Journal of Biomolecular NMR, 5 (1995) 67–78
ESCOM

J-Bio NMR 268

Graph-theoretical assignment of secondary structure in multidimensional protein NMR spectra: Application to the *lac* repressor headpiece*

Elizabeth C. van Geerestein-Ujah, Monique Slijper, Rolf Boelens and Robert Kaptein**

Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Received 6 June 1994
Accepted 23 January 1995

Keywords: Graph theory; Pattern recognition; Secondary structure identification; Computer-assisted assignment; SERENDIPITY

Summary

A novel procedure is presented for the automatic identification of secondary structures in proteins from their corresponding NOE data. The method uses a branch of mathematics known as graph theory to identify prescribed NOE connectivity patterns characteristic of the regular secondary structures. Resonance assignment is achieved by connecting these patterns of secondary structure together, thereby matching the connected spin systems to specific segments of the protein sequence. The method known as SERENDIPITY refers to a set of routines developed in a modular fashion, where each program has one or several well-defined tasks. NOE templates for several secondary structure motifs have been developed and the method has been successfully applied to data obtained from NOESY-type spectra. The present report describes the application of the SERENDIPITY protocol to a 3D NOESY-HMQC spectrum of the ^{15}N-labelled *lac* repressor headpiece protein. The application demonstrates that, under favourable conditions, fully automated identification of secondary structures and semi-automated assignment are feasible.

Introduction

NMR has become an important alternative to X-ray crystallography for biomolecular structure determination. Its principal advantage is the fact that it allows the study of biomolecules in solution, i.e., under circumstances closer to their physiological environment. The process of structure determination by NMR can be divided into four main stages: (i) assignments of the proton resonances; (ii) determination of structural constraints from NOE and J-coupling data; (iii) calculation of structures satisfying these constraints; and (iv) refinement of structures. The first of these stages is manually quite laborious and remains the main bottleneck of structure determination. However, the development of new multidimensional NMR techniques and the use of labelled samples (^{15}N, ^{13}C) (Bax and Grzesiek, 1993; Bax, 1994; Clore and Gronenborn, 1994) have heralded the advent of much ongoing progress in this area. Further, several computational methods have been developed for the automatic assignment of protein NMR spectra (Billeter et al., 1988; Cieslar et al., 1988; Eads and Kuntz, 1989; Kraulis, 1989; Ikura et al., 1990; Kleywegt et al., 1993; Meadows et al., 1994). Most of these methods are based on the analysis of two- and three-dimensional ^{1}H spectra using the sequential assignment strategy (Billeter et al., 1982; Wüthrich, 1986). In general, all of these automated assignment programs contain bookkeeping features and use a stepwise strategy analogous to manual analysis, although the implementation of particular procedures may vary to some degree among them. Due to spectral artefacts and overlap of resonances, the results of the programs are not completely reliable and must be checked at each step. The main-chain-directed (MCD) approach (Nelson et al.,

*Supplementary Material available from the authors: Two tables containing the total number of mappings resulting from the graph search procedure for simulated and experimental NOE data.

**To whom correspondence should be addressed.

Abbreviations: 2D, 3D, two-, three-dimensional; NOESY, nuclear Overhauser enhancement spectroscopy; HMQC, heteronuclear multiple quantum coherence; SSE, secondary structure element; SERENDIPITY, SEcondary structuRE ideNtification in multiDImensional ProteIn specTra analYsis.

1991; Wand and Nelson, 1991) is an alternative to sequential assignment which combines the assignment and identification of secondary structure into a single step. The MCD approach has been automated in a series of programs. With the exception of the triple-resonance technique (Ikura et al., 1990), which bases its assignment protocol on through-bond coupling through adjacent residues, all of the above assignment procedures rely on nuclear Overhauser effects (NOEs) for the identification of proximal residues in the sequence.

In the present report, we describe a novel method that uses graph theory to carry out sequence-specific assignments and to identify the regular secondary structure (α-helix and β-sheet) elements that are present in a protein molecule from NOESY-type spectra. Graph theory is a convenient tool for describing and analysing molecular connectivity and has already found many applications in chemistry (Artymiuk et al., 1991,1994,1995; Rouvray, 1991; Ujah, 1992). Furthermore, graph theory has also been used in pattern recognition algorithms for the assignment of 2D NMR spectra of peptides and proteins (Oschkinat et al., 1988,1991; Pfändler and Bodenhausen, 1988; Liu et al., 1990; Xu et al., 1994). The method of Xu and coworkers (Xu and Sanctuary, 1993; Xu et al., 1993,1994), for instance, uses fuzzy-graph pattern recognition and graph-search methodologies for the assignment of 2D ^{1}H NMR spectra. Generally, the method first identifies spin coupling topologies by comparing a predicted data set based upon the cross peaks of NMR correlation spectra against experimental spectra. Then, with a fuzzy-graph pattern recognition algorithm and applying chemical shift rules (Groß and Kalbitzer, 1988), these spin coupling topologies are mapped to specific amino acids using DQF-COSY and TOCSY/HOHAHA peak sets. Sequence-specific assignments are achieved by creating a forest of spin coupling networks such that each tree in the forest consists of sequential spin graphs. An algorithm then searches this forest and finds the optimum sequence-specific assignment based on a NOESY data set.

In contrast to the pattern recognition methods mentioned above, our approach, known as SERENDIPITY (SEcondary structuRE ideNtification in multiDImensional ProteIn specTra analYsis), first identifies graph representations of NOE patterns in the spectrum that are characteristic of the regular secondary structure. Each graph matched in the spectrum corresponds to independent spin coupling systems, which in turn correspond to individual amino acid residues. Assignment is then achieved, with manual assistance, by connecting these elucidated (sub)-graphs of secondary structure together, thereby matching the connected spin systems to specific segments of the protein sequence. The input data set consists of a cross-peak list of the NOESY spectrum acquired with a peak-picking routine, as implemented in the program ALISON (Kleywegt et al., 1993). Our procedure requires that the NH, C^{α} and C^{β} protons be identified in the spectrum and correctly associated with their respective, albeit unassigned, spin coupling system (henceforth defined in this work as a partial spin system). In other words, we assume that unambiguous *partial* spin systems have been identified, but that these spin systems have not yet been conclusively assigned to amino acid types (multiple possibilities exist for most spin systems), nor have the spin systems been ascribed any sequential order. Formally, the method generates the maximal common subgraphs (Bron and Kerbosch, 1973; Lau, 1989) from among the observed NOE graphs (derived from the NOESY spectrum) and the NOE-graph template of secondary structure. In other words, the procedure determines the set of equivalent protons between the NOE graphs representing the secondary structure and the NOESY data set, respectively. Maximal common subgraph isomorphism algorithms have been used successfully to determine the similarity of specific patterns in small molecules (Brint and Willett, 1988) and also in protein database searches to extract common secondary structural features (Grindley et al., 1993).

In the present paper, the method is applied to the protein *lac* repressor headpiece. This is a small α-helical protein consisting of three helices (residues 6–13, 17–25 and 34–45), where 29 out of its 56 amino acid residues are in an α-helical conformation. The first two helices are part of a helix–turn–helix motif (helix II is the recognition helix), which is a common feature of several DNA-binding proteins (Harrison, 1991; Brennan, 1992). At present, all available structural data are based solely on NMR experiments (Zuiderweg et al., 1983a,b). We first tested the robustness of the algorithm by applying it to simulated NOE data comprising patterns of ^{1}H-^{1}H distances observed in the NMR-determined structure of the *lac* repressor headpiece–DNA complex (PDB code 1LCC) (Chuprina et al., 1993). The method was then directly applied to experimental NOE data pertaining to a 3D NOESY-HMQC spectrum (Marion et al., 1989; Zuiderweg and Fesik, 1989) of the ^{15}N-labelled *lac* repressor headpiece protein. Fully automated identification of the secondary structures was successfully achieved. Sequence-specific assignments were performed with manual assistance.

Materials and Methods

Automated secondary structure identification

SERENDIPITY has been developed in a modular fashion, where each program has one or several well-defined tasks. The core programs in the SERENDIPITY suite utilise a maximum common subgraph isomorphism algorithm (Lau, 1989) to recognise patterns of NOEs pertaining to the regular secondary structures. In the algorithm, short sequential, medium- and long-range ^{1}H-^{1}H distances characteristic of secondary structures and sufficiently short to be observed as NOEs are matched with

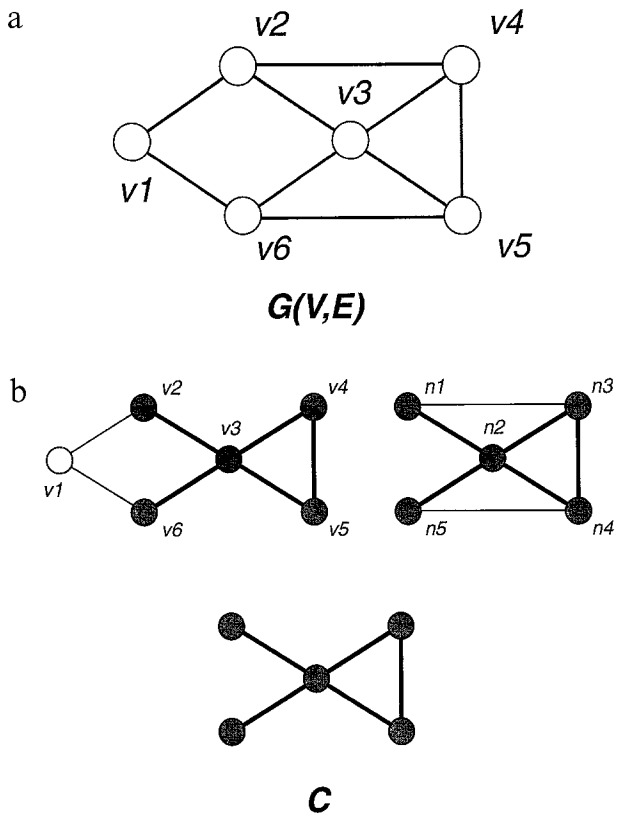

Fig. 1. Schematic representation of a graph and the isomorphism method used in this investigation. (a) A simple graph. (b) Generation of the maximal common subgraph. Two graphs and their maximal common subgraph, C, are represented, delineated by bold lines and shaded nodes. In this example, the maximum common subgraph consists of nodes v_2, v_3, v_4, v_5, v_6 that are isomorphous to the subgraph of nodes n_1, n_2, n_3, n_4, n_5 because their edges are also equivalent (not drawn to scale).

^{1}H-^{1}H distance constraints computed from the NOESY spectrum. The algorithm searches for all incidences of common overlap between two graphs. The method identifies partial (NH, C^α and C^β) spin coupling systems in the prespecified secondary structure conformation. Sequence-specific assignment is then achieved by connecting these graphs of spin systems together and matching them to specific residues in the primary sequence.

A graph-theoretical approach

The term 'graph' can be used to describe any abstract mathematical entity which can be formulated in terms of objects and the connections between them. Thus, if an object or situation can be abstracted as a graph, then graph theory can be used to analyse the problem domain. A graph $G(V,E)$ (see Fig. 1) is a finite set of vertices V (or nodes) which are related to each other by a finite set of edges E (that is, $E \subseteq V \times V$) (Harary, 1972; Deo, 1975; McGregor, 1982). Each edge and each vertex may in addition have a label specifying certain information about the edge or vertex, thus constituting a labelled graph. Two vertices are referred to as adjacent if they are connected by an edge. Two graphs G and H are isomorphic if they have identical graph-theoretical properties, in other words, if there is a one-to-one mapping between the vertices of G and H such that adjacent pairs of vertices in G are mapped to adjacent pairs of vertices in H. A subgraph of G is a subset F of the vertices of G, together with a subset J of the edges connecting pairs of vertices in F ($F \subseteq V$ and $J \subseteq F \times F$). Thus, a common subgraph of two graphs G and K consists of a subgraph g of G and a subgraph k of K, such that g is isomorphic to k. The maximal common subgraph (MCS) is the largest common subgraph between G and K (Fig. 1).

Secondary structure NOE-graph templates

The problem we are confronted with is one of matching idealised secondary structure NOE graphs with NOE graphs pertaining to the NOESY spectrum. Inspection of the regular polypeptide secondary structures reveals a variety of backbone ^{1}H-^{1}H distances that are sufficiently short to be observed as NOEs. For instance, the α-helix is primarily characterised by a close approach between residues i and i+3, and between residues i and i+4, whereas in β-structures the individual strands consist of almost fully extended polypeptide segments, which excludes these medium-range distances and includes a prevalence of

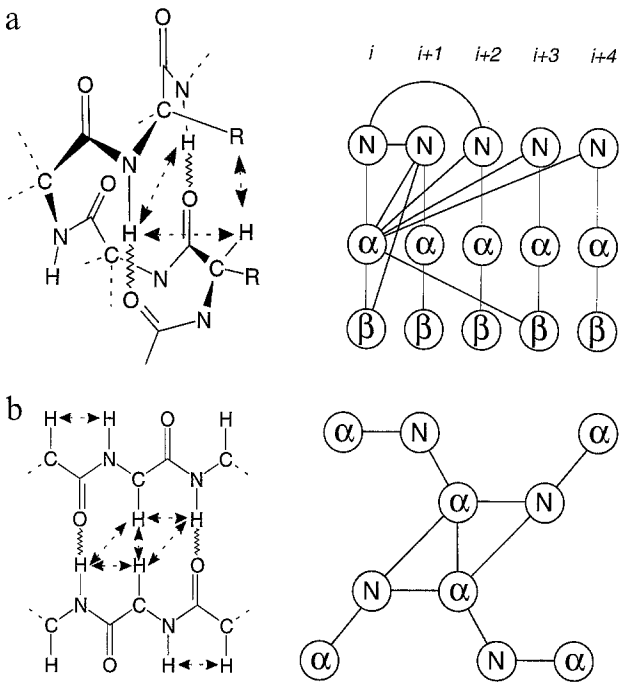

Fig. 2. Schematic of some regular secondary structures and their corresponding graph representations. (a) One turn of an α-helix and the corresponding graph representation; N, α and β refer to NH, C^α and C^β protons, respectively. Connectivities are shown between residues i, i+1, i+2, i+3 and i+4. Bold lines indicate interresidue connectivities. For clarity, the arrows show only some of the characteristic short distances that are likely to be observed as NOEs. (b) Antiparallel β-sheet and the corresponding graph representation; N and α are as for (a).

a

Connectivity	NOE-graph templates: hxI	hxII	hxIII	hxIV	hxV
NN(i,i+1)		2.8		2.8	
αN(i,i+1)		3.5		3.5	
NN(i,i+2)	4.2	4.2	4.2	4.2	4.2
αN(i,i+2)	4.4	4.4	4.4	4.4	4.4
αN(i,i+3)	3.4	3.4	3.4	3.4	3.4
αN(i,i+4)	4.2	4.2	4.2	4.2	4.2
αβ(i,i+3)			3.5	3.5	3.5
βN(i,i+1)				3.3	3.3

b

Connectivity	NOE-graph templates: β-strand	β-hairpin	apβ	pβ
NN(i,i+1)	4.2	4.5	4.5	4.2
αN(i,i+1)	2.2	2.3	2.3	2.2
βN(i,i+1)	4.2	3.8	3.8	4.2
αα(i,i+1)	4.2	4.3	4.5	4.2
NN(i,j)		3.3	3.3	4.0
αN(i,j)		4.5	4.5	3.0
αα(i,j)		2.6	2.6	4.8
Nα(i,j)		4.6	4.6	
NN(i,j+1)		4.3	4.5	
αN(i,j+1)		4.2	4.7	4.2

Fig. 3. ^{1}H-^{1}H connectivity distances (in Å) used in the regular secondary structures. (a) α-Helix and (b) β-sheet NOE-graph templates (apβ = anti-parallel β-sheet; pβ = parallel β-sheet; i and j refer to adjacent β-strands in a β-sheet).

short intrastrand main-chain ^{1}H-^{1}H distances. Overall, helices and tight turns can be characterised by short sequential and medium-range ^{1}H-^{1}H distances, and β-sheets by short sequential and long-range interstrand backbone ^{1}H-^{1}H distances (Wüthrich, 1986).

As shown in Fig. 2, these ^{1}H-^{1}H contacts form a continuous, dense network over the length of the secondary structures. Thus, the geometric arrangement of the backbone protons in secondary structures, which determines the characteristic pattern of observed NOEs, can be described by a labelled graph in which the vertices represent the NH, C^α and C^β protons and the edges their interproton distances. Each secondary structure NOE-graph template has been constructed to incorporate characteristic ^{1}H-^{1}H patterns *between* spin systems, i.e., between spin system i and spin system j, where j is any spin system other than i and where the spin systems correspond to individual amino acid residues. Thus, only interresidue interactions are matched, thereby providing information about specific interactions between hydrogen atoms in sequential and nonadjacent residues in the primary sequence that give rise to a characteristic pattern of cross peaks in a NOESY spectrum.

TABLE 1
SEARCH MATRIX OF NOE GRAPH hxIV CONSISTING OF INTERPROTON DISTANCES (in Å)[a]

	(i+1)HN	(i+2)HN	(i+3)HN	(i+3)HB	(i+4)HN
iHA	3.5	4.4	3.4	3.5	4.2
iHN	2.8	4.2	*[b]	*	*
iHB	3.3	*	*	*	*

[a] Interproton distances were compiled from systematic studies of standard secondary structures (Wüthrich, 1984,1986).
[b] The '*' represents a 'wild card' (i.e., infinite) error tolerance for an undefined distance, e.g. $d_{\beta N}(i,i+2)$ (column three, row three).

The elucidation of sequential residues is inherent in the α-helical connectivity patterns, which display couplings between the *i*th residue and up to and including the i+4th residue (Fig. 3a). However, the situation becomes somewhat more complicated in β-sheet structures because of the long-range nature of β-sheet topology. Nevertheless, sequential information can be derived from intrastrand connectivity patterns, mainly the $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ connectivities (Wüthrich, 1986) (Fig. 3B). The distance values adopted in the NOE-graph templates (Fig. 3) have been deduced following systematic studies with standard secondary structures generated from polyalanine with trans peptide bonds (Schulz and Schirmer, 1979; Richardson, 1981; Wüthrich et al., 1984; Wüthrich, 1986). One of the strengths of our methodology is the fact that it is not necessary to specify all possible distances in the secondary structure NOE-graph template. This feature has been exploited in the present work, where the templates described are disconnected graphs. As an example, the search matrix corresponding to the NOE-graph template hxIV (Fig. 3a) is shown in Table 1. The use of 'wild card' values (i.e., infinite error tolerances) enables the graph matching to be carried out.

SERENDIPITY protocol

A set of routines for processing NOESY cross-peak data into NOE graphs and identifying the secondary structure maximum common subgraphs within them have been grouped under the name SERENDIPITY. The program, which is written in FORTRAN 77 and runs on Silicon Graphics Iris workstations, can be adapted to work with other spectral analysis programs. In the present work, the SERENDIPITY routines were augmented with the program ALISON, our in-house interactive graphics program. The secondary structure assignment strategy is illustrated in Fig. 4.

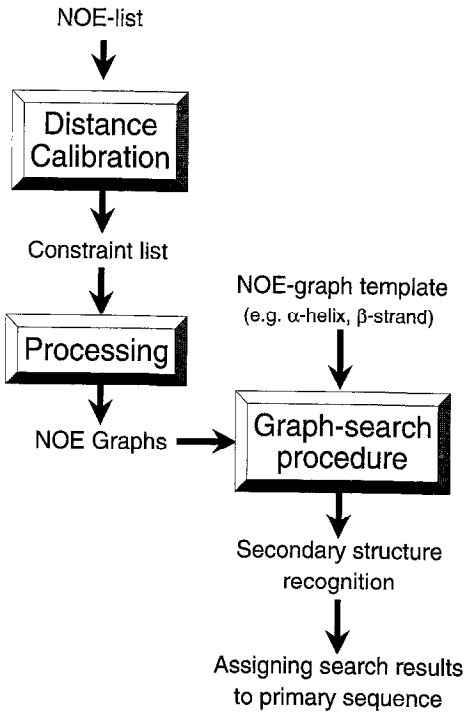

Fig. 4. General outline of the secondary structure assignment strategy of SERENDIPITY.

TABLE 2
EXTRACT OF DISTANCE CONSTRAINTS LIST[a]

ω1	ω3	Distance (Å)
s24:HA	s5:HN	3.5
s24:HB	s5:HN	2.6
s24:HN	s5:HN	2.4
s10:HB	s5:HN	4.7
s10:HB	s9:HN	4.0
s10:HB	s52:HN	2.9
s10:HN	s5:HN	2.7
s10:HN	s7:HN	4.1
s15:HA	s20:HN	3.7
s15:HB	s19:HN	3.4
s15:HB	s21:HN	4.4
s15:HN	s17:HN	3.0
s15:HN	s19:HN	3.3
s15:HN	s21:HN	3.3

[a] The list was compiled by grouping ^{1}H-^{1}H contacts of delineated spin systems together for the heteronuclear 3D [H,N,H] NOESY-HMQC spectrum of *lac* repressor headpiece. The letter 's' refers to a delineated spin system. The numbering of spin systems is entirely arbitrary and bears no relationship to sequential assignment.

The NOE list is a cross-peak list of the NOESY spectrum, acquired with the peak-picking routine implemented in ALISON. This list comprises arbitrarily numbered NOESY cross peaks, their volumes and intensities, and the corresponding ^{1}H-^{1}H contact. As previously mentioned, the spin systems are partial in nature (only the NH, C^α and C^β protons are known) and have been derived from other NMR (namely COSY- and TOCSY-type) methods. In the present case, this information was derived from a TOCSY spectrum of ^{15}N-labelled *lac* headpiece. All intra-spin system NOEs are discarded, so that the NOE list consists only of inter-spin system NOEs. The spin systems have not been ascribed any sequential order and the amino acid type of most of the spin systems remains unidentified at this stage, although multiple 'guesses', consistent with the available data, exist for most cross peaks. These guesses are generated by ALISON on the basis of chemical shift statistics (Richarz and Wüthrich, 1978; Groß and Kalbitzer, 1988; Wishart et al., 1991; Wishart and Sykes, 1994).

For the distance calibration facility (see Fig. 4), an approximate distance r is calculated by comparing the integrated volume v of the cross peak with the volume v_c of a cross peak or a set of cross peaks between protons with a known distance r_c. SERENDIPITY uses all the intra-residue $d_{\alpha N}(i,i)$ distances found in the original NOE list to calibrate the NOE intensities. Due to the constraints imposed by the covalent structure, this distance can vary

only between 2.4 and 2.9 Å (Billeter et al., 1982; Wüthrich, 1986) and, generally speaking, the $d_{\alpha N}(i,i)$ cross peaks should have a narrow intensity distribution consistent with this distance range. The program calculates the mean value of volumes of $d_{\alpha N}(i,i)$ cross peaks and associates this with a distance of 2.7 Å. From this calibration, all other interspin distances in the NOE list are calculated. The output of the distance calibration is a list of distance constraints associated with unidentified spin systems. This listing is then ordered by grouping together a spin system and all its interspin NOEs (Table 2). This ordered list of distance constraints is then processed into a series of NOE graphs (Fig. 5), where each NOE graph corresponds to an individual spin system (amino acid residue) and its partial (NH, C^{α} and C^{β} protons only) interspin system ^1H-^1H contacts. Thus, a NOESY cross peak $p(\omega_i,\omega_j)$ is considered as an edge in the NOE graph connecting protons i and j. In other words, the vertices of the NOE graph represent protons and the edges represent the calculated interproton distances. The NOE graph serves as input for the pattern recognition analysis (see Fig. 4).

Graph search procedure

SERENDIPITY uses a modified maximum common subgraph (MCS) isomorphism algorithm to identify all patterns of secondary structure NOE connectivities present in the NOE graphs derived from the NOESY spectrum. The MCS algorithm allows determination of the largest subgraph common to a pair of graphs. Thus, in the context of spectral analysis, this allows one to identify the extent of the secondary structure connectivity patterns in the NOE graphs. MCS algorithms are computationally expensive; comparing two graphs containing x and y nodes, in order to identify all subgraphs containing u nodes in common, requires up to $x!y!/u(x-u)!(y-u)!$ node-to-node comparisons. The identification of the largest common subgraph is achieved with u equal to 1, 2, etc. until it is no longer possible to identify a larger common subgraph (i.e., the condition ($1 \le u \le$ minimum $\{x,y\}$) is reached). In order to reduce the number of comparisons that need to be carried out, SERENDIPITY uses a graph-theoretic technique known as clique detection. A clique is a subgraph of a graph in which every node is connected to every other node, and which is not contained in any larger subgraph. The input to the clique detection procedure is a correspondence graph, i.e., an intermediate data structure that contains all possible equivalencies between the two graphs being compared. The identification of the largest clique in the correspondence graph gives the MCS between the two graphs. Specifically, given a pair of graphs G and H, a correspondence graph C is formed as follows:

(1) Create the set of all pairs of nodes, one from G and one from H, such that the nodes of each pair are of the same type.

(2) The correspondence graph C is the graph whose nodes are the pairs from (1). Thus, two nodes $(G(I),H(X))$ and $(G(J),H(Y))$ are considered as being connected in C if the values of the edges from $G(I)$ to $G(J)$ and $H(X)$ to $H(Y)$ are the same.

(3) Maximal common subgraphs then correspond to the cliques in the correspondence graph C (Barrow and Burstall, 1976).

Thus, the identification of the maximal common subgraph for a pair of NOE graphs, where the nodes and edges correspond to the protons and to the interproton distances, respectively, is equivalent to the identification of the largest clique in the correspondence graph. In addition, the method may also be used to identify all of the subgraphs in common when matching NOE graphs, not just the largest subgraph.

In the present work, the nodes of the two NOE graphs are paired to form a correspondence graph if the pair of protons in question are of the same type, for instance, either both C^{α} protons or both C^{β} protons. In practical applications, an exact matching of the secondary structure NOE-graph template with the spectrum NOE graph is not possible; therefore it is necessary to allow for some deviation in the matching of the interproton distances. The correspondence graph nodes are considered to be connected if the differences between the interproton distances for the two pairs of protons are below some user-specified tolerance, e.g. ±20% of the NOE-graph template distance. Therefore, the condition to match NOEs as edges is of the form $r_l < r < r_u$, where r is the NOE-graph template distance and r_u and r_l are the upper and lower distance tolerances, respectively.

The correspondence graph is set up as a symmetrical Boolean matrix of size $N \times N$, where N is the number of

Noe-Graph s24

^1H-^1H	s5:HN
s24:HA	3.5
s24:HB	2.6
s24:HN	2.4

Noe-Graph s10

^1H-^1H	s5:HN	s9:HN	s52:HN	s7:HN
s10:HB	4.7	4.0	2.9	–
s10:HN	2.7	–	–	4.1

NOE-Graph s15

^1H-^1H	s20:HN	s19:HN	s21:HN	s17:HN
s15:HA	3.7	–	–	–
s15:HB	–	3.4	4.4	–
s15:HN	–	3.3	3.3	3.0

Fig. 5. NOE graphs processed from a NOE list extract (Table 2) from the 3D NOESY-HMQC spectrum of *lac* repressor headpiece. Delineated spin systems, labelled 's', are arbitrarily numbered. Numbers in tabular representation refer to computed interproton distances (in Å).

nodes in the correspondence graph, C. The diagonal elements of the correspondence graph matrix are set to TRUE, whereas all other elements in the matrix are set to TRUE or FALSE, depending on whether the two correspondence graph nodes involved are connected or not. Cliques within C are located by finding the largest subset of the matrix whose elements are all true. The clique detection algorithm is a depth-first, backtracking tree search. A complete description of the algorithmic procedure can be found elsewhere (Bron and Kerbosch, 1973; Lau, 1989; Grindley et al., 1993).

Optimising search procedures

The introduction of distance tolerances tends to produce multiple mappings for a given NOE-graph template. However, only one of the mappings for a q-sized clique (where q is the number of edges) can be accepted as the correct solution. In order to reduce the number of undesirable solutions, we have allocated weights to the matching of a combination of specific edges. These weights are dependent on the type of secondary structure that is being identified. For instance, the identification of strong d_{NN} NOEs combined with $d_{\alpha N}(i,i+3)$ and $d_{\alpha N}(i,i+4)$ NOE connectivities is highly indicative of the presence of α-helices, as is the combination of interstrand $d_{\alpha\alpha}$ contacts, strong sequential $d_{\alpha N}$ contacts and weak d_{NN} contacts for β-sheet structures. Thus, adopting this approach, we select only those matched common subgraphs comprising the combination of the edges as specified above for α-helix and β-sheet NOE graphs, respectively. The implementation of such weights leads to substantial reductions in the number of mappings. The method can be used to identify just the largest common subgraph (obtained as the result of an exhaustive tree search that identifies all mappings that are common to the two NOE graphs), or all common subgraphs that are larger than some user-defined minimum clique size (Grindley et al., 1993). This restriction is imposed to reduce the number of undesirable mappings consisting primarily of small common subgraphs that have little or no significance.

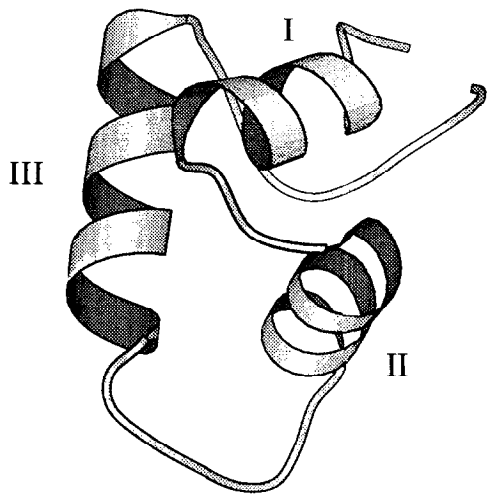

Fig. 6. MOLSCRIPT schematic (Kraulis, 1991) of the tertiary structure of the *lac* repressor headpiece protein.

TABLE 3
PERCENTAGE OF α-HELIX REGIONS IDENTIFIED IN SIMULATED NOE DATA FROM THE *lac* HEADPIECE PROTEIN

NOE-graph template[a]	No. of nodes of detected (sub)graphs							
	4		5		6		7	
hxI (5)	79	79[b,c]	21	21				
	14	12[d]	14	14				
hxII (6)	90	90	76	69	21	21		
	16	11	8	9	14	14		
hxIII (6)	79	72	28	28	7	7		
	21	12	3	3	0	0		
hxIV (8)	90	83	76	69	31	31	10	10
	24	8	19	13	10	10	0	0
hxV (8)	83	83	48	45	14	14		
	25	11	22	13	0	0		

[a] Numbers in brackets give the nodal size of the NOE-graph template.
[b] The first row of numbers for each of the graph templates refers to the percentage of successfully identified α-helical regions.
[c] The second column under each nodal heading gives the results obtained using selective edge retrieval.
[d] The second row of numbers for each graph template refers to the percentage of incorrect solutions obtained.

Applications

The method was applied to NOE data pertaining to the *lac* repressor headpiece protein, which is the N-terminal domain of the monomeric subunit of the *lac* repressor protein. A schematic of the tertiary structure of *lac* headpiece is given in Fig. 6. The structure of *lac* headpiece has been determined by NMR (Kaptein et al., 1985; De Vlieg et al., 1986,1988), as has the structure of the complex of *lac* headpiece and an 11-base-pair *lac* half-operator (Chuprina et al., 1993). This complexed structure has been deposited in the Brookhaven Protein Data Bank (PDB code 1LCC) and is used in the application pertaining to simulated NOE data described below. Hydrogen atoms were added to 1LCC with the Insight program (InsightII, Version 2.2.0, Biosym Technologies, San Diego, CA, 1993) using standard protein bond lengths and geometry.

The SERENDIPITY protocol was applied to experimental NOE data from a 3D NOESY-HMQC spectrum (Marion et al., 1989; Zuiderweg and Fesik, 1989) of the ^{15}N-labelled *lac* repressor headpiece protein, measured in 95% H_2O/5% D_2O, pH 4.5 at 500 MHz and 298 K. The NH, C^{α} and C^{β} protons of the spin systems had previously been assigned using 2D NMR methods (Zuiderweg et al., 1983b). An NOE list was generated by ALISON in which the sequential assignment of the spin systems involved was not known, although multiple guesses existed for most spin systems. The method considers each of these possibilities and attempts to match them with the

TABLE 4
SEARCH MATRIX OF NOE-GRAPH TEMPLATE hxr CONSISTING OF INTERPROTON DISTANCES (in Å)[a]

	(i,i−1)HA	(i,i−2)HA	(i,i−3)HA	(i,i−4)HA	(i,i−1)HN	(i,i−2)HN	(i,i−1)HB
iHN	3.5	4.2	3.4	4.3	2.8	4.4	3.2
iHB	*[b]	*	3.2	*	*	*	*

[a] Interproton distances were compiled from systematic studies of standard secondary structures (Wüthrich, 1984,1986).
[b] The '*' represents a 'wild card' (i.e., infinite) error tolerance for an undefined distance.

primary sequence. This NOE list served as input to SERENDIPITY. The results obtained are presented below.

Results

Simulated NOE data

A computer program was used to extract all main-chain ^{1}H-^{1}H distances and to form a distance constraint list. From this listing, 55 NOE graphs were constructed, where the vertices corresponded to the main-chain protons and the edges to the interproton distances. As previously mentioned, *lac* headpiece is an α-helical protein, where 29 out of its 56 amino acid residues are in α-helical conformation (Chuprina et al., 1993). α-Helix NOE-graph templates (Fig. 3) were used to identify the spin systems in α-helical regions. In order to simulate the situation encountered when dealing with unassigned spin systems, the sequence numbering and residue identity were concealed, hence all results were obtained without prior knowledge of either the sequence ordering or the identity of the amino acid residue. Only the proton type and the interproton distances were used to recognise patterns.

For clarity, these results are presented in Table 3. For each pattern, the first row of numbers refers to the percentage of α-helical regions in the protein that was successfully identified. The second row refers to the percentage of incorrect solutions obtained. From the table we see that both the hxII and hxIV patterns were very successful, identifying up to 90% of the helical regions. Nevertheless, as with all of the patterns, the degree of false identifications (i.e., the percentage of spin systems wrongly identified as α-helical) is quite high, sometimes reaching 25%. Implementing the weighting procedure, where only those subgraphs containing a combination of d_{NN}, $d_{\alpha N}(i,i+3)$ and $d_{\alpha N}(i,i+4)$ distances are considered, reduces the number of false solutions, whilst the percentage of correct solutions remains essentially the same (see Table 3). Once again, the best results were obtained with the hxII and hxIV NOE graphs. However, all of these patterns are somewhat inadequate at identifying terminal α-helical residues, as the $d_{\alpha N}(i,i+3)$ and $d_{\alpha N}(i,i+4)$ connectivities are generally not found at the end of α-helices. This feature was addressed by the use of an additional NOE-graph template, hxr (Table 4), which simply comprises the 'reverse' of the ^{1}H-^{1}H distances found in the α-helix NOE graphs shown in Fig. 3. The reverse template is more suitable for identifying the C-terminal regions of helices. Unmasking the amino acid identities of the 55 NOE graphs, we summarise the results obtained from the

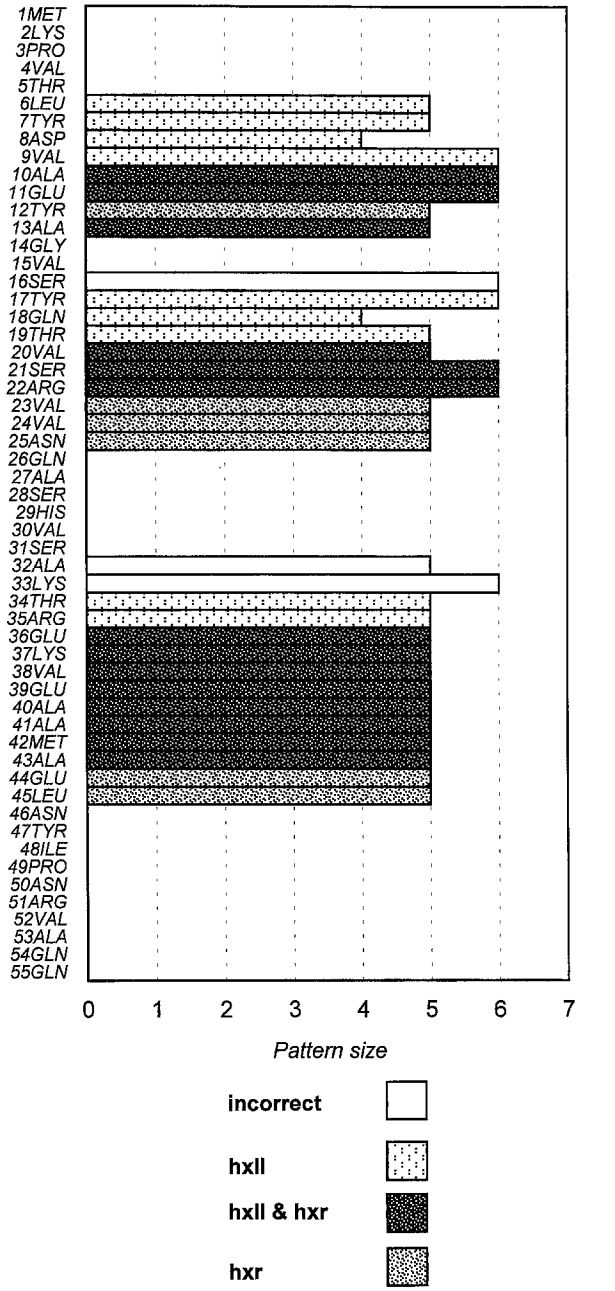

Fig. 7. Results from an analysis of *lac* headpiece data with hxII and hxr NOE-graph templates. As indicated, the patterned bars depict those residues identified by the hxII and/or the hxr NOE graphs. Plain bars depict an incorrect analysis (see text for further details).

combined use of the hxII and hxr NOE-graph templates in Fig. 7. It can be seen that all of the spin systems in α-helical conformations are now completely identified. Only three spin systems were incorrectly identified as being in α-helical conformations. These were unmasked and discovered to be Ser^{16}, Ala^{32} and Ala^{33}. These residues immediately precede the beginning of helix II (Ser^{16}) and helix III (Ala^{32}, Ala^{33}), respectively. Closer inspection of the interproton distances manifested by these residues reveals the presence of the $d_{\alpha N}(i,i+3)$ and $d_{\alpha N}(i,i+4)$ distance ranges. Furthermore, secondary structure classifications using the algorithm developed by Kabsch and Sander (1983) as implemented in the program PROCHECK (Laskowski et al., 1993) classify Ser^{16} ($\phi=-66.8,\psi=151.8$) as an *extension* of an α-helix (i.e., a partial α-helical conformation, helical ψ but different φ and commonly found at the termini of helices) and Ala^{32} ($\phi=-55.4,\psi=-52.4$) and Ala^{33} ($\phi=-48.1,\psi=-43.5$) as α-helical. Thus, although not classified as α-helical in the NMR-determined structure (Chuprina et al., 1993), these residues are found in α-helix-type conformations and hence were duly identified by the α-helix NOE-graph template.

Using the hxII template and selective weighting, where only those cliques mapping a combination of d_{NN}, $d_{\alpha N}(i,i+3)$ and $d_{\alpha N}(i,i+4)$ distances were accepted, resulted in a total of 487 mappings (Supplementary Material; Table 5). Cliques of four and five nodes were generated with an edge tolerance of 25%, and cliques of six nodes with an edge tolerance of ±30%. Linking the correct graph mappings together from the set of generated mappings was done manually by beginning with the largest cliques of six nodes (for which fewer mappings were present), and then forming a sequence with those graphs exhibiting a contiguous pattern of sequential protons. The procedure is exemplified in Fig. 8, where we begin by examining the six-node clique mapped by NOE graph number 3 (s3) (the numbering of NOE graphs is arbitrary). This was the only mapping generated at this size. NOE graph number 28 (s28) also generated one solution at this clique size; however, s3 was chosen as fewer cliques were mapped in s3 in total. According to the graph-matching, s14 is the sequential (i+1) neighbour of s3. NOE graph s14 maps two six-node cliques. Both of these mappings share a common sequence up to and including i+2. The sequence mapped by the first of these overlaps precisely with the sequence described by s3 and hence this mapping is accepted. From these solutions, NOE graph s5 is found to be the sequential neighbour of s14. NOE graph s5 maps four six-node cliques, which all share a common i+1 neighbour, NOE graph s16. This graph maps two six-node cliques, both of which share common sequential neighbours up to and including i+2. Therefore, on the basis of this analysis we accept the sequence

s3 s14 s5 s16 s7 s8

which unmasks to reveal the residues

8asp 9val 10ala 11glu 12tyr 13ala

Continuing with this analysis of the six-node cliques and then looking for overlapping sequences among the five- and four-node cliques, we succeeded in elucidating three contiguous sequences which, when unmasked, comprised the correct sequence of spin systems, residues 6–15, 17–26 and 32–45. The sequential nature of the α-helix templates facilitates the elucidation of further residues that are adjacent to the helical residues (6–13, 17–25 and 34–45).

The present application demonstrates the α-helix identification facility of SERENDIPITY. The matching of

HxII NOE-Graph Template Nodes

Mappings	iHA	iHN	(i+1)HN	(i+2)HN	(i+3)HN	(i+4)HN
s3	3HA	3HN	14HN	5HN	16HN	7HN
s14	14HA	14HN	5HN	16HN	7HN	8HN
s14	14HA	14HN	5HN	16HN	8HN	7HN
s5	5HA	5HN	16HN	7HN	20HN	19HN
s5	5HA	5HN	16HN	7HN	20HN	25HN
s5	5HA	5HN	16HN	8HN	20HN	25HN
s5	5HA	5HN	16HN	8HN	20HN	19HN
s16	16HA	16HN	7HN	8HN	20HN	19HN
s16	16HA	16HN	7HN	8HN	19HN	20HN

Sequence accepted: s3 s14 s5 s16 s7 s8

Fig. 8. The process of linking delineated spin systems from the set of generated mappings to form a sequence, beginning with the largest clique of six nodes of s3 (for which only one mapping was present) and then forming a sequence with those NOE graphs which exhibited an overlapping pattern of sequential protons (see text for further details). The NOE graphs which form the accepted sequence are underlined.

delineated spin systems to the primary sequence was done with NOE data from a 3D NOESY-HMQC spectrum and is described below.

Experimental NOE data

The 3D NOESY-HMQC spectrum was used as basic input for the ALISON program, which was utilised for the automatic picking of cross peaks. The peak data set was manually edited, and spurious cross peaks were deleted. Authentic peaks that were unpicked for one reason or another (e.g. peak overlap) were picked manually and the sizes of their peak boxes were readjusted where necessary to represent the number and size(s) of the peak(s) present. Figure 9 shows a cross section of the 3D NOESY-HMQC spectrum of the ^{15}N-labelled *lac* repressor headpiece protein in H_2O. The characteristic connectivities of Val^{23} (located in helix II) are exemplified in the figure.

ALISON generated an NOE list consisting of 291 cross peaks involving interactions between protons in partial spin systems. This NOE list served as input to SERENDIPITY. In the distance calibration procedure, 42 $d_{\alpha N}(i,i)$ interactions were elucidated from the NOE list and used to compute the distance constraints. These were processed into 48 NOE graphs (Fig. 5) which were numbered accordingly.

The hxII and hxIV NOE-graph templates were used for the subgraph isomorphism and only those cliques containing a combination of d_{NN}, $d_{\alpha N}(i,i+3)$ and $d_{\alpha N}(i,i+4)$ distances were selected automatically. Finally, the hxr NOE-graph template was used to elucidate terminal helix residues. A total of 387 multiple mappings were generated using the hxII, hxIV and hxr NOE-graph templates (Supplementary Material; Table 6). An edge tolerance of ±50% was used in all searches. Following the procedure outlined for the simulated data set, we attempted to join the mapped subgraphs into contiguous sequences. As before, beginning with the largest cliques, we first considered NOE graph s40 (once again, the numbers were arbitrarily generated, referring to delineated spin systems from the NOE list). This NOE graph mapped two solutions for a six-node clique. Both of these mappings shared a common sequence up to and including the third sequential residue. This common sequence is s40, s16, s27, s33, where both NOE graphs s16 and s27 featured in the mapping solutions of NOE graph 40, yet neither of these graphs produced any mappings that satisfied the given search criteria. NOE graph s33 produced seven five-node solutions. Analysing these mappings and also the cliques of their sequentially mapped NOE graphs, we deduced the following contiguous sequence:

s32 s40 s16 s27 s33 s18 s26 s46 s47 s48 s49 (1)

NOE graph s49 did not generate any cliques during the graph-search procedure, but was itself mapped as a sequential neighbour of NOE graph s48. Continuing with the four six-node mappings generated by searching NOE graph s1, and considering all pathways, elucidated the following sequence of NOE graphs (i.e., delineated spin systems):

s1 s10 s3 s12 s5 s6 s14 s7 (2)

Continuing with the four five-node mappings generated by s29 and then looking at smaller four-node cliques, we elucidated the sequence

s15 s4 s9 s28 s20 s29 (3)

Evidently, this procedure of joining the correct sequence of NOE graphs together becomes far more complicated as the number of possible graph solutions increases; thus, further work is necessary to formalise a procedure that can automatically identify the correct

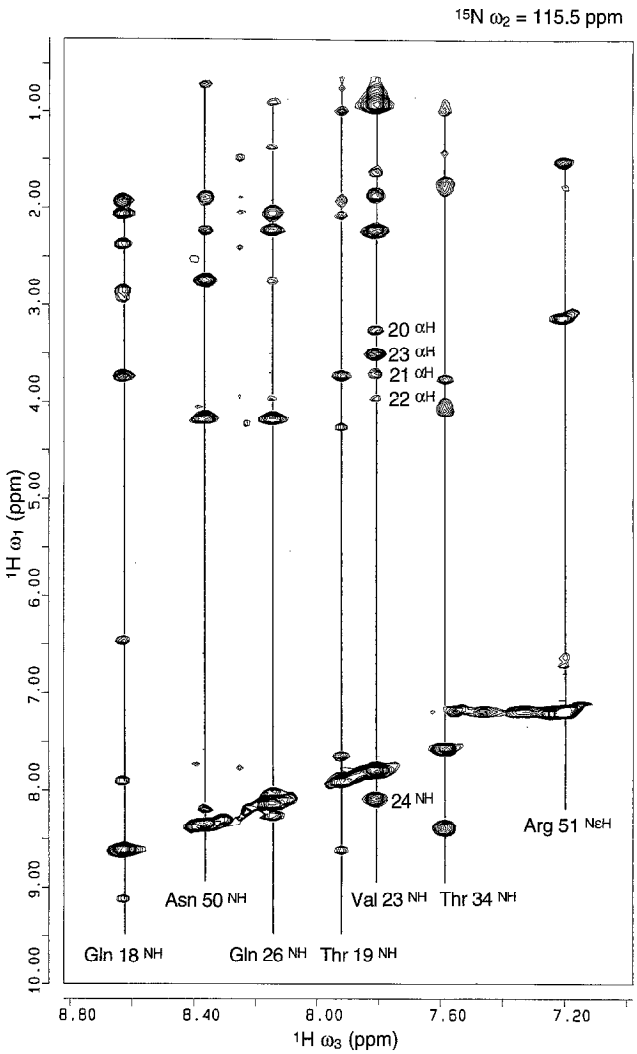

Fig. 9. A cross section of the 3D NOESY-HMQC spectrum of the ^{15}N-labelled *lac* repressor headpiece protein in H_2O. The characteristic connectivities of Val^{23} (located in helix II) are outlined.

sequential assignment from the set of possible candidates. This could be achieved by a combinatorial minimisation methodology (Christofides et al., 1979).

As previously mentioned, the amino acid identity of the delineated spin systems is unknown. However, 'guesses' consistent with the available COSY and TOCSY data can be generated by the program ALISON on the basis of chemical shift statistics. From such an analysis, NOE graph 14 is designated as a glycine residue, NOE graphs 3, 6 and 33 as alanine residues, and NOE graphs 7, 9, 10, 29 and 40 as valine residues. Sequence (2) is then

s1 s10 ala s12 s5 ala gly val

which, upon examination of the primary sequence can map only onto the stretch of residues

8asp 9val **10ala** 11glu 12tyr **13ala** **14gly** **15val**

In the same way, sequence (1) was mapped onto the primary sequence comprising 37lys–48ile, and sequence (3) onto residues 18gln–23val. Manual assignment of the 3D NOESY-HMQC spectrum (M. Slijper, 1994, personal communication) successfully verified these results.

Conclusions

We have described a novel method for the automatic identification of secondary structures in proteins from their corresponding NOE data. Using distance constraints derived from the NMR-determined complex of *lac* headpiece and, subsequently, NOE data from a 3D NOESY-HMQC spectrum of ^{15}N-labelled *lac* headpiece, we have shown that under favourable conditions fully automated identification of the secondary structures is feasible. Then, with manual assistance, the secondary structures can be assigned. In many instances multiple mappings for a given template may occur. However, the number of mappings is drastically reduced when the size of the matched subgraph increases and by selectively retrieving specific combinations of secondary structure-dependent edges. Additional research in further developing the assignment strategy is currently underway in order to alleviate the manual effort involved in linking the correct sequence of NOE graphs together from the set of possible subgraphs generated, and to match these NOE graphs to the primary sequence. However, the SERENDIPITY protocol provides a potential aid to automated assignment and a means to rapidly and comprehensively assess the secondary structure content of a protein from its NOE data set. It is expected that the ongoing development of this protocol and its use in the program ALISON will provide a powerful tool for computer-assisted assignment and secondary structure analysis of protein NMR spectra.

Acknowledgements

This work was supported by an EU HCM grant to E.G.U. M.S. acknowledges the Netherlands Foundation for Chemical Research (SON) and the Netherlands Organisation for Scientific Research (NWO) for financial support.

References

Artymiuk, P.J., Grindley, H.M., Mitchell, E.M., Rice, D.W., Ujah, E.C. and Willett, P. (1991) In *Recent Advances In Chemical Information* (Ed., Collier, H.), Royal Society of Chemistry, Cambridge, pp. 91–106.

Artymiuk, P.J., Grindley, H.M., Poirrette, A.R., Rice, D.W., Ujah, E.C. and Willett, P. (1994) *J. Chem. Inf. Comput. Sci.*, **34**, 54–62.

Artymiuk, P.J., Grindley, H.M., MacKenzie, A.B., Rice, D.W., Ujah, E.C. and Willett, P. (1995) In *Molecular Similarity and Reactivity* (Ed., Carbo, R.) in press.

Barrow, H.G. and Burstall, R.M. (1976) *Inform. Proc. Lett.*, **4**, 83–84.

Bax, A. and Grzesiek, C. (1993) *Acc. Chem. Res.*, **26**, 131–138.

Bax, A. (1994) *Curr. Opin. Struct. Biol.*, **4**, 738–744.

Billeter, M., Braun, W. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 321–346.

Billeter, M., Basus, V.J. and Kuntz, I.D. (1988) *J. Magn. Reson.*, **76**, 400–415.

Brennan, R.G. (1992) *Curr. Opin. Struct. Biol.*, **2**, 100–108.

Brint, A.T. and Willett, P. (1988) *J. Comput.-Aided Mol. Design*, **2**, 311–320.

Bron, C. and Kerbosch, J. (1973) *Commun. Assoc. Comput. Machinery*, **16**, 575–577.

Christofides, N., Mingozzi, A., Toth, P. and Sandi, S. (Eds) (1979) *Combinatorial Optimization*, Wiley-Interscience, London.

Chuprina, V.P., Rullmann, J.A.C., Lamerichs, R.M.J.V., Van Boom, J.H., Boelens, R. and Kaptein, R. (1993) *J. Mol. Biol.*, **234**, 446–462.

Cieslar, C., Clore, G.M. and Gronenborn, A.M. (1988) *J. Magn. Reson.*, **80**, 119–127.

Clore, G.M. and Gronenborn, A.M. (1994) *Methods Enzymol.*, **239**, 349–363.

De Vlieg, J., Boelens, R., Scheek, R.M., Kaptein, R. and Van Gunsteren, W.F. (1986) *Isr. J. Chem.*, **27**, 181–188.

De Vlieg, J., Boelens, R., Scheek, R.M., Van Gunsteren, W.F., Berendsen, H.J.C., Kaptein, R. and Thomason, J. (1988) *Proteins*, **3**, 209–218.

Deo, N. (1975) *Graph Theory with Applications to Engineering and Computer Science*, Prentice Hall, Englewood Cliffs, NJ.

Eads, C.D. and Kuntz, I.D. (1989) *J. Magn. Reson.*, **82**, 467–482.

Grindley, H.M., Artymiuk, P.J., Rice, D.W. and Willett, P. (1993) *J. Mol. Biol.*, **229**, 707–721.

Groß, K.-H. and Kalbitzer, R. (1988) *J. Magn. Reson.*, **76**, 87–99.

Harary, F. (1972) *Graph Theory*, Addison-Wesley, Reading, MA.

Harrison, S.C. (1991) *Nature*, **353**, 715–719.

Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.

Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.

Kaptein, R., Zuiderweg, E.R.P., Scheek, R.M., Boelens, R. and Van Gunsteren, W.F. (1985) *J. Mol. Biol.*, **182**, 179–182.

Kleywegt, G.J., Vuister, G.W., Padilla, A., Knegtel, R.M.A., Boelens, R. and Kaptein, R. (1993) *J. Magn. Reson. Ser. B*, **102**, 166–176.

Kraulis, P.J. (1989) *J. Magn. Reson.*, **84**, 627–633.

Kraulis, P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crystallogr.*, **26**, 283–290.

Lau, H. (1989) *Algorithms on Graphs*, TAB Books Inc., Blue Ridge Summit, PA.

Liu, X., Balasubramanian, K. and Munk, M.E. (1990) *J. Magn. Reson.*, **87**, 457–474.

Marion, D., Driscoll, L.E., Kay, P.T., Wingfield, A., Bax, A., Gronenborn, A.M. and Clore, G.M. (1989) *Biochemistry*, **28**, 6150–6156.

McGregor, J.J. (1982) *Software – Pract. Exp.*, **12**, 23–34.

Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.

Nelson, S.J., Schneider, D.M. and Wand, A.J. (1991) *Biophys. J.*, **59**, 1113–1122.

Oschkinat, H., Griesinger, C., Kraulis, P.J., Sørensen, O.W., Ernst, R.R., Gronenborn, A.M. and Clore, G.M. (1988) *Nature*, **332**, 374–376.

Oschkinat, H., Holak, T.A. and Cieslar, C. (1991) *Biopolymers*, **31**, 699–712.

Pfändler, P. and Bodenhausen, G. (1988) *J. Magn. Reson.*, **79**, 99–123.

Richardson, J.S. (1981) *Adv. Protein Chem.*, **34**, 167–339.

Richarz, R. and Wüthrich, K. (1978) *Biopolymers*, **17**, 2133–2141.

Rouvray, D.H. (Ed.) (1991) *Computational Chemical Graph Theory*, Nova, New York, NY.

Schulz, G.E. and Schirmer, R.H. (1979) *Principles of Protein Structure*, Springer, Berlin.

Ujah, E.C. (1992) Ph.D. Thesis, University of Sheffield, Sheffield.

Wand, A.J. and Nelson, S.J. (1991) *Biophys. J.*, **59**, 1101–1112.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wüthrich, K., Billeter, M. and Braun, W. (1984) *J. Mol. Biol.*, **180**, 715–740.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.

Xu, J., Gray, B. and Sanctuary, B.C. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 475–489.

Xu, J. and Sanctuary, B.C. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 490–500.

Xu, J., Strauss, S.K., Sanctuary, B.C. and Trimble, L. (1994) *J. Magn. Reson. Ser. B*, **103**, 53–58.

Zuiderweg, E.R.P., Kaptein, R. and Wüthrich, K. (1983a) *Proc. Natl. Acad. Sci. USA*, **80**, 5837–5841.

Zuiderweg, E.R.P., Kaptein, R. and Wüthrich, K. (1983b) *Eur. J. Biochem.*, **137**, 279–292.

Zuiderweg, E.R.P. and Fesik, S.W. (1989) *Biochemistry*, **28**, 2387–2391.