INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **17** (2005) S1515–S1522 doi:10.1088/0953-8984/17/18/009

What determines the structures of native folds of proteins?

Antonio Trovato[1], Trinh X Hoang[2], Jayanth R Banavar[3], Amos Maritan[1] and Flavio Seno[1]

[1] INFM and Dipartimento di Fisica 'G Galilei', Università di Padova, Via Marzolo 8, 35131 Padova, Italy
[2] Institute of Physics, NCST, 46 Nguyen Van Ngoc, Hanoi, Vietnam
[3] Department of Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received 30 September 2004, in final form 1 October 2004
Published 22 April 2005
Online at stacks.iop.org/JPhysCM/17/S1515

Abstract

We review a simple physical model (Hoang *et al* 2004 *Proc. Natl Acad. Sci. USA* **101** 7960, Banavar *et al* 2004 *Phys. Rev.* E at press) which captures the essential physico-chemical ingredients that determine protein structure, such as the inherent anisotropy of a chain molecule, the geometrical and energetic constraints placed by hydrogen bonds, sterics, and hydrophobicity. Within this framework, marginally compact conformations resembling the native state folds of proteins emerge as competing minima in the free energy landscape. Here we demonstrate that a hydrophobic-polar (HP) sequence composed of regularly repeated patterns has as its ground state a β-helical structure remarkably similar to a known architecture in the Protein Data Bank.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Proteins are well-tailored chain molecules employed by life to store and replicate information, to carry out a dizzying array of functionalities and to provide a molecular basis for natural selection. A protein molecule is a large and complex physical system with many atoms. In addition, the water molecules surrounding the protein play a crucial role in its behaviour. At the microscopic level, the laws of quantum mechanics can be used to deduce the interactions but the number of degrees of freedom are far too many for the system to be studied in all its detail. When one attempts to look at the problem in a coarse-grained manner [3] with what one hopes are the essential degrees of freedom, it is very hard to determine what the effective potential energies of interaction are. This situation makes the protein problem particularly daunting and no solution has yet been found. Nevertheless, proteins fold into a limited number [4, 5] of evolutionarily conserved structures [6, 7]. The same fold is able to house many different

sequences which have that conformation as their native state and is also employed by nature to perform different biological functions, pointing towards the existence of an underlying simplicity and of a limited number of key principles at work in proteins.

We recently showed that a simple model which encapsulates a few general attributes common to all polypeptide chains, such as the anisotropy implicit in a chain molecules [8, 9], steric constraints [10–12], hydrogen bonding [13–15] and hydrophobicity [16], gives rise to the emergent free energy landscape of globular proteins [1]. The relatively few minima in the resulting landscape correspond to distinct putative marginally-compact native-state structures of proteins, which are tertiary assemblies of helices, hairpins and planar sheets. A superior fit [17, 18] of a given protein or sequence of amino acids to one of these predetermined folds dictates the choice of the topology of its native-state structure. Instead of each sequence shaping its own free energy landscape, we find that the overarching principles of geometry and symmetry determine the menu of possible folds that the sequence can choose from.

Sequence design would favour the appropriate native state structure over the other putative ground states leading to an energy landscape conducive for rapid and reproducible folding of that particular protein. Nature has a choice of 20 amino acids for the design of protein sequences. A pre-sculpted landscape greatly facilitates the design process. We have shown elsewhere [19] that, within our model, a crude design scheme, which takes into account the hydrophobic (propensity to be buried, H) and polar (desire to be exposed to the water, P) character of the amino acids, is sufficient to carry out a successful design of sequences with one of the structures shown in figure 1. Here we will show that within the same HP-scheme, a longer (45 residues) sequence, composed of a regular repetition of the same hydrophobicity pattern (PHPHP), has its ground state in an extremely regular β-helix structure, stabilized by the formation of a hydrophobic core characterized by a high degree of geometric regularity, as is the case for a broad class of proteins known as repeat proteins.

2. Model and methods

We model a protein as a chain of *identical* amino acids, represented by their C^{α} atoms, lying along the axis of a self-avoiding flexible tube. The preferential parallel placement of nearby tube segments approximately mimics the effects of the anisotropic interaction of hydrogen bonds, while the space needed for the clash-free packing of side chains is approximately captured by the non-zero tube thickness [8, 9]. A tube description places constraints on the radii of circles drawn through both local and non-local triplets of C^{α} positions of a protein native structure [9, 20].

Unlike unconstrained matter for which pairwise interactions suffice, for a chain molecule, it is necessary to define the context of the object that is part of the chain. This is most easily carried out by defining a local Cartesian coordinate system whose three axes are defined by the tangent to the chain at that point, the principal normal, and the binormal which is perpendicular to both the other two vectors. A study [1, 2] of the experimentally determined native state structures of proteins from the Protein Data Bank reveals that there are clear amino acid aspecific geometrical constraints on the relative orientation of the local coordinate systems due to sterics and also associated with amino acids which form hydrogen bonds with each other. It is interesting to note that similar geometrical constraints had already been introduced in off-lattice polymer models [21, 22] in order to model hydrogen bond formation.

The geometrical constraints associated with the formation of hydrogen bonds and with the tube description within the C^{α} representation of our model are described in detail elsewhere [1]. In our representation of the protein backbone, local hydrogen bonds form between C^{α} atoms separated by three along the sequence with an energy defined to be -1 unit, whereas non-local

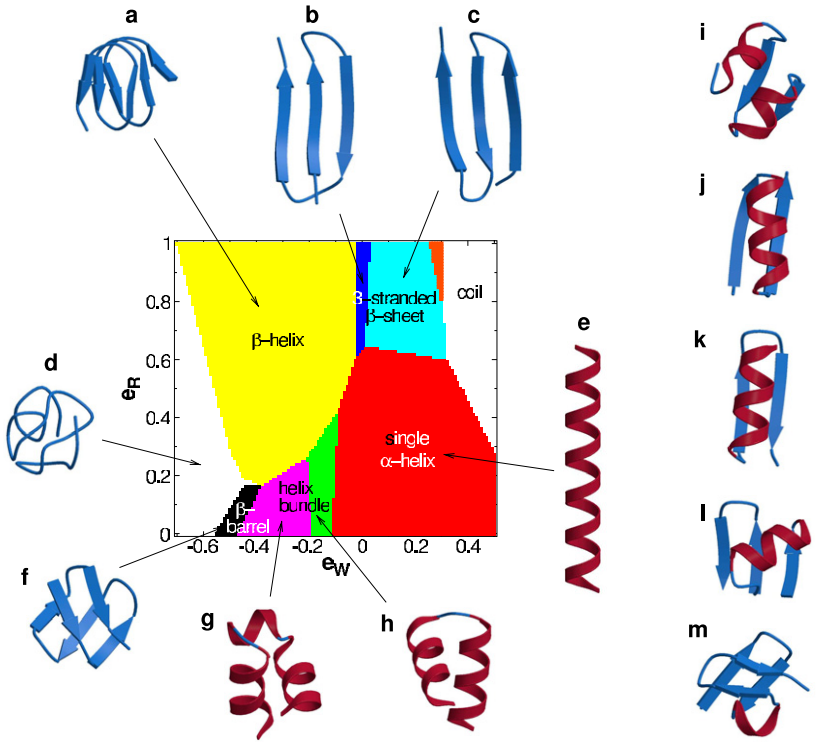


Figure 1. Phase diagram of ground state conformations. The ground state conformations were obtained by means of Monte Carlo simulations of chains of 24 C^{α} atoms. e_R and e_W denote the bending energy penalty and the solvent mediated interaction energy respectively. Over 600 distinct local minima were obtained in our simulations in different parts of parameter space starting from a randomly generated initial conformation. The temperature is set initially at a high value and then decreased gradually to zero. (a)–(c), (e)–(h) are the Molscript representation of the ground state conformations which are found in different parts of the parameter space as indicated by the arrows. The helices and strands are assigned when local or non-local hydrogen bonds are formed according to the rules employed within our model [1]. Conformations (i)–(m) are competitive local minima. In the non-labelled dark grey phase on the top-left of the phase diagram (orange online), the ground state is a two-stranded β-hairpin (not shown). Two distinct topologies of a three-stranded β-sheet are found corresponding to conformations shown in conformations (b) and (c) respectively (dark and light blue online). The white region in the left of the phase diagram has large attractive values of e_W and the ground state conformations are compact globular structures with a crystalline order induced by hard sphere packing considerations [31] and not by hydrogen bonding (conformation (d)).

hydrogen bonds are those that form between C^{α} atoms separated by more than four along the sequence with an energy of -0.7. This energy difference is based on experimental findings that the local bonds provide more stability to a protein than do the non-local hydrogen bonds [23]. Cooperativity effects [24, 25] are taken into account by adding an energy of -0.3 units when consecutive hydrogen bonds along the sequence are formed. There are two other ingredients in the model: a local bending penalty e_R which is related to the steric hindrance of the amino acid side chains and a pair-wise interaction e_W of the standard type mediated by the water [16]. Note that whereas the geometrical constraints associated with the tube and hydrogen bonds are representative of the typical *aspecific* behaviour of the interacting amino acids, the latter properties clearly depend on the *specific* amino acids involved in the interaction.

Constant temperature Monte Carlo simulations have been carried out with pivot and crankshaft moves commonly used in stochastic chain dynamics [26]. A Metropolis procedure

is employed with a thermal weight $\exp(-E/T)$, where E is the energy of the conformation and T is the effective temperature.

3. Results and discussion

Figure 1 shows the ground state phase diagram obtained from Monte Carlo computer simulations using the simulated annealing technique [27], along with the corresponding conformations, for a 24-bead homopolymer [1]. The solvent-mediated energy, e_{W}, and the local bending penalty, e_{R}, are measured in units of the local hydrogen bond energy. When e_{W} is sufficiently repulsive (hydrophilic) (and $e_{\mathrm{R}} > 0.3$ in the phase diagram), one obtains a swollen phase with very few contacts between the C^{α} atoms. When e_{W} is sufficiently attractive, one finds a very compact, globular phase with featureless ground states with a high number of contacts.

Between these two phases (and in the vicinity of the swollen phase), a marginally compact phase emerges (the interactions barely stabilize the ordered phase) with distinct structures including a single helix, a bundle of two helices, a helix formed by β-strands, a β-hairpin, three-stranded β-sheets with two distinct topologies and a β-barrel-like conformation. These structures are the stable ground states in different parts of the phase diagram. Furthermore, other conformations, closely resembling distinct super-secondary arrangements observed in proteins [4], also shown in figure 1, are found to be competitive local minima, whose stability can be enhanced [1, 19] by sequence design after heterogeneity is introduced by means of, for example, non-uniform values of curvature energy penalties for single amino acids and hydrophobic interactions for amino acid pairs. Note that while there is a remarkable similarity between the structures that we obtain and protein folds, our simplified coarse-grained model is not as accurate as an all-atom representation of the polypeptide chain in capturing features such as the packing of amino acid side chains. The lack of detailed side-chain structure causes the conformations depicted in figure 1 to be more compact than real protein native folds.

The common belief in the field of proteins is that given a sequence of amino acids, with all the attendant details of the side chains and the surrounding water, one obtains a funnel-like landscape with the minimum corresponding to its native state structure. Each protein is characterized by its own landscape. In this scenario, the protein sequence is all-important and the protein folding problem, besides becoming tremendously complex, needs to be attacked on a protein-by-protein basis.

In contrast, our model calculations show that the large number of common attributes of globular proteins [9, 28] reflect a deeper underlying unity in their behaviour. At odds with conventional belief, the gross features of the energy landscape of proteins result from the amino acid aspecific common features of all proteins, as is clearly established by the fact that different putative native structures are found to be competing minima for the same homopolymer chain.

The protein energy landscape is *(pre)sculpted* by general considerations of geometry and symmetry (figure 2), which we have utilized as ingredients in our model. Our unified framework suggests that the protein energy landscape ought to have around a thousand broad minima corresponding to putative native state structures. The key point is that for each of these minima the desirable funnel-like behaviour is already achieved at the homopolymer level *in the marginally compact part of the phase diagram*. The self-tuning of two key length scales, the thickness of the tube and the interaction range, to be comparable to each other and the interplay of the three energy scales, hydrophobic, hydrogen bond, and bending energy, in such a way as to stabilize marginally compact structures, also provide the close cooperation between energy gain and entropy loss needed for the sculpting of a funnelled energy landscape. At the same time, relatively small changes in the parameters e_{W} and e_{R} lead to significant differences

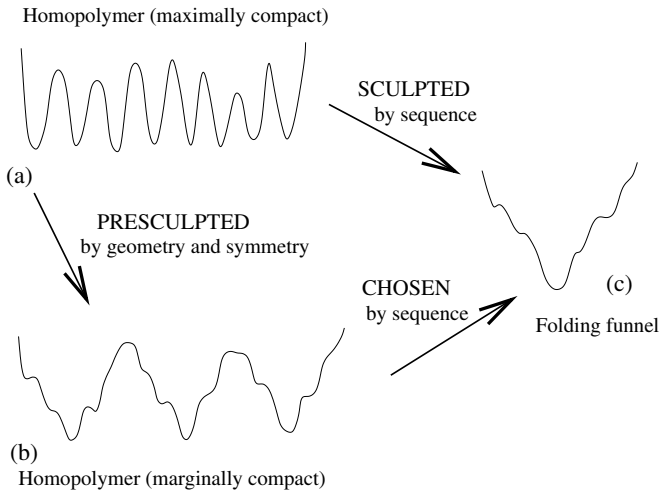
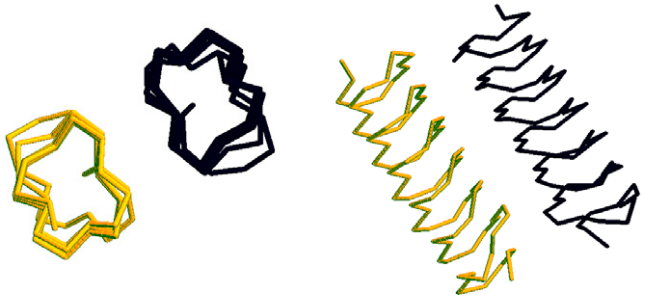


Figure 2. Simplified one-dimensional sketches of energy landscapes. The quantity plotted on the horizontal axis schematically represents a distance between different conformations in the phase space and the barriers in the plots indicate the energy needed by the chain in order to travel between two neighbouring local minima. (a) Rugged energy landscape for a homopolymer chain with an attractive potential promoting compaction as, for example, in a string and beads model. There are many distinct maximally compact ground state conformations with roughly the same energy, separated by high energy barriers (the degeneracy of ground state energies would be exact in the case of both lattice models and off-lattice models with discontinuous square-well potentials). (b) Presculpted energy landscape for a homopolymer chain in the marginally compact phase. The number of minima is greatly reduced and the width of their basin increased by the introduction of geometrical constraints. (c) Funnel energy landscape for a protein sequence. As folding proceeds from the top to the bottom of the funnel, its width, a measure of the entropy of the chain, decreases cooperatively with the energy gain. Such a distinctive feature, crucial for fast and reproducible folding, arises from careful sequence design in models whose homopolymer energy landscape is similar to (a). In contrast, funnel-like properties already result from considerations of geometry and symmetry in the marginally compact phase (b), thereby making the goals of the design procedure the relatively easy task of stabilization of one of the pre-sculpted funnels followed by the more refined task of fine-tuning the putative interactions of the protein with other proteins and ligands.

in the emergent ground state structure, underscoring the sensitive role played by chemical heterogeneity in selecting from the menu of native state folds.

A design scheme able to select and stabilize one of such folds is straightforward within our model. The primitive scheme of introducing sequence heterogeneity at the level of differentiating hydrophobic and polar residues and designing the hydrophobicity profile of the sequence by hand performs adequately. For example, the β–α–β motif shown as (j) in figure 1 (which is a local energy minimum for a homopolymer) can be stabilized into a global energy minimum for the sequence HPHHHPPPPHHPPHHPPPPHHHPP, with $e_{\mathrm{W}} = -0.4$ for HH contacts and $e_{\mathrm{W}} = 0$ for other contacts, and $e_{\mathrm{R}} = 0.3$ for all residues [19].

The same ground state stabilization can be achieved for a longer chain, composed of 45 beads, a size already comparable to that of small proteins, within the same simplified HP scheme. We will show this, by investigating at the same time the capability of our model to reproduce the basic sequence–structure relationship underlying a particular class of proteins, namely *repeat proteins*. An interesting class of naturally occurring proteins contain homologous segments which repeat [29]. In the absence of an interaction between these repeat units, the segments remain unfolded. However, in the folded state, they utilize their near



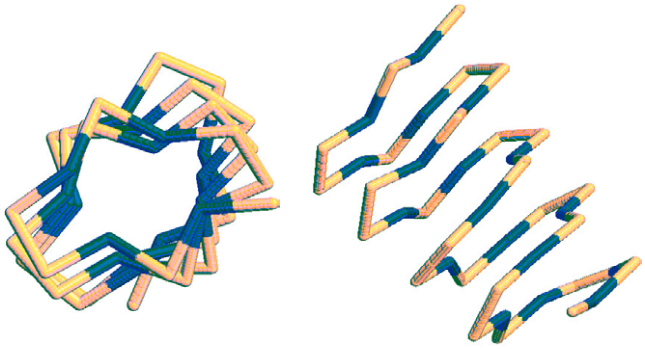
(a) (b)

Figure 3. Backbone trace representation of the antifreeze protein 1EGZ from the beetle, viewed from the top (a) and from the side (b). The two different chains in the complex are shown in different shades of grey (different colours online). The sequence of the two chains is identical and the repetition of the following regular 12-residue pattern can be observed: CYS, THR, X, SER, X, X, CYS, Y, X, ALA, X, THR, where X is almost always a polar (the exception is always glycine) or charged residue and Y is usually a hydrophobic residue (X does not need to be the same residue in all places).

similarity to stack up and create a stable hydrophobic core. The resulting stacks are commonly elongated structures, and circular shapes resembling a propeller are less often observed [30]. Nature exploits the modular (and somewhat flexible) structures adopted by repeat proteins to bind molecules in a variety of situations and organisms.

An example of a repeat protein elongated architecture is shown in figure 3, where the same conformation (Protein Data Bank code 1EZG) is shown viewed from the top and from a side. The two chains forming the complex (which is an antifreeze protein from the beetle) share the same β-helical architecture known as *3 solenoid*. There is a clear, albeit not exact, repeating pattern of 12 residues in the sequence, marked by the presence of cysteines (see the caption of figure 3 for details). Note that a very similar topology was obtained as one of the ground states for the 24-bead homopolymer (see figure 1 conformation (a)). The same β-helical topology is instead much harder to recover, even as a local energy minimum, for a 45-bead homopolymer, consistent with the fact that the stabilization of such an elongated architecture needs the introduction of a repeat sequence pattern.

Remarkably, the β-helical architecture can indeed be stabilized as a ground state in our minimal model, using a five-bead repeat sequence PHPHP (repeated nine times to yield a total length of 45 beads), within the same simplified HP scheme as above ($e_{\mathrm{W}} = -0.4$ for HH contacts and $e_{\mathrm{W}} = 0$ for other contacts, but now $e_{\mathrm{R}} = 0.2$ for all residues). The resulting minimum energy conformation, having an energy of -49.5 (recall that in our model energy is defined in units of local hydrogen bond energy), is shown in figure 4. Hydrophobic residues are located in a very tight and hydrophobic core, which evidently contributes the most to the stability of the conformation, which is itself strikingly similar to the real 3 solenoid fold (figure 3). We remark, though, that structural similarity at a finer level than overall fold architecture is not a goal of our present approach, given the lack of details such as side chain atomic structure in our model. For the same reason, one cannot expect that the stabilization of the real 1EZG conformation is explained only on the basis of the hydrophobic-polar scheme employed by us. Indeed, the regular pattern of cysteines and the presence of several charged and polar residues in the 1EZG sequence (see figure 3 caption), strongly suggest that disulfide bonds and hydrogen bonds between polar side chain and backbone groups are quite important in the overall energy balance, being most likely involved in the stabilization of turns. Even



(a) (b)

Figure 4. Ground state conformation, with energy -49.5 for the sequence obtained by repeating the PHPHP pattern nine times, viewed from the top (a) and from the side (b). The energy parameters are $e_{\mathrm{W}} = -0.4$ for HH interactions, $e_{\mathrm{W}} = 0$ for other interactions, and $e_{\mathrm{R}} = 0.2$ for all residues. Hydrophobic residues are shown as darker (blue online) than polar ones (yellow online).

though our model cannot evidently capture such details, it nevertheless reproduces correctly the basic qualitative fact that a sequence consisting of a repeating pattern stabilizes a modular elongated architecture, via the formation of a tight hydrophobic core.

4. Conclusion

In summary, within a simple, yet realistic, framework, we demonstrate [1] that protein native-state structures can arise from considerations of symmetry and geometry associated with the polypeptide chain. The aggregation of different polypeptide chains in cross-β structures which resemble amyloid fibrils is also a by-product of the same principles [2]. The sculpting of the free energy landscape with relatively few broad minima is consistent with the fact that proteins can be designed to enable rapid folding to their native states, while avoiding aggregation. We have shown [19] that the introduction of heterogeneity within the simplest hydrophobic-polar scheme is sufficient to design a sequence that is able to fold cooperatively into one of the presculpted minima in the energy landscape. Here we show that within the same scheme a sequence consisting of the same simple pattern repeated several times has its global energy minimum in a β-helical architecture characterized by the formation of a tight hydrophobic core, mimicking the generic behaviour of repeat proteins.

Acknowledgments

This work was supported by PRIN 2003, FISR 2001, NASA, NSF IGERT grant DGE-9987589, NSF MRSEC and VNSC.

References

[1] Hoang T X, Trovato A, Seno F, Banavar J R and Maritan A 2004 *Proc. Natl Acad. Sci. USA* **101** 7960
[2] Banavar J R, Hoang T X, Maritan A, Seno F and Trovato A 2004 *Phys. Rev.* E **70** 041905
[3] Banavar J R and Maritan A 2001 *Proteins* **42** 433
[4] Chothia C and Finkelstein A V 1990 *Annu. Rev. Biochem.* **59** 1007
[5] Chothia C 1992 *Nature* **357** 543

[6] Denton M and Marshall C 2001 *Nature* **410** 417
[7] Chothia C, Gough J, Vogel C and Teichmann S A 2003 *Science* **300** 1701
[8] Maritan A, Micheletti C, Trovato A and Banavar J R 2000 *Nature* **406** 287
[9] Banavar J R and Maritan A 2003 *Rev. Mod. Phys.* **75** 23
[10] Ramachandran G N and Sasisekharan V 1968 *Adv. Protein Chem.* **23** 283
[11] Pappu R V, Srinivasan R and Rose G D 2000 *Proc. Natl Acad. Sci. USA* **97** 12565
[12] Baldwin R L and Rose G D 1999 *Trends Biochem. Sci.* **24** 26
[13] Pauling L, Corey R B and Branson H R 1951 *Proc. Natl Acad. Sci. USA* **37** 205
[14] Pauling L and Corey R B 1951 *Proc. Natl Acad. Sci. USA* **37** 729
[15] Eisenberg D 2003 *Proc. Natl Acad. Sci. USA* **100** 11207
[16] Kauzmann W 1959 *Adv. Protein Chem.* **14** 1
[17] Bryngelson J D and Wolynes P G 1987 *Proc. Natl Acad. Sci. USA* **84** 7524
[18] Brenner S A 2001 *Nature* **409** 459
[19] Hoang T X, Trovato A, Seno F, Banavar J R and Maritan A 2004 *Biophys. Chem.* submitted
[20] Banavar J R, Maritan A, Micheletti C and Trovato A 2002 *Proteins* **47** 315
[21] Kemp J P and Chen Z Y 1998 *Phys. Rev. Lett.* **81** 3880
[22] Trovato A, Ferkinghoff-Borg J and Jensen M H 2003 *Phys. Rev.* E **67** 021805
[23] Shi Z, Krantz B A, Kallenbach N and Sosnick T R 2002 *Biochemistry* **41** 2120
[24] Liwo A, Kazmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Rackovski R J, Pincus M R and Scheraga H A 1998 *J. Comput. Chem.* **19** 259
[25] Fain B and Levitt M 2003 *Proc. Natl Acad. Sci. USA* **100** 10700
[26] Sokal A D 1996 *Nucl. Phys.* B **47** (Suppl.) 172
[27] Kirkpatrick S, Gelatt C D Jr and Vecchi M P 1983 *Science* **220** 671
[28] Bernal J D 1939 *Nature* **143** 663
[29] Forrer P, Binz K H, Stumpp M T and Pluckthun A 2004 *Chem. Bio. Chem.* **5** 183
[30] Paoli M 2001 *Prog. Biophys. Mol. Bio.* **76** 103
[31] Zhou Y, Hall C K and Karplus M 1996 *Phys. Rev. Lett.* **77** 2822