NMR recipe for sequencing short DNA fragments

R. Ajay Kumar, K. Majumdar and R. V. Hosur
Chemical Physics Group, Tata Institute of Fundamental Research, Bombay 400 005, India

A new recipe has been described for determination of the base sequence in short DNA segments by two-dimensional NMR spectroscopy. The recipe is based on (i) A,T,G,C-distinguishing criteria obtained by analysis of chemical shifts of the non-exchangeable protons and (ii) cross-peak patterns in two-dimensional COSY and NOESY spectra. The base H8 and sugar H2″ chemical shifts have been found to be characteristically dependent on the base type to which they belong and the patterns of H8–H2″ cross-peaks in NOESY allow determination of sequence of bases in DNA segments.

We describe here a novel application of two-dimensional NMR spectroscopy[1,2], namely sequencing of DNA segments. The 2D-NMR methods of significance here are J-correlated spectroscopy (COSY)[3,4] and nuclear Overhauser effect correlated spectroscopy (NOESY)[5,6]. The 'cross-peaks' in COSY display J-coupling (through bond) correlations while those in NOESY display dipolar coupling (through space) correlations and carry proximity information (interproton distance less than 5 Å).

There are basically two steps in the sequencing procedure: (i) nucleotide units A, T, G and C must be distinguished in the two-dimensional spectrum, and (ii) adjacent nucleotide units in the sequence must be identified.

The nucleotide units C and T are readily distinguishable from the 2D COSY and NOESY spectra via the characteristic CH6–CH5 and TH6–TCH_3 cross-peaks, which appear in distinct and identical regions in both the spectra. Further, due to the relatively large CH6–CH5 coupling constant ($\approx$ 7 Hz) the cross-peaks originating from CH6 protons often appear as doublets in the NOESY spectrum. As regards A and G, although these units do not display such characteristic features in the 2D spectra, we observed from a statistical analysis of published chemical shift data on various right-handed double-helical DNA segments[7–22], that they can be distinguished on the basis of chemical shifts of base H8 and sugar H2″ protons (Figure 1). In Figure 1,*a*, the chemical shifts of H8/H6 protons of the bases A, G, T and C in different DNA segments have been plotted against their positions along the sequence; Figure 1,*b* shows a similar plot for the H2″ protons. In both these cases, the A and G nucleotides have markedly distinct chemical shifts, irrespective of position along the sequence or nature of the sequence. A few terminal units deviate from this rule but they can be readily identified (see below).

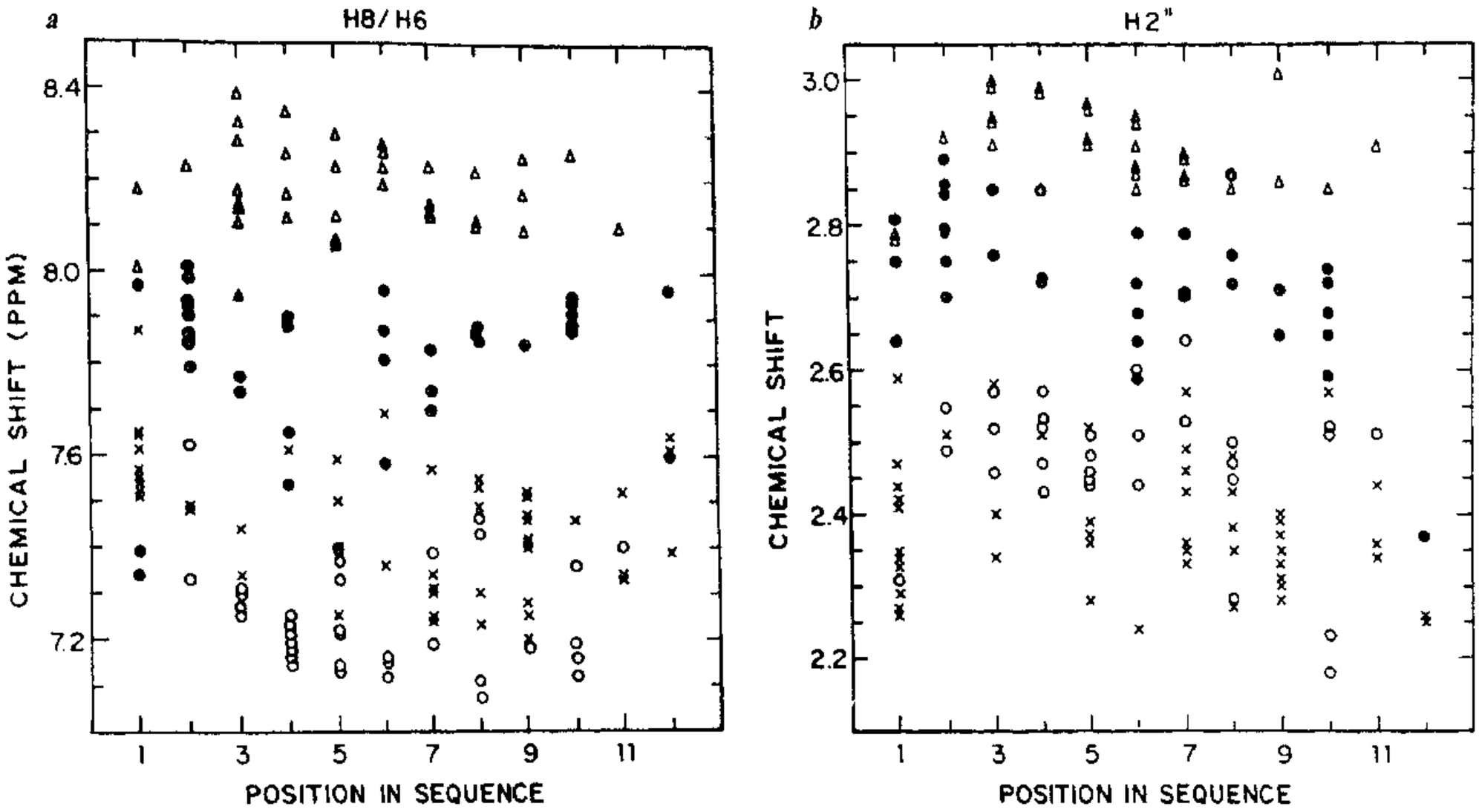

Figure 1. *a*, Chemical shifts of H8 or H6 protons of bases A, G, T, C, plotted against the position of the respective nucleotide unit along the sequence in various DNA segments. Secondary structure studies of these DNA segments have been published during the last five years[7–22]. The selected DNA segments varied from 6 to 19 nucleotides in length, and the spectra from which the chemical shift data base was prepared, were recorded under roughly similar experimental conditions of pH ($\approx$7.0), salt concentration (0.01–0.1 M), temperature (20–35°C, i.e. below melting temperature), etc. All the DNA segments had duplex structure but exhibited localized sequence-specific variations. All of them were right-handed double helices with overall topology ranging from B DNA to midway between A and B DNA structures. The bases are discriminated by different symbols: A (△), G (●), C (×) and T (○). *b*, Chemical shifts of H2″ protons, plotted in a similar fashion as in *a*.

Having thus obtained a discriminating criteria for A, G, T and C, we have devised a strategy based on 2D-NOESY spectra, for identifying adjacent nucleotide units in a given molecule. The useful spectral region covers the chemical shifts of H8/H6 protons along the ω_2 axis and those of H2′/H2″ and CH_3 protons along the ω_1 axis. However, peaks originating from H2′ protons can be clearly identified[17], and can be excluded from the analysis. We shall label the above region as the RISD (region of interest for sequence determination). When a NOESY spectrum is recorded with a sufficiently long mixing time ($\approx$ 300–400 ms), the RISD will contain the cross-peaks $(H8/H6)_i \rightarrow (H2'')_i$, $(H8/H6)_i \rightarrow (H2'')_{i-1}$, $(TCH_3)_i \rightarrow (H6)_i$ and $(TCH_3)_i \rightarrow (H8/H6)_{i-1}$ (i increases from the 5′-end to the 3′-end). From this, the 5′-terminal (H8/H6) proton and the 3′-terminal H2″ proton can be readily identified, since each of these protons produces only one H8/H6–H2″ cross-peak. Every other H2″ proton generates two such peaks, namely $(H2'')_i \rightarrow (H8/H6)_i$ and $(H2'')_i \rightarrow (H8/H6)_{i+1}$ and thus allows identification of adjacent nucleotide units.

Figure 2 shows the approximate positions of H8/H6→H2″ and H8/H6→TCH_3 cross-peaks in the RISD for all possible dinucleotide segments. Clearly, in any given RISD all dinucleotide stretches can be readily identified; CT/TC discrimination which is not possible from H8/H6→H2″ peaks is possible from TH6→TCH_3 cross-peaks. Further, the RISD also allows identification of adjacent dinucleotide segments with a common partner; this is illustrated in the Figure 2 by dashed vertical lines joining the peaks belonging to the common units in triplets ATA and ACG. Thus, in a given molecule the complete sequence can be read out from H8/H6→H2″ peaks is possible from TH6→TCH_3 cross-peaks. Further, the RISD also allows identification joining the adjacent pairs with a common partner by vertical lines as indicated above. Such a pattern of horizontal and vertical lines may be abbreviated as SDCP (sequential dinucleotide connectivity pattern). In a given RISD, there will be only one SDCP if the molecule is self-complementary, and two patterns—of the same length—if the two strands of the duplex have different sequences. In the latter case, the sequence derived from the two patterns will be complementary. In both cases, the total number of As equals the total number of Ts and the total number of Gs equals the total number of Cs. These factors provide additional checks for the validity of the derived sequence and any error in A, G discrimination can be immediately identified.

Figure 3 illustrates the above methodology with an experimental spectrum taken from the literature[15]. We

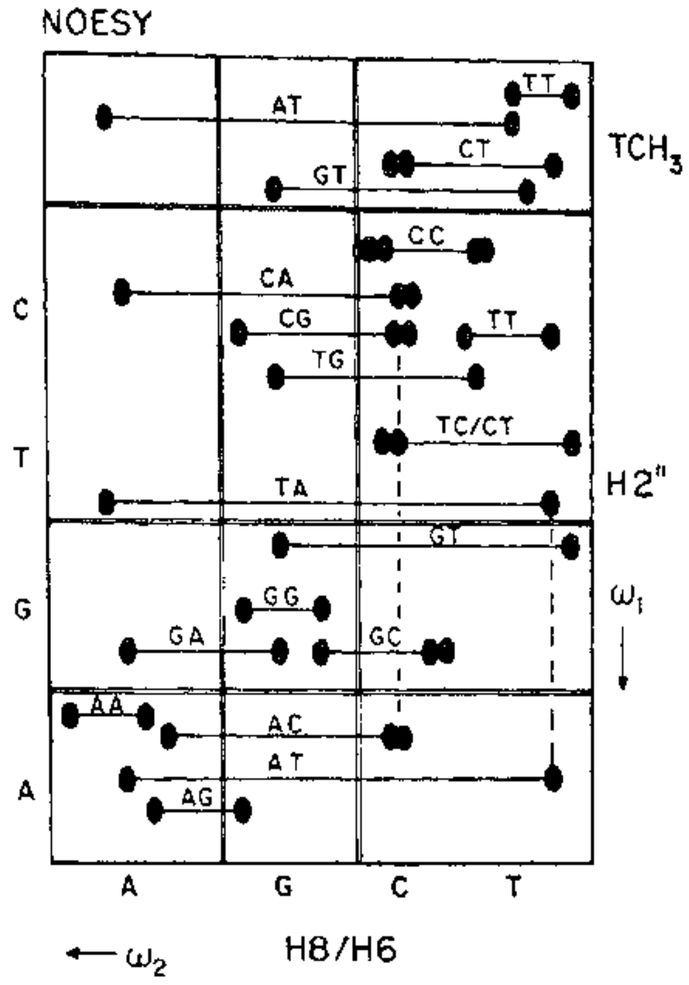

Figure 2. Schematic display of H8/H6 → H2″ and H8/H6 → TCH_3 cross-peaks in the RISD for all possible dinucleotide segments. The spectral region is divided into subregions in accordance with the chemical shift information in Figure 1, pairs of peaks belonging to individual dinucleotide segments are joined by horizontal lines. In reality the RISD also contains peaks originating from H2′ proton and these must be clearly distinguished. This discrimination can be easily obtained from NOESY spectra recorded with short mixing time (< 100 ms)[19]. In each box, the relative positions of the peaks is not significant (they have been chosen arbitrarily for clarity). For example, the A peak of AT could be above or below the A peak of the AG segment. Accordingly, the positions of the two horizontal lines may get reversed, but will remain in their respective domains. Dashed vertical lines illustrate identification of adjacent dinucleotides with a common partner. Peaks originating from CH6 are shown as doublets.

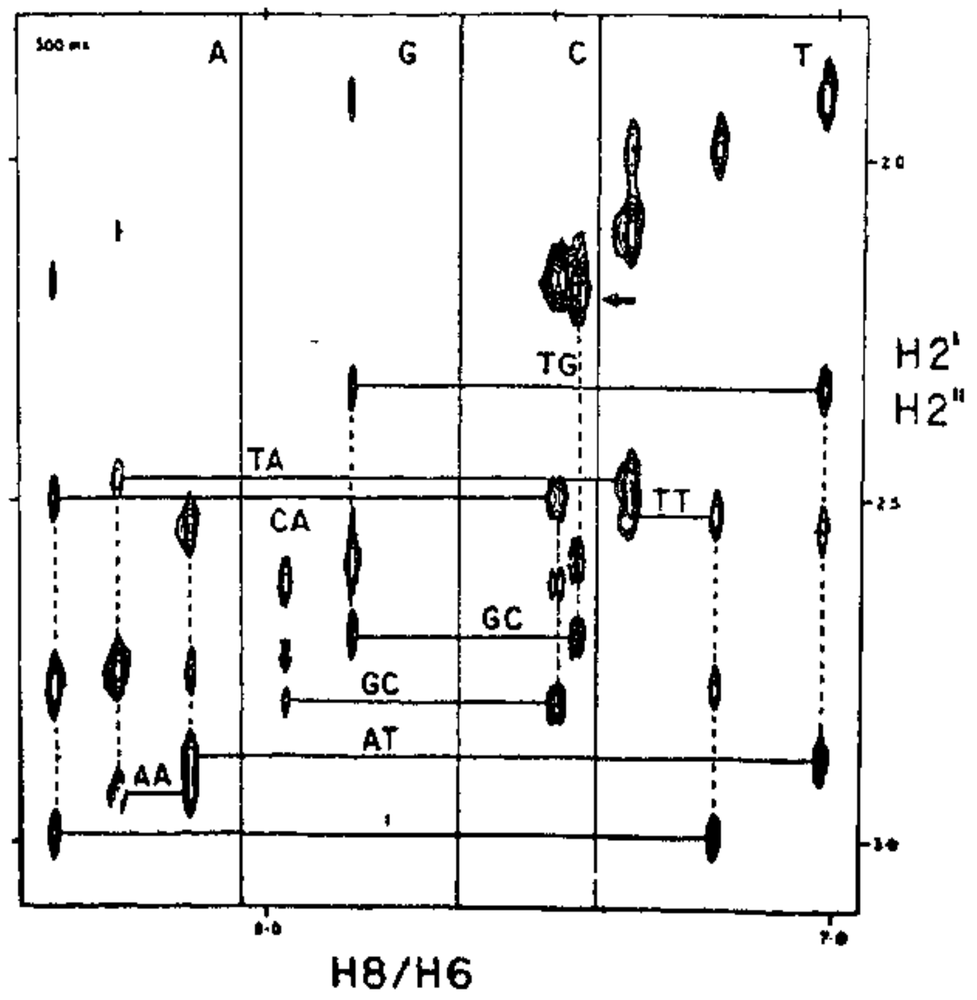

Figure 3. Experimental demonstration of the sequencing procedure. The 2D NOESY spectrum shows NOE cross-peaks between H8/H6 protons and H2′/H2″ protons. Peaks originating from H2″ protons are joined by horizontal lines and their assignment to specific dinucleotide segments are shown above the lines. Adjacent dinucleotide pairs with common partners are joined by dashed vertical lines at the chemical shift position of the common base proton. The thick vertical arrow identifies the H8 proton of 5′-terminal nucleotide and indicates the beginning of SDCP. The thick horizontal arrow identifies the H2″ proton of the 3′-terminal nucleotide and indicates the end of SDCP in the spectrum. The NOESY spectrum has been taken from the literature[17]. Experimental conditions are: mixing time = 300 ms; temperature = 28°C, pH = 7.0.

have purposely taken a spectrum from work belonging to another laboratory to minimize our bias about the molecule. For the reader, the sequence is unknown and therefore, there is no bias about the analysis. The spectrum in Figure 3 represents the RISD without the peaks arising from the TCH_3 protons. Two cytosines can be identified straightaway from their doublet patterns. A total of ten base protons can be easily counted and, in every case, all the expected peaks to H2′/H2″ are clearly distinguishable. The thick vertical arrow identifies the base proton (H8) of the 5′-terminal nucleotide and indicates the beginning of SDCP. From another set of COSY and NOESY spectra it was concluded that the H2″ proton resonates downfield of the H2′ proton in all the cases. Using this information, nine dinucleotide pairs originating from H2″ proton connectivities have been identified and joined by horizontal lines in the figure. They have been assigned in accordance with Figure 2 and their assignments are indicated above the horizontal lines. The adjacent dinucleotide pairs with a common partner have then been identified and joined by vertical dashed lines. From these, the sequence can be readily obtained as d-GCATTAATGC. This is indeed the true sequence of the molecule[15].

The above approach has also been successfully tested on a few DNA segments. An important requirement for the successful application of the proposed procedure is that the cross-peaks should be well separated in the NOESY spectrum. In this context use of the modern NMR techniques such as selective pulse techniques[23] and three-dimensional techniques[24,25] may be envisaged. Experimental conditions must also be suitably adjusted to ensure right-handed duplex structure for the molecule; other physical techniques such as circular dichroism will be helpful in this regard. Observation of imino proton resonances in H_2O spectra will help confirm the duplex state of the DNA segment.

The NMR method of sequencing DNA segments also has some limitations. (i) At present the size of DNA segment has to be less than 20 base pairs; this arises due to large line width (short T_2) problems and insufficient resolution in the spectra. (ii) Repetitive sequences such as AAAAATTTTT or GGGGGCCCCC are difficult to handle. (iii) Abnormal synthetic sequences such as hair pins or loop structures are not

amenable to analysis because of their modified chemical shift patterns.

1. Ernst, R. R., Bodenhausen, G. and Wokaun, A., *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Claredon Press, Oxford, 1987.
2. Wüthrich, K., *NMR of Proteins and Nucleic Acids*, Wiley, New York, 1986.
3. Jeener, J., *Ampere Summer School on Magnetic Resonance and Related Phenomena*, Basko Polje, Yugoslavia, 1971.
4. Aue, W. P., Bartholdi, E. and Ernst, R. R., *J. Chem. Phys.*, 1976, **64**, 2229.
5. Jeener, J., Meier, B. H., Bachmann, P. and Ernst, R. R., *J. Chem. Phys.*, 1979, **71**, 4546.
6. Anil Kumar, Wüthrich, K. and Ernst, R. R., *Biochem. Biophys. Res. Commun.*, 1980, **95**, 1.
7. Chary, K. V. R., Hosur, R. V., Govil, G., Tan, Z. and Miles, H. T., *Biochemistry*, 1987, **26**, 1315.
8. Chary, K. V. R., Hosur, R. V., Govil, G., Tan, Z. and Miles, H. T., *Biochemistry*, 1988, **27**, 3858.
9. Hare, D. R., Wemmer, D. E., Chou, S. H., Drobny, G. and Reid, B. R., *J. Mol. Biol.*, 1983, **171**, 319.
10. Ravikumar, M., Hosur, R. V., Roy, K. B., Miles, H. T. and Govil, G., *Biochemistry*, 1985, **24**, 7703.
11. Sheth, A., Ravikumar, M., Hosur, R. V., Govil, G., Tan, Z. and Miles, H. T., *Biochem. Biophys. Res. Commun.*, 1987, **144**, 26.
12. Sheth, A., Ravikumar, M., Hosur, R. V., Govil, G., Tan, Z., Roy, K. B. and Miles, H. T., *Biopolymers*, 1987, **26**, 1301.
13. Chary, K. V. R., Modi, S., Hosur, R. V., Govil, G., Chen, C. Q. and Miles, H. T., *Biochemistry*, 1989, **28**, 5240.
14. Sheth, A., Hosur, R. V., Govil, G., Hosur, M. V., Kannan, K. K., Tan, Z. and Miles, H. T., *Biochemistry*, 1989, **28**, 7275.
15. Chazin, W. J., Wüthrich, K., Hyberts, S., Rance, M., Denny, W. A. and Leupin, W., *J. Mol. Biol.*, 1986, **190**, 439.
16. Lown, J. W., Hanstock, C. C., Bleackley, R. C., Imbach, J. L., Rayner, B. and Vasseur, J. J., *Nucl. Acids. Res.*, 1984, **12**, 2519.
17. Hosur, R. V., Govil, G. and Miles, H. T., *Magn. Reson. Chem.*, 1988, **26**, 927.
18. Rinkel, L. J., Sanderson, M. R., Van der Marel, G. A., Van Boom, J. H. and Altona, C., *Eur. J. Biochem.*, 1986, **159**, 85.
19. Scheek, R. M., Boelens, R., Russo, N., Van Boom, J. H. and Kaptein, R., *Biochemistry*, 1984, **23**, 1371.
20. Wemmer, D. E., Chou, S. H., Hare, D. R. and Reid, B. R., *Biochemistry*, 1984, **23**, 2262.
21. Feigon, J., Leupin, W., Denny, W. A. and Kearns, D. R., *Biochemistry*, 1983, **22**, 5943.
22. Shindo, H., Okhubo, S., Matsumuto, U., Giessner-Prettre, C. and Zon, G., *J. Biomol. Struct. Dyn.*, 1988, **5**, 913.
23. Bermel, W., Kessler, H., Griesinger, C. and Oschkinat, H., *BRUKER Report*, 1986, **1**.
24. Griesinger, C., Sorensen, O. W. and Ernst, R. R., *J. Magn. Reson.*, 1987, **73**, 574.
25. Griesinger, C., Sorensen, O. W. and Ernst, R. R., *J. Magn. Reson.*, 1989, **84**, 14.

ACKNOWLEDGEMENT. We thank Prof. G. Govil for useful suggestions.

9 April 1990; revised 15 June 1990

National Symposium on Mushrooms

Place: Thiruvananthapuram
Date: 22–24 January 1991
Contact: Dr M. C. Nair
Department of Plant Pathology
College of Agriculture
Vellayani 695 022

International Conference on Neutron Scattering

Place: Bombay, India
Date: 21–25 January 1991
Contact: Dr K. R. Rao
Organizing Committee–NS '91
Nuclear Physics Division
Bhabha Atomic Research Centre
Bombay 400 085

Topics to be covered include Interferometry; Neutron optics; High-temperature superconductors; Magnetic structures; Elementary excitations; Phase transitions; Molecular spectroscopy; Low-dimensional systems; Incommensurate and quasiperiodic systems; Liquids and amorphous systems; Surfaces and interfaces; Micelles, microemulsions and membranes; Polymers and macromolecules; Structures of biological interest; Emerging trends in instruments and techniques; Pulsed and steady-state sources.

International Symposium on Oceanography of the Indian Ocean

Place: Goa, India
Date: January 1991

The symposium is aimed to provide a forum to present and discuss the advances in Oceanography of the Indian Ocean and its adjacent seas, especially in the post International Indian Ocean Expedition years.

Contact: Dr A. H. Parulekar
Convener, Organizing Committee
ISOIO, National Institute of Oceanography
Goa 403 004

International Conference on Brazil Gold '91

Place: Belo Horizonte, Minas Gerais, Brazil
Date: 7–22 May 1991
Contact: Prof. V. K. Nayak
Department of Applied Geology
Indian School of Mines
Dhanbad 826 004

The conference aims to provide a forum to present and discuss matters related to various aspects of geology, exploration and genesis of gold ores.

Abstracts of talks given at the fiftysixth annual meeting of the Indian Academy of Sciences at Bhubaneswar, 8–11 November 1990

Spin–statistics connection in two dimensions

R. Ramachandran, *Institute of Mathematical Sciences, Madras 600 113*

The spin-statistics connection—that species with integral angular momentum obey Bose statistics and those with half odd integral spin obey Fermi statistics—is an important ingredient of field theory in the familiar (3 + 1)-dimensional space–time. For a two-dimensional space, which admits fractional statistics, we establish such a connection and study the minimal set of constraints necessary to achieve this in various two-dimensional manifolds. The minimum assumption that we make is the existence of antiparticles and the possibility of pair creation and annihilation. With these we find that the dimensional representations correspond to familiar bosons and fermions and the fractional statistics is sustainable only for multicomponent representations.

Self-organized criticality in nature

Deepak Dhar, *Tata Institute of Fundamental Research, Bombay 400 005*

Many natural objects and phenomena, such as soot particles, mountain ranges, and $1/f$ noise in electrical networks, display power-law correlations in space or in time. In the conventional theory of many-body systems, one can get such long-range correlations only for very special values of system parameters when the system is at a phase-transition point. Recently, Bak *et al.* have introduced theoretical models in which an open dissipative system organizes itself into a steady state showing power-law correlations without need to fine-tune any parameters. These are said to show self-organized criticality. The idea has attracted much attention as one providing a framework for the description of the ubiquitous power laws in nature.

The simplest of these models is of a sand-pile. Sand is dropped on a flat table at a slow, steady rate. For long times, one gets a sand-pile whose mass has small fluctuations about an average constant value, excess added sand coming out from the edges. The quantities of interest are the relative frequencies of different-sized avalanches caused by adding a tiny amount of sand. This model has a very interesting mathematical structure, and we have obtained several exact results about its behaviour. Earthquakes can be described by a similar model. Stresses build up at a steady rate along fault lines owing to movement of continental plates. Relaxation of stress occurs in irregular bursts, with larger bursts less frequent. This gives a qualitative explanation of the well-known Gutenberg–Richter law in geophysics.

These models give power laws without fine-tuning any parameters. Though highly simplified, they capture some essential features of the phenomenon. Models taking account of other complicating factors, or describing other phenomena, are also being studied.

Pressure-induced amorphization

S. K. Sikka, *High Pressure Physics Section, Bhabha Atomic Research Centre, Bombay 400 085*

Amorphous solids are traditionally prepared by quenching a melt sufficiently rapidly to prevent crystallization, the result being that they are a metastable, kinetically frozen state of the liquid. However, several new methods have been discovered that induce a direct solid-state transition from a crystalline to an amorphous state[1] (interdiffusion, radiation damage, mechanical alloying, cold rolling, etc.). During the last five years, the pressure variable has been added to this list. Mishima *et al.*[2] were the first to obtain a glass through compression of hexagonal ice. They noted that the melting temperature of ice Ih falls with increasing pressure and an extrapolation suggested that ice should melt at 77 K between 0.5 and 1 GPa. As this temperature is well below the glass-transition temperature, the melt would be an amorphous solid. Following this, Jeanloz[3] suggested that similar transformations might occur when solids with melting curves having positive slopes are decompressed at a suitable temperature. Since the publication of ideas, there has been a rapid increase in the number of materials that have been shown to undergo the crystalline–amorphous transition with pressure. We review the experimental