Fold and flexibility: what can proteins’ mechanical properties tell us about their folding nucleus?

Sophie Sacquin-Mora

Laboratoire de Biochimie Théorique, CNRS UPR9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France

The determination of a protein’s folding nucleus, i.e. a set of native contacts playing an important role during its folding process, remains an elusive yet essential problem in biochemistry. In this work, we investigate the mechanical properties of 70 protein structures belonging to 14 protein families presenting various folds using coarse-grain Brownian dynamics simulations. The resulting rigidity profiles combined with multiple sequence alignments show that a limited set of rigid residues, which we call the consensus nucleus, occupy conserved positions along the protein sequence. These residues’ side chains form a tight interaction network within the protein’s core, thus making our consensus nuclei potential folding nuclei. A review of experimental and theoretical literature shows that most (above 80%) of these residues were indeed identified as folding nucleus member in earlier studies.

1. Introduction

It is a truth universally acknowledged (at least among biochemists) that most proteins must fold into a specific and functionally relevant three-dimensional structure. However, after decades of research, understanding how this spontaneous folding process can be achieved still remains a fundamental problem and a grand challenge in biochemistry [1–6].

Numerous experimental and theoretical approaches suggest that most protein folding processes follow a nucleation–condensation mechanism, where the rate-limiting step is a formation of a specific folding nucleus presenting a set of native contacts, a concept introduced in a seminal paper by Abkevich et al. [7–9]. Following this initial step, the transition towards the fully folded state is fast and efficient [10,11]. Other proteins will follow a hierarchical mechanism, which is characterized by the successive condensation of specific secondary structural units (termed foldons) [12–14], while in some protein families (such as the homeodomain) the different members will present mixed folding mechanisms ranging from pure nucleation–condensation to pure hierarchical [15].

In the last 20 years, considerable efforts, both on the experimental and theoretical sides, have been devoted to the identification of the folding nuclei in numerous proteins of various folds [16]. Experimentally, $\phi$-value analysis [8] and hydrogen–deuterium exchange [17] (H–X) are among the classical tools used to identify core residues, i.e. fully buried residues with no solvent accessibility, playing a key part in the protein folding process. Theoretical works include, among others, simulations ranging from simple lattice models [18–20] to fully atomistic molecular dynamics [21–24], various network approaches [25–28] and evolutionary methods based on sequence analysis within a protein family, which showed that the folding nucleus is significantly more conserved than the rest of the protein [29–32].

Protein mechanical properties on the other hand are known to be a key aspect of protein function [33–36]. In order to investigate them on a residue level, we developed the ProPhET (Probing Protein Heterogeneity) program [37], which combines a coarse-grain (CG) elastic network representation and Brownian dynamics (BD). Studies using ProPhET on various proteins indicate that some
residues present specific mechanical properties that can be related to their biological activity, such as catalytic residues [38], residues controlling ligand migration within protein cavity networks [39,40], or residues located on the inter-domain interfaces from multi-domain proteins [41,42]. In particular, an early work focusing on haemoproteins [43], and more specifically proteins presenting domains with globin or cytochrome c folds, showed that residues forming the folding nucleus in these two families could be identified via their remarkably high force constants within the protein structure. Following this first observation, we applied our force constant analysis on a set of 70 single protein domains corresponding to 14 protein folds. The CG BD calculations were combined with COBALT [44] sequence alignment of proteins from the same family (and presenting similar folds) in order to identify another experimental and theoretical data shows that our method can sequence. The comparison of our results with earlier experimental and theoretical data shows that our method can

2. Material and methods

2.1. Coarse-grain simulations

CG BD simulations were performed using the ProPHET program [37,38,43] on a set of 70 protein crystallographic structures from the protein data bank [45] belonging to 14 protein folds (4 all α, 2 all β, and 8 α/β; table 1).

Diverging from most common CG models, where each residue is described by a single pseudo-atom [46], ProPHET uses a more detailed model, originally developed by Zacharias [47], enabling different residues to be distinguished. The amino acids are represented by one pseudo-atom located at the Co position, and either one or two (for larger residues) pseudo-atoms replacing the side chain (with the exception of Gly) [47], thus leading to a reduction of the number of heavy atoms (non-hydrogen atoms) in the system by a factor of 3 approximately. Interactions between the pseudo-atoms are treated according to the standard elastic network model (ENM) [46], that is, pseudo-atoms closer than pseudo-atoms are treated according to the standard elastic network model (ENM) [46], that is, pseudo-atoms closer than

residue

Table 1. Protein folds studied in this work. Residues already identified as potential folding nucleus members via earlier experimental or theoretical studies are highlighted in bold. Functional residues (like catalytic residues) are highlighted in italic.

<table>
<thead>
<tr>
<th>protein family, fold type</th>
<th>reference PDB</th>
<th>other PDB structures</th>
<th>consensus nucleus (residue numbers from the reference PDB)</th>
<th>sequences identity range</th>
</tr>
</thead>
<tbody>
<tr>
<td>cytochrome c (all α)</td>
<td>1HRC</td>
<td>1CRI, 1CRY, 1COT, 3C2C</td>
<td>His18, Leu32, Leu35, Trp59, Leu64, Tyr67, Leu68, Leu94, Leu98</td>
<td>32 – 61%</td>
</tr>
<tr>
<td>calmodulin (all α)</td>
<td>3IF7</td>
<td>2SCP, 2SAS, 1RRO, 2PAL, 1REC</td>
<td>Ala88, Phe92, Val108, Leu112, Met124, Phe141</td>
<td>11 – 45%</td>
</tr>
<tr>
<td>myoglobin (all α)</td>
<td>1MBA</td>
<td>1MYT, 1DXT_A, 1DXT_B, 1ECA, 1GDI</td>
<td>Phe28, Leu71, Val108, Phe112, Phe134</td>
<td>14 – 44%</td>
</tr>
<tr>
<td>spectrin (all α)</td>
<td>1U4Q-R15</td>
<td>1U4Q-R16, 1U4Q-R17, 1QUU-R1, 1QUU-R2</td>
<td>Phe1675, Leu1679, Leu1719, Ile1747, Phe1751, Ile1754</td>
<td>10 – 34%</td>
</tr>
<tr>
<td>immunoglobulin (all β)</td>
<td>1TEN</td>
<td>1TTT, 1TTF, 1K85</td>
<td>Ile821, Ile833, Leu835, Val871, Leu873</td>
<td>8 – 33%</td>
</tr>
<tr>
<td>SH3 domain (all β)</td>
<td>1SR1</td>
<td>1SHG, 1FYN, 1J08, 2PNI</td>
<td>Phe26, Leu32, Ala45, Ile56</td>
<td>24 – 32%</td>
</tr>
<tr>
<td>acylphosphatase (α/β)</td>
<td>1APS</td>
<td>2GV1, 1Y90, 2ACY</td>
<td>Tyr11, Ala26, Val39, Val51, Leu67</td>
<td>17 – 54%</td>
</tr>
<tr>
<td>α-lactalbumin (α/β)</td>
<td>1HML</td>
<td>3BOK, 1F6S, 2HL, 7EQL, 3EL1</td>
<td>Met30, Leu52, Phe53, Ile55, Ala92, Trp104</td>
<td>35 – 94%</td>
</tr>
<tr>
<td>barnase (α/β)</td>
<td>1BSE</td>
<td>1GMP, 1RNT, 1AQZ</td>
<td>Ala74, Arg87, Ile88, Val89, Tyr90</td>
<td>8 – 31%</td>
</tr>
<tr>
<td>chymotrypsin inhibitor II</td>
<td>2C12</td>
<td>2SEC_I, 1TIN, 1CIS</td>
<td>Ile39, Val66, Leu68, Pro80</td>
<td>28 – 83%</td>
</tr>
<tr>
<td>protein S6 (α/β)</td>
<td>2KJV</td>
<td>2J5A, 1VMB, 106X, 1FHT, 2KW</td>
<td>Val6, Ile8, Ile26, Leu61</td>
<td>15 – 32%</td>
</tr>
<tr>
<td>ribonuclease A (α/β)</td>
<td>5RSA</td>
<td>2K11, 1ONC, 1ANG</td>
<td>Phe8, His12, Asn44, Thr45, Phe46, Val47, Phe120</td>
<td>24 – 70%</td>
</tr>
<tr>
<td>ribonuclease H (α/β)</td>
<td>2RN2</td>
<td>1RIL, 1HRH_A, 3H08_A</td>
<td>Thr9, Tyr22, Ala24, Ala51, Ala52, Thr69</td>
<td>19 – 53%</td>
</tr>
<tr>
<td>ubiquitin (α/β)</td>
<td>1UBI</td>
<td>1FRD, 1G0, 1PUT, 1RFA, 1STE-D2, 2PIA-D2</td>
<td>Ile3, Val5, Val26, Ile30, Leu43, Leu67</td>
<td>nan-31%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The inverse of these fluctuations yields an effective force constant \( k_i \) for each pseudo-atom belonging to the remaining protein residues. Each pseudo-atom interacts with all the other pseudo-atoms belonging to a given amino acid via the following criterion:

\[
\text{dist}(\mathbf{r}_i, \mathbf{r}_j) = \sum_{k} \frac{1}{2k_i} \left( \frac{1}{r_{ij}^2} - \frac{1}{r_{ij}^3} \right)
\]

where \( r_{ij} \) is the distance between the pseudo-atoms and \( k_i \) is a force constant for each pseudo-atom \( i \). The total force \( \mathbf{F}_i \) on pseudo-atom \( i \) is the sum of the forces on it from all other pseudo-atoms in the simulation.

The simulations are performed using a desktop computer. For a complete BD simulation of all the protein residues, a system of 200,000 BD steps at 300 K takes around 10 hours on a standard desktop computer.

Following earlier studies, which showed that some small ligands had little influence on calculated force constants [38, 43], we chose not to include ligands or prosthetics groups (such as haeme groups) in the protein representations.

Finally, for each protein, we define rigid residues as the residues presenting a force constant \( k_i \) that satisfies the following criterion:

\[
k_i \geq (k_p + \sigma(k_p))_p
\]

where \( k_p \) is the mean force constant for each residue and \( \sigma(k_p) \) its variance. As shown in electronic supplementary material, table SI1, in our dataset, this criterion leads to the selection of around 10% of each protein’s residues.

### 2.2 Sequence alignments

Sequences from proteins (or protein domains) belonging to the same protein fold family were aligned using the online COBALT [44] tool (http://www.st-va.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?CMD=Web) with the defaults parameters.
Rigid residues that were identified using the definition given in the previous section are highlighted in bold and colour along the protein sequences in Figure 1a and all throughout electronic supplementary material, figure SI1.

3. Results and discussion

As shown in previous mechanical studies done on members of the globin family [39] and homologous hydrogenases [40], proteins with similar folds also present similar rigidity profiles. For each protein fold that was studied, we define the consensus nucleus as the subset of conserved positions along the protein sequence that present rigidity peaks in all (or most of, in the case of largest alignments) the rigidity profiles of the structures that were studied. Figure 1a illustrates our approach in the case of cytochrome c, with rigid residues shown in bold red and conserved rigid residues, i.e. consensus nucleus residues, framed and shown in bold green along the aligned sequences. One should note that in this consensus nucleus, it is the positions of the rigid residues that are conserved along the sequence and not necessarily the amino acids themselves (although it is often the case). For example, in cytochrome c, the position of Leu64 can be also occupied by a methionine or a phenylalanine in the alignment. The same colour code applies for the complete set of sequence alignments that is displayed in electronic supplementary material, figure SI1. All the rigidity profiles generated for the 70 protein structures that were studied are also shown in electronic supplementary material, figure SI2.

We systematically compared consensus nuclei obtained from our joint mechanistic and evolutionary approach with earlier theoretical and experimental predictions of the folding nucleus for those same protein folds. Table 1 gives a summary of the 14 protein folds that were studied and the resulting consensus nuclei.

3.1. All α folds

3.1.1. Cytochrome c

Cytochrome c (104 residues) sequences from horse heart (1HRC) and Saccharomyces cerevisiae (1CRD) were aligned with cytochrome c2 sequences from Blastochloris viridis (1CRY), Paracoccus denitrificans (1COT) and Rhodospirillum rubrum (3C2C). After the sequence alignment, nine positions are systematically occupied by rigid residues (Figure 1). Three of these (shown in blue in figure 1c), namely His18, Trp59 and Tyr67, correspond to highly conserved functional residues bound to the haeme prosthetic group in the protein’s core [55]. The remaining six, Leu32, Leu35, Leu64, Leu68, Leu94 and Leu98, are likely to contribute to the fold stability. Leu32, Leu64, Leu86, Leu94 and Leu98 occupy topohydrophobic positions (i.e. positions occupied by hydrophobic residues in each member of the protein family), which are known to be good candidates for involvement in the folding nucleus [30], and H–X experiments showed that Leu32, Leu64, Leu68, Leu94 and Leu98 are in rapidly protected positions [17,56]. Finally, kinetic [57] and mutagenesis [55] experiments highlighted the importance of Leu94 for proper folding and this residue also belongs to the cytochrome c folding nucleus defined by Ptitsyn [58].

3.1.2. Calmodulin

Calmodulin (3IF7, 143 residues) was compared with oncomodulin (1RRO), recoverin (1REC), parvalbumin (2PAL) and sarcoplasmic calcium-binding proteins from sandworm (2SCP) and amphioxus (2SAS). Due to the larger number of sequences in this alignment, our selection criterion was lightened, and we selected for the consensus nucleus those rigid residues which occupy the same position in at least four out of the six aligned sequences. This criterion leads to a consensus nucleus made of six residues: Ala88, Phe92, Val108, Leu112, Met124 and Phe141 (Figure 2a,b) and for each aligned sequence, at least four residues from this nucleus are rigid (electronic supplementary material, figure SI1). Within the consensus nucleus, Phe92, Val108, Leu112 and Phe141 occupy topohydrophobic positions [30].

3.1.3. Myoglobin

The globin fold (around 150 residues) was investigated using the structures and sequence from myoglobin (1MBA), tuna metmyoglobin (1MYT), the α and β chains from haemoglobin (1DXT, chains A and B), erythrocruorin (1IEA) and leghaemoglobin (1GDI). The consensus nucleus resulting from the rigidity profiles and sequence alignment comprises five residues: Phe28-B9, Leu71-E15, Val108-G8, Phe112-G12 and Phe134-H12 (Figure 2c,d), with B9, E15, G8, G12 and H12 indicating residue positions along helices B, E, G and H, respectively [59]. NMR experiments have shown that α-helices A, B, G and H are folded at an early stage of the protein folding process [60,61]. More specifically, Val108 was identified as a rapidly protected position in H–X experiments [17]. Leu71, Val108 and Phe112 occupy topohydrophobic positions [30] and positions G12-Phe112 and H12-Phe134 belong to the folding nucleus defined by Ptitsyn & Ting [62] after aligning over 700 globin sequences.

3.1.4. Spectrin

The R15, R16 and R17 α-helical repeats (around 100 residues) from chicken brain spectrin (1U4Q) were compared with the R1 and R2 repeats from actinin (1QUU). The spectrin repeat rigidity profile presents three well-defined groups of peaks, reflecting the repeat’s three helix bundle regular structure (Figure 2e). The sequence alignment leads to the identification of a six residues consensus nucleus: Phe1675, Leu1679, Leu1719, Ile1747, Phe1751 and Ile1754 (Figure 2f). d-value analysis studies have shown that Phe1675, Leu1679 and Ile1754 belong to the minimal fast folding core of the R15 repeat [63–65]. Additionally, Ile1747 and Phe1751 lie at the heart of the putative nucleating region in the C-helix [64].

3.2. All β folds

3.2.1. Immunoglobulin

The immunoglobulin fold (around 90 residues) was studied by comparing the sequences and rigidity profiles of fibronectin type III domains from tenascin (1TEN), and chitinase A1 (1K85), the immunoglobulin repeat from titin (1TTT), and the type III module of fibronectin (1TTP). The mechanical alignment leads to a consensus nucleus formed by five residues: Ile821, Ile833, Leu835, Val871 and Leu873 (Figure 2g,h), which are all located in highly conserved positions [31]. Additionally, positions corresponding to Ile821, Leu835 and Leu873 correspond to most interacting residues, that is, amino acid presenting a maximum number of neighbours during the early steps of the folding process, and which are likely to be important for the formation of the protein structure [66,67].
Figure 2. Rigidity profiles (left column, in kcal mol\(^{-1}\) A\(^2\)) and cartoon representations with the consensus nucleus shown in red (right column), for the following protein folds: (a–b) calmodulin, (c–d) myoglobin, (e–f) spectrin, (g–h) immunoglobulin, (i–j) SH3 domain, (k–l) acylphosphatase, (m–n) \(\alpha\)-lactalbumin, (o–p) barnase, (q–r) chymotrypsin inhibitor II, (s–t) protein S6, (u–v) ribonuclease A, (w–x) ribonuclease H and (y–z) ubiquitin. (Online version in colour.)
Figure 2. (Continued.)
Figure 2. (Continued.)
3.3.2. src-SH3 domain

The chicken src-tyrosine kinase SH3 domain (1SRP, 56 residues) was compared with SH3 domains from α-spectrin (1SHG), phosphotransferase fyn (1FYN), actin-binding protein (1J08) and phosphatidylinositol 3-kinase (2PNI). We selected for the consensus nucleus positions occupied by rigid residues in at least three out of the five sequences, namely Phe26, Leu32, Ala45 and Ile66 (figure 2i). All four positions belong to the domain’s hydrophobic core in α-spectrin [72] and fyn [73], and they correspond to folding nucleus residues obtained from Monte Carlo simulations using restraints based on experimental ϕ-values [16,74].

3.3. α/β folds

3.3.1. Acylphosphatase

The alignment and nucleus determination for acylphosphatase (98 residues) was made using four sequences from Equus caballus (1APS), Escherichia coli (2GV1), the hyper-thermophile Sulfobolus solfataricus (1Y90) and Bos taurus (2ACY), leading to the final selection of five rigid residues: Tyr11, Ala26, Val39, Val51 and Leu65 (figure 2k), which all belong to the hydrophobic core, and whose side chains form a tight network binding the α-helices and the β-sheets in the centre of the protein (figure 2l). Experimentally, Tyr11 was shown to be the residue with the largest ϕ-value (ϕ = 0.93, see electronic supplementary material, table S12, for the available experimental ϕ-values for the consensus nucleus residues) [75] along the sequence. Later on, biased MD simulations using these same experimental ϕ-values showed that Tyr11 is one of the three key residues (with Pro54 and Phe94) forming a critical contact network during folding [76], and in the transition state ensemble Tyr11 strongly interacts with Val51 and Leu65 [77]. Interestingly, another key residue for folding [76], Phe94, appears to be highly rigid in three (1APS, 1Y90 and 2ACY) out of the four structures that were studied here (see figure 2k and electronic supplementary material, figure S12e). Finally, evolutionary approaches also highlighted Tyr11, Ala26 and Leu65 as potential folding nucleus members, due to their important conservation within this structural family [31].

3.3.2. α-Lactalbumin

α-Lactalbumin comprises 123 residues. The rigidity and sequence comparison was done on human (1HML), goat (3B80 K) and bovine (1F65) α-lactalbumin and lysozyme from Cattunix japonica (2IHL), E. caballus (2EQL) and canine milk (1E1L), and led to the selection of six consensus nucleus residues: Met30, Leu52, Phe53, Ile55, Ala92 and Trp104 (figure 2m,n). Met30 and Ala92 have already been listed as slow exchanging core members, i.e. potential folding core members, from hydrogen-exchange experiments [17,78]. Met30 role in the protein’s hydrophobic core was also highlighted in mutagenesis experiments [79]. Combined ϕ-value analysis and hydrogen exchange showed that Phe53 and Trp104 belong to a group of strongly interacting residues involved in the protein’s folding initiation site of goat α-lactalbumin [80]. Finally, Leu52, Phe53, Ile55 and Trp104 occupy topohydrophobic positions [30]. Another notable residue is Asp88, which is well conserved (electronic supplementary material, figure S11), corresponds to a local rigidity peak in all six structures (electronic supplementary material, figure S12), but whose force constant remains the rigidity threshold in only three out of six cases. As a consequence, it was not selected for the final consensus nucleus. But it is still interesting to note that this residue is involved in the Ca2+ binding site [81], which is considered to be a folding initiation site for bovine and goat α-lactalbumin, but not in the case of canine milk lysozyme [82].

3.3.3. Barnase

The sequence and mechanical properties from barnase (1BSE, 110 residues) were compared with those from ribonuclease SA (1GMP), ribonuclease T1 (1RNT) and restrictocin (1AQZ), leading to the selection of five rigid residues in conserved position: Ala74, Arg87, Ile88, Val89 and Tyr90 (figure 2o,p). Arg87 excepted, all these residues occupy topohydrophobic positions [30]. In addition, Ile88 lies in the centre of the protein’s hydrophobic core, which also includes Ala74 and Tyr90 [83,84]. H–X experiments highlighted Ala74, Ile88, Val89 and Tyr90 as rapidly protected positions [17,85], while Ile88 and Val89 also present important ϕ-values (0.7 and 0.5, respectively; electronic supplementary material, table S12) [86].

3.3.4. Chymotrypsin inhibitor 2

The CI2 (2CI2, 83 residues) from barley seeds sequence was aligned with a CI2/subtilisin hybrid (1CIS), trypsin inhibitor...
sequences, namely Val6, Ile8, Ile26 and Leu61 (figure 2). All these residues belong to the protein’s hydrophobic core, with $\phi$-value analysis highlighting Ile39 and Leu68 as members of the nucleation site (electronic supplementary material, table SI2) [8]. H–X experiment signals that Ile39, Val66 and Leu68 are slow exchanging NHs [17] and these residues also occupy topohydrophobic positions [30]. Finally, Shakhnovich et al. [29] selected Ile39 and Leu68 as nucleus residues due to their important conservation.

### 3.3.5. Protein S6

Protein S6 (101 residues) sequences and structures from *Thermus thermophilus* (2KJV), *Aquificus aeolicus* (2FA) and *Thermotoga maritima* (1VMB) were compared with the activation domain of human procarboxypeptidase A2 (1O6X) and the N-terminal domain of the U1A protein (1FHT). In addition, the rigidity profile of the P54–55 circular permutant from protein PS6 (2KJW) was also calculated. The sequence alignment leads to a consensus nucleus formed by four residues occupying rigid positions in at least four out of the five aligned sequences, namely Val6, Ile8, Ile26 and Leu61 (figure 2). All these residues belong to the protein hydrophobic core and kinetic studies have shown that they form early contacts during the folding process [87,88]. Furthermore, several $\phi$-value analysis studies highlighted Val6, Ile8 and Ile26 as potential folding nucleus members (electronic supplementary material, table SI2) [16,87,89–92]. Interestingly, the residues forming this consensus nucleus are also rigid in the P54–55 circular permutant (see the alignment in electronic supplementary material, figure S11), which presents an overall structure similar to the original S6 wild-type protein [93] and a conserved folding nucleus [91], despite its different connectivity of secondary structure elements and different folding pathway [94].

### 3.3.6. Ribonuclease A

Bovine ribonuclease A (5RSA, 124 residues) sequence and structure were compared with those from human pancreatic ribonuclease (2K11), ribonuclease P30 (IONC) and angiogenin (1ANG), leading to the following seven residues consensus nucleus: Phe8, His12, Asn44, Thr45, Phe46, Val47 and Phe120 (figure 2). Phe8, Phe46, Val47 and Phe120 belong to the protein’s hydrophobic core [95], in addition, Phe46 and Val47 occupy topohydrophobic positions [30] and were identified as slow exchanging positions in H–X experiments [17].

### 3.3.7. Ribonuclease H

Ribonuclease H from *E. coli* (2RN2, 155 residues) and *T. thermophilus* (1RL) were compared with ribonuclease H1 from *Chlorobium tepidum* (SH08, chain A) and the ribonuclease H domain from HIV-1 reverse transcriptase (1HRH), leading to a five-residue consensus nucleus: Thr9, Ala24, Ala51, Ala52 and Thr69 (figure 2). Ala51, Ala52 and Thr69 belong to $\alpha$-helix A and $\beta$-strand 4, which are the earliest folding elements of the protein [96]. In addition, Ala51 and Ala52 have been highlighted as rapidly protected positions in H–X experiments [17], and Ala52 occupies a topohydrophobic position [30]. Interestingly, Thr9 and Ala24 belong to $\beta$-strands 1 and 2, which are some of the protein last folding segments [96].

### 3.3.8. Ubiquitin

Human ubiquitin (1UBI, 76 residues) was compared with ferredoxin (1FRD), putidaredoxin (1PUT), an IgG-binding domain from protein G (1IGD), the Ras-binding domain from RAF1 (1RFA), the ferredoxin domains from the enterotoxin C2 (1STE) and phthalate dioxygenase reductase (2PFA). Because of the important sequence variability within this large alignment, we selected for the consensus nucleus positions occupied by rigid residues in at least four out of the seven sequences, namely Ile3, Val5, Val26, Ile30, Leu43 and Leu67 (figure 2). Except for Leu43, all these residues occupy topohydrophobic positions [30], correspond to slow exchanging positions in H–X experiments [17] and were already identified as folding nucleus members from evolutionary approaches [20]. In addition, $\phi$-value analysis highlighted the role of residues Ile3 and Val26 (electronic supplementary material, table SI2) [16], while molecular simulations indicate that Ile3, Val5, Ile30 and Leu43 belong to the protein’s folding nucleus [16,97].

### 4. Conclusion

Several experimental and theoretical techniques have been developed to investigate protein folding pathways. Our method combining structural data from the protein native state and evolutionary information leads to the selection of a small subset from the deeply buried residues found in the protein’s core, and which might coincide with the protein’s folding nucleus. More precisely, in this work, the rigidity of proteins from 14 different fold families (including four to seven structures in each family) was investigated using CG BD simulations. The comparison of rigidity profiles for proteins belonging to the same family (which are all shown in electronic supplementary material, figure S12) and multiple sequence alignments show an important conservation of these mechanical properties for proteins with similar folds. In particular, the position of the most rigid residues along the sequence is often conserved, which permits one to identify for each family a subset of highly rigid residues comprising between four and seven residues, which we call the consensus nucleus, as it is common to all members of a given family. These residues are usually located in the protein hydrophobic core, and their inward pointing side chains form a tight interaction network. Interestingly, a survey of earlier literature regarding protein folding shows that the vast majority (over 80%) of these residues has already been identified as folding nucleus members, either via experimental techniques such as $\phi$-value analysis or hydrogen–deuterium exchange (H–X) or via theoretical methods like molecular modelling (with or without restraints based on experimental data) or evolutionary approaches. Regarding $\phi$-values, we can note however that in some specific cases (such as CI2 and ubiquitin; electronic supplementary material, table SI2), our method highlights as potential folding nucleus members, residues with low (less than 0.3) or undetermined $\phi$-values. This concurs with earlier observations made from all-atom computational investigations that the $\phi$-value of a residue does not necessarily translate the residue’s kinetic significance [16,32,98]. In particular, it seems that in the case of proteins folding via diffuse transition states involving a number of partially formed interactions (like CI2 or RNA), folding cores should be assessed by methods other than $\phi$-values [99]. The case of spectrin and its three-helix
bundle is also of special interest. First, because it is not clear whether this protein might be considered as globular, and we do not know if our approach would work on non-globular proteins, since the concept of folding nucleus has initially been proposed for compact globular proteins, so the spectrin fold somehow represents a limit case for this method. Nevertheless, the folding nucleus of the spectrin-R15 repeats has been well documented; experimental studies have shown that repeats R15 on the one hand, and R16 and R17 on the other hand, present different folding mechanisms despite their similar structure. R15 folds by nucleation-condensation, with its secondary and tertiary structures forming simultaneously, while R16 and R17 follow a framework-like mechanism, where the helices form before packing, and these different pathways lead to the determination of different folding nuclei via $\phi$-value analysis [63–65]. Our approach however leads to the determination of a unique consensus nucleus that coincides with the R15 repeat’s folding nucleus. As a consequence, the consensus nucleus resulting from the mechanical/evolutionary method might be more of a structural nucleus, i.e. a group of key residues forming interactions that are essential for the stability of the protein’s fold once the protein has reached its native state, than a folding one. This structural nucleus will often coincide with the folding nucleus, the group of residues forming the first set of native interactions along the folding pathways, but not systematically, and the superposition of these two nuclei in a protein might depend on its folding mechanism. This agrees with the fact that kinetic experiments on the R15 spectrin repeat with a mutated core show a strong destabilization of the protein, which refolds significantly more slowly and unfolds more rapidly than the native R15 [65]. This hypothesis would also explain why, in the case of ribonuclease H, some members of our consensus nucleus are located in late foldon units (namely Thr8, Tyr22 and Ala24 in strands 1 and 2, respectively) [96]. Ribonuclease H folds through a hierarchical mechanism [100,101] that can be compared to the framework mechanism of the R16 and R17 spectrin repeats. Note that Tyr22 and Ala24, however, still belong to the most stable elements of strand 2, as shown by amide hydrogen exchange experiments [102]. Another example is the $\alpha$-lactalbumin fold; while goat $\alpha$-lactalbumin and canine milk lysozyme take different folding pathways, our consensus nucleus coincides with the folding initiation site of goat $\alpha$-lactalbumin (the G8 cluster of strongly interacting residues in [80]).

More generally, since our method somehow ‘averages’ the residues mechanical properties over a whole protein family, by selecting only residues in positions that are rigid for all (or most of) the sequences of the alignment, some of the folding specificities of the different family members are bound to be lost in the process, especially for protein families whose members present various folding pathways. Therefore, further work will be needed to see how the identity between a protein’s structural nucleus (which depends on its fold family) and its particular folding nucleus might be related to its forming mechanism. A possibility would be to compare the contact order within a full protein and its structural and folding nuclei, since earlier work has shown that this specific metric is strongly correlated with the folding rate of single domain proteins [103,104]. Another issue in the case of larger proteins presenting an active site is the discrimination between structural and functional residues (such as catalytic site residues), since several modelling approaches (including ProPhet) have already shown that the latter are, on average, also more rigid than the rest of the protein [38,105]. An analysis of these residue physico-chemical properties based on large datasets of catalytic residues on the one hand, and structural nucleus residues on the other hand, might be useful to predict whether a residue is important for function or structure.

What was previously suggested in the individual cases of cytochrome c and globins appears to be a general feature of globular proteins independently of their fold. In addition to occupying conserved positions along the protein sequence, ProPhet calculations based on protein native structures show that for most proteins, folding nucleus residues form a rigid network in the protein core, and that their mechanical properties are also conserved within a fold family. While non-native contacts can influence folding dynamics [106,107], native contacts still play a key role for the determination of the folding mechanism [23]. As a consequence, the native state structure encodes information regarding the protein folding process, as was shown by our results. Despite the simplicity of the ENM used by ProPhet, rigidity profiles calculations appear to be an efficient tool for the determination of the structural nucleus, thus bringing useful information regarding the protein folding process and the potential impact of point mutations on protein structural stability.

**Competing interests.** I declare I have no competing interests.

**Funding.** I received no funding for this study.

**Acknowledgements.** S.S.M. wishes to thank Jacques Chomilier for the inspiring scientific discussion that led to this work.

**References**


