doi:10.1016/j.jmb.2007.02.049 J. Mol. Biol. (2007) 368, 1231–1248

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Multivariate Analysis of Conserved Sequence–Structure Relationships in Kinesins: Coupling of the Active Site and a Tubulin-binding Sub-domain

Barry J. Grant[1]*, J. Andrew McCammon[1,2], Leo S. D. Caves[3] and Robert A. Cross[4]

[1]*Department of Chemistry and Biochemistry, University of California, San Diego La Jolla, CA 92093, USA*

[2]*Howard Hughes Medical Institute, University of California, San Diego La Jolla, CA 92093, USA*

[3]*Department of Biology University of York, York YO10 5YW, UK*

[4]*Molecular Motors Group Marie Curie Research Institute Oxted, RH8 0TL, UK*

**Corresponding author*

An extensive computational analysis of available sequence and crystal structure data was used to identify functionally important residue interactions within the motor domain of the kinesin molecular motor. Principal component analysis revealed that all current kinesin crystal structures reside in one of two main conformations, which differ at the active site, and in the position of a microtubule-binding sub-domain relative to a rigid central core. This sub-domain consists of secondary structure elements α4-loop12-α5-loop13 and contains a conserved hydrophilic surface patch that may be involved in strong binding to microtubules. A hinge point for the sub-domain motion lies near a conserved glycine at position 292. Statistical coupling analysis revealed a network of co-evolving positions that link this region to the nucleotide-binding site, *via* a highly conserved histidine in the switch I loop. The data are consistent with a model in which the nucleotide status of the active site shifts kinesin between weak and strong binding conformations *via* reconfiguration of the identified sub-domain. Our data provide a statistically supported framework for further examination of this and other structure–function relationships in the kinesin family.

Keywords: kinesin; molecular motors; sequence analysis; structure analysis; structure–function relationships

Introduction

Comparing multiple structures of homologous proteins and carefully analysing large multiple sequence alignments can help identify patterns of sequence and structural conservation and highlight conserved interactions that are crucial for protein stability and function. Conserved amino acids are most often located in structurally important cores, or reside at functionally important active sites and protein–protein interaction sites.[1–5] Additionally, subtle patterns of conservation have the potential to highlight positions that co-evolve to maintain structural and dynamic features essential for allosteric modulation.[6] This study aims to analyse the natural variation in available sequences and structures of the molecular motor protein kinesin and thereby infer which residues are likely to be functionally and structurally important.

Kinesins are molecular motors responsible for the ATP dependent transport of cellular cargo along microtubules. Kinesin family members have been found in all eukaryotic organisms, where they contribute to the transport of molecules and organelles, organisation and maintenance of the cytoskeleton, and the segregation of genetic material during mitosis and meiosis. The defining attribute of kinesin family members is the possession of one or more globular motor domains. These ~350 residue domains display high sequence conservation, and are responsible for ATP hydrolysis, microtubule binding and force production. Outside the motor domain, both the domain structure and sequence of kinesin family members can be quite diverse, reflecting a variety of functional roles including the binding of molecular cargo.

Abbreviations used: cryoEM, cryo-electron microscopy; PCA, principal component analysis; HMM, hidden Markov model; SCA, statistical coupling analysis; RMSF, root mean square fluctuation; PDB, protein data bank.

E-mail address of the corresponding author: bgrant@mccammon.ucsd.edu

The structure of the motor domain has been solved by X-ray diffraction and offers the exciting possibility of dissecting the motor mechanism at atomic resolution. There are currently 37 published atomic structures of recombinant kinesin motor domains[7–30] (see Supplementary Data for details). In every case the motor region is an "arrow-head" shaped *aba* sandwich domain, composed of an eight-stranded β-sheet that is flanked on either side by three major α-helices. The head measures approximately 70 Å by 45 Å by 45 Å. The nucleotide binding site is similar to that found in other P-loop containing NTPases such as myosins and G-proteins. Alanine scanning mutagenesis and limited proteolysis indicate that a putative microtubule-interacting region lies on the opposite side of the motor domain from the nucleotide binding site.[31,32] Recent high resolution 3D cryo-electron microscopy (cryoEM) imaging of the kinesin–microtubule complex has allowed well-constrained fitting of crystal structures into EM mass density maps.[33,34] These fits predict the microtubule contact surface in some detail, but have the limitation that the crystallographically determined structures may be in a different conformation to that visualised in the EM experiments.[33]

Kinesin is thought to undergo nucleotide-induced conformational changes that resemble those known to occur in the structurally related myosin and G-protein families.[35–37] The general principle by which conformational changes that occur during nucleotide hydrolysis are translated and amplified into larger movements is known as mechanochemical coupling.[38] Nucleotide turnover in the active site drives changes in the global conformation of the head that modulate kinesin's affinity for microtubules, generate force, and coordinate processive motility.[10,37,39–41] Deciphering the nature and extent of allosteric pathways and probing possible mechanisms for conformational changes in kinesin is thus essential for understanding how ATP binding, hydrolysis, ADP release and phosphate release are linked to microtubule interaction and force production.

Approach

The kinesin motor domain is well suited to a systematic analysis of sequence and structural conservation: hundreds of diverse kinesin motor domain sequences are available, providing a broad sampling of sequences that are consistent with the motor domain fold. Crucially, 37 structures, from at least five different subfamilies[42–44] are also available for analysis (see Supplementary Data for details). Here, we undertake a detailed analysis of all available kinesin sequence and structural information to extract evidence for possible sub-domain movements, and highlight conserved positions that may connect the active site with remote sub-domains. In particular, we focus on a specific sub-domain that lies at the centre of the microtubule-binding interface. Conformational changes in this region may modulate the microtubule binding affinity of kinesin.

Initial work entailed the alignment and quantitative assessment of sequence and structure conservation at each residue position within the motor domain, so as to catalogue residues and structural features that are strongly conserved and therefore important for the structure and function of the kinesin head. New sequence motifs are identified and a common core, whose structure is invariant between all extant kinesin crystal structures, is defined. The kinesin crystal structures are further analysed with principal component analysis (PCA). This analysis highlights intrinsic sub-domains that change relative conformation during kinesin's ATPase cycle. Results are presented as a conformer plot that succinctly displays the relationship between structures in terms of major conformational differences. This plot affords the possibility to standardise discussions of kinesin's various conformational states and provides a means to interpret new crystallographic structures. Finally, sequence and structural approaches are combined and synthesised to reveal relationships between conserved patterns of sequence and conserved sub-domain motions. Our results identify two major conformational classes, and highlight positions that link the configuration of the active site to that of remote sub-domains.

Results and Discussion

Sequence alignment

To ensure the accurate alignment of a diverse range of sequences consistent with the motor domain fold, an iterative supervised profile training and sequence collection procedure was employed. A hidden Markov model (HMM)[45] was first constructed from a structure-based sequence alignment of the available kinesin motor domain crystal structures (see Supplementary Data). This initial model was utilized to guide the alignment of 143 sequences from the Kim & Endow dataset[46] that contains representatives from all previously identified kinesin sub-families. The resultant alignment formed the basis of a new HMM that was used to identify and align kinesin motor domain sequences in the SWISSPROT and TrEMBL databases.[47] The final alignment contained 496 non-redundant sequences (no two sequences are more than 90% identical) whose length ranged from 287 residues to 524 residues, with an average length of 351 residues. No gaps were introduced within regions of secondary structure and the variability in length of motor domain sequences was found to result from insertions and deletions in surface exposed loop regions (predominantly loop 6, loop 10, loop 2, loop 5 and loop 12).

To uncover functionally important residues, it is essential that the observed amino acid distribution at each position in the alignment is representative of the allowed variation at those positions. Therefore the alignment should be sufficiently large and di-

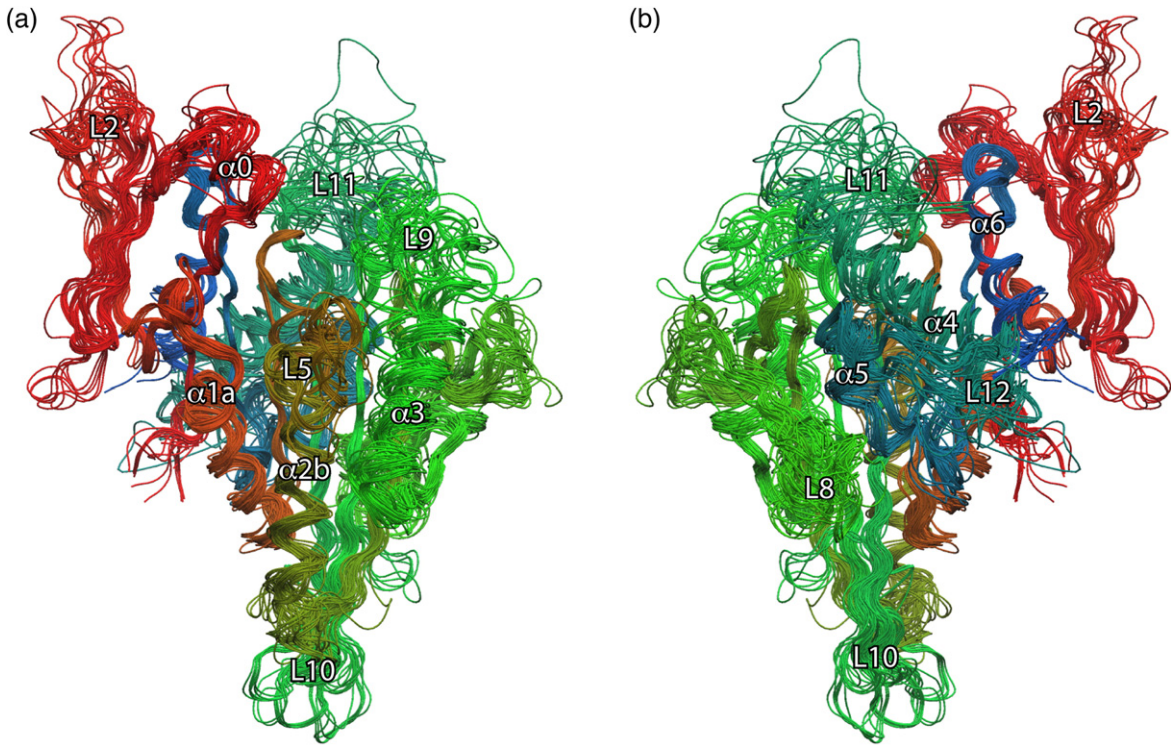

Figure 1. Structural superposition of all kinesin motor domain crystallographic structures. (a) Front view and (b) back view; see text for details. The strand order, as viewed from left to right in (a) is β5, β4, β6, β7, β3, β8 and β1. Structures are coloured from red, N-terminal, to blue, C-terminal. Superposition was performed with the Bio3d package[84] and rendered with VMD.[51]

verse so as to reflect the constraints on the entire protein family during evolution.[48,49] Pairwise identity analysis of the aligned sequences revealed a close to normal distribution of sequence identities (with a mean value of 37% and range of 21% to 89%) indicating a diverse alignment.[3] Furthermore, a number of unconserved positions within the alignment were found to have an amino acid distribution close to the mean in all natural proteins (i.e. the mean amino acid frequency distribution in the SWISS-PROT database) (data not shown), indicating that the sequences in the alignment have experienced substantial evolution. In addition, the random elimination of sequences from the alignment did not drastically alter the amino acid distribution at all positions (Supplementary Data). On this basis it was concluded that the alignment was representative of the sequence variations permitted within the structural and functional constraints of the motor domain.

Structural alignment

To complement the sequence alignment analysis, a thorough structural comparison of the kinesin motor domain was undertaken. Fifty-six motor domain structures, extracted from the 37 available monomeric and dimeric kinesin structures deposited in the protein data bank (PDB),[50] were structurally aligned and used to infer general structural and functional properties. Overall, the structures were observed to be extremely similar (Figure 1), with a global average root-mean-square deviation of 1.59(±0.61) Å measured over the alpha carbon positions of 218 equivalent residues present in all structures.

It should be noted that the 37 available atomic structures are representative of only five of the 14 currently recognized kinesin sub-families.[44] This redundancy is reflected in the pairwise sequence identity of the structural dataset, which has a more skewed distribution (mean value of 51% and range of 29% to 100%) than that of the sequence alignment dataset. An additional limitation is that the available structures may represent only a small sampling of the conformational space available to the motor domain. As such, all inferences are subject to these limitations. We believe that the size and diversity of the structural data set is sufficient to support analyses probing sequence, structure and functional relationships in the kinesin motor domain.

Conservation analysis

To assess the level of conservation at each position in the sequence alignment the similarity, identity, class identity and entropy per position were calculated (detailed in Materials and Methods). Each of these conservation scores has particular strengths and weaknesses.[52] For example, entropy elegantly captures amino acid diversity but fails to account for stereochemical similarities. By employing a combination of scores and taking the union of their respective conservation signals we expect to

achieve a comprehensive analysis of sequence conservation. A position was defined as conserved if the similarity, identity, or entropy scores for that position exceed 0.6. Positions in which more than 30% of the sequences had gaps were excluded from all sequence conservation analysis. The conservation data from the sequence alignment was assessed in light of the analysis of structural and physical properties of the 56 kinesin motor domain structures.

The structural role of a residue is determined by the nature and extent of the interactions it makes with other residues.[2] In this work the extent of residue–residue interaction was assessed by calculating the number of contacts that every residue makes with all other residues in each of the available structures. Two residues were assumed to be in contact if any two heavy atoms of these residues are closer than 5.0 Å. The extent to which a residue is exposed to solvent can potentially provide further insight into its structural or functional role. For example, residues that are buried in the interior of a protein are unlikely to directly contact protein partners or form substrate binding interfaces. The

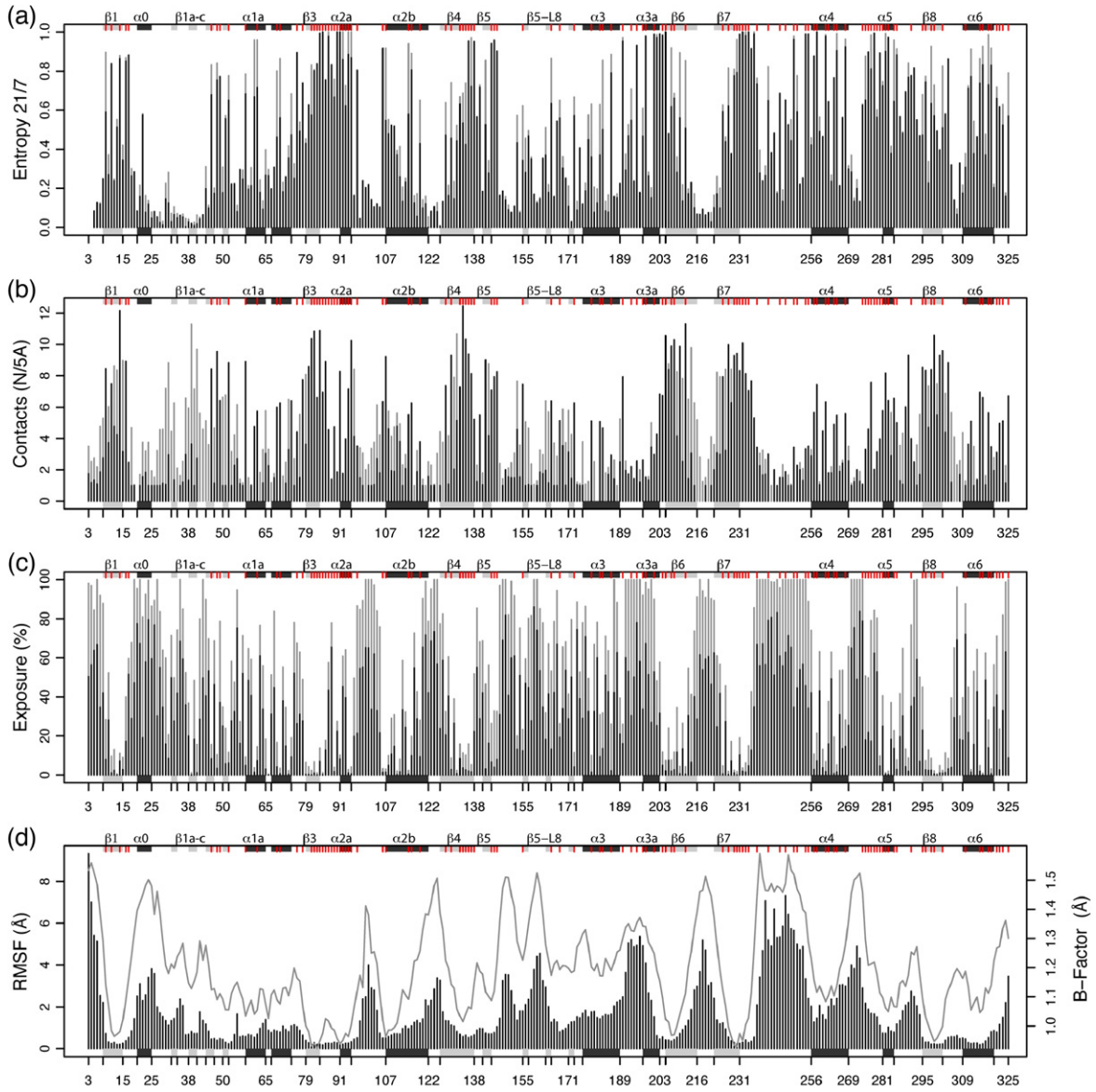

Figure 2. Sequence and structural conservation in the kinesin motor domain. (a) Sequence entropy scores for both a 21-letter alphabet (20 amino acids and a gap) and seven-letter alphabet (where amino acids are grouped into six classes based on their physicochemical properties) are plotted per position in dark gray and light gray, respectively. (b) The average number of contacts per position in all motor domain structures. Total contacts are shown in light grey whilst contacts to positions that are conserved in sequence are plotted in dark grey. (c) The mean and maximum solvent exposure per position in all structures (in dark and light gray, respectively). (d) The RMSF (bars) and mean *B*-factor (line) per position over all structures. The major elements of secondary structure (shaded rectangles) and positions with a high degree of sequence conservation (red ticks) are indicated in the marginal areas of each plot to facilitate comparison. Both the secondary structure and residue numbering are according to kinesin-1 from *Homo sapiens* (also known as conventional kinesin, HsK or Kif5B, PDB code 1bg2).

degree of solvent exposure was evaluated by comparing each residue's solvent accessible surface area to that of the same residue when found in an isolated G-X-G tripeptide context (where X is the residue in question).

Functionally important internal motions can be highlighted through protein structure comparison. To assess the average structural variability per position the root-mean-square fluctuation (RMSF) for backbone atoms of equivalent positions in the available structures was assessed. Finally, PCA was used to examine the relationships between the available kinesin crystal structures in terms of large-scale concerted atomic displacements. The principal components were obtained from diagonalisation of the covariance matrix built from the Cartesian coordinates of the superposed motor domain structures. The principal components are orthogonal eigenvectors that describe concerted atomic displacements and thus serve as a basis in which to highlight the major conformational differences between the kinesin structures.

Sequence and structural variability

Conservation data from the analysis of sequence and structural alignments is summarised in Figure 2. A total of 147 positions (45.5% of all positions) were identified as conserved in sequence based on a combination of similarity, identity, class identity and entropy scores. Of these, 57 positions (17.8%) were found to have a conserved hydrophobic character, whilst 39 positions (12.1%) have a conserved hydrophilic character and 50 positions (15.6%) have a conserved neutral character. In relation to the primary structure, the majority (80%) of conserved positions were found to cluster into eight contiguous motifs. Outside of these motifs (detailed in Table 1), conservation was confined to positions 48, 49, 106 and 116 along with the more subtle class conservation of positions in loop 2, α2a and loop 8. As with other protein families, a greater degree of sequence variability was evident within loop regions (61% of conserved positions are located within secondary structure elements).

The most striking structural differences were observed in regions that do not appear in all compared structures. These regions include loop 11 (between β7 and α4) and the regions following α6 (i.e. the neck region). In kinesin-14 structures (PDB codes 1cz7, 1n6m and 2ncd) the neck helix enters β1 from the N-terminal side through a short (three residues long) neck-linker region, whereas in kinesin-1 and kinesin-5 (2kin, 3kin and 1mkj) the neck helix enters α6 from the C-terminal side through a longer (approximately 15 residue) neck-linker, composed of two short secondary structure elements (β9 and β10). This neck-linker region is disordered and therefore unresolved in the majority of structures. These regions, as with other structurally variable loops have a high solvent exposure and low number of contacts.

Positions comprising the central beta sheet (β1, β3, β4, β6, β7 and β8) show very little structural variation. These positions have a low RMSF value, low solvent exposure and a relatively high number

Table 1. Conserved sequence motifs in the kinesin motor domain

Motif	Name	Positions	Location	Details
1	X	9-22	β1-α0	[IV]^{89}- [Rkqc]^{76}-**V**^{92}-X- [Vcil]^{91}-**R**^{93}- [Vcikfl]^{93}-**R**^{93}-**P**^{94}- [Lfmpq]^{77} - [Nlst]^{77}-X-X- [Eq]^{78}
2	A	74-97	L3-β3-L4-α2a	[LFivm]^{98}- [EDqnksa]^{88}-**G**^{94}- [Yfgi]^{87}- [Nk]^{90}- [GAvcs]^{95}- [TCs]^{95} - [IVlc]^{98}- [Fl]^{95}- [At]^{95}-**Y**^{100}-**G**^{100}- [Qv]^{89}-**T**^{99}- [Gs]^{98}- [Sta]^{98} -**G**^{100}-**K**^{100}- [Ts]^{100}- [YFh]^{100}- [Ts]^{100}- [Mi]^{98}-X- [Gt]^{94}
3	B	136-146	β4-L7-β5	**E**^{98}- [Ilv]^{99}-**Y**^{98}- [Nqm]^{83}- [Egd]^{94}-X- [IVl]^{95}- [Ryfln]^{78}-**D**^{97}-**L**^{98}-**L**^{95}
4	C	186-212	α3-L9-β6	[Ga]^{95}- [Nsea]^{65}-X- [Nsakqrh]^{83}-**R**^{98}- [Tsahrkq]^{95}- [VTis]^{90}- [AGs]^{91} - [Aseqp]^{78}-**T**^{96}- [Nakslq]^{78}- [Mlav]^{88}-**N**^{99}- [Eadksq]^{85}-X-**S**^{100} -**S**^{99}-**R**^{100} -**S**^{99}-**H**^{100}- [AStc]^{90}- [IV]^{90}- [Fl]^{90}- [Tqsirl]^{85}- [Ilv]^{94}- [Tivhkgnr]^{82} - [ILVf]^{97}
5	D	225-237	β7-L11	[GSa]^{97}- [Kqrt]^{83}- [Lifm]^{97}- [NShy]^{87}- [Lf]^{95}- [Vi]^{98}-**D**^{100}-**L**^{100}-**A**^{99}-**G**^{100} -**S**^{100}-**E**^{100}- [Rkn]^{95}
6	E	241-270	L11-α4	[TS]^{91}- [Gkqe]^{77}- [Asnv]^{88}-X- [Gr]^{88}-X- [Rtq]^{91}- [Larf]^{92}- [Kre]^{78}-**E**^{98} - [GAt]^{95}-X- [Nshkeay]^{92}-**I**^{100}-**N**^{100}- [Krlqs]^{92}- [**S**g]^{99}-**L**^{100}- [LStm]^{93} - [Atc]^{85}-**L**^{99}- [Gk]^{84}- [Ndkertq]^{88}- [Vc]^{97}-**I**^{95}- [Snra]^{70}- [Aks]^{88}-**L**^{96} - [Asgvrt]^{92}- [Desqkrn]^{92}
7	F+Y	274-304	L12-α5-L13-β8	[Hfy]^{91}- [VIt]^{98}-**P**^{94}- [Yf]^{99}-**R**^{100}- [DNe]^{91}-**S**^{95}- [Kv]^{92}-**L**^{94}-**T**^{99}- [Rwqyh]^{86} - [Liv]^{98}-**L**^{93}- [Qkr]^{94}- [Denp]^{87}- [Sn]^{85}- [Li]^{92}- [Gs]^{93}- [Ge]^{95}- [Nrsd]^{85} - [Sact]^{91}- [KRq]^{85}- [Tv]^{91}-X- [Mil]^{94}- [Ivfl]^{97}- [Avci]^{92}- [TNca]^{87} - [IVlc]^{95}- [Sngt]^{91}-**P**^{93}
8	Z	311-325	α6-L15	E^{82}- [Ts]^{97}- [Lik]^{72}- [SN]^{80}- [Ts]^{94}-**L**^{91}- [Rk]^{69}- [YF]^{99}- [Ag]^{97}-X- [Rk]^{87} - [AV]^{89}- [Krn]^{85}-X- [Ivlc]^{94}

Motifs are numbered 1 to 8 in order of their position in the primary structure. Conservation in seven of these regions has been previously noted[53] (see http://www.proweb.org/kinesin/, where they are named X, A, B, C, D, E, F and Y). The details of each motif characterize the residue occurrences for each position. Residues are listed in descending order of occurrence (i.e. the consensus residue for each position is given first). Capital letters represent residues that occur in at least 25% of sequences, whilst bold letters represent residues that occur in more than 90% of sequences. Unconserved positions, where no single residue occurs in 25% of sequences, are assigned an X character. In superscript is the percentage of sequences that are successfully matched by the position's regular expression. Underlined positions are within 10 Å of the bound nucleotide in all kinesin crystal structures. For example, in motif 1 the pattern [Rkqc]^{76}-**V**^{92}-X summarises the residue occurrence at three connected positions. An arginine residue occupies the first position in 90% to 25% of sequences, whilst lysine, glutamine and cysteine occur in less than 25% of sequences. The superscript 76 indicates that 76% of sequences contain one of the afore-mentioned residues at this position. Valine predominates at the second position occurring in 92% of sequences. The third position displays little or no conservation. Note that positions are separated with a dash and alternatives are grouped in square brackets.

of contacts. The most rigid zone identified outside the beta sheet was the eight residue long loop 4 region (residues 85 to 92 with an average root mean square deviation of 0.95(±0.84) Å). This structurally invariant loop (known as the P-loop) is found in many nucleotide binding proteins and is characterised by the conserved sequence motif G-x(4)-G-K-[ST] (with two glycine residues separated by four un-conserved residues, followed by a lysine and either a serine or threonine). In the kinesin family this motif can be extended into the flanking β3 and α2a regions (motif 2 in Table 1). Of the other conserved sequence elements, motifs 1, 3 and 5 show relatively low variability whilst motifs 4, 6, 7 and 8 show more. The current alignment reveals different degrees of conservation within and between these regions. However, its principal role is to facilitate the more extensive synthesis of sequence and structural conservation described below.

Principal component analysis

PCA was used to examine the major conformational differences between available kinesin structures. Over 91% of the total mean-square displacement (or variance) of atom positional fluctuations was captured in eight dimensions, over 62% in two dimensions and over 74% in three dimensions. The first few principal components retain most of the variance in the original distribution and thus provide a useful description of the conformational space of the system (see Figure 3 for details). Projecting the original structures into the sub-space defined by the principal components with largest associated variance resulted in a low dimensional graphical representation that succinctly displays the relationship between structures (Figure 3).

Figure 3(a) displays the relationship between structures in terms of the conformational differences described by the first two principal components (PC1 and PC2). The contribution of each residue to PC1 is displayed in Figure 4. The height of each bar represents the relative displacement of each residue described by a given PC and are depicted as atomic displacements from the mean structure in Figure 4(c) and (d). The dominant feature described by PC1 is the concentrated displacement of a sub-domain comprising secondary structure elements α4-loop12-α5-loop13 (residues 254 to 291). The entire region undergoes a significant displacement relative to the central β-sheet. Interestingly, helices α4 and α5 make conserved (in terms of the static structures) and persistent (throughout the extent of sub-domain movement) contacts to the back of the active site through a hyper conserved histidine in the β6-loop9 region (Figure 5). The conformation of H205 is directly correlated with sub-domain orientation. Coincident displacements are evident in surrounding positions at the top of β6 and neighbouring β7 and β4 strands (Figure 5). Also of note are the concerted displacements of residues in the α3 helix, which alters its angle relative to the neighbouring α2b helix. The remaining displacements captured by PC1 are confined to loop 8 and

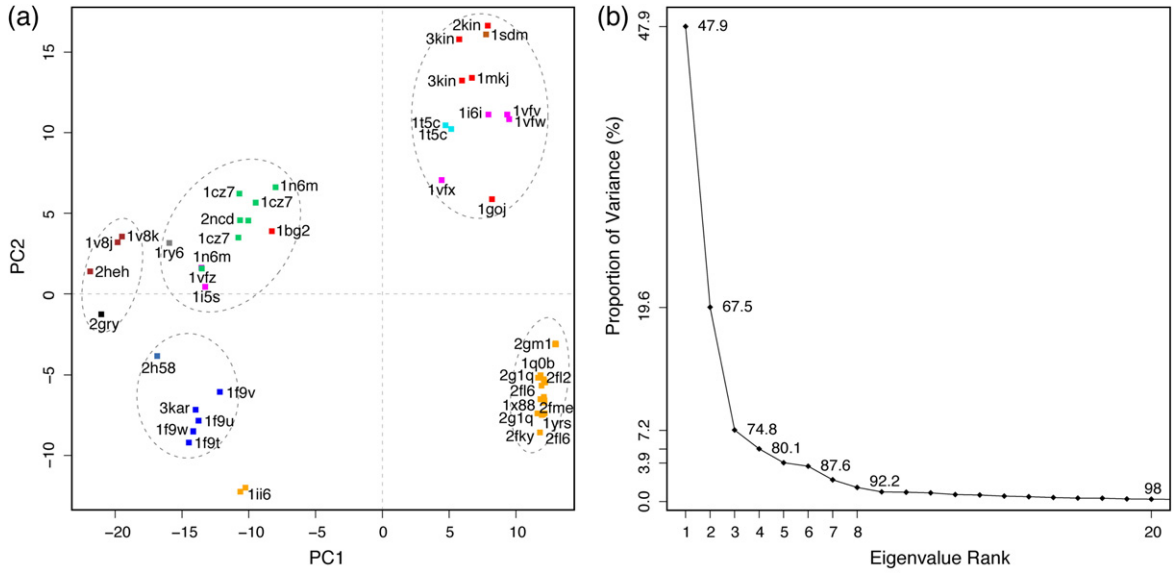

Figure 3. Results of PCA on the kinesin motor domain. (a) Conformer plot: projection of all kinesin X-ray structures onto the principal planes defined by the two most significant principal components (termed PC1 and PC2). Structures are coloured by sequence group (structures that have a sequence identity value of more than 60% are assigned the same colour) and labelled with their PDB code where space permits. Dashed ovals represent the grouping obtained from hierarchical clustering of the projected structures in the PC1 to PC3 planes. (b) Eigenvalue spectrum: results obtained from diagonalisation of the covariance matrix of C^{α} atom coordinates from the kinesin crystal structures. The magnitude of each eigenvalue is expressed as the percentage of the total variance (mean-square fluctuation) captured by the corresponding eigenvector. Labels beside each point indicate the cumulative sum of the proportion of the total variance accounted for in all preceding eigenvectors.

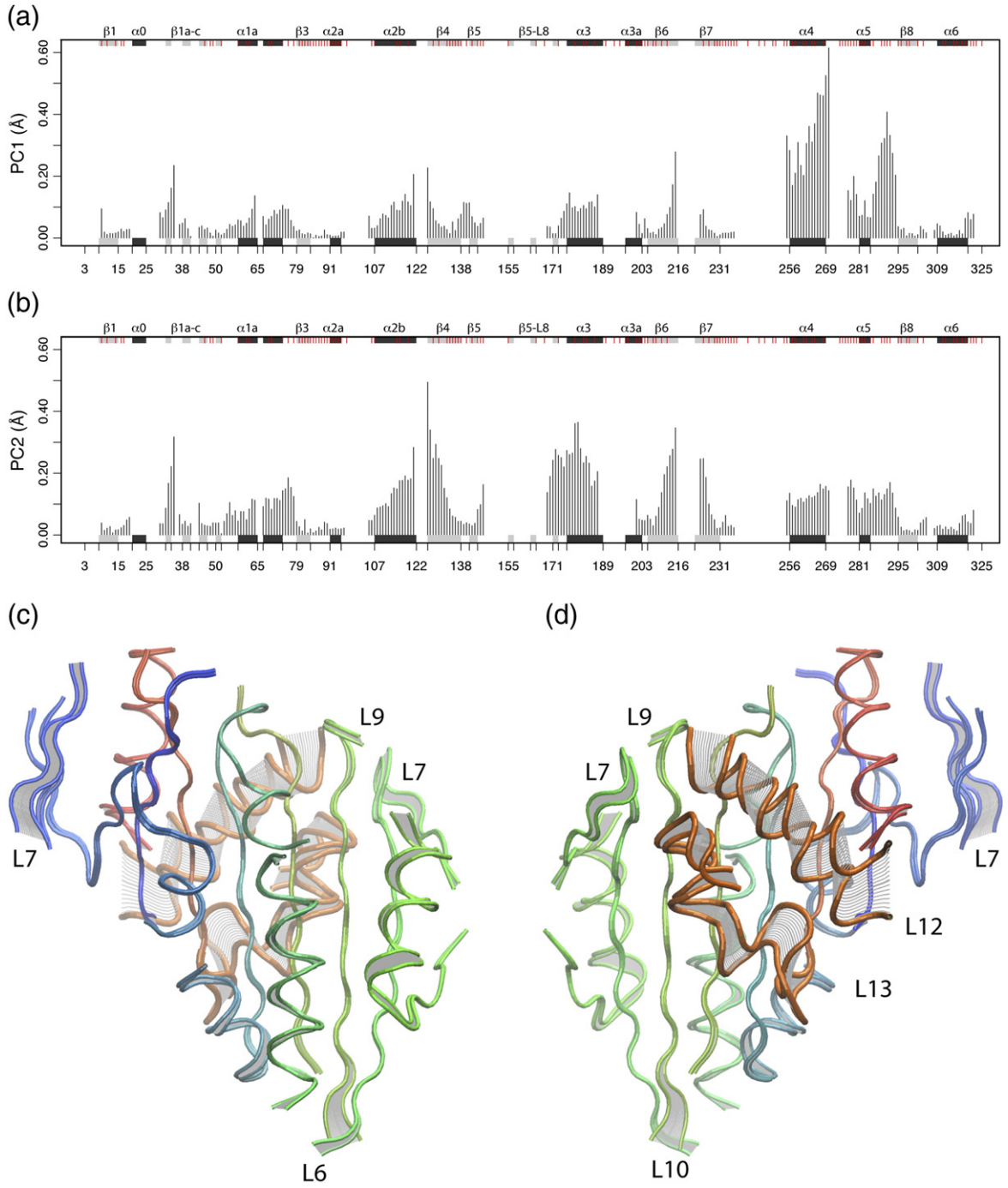

Figure 4. Results of PCA on the kinesin motor domain. (a) and (b) The contribution of each residue to the first two principal components. Secondary structure and sequence conservation is displayed as in Figure 2. (c) and (d) Front and back views of the kinesin motor domain, with the first principal component represented as equidistant atomic displacements from the mean structure. Displacements are scaled by the standard deviation of the distribution along the first principal component. The figure was generated using VMD.[51]

the flexible tip of the motor domain, namely loop 6 and loop 10. PC2 and PC3 characterise displacements of the α0-loop1-β1-loop2 region, which together with α6 form a lobe or sub-domain that can undergo significant conformational changes. The flexibility of the motor domain tip is again portrayed in PC2 and PC3 with large displacements evident for residues in the vicinity of loop 10 and loop 6 regions.

Interconformer relationships

The current kinesin structures can be divided into two major groups in the PC1 plane. Members of each cluster differ in the relative orientation of their α4-loop12-α5-loop13 sub-domain. The average angle between the principal axis of helix α4 and strand β8 is 50.4(±4.8)° for one group of structures (including 1cz7, 1bg2 1i5s and 3kar) and 66.5(±1.6)°

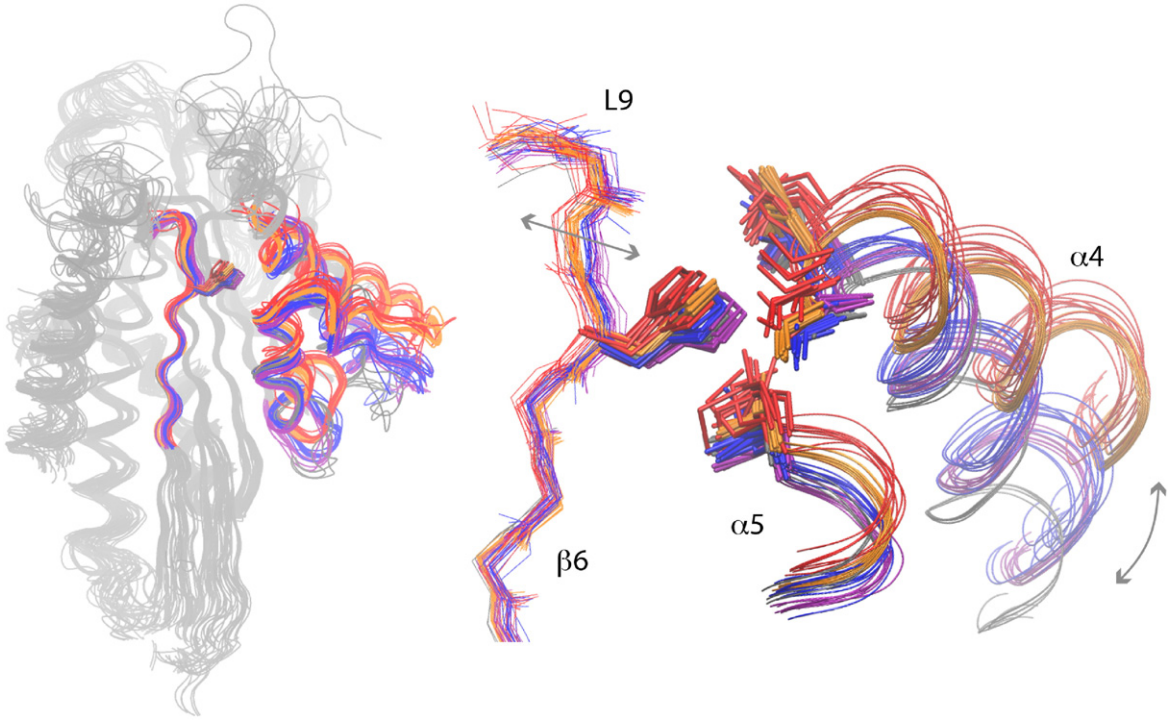

Figure 5. The conserved interaction of switch I histidine 205 in the β6-loop9 region, with serine 257 and leucine 258 in helix α4, and leucine 282 in helix α5. All available kinesin structures are displayed and coloured according to their membership in the five conformational sub-clusters obtained from PCA (see Figure 3). The figure was generated using VMD.[51]

for the other group of structures (including 2kin, 1mkj 1i6i, 1goj and 2gm1). Furthermore, comparing C^{α} atom torsion values (the torsion angle defined by every set of four consecutive C^{α} atoms[54]) of structures residing in either cluster indicates that residues G291 and G292 are the hinge/pivot point for this sub-domain motion relative to the rigid central core. Mutation of these residues would likely affect sub-domain dynamics. In kinesin-1, these glycine residues form the upper part of the docking site for the kinesin neck linker. Mutation of G291 is known to affect microtubule activated ATPase activity, microtubule sliding velocity in motility assays, and microtubule affinity.[55]

PC2 serves to discriminate the structures into additional sub-groups (Figure 3). This further subdivision captures distinct states in the vicinity of loop 6 and loop 10, towards the bottom tip of the motor domain, as well as different orientations of helix α3.

It is important to note that the conformational clustering we observe does not coincide with the nature of the bound nucleotide in the various structures. This is consistent with the current consensus in the myosin field, where the observed crystallographic conformations do not correlate with the nucleotide occupying the active site. Crystallography has identified at least three different myosin conformations, defined by distinctive orientations of certain sub-domains, such as the lever-arm and converter domains. Hence, the particular conformation observed in any one crystal depends only weakly, if at all, on the active site nucleotide.[56] The kinesin clusters we identify do not represent particular chemical species in the active site but rather distinct global conformations (characterized by the relative orientation of the α4-loop12-α5-loop13 sub-domain) that are correlated with small changes in the active site (most notably, H205).

The current results suggest that regions of the motor domain can be usefully described as semi-rigid sub-domains, in the sense that main-chain displacements internal to a particular sub-domain region are small in comparison to displacements of the sub-domain relative to the rest of the motor domain. Furthermore, the separation of all structures into two main clusters indicates that there are two predominant motor domain conformations differing in terms of a large-scale displacement of a dynamic sub-domain relative to a rigid core, which is likely to be of functional significance for microtubule-bind release cycles.

Roles of conserved residues

The possible roles of conserved positions in mediating the structure and function of the motor domain are discussed in this section.

Nucleotide binding

The available kinesin structures position the nucleotide (ADP or AMP.PCP) and its magnesium co-factor in an approximately 125 $Å^3$ cavity that is lined with highly conserved residues from motifs 1, 2, 4 and 5. These regions are commonly referred to, in the kinesin literature, as the N4 region (positions 14–

17 from motif 1), the N1 or P-loop region (positions 85–92 from motif 2), the N2 or switch I region (positions 201–205 from motif 4) and the N3 or switch II region (positions 231–236 from motif 5). Conserved residues in these regions are involved in binding adenine nucleotides, in the hydrolysis of ATP, in controlling the release of the hydrolysis products ADP and Pi, and in controlling conformational switching of the microtubule-binding interface.

Inspection of the structures reveals that positions 14 (R^{93}, see Table 1 for nomenclature), 16 (R^{93}), 17 (P^{94}) and 93 ($[YFh]^{100}$) have hydrophobic interactions with the nucleotide base and may be responsible for nucleotide specificity[57] (see Table 1). Position 93 together with positions 203 (R^{100}), 236 (E^{100}), 92 ($[Ts]^{100}$) and 231 (D^{100}) link the various regions of the nucleotide-binding site. Positions from motif 2 (the P-loop or N1 region; see Table 1) cradle the charged phosphate groups of ADP/ATP and appear to provide a major contribution to nucleotide binding.[57] The backbone nitrogen atoms of residues 86–89 point towards the negatively charged β-phosphates and, together with the side-chain of an invariant lysine in position 91, create a positively polarised environment. The neighbouring position 92 ($[Ts]^{100}$), along with positions 231 (D^{100}) and 202 (S^{99}) are involved in coordinating the important Mg cofactor. Positions from motif 4 and 5 correspond to the switch I and switch II regions in G proteins and myosin. In these proteins, analogous positions are sensitive to the presence or absence of γ-phosphate. Following nucleoside triphosphate hydrolysis in the small G-proteins, there is a structural rearrangement associated with the loss of interactions to residues that are equivalent to positions 202 (S^{99}) and 234 (G^{100}) in kinesin. The binding of these positions to the γ-phosphate in the small G-protein Ras has been likened to the loading of a spring that is released after triphosphate hydrolysis.[58,59] In kinesin, formation of a salt bridge between the glutamate of DLAGSE (switch II) and the arginine of SSRSH (switch I) is required for the hydrolysis step of ATP turnover.[60]

Structural invariance and the common core

Comparing the coordinates of the backbone atoms in different motor domain structures and implementing the core-finding method of Gerstein & Altman[61] highlights 72 positions with a low structural variance (less than 1 Å). These geometrically conserved positions are characterised by a high number of contacts (more than ten contacts per position) and low solvent exposure (typically less than 40% exposure per position).

The common core includes portions of the central β-sheet, loop 4 (P-loop), α2a and several turns of α2b. The core corresponds to 22% (72/321) of the total motor domain and contains 37% of all the conserved positions (55/147). The majority (82%) of these conserved positions are either hydrophobic or neutral. However, ten positions were found to have a conserved hydrophilic character. Of these positions, five (positions 14, 91, 231, 203 and 136) are restricted to a single residue type in over 90% of kinesin sequences. Four of these, together with the highly conserved position 86, are involved in directly contacting the ligand and are likely to be critical for binding and catalysis. The remaining hydrophilic positions have a greater degree of variability; position 110, for example, is located on the outer surface of α2b and does not directly contribute to core packing. Positions 295 and 78 define the first residue of β8 (position 295 to 302) and β3 (position 78 to 84), and together with position 226 have a moderate solvent exposure, and are thus not expected to be involved in core packing. The precise role of position 136 remains unclear, though its proximity to the oppositely charged position 190 in several structures may be of significance.

Thus it appears that core motor domain positions contribute both to the active site and to the maintenance of the structural integrity of the domain. This juxtaposition of functional and structural sites in kinesin differs from that found in the study of immunoglobulin domains[2] where structural and functional sites are clearly separated (surface exposed loops are responsible for ligand binding whilst invariant core positions provide stability[2]). A key difference may be that no enzymatic activities are carried out by the majority of immunoglobulin domains. In the case of kinesin, locating the nucleotide-binding site in close proximity to the most rigid part of the structure may help to minimise the effect of thermal fluctuations on ligand coordination and ensure the precise spatial positioning of functional groups essential for catalysis.[62]

Conserved residue–residue contacts and the peripheral region

Knowledge of the nature and extent of conserved residue–residue interactions helps to characterise the structural role of positions in the motor domain fold. A systematic analysis of residue-residue contacts (see Materials and Methods) in all motor domain structures was performed. Here, contacts are considered conserved if they are observed in all motor domain crystal structures. The nature of conserved contacts was assessed in light of the sequence and structural conservation data detailed above.

A total of 42 positions were found to have a large number of conserved contacts to positions that are conserved in sequence. The majority of these positions (listed in Table 2) reside in the motor domain core (discussed above), and are largely involved in the packing of the core β-strands. Additional positions in loop 11 and helices α4, α5 and α6 (positions 235, 257, 258, 282 and 315) were also found to possess a high number of conserved contacts. These non-core positions have a low solvent exposure and pack together with core and nucleotide-binding site positions to form a structural cluster that we refer to as the peripheral region. Hydrophobic residues predominate in the peripheral region and the apparent maintenance of a closely packed environment suggests structural constraints that might be expected to limit the range of residue combinations at these positions.

Table 2. Positions with a high number of conserved contacts

Position	Details	Location	Number of contacts			Exposure (%)
			Conserved in all structures	Average per structure	Average to conserved positions	
84	**Y**^{100}	L4	10	10.89	8.62	13.8
205	**H**^{100}	β6	9	10.21	10.2	7.4
232	**L**^{100}	L11	8	10.05	9.32	3.5
230	[Vi]^{98}	L11	8	9.38	8.3	1.8
95	[**M**i]^{98}	L5	8	10.18	8.16	4
91	**K**^{100}	α2a	8	8.27	7.27	10.2
86	[Qv]^{90}	L4	8	8.91	5.93	17.6
301	[TNca]^{87}	β8	7	9.32	9.14	4.7
233	**A**^{99}	L11	7	8.02	7.98	6.1
81	[IVcl]^{98}	β3	7	10.38	6.68	2.4
231	**D**^{100}	L11/β7	7	8.25	6.5	19
130	[Vli]^{83}	β4	7	8.96	5.55	8.7
315*	[Ts]^{94}	α6	7	6.95	4.96	4.7
82	[Fl]^{95}	β3	6	10.86	9.77	6.7
235*	[**S**n]^{100}	L11	6	7.79	7.59	42.7
234	**G**^{100}	L11	6	7.11	7.04	15
208	[Fl]^{90}	β6	6	9.98	6.46	11.7
204	**S**^{99}	β6	6	6.46	6.36	11.7
212	[ILVf]^{97}	β6	6	11.04	6.23	4.1
228	[NShy]^{87}	β7	6	8.34	6.2	17
83	[At]^{95}	L4	6	6.61	6.18	1.2
136	**E**^{98}	β4	6	9.07	6.02	15.9
85	**G**^{100}	L4	6	6.95	5.95	1.2
132	[Vaci]^{85}	β4	6	10.29	4.93	2.4
131	X	β4	6	7.61	1.96	39
134	[YFm]^{90}	β4	5	11.96	8.25	11.3
299	[Ivfl]^{97}	β7	5	10.54	8.2	2.2
206	[AStc]^{90}	β6	5	8.48	8.11	24
282*	**L**^{94}	α5	5	8.14	8.07	1.2
258*	**L**^{100}	α4	5	7.41	7.23	21.4
80	[TCs]^{95}	β3	5	8.59	5.98	4.1
257*	[**S**g]^{99}	α4	5	5.66	5.59	10.7
229	[Lf]^{95}	β7	5	9.14	5.43	2.5
227	[Lifm]^{97}	β7	5	9.82	5.43	1.9
210	[Ilv]^{94}	β6	5	9.62	4.82	4.3
106	**G**^{94}	α2b	5	6.7	4.82	3.3
108	[Itynml]^{94}	α2b	5	5.57	4.46	9.9
209	[Tqsirl]^{85}	β6	5	7.86	3.98	34.5
225	[GSa]^{97}	β7	5	7.73	3.95	4.6
79	[GAvcs]^{97}	β3	5	8.07	3.62	6.4
226	[Kqtr]^{83}	β7	5	8.3	3.18	45.6
211	[Tivhkynr]^{82}	β6	5	7.55	2.46	26.5

Also listed are the average number of contacts per position in all structures, the average number of contacts to positions that are conserved in sequence, and the maximum percent solvent exposure per position in all structures. Positions marked with an asterisk reside outside the motor domain core (see the text for details). Refer to Table 1 for an explanation of the sequence conservation notation used in the details column.

In contrast, the majority of unconserved positions are located outside the core and peripheral regions. In these unconserved positions, mutations can often result in a variation in the size or volume occupied in relation to the native residue. It appears that these changes can be accommodated by small conformational rearrangements that do not affect the packing of core residues or the protein structure as a whole. However, the pattern of mutation in these apparently unconserved positions may be of interest and is addressed below.

Solvent exposure and the conserved surface patch

Analysis of solvent exposure suggests the conservation of 42 surface positions (Table 3). The majority of these positions, which are conserved in sequence, possess a high solvent exposure (more than 40% exposure per position) and are structurally clustered on the rear face of the motor domain (Figure 6). Exceptions to this cluster are several residues in the switch I region (α3a loop 9), three residues at the C terminus (loop 15), and one residue at the boundary of loop 5. There is no apparent structural reason for conservation at any of these positions. Their striking spatial localisation suggests a functional role all but two of these positions (positions 254 and 275) have a conserved hydrophilic or neutral character.

The identified positions form a broad, relatively flat hydrophilic surface patch (Figure 6b). Alanine scanning mutagenesis and limited proteolysis studies have pointed to this region of the motor domain as a likely microtubule-binding site.[31,32] CryoEM studies place the loop11-α4-loop12 region

Table 3. Conserved positions with a high solvent exposure

Position	Location	Details	Class	Exposure (%)	Entropy 21
49	β1c	D[87]	s	78.7	0.77
76	L3	**G**[94]	n	67.6	0.89
78	β3	[Nk][90]	s	48.9	0.74
97	L5	[Gt][94]	n	86.6	0.8
138	β4	**Y**[98]	n	51.8	0.95
140	L7	[Egd][94]	s	68.3	0.72
142	β5	[IVl][95]	b	46.5	0.53
165	β5b	[Vil][94]	b	53.2	0.63
193	L9	[AGs][91]	n	100	0.55
195	L9	**T**[96]	n	100	0.93
197	α3a	[Mlav][88]	b	92.1	0.47
198	α3a	**N**[99]	s	98.6	0.98
203	α3a	**R**[100]	s	57.7	0.99
235	L11	[**S**n][100]	n	42.7	0.92
236	L11	**E**[100]	s	85.6	0.99
237	L11	[Rkn][95]	s	100	0.73
241	L11	[TS][91]	n	100	0.65
245	L11	[Gr][88]	n	96	0.68
247	L11	[Rtq][91]	s	100	0.65
249	L11	[Kre][78]	s	100	0.44
250	L11	**E**[98]	s	100	0.96
251	L11	[Gat][95]	n	100	0.59
254	L11	**I**[100]	b	100	0.99
255	L11	**N**[100]	s	88.8	0.99
263	α4	[Ndrtkeq][88]	s	62.9	0.24
270	α4	[Desqnrk][92]	s	100	0.28
274	L12	[Hfy][91]	n	100	0.63
275	L12	[Vlt][98]	b	50.7	0.65
276	L12	**P**[94]	n	49.8	0.90
277	L12	[Yf][99]	n	79.1	0.81
278	L12	**R**[100]	s	84.3	0.99
279	L12	[DNe][91]	s	80.4	0.56
281	α5	[Kv][92]	s	40.6	0.72
291	L13	[Gs][93]	n	76.8	0.78
292	L13	[Ge][95]	n	99.6	0.81
293	L13	[Nrsd][95]	s	100	0.55
295	β8	[Krq][85]	s	44.9	0.47
311	α6	E[82]	s	44.9	0.71
317	α6	[Rk][69]	s	62.4	0.37
321	L15	[Rk][87]	s	71.1	0.66
323	L15	[Krn][85]	s	82	0.52
325	L15	[Ivlc][94]	b	100	0.57

The class of each position indicates the conserved hydrophobic (b), hydrophilic (s) or neutral (n) nature of conserved residues. Refer to Table 1 for an explanation of the sequence conservation notation used in the details column.

at the centre of the interaction surface between kinesin and the microtubule, flanked by other potential microtubule interaction sites in loop 2 and loop 8.[10,33,34,63–71] These regions are absent from the identified cluster, indicating that their conformations vary widely. Interestingly, residues in loop 2 and loop 8 were found to display more subtle, sub-family-specific conservation, suggesting that they may be related to the distinct properties of different kinesin classes.

Statistical coupling analysis

How are these putative microtubule-binding positions linked to the nucleotide-binding site that is located on the opposite face of the motor domain? Although conserved peripheral contacts provide physical connectivity between these sites, it is also likely that additional regions, such as loop 2 and loop 8, may interact with the microtubule. Examining more subtle conservation signals embedded in the kinesin family has the potential to provide valuable insights into possible positions of communication. We therefore applied statistical coupling analysis (SCA) to detect positions within the motor domain that are likely to mutate in concert. The hypothesis with the SCA approach is that co-evolution may be expected for residues that are allosterically coupled, or that otherwise share a role in structure-function.[6,48,72–74] The ultimate goal of such an analysis is to trace networks of positions that link possible functional sites.

The SCA procedure assesses the change in the amino-acid distribution at one position (i) in a multiple sequence alignment, given a perturbation at another position (j), as a statistical coupling energy between the two positions ($\Delta\Delta G_{i,j}$). Evaluating the statistical coupling between all positions and all possible perturbations yielded a 321 by 242 matrix (Supplementary Material). This matrix details for each position (321 columns from N- to C-terminus) the effect of all perturbations exhibited by the evolutionary ensemble of naturally occurring sequences (242 rows). Iterative rounds of two-dimensional clustering were used to extract the submatrix that contained positions and perturbations with similarly high patterns of $\Delta\Delta G_{i,j}$ (Supplementary Material). Thus, each iterative step is an attempt to refine the assignment of sets of co-evolving residues by focusing the clustering algorithm around regions of positions and perturbations that show significant values, eventually resulting in the identification of co-evolving networks of positions.

The SCA revealed a subset of 30 positions with a similar pattern of significant $\Delta\Delta G_{i,j}$ values (Table 4). This subset of positions forms a self consistent network of concerted mutations (that is, perturbations at these 30 positions redundantly identified other positions within the subset). These results indicate that the motor domain contains a small set of coevolving positions. Structurally, the 30 positions were found to comprise a web of interconnected residues rather than a single pathway. Nevertheless, the physical connectivity of many of these residues is striking, given that it comprises only 10% of the total residues in the motor domain, and that no structural data was employed for their identification.

The primary hypothesis underlying the SCA approach is that there exists a mutual physical, energetic or functional constraint that can only be satisfied by the correlated mutation of certain positions. Previous application of the SCA method has largely highlighted positions that are localised along a putative allosteric pathway that links known functional sites.[6,48,49,72–74] However, the positions highlighted in the current analysis do not form a single obvious pathway when examined in the context of the available crystal structures. Rather, the identified subset of coevolving positions may reflect a plurality of overlapping structural/

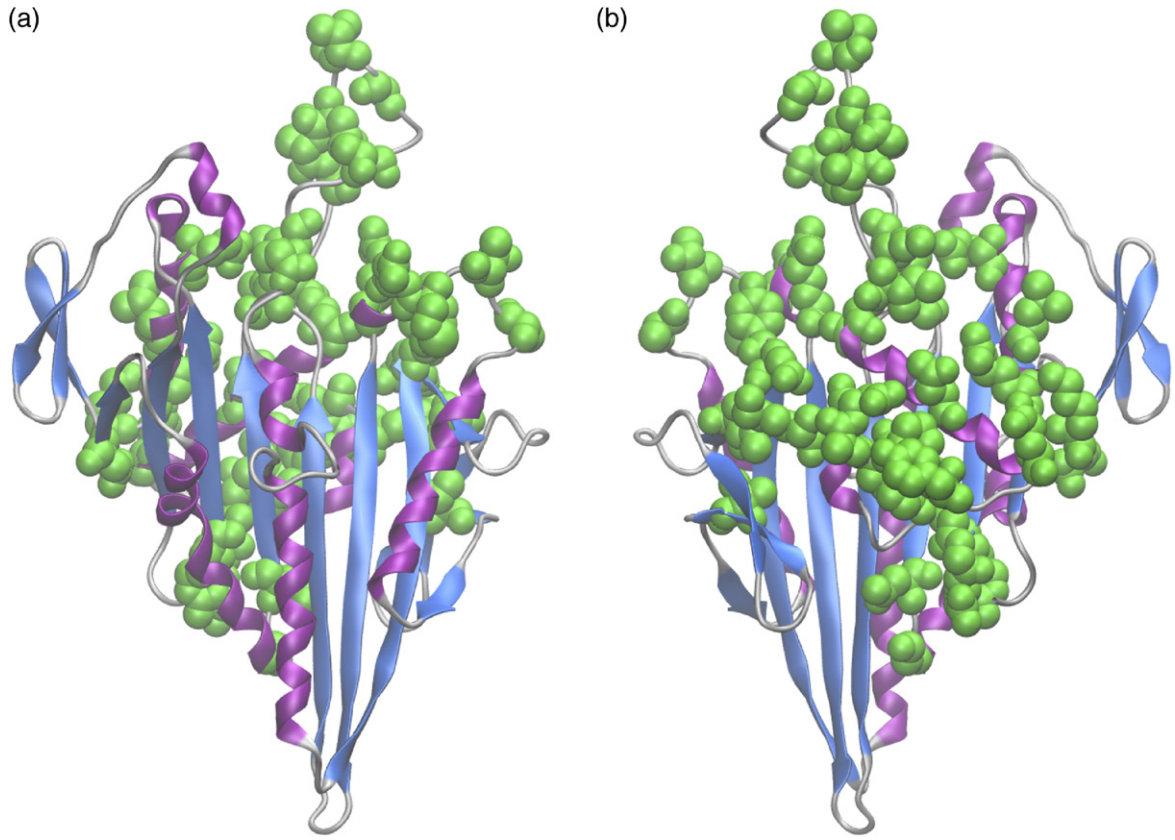

Figure 6. Solvent exposed conserved residues displayed on the front (a) and back (b) of the motor domain as green VDW spheres. The figure was generated using VMD.[51]

functional requirements that include, but are not limited to, allosteric communication. The majority of identified positions are somewhat localised around the putative microtubule binding site and the edge of the peripheral region (Supplementary Data). Interesting contacts are also noticeable between residues in loop 8, α4, loop 12 and α5, as are the dense interactions towards the bottom of α1b, β8, β1 and loop 3. The latter of these regions may be of relevance to the proposed neck-linker docking and undocking cycles,[37,75–77] and the former for microtubule interaction. However, such links without experimental support remain highly speculative.

The SCA approach does not provide a physical mechanism that explains how the identified positions might couple. We suggest that the possible coupling between positions could be addressed further by using time-dependent cross-correlation analyses from suitable molecular dynamics simulations, which might reveal out-of-phase (time-lagged) effects.[78] Such an approach coupled with targeted mutagenesis should help provide further insight into the detailed mechanisms of allosteric communication in kinesin.[79]

Proposed implications: evidence for a web of conformational communication

It should be noted that highly conserved positions, by definition, will not yield information on covariation and thus will not be highlighted by the SCA method. However, such positions (e.g. the switch regions) may be involved in allosteric coupling and should be considered in conjunction with the results of the current analysis. In this respect the spatial locations of both highly conserved and co-evolving positions in the various crystal structures are consistent with a potential pathway of communication that leads from the nucleotide binding site to the α4-loop12-α5-loop13 sub-domain. Of particular note is a conserved histidine in the switch I region (residue H205 from the SSRSH motif) that contacts S257 and L258 in helix α4 and L282 in helix α5. The possible importance of these conserved contacts is emphasised by the PCA results. These indicate that the conformation of H205 and neighbouring positions is directly correlated with sub-domain orientation: different sub-domain conformations are associated with a small (~1.6 Å) displacement of the C-terminal portion of the switch I backbone towards the nucleotide binding site (Figure 5).

Sequence and structure analysis have highlighted a conserved surface patch that encompasses the dynamic α4-loop12-α5-loop13 sub-domain. The hydrophilic character of the identified surface patch is intriguing as studies have indicated that strong binding of kinesin to microtubules (i.e. when kinesin is in the ATP bound state) is predominantly hydrophobic (i.e. gets stronger with increasing ionic strength). This suggests that in the crystal structures

Table 4. Co-evolving positions identified by SCA

Position	Location	Details	Motif
325	L15	[Ivlc][94]	8
323	L15	[Krn][85]	8
315	α6	[Ts][94]	8
311	α6	E[82]	8
300	β8	[Avci][92]	7
299	β8	[Ivfl][97]	7
298	β8	[Mil][94]	7
295	β8	[KRq][85]	7
284	α5	[Rwqyh][86]	7
279	L12	[DNe][91]	7
263	α4	[Ndkertq][88]	6
259	α4	[LStm][93]	6
252	L11	X	6
241	L11	[Ts][91]	6
157	L8	[Ed][72]-	–
110	α2b	[Rln][83]	–
81	β3	[IVlc][98]	2
80	β3	[TCs][95]	2
79	β3	[GAvcs][95]	2
77	L3	[Yfgi][87]	2
75	L3	[EDqnksa][88]	2
72	α1b	[Sdgnhtea][91]	–
62	α1a	[YF][98]	–
58	α1a	[Qnt][96]	–
41	β1b	X	–
13	β1	[Vcil][91]	1
10	β1	[Rkqc][76]	1
9	β1	[IV][89]	1
7	L0	X	–
6	L0	X	–

Refer to Table 1 for an explanation of the sequence conservation notation used in the details column.

we may be looking at the weak binding state, with hydrophobic recognition elements predominantly surface retracted or occluded. Interestingly, several conserved hydrophobic positions from motif 7, [Hfy][91]-[VIt][98]-P[94]-[Yf][99]-P[100] (residues 274 to 278), which are in the immediate vicinity of the identified cluster, have relatively low solvent exposure values but have been linked to microtubule binding, *via* biochemical mutagenesis work.[32] The data suggest that the microtubule binding interface may oscillate between display of such positions (giving strong hydrophobic binding to microtubules), and retraction or concealment of such residues, giving weaker (ADP state) binding based on electrostatic interactions with the microtubule across the identified broad surface of conserved hydrophilic and neutral residues.

A limitation of the current approach is that not all relevant conformations might be represented in the available structures. Indeed it has been proposed, based on extensive structural studies of related G-protein and myosin families, that a conserved switch I serine (SSRSH) and switch II glycine (DLAGSE) should form hydrogen bonds with the γ-phosphate of a bound ATP. However, all current kinesin structures, regardless of sub-domain orientation, have the backbone of these residues hydrogen bonded to each other in an orientation that is incapable of contacting the γ-phosphate of a bound ATP. In addition, the catalytically competent ATP conformation of both myosin and kinesin is likely to require the formation of an inter-switch salt-bridge between the conserved arginine of switch I (SSRSH) and glutamic acid of switch II (DLAGSE).[60] However, only a small subset of kinesin structures possess an inter-switch salt-bridge, all with sub-optimal geometry (PDB codes: 1vfv, 1vfw, 3kar, 1f9t and 1f9u). The presence or absence of this salt-bridge shows no correlation with sub-domain orientation or a displacement of the switch regions towards the nucleotide cleft. In addition the N-terminal portion of the switch I loop and C-terminal portion of the switch II loop displays significant variability between structures with no obvious correlation to sub-domain orientation. This is most pronounced for loop 11, which joins switch II to helix α4. This loop is disordered in the majority of kinesin structures and is unlikely, in this disordered state, to provide a direct mechanical link between these regions. However, it has been suggested that microtubule binding may act to rigidify loop 11 and the switch I containing loop 9, thereby enhancing the coupling between the nucleotide and microtubule binding sites.[40,80]

Conclusions

Molecular evolution provides a natural site-directed mutagenesis experiment that can provide insight into the structure, function, folding and kinship of proteins. Here, information regarding the possible functional and structural roles of particular residues in the kinesin motor domain was inferred from a large dataset of sequences and crystal structures by making an alignment based on both sequence and structure, and then computing the distributions and correlations of various sequence and structural properties. The analysis reveals conserved positions that are subject to strong evolutionary constraints, such that a particular residue or class of residues must be present in a particular spatial context.

The present analysis indicates that all current kinesin crystal structures reside in one of two main conformations that differ in the position of a microtubule-binding sub-domain relative to a rigid central core. This sub-domain consists of secondary structure elements α4-loop12-α5-loop13 and contains a conserved hydrophilic surface patch thought to be involved in strong binding to microtubules. A triple alanine substitution of positively charged residues in loop 12 entirely abolishes microtubule activated ATPase activity.[32]

Our analysis reveals that the α4-loop12-α5-loop13 sub-domain is conformationally associated with the active site, and connected to it *via* a chain or web of residue conservation. The reciprocal, allosteric communication of these two sites is central to the mechanochemical function of the kinesin motor. The apo (empty) ATP and ADP.Pi states of the active site lead to strong (stable) microtubule binding, whilst the ADP state leads to weak (unstable) microtubule binding.[60] Conversely, microtubules activate (accelerate) ADP release from kinesin by

switching its active site conformation. Our analysis suggests that changes in the active site are capable of cyclically reconfiguring the α4-loop12-α5-loop13 sub-domain so as to switch kinesin between weak and strong microtubule binding.

The mobility of helix α4 has been discussed by a number of authors who have noted that the equivalent helix in myosin (termed the "relay helix") also displays a nucleotide dependent rearrangement.[10,14,37,40,81,82] Furthermore, the movement of helix α4 and adjacent loops has been proposed to modulate docking interactions of the neck linker region with the main body of the motor domain.[10,37,76,83] In these models, helix α4 resides in either a conformation stabilised by bound ATP that permits the formation of contacts between the neck-linker and the side of the motor domain, or a conformation stabilised by no nucleotide or bound ADP that sterically hinders this docking.[37,81] It has been further suggested that the transition between neck-linker conformations of a microtubule-bound head could actively translate the second unbound head of a kinesin dimer forward, in the progress direction of the motor.[76] The neck-linker region was excluded from the current analysis as it is divergent in sequence and structure between kinesin sub-families and is absent or unresolved in the majority of high-resolution structures. Nevertheless, in partial agreement with the neck-linker docking model, we note that all structures where the C-terminal neck-linker region contacts the side of the motor domain possess a similar α4-loop12-α5-loop13 sub-domain orientation (structures 1mkj, 1sdm, 2kin, 3kin, 1t5c, 1vfv and 1vfw: see clustering in Figure 3). However, we also observe several structures with this sub-domain orientation, which have an undocked (1goj) or unresolved (1i6i and 1vfx) neck-linker region.

More generally, our results extend previous analyses of residue conservation in the kinesin motor domain and place these on a wider and firmer statistical footing. They provide a framework for directing experiments to sites that are likely to have a functional role, such as mediating microtubule bind-release cycles. We suggest that the approach used here may be applicable to any protein for which sufficient sequence and structural homologues are available. Successful application of the current approach depends critically on the accurate alignment of a diverse range of sequences and structures. In this regard, our work highlights the value of employing structural information to both increase sequence detection sensitivity and alignment accuracy once a relationship has been detected.

Materials and Methods

Unless otherwise noted, all analyses were performed with the recently described Bio3D package†.[84] Atomic coordinates for all available kinesin structures (see Supplementary Data) were obtained from the RCSB Protein Data Bank.[50] Multiple structural alignments were performed with the MUSTANG program[85] and refined with utilities within the Bio3D package. The resulting structure based sequence alignment was used to construct an initial HMM with the HMMER 2.2 package.[45] Alignment of the 143 kinesin motor domain sequences in the Kim and Endow dataset‡[46] to this initial HMM was carried out with the HMMALIGN program in HMMER 2.2.[45] The resultant alignment formed the basis of a new HMM that was used to identify and align available kinesin motor domain sequences in the SWISS-PROT and TrEMBL databases.[47]

The rational for this procedure stems from the observation that structure based alignments outperform multiple sequence alignments for diverse kinesin motor domain structures. A further advantage of the procedure is that the first structure based HMM was searched only against known members of the kinesin family. Since there were no unrelated sequences in this first dataset there was no danger that false positives could be included in the second more sensitive HMM. Therefore, the current procedure creates a clean HMM for database searching that reliably represents the kinesin family and is less error-prone than profiles generated by iterative PSI-BLAST searches of large sequence datasets. The pairwise identity of the aligned sequences follows a close to normal distribution (see Supplementary Data) and the random elimination of sequences from the alignment does not drastically alter the amino acid distribution at all positions (see Supplementary Data).

Sequence conservation analysis

To assess the level of sequence conservation at each position in the alignment, the similarity, identity, class identity and entropy per position were calculated. The "similarity" was defined as the average of the similarity scores of all pairwise residue comparisons for that position in the alignment (where the similarity score between any two residues is the score value between those residues in the BLOSSUM 62 substitution matrix[86]) The "identity" (i.e. the preference for a specific amino acid to be found at a certain position) was assessed by averaging the identity scores resulting from all possible pairwise comparisons at that position in the alignment (where all identical residue comparisons are given a score of 1 and all other comparisons are given a value of 0). The "class identity" was calculated in a similar manner to the "identity". The exception being that amino acids were considered class identical (i.e. assigned a score of 1) if they possessed similar physicochemical properties. For this analysis residues were partitioned into three classes based on their relative hydrophobicity and the extent to which they are distributed between the surface and interior of known globular aqueously soluble protein structures.[2,87] The first class contains hydrophobic residues (C, V, L, I, M, F and W) that have a high probability of residing within protein interiors. The second class contains hydrophilic residues (R, K, E, D, Q and N) that are most likely to be found on the surface of proteins. Finally, the third class is comprised of neutral residues (P, H, Y, G, A, S and T) that have a roughly equal chance of being on the surface or in the interior.

"Entropy" is based on Shannon's information entropy for both a 21-letter alphabet (20 amino acids and a gap

† Available from http://mccammon.ucsd.edu/~(grant/bio3d/

‡ Available from http://www.proweb.org/kinesin/

character) and a seven-letter alphabet (six groups of amino acids and a gap character)[88–90] (equation (1)):

$$S = -\sum_{i}^{N} p_i \log_2 p_i \quad (1)$$

where S is Shannon's entropy, p_i is the frequency of each alphabet's letter at position i and N is the alphabet's size (7 or 21). The six groups chosen were aliphatic (A, V, L, I, M and C), aromatic (F, W, Y and H), polar (S, T, N and Q), positive (K and R), negative (D and E), and finally special conformations (G and P). Entropy scores plotted in Figure 2 are normalized so that conserved (low entropy) columns score 1 and diverse (high entropy) columns score 0 (equation (2)):

$$C = \frac{-\sum_{i}^{N} p_i \log_2 p_i}{\log_2(\min(N_{\text{seq}}, N))} \quad (2)$$

where, C is the normalized entropy, p_i is the frequency of each alphabet's letter at position i, N is the alphabet's size and N_{seq} is the length of the sequence. We define a position to be conserved if the similarity, identity, class identity, entropy 21 or entropy 7 at a position is >0.6. Positions in which more than 30% of the sequences have gaps were excluded from all sequence conservation analysis.

Structural conservation analysis

To complement sequence alignment analysis the conservation of various structural properties at equivalent residue positions in available motor domain structures were examined. The number of residue–residue contacts were calculated in all available structures. Two residues were assumed to be in contact if any two heavy atoms from these residues were closer than 5.0 Å. Percent solvent exposure per position was calculated with the NACCESS program§. A residue was considered to be exposed when the accessible surface area of the residue was more than 40% of the measured accessible surface area of that residue in an extended G-X-G tripeptide context.

Prior to assessing structural variability iterated rounds of structural superposition were used to identify the most structurally invariant region of the motor domain. This procedure entailed excluding those residues with the largest positional differences, before each round of superposition, until only the invariant "core" residues remained.[84]

The structurally invariant core was used as the reference frame for structural alignment of the dataset and the RMSFs of equivalent backbone atoms around their average position were evaluated. PCA was employed to further examine the conformational relationships between the different superposed structures. The application of PCA to both distributions of experimental structures and molecular dynamics trajectories, along with its ability to provide considerable insight into the nature of conformational differences in a range of protein families has been previously discussed.[91–95] Briefly, PCA is based on the diagonalization of the covariance matrix, $\boldsymbol{C}$, with elements C_{ij}, built from the Cartesian coordinates, $\boldsymbol{r}$, of the superposed motor domain structures (equation (3)):

$$C_{ij} = \langle (r_i - \langle r_i \rangle) \cdot (r_j - \langle r_j \rangle) \rangle \quad (3)$$

where i and j represent all possible pairs of $3N$ Cartesian coordinates (where N is the number of atoms) being considered. The eigenvectors of the covariance matrix correspond to a linear basis set of the distribution of structures, referred to as PCs, whereas the eigenvalues provide the variance of the distribution along the corresponding eigenvectors. Projecting the kinesin structures into the sub-space defined by the largest principal components (along which the sample variance is largest) resulted in a lower dimensional representation of the structural dataset (see Figure 3 for details). The resulting low-dimensional "conformer plots", succinctly display the major differences between structures, highlight relationships between different specific conformers and thus enable the interpretation and characterization of multiple inter-conformer relationships.[84]

Statistical coupling analysis

Statistical coupling energies were calculated as described in Lockless & Ranganathan[6] with code provided by the authors. The calculation is based on selecting a subset of sequences (i.e. a sub-alignment) from a multiple sequence alignment and comparing the characteristics of the sub-alignment with the characteristics of the full alignment. Consider two positions i and j; from the full alignment, a subset of sequences are chosen by placing a constraint on the identity of the residue occupying position i. For example, choosing all sequences that contain a glycine residue at position i in the original alignment. Next, the degree of bias present at a second position j in the chosen sub-alignment is assessed in relation to the distribution of residues at position j in the full alignment. This involves calculating a conservation parameter (termed the positional energy, ΔGstat) for column j in both the full and sub-alignments. If substitutions at positions i and j occur independently, then their amino acid distribution (and hence ΔGstat value) should remain similar in both the sub-alignment and the full alignment. However, if positions i and j co-vary, then the composition at position j in the subset may be biased by the constraint imposed upon position i, thus yielding a difference in ΔGstat. Such differences are quantified by a $\Delta\Delta G$stat parameter that the original authors refer to as the "statistical coupling energy". Calculation of $\Delta\Delta G$stat for all positions given a perturbation at position i, is a mapping of how all positions in the protein feel the effect of perturbing position i (see Lockless & Ranganathan[6] for further details). Iterative rounds of two-dimensional clustering (of $\Delta\Delta G$stat values determined for all positions and all perturbations) is then used to extract the sub-matrix that contains positions and perturbations with similarly high patterns of $\Delta\Delta G$stat.

§ Hubbard, S. & Thornton, J. M. (1993). NACCESS, computer program. Department of Biochemistry and Molecular Biology. University College London, London UK.

Acknowledgements

We thank Ana Rodrigues and members of the Cross, Caves and McCammon groups for fruitful and entertaining discussions. This work was supported in part by Marie Curie Cancer Care, the University of York, the National Institutes of Health, National Science Foundation, the Howard Hughes Medical Institute, the National Biomedical Computation Resource, and the National

Science Foundation Center for Theoretical Biological Physics.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.02.049

References

1. Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.
2. Chothia, C., Gelfand, I. & Kister, A. (1998). Structural determinants in the sequences of immunoglobin variable domain. *J. Mol. Biol.* **278**, 457–479.
3. Larson, S. M. & Davidson, A. R. (2000). The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Sci.* **9**, 2170–2180.
4. Lesk, A. M. & Fordhman, W. D. (1996). Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *J. Mol. Biol.* **258**, 501–537.
5. Michnick, S. W. & Shakhnovich, E. (1998). A strategy for detecting the conservation of folding nucleus residues in protein superfamilies. *Fold. Des.* **3**, 239–251.
6. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
7. Kull, F. J., Sablin, E. P., Lau, R., Fletterick, R. J. & Vale, R. D. (1996). Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature*, **380**, 550–555.
8. Kozielski, F., Sack, S., Marx, A., Thormahlen, M., Schonbrunn, E., Biou, V. *et al.* (1997). The crystal structure of dimeric kinesin and implications for microtubule-dependent motility. *Cell*, **91**, 985–994.
9. Kozielski, F. K., De Bonis, S., Burmeister, W., Cohen-Addad, C. & Wade, R. (1999). The crystal structure of a minus-end directed microtubule motor protein ncd reveals variabledimer conformations. *Structure*, **7**, 1407–1416.
10. Kikkawa, M., Sablin, E. P., Okada, Y., Yajima, H., Fletterick, R. J. & Hirokawa, N. (2001). Switch-based mechanism of kinesin motors. *Nature*, **411**, 439–445.
11. Turner, J., Anderson, R., Guo, J., Beraud, C., Fletterick, R. & Sakowicz, R. (2001). Crystal structure of the mitotic spindle kinesin Eg5 reveals a novel conformation of the neck-linker. *J. Biol. Chem.* **276**, 25496–25502.
12. Sablin, E. P., Kull, F. J., Cooke, R., Vale, R. D. & Fletterick, R. J. (1996). Crystal structure of the motor domain of the kinesin-related motor ncd. *Nature*, **380**, 555–559.
13. Gulick, A. M., Song, H., Endow, S. A. & Rayment, I. (1998). X-ray crystal structure of the yeast Kar3 motor domain complexed with Mg.ADP to 2.3 A resolution. *Biochemistry*, **37**, 1769–1776.
14. Yun, M., Zhang, X., Park, C. G., Park, H. W. & Endow, S. A. (2001). A structural pathway for activation of the kinesin motor ATPase. *EMBO J.* **20**, 2611–2618.
15. Song, Y. H., Marx, A., Muller, J., Woehlke, G., Schliwa, M., Krebs, A. *et al.* (2001). Structure of a fast kinesin: implications for ATPase mechanism and interactions with microtubules. *EMBO J.* **20**, 6213–6225.
16. Yun, M., Bronner, C. E., Park, C. G., Cha, S. S., Park, H. W. & Endow, S. A. (2003). Rotation of the stalk/neck and one head in a new crystal structure of the kinesin motor protein, Ncd. *EMBO J.* **22**, 5382–5389.
17. Shipley, K., Hekmat-Nejad, M., Turner, R., Moores, C., Anderson, R., Milligan, R. *et al.* (2004). Structure of a kinesin microtubule depolymerization machine. *EMBO J.* **23**, 1422–1429.
18. Yan, Y., Sardana, V., Xu, B. B., Homnick, C., Halczenko, W., Buser, C. A. *et al.* (2004). Inhibition of a mitotic motor protein: where, how, and conformational consequences. *J. Mol. Biol.* **333**, 547–556.
19. Sindelar, C. V., Budny, M. J., Rice, S., Naber, N., Fletterick, R. & Cooke, R. (2002). Two conformations in the human kinesin power stroke defined by X-ray crystallography and EPR spectroscopy. *Nature Struct. Biol.* **9**, 844–848.
20. Vinogradova, M. V., Reddy, V. S., Reddy, A. S., Sablin, E. P. & Fletterick, R. J. (2004). Crystal structure of kinesin regulated by Ca(2+)-calmodulin. *J. Biol. Chem.* **279**, 23504–23509.
21. Garcia-Saez, I., Yen, T., Wade, R. H. & Kozielski, F. (2004). Crystal structure of the motor domain of the human kinetochore protein CENP-E. *J. Mol. Biol.* **340**, 1107–1116.
22. Ogawa, T., Nitta, R., Okada, Y. & Hirokawa, N. (2004). A common mechanism for microtubule destabilizers-M type kinesins stabilize curling of the protofilament using the class-specific neck and loops. *Cell*, **116**, 591–602.
23. Nitta, R., Kikkawa, M., Okada, Y. & Hirokawa, N. (2004). KIF1A alternately uses two loops to bind microtubules. *Science*, **305**, 678–683.
24. Cox, C. D., Breslin, M. J., Mariano, B. J., Coleman, P. J., Buser, C. A., Walsh, E. S. *et al.* (2005). Kinesin spindle protein (KSP) inhibitors. Part 1: the discovery of 3,5-diaryl-4,5-dihydropyrazoles as potent and selective inhibitors of the mitotic kinesin KSP. *Bioorg. Med. Chem. Leters*, **15**, 2041–2045.
25. Sablin, E. P., Case, R. B., Dai, S. C., Hart, C. L., Ruby, A., Vale, R. D. & Fletterick, R. J. (1998). Direction determination in the minus-end-directed kinesin motor ncd. *Nature*, **395**, 813–816.
26. Sack, S., Muller, J., Marx, A., Thormahlen, M., Mandelkow, E. M., Brady, S. T. & Mandelkow, E. (1997). X-ray structure of motor and neck domains from rat brain kinesin. *Biochemistry*, **36**, 16155–16165.
27. Fraley, M. E., Garbaccio, R. M., Arrington, K. L., Hoffman, W. F., Tasber, E. S., Coleman, P. J. *et al.* (2006). Kinesin spindle protein (KSP) inhibitors. Part 2: the design, synthesis, and characterization of 2,4-diaryl-2,5-dihydropyrrole inhibitors of the mitotic kinesin KSP. *Bioorg. Med. Chem. Letters*, **16**, 1775–1779.
28. Tarby, C. M., Kaltenbach, R. F., 3rd, Huynh, T., Pudzianowski, A., Shen, H., Ortega-Nanos, M. *et al.* (2006). Inhibitors of human mitotic kinesin Eg5: characterization of the 4-phenyl-tetrahydroisoquinoline lead series. *Bioorg. Med. Chem. Letters*, **16**, 2095–2100.
29. Kim, K. S., Lu, S., Cornelius, L. A., Lombardo, L. J., Borzilleri, R. M., Schroeder, G. M. *et al.* (2006). Synthesis and SAR of pyrrolotriazine-4-one based Eg5 inhibitors. *Bioorg. Med. Chem. Letters*, **16**, 3937–3942.
30. Cox, C. D., Torrent, M., Breslin, M. J., Mariano, B. J., Whitman, D. B., Coleman, P. J. *et al.* (2006). Kinesin spindle protein (KSP) inhibitors. Part 4: structure-based design of 5-alkylamino-3,5-diaryl-4,5-dihydropyrazoles as potent, water-soluble inhibitors of the mitotic kinesin KSP. *Bioorg. Med. Chem. Letters*, **16**, 3175–3179.
31. Alonso, M. C., van Damme, J., Vandekerckhove, J. &

Cross, R. A. (1998). Proteolytic mapping of kinesin/ncd-microtubule interface: nucleotide-dependent conformational changes in the loops L8 and L12. *EMBO J.* **17**, 945–951.
32. Woehlke, G., Ruby, A. K., Hart, C. L., Ly, B., Hom-Booher, N. & Vale, R. D. (1997). Microtubule interaction site of the kinesin motor. *Cell*, **90**, 207–216.
33. Hirose, K., Akimaru, E., Akiba, T., Endow, S. A. & Amos, L. A. (2006). Large conformational changes in a kinesin motor catalyzed by interaction with microtubules. *Mol. Cell*, **23**, 913–923.
34. Kikkawa, M. & Hirokawa, N. (2006). High-resolution cryo-EM maps show the nucleotide binding pocket of KIF1A in open and closed conformations. *EMBO J.* **25**, 4187–4194.
35. Kull, F. J., Vale, R. D. & Fletterick, R. J. (1998). The case for a common ancestor: kinesin and myosin motor proteins and G proteins. *J. Musc. Res. Cell Motil.* **19**, 877–886.
36. Vale, R. D. (1996). Switches, latches, and amplifiers: common themes of G proteins and molecular motors. *J. Cell Biol.* **135**, 291–302.
37. Vale, R. D. & Milligan, R. A. (2000). The way things move: looking under the hood of molecular motor proteins. *Science*, **288**, 88–95.
38. Cross, R. A. & Carter, N. J. (2000). Molecular motors. *Curr. Biol.* **10**, R177–R179.
39. Sack, S., Kull, F. J. & Mandelkow, E. (1999). Motor proteins of the kinesin family. Structures, variations, and nucleotide binding sites. *Eur. J. Biochem.* **262**, 1–11.
40. Kull, F. J. & Endow, S. A. (2002). Kinesin: switch I and II and the motor mechanism. *J. Cell Sci.* **115**, 15–23.
41. Song, Y. H., Marx, A. & Mandelkow, E. (2003). Structures of kinesin motor domains: implications for conformational switching involved in mechanochemical coupling. In *Molecular Motors* (Schliwa, M., ed), pp. 287–303. Wiley-vch, Weinheim, Germany.
42. Wickstead, B. & Gull, K. (2006). A "holistic" kinesin phylogeny reveals new kinesin families and predicts protein functions. *Mol. Biol. Cell*, **17**, 1734–1743.
43. Miki, H., Okada, Y. & Hirokawa, N. (2005). Analysis of the kinesin superfamily: insights into structure and function. *Trends Cell Biol.* **15**, 467–476.
44. Lawrence, C. J., Dawe, R. K., Christie, K. R., Cleveland, D. W., Dawson, S. C., Endow, S. A. *et al.* (2004). A standardized kinesin nomenclature. *J. Cell Biol.* **167**, 19–22.
45. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
46. Kim, A. J. & Endow, S. A. (2000). A kinesin family tree. *J. Cell Sci.* **113**, 3681–3682.
47. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
48. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. (2002). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.* **10**, 59–69.
49. Hatlry, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G. & Ranganathan, R. (2003). Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl Acad. Sci. USA*, **100**, 14445–14450.
50. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2002). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
51. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD–visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38.
52. Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins: Struct. Funct. Genet.* **48**, 227–241.
53. Kull, F. J. (2000). Motor proteins of the kinesin superfamily: structure and mechanism. *Essays Biochem.* **35**, 61–73.
54. Flocco, M. M. & Mowbray, S. L. (1995). C alpha-based torsion angles: a simple tool to analyze protein conformational changes. *Protein Sci.* **4**, 2118–2122.
55. Case, R. B., Rice, S., Hart, C. L., Ly, B. & Vale, R. D. (2000). Role of the kinesin neck linker and catalytic core in microtubule-based motility. *Curr. Biol.* **10**, 157–160.
56. Houdusse, A., Szent-Gyorgyi, A. G. & Cohen, C. (2000). Three conformational states of scallop myosin S1. *Proc. Natl Acad. Sci. USA*, **97**, 11238–11243.
57. Muller, J., Marx, A., Sack, S., Song, Y. H. & Mandelkow, E. (1999). The structure of the nucleotide-binding site of kinesin. *Biol. Chem.* **380**, 841–992.
58. Vetter, I. R. & Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. *Science*, **294**, 1299–1304.
59. Wittinghofer, A. & Nassar, N. (1996). How Ras related proteins talk to their effectors. *Trends Biochem. Sci.* **21**, 488–491.
60. Cross, R. A. (2004). The kinetic mechanism of kinesin. *Trends Biochem. Sci.* **29**, 301–309.
61. Gerstein, M. & Altman, R. B. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175.
62. Williams, R. J. (1993). Are enzymes mechanical devices? *Trends Biochem. Sci.* **18**, 115–117.
63. Hirose, K., Lowe, J., Alonso, M., Cross, R. A. & Amos, L. A. (1999). Congruent docking of dimeric kinesin and ncd into three-dimensional electron cryomicroscopy maps of microtubule-motor ADP complexes. *Mol. Biol. Cell*, **10**, 2063–2074.
64. Hirose, K., Lockhart, A., Cross, R. A. & Amos, L. A. (1996). Three-dimensional cryoelectron microscopy of dimeric kinesin and ncd motor domains on microtubules. *Proc. Natl Acad. Sci. USA*, **93**, 9344–9539.
65. Hirose, K., Lockhart, A., Cross, R. A. & Amos, L. A. (1995). Nucleotide-dependent angular change in kinesin motor domain bound to tubulin. *Nature*, **376**, 277–279.
66. Hirose, K., Henningsen, U., Schliwa, M., Toyoshima, C., Shimizu, T., Alonso, M., Cross, R. A. & Amos, L. A. (2000). Structural comparison of dimeric Eg5, *Neurospora* kinesin (Nkin) and Ncd head-Nkin neck chimera with conventional kinesin. *EMBO J.* **19**, 5308–5314.
67. Hirose, K., Amos, W. B., Lockhart, A., Cross, R. A. & Amos, L. A. (1997). Three-dimensional cryo-electron microscopy of 16-protofilament microtubules: structure, polarity, and interaction with motor proteins. *J. Struct. Biol.* **118**, 140–148.
68. Hoenger, A., Thormahlen, M., Diaz-Avalos, R., Doerhoefer, M., Goldie, K. N., Muller, J. & Mandelkow, E. (2000). A new look at the microtubule binding patterns of dimeric kinesins. *J. Mol. Biol.* **297**, 1087–1103.
69. Hoenger, A., Sack, S., Thormahlen, M., Marx, A., Muller, J., Gross, H. & Mandelkow, E. (1998). Image reconstructions of microtubules decorated with monomeric and dimeric kinesins: comparison with X-ray structure and implications for motility. *J. Cell Biol.* **141**, 419–430.
70. Hoenger, A., Sablin, E. P., Vale, R. D., Fletterick, R. J. & Milligan, R. A. (1995). Three-dimensional structure of a tubulin-motor-protein complex. *Nature*, **376**, 271–274.

71. Sosa, H., Dias, D. P., Hoenger, A., Whittaker, M., Wilson-Kubalek, E., Sablin, E. *et al.* (1997). A model for the microtubule-Ncd motor protein complex obtained by cryo-electron microscopy and image analysis. *Cell*, **90**, 217–224.
72. Shulman, A. I., Larson, C., Mangelsdorf, D. J. & Ranganathan, R. (2003). Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell*, **116**, 429–517.
73. Dekker, J. P., Fodor, A., Aldrich, R. & Yellen, G. (2004). A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.
74. Kass, L. & Horovitz, A. (2002). Mapping pathways of allosteric communication in GROEL by analysis of correlated mutations. *Proteins: Struct. Funct. Genet.* **48**, 611–617.
75. Romberg, L., Pierce, D. W. & Vale, R. D. (1998). Role of the kinesin neck region in processive microtubule-based motility. *J. Cell Biol.* **140**, 1407–1416.
76. Rice, S., Lin, A. W., Safer, D., Hart, C. L., Naber, N., Carragher, B. O. *et al.* (1999). A structural change in the kinesin motor protein that drives motility. *Nature*, **402**, 778–784.
77. Vale, R. D., Case, R., Sablin, E., Hart, C. & Fletterick, R. (2000). Searching for kinesin's mechanical amplifier. *Phil. Trans. Roy. Soc. ser. B*, **355**, 449–457.
78. Sharp, K. & Skinner, J. J. (2006). Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. *Proteins: Struct. Funct. Genet.* **65**, 347–361.
79. Kuriyan, J. (2004). Allostery and coupled sequence variation in nuclear hormone receptors. *Cell*, **116**, 354–356.
80. Naber, N., Minehardt, T. J., Rice, S., Chen, X., Grammer, J., Matuska, M. *et al.* (2003). Closing of the nucleotide pocket of kinesin-family motors upon binding to microtubules. *Science*, **300**, 798–801.
81. Sablin, E. P. & Fletterick, R. J. (2001). Nucleotide switches in molecular motors: structural analysis of kinesins and myosins. *Curr. Opin. Struct. Biol.* **11**, 716–724.
82. Woehlke, G. (2001). A look into kinesin's powerhouse. *FEBS Letters*, **508**, 291–294.
83. Asenjo, A. B., Weinberg, Y. & Sosa, H. (2006). Nucleotide binding and hydrolysis induces a disorder-order transition in the kinesin neck-linker region. *Nature Struct. Mol. Biol.* **13**, 648–654.
84. Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A. & Caves, L. S. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.
85. Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins: Struct. Funct. Genet.* **64**, 559–574.
86. Henikoff, S. & G., H. J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
87. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656.
88. Shehkin, P. S., Erman, B. & Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins: Struct. Funct. Genet.* **11**, 297–313.
89. Shannon, C. E. (1948). The mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423.
90. Shannon, C. E. (1948). The mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656.
91. Elsawy, K. M., Hodgson, M. K. & Caves, L. S. D. (2005). The physical determinants of the DNA conformational landscape. *Nuclei. Acids Res.* **33**, 5749–5762.
92. van Aalten, D. M. F., Conn, D. A., de Groot, B. L., Berendsen, H. J., Findlay, J. B. & Amadei, A. (1997). Protein dynamics derived from clusters of crystal structures. *Biophys. J.* **73**, 2891–2896.
93. Abseher, R., Horstink, L., Hilbers, C. & Nilges, M. (1998). Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins: Struct. Funct. Genet.* **31**, 370–382.
94. Caves, L. S. D., Nguyen, D. T. & Hubbard, R. E. (1991). Conformational variability of insulin: a molecular dynamics analysis. In *Molecular Dynamics: Applications in molecular biology* (Goodfellow, J. M., ed), pp. 27–68, The Macmillan Press Ltd, London.
95. Caves, L. S. D., Evanseck, J. D. & Karplus, M. (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.* **7**, 649–666.

Edited by M. Levitt

(Received 13 December 2006; received in revised form 30 January 2007; accepted 6 February 2007)
Available online 30 March 2007