Prediction of a novel topology in the N-terminal, 14 kDa fragment of Ada protein

Mark B. Swindells

Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan

Previously determined protein structures have been analysed, in order to find folding motifs similar to that proposed by NMR spectroscopy, for the N-terminal, 14 kDa fragment of the Ada protein. The analyses reveal only limited similarities with the NMR-derived structural data and strongly suggest that this region of the Ada protein adopts a previously unobserved topology. Characteristic structural features, which arise from the inferred chain connectivity, are examined through comparisons with other structures. Using this information, the topology of the Ada protein 14 kDa fragment has been predicted in order to provide structural data not yet attainable from NMR experiments.

Ada protein; Topology; Analysis; Prediction

1. INTRODUCTION

The Ada protein, which is found in *Escherichia coli*, plays a central role in removing alkylated nucleotides from damaged DNA duplexes [1,2]. This protein has previously been shown to have two functionalities. The first is as a methyl transferase, in which a methyl group is transferred from the DNA duplex to itself, whereas the second is as a transcriptional regulator [3–6]. In order to investigate these processes at the molecular level, NMR analyses of 14 kDa, 16 kDa and 20 kDa N-terminal fragments have been initiated by Sakashita et al. [17]. The results of these analyses are presented in an accompanying paper, but in summary, the assignments are limited to the N-terminal, 14 kDa fragment (from now referred to as N-ada 14k) and these identify a fold which is based on a mixed β-sheet consisting of four β-strands (Fig. 1). As the complete three dimensional structure of N-ada 14k is currently unknown, and given that proteins with no sequence similarities frequently adopt familiar topologies [7,8], information relevant to the Ada topology might be inferred by comparing the current NMR data with topologies which are already available in the Protein Data Bank [9]. The following paper uses the NMR results of Sakashita et al. [17] to search for other proteins with complementary topologies and propose a simple model for the N-ada 14k topology.

2. MATERIALS AND METHODS

The NMR data used in this work were kindly made available by the

authors of the accompanying paper (Sakashita et al. [17]) prior to publication. Structural analyses concentrated on detecting similarities with 112 unique topologies observed in the Protein Data Bank (list provided by C. Orengo, personal communication). Two comprehensive publications on protein structure were also used to assist these analyses [10,11]. The PDB codes of proteins specifically referred to in this paper are: ribonuclease H (2rn2), ubiquitin (1ubq), histidine-containing phosphocarrier protein (1hid), ovomucoid third domain (1ovo) and rubredoxin (5rxn). Secondary structure assignments were made using the method of Kabsch and Sander [12]. Ribbon diagrams were made using Molscript [13].

In the following work, fold and topology are used to represent different aspects of protein structure. Fold simply describes a spatial arrangement of regular secondary structural elements, whereas topology also includes the connectivity of these elements. By using this definition, proteins with the same topology must have the same fold. This relationship is explained in more detail in Fig. 2.

Descriptions of strand connectivity follow the procedure of Richardson [11]. Details of this procedure can be found in the appropriate reference, but a summary is given here for clarity. A sequential connection between two β-strands will always fall into one of two simple categories. In the first case the main-chain leaves and re-enters the sheet at the same end. This is known as a hairpin connection. If the main-chain crosses the sheet in order to re-enter at the opposite side this is known as a crossover connection and is represented by an 'x'. Thus in Fig. 2b, the four strands have a −1,+2x,−3x connectivity whereas in Fig. 2c it is +1,+1,−3x. Numerals represent the number of strands traversed, whereas the '+' and '−' signs indicate the relative (not absolute) directions of sheet traversal. (In this manner, −1,+2x,−3x and +1,−2x,+3x describe the same connectivity.)

Crossover connections can be either right-handed or left-handed. The difference between these connections is highlighted in Fig. 3. There is a strong preference for right-handed crossover connections [14] and in the absence of evidence to the contrary, it is reasonable to assume that a connection will be right-handed.

3. RESULTS

A summary of the NMR-derived structural data is shown in Fig. 1. Although an α-helix is assigned be-

Correspondence address: M.B. Swindells, Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan. Fax: (81) (6) 872 8210.
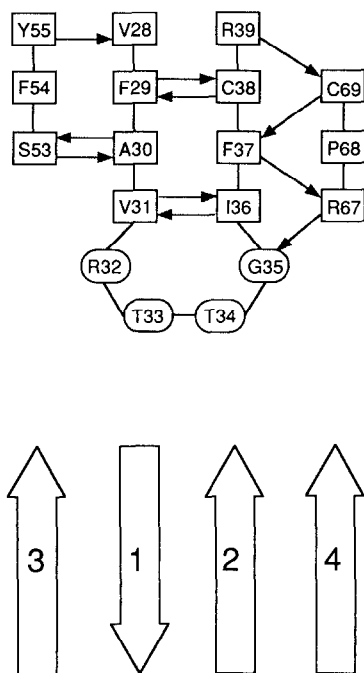
Fig. 1. Summary of NMR-derived structural data for N-ada 14k between residues 27 and 70. The diagram shows the residues involved in a mixed four-stranded β-sheet. The first two strands are located in the centre of the sheet, where they form a β-hairpin. An α-helix (not shown) is also detected in the loop connecting strands three and four. Arrows indicate main-chain hydrogen bonds which are inferred from the NMR data.
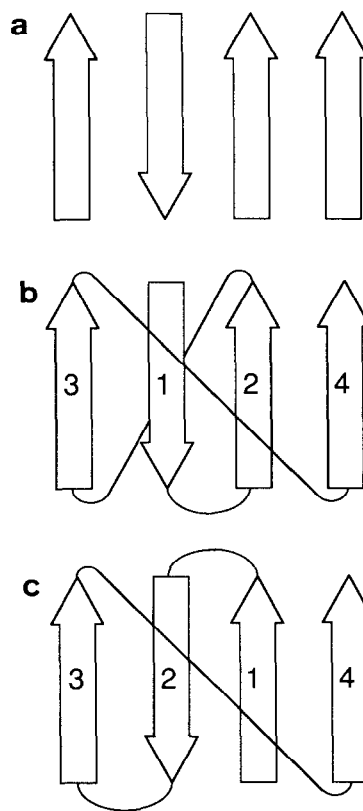

Fig. 2. Representation of the different meanings for the terms fold and topology (a) Represents the fold observed in both N-ada 14k and the N-terminal region of RNase H. (b) Depicts the topology predicted for N-ada 14k, whereas (c) portrays the topology observed in RNase H. Using this definition, one fold may give rise to many topologies.

tween residues 9 and 15, there are no long distance constraints available for this region. As a result, the region for which topology might be inferred, occurs between residues 27 and 70. In this region a mixed four stranded β-sheet is detected together with a short α-helix in the loop connecting the edge strands. The first two strands are located in the centre of the sheet, where they form a β-hairpin. Other β-strand connectivities may be inferred from the NMR data. As the preference for right-handed connections is very strong [14] it is also reasonable to assume that the N-ada 14k crossover connections are right-handed. This assumption complements the proposed connectivity of N-ada 14k, because it will allow the two crossover connections to be situated on opposite faces of the β-sheet, thus avoiding steric clashes.

Using the nomenclature of Richardson, the connectivity of the β-sheet in N-ada 14k may be assigned as −1,+2,−3x (or +1,−2x,+3x). Surprisingly, in the 112 topologies surveyed, there are no other examples of proteins which adopt this entire connectivity. In fact, there are only two β-sheet topologies in which a −3x connection also includes an α-helix. Analyses of ubiquitin and the first seventy residues of ribonuclease H (RNase H), which contain this −3x connectivity, reveal that only RNase H has an arrangement of parallel and antiparallel β-strands (i.e. a fold) which is compatible

with the mixed β-sheet of N-ada 14k (Figs. 2 and 4). Although a −3x connection containing an α-helix also occurs in the oligonucleotide binding fold [15], it exists as part of a five-stranded β-barrel. As neither a fifth strand nor a β-barrel are apparent from the NMR data for N-ada 14k, this is an unlikely candidate.

Although N-ada 14k and RNase H both have similar arrangements of parallel and antiparallel β-strands (i.e. their folds are similar) their topologies are different. In fact the N-ada 14k, −1,+2x connectivity is only observed in a small fraction of the 112 proteins surveyed. Structures containing this motif include the ovomucoid third domain, rubredoxin and histidine-containing phosphocarrier protein (HPr). In all three cases the crossover connections are right-handed, although the loops connecting strands two and three are larger than proposed for N-ada 14k and contain short elements of regular secondary structure.

With similarities in the strand connectivities analysed, a more detailed investigation of the three connecting loops can be implemented. The first loop in N-ada 14k connects strands 1 and 2, thereby generating a β-hairpin structure. A comparison of the NMR data for this region with previous analyses of β-hairpins [16]
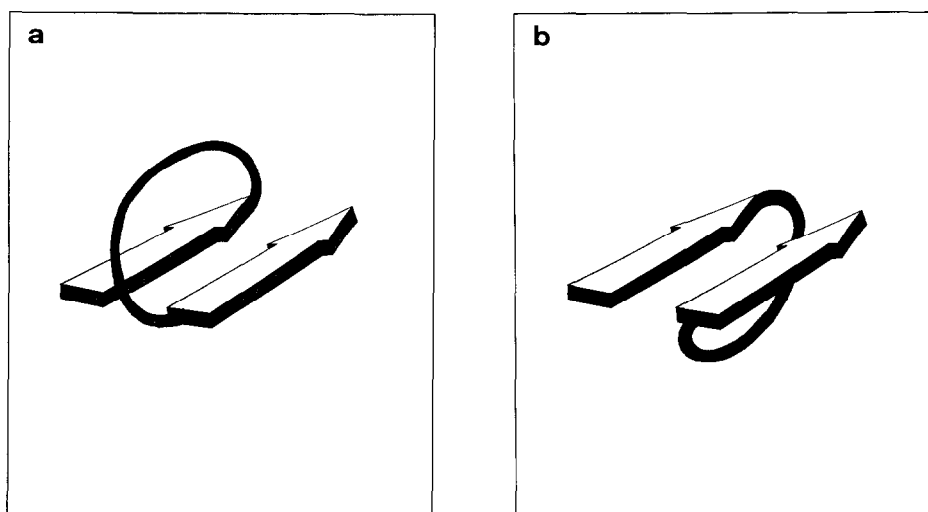
Fig. 3. Representations of (a) right-handed and (b) left-handed crossover connections in two parallel β strands.

reveals that this loop is consistent with a type 4:4 β-hairpin structure. It is also likely that the loop will adopt a type I β-turn, as 4:4 β-hairpins and the sequence of this N-ada 14k loop (Arg-Thr-Thr-Gly) both prefer this conformation.

The second loop of N-ada 14k consists of thirteen residues and connects the second and third strands with an unusual $-2x$ connectivity which has already been described. Due to the limited structural data, detailed predictions are not possible. However, the loop will probably adopt a right handed connection between the two strands.

Loop three, which performs the $+3x$ crossover connection, consists of ten residues in the Ada protein, of which approximately six will be in a helix. This is significantly different to the equivalent $+3x$ connections in both RNase H and ubiquitin as they contain larger helices within loops which are 25–30 residues long. Analyses of RNase H and ubiquitin reveal that in both cases the helix lies at an angle across the β-sheet. However, in neither structure does the helix orient in a manner which would enable a short loop connection between the two edge strands. In fact, the direction of the helix is opposite to the direction required by the chain topology. Consequently, long loops are required at both termini in order to complete the topological connection (Fig. 4). As the N-ada 14k connection consists of only ten residues, of which at least six form an α-helix, the orientation found in RNase H and ubiquitin is not possible. In fact, visual inspection suggests that the loop is only just long enough to complete such a connection. The presence of a glycine residue at position 65 will probably assist these connections by occupying an unusual region of ϕ,ψ space. In addition, a larger twist in the β-sheet (similar to that observed in ubiquitin) may assist the connection, but this cannot be predicted from the information available.

Even though there are no structures in the Protein Data Bank which agree entirely with the proposed topology of N-ada 14k, a low-level model of the N-ada 14k topology can be predicted using the information reported above. Firstly, we know that strands in the N-terminal region of RNase H and N-ada 14k adopt similar folds (but different topologies). Onto this the approximate orientations of the loops and helices can
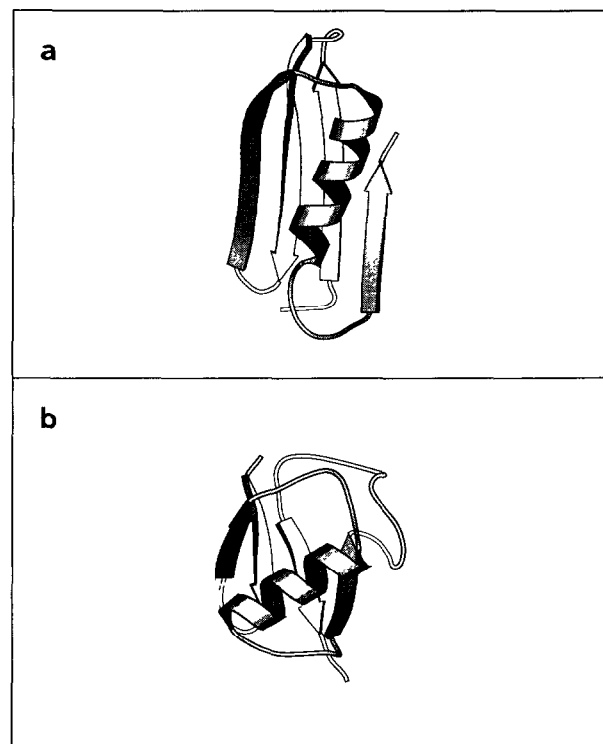

Fig. 4. Topology diagrams of (a) the N-terminal 70 residues of RNase H and (b) ubiquitin. The β-α-β units which take part in the $+3x$ connections are emphasised by shading.
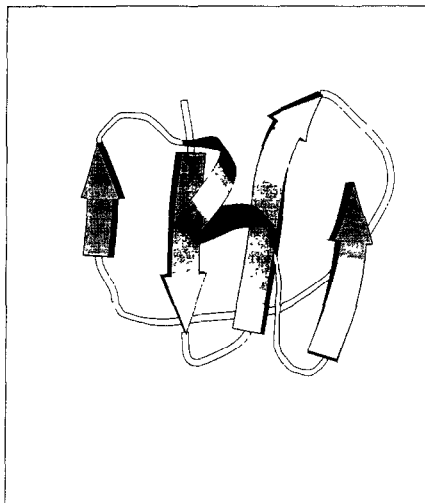
Fig. 5. Prediction of the N-ada 14k topology. In this orientation, the helix between strands three and four is lying on top of the sheet, whereas the loop connecting the second and third strands is below.

be predicted based on the assumption that the crossover connections will be right-handed. A cartoon of the predicted N-ada 14k topology is shown in Fig. 5. It should be emphasized that this drawing is designed only to illustrate the right-handed crossover loop connections and the orientation of the α-helix. Specific details regarding loop orientations are unpredictable from current data and their allocated positions have no particular significance in this figure.

4. DISCUSSION

In this paper, knowledge-based techniques have been described which maximise the information obtained from preliminary NMR data. These investigations were enabled by the NMR assignments for β-strands, as they automatically provide information about the tertiary structure. Equivalent analyses would not have been possible had the structure been α-helical because side-chain

interaction data would also be required. This approach facilitated a prediction of the N-ada 14k topology which is consistent with current NMR data and compatible with our present knowledge of protein structures. As the number of protein structures increases, the methods described in this paper will become increasingly useful for prediction. Furthermore, these results provide hope that similar approaches will assist future structural determinations.

REFERENCES

[1] Lindahl, T., Sedgwick, B., Sekiguchi, M. and Nakabbepu, Y. (1988) Annu. Rev. Biochem. 57, 133–157.
[2] Volkert, M.R. (1988) Environ. Mol. Mutagen. 11, 241–255.
[3] Teo, I., Sedgwick, B., Demple, B., Li, B. and Lindahl, T. (1984) EMBO J. 3, 2151–2157.
[4] Nakabeppu, Y., Kondo, H., Kawabata, S., Iwanaga, S. and Sekiguchi, M. (1984) J. Biol. Chem. 259, 13723–13729.
[5] Nakabeppu, Y. and Sekiguchi, M. (1986) Proc. Natl. Acad. Sci. USA 83, 6297–6301.
[6] Teo, I., Sedgwick, B., Kilpatrick, M.W., McCarthy, T.V. and Lindahl, T. (1986) Cell 45, 315–324.
[7] Swindells, M.B. (1992) Science 258, 1160–1161.
[8] Murzin, A G. and Chothia, C. (1992) Curr. Opin. Struct. Biol. 2, 895–903.
[9] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 122, 535–542.
[10] Branden, C. and Tooze, J. (1991) An Introduction to Protein Structure, Garland Publishing Inc.
[11] Richardson, J.S. (1981) Adv. Prot. Chem. 34, 167–339.
[12] Kabsch, W and Sander, C. (1983) Biopolymers 22, 2577–2637.
[13] Kraulis, P.J. (1991) J. Appl. Cryst. 24, 946–950.
[14] Richardson, J.S. (1976) Proc. Natl. Acad. Sci. USA 73, 2619–2623.
[15] Murzin, A.G. (1993) EMBO J. 12, 861–867
[16] Sibanda, B.L., Blundell, T.L. and Thornton, J.M. (1989) J. Mol. Biol. 206, 759–777.
[17] Sakashita, H., Sakuma, T., Ohkubo, T., Kainosho, M., Sakumi, K., Sekiguchi, M. and Morikawa, K. (1993) FEBS Lett.