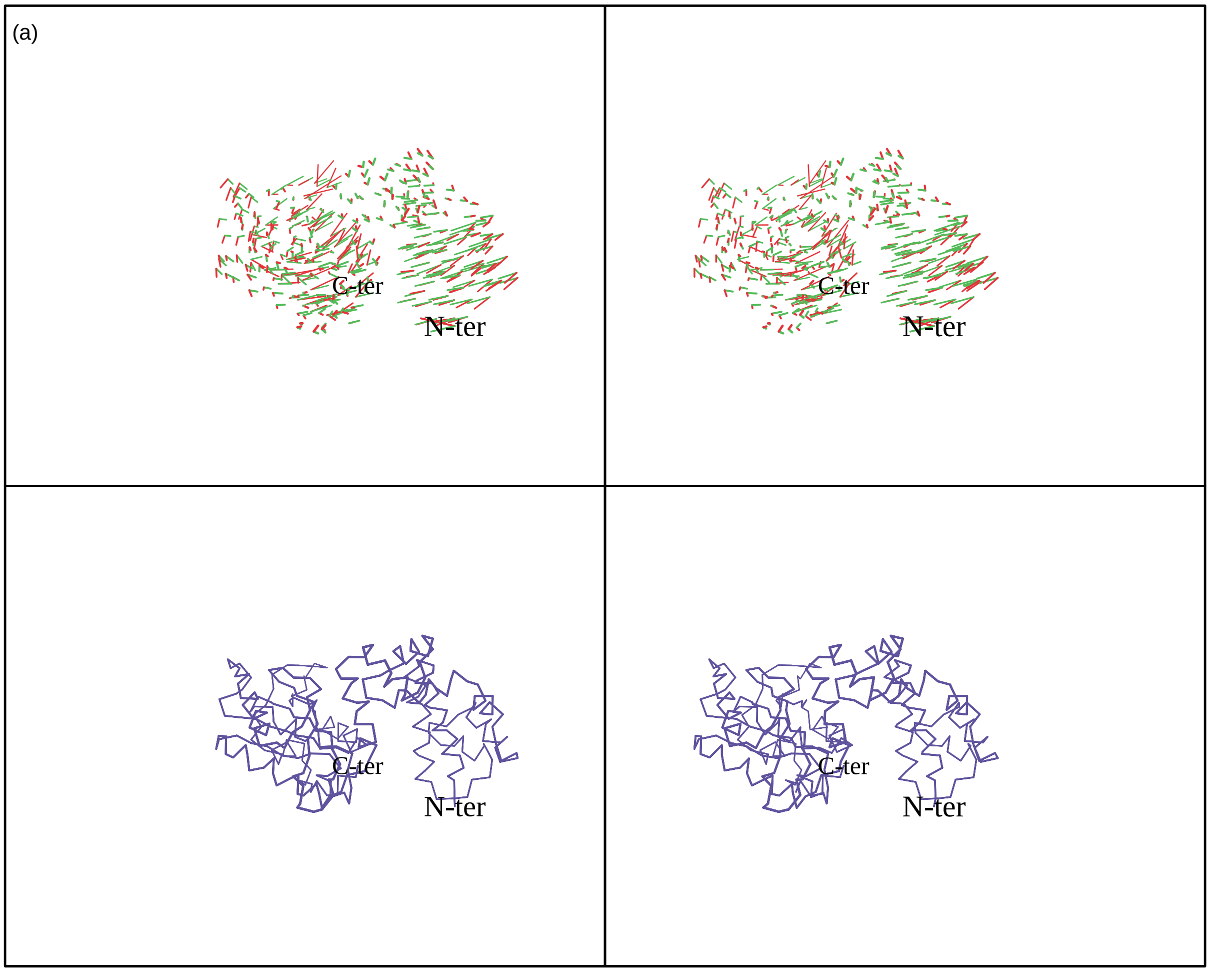

Figure 3 (*legend opposite*)

onto the experimental difference vector between the two forms. The result is quite impressive: in most cases, the lowest frequency mode has a mean cosine of about 0.50 or more (see Table 1), while most of the other modes have a negligible correlation coefficient (Figures 2 and 4). This mean cosine is a generalized correlation coefficient, averaged over all C^{α} displacement vectors along the sequence. Note that the sum over all modes of the square of the mean cosine of each mode is one, by definition, as the normal modes form a basis set. The two sets of vectors (calculated and observed) are plotted for each C^{α} position in Figure 3 both for pol β and one member of the pol α family.

The normal modes have been calculated from either the closed or the open forms. It was found, in accordance with earlier studies,[47] that the projection onto the difference vectors was always better if the normal modes were calculated from the open form (Table 1). Also, the lowest frequency mode of the open form is in general lower (e.g. 7 cm^{-1} instead of 16 cm^{-1} for pol α) than the one of the closed form, leading to larger amplitude movements, according to normal mode theory.

In two cases, pol β and pol I, the influence of including the DNA in the model was tested and normal mode displacement vectors were calculated from the protein + DNA complex, instead of from the protein alone, but the results were found to be slightly inferior (data not shown).

For pol β (human polymerase β, pol X family), it is the C-terminal domain that moves the most in the open/closed conformational transition, around an axis that runs parallel to the axis of helix M[16]. In this case, the result is especially striking: the lowest frequency mode of the open structure is almost sufficient to describe this transition (see Figure 2(a)).

For pol I (*Thermus aquaticus* pol I, pol A family), the tip of the finger domain rotates about 45° inwards towards the active site in a combined hinge-like motion involving the repacking of several helices.[13,15] Among them, helix O (containing sequence motif B[6]) is of utmost importance to form the dNTP binding site. This overall movement is well described by the combination of several (three to four) of the lowest normal modes calculated from the open form (see Figure 2(b)).

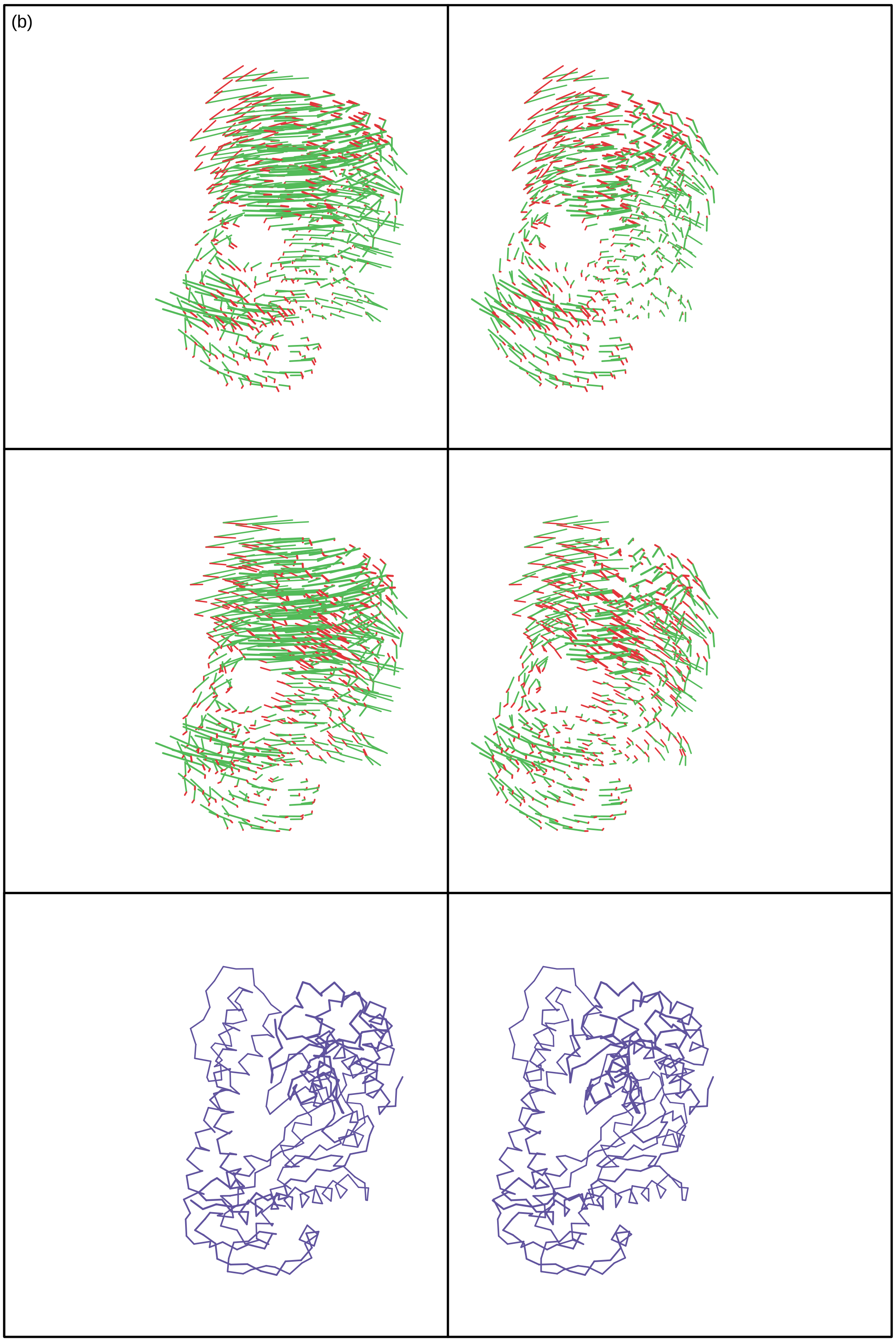

Figure 3. Stereo representation of the observed difference vectors (in red) and the eigenvectors (in green) of the lowest normal mode for pol β (1BPX) in (a) and of the two lowest normal modes of pol α (1IH7) in (b). Drawn with Molscript.[63] The C^{α} trace of the molecule is also shown at the bottom of the molecule in blue. An approximate scale factor was applied to the calculated displacement vectors, so as to match the mean magnitude of the experimental ones.

For pol α (phage RB69 polymerase, pol B family), a direct comparison between the binary complex of the protein with DNA and the ternary complex with DNA and incoming dNTP is not possible since only the structures of the free polymerase and the (replicating) ternary complex are available. Both a movement of the thumb domain to hold a solid grip on the DNA and a transition to the closed conformation of the fingers domain (similar to what is observed in pol A family) are observed[48] and they are altogether very well described by just the two lowest frequency modes of the free polymerase (Figure 2(c)).

We have also performed NMA analysis on phage RB69 polymerase in the editing mode (PDB code 1CLQ[49]). The best correlation coefficients between difference vectors with the apo form and predicted displacement vectors are observed for modes 2, 7, 14 and 22 (0.36, 0.41, 0.31 and 0.34, respectively). In this case, it seems that the transition between the two forms is a complicated one, with the superposition of at least four different normal modes contributing roughly equally with no dominant mode. The same is true when comparing the replicating and editing complexes (rmsd 6.5 Å) where the dominant modes are modes 5, 6, 7, 15, with correlation coefficients of 0.32, 0.36, 0.35 and 0.42.

For the monosubunit T7 DNA-dependent RNA polymerase, however, the transition between the free and the bound forms (a binary complex with a 17 bp promoter[50]) is again extremely well predicted by just the lowest mode (see Table 1 and Figure 4(a)).

It is especially impressive to see the result of the Elastic Network Model for a very large assembly of atoms such as the yeast multisubunit RNA pol II. Most of the transition between the two crystallized eukaryotic forms[9] is embedded in only two of the lowest frequency modes (see Table 1 and Figure 4(b)). For bacterial multisubunit DNA-dependent RNA polymerase it has recently become possible to apply NMA using both the *T. thermophilus* crystal structure[51] (at 3.3 Å resolution) and a model fitted into the electron density deduced from electron microscopy data at 15 Å of the *E. coli* enzyme.[52] The two structures show large structural rearrangements that are again well fitted by just the lowest frequency normal mode (Table 1 and Figure 4(c)).

Are there some polymerases locked into a closed form: insight into low-fidelity, distributive polymerase structure and function

In the recent description of the structure of TdT,[29] which was found to resemble the closed form of pol β in its crystal form, we hypothesized that TdT might be permanently locked in the closed form in solution. There were two main reasons for this proposal: (i) the existence of a 16-residue long N-terminal extension that touches upon the C-terminal domain, thereby sealing the closed form and (ii) the non-conservation in TdT of an arginine residue which is making a salt bridge with a conserved aspartate of motif A in the open form of pol β. Here we provide a more quantitative measure of the first argument: the lowest frequency mode of TdT is about twice higher (44 cm^{-1}), and therefore, according to normal mode theory, of smaller amplitude than the lowest frequency mode of the closed form of pol β (28 cm^{-1}), while the number of residues in both structures is comparable.

Two members of the pol Y family have been solved recently in different crystal forms.[26–28] Because they are error-prone polymerases, it has been hypothesized that the induced-fit transition between the closed and open forms might not occur in these polymerases. Indeed, their finger domains were both found in the closed form in an unliganded state.

The full-length free dinB protein structure shows rather small thumb and finger domains, indicating that the contact with the DNA may not be very extensive. In particular, the finger domain lacks the essential helix O of pol A polymerases and possesses a topology that is typical of the N-terminal domain of pol β instead, i.e. the structural motif HhH. dinB has an additional C-terminal "little finger" domain (also called the "wrist" domain) that touches upon the finger domain.[27] The finger domain is in a closed conformation, and this observation is also true in both the truncated dinB[28] and the free yeast pol η structures.[26] Experimentally, the recent crystal structure of a ternary structure of Dpo4 (dinB) polymerase with DNA and the incoming nucleotide[53] shows that the finger domain does not close any further upon binding the DNA substrate, with limited and non-specific contacts with the replicating base-pair. However, the thumb and the wrist (little finger) domains do move a lot upon DNA and incoming dNTP binding.

Examination of the movements corresponding to the lowest frequency normal modes of the unliganded full length dinB structure shows no movement of the fingers domain but, interestingly, more global movements of the other domains, such as the ones observed upon binding of the DNA substrate, resulting in a very good overlap of the predicted and observed difference vectors between the liganded and unliganded forms (see Table 1).

Extension to other polymerases

Recently, the crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus was solved.[54–57] This form was described as a closed form, especially because of the presence of an unusual piece of the polypeptide chain, which joins the thumb and finger domains. Soaking these crystals into a solution containing dNTPs shows density in the dNTP binding site without major rearrangement of the domains, which indicates

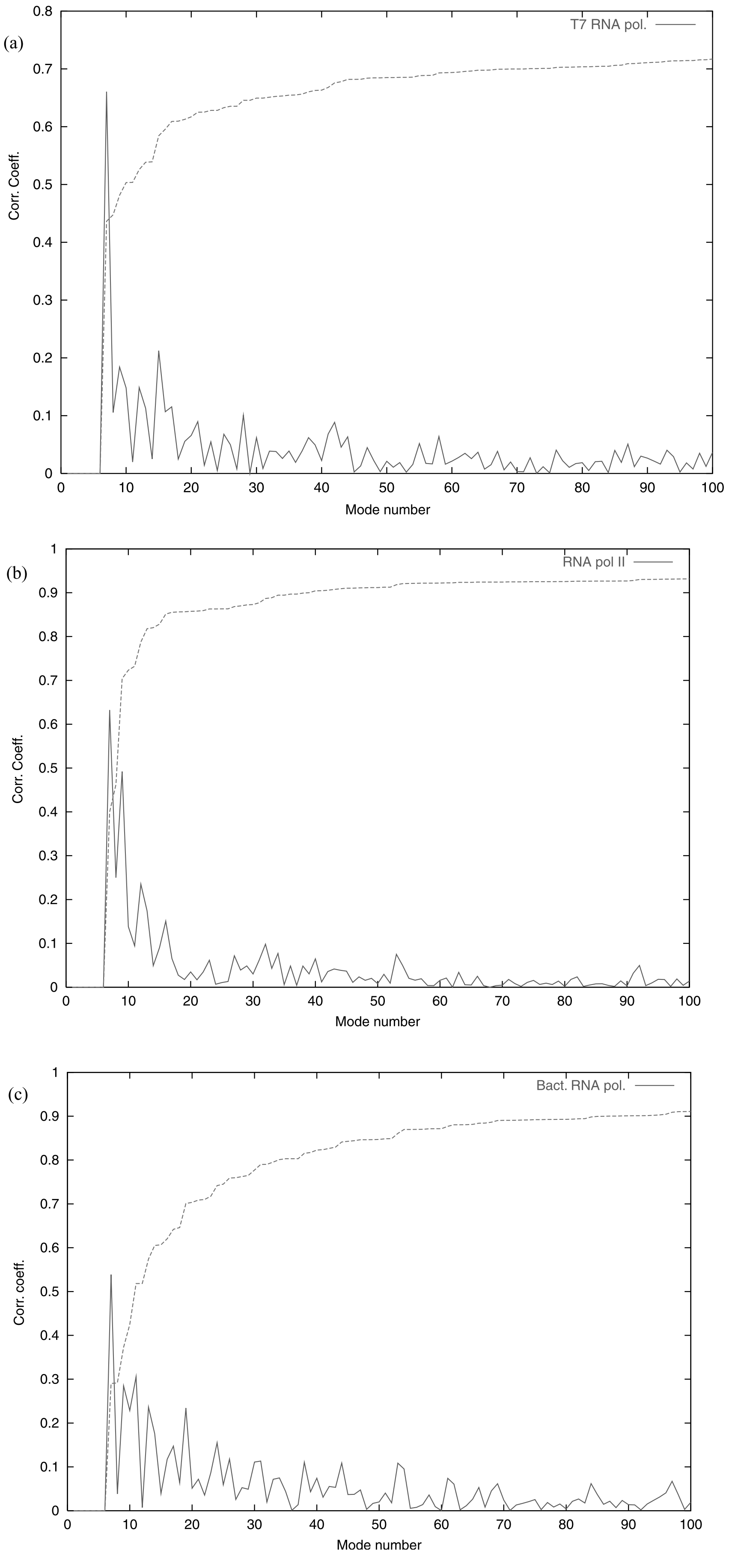

Figure 4. Mean cosine of the projection of the displacement vectors of each of the 100 lowest normal modes onto the open/closed difference vectors for DNA-dependent RNA polymerases. The dotted line is the cumulated square cosine. In this representation, modes 1–6 should be ignored: they represent overall translation and rotation movements to superpose the two forms of the protein, through a rigid body movement. All pairs of structures were superimposed with Profit (see the text) prior to calculations. (a) T7 DNA-dependent monosubunit RNA polymerase (PDB code 1ARO). (b)Yeast DNA-dependent RNA polymerase II (PDB code 1I6H). (c) Bacterial DNA-dependent RNA polymerase (PDB code 1HQM).

that the dNTP binding site is already preformed in this crystal form.[57]

This polymerase is very processive and has a fidelity that is comparable to other polymerases lacking the 3′-5′ exonuclease activity. These two characteristics are classically associated to the presence of at least two forms, alternating during the elongation process.

We used the Elastic Network Model to calculate the lowest frequency modes of this protein. The two lowest modes (65 and 70 cm^{-1}) were found to be well detached from all the others (next ones at 99 cm^{-1}, 115 cm^{-1},…) and to have a high degree of collectiveness[47] (0.65 and 0.56).

Of these two modes, the second one gives a more open form of the polymerase (or an even more closed form since the polarity of the movement cannot be decided by the method, but then this would apparently prevent the binding of the template strand and therefore is highly unlikely), by a concerted movement of both the finger and thumb domains, the latter contributing more than the former to the opening transition.

Discussion

Why does the simplified NMA method work so well?

It is surprising that such a crude and low-resolution model gives good results for the prediction of large amplitude movements in polymerases.

Stereochemistry is not taken into account in this model, neither the excluded volume effect nor electrostatics. Rather, the protein is modelled as a solid at zero temperature, where solvent effects are effectively ignored. The range of frequencies predicted by the model span the range of acoustic modes in a solid and agrees with experimental values measured by inelastic neutron scattering experiments below the so-called dynamical transition.[58–59] However, it must be pointed out here that all frequencies are calculated up to a single scaling constant, which has been adjusted by Tirion[36] to fit results obtained by other methods and is set to 10 kcal/(mol Å^2) here.

If one attempts to relate the absolute value of the calculated lowest frequency (about 10–30 cm^{-1}) to the time scale of the open/closed transition in polymerases (100–300 s^{-1}), one ends up with a value of about 30 m s^{-1} for the speed of propagation of this wave, which is typical for acoustic modes in a solid.

Modelling proteins as solids is perhaps not such a bad approximation given the packing of side-chains in proteins, which are as tightly packed as atoms in crystals of organic molecules.[60]

The potential energy that is used in the Elastic Network Model is essentially an *ad hoc* potential that tends to preserve the relative distances between close residues in space. Therefore, the expected movements described by this model are collective ones. As such, it is maybe not so surprising that whole domain movements are well predicted by this model. But the high correlation between the difference vectors of the two known forms and just a few of the lowest frequency normal modes was not *a priori* expected.

This may be related to the recent observation that biologically relevant movements in proteins form a very tiny subspace of all possible movements.[61] For instance, it is possible to extract from molecular dynamics simulations the so-called “essential modes”, by diagonalization of a displacement correlation matrix in Cartesian coordinate space. Principal component analysis is then used to project the movements onto the lowest (collective) modes of this analysis, which are not necessarily vibrational, and this approach has proved successful in a number of cases where just the ten lowest modes could reproduce most of the molecular dynamics simulation. The resulting low dimensional subspace is sufficient to describe large amplitude movements and biologically relevant transitions between different conformations of proteins. Indeed, low-frequency modes represent an important fraction of these “collective modes”, which are relatively insensitive to solvent effects and initial or other conditions.

However, one intriguing point is why just a handful of modes can be singled out among other highly collective ones, of similar frequencies. One possible explanation, taking into account the fact that most often the “significant” modes are among the three lowest frequency ones, is that these modes are among the few ones needed to describe the rigid-body motions of the largest “collective” domains of the given macromolecular system. Reciprocally, normal mode analysis might be viewed as a tool to characterize collective domains. Our results indicate that biologically relevant large amplitude motions are essentially rigid-body motions of such domains.

For pol β, only one mode is necessary to describe the transition between the open and the closed forms, but the rmsd between them is only 2.7 Å. For pol α, more than one mode (two) is needed to describe the transition that can be derived from crystallographic structures deposited in the PDB, but in this case the transition (rmsd 6–7 Å) is between the apo form and the closed form, and not between a binary and a ternary complex. For pol I, the existence of several modes with a good correlation coefficient may be linked to the necessity of repacking several helices during the transition.

What is the predictive power of the method?

In an attempt to characterize more fully the lowest frequency normal modes, the following approach has been followed: each residue is scanned in turn and its associated mass is increased by a factor of 100 compared to the other residues. Then, the shift in the frequency of each of the ten lowest modes is recorded. In this manner, residues whose mass contributes most to these low frequency modes are highlighted and a residue-by-residue "signature" is being built for each of the ten lowest frequency normal modes. In general, we find that for hinge motions of loosely connected domains, the residues that matter most are the ones at the tip of the distal domains.

We have tried to apply this method to cases were only one form is known, namely the open form. For instance, for the free (unliganded) form of RNA-dependent RNA polymerase of Hepatitis C virus (HCV, PDB code 1CSJ), two well-detached low frequency modes were found. One of them (the second lowest frequency one) clearly points to a large movement of the thumb domain. The structure of the ternary complex of HCV bound with all its substrates is not known but the recent structure of phage Phi6 RNA-dependent RNA polymerase[62] may provide some clues of what is expected upon binding of the oligoribonucleotide substrate, because it belongs to the same family as HCV polymerase and was crystallized both in the absence and in the presence of an oligoribonucleotide template (PDB codes 1HHS and 1HHT). While there is almost no difference between these latter two forms, there is a rotation of the thumb domain by 35° compared to the HCV polymerase, in a direction that is compatible with the one predicted by the second lowest frequency mode of the HCV structure. However, a quantitative correlation coefficient between the predicted movements for this normal mode and the observed difference vectors obtained with the palm domains aligned cannot be easily derived as there is a large rigid body component in the thumb domain movement and the number of structurally aligned and consecutive positions in the thumb domains of HCV and Phi6 polymerases is small. The confirmation of the validity of the normal mode analysis in this case therefore has to await further structural information on the ternary complex of HCV polymerase.

A novel tool for low resolution studies

One particularly attractive feature of this method is its computational speed. In the future, it might be possible to systematically probe the surface of the protein, at low resolution, by randomly adding pseudoatoms on the surface, and see how many of the tried positions dramatically affect the normal modes of the protein. Indeed, if the frequency of the lowest normal mode is increased, this is a sign of a reduced amplitude for the transition between the two forms. In this way, it might be possible to find inhibitors of the dynamical transition necessary to the function of the protein. This would open a new avenue of research to conduct drug design against viral DNA polymerases such as those from the pol α family, which contain, among others, Epstein Barr virus, hepatitis B virus, Herpes Simplex virus, adenovirus and vaccinia virus. Preliminary work indicates that to modify substantially the lowest frequency modes, ligands of the size of oligopeptides (or oligosaccharides) are needed.

One of the most promising outcomes of this work is the application of simplified NMA to help interpret conformational changes seen in low-resolution maps derived from electron microscopy data. Indeed, it has been possible in this study to get good results in the case of the recently released structural results on bacterial DNA-dependent RNA polymerases.[62] Further work is in progress to assess its usefulness in other systems involving large molecular assemblages.

Materials and Methods

The PDB coordinates of the closed and open forms of DNA-dependent polymerases are the following: 1BPX and 1BPY for pol β[16] (pol X family); 2KTQ and 3KTQ for *Taq* pol I[15] (pol A family); 1IH7 and 1IG9 for phage RB69 polymerase, a representative member of the pol α family[5,48] (pol B family); 1I50 and 1I6H for yeast multisubunits RNA polymerase II[9,41]; 1CEZ and 1ARO for the monosububit T7 RNA polymerase[7,50] and 1HQM[51] for *T. thermophilus* multisubunit RNA polymerase. The *E. coli* model fitted into the 15 Å electron microscopy map was obtained from the Web†. The resolution and the number of residues of each structure are indicated in Table 1. The PDB code for the TdT structure is 1JMS.[29] For pol η, it is 1JIH[26] and for the full-length dinB it is 1K1Q or 1JX4 in the free and liganded forms, respectively[27,53] (pol Y family). The open forms 1BPX and 2KTQ of pol β and pol I were preferred to 1BPZ and 4KTQ, respectively, because they were solved at higher resolution.

Only the C^α coordinates were retained and those closer than 10 Å from each other were connected by harmonic springs with the same force constant, in the spirit of Tirion's model.[36] The closed and open forms were superposed using the program Profit prior to all calculations (A.C. Martin. Profit v1.8‡). Residues present in one of the two forms but not the other were ignored for the superposition.

When taking DNA into account, each nucleotide was represented by three atoms: the P of the phosphate, C2 of the base and C4′ of the sugar. This reflects the difference in the average mass for a nucleotide (330) and an amino acid residue (110). The distance between C2 atoms of the same base-pair is about 3.6–4.2 Å, comparing well with the distance between C^α of successive amino acid residues (3.8 Å).

Calculation of normal modes was performed as described.[45] This involves diagonalization of $\mathbf{H}$, the matrix of second derivatives of the potential energy V of the protein described as a set of harmonic springs of the same strength k linking the C^α atoms less than 10 Å from each other, namely $V = \sum_{i<j} k(d_{ij} - d_{ij}^0)^2$ where d_{ij}^0 is the crystallographic distance between atoms i and j.

From the eigenvectors and eigenvalues of $\mathbf{H}$, the mean-square displacements $\langle R^2 \rangle$ were calculated and compared to experimental crystallographic B-values

† ftp://ftp.scripps.edu/pub/wriggers/dars02
‡ http://www.biochem.ac.uk/~martin/programs

through the relation $B = 8\pi^2/3\langle R^2\rangle$, after linear rescaling. Each of the 100 lowest frequency normal mode displacement vectors was projected onto the open/closed difference vectors, for each residue in the protein, and the corresponding mean cosine (correlation coefficient) was calculated. The index of collectiveness of each mode was defined and calculated as described.[47] All programs used here are available upon request (sanejouand@crpp.u-bordeaux.fr).

For large protein cases, the grouping of residues into superresidues (blocks) was used, the rigid-body rotations and translations of each block being used as a new set of coordinates, instead of the Cartesian ones, as described previously.[44,45] Note that for DNA-dependent yeast RNA polymerase II, the calculation would not have been possible without this trick (there are 3500 residues in this protein). In that case, ten consecutive C^α atoms were put into each block. The results were found to be relatively insensitive to this choice, within the range of five to ten residues per block. Care was taken not to include atoms of different subunits in the same block.

Acknowledgments

We thank S. Bressanelli and F. A. Rey for communicating results prior to publication on HCV RNA polymerase complexed with various dNTPs.

This study was initiated in the framework of the G.D.R. "Simulation et Modelisation des molecules biologiques" (CNRS), coordinated by T. Simonson (IGBMC, Illkirch, France). Financial support from Association de la Recherche contre le Cancer (grant no. 5470 to M.D.) is gratefully acknowledged.

References

1. Baker, T. & Bell, S. P. (1998). Polymerases and the replisome: machines within machines. *Cell*, **92**, 295–305.
2. Ollis, D., Brick, P., Hamlin, R., Xuong, N. G. & Steitz, T. A. (1985). Structure of the large fragment of *E. coli* DNA polymerase I complexed with TMP. *Nature*, **313**, 762–766.
3. Steitz, T. A. (1999). DNA polymerases: structural diversity and common mechanisms. *J. Biol. Chem.* **274**, 17395–17398.
4. Ito, J. & Braithwaite, D. K. (1991). Compilation and alignment of DNA polymerase sequences. *Nucl. Acids Res.* **19**, 4045–4057.
5. Wang, J., Sattar, A. K., Wang, C. C., Karam, J. D., Konigsberg, W. H. & Steitz, T. A. (1999). Crystal structure of a pol alpha family replication DNA polymerase from bacteriophage RB69. *Cell*, **89**, 1087–1099.
6. Delarue, M., Poch, O., Tordo, N., Moras, D. & Argos, P. (1990). An attempt to unify the structure of polymerases. *Protein Eng.* **3**, 461–467.
7. Sousa, R., Chung, Y. J., Rose, J. P. & Wang, B. C. (1993). The crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. *Nature*, **364**, 593–598.
8. Pelletier, H., Sawaya, M., Kumar, A., Wilson, S. H. & Kraut, J. (1994). Structures of ternary complexes of rat DNA polymerase β, a DNA template primer and ddCTP. *Science*, **264**, 1891–1903.
9. Cramer, P., Bushnell, D. A. & Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science*, **292**, 1863–1870.
10. Steitz, T. A. & Steitz, J. A. (1993). A general two-metal ions mechanism for catalytic RNA. *Proc. Natl Acad. Sci. USA*, **90**, 6498–6502.
11. Doublié, S., Sawaya, M. R. & Ellenberger, T. (1999). An open and closed case for all polymerases. *Structure*, **7**, R31–R35.
12. Patel, P. H. & Loeb, L. A. (2001). Getting a grip on how polymerases function. *Nature Struct. Biol.* **8**, 656–659.
13. Doublié, S., Tabor, S., Long, A. M., Richarson, C. C. & Ellenberger, T. (1998). Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature*, **391**, 251–258.
14. Kiefer, J. R., Mao, C., Braman, J. C. & Beese, L. S. (1998). Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. *Nature*, **391**, 304–307.
15. Li, Y., Korolev, S. & Waksman, G. (1998). Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of *T. aquaticus* DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.* **17**, 7514–7525.
16. Sawaya, M. R., Prasad, R., Wilson, S. H., Kraut, J. & Pelletier, H. (1997). Crystal structures of DNA polymerase beta complexed with gapped and nicked DNA: evidence for an induced fit mechanism. *Biochemistry*, **36**, 11205–11215.
17. Poch, O., Sauvaget, I., Delarue, M. & Tordo, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J.* **8**, 3867–3874.
18. Madrid, M., Jacobo-Molina, A., Ding, J. & Arnold, E. (1999). Major subdomain rearrangement in HIV-1 reverse transcriptase simulated by molecular dynamics. *Proteins: Struct. Funct. Genet.* **35**, 332–337.
19. Madrid, M., Lukin, J. A., Madura, J. D., Ding, J. & Arnold, E. (2001). Molecular dynamics of HIV-1 reverse transcriptase indicates increased flexibility upon DNA binding. *Proteins: Struct. Funct. Genet.* **45**, 176–182.
20. Bahar, I., Erman, B., Jernigan, R. L., Atligan, A. R. & Covell, D. G. (1999). Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J. Mol. Biol.* **285**, 1023–1037.
21. Wong, I., Patel, S. S. & Johnson, K. A. (1991). An induced fit kinetic mechanism for DNA replication fidelity: direct measurements by single-turnover kinetics. *Biochemistry*, **30**, 526–537.
22. Beard, W. A. & Wilson, S. H. (2001). DNA lesion bypass: polymerases open up. *Structure*, **9**, 759–764.
23. Ellenberger, T. & Silvian, L. F. (2001). The anatomy of infidelity. *Nature Struct. Biol.* **8**, 827–828.
24. Ohmori, H. *et al.* (2001). The Y family of DNA polymerases. *Mol. Cell*, **8**, 7–8.
25. Johnson, R. E., Washington, M. T., Prakash, S. & Prakash, L. (2000). Bridging the gap: a family of novel DNA polymerases that replicate faulty DNA. *Proc. Natl Acad. Sci. USA*, **96**, 12224–12226.
26. Trincao, J., Johnson, R. E., Escalante, C. R., Prakash, S., Prakash, L. & Aggarwal, A. K. (2001). Structure of the catalytic core of *S. cerevisiae* DNA polymerase

η implications for translation DNA synthesis. *Mol. Cell*, **8**, 417–426.

27. Silvian, L. F., Toth, E. A., Pham, P., Goodman, M. F. & Ellenberger, T. (2001). Crystal structure of a DinB family error-prone DNA polymerase from *Sulfolobus solfataricus*. *Nature Struct. Biol.* **8**, 984–989.
28. Zhou, B.-L., Pata, J. D. & Steitz, T. A. (2001). Crystal structure of a DinB Lesion bypass DNA polymerase catalytic fragment reveals a classic polymerase catalytic domain. *Mol. Cell*, **8**, 427–437.
29. Delarue, M., Boulé, J. B., Lescar, J., Expert-Bezançon, N., Jourdan, N., Sukumar, N. *et al.* (2002). Crystal structure of a template-independant DNA polymerase: murine terminal desoxynucleotidyltransferase. *EMBO J.* **21**, 427–439.
30. Aoufouchi, S., Flatter, E., Dahan, A., Faili, A., Bertocci, B., Storck, S. *et al.* (2000). Two novel human and mouse DNA polymerases of the pol X family. *Nucl. Acids Res.* **28**, 3684–3693.
31. Aravind, L. & Koonin, E. V. (1999). DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucl. Acids Res.* **27**, 1609–1618.
32. Go, N., Noguti, T. & Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl Acad. Sci. USA*, **80**, 3696–3700.
33. Brooks, B. & Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine trypsin inhibitor. *Proc. Natl Acad. Sci. USA*, **80**, 6571–6575.
34. Levitt, M., Stern, P. & Sander, C. (1985). Protein normal-mode dynamics. *J. Mol. Biol.* **181**, 423–447.
35. Kidera, A. & Go, N. (1992). Normal mode refinement: crystallographic refinement of protein dynamic structure. *J. Mol. Biol.* **225**, 457–475.
36. Tirion, M. (1996). Large amplitude elastic motions in proteins from a single parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908.
37. Hinsen, K. (1998). Analysis of domain motions by approximate normal mode analysis. *Proteins: Struct. Funct. Genet.* **33**, 417–429.
38. Hinsen, K., Thomas, A. & Field, M. J. (1999). Analysis of domain motions in large proteins. *Proteins: Struct. Funct. Genet.* **34**, 369–382.
39. Bahar, I., Atligan, A. R. & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Design*, **2**, 173–181.
40. Atligan, A. R., Durell, S. D., Jernigan, R. L., Demirel, M. C., Keskin, O. & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**, 505–515.
41. Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A. & Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II elongation complex at 3.3 Å resolution. *Science*, **292**, 1876–1879.
42. van Vlijmen, H. W. & Karplus, M. (2001). Normal mode analysis of large systems with icosahedral symmetry: application to Dialanine(60) in full and reduced basis set implementations. *J. Chem. Phys.* **115**, 691–698.
43. Tama, F. & Brooks, C. L., III (2002). The mechanism and pathway of pH induced swelling of cowpee chlorotic mottle virus. *J. Mol. Biol.* **318**, 733–737.
44. Durand, P., Trinquier, G. & Sanejouand, Y. H. (1994). A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, **34**, 759–771.
45. Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H. (2000). Building blocks approach for determining low-frequency normal modes in macromolecules. *Proteins: Struct. Funct. Genet.* **41**, 1–7.
46. Perahia, D. & Mouawad, L. (1995). Computation of low-frequency normal modes in macromolecules: improvements of the method of diagonalization in a mixed basis and application to hemoglobin. *Comput. Chem.* **19**, 241–246.
47. Tama, F. & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **14**, 1–6.
48. Franklin, M. C., Wang, J. & Steitz, T. A. (2001). Structure of the replicating complex of a Pol alpha family DNA polymerase. *Cell*, **105**, 657–667.
49. Shamoo, Y. & Steitz, T. A. (1999). Building a replisome from interacting pieces: sliding clamp complexed to a peptide from DNA polymerase and a polymerase editing complex. *Cell*, **99**, 155–166.
50. Cheetham, G. M. T., Jeruzalmi, D. & Steitz, T. A. (1999). Structural basis for initiation of transcription from an RNA polymerase–promoter complex. *Nature*, **399**, 80–83.
51. Zhang, G., Campbell, E. A., Minakhin, L., Richter, C., Severinov, K. & Darst, S. A. (1999). Crystal structure of *T. aquaticus* RNA polymerase at 3.3 Å resolution. *Cell*, **98**, 811–824.
52. Darst, S. A., Opalka, N., Chaon, P., Polyakov, A., Richter, C., Zhang, G. & Wriggers, W. (2002). Conformational flexibility of bacterial RNA polymerase. *Proc. Natl Acad. Sci. USA*, **99**, 4296–4301.
53. Ling, H., Boudsocq, F., Woodgate, R. & Yang, W. (2001). Crystal structure of a Y-family DNA polymerase in action: a mechanism for error-prone and lesion-bypass replication. *Cell*, **107**, 91–102.
54. Ago, H., Adashi, T., Yoshida, A., Yakamoto, M., Habuka, N., Yatsunami, K. & Miyano, M. (1999). Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Structure*, **7**, 1417–1428.
55. Bressanelli, S., Tomei, L., Roussel, A., Incitti, I., Vitale, R. L., Mathieu, M. *et al.* (1999). Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc. Natl Acad. Sci. USA*, **96**, 13034–13039.
56. Lesburg, C. A., Cable, M. B., Ferrari, E., Hong, Z., Mannarino, A. F. & Weber, P. C. (1999). Crystal structure of the RNA-dependent RNA polymerase from hepatitis C reveals a fully encircled active site. *Nature Struct. Biol.* **6**, 937–943.
57. Bressanelli, S., Tomei, L., Rey, F. A. & de Francesco, R. (2002). A structural analysis of hepatitis C RNA-dependent RNA polymerase in complex with ribonucleotides. *J. Virol.* **76**, 3482–3492.
58. Jacrot, B., Cusack, S., Dianoux, A. J. & Engelman, D. M. (1982). Inelastic neutron scattering analysis of hexokinase dynamics and its modification on binding of glucose. *Nature*, **300**, 84–86.
59. Cusack, S. & Doster, W. (1990). Temperature dependence of the low frequency dynamics of myoglobin. *Biophys. J.* **58**, 243–251.
60. Richards, F. M. (1977). Areas, volumes, packing and protein structures. *Ann. Rev. Biophys. Bioeng.* **6**, 151–175.
61. Kitao, A. & Go, N. (1999). Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **9**, 164–169.

62. Butcher, S. J., Grimes, J. M., Makeyev, E. V., Bamford, D. H. & Stuart, D. I. (2001). A mechanism for initiating RNA-dependent RNA polymerization. *Nature*, **410**, 235–238.

63. Kraulis, P. (1991). Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Edited by R. Huber

(Received 26 February 2002; received in revised form 22 May 2002; accepted 24 May 2002)