

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/112025>

**Copyright and reuse:**

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)

The protein bodies of the endoplasmic reticulum of the seeds of the castor oil plant contain a variety of storage proteins, along with two lectins, viz. castor bean agglutinin and ricin. Ricin is a heterodimer consisting of an A chain which is toxic to eukaryotic translation systems, and a B chain which has galactose-binding activity; the whole molecule is toxic to cells and animals by virtue of the ability of the A chain to enzymatically inactivate ribosomes after it crosses the cell membrane, this latter being achieved after binding of the molecule to cell surfaces by the B chain. Ricin is a dimer, and is not significantly toxic to cells.

### CASTOR BEAN LECTINS: CONSTRUCTION AND SEQUENCING OF cDNA CLONES

Previous work on the synthesis of the lectins indicated that each subunit had its own precursor, each being constitutively expressed and glycosylated in the endoplasmic reticulum. The lectins were thought to fold after transport to the protein bodies. However, the presence of a single precursor for the mature B chains, and further molecular-based evidence indicated that it contains both A and B chain sequences. The precursor was identified as an albumin which constituted lectin preparations used for various purposes.

A thesis presented by

Francis Ian Lamb

for the Degree of

Doctor of Philosophy

of the

UNIVERSITY OF WARWICK

The work reported here involves the identification of the precursor of the constitutive B chain. The precursor contains A and B chain sequences, by means of the cloning of cDNA complementary to lectin-specific cDNA. Clones of nearly full length have been obtained and sequenced, and the precursor is shown to have an N-terminal signal sequence, which is followed by the A chain sequence, and then by the B chain sequence. A linker of 12 amino acids is shown to be present between the two chains. Differences between these and other lectins are discussed. The sequences are related into context by comparison with other plant nucleotide and protein sequences.

Department of Biological Sciences, University of Warwick

September 1984

Lectins are reviewed, and the use of the cloned sequences is discussed with special reference to their potential use in cancer therapy.



AUTHOR F E

TITLE CARD FOR  
CONTRACT

I agree that the  
regulations govern

I agree that the

Any proposal to  
amend (normally  
be made by the  
Office at the  
"Regulations")

I agree that the  
purpose only)  
(The maximum  
exceed three  
years will be

- (i) I understand
- these will
- (ii) I further
- in any way

DATE

---



---



---



---



---



---



---



---



---



---

CONSTRUCTION AND EXAMINATION OF THE

TO

**Heil, Healey and Ba**

A thesis presented by

Francis Lee East

for the degree of  
Doctor of Philosophy  
of the

UNIVERSITY OF WASHINGTON

Department of Biological Sciences, University of Washington

September 1981

#### SUMMARY

The protein bodies of the endosperm of the seeds of the castor oil plant Ricinus communis contain a variety of storage proteins, along with two lectins, ricin and Ricinus communis agglutinin. Ricin is a heterodimer consisting of an A chain which is toxic to cell-free translation systems, and a B chain which has galactose-binding activity; the whole molecule is toxic to cells and animals by virtue of the ability of the A chain to enzymatically inactivate ribosomes after it crosses the cell membrane, this latter being achieved after binding of the molecule to cell surfaces by the B chain. The agglutinin consists of two ricin-like species linked non-covalently, is divalent, and is not significantly toxic to cells.

Previous work on the synthesis of the lectins indicated that each subunit had its own precursor, each being cotranslationally segregated and glycosylated, and assembly of the lectins was thought to occur after transport to the protein bodies. However, the putative B chain precursor was far larger than the mature B chains, and further protein-based evidence indicated that it contains both A and B chain sequences. The former A chain precursor was identified as an albumin which contaminated lectin preparations used for raising antisera.

The work reported here confirms that the putative B chain precursor contains A and B chain sequences, by means of the cloning of cDNA complementary to lectin-specific mRNA. Clones of nearly full length have been obtained and sequenced, and the precursor is shown to have an N-terminal signal sequence, which is followed by the A chain sequence, and then by the B chain sequence. A linker of 12 amino acids is shown to be present between the two chains. Sequences corresponding to both lectins are reported, and the similarities and differences between them are discussed. The sequences are placed into context by comparison with other plant nucleotide and protein sequences.

Literature on the castor bean lectins is reviewed, and the uses of the clones are discussed, with special reference to their possible use in immunotoxins for cancer therapy.

## CONTENTS

Acknowledgements	i
Declaration	ii
List of figures	iii
Abbreviations	vi
<u>Chapter 1: Introduction</u>	1
1A Preamble	1
1B Isolation and structure of castor bean lectins	7
1C Synthesis of the lectins	22
1D Biological functions of the lectins	25
1D1 General features	25
1D2 Properties of the A chains	27
1D3 Properties of the B chains	32
1D4 Entry of ricin into cells	36
1E Uses of castor bean lectins	44
1F Castor bean lectin cDNA cloning	50
<u>Chapter 2: Materials and Methods</u>	52
2A Materials	52
2B General methods	54
2B1 Growth and harvesting of castor beans	54
2B2 Growth and maintenance of <u>E coli</u>	54
2B3 Preparation of competent cells & transformation	54
2B4 Extraction and purification of plasmid DNA	56
2B5 Restriction enzymes and mapping	57
2B6 <u>In vitro</u> protein synthesis and immunoprecipitation	58
2B7 End-labelling of DNA fragments	59

2C	Gel electrophoresis	61
2C1	Neutral agarose gels	61
2C2	Alkaline agarose gels	61
2C3	Formamide agarose gels	62
2C4	SDS-Polyacrylamide gels	62
2C5	DNA sequencing gels	62
2C6	Elution of DNA from gels	63
2D	Extraction and fractionation of castor bean mRNA	64
2E	Construction of cDNA library in pBR322	66
2E1	Synthesis of double-stranded cDNA	66
2E2	Construction of recombinants	67
2F	Screening of the cDNA library	69
2F1	Differential hybridisation	69
2F2	Oligomer hybridisation	70
2F3	Translation of hybridisation-selected mRNA	72
2G	Subcloning into pUC8	74
2H	M13 cloning and dideoxy sequencing	75
2H1	Cloning into M13	75
2H2	Dideoxy sequencing	76
2I	Maxam and Gilbert sequencing	78
2I1	Preparation of DNA fragments	78
2I2	Modification and cleavage reactions	79
2J	Primer extension	82
2J1	Cloning of the primer	82
2J2	Primer extension reaction	84
2K	Nucleic acid transfer and hybridisation	86
2K1	Southern blotting	86
2K2	Northern blotting	86

<b>Chapter 3: Results and Discussion</b>	<b>87</b>
3A Isolation and characterisation of mRNA	87
3B cDNA synthesis	93
3C Construction of clones	98
3D Screening of the cDNA library	100
3E Characterisation of clones	108
3F Subcloning into pUC8	115
3G Sequencing: Introduction	119
3G1 Summary of results	119
3G2 M13 cloning and dideoxy sequencing	126
3G3 Maxam & Gilbert sequencing	129
3H The preproricin sequence	135
3I The preproagglutinin sequence	137
3J Analysis of protein sequences	139
3J1 Comparison with protein-determined sequences	139
3J2 Comparison of deduced amino acid sequences	149
3J3 The signal and linker peptides	158
3K Analysis of nucleotide sequences	165
3K1 Comparison of prepro-lectin sequences	165
3K2 Base composition, dinucleotides and codon usage	168
3K3 The 5' non-coding region	173
3K4 The 3' non-coding region	185
3L Primer extension	193
3M Northern blotting	195
<b>Chapter 4: Final Discussion</b>	<b>199</b>
<b>Appendix 1: Computer programmes</b>	<b>213</b>
<b>Appendix 2: Restriction data</b>	<b>237</b>
<b>References</b>	<b>243</b>

ACKNOWLEDGEMENTS

I wish to thank my Supervisor, Dr JM Lord, for continual help, advice and encouragement, and Dr IM Roberts and many others at Warwick for information, suggestions and materials, and to Professor H Woodland for performing the oocyte injections.

I am indebted to Celltech Ltd for the provision, free of charge, of the oligodeoxynucleotide with which the cDNA library was screened, and especially to Dr T Harris and Dr M Bodnar, for two weeks spent at Celltech learning M13 cloning and dideoxy sequencing.

My thanks also to the Agriculture and Food Research Council for the grant which supported this work.

I am very grateful to my wife Helen for moral support and tolerance, and for assistance with the manuscript, and to my mother for helping with access to the books and journals of the British Library.

Fig	Title	Page
1B-1	Structure of castor bean lectins	11
1B-2	Sequences of A and B chains	15
1B-3	Disulphide bonds in the ricin B chain	17
1E-1	Uses of the castor bean lectins	42
2J-1	Cloning of primer extension primer free cDNA	43
3A-1	Stimulation of reticulocyte lysosomes by mRNA	44
	5' / PAGE analysis of mRNA translation products	47

**DECLARATION**

The conclusions reached in this Thesis are my own, and are based on the experiments reported, on many discussions with my Supervisor, Dr JM Lord, with Dr LM Roberts, and others, and on published works (opp cit).

The experiments were performed in the Department of Biology, University of Bradford (first year) and the Department of Biological Sciences, University of Warwick. They were carried out by myself, though Dr Roberts provided much assistance with cDNA synthesis; Dr Lord performed the experiments described in fig 3A-2, and Professor H Woodland carried out the oocyte injections and labelling.

*F. I. Lang*

3B-1	Characterisation of the mRNA	117
3B-2	Southern blot of DNA	121
3B-3	Characterisation of the mRNA	123
3B-4	Presence of 5' cap in the mRNA	124
3B-5	Characterisation of the mRNA	125
3B-6	Characterisation of the mRNA	126
3B-7	Characterisation of the mRNA	127
3B-8	Characterisation of the mRNA	128
3B-9	Characterisation of the mRNA	129

LIST OF FIGURES (continued)

<u>Fig</u>	<u>Title</u>	<u>Page</u>
1B-1	Structure of castor bean lectins	11
1B-2	Sequences of ricin A and B chains	14
1B-3	Disulphide bonds in the ricin B chain	17
1E-1	Uses of the castor bean lectins	45
2J-1	Cloning of primer extension primer from pRCL6	83
3A-1	Stimulation of reticulocyte lysates by mRNA	88
3A-2	SDS/PAGE analysis of mRNA translation products	89
3A-3	mRNA fractionation: translation products	91
3A-4	mRNA fractionation: immunoprecipitates	92
3B-1	Effect of temperature on cDNA synthesis	94
3B-2	Synthesis of cDNA used for cloning	95
3B-3	dC-tailing of cDNA	97
3D-1	Translation of early and late castor bean mRNAs	101
3D-2	Screening by differential hybridisation	102
3D-3	Screening by oligodeoxynucleotide hybridisation	104
3D-4	Digestion of plasmids with EcoRI	106
3D-5	Translation of hybridisation -selected mRNA	107
3E-1	Digestion of plasmids with PstI	109
3E-2	Digestion of plasmids with BamHI	110
3E-3	Southern blot of BamHI digests	111
3E-4	Extents of eight lectin-specific clones	113
3E-5	Presence of BglII sites in fragment BamHI-D	114
3F-1	Subcloning into pUC8: recombinants	116
3F-2	Determination of orientations of pUC8 subclones	117
3F-3	Orientations of pUC8 subclones	118
3G-1	Comparison of cDNA and Funatsu sequences	121

Continued.... 181

Continued....

List of figures (continued)

<u>Fig</u>	<u>Title</u>	<u>Page</u>
3G-2	Map of plasmid pRCL17	124
3G-3	Junction region in pRCL17	125
3G-4	Extent of M13 clone libraries	127
3G-5	Maxam & Gilbert strategy: clone pRCL6	131
3G-6	Maxam & Gilbert strategy: clone pRCL17	132
3G-7	Maxam & Gilbert strategy: clone pRCL52	133
3G-8	Maxam & Gilbert strategy: clone pRCL57	134
3H-1	Preprorizin cDNA sequence	136
3I-1	Preproagglutinin cDNA sequence	138
3J-1	Amino acid compositions and molecular weights	140
3J-2	Differences between ricin cDNA and Funatsu sequences	141
3J-3	Plot of 'genuine' differences between two ricins	143
3J-4	Amino acid sequences of prepro-lectins	150
3J-5	Amino acid differences between the lectins	151
3J-6	A chain hydropathy plots	154
3J-7	B chain hydropathy plots	155
3J-8	Hydropathy difference plots	156
3J-9	Signal sequence hydropathy plot	159
3J-10	Plant signal sequence cleavage sites	161
3J-11	Plant precursor protein processing sites	163
3K-1	Prepro-lectin cDNA sequences	166
3K-2	Plant mRNA base compositions	169
3K-3	Preprorizin dinucleotide frequencies	170
3K-4	Preprorizin codon usage	171
3K-5	Preprorizin cDNA 5' end: sequencing gel	174
3K-6	Plant mRNA 5' non-coding regions	178
3K-7	Plant mRNA initiation environments	181

Continued....

List of figures (continued)

<u>Fig</u>	<u>Title</u>	<u>Page</u>
3K-8	Plant mRNA initiation environments: selected sequences	182
3K-9	Plant mRNA termination environments	186
3K-10	Bases flanking plant termination codons	187
3L-1	Primer extension	194
3M-1	Northern blot of castor bean mRNA	196
3N-1	Oocyte injection products	198
4-1	Homology of conA and favin	203
4-2	B chain halves: Funatsu data	204
4-3	B chain halves: Ricin cDNA data	205
4-4	B chain halves: Ricin hydropathy plots	207
4-5	B chain halves: Agglutinin cDNA data	208
4-6	Putative B chain signal sequence	210

ABBREVIATIONS

<b>Amino acids:</b>	Standard three-letter codes are used, along with one-letter codes according to the IUPAC-IUB Commission (Biochem J 113:1-4, 1969).
<b>b</b>	bases
<b>bp</b>	base pairs
<b>BCIG</b>	5-bromo-4-chloro-indolyl-beta-galactoside
<b>cDNA</b>	Complementary deoxyribonucleic acid
<b>ConA</b>	Concanavalin A
<b>CR</b>	Coding region
<b>Dal</b>	Daltons
<b>DMS</b>	Dimethyl sulphate
<b>DNA</b>	Deoxyribonucleic acid
<b>DNase</b>	Deoxyribonuclease
<b>dNTP</b>	Deoxyribonucleoside triphosphate (A = adenosine, T = thymidine, C = cytosine, G = guanosine)
<b>DTT</b>	Dithiothreitol
<b><u>E coli</u></b>	<b><u>Escherichia coli</u></b>
<b>EDTA</b>	Ethylenediaminetetraacetic acid
<b>EF-1, EF-2</b>	Elongation factors 1 and 2
<b>EMBL</b>	European Molecular Biology Laboratory
<b>EndoH</b>	Endo-N-acetylglucosaminidase H
<b>EtBr</b>	Ethidium bromide
<b>IPTG</b>	Isopropyl-beta-thiogalactoside
<b>kDal</b>	KiloDaltons
<b>LD<sub>50</sub></b>	Lethal dose for 50 % of animals
<b>M<sub>r</sub></b>	Apparent relative molecular weight
<b>NCR</b>	Non-coding region
<b>PAGE</b>	Polyacrylamide gel electrophoresis
<b>PMSF</b>	Phenylmethylsulphonylfluoride

ABBREVIATIONS

pRCL	Castor bean lectin cDNA clone in pBR322
RCA	<u>Ricinus communis</u> agglutinin
RNA	Ribonucleic acid
RNase	Ribonuclease
SDS	Sodium dodecyl sulphate
SSC	Standard saline citrate
TCA	Trichloroacetic acid
SΔQ	See reference 62. The amino acid compositions of two proteins are compared, after correction for different numbers of residues, by subtracting the number of residues of each amino acid in one sequence from its counterpart in the other sequence, squaring the result, and summing all such results.

... (7) has collected descriptions of over 100 cases of ...  
... in man, though by far the most noteworthy case ...  
... of Georgi Vachon, a Bulgarian physician who was ...  
... by the injection of an extract of a ...  
... aggregated ... in 1978, and ...  
... The possibility of ricin poisoning ...  
... is followed by the granting of a patent in the US for a ...

CHAPTER 1: INTRODUCTION

1A Preamble

The endosperm cells of the castor bean plant Ricinus communis contain protein bodies in which a variety of proteins are stored prior to germination. These include the nitrogen-rich albumins, the 11 S globulins, and two lectins: the toxic ricin, and the related agglutinin RCA (Ricinus communis agglutinin) (1,2).

The toxic and medical actions of the beans have been known since ancient times, and were used in classical Greek and Sanskrit medicine (3); the oil is noted for its purgative effects. More recently, attention has focussed on the anti-tumour properties of ricin (4), and as recently as 1984 a report was published describing clinical trials using ricin in the treatment of terminal cancer (5). However, since ricin is a lectin, and consequently displays relatively non-specific binding to cell surfaces, most cancer work is now directed towards the exploitation of conjugates between ricin and antibodies. It is hoped that such conjugates will have specificity for tumour cells, and will represent the ultimate development of Ehrlich's 'magic bullet' concept (6).

Balint (7) has collected descriptions of over 700 cases of ricin poisoning in man, though by far the most notorious case is that of Georgi Markov, a Bulgarian expatriate who was assassinated by the injection from an umbrella of a ricin-impregnated metal ball in 1978, near Waterloo Bridge (8). The possibility of ricin poisoning on a vastly larger scale is indicated by the granting of a patent to the US Army for a

method of purifying very large quantities of ricin from castor beans (9). All work on ricin in Bulgaria is apparently classified (10).

Considerations of such possibilities have led to controversy regarding more contemporary approaches to the study of ricin and other toxic proteins: a letter published in Nature (11) advocated the outright banning of cloning of toxin nucleotide sequences, though a later correspondent pointed out that large supplies of diphtheria toxin are available without cloning (12). It is of note that the anti-cloning letter was from an employee of the Cetus Immune Corporation of California - eight months later, Cetus published the sequence of cloned diphtheria toxin (13). This latter sequence has been expressed in E coli (14,15,16), and a number of other toxin DNA sequences have also been cloned. Examples are bee venom melittin (17), cholera toxin (18,19), E coli heat-labile enterotoxin (20,21) and Pseudomonas exotoxin (22). Cetus hold a patent on the E coli toxin, and claim expression (23).

The extreme toxicity of ricin is indicated by its LD<sub>50</sub> values: one estimate gave 12 µg per kg in mice, while more recent determinations give 2.7 µg per kg in mice and 1.75 µg per kg in dogs (see ref 5). Fodstad et al (5) believe that the maximum dose tolerable by man is some 23 µg per sq m, given intravenously. Doses of 1 ng/ml of ricin are adequate to inhibit the activity of cell-free protein synthesising systems, and it is believed that a single molecule can kill a cell (24).

The castor bean plant is a native of tropical Africa, but has been introduced and naturalised in many other climates.

European varieties tend to grow as small bushes - often a bare stem carries just a few leaves at the top - and flower only with difficulty. Tropical specimens can grow rapidly to a height of 30 - 40 feet (25): it is sometimes named the 'miracle tree', one of which was used by God to teach Jonah the virtue of mercy (26). Etymologically, the name Ricinus probably derives from the Hebrew kikar, which means 'round', via the Greek kikinon (3).

A brief history of scientific work on ricin follows, and is condensed from Balint's thorough review (7).

The proteinaceous nature of the toxic component of castor beans was noted in 1878 by Ritthausen, and confirmed in 1887 by Dixon. In that year, Stillmark named this component ricin. The first detailed description of human ricin poisoning was published in 1899 by Muller. The resistance of ricin to proteolytic enzymes was noted around the turn of the century, and was long thought to be related to the protease activity attributed to it - as late as 1973, Funatsu's group described experiments demonstrating this function (27). Ricin is no longer believed to have proteolytic activity (eg ref 28).

These early experiments were performed using crude extracts of castor beans; better preparations were obtained in 1905 by Osborne. In 1913 Kobert improved the methodology by the introduction of ammonium sulphate precipitation, and in 1923 Karrer introduced inorganic adsorbents. Affinity chromatography on sepharose, exploiting the galactose specificity of ricin, was first used in 1965, by Dirheimer. This is now the most widely used method of preparation.

The toxicity of ricin was for many years attributed to a haemagglutinating activity, but with the advent of more discriminating fractionation techniques, the toxic and agglutinating activities were assigned to two different proteins, referred to as ricin and Ricinus communis agglutinin respectively. At the same time, the existence of multiple forms of the lectins became apparent - for example, Schone, in 1958, obtained five bands on paper electrophoretograms. Most contemporary workers report two ricins and two agglutinins (eg refs 29,30), though occasionally more are reported (31). It was shown in the early 1970s that both the toxin and the agglutinin are composed of two dissimilar types of subunit - the agglutinin being essentially a dimer of ricin. Although the A chains and the B chains of the two proteins cross-react immunologically (though A chains and B sera do not react, and vice versa), they must clearly be different, in order to account for the differences in structure, toxicity and sugar specificity. The toxic and sugar-binding activities of the proteins have been assigned to different subunits, the A chains being toxic, while the B chains have lectin activity. Although the agglutinin A chain is toxic to cell-free systems, its activity is only a fraction of that of the ricin A chain, and the whole protein is virtually non-toxic to animals. It is likely that this is due to haemagglutination brought about only by the agglutinin, and a failure of its A chain to penetrate cell membranes.

The sequences of both chains of ricin have been determined, by Funatsu's group in the late 1970s (32 - 36), though only partial sequences are available for the agglutinin (31). N-terminal sequences of the subunits of both proteins show almost complete identity.

The toxicity of ricin was for many years attributed to a haemagglutinating activity, but with the advent of more discriminating fractionation techniques, the toxic and agglutinating activities were assigned to two different proteins, referred to as ricin and Ricinus communis agglutinin respectively. At the same time, the existence of multiple forms of the lectins became apparent - for example, Schone, in 1958, obtained five bands on paper electrophoretograms. Most contemporary workers report two ricins and two agglutinins (eg refs 29,30), though occasionally more are reported (31). It was shown in the early 1970s that both the toxin and the agglutinin are composed of two dissimilar types of subunit - the agglutinin being essentially a dimer of ricin. Although the A chains and the B chains of the two proteins cross-react immunologically (though A chains and B sera do not react, and vice versa), they must clearly be different, in order to account for the differences in structure, toxicity and sugar specificity. The toxic and sugar-binding activities of the proteins have been assigned to different subunits, the A chains being toxic, while the B chains have lectin activity. Although the agglutinin A chain is toxic to cell-free systems, its activity is only a fraction of that of the ricin A chain, and the whole protein is virtually non-toxic to animals. It is likely that this is due to haemagglutination brought about only by the agglutinin, and a failure of its A chain to penetrate cell membranes.

The sequences of both chains of ricin have been determined, by Funatsu's group in the late 1970s (32 - 36), though only partial sequences are available for the agglutinin (31). N-terminal sequences of the subunits of both proteins show almost complete identity.

It is striking that the castor bean lectins are related to a variety of other plant proteins, in terms of the structural distribution of the toxic and sugar-binding activities, and the mechanism of toxicity. For example the jequirity bean (Abrus precatorius) contains a toxin, abrin, with the same sugar specificity and toxicity as ricin, along with an agglutinin which is related to abrin in the same way as the castor bean agglutinin is related to ricin (3). A number of other examples are known, and several plants contain proteins which resemble the A chains of abrin and ricin, but lack sugar-binding properties (and B chains). For a recent review of plant toxin proteins, see reference (37), in which all such proteins known up to 1982 are listed. More such proteins are regularly discovered, for example volkensin, found in June 1984 (38).

It is curious that all the highly toxic two-chain lectins have galactose as their sugar specificity - including ricin, abrin, modeccin, viscumin and volkensin. This may imply that only cell surface glycoproteins and glycolipids bearing terminal galactose residues are able to transport such molecules into the interior of the cell.

A number of bacteria also contain toxic proteins which separate their cell-binding and enzymatic activities onto different polypeptide chains - these include diphtheria toxin (39), E coli heat-labile enterotoxin (20), Pseudomonas exotoxin A (40), pertussis toxin (41), cholera toxin (42) and tetanus toxin (43). Of these, diphtheria toxin and exotoxin A interfere with protein synthesis; all catalyse ADP-ribosylation of cell components.

The synthesis and intracellular assembly of the castor bean lectins have not been widely studied. Roberts & Lord (44)

proposed that the A and B chains have separate precursors, on the basis of detecting two precursors with anti-agglutinin sera. However, Butterworth & Lord (45) compared the immunological properties and enzymatic cleavage products of both precursors with the mature lectin subunits, and found that both chains are contained within a single precursor. The other, smaller, precursor was identified as a pre-albumin. This is believed to copurify with the lectins (in commercial preparations as well as those carried out in the laboratory) - and since it is an extremely allergenic protein, antisera raised against the agglutinin contain high anti-albumin titres.

A fairly large number of plant proteins appear to be synthesised by comparable mechanisms, including a variety of lectins and storage proteins. Examples are also known in animals, such as insulin, and more are emerging.

Interest in ricin continues to flourish, not only because of its own complex and perplexing nature, but also because of its great potential for immunotoxins. This report describes the application of recombinant DNA methods to the castor bean lectins: cDNA clones for both are described, along with their nucleotide sequences, and the future uses of these clones will be discussed.

Note that a number of reviews of the extensive literature on ricin are available, including reference 46 (1982), other articles in the same volume, reference 7 (1974), reference 3 (1976), and reference 47 (1984), and reference 48 (1979).

### 1B Isolation and Structure of the Lectins

The lectins are generally isolated and purified by adsorption chromatography on columns containing materials such as CM-cellulose followed by affinity chromatography on sepharose. The properties of the proteins in the various fractions may be assayed by haemagglutination, toxicity and gel electrophoresis (30,31,45,49). Multiple forms of the lectins are usually reported, the properties and number of which appear to depend on the variety of beans in use. Although at least 21 varieties of the castor oil plant are recognised (7), very few workers specify which they use. At best, the beans are described as 'small', grown in Texas or Japan, or as 'large', grown in Thailand (50). The beans used in this laboratory are probably a heterogeneous mixture, and they vary in size and patterning.

Funatsu's group reports one toxin (ricin D) from large beans (51) and two (ricins D and E) from small Japanese beans (52); Lin & Li (30) report two toxins and two agglutinins from small beans, and Olsnes et al (49) also find two forms of each lectin, though they do not specify bean size. Cawley et al (28,31) find three ricins and two agglutinins, again from undescribed beans.

Toxins and agglutinins may be distinguished by elution from CM-cellulose at different salt concentrations (49) or by successive elution from sepharose columns with N-acetylgalactosamine and galactose respectively (31,45). The molecular weights of the two lectins, as determined by sedimentation analysis, gel filtration and gel electrophoresis, generally fall within the ranges 54 - 65 kDal for ricins, and 110 - 140 kDal for agglutinins (30,31,45,49,53,54), though Nicolson points out (29) that estimates depend on the conditions used in the measurements.

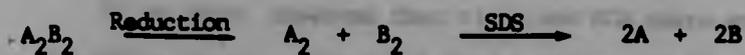
The molecular weight of ricin, calculated from the complete sequence data, is 62,057 Dal, including carbohydrate (33,35). There is reason to believe that this estimate is in error, since evidence to be presented indicates that the sugars are more complex than Funatsu supposed, and the sequences to be reported here imply that those of Funatsu may contain errors.

Funatsu investigated the similarities and differences between ricins D and E (52): in terms of molecular weight, N-terminal amino acids, toxicity and agglutinating activity, the two proteins were very similar, though they had different isoelectric points, amino acid compositions, affinities for sepharose, and toxicities for certain malignant cells.

It is likely that the castor bean lectins represent a family of related proteins, differing within and between varieties. Differences between the primary sequences, while probably small, must nonetheless account for the quite gross differences in quaternary structure of toxins and agglutinins. The existence of gene families in plants is well documented; for example there are believed to be over 100 genes for zain in Zea mays (55).

Boiling of the lectins in the presence of SDS reduces the size of the agglutinin to about that of ricin, while ricin itself is unchanged (28). Along with the evidence of the similarity of the subunits between the two, this supports the idea that the agglutinin contains two ricin-like heterodimers linked non-covalently. While the mechanisms which hold the two dimers together are unclear, it has been suggested that proton donation or sulphhydryl-catalysed disulphide exchange are possible mechanisms (28).

Reduction of the agglutinin prior to dissociation with SDS produces bands which correspond to dimers of two A chains, and of two B chains: the non-covalent forces holding the ricin-like heterodimers together may thus act both on the A chain half of the molecule, and on the B chain half (56). This is presented in schematic form:



Although studies on viscumin indicate that tetrameric forms comparable to Ricinus communis agglutinin form by aggregation of dimers at high concentration (57), such concentration-dependent agglutination of ricin is not observed to an extent sufficient to explain the structure of RCA (58).

Reduction of ricin with 2-mercaptoethanol or dithiothreitol, or oxidation with performic acid, results in the appearance of two types of subunit: the smaller is designated the A chain, and the larger is the B chain. The corresponding subunits of the agglutinin are designated the A' and B' chains. The molecular weights of the subunits as generally determined are shown below, along with those determined in this laboratory (45); all sizes are in kDal:

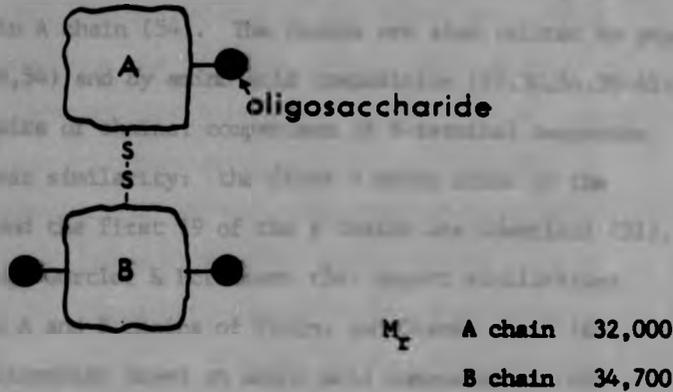
(M. Lord, personal communication). The structural model currently accepted for the two proteins using size data from references (45), for the major variants, is shown in fig. 1B-1.

	<u>Range</u>	<u>From Ref (45)</u>
Ricin A chain	29.5 - 33.0	32.0 (34.0)
Ricin B chain	32.0 - 34.7	34.7
RCA A' chain	27.5 - 33.0	32.0 (34.0)
<u>RCA B' chain</u>	<u>33.0 - 37.0</u>	<u>36.0</u>

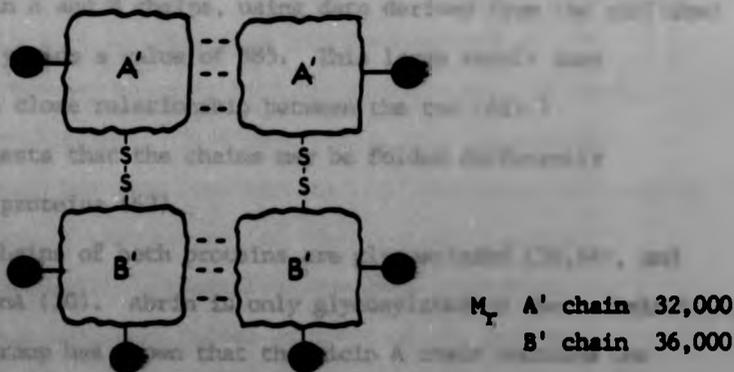
Figures in brackets correspond to ricin<sub>2</sub> and RCA<sub>2</sub>, the less abundant variants found in reference (45).

It is often observed that ricin and RCA share an A chain of similar length, while the B chains differ (29-31,45,52), though other workers have found the opposite (49,54). However, isoelectric points determined for the subunits of the two proteins found differences for both chains (31). When the B chains are of different lengths, the ricin chain is generally 1300 - 3000 Dal smaller than that of RCA. The significance of these differences may sometimes lie in the systems used for measurement (see reference 29), or may reflect heterogeneity between seeds. One group (31) suggested that the agglutinin B chain may have an extra 30 or so residues at the C-terminus (the N-terminus amino acids were identical), and deglycosylation of B chains with EndoH (45) indicated that the agglutinin B chain remained larger than the ricin B chain. However, the significance of this is unclear, since in vivo synthesis of the lectins in the presence of tunicamycin, which blocks N-glycosylation, produces only two bands. One of these corresponds to A chains, while the other corresponds to B chains, implying that they may be of very similar size (JM Lord, personal communication). The structural model currently accepted for the two proteins using size data from reference (45), for the major variants, is shown in fig 1B-1.

**Ricin**



**agglutinin**



**Fig 1B-1: Structures of ricin and RCA.**

$M_r$  values are derived from glycosylated chains.

The two castor bean lectins are related immunologically - that is, the two A chains cross-react, as do the two B chains (3), though sera specific to each protein have been made (59). The A' chain apparently lacks antigenic determinants present on the ricin A chain (54). The chains are also related by peptide mapping (29,54) and by amino acid composition (29,30,54,59-61). For both pairs of chains, comparison of N-terminal sequences reveals great similarity: the first 9 amino acids of the A chains, and the first 19 of the B chains are identical (31).

Although Gurtler & Horstmann (54) report similarities between the A and B chains of ricin, and Olsnes et al (61) claim relationships based on amino acid compositions, the A and B chains do not cross-react immunologically (3). (Calculation of SQ values from the amino acid compositions of the ricin A and B chains, using data derived from the published sequences, yields a value of 585. This large result does not imply a close relationship between the two (62).) Olsnes suggests that the chains may be folded differently in the two proteins (63).

Both chains of both proteins are glycosylated (30,54), and all bind conA (30). Abrin is only glycosylated on the B chain (46). Funatsu's group has shown that the ricin A chain contains one carbohydrate group, attached to residue number 10 (asparagine), while the B chain carries two carbohydrate groups, attached to asparagine residues 93 and 133 (33,35). The sequences around these sites are characteristic of N-glycosylation sites in proteins (64,65). The compositions of the carbohydrate moieties have been determined (35,61), and are consistent with the classical structures of N-linked oligosaccharides (65). More recent

work (JM Lord, personal communication) indicates that both lectins contain fucose: both A chains and the B chain of the agglutinin can be labelled in vivo with tritiated fucose, though the sugar is not incorporated into the ricin B chain. Although the detailed structures of the castor bean lectin oligosaccharides have not been elucidated, other examples of fucosylation indicate that this sugar is likely to be attached to the proximal N-acetylglucosamine unit of the classical N-linked oligosaccharide (65-67). The ricin B chain may contain xylose (B Foxwell, personal communication), as has been reported for Phaseolus vulgaris phytohaemagglutinin (68), and lima bean lectin (69). However, these latter two proteins contain fucose as well as xylose.

It is noteworthy that fucosylation generally results in insensitivity to cleavage by EndoH, as is the case with pineapple stem bromelain (67) and Phaseolus phytohaemagglutinin (70), but this appears not to occur with the RCA B' chain (JM Lord, personal communication). Such sensitivity of a fucosylated protein to EndoH has also been reported for rat  $\alpha_1$ -antitrypsin (66): in the presence of swainsonine (an inhibitor of lysosomal  $\alpha$ -mannosidase II), a partially processed product was obtained. This contained fucose attached to the proximal N-acetylglucosamine, and was EndoH-sensitive - the completed RCA B' chain may perhaps resemble this product.

The complete amino acid sequences of the A and B chains of ricin D have been determined by Funatsu's group (32-36), and are presented in fig 1B-2. Other workers have sequenced parts of the two chains, mainly at the N-termini. Bull et al (50) sequenced three cyanogen bromide fragments of the ricin A chain, covering

1 10 20 30 40  
I P P K Q Y P I I N F T T A G A T V S Y T N F I R A V R G R L T T G A D V R H  
carb  
41 50 60 70 80  
K I P V L P N R V G L P I N O R F I L V E L Q N H A E L S V T L A L S V T R A Y  
81 90 100 110 120  
V V G V R A G N S A Y P F P N D N O E D A E A I T H L P T D V Q N R Y T F A P S  
121 130 140 150 160  
E R Y D R L E Q L A G N L R E N I E L O N G P L E E A I S A L Y V Y S T G T Q  
161 170 180 190 200  
L P T L A R S P I I C I Q N I S E A A R F Q Y I E G E N R Y R I R V H R R S A P  
201 210 220 230 240  
D P S V I T L E N S W G R L S T A I Q E S N Q G A F A S P I Q L Q R D S K F S  
241 250 260 265  
V Y D V S I L L P I A N V Y R C A P P P S S Q F  
Ala-chain  
1 10 20 30 40  
A D V C M D P E P I V R I V R N G L C V N V R D G R F N H O H A I Q L M P C K  
41 50 60 70 80  
S N T B A N Q L T L K R D N T I R S N G K L T T Y G Y P S G V Y V N I Y D C R  
81 90 100 110 120  
T A A T T A D R E I W N N S T I I N P R S S L Y L A A T S O N S O T T L Y Q T  
carb  
121 130 140 150 160  
N I Y A V S Q G P L F T H N T Q P M V T T I V G L Y G L C L Q A N S G Q V V I E  
carb  
161 170 180 190 200  
D S C S E Y A E Q Q A L Y A S G N I P Q O R R O N C L Y S D S H I R E T V V  
201 210 220 230 240  
K I L S C P A S S G E R M H P K N D O T I L N L Y S G L V L D V R A S D P S L  
241 250 260  
K Q I I L Y P L W N D P N Q L I L P F

**Fig 1B-2**  
Amino acid sequences of the A and B chains of ricin D, as determined by Funatsu's group.

74 residues, one of which was not identified. Apart from this one, and one glutamine/glutamic acid uncertainty, the results were in complete accord with those of Funatsu's group. In the case of the B chain, however, a number of differences were noted. These will be discussed in detail in the Results and Discussion section, where it will be shown that the results of Bull et al are in closer agreement with the sequences deduced from the cDNA clones than with the Funatsu sequences. Bull et al point out that these differences are likely to represent heterogeneity between seed types - Funatsu uses large beans, whereas Bull et al use small beans. However, for one of the differences, they suggest that Funatsu's group mistakenly transposed two residues.

Cawley et al (31) obtained N-terminal sequences for both chains of ricin and of the agglutinin, and found that both were extremely similar:

A chain

Ricin (34)\* I F P K Q Y P I I N F T T A G A T V Q

RCA (31) I F P K Q Y P I I (I) F T Y A D A (T) V Q

B chain

Ricin (31) A D V Q M D P E P I V R I V G R N G L

RCA (31) A D V ? M D P E P I V R I V G R N G L

\*The A chain sequence of ricin is that of Funatsu, since Cawley et al determined only the first seven residues, which are identical to those of Funatsu.

Residues underlined in the B chain were not identified by Cawley et al, so the results of Funatsu's group have been inserted.

The sequence of the ricin A chain reveals that it contains two cysteine residues. Funatsu (34) has shown that cysteine - 257 is involved in the disulphide linkage with the B chain, the other cysteine not being involved in any disulphide bridge. This residue is apparently extremely inert, perhaps because of its very hydrophobic environment (34). The B chain contains nine cysteine residues, which were shown to participate in four intrachain disulphide bonds, and one interchain bond. Thus, cysteine - 4 of the B chain is linked to cysteine - 257 of the A chain (35). Fig 1B-3 shows the orientations of the intrachain bonds in the B chain.

The interchain bond is apparently more susceptible to reduction than the intrachain bonds (61), and has been examined in more detail by Lappi et al (71). They found that the bond required some 50 times as much 2-mercaptoethanol to reduce it in native ricin than in the SDS-denatured protein, suggesting that the bond is in the hydrophobic interior of the molecule. Dithioerythritol, normally a more efficient reducing agent, was less effective than 2-mercaptoethanol: it is a larger molecule, and is presumably unable to gain access to a buried disulphide bond. A preliminary X-ray analysis was unable to define the position of the bond unambiguously (72).

Good X-ray diffraction data for ricin are not yet available. Although ricin crystals were obtained as long ago as 1964 (73), these were too small for analysis. Larger crystals made more recently (53,74) were unstable on irradiation. Villafranca & Robertus in 1981 (72) obtained better results, finding that ricin is a compact molecule with two elongated subunits, roughly parallel.



The A chain had an asymmetrical wedge shape, tightly associated with the B chain at its wider end, the probable position of the disulphide bond. The B chain consisted of two similarly shaped domains, each with one lactose binding site, one of which was more highly occupied than the other. A prominent groove extended between the sites, implying an extended binding site for oligosaccharides. Although other workers, using circular dichroism spectrometry (75) claimed that a conformational change occurred on binding of lactose, which was weaker for ricin than for the agglutinin, Villafranca & Robertus found no significant difference between ricin alone and ricin with bound galactose. This will be discussed in more detail in section 1D.

The size of the ricin molecule determined by Villafranca & Robertus was  $73 \times 58 \times 35 \text{ \AA}$ ; this compares with a rounded shape with a surface of  $80 \times 90 \text{ \AA}$  seen under the electron microscope for viscumin (57). Dimeric forms of viscumin were also seen, with dimensions of  $175 \times 80 \times 60 \text{ \AA}$ , which were claimed to be comparable to those of Abrus agglutinin. However, another study, using X-ray diffraction and electron microscopy (76), found that Abrus agglutinin consisted of four similar domains grouped at the vertices of a roughly planar rhomboid. The average diameter of each domain was some  $25 \text{ \AA}$ :

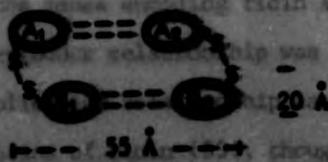


Fig reproduced from ref (76).  
Dimensions are approximate.

Similar results were obtained for RCA by Robertus & Oliver - these

be made for ricin and a variety of other proteins, including  
are unpublished, but are quoted in reference (76).  
the plant toxins from species other than Abrus, the single

Funatsu carried out a prediction of the secondary structures  
of the ricin chains (34,35) according to the rules of Chou and  
Fasman. They found that (taking averages of the A and B chains) (78).  
ricin was 16 -17 % alpha helix and 32 - 33 % beta sheet. Values  
determined by circular dichroism were 13 % and 51 % respectively (75).  
The authors of the latter report pointed out that the X-ray structure  
of conA also has a high beta content, as do circular dichroism  
results for lectins from two other plants. However, they also  
note that their data are subject to considerable inaccuracies  
due to the dimensions of pleated sheet structures, and the  
presence of aperiodic bends and kinks.

A computer search of Funatsu's ricin sequences has been  
performed (72), and revealed one region of possibly significant  
homology between the chains - for two regions of 20 amino acids,  
the two coding sequences could differ by a minimum of 17  
nucleotide changes. This will be considered in the light of the  
present results in the Results and Discussion section. The search  
also found that the ricin B chain is composed of two related domains,  
with several regions of close homology, including two pairs of  
disulphide bridges. It was suggested that the B chain arose in  
its present form by gene duplication: such relationships may also  
hold for the genes encoding ricin and RCA (31), though a  
precursor-product relationship was also mooted by the same authors.  
A gene duplication relationship has also been suggested for the  
A and B chains of ricin (59), though the lack of homology  
argues against this. Some evolutionary relationship may also exist  
between ricin and abrin genes (77), although Abrus is not  
taxonomically close to Ricinus (3). Similar propositions could

be made for ricin and a variety of other proteins, including the plant toxins from species other than Abrus, the single - chain ribosome inhibitors, such as galonin, and other lectins. One report suggests that two lectins from Momordica charantia may be homologous to ricin, starting at residue 9 of the A chain (78). Although a number of gaps must be included to show the homology to full effect, 12 out of 30 residues can be aligned (the 30 residues represent the N-terminal 27 which were sequenced, plus 3 gaps relative to the ricin A chain):

```
I: D V S F R L S G A D P R S - - - Y G M F I K D L R N A L P F
   | | | | | | | | | | | | | | | | | | | | | | | | | | | |
II: D V N F D L S T A T A K T - - - Y T K F I E D F R A T L P F
   | | | | | | | | | | | | | | | | | | | | | | | | | | | |
R:  I I N F T - - T A G A - T V Q S Y T N F I - - - R A V R G R
```

I - Momordica charantia lectin I.

II - Momordica charantia lectin II.

R - Ricin A chain from residue 9.

Lectins I and II are identified as an agglutinin and a toxin respectively (78); both are apparently dimers linked by disulphide bonds, all chains being approximately 26 kDal in size. SaQ values determined from the amino acid compositions do not indicate a close relationship with ricin (comparison of either Momordica lectin with ricin gives a value of 950, whereas comparison of the two Momordica lectins with each other gives 350, which may imply homology. Values were calculated as in ref 62; see List of Abbreviations, p vii).

The lack of immunological cross-reactivity between the castor bean lectins and these other proteins does not absolutely rule out the possibility of sequence relationships - phaseolin and conglycinin, which are related by sequence, do not cross-react (79).

On the other hand, broad structural and functional similarities do not guarantee evolutionary relationships: if a given function is required by a group of otherwise distantly related plants, each may evolve its own protein to fulfil that need. Such convergent evolution may indeed produce proteins which superficially appear to be related, but in sequence terms may be entirely distinct.

proteins of the same level, but evolutionary relationships are not necessarily indicated by the presence of similar amino acid sequences. It is possible that some amino acids may be developmentally conserved by selection of conserved amino acids (13,14).

In a similar report describing the synthesis of the carrier bean lectin, we suggested that the 4 and 8 chains are derived from separate precursors. In vivo labelling of endogenous lectin as well as in vitro translation of RNA, and then to provide the analysis of molecular weights of 4 and 8 kDa, which were obtained from electron-microscopic studies of the preparation of the precursor molecules in the lectin system yielded an average of a group of particles of 40 - 45 kDa. These were obtained by cleavage of the 45 kDa species by cleavage of a single asparagine and by glycosylation. Both apparent precursors were not evolutionarily conserved. On digestion of the larger group with trypsin, the multiple forms were replaced by a single band of 17 kDa, leading to the final sequence of 4 and 8 kDa in vivo. In the in vitro studies we demonstrated the cleavage of the larger precursor from the endogenous particles together with the associated elimination of the native sugar oligosaccharide in the protein part (fraction 45). The larger precursor was tentatively identified as a 3 chain precursor while the smaller one was identified as an 4 chain precursor - on the basis of amino acid composition (14).

### 1C Synthesis of the lectins

The lectins are synthesised during seed development, first appearing about 20 days after pollination (80), corresponding to the first stages of testa formation (81). It is not known whether the developmental regulation is mediated at the transcriptional or translational level, but translatable mRNA only appears at this time (81,82). Certain other storage proteins in plants have been shown to be developmentally regulated by alteration of transcriptional activity (83,84).

An initial report describing the synthesis of the castor bean lectins (44) indicated that the A and B chains are derived from separate precursors: in vivo labelling of endosperm tissue, as well as in vitro translation of mRNA, was shown to produce two proteins of molecular weights 33.5 kDal and 59 kDal, which were reactive with antisera directed against the agglutinin. Addition of dog pancreas membranes to the in vitro system resulted in the appearance of a group of proteins of 66 - 69 kDal. These were presumed to derive from the 59 kDal species by cleavage of a signal sequence and by glycosylation. Both apparent precursors were cotranslationally segregated. On digestion of the larger group with EndoH, the multiple forms were replaced by a single band of 57 kDal, implying that the signal sequence is some 2 kDal in size. In vivo pulse-chase experiments demonstrated the disappearance of the larger precursor from the endoplasmic reticulum fraction, with the concomitant accumulation of the mature subunit polypeptides in the protein body fraction (45). The larger precursor was tentatively identified as a B chain precursor, while the smaller one was identified as an A chain precursor - on the basis of size and immunoreactivity (44).

These identifications were not confirmed by any other means at this stage.

The anomalously high molecular weight of the putative B chain precursor led to further work: the A and B subunits of both lectins were purified and antibodies were raised against them (45). Both A and B chain sera recognised the 59 kDal precursor and its glycosylated derivatives, while neither reacted with the putative A chain precursor. This suggests that the larger molecule contains sequences corresponding to both chains, while the smaller one is a contaminant. To support this, partial proteolysis of the subunits with papain (45) showed that similar patterns were obtained with all A-type subunits, and that a comparison of the products obtained from the agglutinin A' chain and from the putative A chain precursor were grossly different. This latter protein was then identified by its immunoreactivity with antisera raised against castor bean 2 S albumins (45).

It is suggested that these albumins copurify in small quantities with the lectins unless very careful separations of the A and B chains are performed, and, as highly immunogenic proteins (85), generate disproportionately large amounts of antibodies in sera. The problem was also encountered using commercially prepared sera. The complete amino acid sequence of an 11 kDal castor bean storage protein has been determined (86), and its amino acid content strongly suggests that it is a member of the 2 S albumin group. It is a heterodimer of 4 kDal and 7 kDal subunits linked by a disulphide bridge: if it is a product of the 32 kDal precursor, considerable post-translational processing must occur.

The pulse-chase studies already mentioned indicate that the castor bean lectins are synthesised on the endoplasmic reticulum membrane; in support of this is the exclusive location there of the enzymes responsible for glycosylation (44). Although cereal protein bodies often have ribosomes attached to them (87,88), this is not observed in other plants (89), and is unlikely to be the case in castor beans.

It has not been demonstrated that the lectins pass through the Golgi apparatus, though their glycosylated nature implies that they do - this organelle is characteristically the site of modification of core sugars (90,91). Similarly, the mode of transport of the lectins is ill-defined. Sucrose density gradient fractionation of castor bean endosperm yields a smooth membrane fraction distinct from the endoplasmic reticulum (JM Lord, personal communication), which may represent transport vesicles. Phaseolus vulgaris lectin has been shown to pass from the Golgi apparatus to the protein bodies in 'dense vesicles' (92).

A number of other heterodimeric plant proteins are known to be synthesised as single-chain precursors, including pea legumin (93), pea seed lectin (94), rapeseed napin (95), and others. Examples are also known in animals, the best known of which is insulin (96), and a number of intestinal microvillar enzymes are assembled by cleavage of a single precursor (97).

## 1D Biological functions of ricin and RCA

### 1D1 General features

As will be discussed, the two chains of the lectin molecules are responsible for different facets of the function of the whole molecule. Thus, the B chain functions as a lectin, attaching the dimer to cell surfaces and, in the case of ricin, promotes the transfer of the A chain to the cytosol. The A chain of ricin then enzymatically inactivates ribosomes, leading to cell death. The agglutinin A' chain is also capable of inactivating ribosomes, but since it does not gain access to the cell interior, this is presumably not biologically important.

Many plants contain toxic proteins which, like ricin, act by inhibition of protein synthesis. A recent review (37) shows that they can be grouped into two classes. First, the ricin-like proteins, which have two chains, and secondly the single-chain proteins, in which the chain is analogous to the ricin A chain. The range of plants containing one or other class of protein is wide, including cereals, legumes, Euphorbiaceae, Solanaceae, etc. They are present in various parts of the plants, including leaves, seeds and stems, though in most cases information on the subcellular localisation is sadly lacking. In castor beans, the toxin and the agglutinin are sequestered in protein bodies (1,80), and a similar situation is likely to occur with the other proteins - in order to protect the cell from its own toxins. However, as will be discussed, the sensitivity of a given cell to its own ribosome inhibitors is a controversial point.

The mechanisms of action have in most cases not been determined, but it is probable that most will be very similar to ricin, abrin and modeccin, the best characterised of the toxins. For example, one single chain ribosome inhibitor, gelonin, is known to inhibit ribosomes by the same mechanism as ricin (98).

Although one study (99) found that of 27 lectins, only 6 were highly toxic (including one from the roe of the fish Rutilus rutilus), most lectins are considered toxic in high concentrations (100), and it has been suggested that toxicity should be considered a general property of lectins (99).

A variety of other organisms contain toxic proteins which, like the plant toxins, display structural segregation of toxic and binding functions. Most notable among these are bacteria (for example, diphtheria toxin, cholera toxin, pertussis toxin), though others have comparable toxins.

Historically, ricin was long considered to be a proteolytic enzyme - on this basis, Funatsu (27,101) suggested a role in the degradation of storage protein during germination. This may now be discounted as ricin is no longer considered to be a protease (eg 28).

The most obvious possible biological role of ricin is as a protective agent, preventing predation either by grazing animals, or by burrowing insects, and against infections (see reference 102 for a discussion of plant defence). However, the large quantities present (103) might argue for a role as a storage protein. Of course, there is no reason why, once a plant's basic storage requirements are fulfilled by other appropriate proteins, other useful proteins should not be co-opted for this purpose.

### 1D2 Properties of the A chains

The A chains are enzymatic inhibitors of protein synthesis, acting on the 60 S subunit of eukaryotic ribosomes. Prokaryotic ribosomes are highly resistant (104), though one study found that a partial tryptic digest of ricin inhibited a cell-free system from E coli (105). Animal ribosomes are more sensitive than those from plants: the concentration of ricin required to inhibit plant ribosomes is some 23,000 times that needed for animal ribosomes (106).

Castor bean ribosomes have been reported to be less sensitive to ricin than those of other plants (107): resistance of plant ribosomes to their own toxins or single-chain inhibitors has been observed in other systems (108,109). However, one study (106) found that castor bean ribosomes were as sensitive as other plant ribosomes, but Stirpe (108) points out that in these experiments soluble factors were provided by a wheatgerm extract, which might contain tritin - the single chain inhibitor of this tissue. Resistance of plant ribosomes to toxins from the same plant may well be a general feature.

Reports of sensitivity of mitochondria to ricin are in conflict (110,111): the general resistance of prokaryotic ribosomes suggests that they should also be resistant. Nuclear protein synthesis also appears to be resistant to ricin, though since this system is relatively ill-defined, its significance is unclear (111). One report (112) suggests that protozoan ribosomes may be resistant.

Although ricin is notoriously the toxic lectin from castor beans, the A chain of the agglutinin is also toxic to cell-free systems: Harley & Beevers (106) found that in plant systems 50 % inhibition of protein synthesis required some 5 - 10 times the amount of RCA A' chain as was needed with ricin A chain. Other workers have obtained similar results (31,59).

It is believed that the ricin A chain can inactivate ribosomes at a rate of some 1,500 per minute per A chain (113), confirming the catalytic nature of its action. The effect can be stopped at any time by the addition of anti - A-chain serum. The extreme toxicity of ricin, whereby a single molecule can, under appropriate conditions, kill a cell (24), may be partly accounted for by the probability that only one or two ribosomes per polysome may need to be affected. Once these have been inactivated, they would stop migrating along the mRNA, thus blocking all other ribosomes in the polysome (114).

The involvement of the 60 S subunit has been indicated by reconstitution experiments: ricin-treated ribosomes mixed with untreated soluble factors (EF-1 and EF-2) did not support protein synthesis, while the reverse experiment produced a competent system (115). Similarly, when 40 S and 60 S subunits were separately treated and reconstituted with untreated partners, proteins were synthesised except when ricin-treated 60 S subunits were involved (115,116). Also, subunits from ricin-treated cells were mixed with those from untreated cells: 40 S subunits were active from treated cells, but 60 S subunits were not (117). This latter is direct evidence that the 60 S subunit is the target of ricin in vivo as well as in vitro.

There is evidence that ricin inhibits the initiation phase of protein synthesis: a reduction in the amount of mRNA associated with polysomes was observed after treatment of cell-free systems with abrin (which acts in the same way as ricin) (118). The results indicated inhibition of binding of the 40 S subunit to the 60 S subunit in the formation of the 80 S initiation complex.

More evidence exists for the effect of ricin on elongation. Addition of A chain to cell-free systems, in the presence of aurin tricarboxylic acid (an inhibitor of initiation) produced no change in the polysome profile (119,120) - thus, ricin does not result in premature termination of protein synthesis by polysome degradation. However, polysomes are eventually broken down in vivo (46).

Ricin appears to inhibit the EF-1 - catalysed ability of the ribosome to hydrolyse GTP (115), and decreases the binding of aminoacyl-tRNA (121,48). Some authors have not observed this (122,123) - it has been suggested that this reflects excessively high concentrations of EF-1 in their reactions, protecting the ribosomes from ricin (46).

Peptidyltransferase activity, also associated with the 60 S subunit, is unaffected, as is evidenced by the failure of ricin to decrease the incorporation of puromycin (115,122). This antibiotic binds to the ribosomal A site, and the growing polypeptide chain is transferred onto it by peptidyltransferase (123). Since the resulting puromycyl-peptide is no longer a substrate for translocation to the P site, it detaches from the ribosome.

Many reports point to the translocation step as the major site of ricin action. Binding of EF-2 to the ribosome is decreased (122,124), and GTP, GDP and their non-hydrolysable analogues are also prevented from binding (125-128). Again, the concentration of EF-2 appears to be important, possibly because prebound EF-2 protects the site of action against ricin, or because modification of the ribosome does not completely abolish the affinity of the ribosome for EF-2, merely diminishing it. Prebound aminoacyl-tRNA also protects against ricin (see 46). The incompleteness of the inhibition of EF-2 binding, along with the partial reversibility of ricin-mediated inactivation by high concentrations of magnesium ions (118,129) or high EF-2 concentrations (130) has been interpreted to suggest that in ricin-treated ribosomes, some functional groups are spatially disturbed, and that under the conditions described, these changes can be reversed (46).

The exact site on the 60 S subunit which is affected by ricin is unknown. The effect on the EF-2 - dependent GTPase activity led one group to examine the effects of ricin on a group of ribosomal proteins as follows (131). Prokaryotic GTPase is associated with proteins L7 and L12, and these can be selectively removed by washing with ammonium chloride and ethanol. Untreated and treated eukaryotic ribosomes were thus washed, and assayed by reconstitution: the proteins which were removed were able to complement untreated ribosomal cores, but not treated cores. Analysis of the proteins removed indicated that at least 10 were involved, and no differences were detected in those from treated ribosomes. This is not unexpected, since the effect appears to be on the ribosomal cores, and it was not shown that they included the EF-2 - stimulated GTPase activity.

Cleavage of ribosomal RNA has been suggested by analogy with the action of colicin E3 (104) and with that of alpha-sarcin (132). Thus, an 8 S complex was released from the 60 S subunit by washing with EDTA, and this contained the 5 S rRNA and one protein. Neither of these species was decreased in size by ricin, even though the GTPase and ATPase activities associated with the 8 S complex were both inhibited (115). Houston (130) searched for dephosphorylation of the major phosphoprotein of the 60 S subunit, but failed to detect this. It has also been suggested that, since ricin A chain can act in simple buffer solutions without added cofactors, it may function as a hydrolytic enzyme (46). In contrast, an elongation inhibitor from wheatgerm requires added ATP and tRNA (133).

The N-terminal 18 amino acids were removed from the ricin A chain with nagarase, which had no effect on the toxicity (134). Since the carbohydrate of the A chain is attached within this region, it is clear that this is not involved in the toxicity. Peptides obtained from partial hydrolysis have also been shown to be toxic in cell-free systems, though insufficient characterisation was carried out to localise the active site upon the A chain (see 46).

In summary, although the broad location of the ricin A chain target is known, the details are proving elusive. It is possible that ricin affects a ribosomal component in a particularly subtle way - and that the component is as yet relatively poorly characterised.

1D3 Properties of the B chains

The B chains of both lectins confer upon them the ability to bind to the surfaces of a wide variety of cell types (see 46). This binding is prevented by the presence of lactose or galactose, as is discussed below. Thus, binding is to the oligosaccharide moieties of glycoproteins and glycolipids - treatment of cells with neuraminidase, which removes terminal sialic acid, thus revealing subterminal sugars, often increases the number of binding sites (46).

The sugar specificities of both lectins are similar in that both bind terminal galactose residues - for example galactose, lactose and melibiose. However, the ricin B chain is also able to bind N-acetylgalactosamine (29,49). The agglutinin binds fucose more strongly than does ricin (49). The importance of the terminal location of the galactose is indicated by the stronger binding of galactose than of lactose (135). Ricin equally binds alpha- and beta-linked sugars, but the agglutinin binds beta-linked sugars more strongly (29). The inhibition of agglutination by both lectins by rhamnose has been interpreted to imply that the important attribute of the binding sugars is the arrangement of the hydroxyl groups at the C-1, C-2 and C-4 positions (numbered as in galactose) (29). However, it has been suggested that rhamnose binds at a site other than the galactose-binding site, because its effect on the circular dichroism spectrum was grossly different from those of other hapten sugars (75).

The biologically meaningful targets of the lectins, as far as binding to cell surfaces is concerned, are oligosaccharides linked either to proteins or to lipids. One study has investigated the binding of both lectins to a variety of glycopeptides (136). The agglutinin bound dibranched oligosaccharides more

tightly than did ricin, while the association constants for tribranched sugars were indistinguishable (all these sugars were attached to core mannose structures). In contrast, ricin bound oligosaccharides consisting of galactosyl-N-acetylgalactosamine - seryl moieties some 2 - 10 fold better than RCA, and the affinity depended on the spatial organisation of the disaccharides along the peptide backbone. This report also noted that the addition of sialic acid residues to the terminal sugars generally reduced binding, while removal of terminal galactose, leaving terminal N-acetylglucosamine, abolished binding. Only ricin was able to bind to structures having terminal N-acetylgalactosamine. Thus, the results obtained with simple sugars are consistent with those obtained with biologically more significant haptens.

The number of saccharide binding sites on each B chain has been the subject of some disagreement: equilibrium dialysis seems usually to point to one site per chain (49,137), but the X-ray structure (72) implies two binding sites, one "more highly occupied" than the other. A number of workers (eg 60) point out that since ricin is able to agglutinate cells, albeit weakly, it should have two sites. Funatsu (60) also considers the existence of two sugar specificities for ricin (that is, for galactose and for N-acetylgalactosamine) to imply two sites. Perhaps in the case of the agglutinin, the single observed specificity indicates that two sites exist, and are identical. A study of sugar binding by microcalorimetry (135) found that there may be one or two sites - if there is only one, then it is an extended one, able to bind more than one sugar. This is consistent with the X-ray structure with its two sites linked by

a prominent groove (72).

The location of the sites on the B chains have not yet been determined; although the duplicate nature of the B chain (72) might indicate conservation of sequences surrounding the sites, it is just as likely that the sites are formed by the apposition of different parts of the polypeptide chain, so that conserved residues are not necessarily reflected in conserved sequences. Attempts to identify the binding site have involved the modification of specified amino acids in the chain, followed by assay of sugar binding. In one study, modification of charged amino groups, and (separately) modification of tyrosine residues, resulted in the inhibition of agglutination by RCA (138); the tyrosine results have been repeated for ricin (75,139). Analysis of these results implicate two lysine residues and one tyrosine. Other modifications, to tryptophan, arginine and carboxy-group - containing residues, had no effect on binding (138).

The involvement of the sugar moieties attached to the B chains in their binding functions have been ruled out, by demonstration of the ability of RCA with conA attached to its oligosaccharides to bind lactose (137), and oxidation of mannose residues with periodate did not affect binding activity (140).

The binding of sugars to the lectins appears to cause conformational changes: alterations in the circular dichroism spectra were reported (75), indicating that binding resulted in the alteration of the environment mainly of tyrosine residues exposed at the molecule's surface. These changes were more marked in RCA than in ricin. In the case of the Abrus lectins,

galactose caused changes involving the burying of tryptophan residues which were otherwise partially exposed (141).

In the castor bean lectins, tryptophan residues were more buried to start with, and no change was seen. Conformational changes on sugar binding have been observed with several other lectins (142).

Spectral and immunological evidence indicates that release of the A chain from the B chain, at least in abrin, results in a conformational change in the A chain (143). Such a change might facilitate the interaction of either the A or the B chain, or both, with membrane components, and could be an obligate step in the uptake of the A chain. Lappi et al (71) suggested that a conformational change might occur prior to dissociation of the chains, that is, on binding to cell surfaces, which might be required to allow reduction of the otherwise quite stable interchain disulphide bond.

#### 1D4 Entry of ricin into cells

It is well documented that the onset of symptoms of ricin poisoning is delayed by several hours (3,7); a lag phase, albeit much shorter, is also observed in cultured cells, though not in cell-free systems (144). Much evidence indicates that the lag is associated with the transfer of the A chain from the surrounding medium to the cytosol - ie the site of action. The process is divided into three distinct stages: binding to the cell surface, internalisation, and crossing of a cell membrane.

#### Binding to cell surfaces

This is a rapid process: even at 4°C extensive binding of ferritin-labelled ricin was observed within 10 minutes (145,146). Similarly, at 20°C (at which the whole lag phase takes about 6 hours), anti - B chain serum was able to protect cells for only 10 - 15 minutes (144,147), after which free toxin was no longer accessible to the serum. However, the continued sensitivity to anti - A chain serum indicates that the bound toxin remained at the cell surface for a further period after binding.

Similarly, lactose added to ricin-treated cells can only displace the toxin in the early phases (146) - that is, while the toxin is present at the cell surface. Most of the lag phase is thus involved with steps subsequent to binding. These experiments also point out the importance of cell surface attachment, in that non-bound or displaced toxin has no effect on the cells.

A number of studies have shown the extent and affinity of binding to several cell types (see 46 for details and table). HeLa cells had some  $3 \times 10^7$  sites per cell, and the affinity

constants indicated that ricin bound to cells more strongly than to free lactose - note that blocking of binding with lactose is generally done at concentrations of 100 - 200 mM.

Quantitative analysis of binding demonstrated that all binding sites have the same affinity, and that the affinity of free B chain is comparable to that of whole ricin. In fact, isolated B chains are able to bind to cells and enter in the same way as intact ricin (148). Prebound B chain is also able to promote the uptake of subsequently added A chain, as long as the latter is added before the B chains are internalised (148). Free A chain is not bound and does not enter cells in the absence of B chain.

Although the receptor for ricin clearly involves oligosaccharides with terminal galactose residues, the functionally relevant binding molecule has not yet been identified. Ricin can bind to glycolipids such as GM<sub>1</sub> (149,150), the association constants being similar (151) or slightly less (46) than those for binding to HeLa cells. The importance of GM<sub>1</sub> was suggested because it is almost certainly the receptor for other membrane-penetrating toxins (eg cholera toxin, ref 152), but since these enter by a mechanism different from that of ricin, their relevance must be treated with great caution.

Most data show that ricin binds to membrane glycoproteins. There appears to be little consistency in the number and sizes of ricin-binding proteins between different cell types (see ref 46), and, as the lectin will bind any structure with terminal galactose, the identification of the relevant species is likely to be extremely difficult. This difficulty is aggravated by the probability that, although only one molecule is needed for complete toxicity, it appears that several thousand must bind

to the cell to ensure entry of that one molecule (24).

#### Internalisation

This second part of the lag phase is characterised by the inaccessibility of bound ricin to antisera (144,147) or to lactose (146): the toxin is presumably sequestered somewhere within the cell. Several groups have followed the fate of ricin by labelling it with ferritin, horseradish peroxidase or radioactive iodine. Nicolson (145,146) found that after binding randomly over the cell surface, ricin clustered, and was then endocytosed into vesicles, subsequently appearing in the cytosol. Gonatas (153-155) found that after internalisation by endocytosis, horseradish peroxidase - labelled ricin accumulated in the GERL system of membranes ("Golgi-Endoplasmic Reticulum-Lysosome"). The ricin eventually found its way to the trans face of the Golgi apparatus. Some was recycled to the exterior of the cells, as might be expected from the receptor- and membrane-recycling roles of the GERL system. These results are strongly suggestive of internalisation by receptor-mediated endocytosis, but differ from the classical version in that other proteins, such as epidermal growth factor and low density lipoprotein enter very rapidly via coated vesicles and are transferred to lysosomes, the receptors being recycled (155). Most studies do not report ricin in lysosomes.

The importance of endocytosis is further indicated by the resistance to ricin of cells deficient in this process, either by mutation (156) or by chemical blockade (157). Reticulocytes, which bind large quantities of ricin, but display very little endocytosis, are resistant to ricin (144).

Thus, ricin enters cells by bulk endocytosis. However, the requirement to bind much more ricin than is needed for toxicity, implies that this is not the only process of internalisation - and that it is not the 'efficiency-limiting' step. Diphtheria toxin, which also requires endocytosis, is also taken up in bulk, but one study showed that only 300 or so molecules gain entry to the cytosol (See 159). Further evidence for another stage involves ricin-resistant cell lines which are endocytosis-competent, and have sensitive ribosomes (158). As will be described, the nature of this final step in the transport of ricin into the cell is that of a membrane-crossing process.

#### Membrane crossing

Early models for this stage were simple: Nicolson (145) suggested that it is mediated by the rupture of a minority of endocytotic vesicles. However, certain ricin-resistant cell lines, which endocytosed ricin, and had sensitive ribosomes, were sensitive to abrin and modeccin (160). If vesicle rupture is important, these cell lines would be unlikely, unless different toxins are selectively endocytosed into different classes of vesicle. It seems unlikely that different toxins with the same sugar specificity should use distinct vesicles, especially since other proteins which enter cells by endocytosis have been observed to share vesicles, for example  $\alpha_2$ -macroglobulin, insulin and epidermal growth factor (161).

A direct interaction of endocytosed ricin with internal membranes is possible, by analogy with diphtheria toxin. This toxin is endocytosed, and when it reaches an acidic compartment it is nicked to separate the toxic A fragment from the binding B fragment (116), and undergoes a conformational change such that

a previously concealed hydrophobic portion at the N-terminus of the B fragment is exposed (162-164). This inserts into the lipid bilayer, and may form a channel through which the toxic A fragment passes (165), perhaps driven by membrane potential (166).

Ricin can readily penetrate protein-free liposomes (167) as long as some galactose-containing receptor is present (such as ganglioside GM<sub>1</sub> (151)). Interactions of ricin and its isolated subunits indicate that all can insert into simple biological membranes (in this case Newcastle Disease Virus) (168), but only the interactions of whole toxin and B chain with the membrane were inhibited by lactose, and penetration was greatly enhanced by reduction of the toxin. Another study found that ricin A chain was able to bind to lipid vesicles much more efficiently than the A fragment of diphtheria toxin (169), but in this case most of the bound material was accessible to proteinase K. The same study reported a conjugate of the ricin A chain and the binding subunit of Wisteria floribunda lectin, which was toxic, while a similar conjugate involving diphtheria toxin fragment A was not. It was suggested that the hydrophobic C-terminus of the ricin A chain was responsible for the difference - note that the hydrophobic portion of diphtheria toxin is at the N-terminus of the B fragment, a nearly analogous position in terms of the unreduced forms of the toxins.

Thus it was proposed (169) that the ricin A chain crosses the membrane by insertion of this hydrophobic domain, which is followed by an as yet unclear stage. However, another conjugate, of diphtheria toxin fragment A with conA (170) was toxic - though in this case an analogous ricin A chain conjugate was not made. ConA may confer the membrane-crossing function to diphtheria toxin

fragment A which is supplied by ricin A chain in the Wisteria conjugates. Alternatively, the ricin A chain may not play a direct role in penetration at all.

A secondary membrane-crossing function may also be associated with the B chain: acetylation of ricin (139) prevents its binding to cells via galactose residues. Similar treatment of ricin conjugated with monophosphopentamannose produced a molecule which was non-toxic by the galactose-binding route, but which could enter fibroblasts via the lysosomal hydrolase pathway. However, the product was still non-toxic, and it was suggested that a second galactose-binding stage is required for uptake. Non-acetylated ricin - monophosphopentamannose is toxic, which validates its use (54). However, upon chemical blockade of the galactose-binding site of abrin, conjugates with antibodies were still toxic: this argues that the second B chain function may not involve the galactose-binding site (171). Also, antibody-ricin conjugates are toxic in the presence of lactose (171), so again, the second B chain function may not involve galactose binding. It is important that the nature of this second function be clarified, as immunotoxins made with ricin A chain are insufficiently toxic; development of conjugates using whole ricin, with the galactose binding site modified in some way, will depend very strongly on this A chain - uptake - promoting activity.

Ricin, unlike diphtheria toxin, does not require an acidic compartment for membrane penetration. Lysosmotropic amines (160,172) and certain carboxylic ionophores (157), all of which increase the vesicular and lysosomal pH, all stimulate ricin toxicity, whereas acidification of endocytotic vesicles by incubation of

cells at pH of 6 or less prevents toxicity (157). In contrast, under the latter conditions, diphtheria toxin is able to penetrate the plasma membrane directly (116). Perhaps an acid - mediated mechanism would be inappropriate for ricin, a priori, since protein bodies contain acid proteases (173).

Conformational change may nonetheless still be important, and has been observed on dissociation of A chains from B chains (61,143,174).

Another membrane-crossing mechanism has been proposed by Olsnes et al (175). They find that calcium deprivation decreases the sensitivity of cells to ricin. Thus, the penetration of membranes by ricin may be linked to calcium ion influx, which may in turn be regulated by vesicular pH or calcium ion concentration - membrane channels controlled by the latter have been shown to change in pore diameter in rat liver gap junction channels, whose diameters are reduced by calcium (176). It was suggested (160) that the ricin A chain might pass through such pores, perhaps in an extended conformation. In an attempt to substantiate this hypothesis, various other treatments were investigated for their effects on ricin sensitivity: the calcium ionophore A23187 was added, and this protected against abrin, but not against ricin (175). It was suggested that the ricin A chain can pass through either type of calcium channel, but that abrin A chain can only utilise natural channels. However, the addition of lanthanum ions, and of other lanthanides (all of which inhibit calcium influx), and iron ions (which do not), all protected against abrin and ricin. These may bind to calcium channels in such a way as to close them - the iron ions having a slighter effect than the lanthanides, resulting in a change in selectivity, rather than outright blockade. It is of note that lanthanum ions alter membrane fluidity and permeability in general -

and that they killed the cells after overnight incubation!

A mechanism of this type might explain the inefficiency of ricin entry, in that relatively few A chain molecules would be able to penetrate a channel designed for smaller species.

A number of retinoids also affect the uptake process, in that they protect cells against modeccin, and to a lesser extent against abrin and ricin. Some phorbol esters have an opposite effect. The effects appear to be on trans-membrane transport, but their mechanisms of action are unclear (160). Phorbol esters have been shown to stimulate the intracellular transport of proteins, and to stimulate protein secretion (177).

In summary, the means by which ricin and other toxins cross cell membranes is unclear. The selective resistance of certain cell lines, along with the different effects of various agents on the toxicities of different toxins implies that a number of routes are employed, which may have some steps, but not all, in common.

and that they killed the cells after overnight incubation!

A mechanism of this type might explain the inefficiency of ricin entry, in that relatively few A chain molecules would be able to penetrate a channel designed for smaller species.

A number of retinoids also affect the uptake process, in that they protect cells against modeccin, and to a lesser extent against abrin and ricin. Some phorbol esters have an opposite effect. The effects appear to be on trans-membrane transport, but their mechanisms of action are unclear (160). Phorbol esters have been shown to stimulate the intracellular transport of proteins, and to stimulate protein secretion (177).

In summary, the means by which ricin and other toxins cross cell membranes is unclear. The selective resistance of certain cell lines, along with the different effects of various agents on the toxicities of different toxins implies that a number of routes are employed, which may have some steps, but not all, in common.

1E Uses of castor bean lectins

The lectins have been exploited for a variety of purposes, including medical, scientific, and some more sinister uses. A selection of these are listed in fig 1E-1. Current interest is directed mainly towards cancer therapy, especially with regard to conjugates of toxins with tumour cell - specific antibodies.

The anti-cancer effects of abrin and ricin have been known for some time; Lin et al (4) reported dramatic effects with both proteins on Ehrlich ascites tumours in mice. A review of the subject is provided in reference 46, the authors of which point out that the plant toxins are as effective as many of the currently used alkaloids, and in some cases have fewer side effects. Most notably, combinations of sub-therapeutic doses of conventional drugs along with small doses of toxins often have synergistic effects on cancer cells, the effects on other cells being less than additive. For example, the combination of ricin and adriamycin increased the lifespan of animals given intracerebral injections of leukaemic cells, while either agent alone had no effect. It is suggested that the access of one of the drugs to the site of action is somehow facilitated by the presence of the other, but details are unknown.

More recently, phase I clinical trials were reported for abrin and ricin in terminal cancer patients (5). Ricin was given intravenously, the highest tolerable dose corresponding to 23 µg per sq m. Of 38 evaluable patients, one with lymphoma had a partial response (treatment with ricin was stopped because of a developing immune response. Continued therapy with abrin produced complete remission). In a further eight patients, advancement of disease was stopped. No toxic deaths occurred with ricin, but two were admitted with abrin.

The problem with using toxins as such, as with all other cancer drugs, lies in the non-specificity of their targets of action. As long ago as 1936, Paul Ehrlich (6) suggested that the specificity of antibodies could be used to direct drugs to specific locations within the body. Early attempts to do this involved joining conventional anti-cancer drugs to antibodies (see 187-189), and these continue. Although some success has been achieved with this approach, the stoichiometric nature of the actions of these drugs necessitates the necessity to deliver large amounts of such conjugates to the affected cells, with the result that the conjugates are often no as effective as the drugs given alone.

An obviously important factor in the use of any antibody-directed therapeutic agent is the availability of a suitably specific antibody. Much knowledge is currently being accumulated on the synthesis of tumour-specific antibodies (for example, Immunology, Vol 62). It has been suggested that antigens induced by tumours can be used to produce antibodies (190), and it has been suggested that tumour cells (191) and tumour cells (192) have provided a means of producing antibodies. The development of antibody therapy has provided a means of delivering drugs to specific sites in the body. It is suggested that the use of antibodies in the treatment of cancer is a promising field.

**Fig 1E-1: Some uses of castor bean lectins**

- Cytochemical marker for intracellular membranes (178)**
- Marker for Duchenne muscular dystrophy membranes (179)**
- Developmental analysis of alpha-fetoprotein variants (180)**
- Purification of hog cholera virus (181)**
- Purification of glycoproteins (See 46)**
- Characterisation of thymocytes (See 46)**
- Analysis of ageing in hepatocytes (See 46)**
- Separation of glycolipids from glycoproteins (182)**
- Selective toxicity against spleen T and B cells (183)**
- Toxin for assassination (8)**
- Potential warfare agent (9)**
- Characterisation of metastatic and non-metastatic tumour cells (184)**
- Possible marker of preneoplastic tissues (185)**
- Therapy of insulin receptor deficiencies (as hybrid of B chains + insulin) (186)**

The problem with using toxins as such, as with all other cancer drugs, lies in the non-specificity of their targets of action. As long ago as 1906, Paul Ehrlich (6) suggested that the specificity of antibodies could be used to direct drugs to specific locations within the body. Early attempts to do this involved joining conventional anti-cancer drugs to antibodies (See 187-189), and these continue. Although some success has been achieved with this approach, the stoichiometric nature of the actions of these drugs generates the necessity to deliver large amounts of such conjugates to the affected cells, with the result that the conjugates are often not as effective as the drugs given alone.

An obviously important factor in the use of any antibody-directed therapeutic agent is the availability of suitably specific antibodies. Much knowledge is currently being accumulated on the expression of tumour-specific antigens (See for example, Immunological Reviews Vol 62). It has been suggested that antigens could be induced by infection with viruses which show tumour specificity (190), and work on immunotoxins has involved SV40 - transformed cells (191) and mumps virus - infected cells (192). The most significant advance in recent years, which has provided a major stimulus in this field, is of course the development of monoclonal antibodies (193). The important question now is whether antigens expressed by a given tumour will also be expressed by all other tumours of the same type. The importance of this is that it is impractical to generate a special monoclonal antibody and link it to a toxin, for every patient requiring treatment.

The enzymatic character of toxins such as ricin and diphtheria toxin has made them extremely attractive as the toxic agents in conjugates - as is indicated by the large number of recent publications in leading cancer journals.

It was at first supposed that immunotoxins should contain only the A chains of the toxins, in order to avoid non-specific binding by the B chains. This was hoped to prevent general toxicity, as well as attachment to plasma proteins and erythrocytes, forming complexes which cannot leave the bloodstream (194). For examples of this approach see references 195 - 197. However, conjugates containing only A chains tend to be erratic in their toxicity, and less effective than expected from the amounts of A chain present (171,191), in agreement with the evidence described earlier that the B chain plays a role other than initial binding in the transport of the A chain to the cytosol.

A number of attempts have been made to circumvent this problem. For example, one group (198) treated cultured cells with a non-toxic A chain conjugate, and added free B chain. This resulted in intoxication of the cells, again demonstrating the importance of the B chain. However, this is unlikely to be of use in vivo because of the problem in getting the B chains to the target cells.

Another approach (199) involved the simultaneous use of two conjugates - one containing the A chain, and the other containing the B chain. In this case, both contained the same antibody, but it might be advantageous to use antibodies with two different specificities (both for the target cell) to further reduce non-specific toxicity. Again, the results obtained in vitro were favourable, but the B chain binding problem remains.

Yet another possibility is to use antibodies of dual specificity: separate antibodies are raised against the toxin component and the target cells. These are reduced, and then mixed, to form molecules containing one 'arm' from each of the original antibodies. These are bound to the toxin and then used as immunotoxins (200).

The most obvious way to abolish the galactose-binding property of an immunotoxin, while retaining the secondary function of the B chain is to use chemical blockade. Abrin has been treated with an agent which modifies lysine residues, and in so doing abolishes galactose binding. Immunotoxins constructed with this product were toxic to their target cells, but considerable residual non-specificity was noted, which was shown not to be due to binding via galactose residues (171). It is possible that the antibody used was insufficiently specific.

Perhaps the most promising approach so far has been the construction of the conjugates in such a way as to physically interfere with the galactose binding site - so-called steric hindrance. Preliminary results indicate that the conjugates are effective in mice, even when given intravenously (194), and they have been used to remove cancer cells from bone marrow which was subsequently reintroduced into an irradiated patient (PE Thorpe, personal communication).

The current interest in immunotoxins arises from the remaining method of abolishing the galactose-binding site: once cDNA clones which express ricin are available, it may be possible to introduce mutations which result in a protein with all the requisite properties except galactose binding. A second potential advantage of this approach is that present conjugates are less effective than the expected maximum because of removal from the

circulation by the reticuloendothelial system (201). Since this is likely to be largely a consequence of the sugar moieties attached to ricin, the production of the mutated product in a non-glycosylating system may well avoid this problem.

1F Castor bean lectin cDNA cloning

It will be clear from the discussion thus far that a great many problems remain to be solved, including purely academic points, as well as more practical aspects. Many of these are amenable to recombinant DNA approaches, inasmuch as structural and functional information can be obtained; hybridisation probes can be made available, and the starting material for immunotoxins is provided.

The structure of ricin and the agglutinin can be approached by determination of their sequences - the differences between them may help explain the characteristics of their different structures and functions. Protein sequencing data are of course applicable, but these are only available for ricin. Also, protein sequencing fails to provide any information on the synthesis of the lectins, in terms of the nature of the precursor. However, cDNA sequences define the initial translation product and reveal features such as signal sequences and the organisation of subunits within precursors.

The construction of expressing derivatives of cDNA clones will help in elucidating the mechanisms of action of the two chains. The introduction of mutations can be correlated with the effects these have on the expressed products - for example, A chain mutants may be obtained which attach to the ribosome, but go no further. Analysis of these would help locate the target of ricin action.

Similarly, the site within the B chain of the galactose-binding function should be clarified.

From the practical point of view, the overriding importance of any expressing toxin clone is its potential for making immunotoxins. Mutation of the galactose binding site has already

been discussed, and the removal of glycoprotein immunotoxins from the circulation has been mentioned. The latter problem may be overcome by expression in bacterial systems, which do not glycosylate proteins, though bacteria may be unable to allow the correct folding and processing of the molecule. However, ricin can be reformed from its reduced subunits (71), so the important point will probably be the preparation of separate subunits. This could be achieved by an in vitro cleavage of an expressed ricin precursor (the castor bean enzyme which carries this out is presently being characterised, JM Lord, personal communication), or by separate expression of both subunits, followed by reassociation.

In any event, non-glycosylating yeast mutants have been described (339), and for these reasons, the clones are currently being recloned in a yeast expression vector (see Results and Discussion section), as well as into E coli expression systems.

The work described here involves the cloning of cDNA sequences encoding both ricin and RCA, along with the determination of their nucleotide sequences. The protein sequences are compared with those of Funatsu, and in a limited survey, the nucleotide sequences are compared with other plant sequences.

## CHAPTER 2: MATERIALS AND METHODS

### 2A Materials

Castor beans were obtained from Croda Premier Oils. The beans supplied are of uncertain origin, and may represent a mixture of strains. The patterning of the testa differs considerably between the seeds; those of even reddish-brown colour being sorted for use here, to try to minimise sequence variation. All bacterial growth reagents were from Difco, and plasmids pUC8 and pBR322, and the latter, cut with PstI and tailed with G residues, were from BRL. An M13 cloning / dideoxy sequencing kit, including M13 mp8 and E coli JM101 cells was obtained from BRL, and a Universal M13 sequencing primer was from New England Biolabs. Rabbit reticulocyte lysate kits were purchased from New England Nuclear, and all isotopes were from Amersham International.

The 20-mer oligonucleotide used for screening the cDNA library was a generous gift from Celltech Ltd. Restriction and other enzymes were from BRL, Boehringer and NBL; terminal transferase was obtained from Miles and reverse transcriptase from Life Sciences Inc. Anti-RCA serum was from Vector Labs, and is known to react with the 2S albumins as well as with the lectins (see p 22).

General chemicals, solvents and antibiotics were obtained from BDH, Sigma and Fisons, as were agarose and acrylamide. Nitrocellulose was Schleicher & Schuell type BA85/1. Oligo(dT)-cellulose was obtained from Collaborative Research.

### Addresses

Croda Premier Oils, Hull, UK

BRL, PO Box 145, Science Park, Cambridge, UK

New England Biolabs, 32, Tozer Rd, Beverly, MA, USA

New England Nuclear, 2 New Road, Southampton, UK

Amersham International, Amersham, Bucks, UK

Celltech Ltd, 244-250 Bath Road, Slough, Berks, UK

Boehringer Corp Ltd, Bell lane, Lewes, E Sussex

NBL Enzymes Ltd, S Nelson Industrial Estate, Cramlington, Northumberland, UK

Miles Labs, PO Box 37, Stoke Poges, Slough, Berks, UK

Life Sciences Inc, St Petersburg, Florida, USA

Vector Labs, Burlingame, California, USA

Collaborative Research, Waltham, Massachusetts, USA

BDH Chemicals Ltd, Broom Rd, Poole, Dorset, UK

Sigma, Fancy Rd, Poole, Dorset, UK

Fisons, Loughborough, Leics, UK

## 2B General methods

### 2B1 Growth and harvesting of castor beans

Castor bean plants (Ricinus communis) were grown from seed in John Innes No 1 compost in the greenhouse. They were maintained at 20°C and illuminated with sodium lamps (10,000 -12,000  $\text{lm m}^{-2}$ ) with a light:dark regime of 16 hr : 8 hr (ref 81).

The development of castor bean plant seeds has been divided into seven stages based on size, formation of the testa and state of hydration (81). Seeds were harvested at stages D-E for cDNA cloning, and at earlier stages for screening by differential hybridisation (see Results and Discussion section 3D).

### 2B2 Growth and maintenance of E coli

Details for E coli JM101 appear in section 2H1. Strain DH1 (202) was stored at -20°C in 50 % glycerol, 0.5 x LB (LB is 1 % tryptone, 1 % NaCl, 0.5 % yeast extract), and were replated on LB plates containing 1.5 % agar each week. Cells containing plasmids were grown in, or plated on, LB containing tetracycline at 14  $\mu\text{g/ml}$ . Ampicillin sensitivity was determined by plating on LB containing ampicillin at 50  $\mu\text{g/ml}$ .

The identity of DH1 cells was periodically checked by plating on LB containing 100  $\mu\text{g/ml}$  nalidixic acid, to which they are resistant (202).

### 2B3 Preparation of competent E coli cells and transformation

DH1 cells to be transformed with pBR322 and its recombinants were rendered competent with RbCl and  $\text{MnCl}_2$ , as this gave transformation frequencies considerably higher than the simple  $\text{CaCl}_2$  method ( $10^7 - 10^8$  transformants per  $\mu\text{g}$  of supercoiled plasmid, and

$10^3 - 10^4$  transformants per  $\mu\text{g}$  respectively). Recombinants in pUC8 were transformed into cells prepared with  $\text{CaCl}_2$ , as this simpler method was found to give satisfactory results ( $10^6 - 10^7$  transformants per  $\mu\text{g}$ ).

The  $\text{RbCl} / \text{MnCl}_2$  method was as follows: cells were grown in 10 ml cultures of psi broth (2 % tryptone, 0.5 % yeast extract, 10 mM  $\text{NaCl}$ , 20 mM  $\text{MgCl}_2$  pH 7.6), and grown at  $37^\circ\text{C}$  in a shaking incubator until the absorbance at 530 nm was 0.3 units. A one ml aliquot was diluted to 25 ml with fresh psi broth and the culture was grown to an absorbance of 0.48  $A_{530}$  units. The cells were chilled on ice for 15 minutes and then harvested at 5,000 rpm for 5 minutes at  $4^\circ\text{C}$ . They were resuspended in 10 ml of 100 mM  $\text{RbCl}$ , 50 mM  $\text{MnCl}_2$ , 10 mM  $\text{CaCl}_2$ , 35 mM  $\text{NaAc}$  pH 5.8, 15 % glycerol and kept on ice for 15 minutes. The cells were again harvested, and resuspended in 1 ml of 10 mM  $\text{RbCl}$ , 10 mM  $\text{MOPS}$  pH 5.8, 75 mM  $\text{CaCl}_2$ , 15 % glycerol and kept on ice for a further 15 minutes.

Aliquots of 100  $\mu\text{l}$  of cells were mixed with the DNA samples and were incubated on ice for 30 minutes, after which they were heat-shocked at  $42^\circ\text{C}$  for 90 - 120 seconds. The mixtures were diluted to 1 ml with fresh psi broth and grown at  $37^\circ\text{C}$  for one hour. Cells were then pelleted, resuspended in 100  $\mu\text{l}$  of psi broth and plated on LB containing tetracycline at 14  $\mu\text{g}/\text{ml}$ . After 18 - 24 hours growth, colonies were counted and spotted onto LB plates containing ampicillin at 50  $\mu\text{g}/\text{ml}$  to identify colonies harbouring recombinant plasmids.

The calcium chloride method was as follows: cultures of 10 - 15 ml of cells were grown in LB until the  $A_{550}$  was 0.6 units, at which point the cells were harvested and resuspended in 10 ml of 50 mM  $\text{CaCl}_2$ . After 40 minutes on ice they were again harvested and resuspended in 1 ml of 50 mM  $\text{CaCl}_2$ , and kept on ice for a further

40 minutes. Aliquots of 100  $\mu$ l of cells were mixed with DNA samples and incubated on ice for another 40 minute period, and were then heat-shocked at 45°C for 2 minutes and plated on LB containing 50  $\mu$ g/ml ampicillin, and grown overnight at 37°C.

#### 2B4 Extraction and purification of plasmid DNA

Plasmids were prepared from E coli by a modification of the alkaline lysis method of Birnboim & Doly (203). Ten ml overnight cultures were grown at 37°C with shaking, in LB containing the appropriate antibiotic. This was diluted to 1 litre with fresh LB and growth was continued until the  $A_{590}$  was approximately 0.8 units, at which point chloramphenicol was added to 175  $\mu$ g/ml to amplify the plasmids (204), and the cultures were grown for a further 15 - 18 hours. Amplification was not carried out with pUC8 clones, since adequate yields were obtained from overnight cultures.

Bacteria were harvested by centrifugation at 5,000 rpm for 5 minutes at 4°C, and resuspended in 3.4 ml of 50 mM glucose, 10 mM EDTA, 25 mM tris-HCl pH 8, 2.5 mg/ml lysozyme. After incubation on ice for 10 minutes, 6.6 ml of 200 mM NaOH, 0.1 % SDS was added, mixed thoroughly and the mixture was kept on ice for a further 10 minutes to allow lysis. Chromosomal DNA and high molecular weight RNA were precipitated by the addition of 5 ml of 3 M NaAc pH 4.8, with a 30 minute incubation on ice. After removal of this material by centrifugation at 10,000 rpm for 10 minutes, nucleic acids were precipitated from the supernatant with 0.6 volumes of isopropanol, and an incubation of 10 minutes at room temperature, and harvested by centrifugation at 10,000 rpm for 10 minutes. Isopropanol does not efficiently precipitate proteins under these conditions (205). Plasmid DNA was washed out of the pellet by thorough resuspension in 3.2 ml of 2 M  $NH_4Ac$ , followed by a further round of centrifugation as described above - the pellet being discarded. Nucleic acids were again precipitated with isopropanol, as above, and were washed with

40 minutes. Aliquots of 100  $\mu$ l of cells were mixed with DNA samples and incubated on ice for another 40 minute period, and were then heat-shocked at 45°C for 2 minutes and plated on LB containing 50  $\mu$ g/ml ampicillin, and grown overnight at 37°C.

#### 2B4 Extraction and purification of plasmid DNA

Plasmids were prepared from E coli by a modification of the alkaline lysis method of Birnboim & Doly (203). Ten ml overnight cultures were grown at 37°C with shaking, in LB containing the appropriate antibiotic. This was diluted to 1 litre with fresh LB and growth was continued until the  $A_{590}$  was approximately 0.8 units, at which point chloramphenicol was added to 175  $\mu$ g/ml to amplify the plasmids (204), and the cultures were grown for a further 15 - 18 hours. Amplification was not carried out with pUC8 clones, since adequate yields were obtained from overnight cultures.

Bacteria were harvested by centrifugation at 5,000 rpm for 5 minutes at 4°C, and resuspended in 3.4 ml of 50 mM glucose, 10 mM EDTA, 25 mM tris-HCl pH 8, 2.5 mg/ml lysozyme. After incubation on ice for 10 minutes, 6.6 ml of 200 mM NaOH, 0.1 % SDS was added, mixed thoroughly and the mixture was kept on ice for a further 10 minutes to allow lysis. Chromosomal DNA and high molecular weight RNA were precipitated by the addition of 5 ml of 3 M NaAc pH 4.8, with a 30 minute incubation on ice. After removal of this material by centrifugation at 10,000 rpm for 10 minutes, nucleic acids were precipitated from the supernatant with 0.6 volumes of isopropanol, and an incubation of 10 minutes at room temperature, and harvested by centrifugation at 10,000 rpm for 10 minutes. Isopropanol does not efficiently precipitate proteins under these conditions (205).

Plasmid DNA was washed out of the pellet by thorough resuspension in 3.2 ml of 2 M  $NH_4Ac$ , followed by a further round of centrifugation as described above - the pellet being discarded. Nucleic acids were again precipitated with isopropanol, as above, and were washed with

70 % ethanol. After brief drying in vacuo the DNA was dissolved in 10 mM tris-HCl pH 7.6, 1 mM EDTA, and RNase T1 (to 100 units/ml) or RNase A (to 5 µg/ml) was added, and the mixture was incubated at 37°C for 30 - 60 minutes. The resulting DNA was twice extracted with phenol equilibrated with 10 mM tris-HCl, 1 mM EDTA, and precipitated from 0.3 M NaAc pH 6 with 2 volumes of ethanol at -80°C for 30 minutes. The DNA was dissolved in up to 500 µl of 10 mM tris-HCl pH 7.6, 1 mM EDTA and an aliquot was analysed on a neutral agarose gel.

Supercoiled plasmid DNA was further purified on CsCl density gradients (206) consisting of 31.8 ml of 10 mM tris-HCl pH 7.6, 1 mM EDTA containing 28.9 g CsCl, 18 mg EtBr and the DNA sample. This was centrifuged in a VT150 rotor at 45,000 rpm for 16 hours. The lower (plasmid) band was removed with a syringe and extracted three times with amyl alcohol to remove the EtBr. The final aqueous phase was diluted with 4 volumes of 10 mM tris-HCl pH 7.6, 1 mM EDTA to prevent precipitation of CsCl, and precipitated with ethanol. After centrifugation, the pellet was rinsed several times with 70 % ethanol, dried briefly, redissolved in 10 mM tris-HCl pH 7.6, 1 mM EDTA and an aliquot was analysed on a neutral agarose gel.

#### 2B5 Restriction enzymes and mapping

Restriction enzymes were used in accordance with the manufacturer's instructions. Where possible, the 10 x core buffer provided by BRL was used; otherwise buffers were prepared from sterile stocks. DNA concentrations were typically 10 - 100 ng/µl in the final reactions which were incubated at 37°C for 1 - 16 hours, with the exceptions of TaqI which was incubated at 65°C and SmaI, which was incubated at 30°C.

The products of digestion were analysed on 0.7 - 1.5 % neutral

agarose gels, and were stained after running with a solution of 0.5 µg/ml EtBr in water for 30 minutes. If necessary, gels were destained in water for 10 - 30 minutes. Stained DNA was visualised under UV light and gels were photographed with a Polaroid camera. Gels with radioactively labelled samples were dried and exposed to X-ray film.

No complex restriction mapping was required for this work, so straightforward approaches were used. Mapping was done by single and double digestion, or by sequential digestion (See ref 205).

#### 2B6 In vitro protein synthesis and immunoprecipitation

Rabbit reticulocyte lysate kits were obtained from New England Nuclear and used in accordance with the manufacturer's instructions, except that reactions were carried out in the presence of 75 mM KAc and 0.62 mM MgAc<sub>2</sub> (207). Reactions were incubated in 25 µl volumes at 37°C for 40 minutes and protein synthesis was assayed by uptake of (<sup>35</sup>S)methionine into hot TCA - insoluble material, samples being processed by the method of Mans & Novelli (205) and counted in Bray's scintillant (209) or in Beckman commercially supplied scintillant cocktail.

Samples for immunoprecipitation were mixed with an equal volume of 1 % Nonidet P40 in 20 mM Tris-HCl pH 7.4, 2 mM EDTA, 150 mM NaCl, 200 mM lactose and 40 µg/ml PMSF (81). After a 45 minute incubation at room temperature insoluble material was pelleted in a Beckman airfuge at 20 psi for 5 minutes and 2 µl of null serum was added to the supernatant. After 15 minutes at room temperature 30 µl of protein A - Sepharose was added and the mixture was reincubated at room temperature for 30 minutes. The beads were pelleted and 2.5 µl of anti-RCA serum (Vector Labs) was added to the supernatant. After one hour at room temperature 50 µl of protein A - Sepharose was added

and the mixture was incubated at room temperature for 30 minutes. The beads were then pelleted and washed three times with 20 mM tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.2 % NP40, then twice in the same buffer but containing 500 mM NaCl, and once in 10 mM tris-HCl pH 7.4. Material eluted from the beads with SDS/PAGE sample buffer was analysed on polyacrylamide gels as described in section 2C4.

#### 2B7 End-labelling of DNA fragments

DNA fragments were labelled at their 3' ends after cutting with restriction enzymes which produce protruding 5' termini. Labelling was with  $\alpha$ -( $^{32}\text{P}$ )dNTPs in the presence of the large fragment of E coli DNA polymerase I (Klenow enzyme). The reactions were carried out in any restriction enzyme buffer (205), using 0.1 - 0.5 Units of enzyme and 5-10  $\mu\text{Ci}$  of isotope (3000 Ci/mmol) per  $\mu\text{g}$  of DNA. After ethanol precipitation the DNA was thoroughly washed with 70 % ethanol and further purified as required. Non-radioactive dNTPs were added as appropriate, to 100  $\mu\text{M}$  each.

This procedure was satisfactory for sequencing fragments labelled at BamHI, BglII, Sau96I and TaqI sites. For AvaI sites, which require, at least in the case of the site in pUC8, the incorporation of two adjacent C residues, a second aliquot of enzyme, along with non-labelled dCTP to 100  $\mu\text{M}$ , was added, and the reaction was reincubated as above.

Oligomers and some DNA fragments were labelled at their 5' ends with polynucleotide kinase. In the case of DNA fragments, the termini were dephosphorylated with calf intestinal alkaline phosphatase prior to labelling.

DNA was dephosphorylated at 10 - 100 ng/ $\mu\text{l}$  in 50 mM tris-HCl pH 9, 1 mM  $\text{MgCl}_2$ , 0.1 mM  $\text{ZnCl}_2$ , 1 mM spermidine, using 1 - 5 units of phosphatase per  $\mu\text{g}$  of DNA (205). For treatment of protruding 5'

termini, the reactions were incubated at 37°C for 30 minutes, and a second aliquot of enzyme was added and the incubation was repeated. For other types of terminus, each incubation was split into a 15 minute part at 37°C followed by 15 minutes at 56°C (205).

Reactions were stopped by adding an equal volume of 20 mM tris-HCl pH 8, 200 mM NaCl, 2 mM EDTA and SDS to 0.5 %, and heating to 65°C for 15 minutes. DNA was recovered by ethanol precipitation after three phenol and one chloroform extractions.

Dephosphorylated 5' termini were labelled in the presence of 1 - 10 units of polynucleotide kinase and 2 - 5  $\mu$ Ci of  $\gamma$ -(<sup>32</sup>P)ATP (5000 Ci/mmol per  $\mu$ g of DNA, and 50 mM tris-HCl pH 8.5, 10 mM MgCl<sub>2</sub>, 5 mM DTT, 0.1 mM spermidine, 0.1 mM EDTA, with a 30 minute incubation at 37°C. The reactions were phenol extracted and the DNA was purified by ethanol precipitation or gel purification.

Oligomers were obtained with 5'-hydroxyl groups, so were not dephosphorylated. They were 5'-end-labelled at a concentration of 10 ng/ $\mu$ l in the presence of 50 mM tris-HCl pH 8.5, 10 mM MgCl<sub>2</sub>, 5 mM DTT, 0.1 mM spermidine, 0.1 mM EDTA. 2.5  $\mu$ Ci/ $\mu$ l ( $\gamma$ -<sup>32</sup>P)ATP (5,000 Ci/mmol) and 0.2 units/ $\mu$ l of polynucleotide kinase. After incubation at 37°C for 30 minutes the reaction was stopped by the addition of an equal volume of 600 mM NH<sub>4</sub>Ac and the labelled oligomers were separated from unincorporated isotope by centrifugation through a spinning column of Biogel P60 in 20 mM tris-HCl pH 7.6, 140 mM NaCl, 5 mM EDTA and 0.1 % SDS.

## 2C Gel electrophoresis

### 2C1 Neutral agarose gels

Samples were mixed with 2 - 10 volumes of 40 mM tris-acetate pH 7.8, 4 mM EDTA, 1 % glycerol, 0.01 % bromophenol blue, and analysed on 0.7 - 1.5 % (depending on the size range to be separated) agarose gels (Sigma medium EEO agarose) in TAE (40 mM tris-acetate pH 7.8, 5 mM NaAc, 1 mM EDTA) or in TBE (89 mM tris-borate, 20 mM EDTA). The buffer system depended on the purpose of the gel: where clear resolution of topological isomers or accurate size estimations were required, TAE was used. For rapid checking of restriction digests, and for preparative gels, TBE was used.

When accurate sizing was required, gels were stained with 0.5 µg/ml EtBr for 30 minutes, with, if necessary, a 30 minute destaining period in water. Otherwise, EtBr was included in the gel buffer at 0.5 µg/ml. Nucleic acids were visualised under UV illumination and photographed with a Polaroid camera. For autoradiography, gels were dried and exposed to X-ray film.

### 2C2 Alkaline agarose gels

cdNA synthesis was followed on denaturing agarose gels (210) made in 30 mM NaOH, 2 mM EDTA. Sample buffer was as for neutral gels, but bromocresol green replaced bromophenol blue. Running buffer was 30 mM NaOH, 2 mM EDTA, and gels were soaked for 30 minutes in 100 mM NH<sub>4</sub>Ac and stained for 30 minutes in 0.5 µg/ml EtBr prior to examination under UV light. Gels were dried and exposed to X-ray film as appropriate.

### 2C3 Formamide agarose gels

RNA samples were analysed under denaturing conditions in the presence of formamide (211). 10 x E buffer is 36 mM tris base, 30 mM  $\text{NaH}_2\text{PO}_4$ , 2 mM EDTA. Samples were made 50 % in formamide, 1 x in E buffer, 5 % in glycerol and 0.01 % in bromophenol blue, and were denatured by heating to 60°C for 3 minutes and cooled in ice-water. They were run on 1.2 % agarose gels in 50 % formamide, 1 x E buffer. After running, gels were stained and photographed as above. Prior to blotting, they were soaked for 1 - 2 hours in 10 % formaldehyde, and then in 20 x SSC for 30 minutes.

### 2C4 SDS-Polyacrylamide gels

Products of in vitro translation and immunoprecipitated samples were run on SDS-polyacrylamide gels (212) with a 10 % acrylamide, 0.26 % bisacrylamide resolving gel in 400 mM tris-HCl pH 8.8, 0.1 % SDS, and a stacking gel of 5 % acrylamide, 0.13 % bisacrylamide in 60 mM tris-HCl pH 6.8, 0.1 % SDS. Gels were run in 60 mM tris-HCl pH 8.8, 500 mM glycine, 0.1 % SDS.

Samples were mixed with 5 - 10 volumes of 200 mM tris-HCl pH 8.4, 500 mM sucrose, 2 mM EDTA, 0.01 % bromophenol blue, 1 % methionine, 4 % SDS, 10 mM DTT, and were heated to 95°C for 2 minutes. They were then alkylated by the addition of iodoacetamide to 50 mM, with a 15 minute incubation at room temperature.

Gels were fixed in cold 10 % TCA for 30 minutes and fluorographed by impregnation either with diphenyloxazole (213) or with salicylate (214).

### 2C5 DNA sequencing gels

DNA sequencing samples were analysed on 40 cm long, 0.4 mm thick, 6 % polyacrylamide, 0.3 % bisacrylamide gels in 7 M urea,

89 mM tris-borate, 20 mM EDTA (215). Running buffer was 89 mM tris-borate, 20 mM EDTA.

Samples were mixed with 0.4 - 1.0 volume of 90 % formamide, 10 mM EDTA, 0.01 % bromophenol blue, 0.01 % xylene cyanol and heated to 95°C for 2 minutes prior to loading on gels which had been pre-run for 30 minutes.

For short runs with dideoxy sequencing samples, buffer gradient gels (216) were used, samples running from 1 x buffer at the top towards 2.5 x buffer at the bottom.

Gels were cast between glass plates, the notched plate having been treated with Repelcote and the back plate treated with gamma-methacryloxypropyltrimethoxysilane as described by Ansorge & DeMaeyer (217) to covalently bind the gel to the plate. After running gels were fixed in 10 % acetic acid for 15 minutes, dried, and exposed to X-ray film.

To obtain sequence close to the labelled site in Maxam & Gilbert samples, 16 % acrylamide, 0.8 % bisacrylamide gels were used, these samples being mixed with 90 % formamide, 10 mM EDTA lacking dyes, as these were found to distort the DNA banding pattern. When marker bromophenol blue in an unused lane was approximately one - third to one - half of the way down the gel, the plates were separated and the gel was covered in clingfilm and exposed wet.

#### 2C6 Elution of DNA from agarose gels

The DNA band to be purified was visualised under UV light, and electrophoresed onto a piece of Whatman No 1 filter paper backed with a piece of dialysis membrane which had been boiled for 5 minutes in 20 mM EDTA (205). DNA was recovered from this by centrifugation through a hole in a 0.4 ml tube placed inside a 1.5 ml tube. The DNA was phenol extracted three times and chloroform extracted once, prior to ethanol precipitation.

2D Extraction and fractionation of castor bean mRNA

Castor bean mRNA was prepared essentially as described by Bowden-Bonnett & Lord (218): 100 -200 g of beans were frozen and ground to a powder in liquid nitrogen, and homogenised in a Waring blender for 1 - 2 minutes in 50 mM tris-HCl pH 9, 150 mM NaCl, 5 mM EDTA, 5 % SDS. The homogenate was extracted with an equal volume of phenol:chloroform (1:1) and the phases were separated by centrifugation. The organic phase and residue were reextracted with 0.5 volume of 20 mM tris-HCl pH 9, 2 mM EDTA and the two aqueous phases were combined. The aqueous material was reextracted with phenol:chloroform until no material was present at the interface. The final RNA solution was made 200 mM in NaCl and precipitated with 2 volumes of ethanol at  $-20^{\circ}\text{C}$  overnight. All glassware was siliconised and autoclaved, and all solutions were autoclaved.

RNA was recovered by centrifugation and was washed several times with 3 M NaAc pH 5.5 until no polysaccharide was detected in the supernatant by ethanol precipitation.

RNA molecules bearing poly(A) tails were extracted by affinity chromatography on oligo(dT)-cellulose: after hybridisation for 30 minutes at room temperature in 400 mM NaCl, 20 mM tris-HCl pH 7.6, 0.1 % SDS the beads were pelleted and washed three times in the above buffer, and twice in 200 mM NaCl, 20 mM tris-HCl pH 7.6, 0.1 % SDS. The slurry was poured into a column and washed further with the latter buffer until the  $A_{260}$  of the eluate (monitored with an ISCO continuous flow UV cell) reached the background level. Poly(A) - containing RNA was then eluted with 20 mM tris-HCl pH 7.6 at  $50^{\circ}\text{C}$ , precipitated with ethanol, dried and dissolved in 10 mM tris-HCl pH 7.6 to approximately 1  $\mu\text{g}/\mu\text{l}$ .

Approximately 400  $\mu\text{g}$  of poly(A)<sup>+</sup> mRNA was layered on top of a

10 - 30 % ribonuclease-free sucrose (Sigma) density gradient in 100 mM tris-HCl pH 7.5, 0.5 % SDS, 1 mM EDTA, and centrifuged in an SW27 rotor at 25,000 rpm at 17°C for 14 hours (219). Fractions of 400  $\mu$ l were collected using an ISCO density gradient fractionator and monitored with the UV cell.

Each fraction was precipitated with ethanol, washed with 70 % ethanol, and redissolved in 10  $\mu$ l of 10 mM tris-HCl pH 7. Fractions were subsequently translated in rabbit reticulocyte lysates, immunoprecipitated and analysed on SDS-polyacrylamide gels.

## 2E Construction of cDNA library in pBR322

### 2E1 Synthesis of double-stranded cDNA

Poly(A) - containing mRNA species encoding the lectin precursor were identified by in vitro translation and immunoprecipitation. Appropriate fractions were reverse transcribed (220,221) at 50 ng/ $\mu$ l in the presence of 50 mM tris-HCl pH 8.3, 10 mM MgCl<sub>2</sub>, 100 mM KCl, 1 mM each of dATP, dTTP, and dGTP, 250  $\mu$ M dCTP, 60 ng/ $\mu$ l oligo(dT)<sub>12-18</sub>, 10 mM DTT and 0.4 units/ $\mu$ l reverse transcriptase. (<sup>3</sup>H)dCTP or  $\alpha$ -(<sup>32</sup>P)dCTP were included in the reaction as appropriate.

The reaction mixture was incubated at 42°C for 45 minutes, at which point an equal volume of 5 mM tris-HCl pH 8.3, 5 mM DTT, 250  $\mu$ M dCTP was added along with the same amount of reverse transcriptase as previously. The reaction was incubated for a further 45 minutes at 45°C, and was terminated by freezing. Aliquots were analysed on a 1% alkaline agarose gel along with the products of second strand synthesis and S<sub>1</sub> nuclease reactions.

The mRNA-cDNA hybrids were denatured by boiling for 3 minutes and cooling rapidly (220). After pelleting insoluble material and transferring the supernatant to a fresh tube, the following reagents were added, ignoring elements already present: dATP, dGTP and dTTP to 100  $\mu$ M each, HEPES-KOH pH 6.9 to 105 mM, KCl to 92 mM, dCTP, labelled as appropriate, to 80  $\mu$ M and 0.1 units / $\mu$ l of DNA polymerase I from E coli (222,223). The reaction was allowed to proceed for 6 hours at 20°C, at which time nucleic acids were purified by gel filtration on 1 ml columns of Biogel P60 in 10 mM tris-HCl pH 7.6, 20 mM NaCl, 1 mM EDTA. Fractions were monitored by Cerenkov or liquid scintillation counting, and peak excluded fractions were pooled and precipitated from 0.3 M NaAc pH 6

with 2 volumes of cold ethanol. Precipitates were recovered by centrifugation and dissolved in water to about 2.5  $\mu\text{g}$  of RNA-equivalent material per  $\mu\text{l}$ .

The double-stranded molecules were then digested with nuclease  $S_1$  from Aspergillus oryzae to remove single-stranded regions downstream and the hairpin loop upstream (223). The reaction was performed in the presence of 300 mM NaCl, 30 mM NaAc pH 4.5, 3 mM  $\text{ZnCl}_2$ , 100 ng/ $\mu\text{l}$  of cDNA (RNA-equivalent) and 0.05 units/ $\mu\text{l}$  of  $S_1$  nuclease. Incubation was for 15 minutes at 37°C, and then for 15 minutes at 15°C, and the reaction was terminated by the addition of tris-HCl pH 7.6 to 130 mM and EDTA to 10 mM. The reaction was then extracted with phenol:chloroform (1:1) and DNA was precipitated with 2 volumes of ethanol. The DNA was dissolved in 10 mM tris-HCl pH 8, 0.1 mM EDTA to 250 ng/ $\mu\text{l}$  of RNA-equivalent material.

#### 2E2 Construction of recombinants

Double-stranded cDNA was tailed with dCTP (224) at 1 - 10 ng/ $\mu\text{l}$  in the presence of 140 mM potassium cacodylate pH 7.6, 30 mM tris base, 0.1 mM DTT, 1 mM  $\text{CoCl}_2$  and  $\alpha$ -( $^{32}\text{P}$ )dCTP in 75-150 - fold excess over 3' termini at 37°C for 6 minutes with a concentration of terminal transferase of 0.5 - 1.0 units/ $\mu\text{l}$ .

The extent of incorporation of label was followed by assaying the TCA-insoluble radioactivity as a proportion of the total by liquid scintillation counting. The reaction was stopped by chilling and adding EDTA to 10 mM, after which the DNA was purified by gel filtration. DNA was precipitated with ethanol and dissolved in 1 M NaAc pH 8, 10 mM tris-acetate pH 8, 1 mM EDTA.

Tailed DNA was fractionated on 5 - 20 % linear sucrose density gradients in 1 M NaAc pH 8, 10 mM tris-acetate pH 8, 1 mM EDTA and centrifuged overnight at 39,000 rpm in an SW50.1 rotor.

DNA sedimentation was checked on a parallel gradient loaded with a mixture of *Hinf*I and *Pst*I digests of pBR322 DNA. Fractions from the gradient were diluted with an equal volume of water and precipitated with ethanol, and pooled to give three final fractions: a large cDNA fraction (larger than 2200 bp), an intermediate fraction (1000 - 2200 bp) and a small cDNA fraction (less than 1000 bp). Smaller cDNA molecules were discarded. The three final fractions were dissolved to approximately 5 ng/ $\mu$ l in 150 mM RbCl, 10 mM tris-HCl pH 7.6, 0.1 mM EDTA.

dC-tailed cDNA was mixed with equimolar quantities of dG-tailed *Pst*I-cut pBR322 DNA (225) at a concentration of 0.4 ng/ $\mu$ l of vector DNA. The mixtures were heated to 70°C for 3 minutes and cooled overnight to room temperature. *E coli* DH1 cells were rendered competent and transformed as described in section 2B3; recombinant clones were identified by plating tetracycline-resistant colonies on ampicillin-containing LB plates.

## 2F Screening of the cDNA library

### 2F1 Differential hybridisation

The cDNA library was initially screened by colony hybridisation (226) using single-stranded cDNA probes prepared from mRNA extracted from immature and mature castor beans (81). It was expected that lectin clones would hybridise with 'late' probes but not with 'early' probes.

The colonies were grown on nitrocellulose filters (Schleicher & Schuell type BA85/1) on LB containing tetracycline at 14 µg/ml until the colonies were 1 - 3 mm in diameter. The filters were then transferred to LB containing chloramphenicol at 150 µg/ml and tetracycline, to amplify the plasmids.

The filters were processed by placing them on pieces of Whatman 3MM paper wetted with the appropriate reagents. They were twice treated with 0.5 M NaOH for 10 - 15 minutes - when complete lysis was observed by the change in appearance of the colonies, the alkali was neutralised by two treatments with 1 M tris-HCl pH 8, followed by two treatments with 1M tris-HCl pH 8 containing 1.5 M NaCl. Protein was then digested by floating the filters on minimal volumes of 0.1 x SSC containing 0.33 µg/ml of proteinase K, both sides of the filter being treated. They were then rinsed in 2 x SSC and dried. After a further rinse in 300 mM NaCl, the filters were dried and baked in vacuo for 2 hours at 80°C.

The filters were prehybridised in a minimal volume of 3 x SSC, 50 % formamide, 50 mM HEPES-NaOH pH 7, 5 x Denhardt's solution, 500 µg/ml total yeast RNA, 10 µg/ml sheared denatured salmon sperm DNA at 43°C for 4 hours. Radioactive cDNA was added to 10<sup>6</sup> cpm per ml, and the filters were incubated for 60 hours at 43°C. After rinsing several times in 2 x SSC at room temperature,

the filters were washed four times in 2 x SSC for 15 minutes each time at room temperature, then for two 15 minute periods in 0.1 x SSC at 65°C, and finally rinsed in 0.1 x SSC, dried, and exposed to X-ray film at -80°C with intensifying screens.

The cDNA probes were prepared by priming synthesis with oligomeric calf thymus DNA (227,228). Two µg of poly(A)<sup>+</sup> mRNA were reverse transcribed at 37°C for 45 minutes in the presence of 250 µCi of α-(<sup>32</sup>P)dCTP, 100 ng/µl of oligomeric calf thymus DNA (denatured by heating), 1 unit/µl reverse transcriptase, 50 mM tris-HCl pH 8.3, 40 mM KCl, 8 mM MgCl<sub>2</sub>, 200 µM dATP, dGTP and dTTP, and 0.5 mM DTT.

After the incubation 5 µg of denatured salmon sperm DNA was added and the mixture was extracted with phenol:chloroform (1:1) and ethanol precipitated. The material was resuspended in 300 mM NaOH and incubated at 60°C for 50 minutes to hydrolyse the RNA. Finally the probe was reprecipitated with ethanol, thoroughly washed and used as described above.

#### 2F2 Oligomer hybridisation

The library was again screened by colony hybridisation, this time using an oligomer of 20 bases, in a mixed synthesis of 16 sequences, provided by Celltech Ltd.

The oligomer was designed from the published amino acid sequences of the ricin A and B chains (33,35): the region whose coding sequence is least ambiguous is in the B chain, and encodes residues 214 - 220:

Amino acids:	N	-	Trp	Met	Phe	Lys	Asn	Asp	Gly	-	C
mRNA sequence:	5'	-	UGG	AUG	UU <sup>U</sup> <sub>C</sub>	AA <sup>A</sup> <sub>G</sub>	AA <sup>U</sup> <sub>C</sub>	GA <sup>U</sup> <sub>C</sub>	GGN	-	3'

The oligomer was complementary to the mRNA sequence:

Oligomer sequence: 5' - CC  $\begin{matrix} A \\ G \end{matrix}$ TC  $\begin{matrix} A \\ G \end{matrix}$ TT  $\begin{matrix} T \\ C \end{matrix}$ TT  $\begin{matrix} A \\ G \end{matrix}$ AA CAT CCT - 3'

The conditions for hybridisation and washing were determined by the '2-4' rule, by which, starting with 0°C, 2°C are added for every A or T residue, and 4°C for every G or C residue. The resulting temperature is the apparent melting point in 6 x SSC (229): hybridisation is carried out 10 - 15°C below this, and the most stringent wash is done at this temperature.

In this case, the '2-4' rule gives apparent melting temperatures with a range of 52 - 60°C. To cover this range, the screening was done in triplicate, with hybridisation at 37°C and the stringent washes at 52, 56 and 60°C.

The oligomer was labelled with polynucleotide kinase as described in section 2B7, with the oligomer at 10 ng/μl, enzyme at 0.2 units/μl, and  $\gamma$ -(<sup>32</sup>P)ATP at 2.5 μCi/μl. After incubation at 37°C for 30 minutes, an equal volume of 600 mM NH<sub>4</sub>Ac was added and the oligomer was purified by gel filtration on a 1 ml spinning column (205) of Sephadex G-100. An aliquot was assayed for incorporation of radioactivity by counting in Beckman aqueous scintillant cocktail.

The library was grown up on nitrocellulose filters and prepared as described in the previous section. They were prehybridised for 2 hours at 55°C in 900 mM NaCl, 90 mM tris-HCl pH 7.4, 6 mM EDTA, 0.5 % NP40, 2 x Denhardt's solution, 100 μg/ml denatured E coli DNA, and 70 μg/ml yeast tRNA (T Harris, personal communication). This buffer was removed and a minimal volume of the same solution was added, but containing 10<sup>6</sup> cpm per ml of radioactive oligomer. Hybridisation was at 37°C for 16 hours.

The filters were washed in 6 x SSC: after three 30 minute washes at room temperature, stringent washes were performed at the temperatures described for 15 minutes, and the filters were finally rinsed at room temperature, and exposed to X-ray film at  $-80^{\circ}\text{C}$  with intensifying screens.

### 2F3 Hybridisation - selected translation

All colonies which were positive by hybridisation with the oligomer were subjected to small-scale plasmid preparations, and the eight largest were confirmed by translation of hybridisation-selected mRNA (230).

The plasmids were linearised (or in some cases were cut into two fragments) with EcoRI, phenol extracted and ethanol precipitated. Aliquots of 10 - 15  $\mu\text{g}$  of each plasmid were denatured in 1 ml of 0.5 M NaOH in 0.05 x SSC at room temperature for 15 minutes, after which the alkali was neutralised by the addition of 5 ml of 250 mM tris-HCl pH 8, 250 mM HCl, 1.5 M NaCl, and the samples were chilled. The following steps were carried out in the cold. A nitrocellulose filter (S&S type BA85/1) was wetted in water and placed in a suction tower with a negative pressure of 20 cm of water and washed with 5 ml of water. The sample was then applied and run through, after which the filter was washed with 5 ml of 6 x SSC. The filters were then dried and baked in vacuo for 2 hours at  $80^{\circ}\text{C}$ , and then prehybridised in 0.5 ml of 50 % formamide, 400 mM NaCl, 10 mM PIPES-NaOH pH 6.4, 4 mM EDTA, 500  $\mu\text{g}/\text{ml}$  E coli tRNA, and 10  $\mu\text{g}/\text{ml}$  poly(A) for 4 hours at  $41^{\circ}\text{C}$ . Poly(A)<sup>+</sup> mRNA from maturing castor beans was added to fresh hybridisation buffer to a concentration of 40  $\mu\text{g}/\text{ml}$  and 0.5 ml of this replaced the prehybridisation buffer. After incubation at  $41^{\circ}\text{C}$  for

16 hours, the buffer was removed and kept for recycling of the mRNA. The filters were washed in 2 ml quantities of 1 x SSC, 0.1 % SDS (room temperature, 30 minutes, twice), then in 0.1 x SSC, 0.1 % SDS (room temperature, 15 minutes, twice), then for 15 minutes in 0.1 x SSC, 0.1 % SDS at 50°C. They were finally rinsed in 0.1 x SSC at room temperature for 15 minutes.

RNA was released from the filters at 40°C for 30 minutes in 200 µl of 10 mM PIPES-NaOH pH 6.4, 1 mM EDTA, 0.5 % SDS, 90 % formamide, and rinsed in 200 µl of water, which was added to the release buffer. Salt was added to 200 mM and the RNA was precipitated with ethanol. An aliquot was translated and immunoprecipitated as described in section 2B6 and products were analysed on SDS-polyacrylamide gels.

16 hours, the buffer was removed and kept for recycling of the mRNA. The filters were washed in 2 ml quantities of 1 x SSC, 0.1 % SDS (room temperature, 30 minutes, twice), then in 0.1 x SSC, 0.1 % SDS (room temperature, 15 minutes, twice), then for 15 minutes in 0.1 x SSC, 0.1 % SDS at 50°C. They were finally rinsed in 0.1 x SSC at room temperature for 15 minutes.

RNA was released from the filters at 40°C for 30 minutes in 200 µl of 10 mM PIPES-NaOH pH 6.4, 1 mM EDTA, 0.5 % SDS, 90 % formamide, and rinsed in 200 µl of water, which was added to the release buffer. Salt was added to 200 mM and the RNA was precipitated with ethanol. An aliquot was translated and immunoprecipitated as described in section 2B6 and products were analysed on SDS-polyacrylamide gels.

## 2G Subcloning into pUC8

The inserts of a number of the original pBR322 clones were subcloned into pUC8 (231) to facilitate the sequencing of the insert ends, the construction of entire coding sequences from overlapping clones, and construction of expression plasmids.

The inserts were excised with PstI and purified from agarose gels. pUC8 DNA was also cut with PstI, and was then dephosphorylated with calf intestinal alkaline phosphatase as described in section 2B7, and purified from an agarose gel.

The two DNA preparations were mixed in approximately equimolar quantities and ligated either overnight at 14°C or for 2 hours at room temperature in 50 mM tris-HCl pH 7.5, 5 mM MgCl<sub>2</sub>, 5 mM DTT, 10 mM ATP, 0.1 - 0.3 units/μl of T4 DNA ligase, with the DNA at a concentration of 10 - 100 ng/μl (total DNA concentration).

Competent E coli DH1 cells were prepared and transformed as described in section 2B3. Since the vector preparation was rigorous, and preparations were only used when transformation backgrounds were less than 100 clones per microgramme, no screening of the subclones was required, except by minimal restriction mapping. Important clones were confirmed by their subsequent use for sequencing.

## 2H M13 cloning and dideoxy sequencing

### 2H1 Cloning into M13

DNA fragments to be sequenced were excised from pBR322 with PstI and were gel purified. They were then digested, in separate reactions, with AluI and Sau3AI. M13 mp8 (232) was cut with SmaI and BamHI, for the AluI and Sau3AI products respectively. Cleaved M13 DNA was dephosphorylated as described in section 2B7, and gel purified. Vector DNA (100 ng) was mixed with 10 - 100 ng of target DNA and ligated overnight at 14°C as described in section 2G. Competent E coli JM101 cells (233) were rendered competent as described in section 2B3, using the simple CaCl<sub>2</sub> method. After the heat-shock, the cells were mixed with 3 ml of HTOP agar (1 % tryptone, 0.8 % NaCl, 0.8 % agar), containing 200 µg/ml BCIG, 190 µg/ml IPTG and a 1 in 15 dilution of log phase JM101 cells, and poured onto MM plates (1.5 % agar, 10.5 g/l K<sub>2</sub>HPO<sub>4</sub>, 4.5 g/l KH<sub>2</sub>PO<sub>4</sub>, 1 g/l (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 0.5 g/l sodium citrate, 0.01 % MgSO<sub>4</sub>, 0.1 % glucose, 0.00025 % vitamin B<sub>1</sub>). After overnight incubation at 37°C, white (recombinant) plaques were counted.

Single-stranded DNA was prepared from recombinant plaques by inoculating 2 ml cultures of JM101 cells (1 in 100 dilutions of stationary phase cultures) from plaques, and growing with vigorous shaking at 37°C for 5 - 6 hours in 2TY medium (1.6 % tryptone, 1 % yeast extract, 5 g/l NaCl). The cells were removed by centrifugation and virus particles were precipitated by the addition of NaCl to 300 mM and polyethylene glycol 6000 to 2.5 %, with a 10 minute incubation at room temperature. Phage particles were pelleted in a microfuge for 10 minutes, and resuspended in 100 µl of 10 mM

tris-HCl pH 8, 0.1 mM EDTA. After phenol extraction, DNA was precipitated with ethanol, dried, and dissolved in 25  $\mu$ l of 10 mM tris-HCl pH 8, 0.1 mM EDTA.

Most of the results presented were obtained with these libraries, though additional data were obtained with an AluI - HaeIII double digest, cloned into M13 mp8 cut with SmaI. Additionally, a library of DNase I - generated fragments was constructed as described in reference (234).

#### 2H2 Dideoxy sequencing

M13 recombinants were sequenced by the dideoxy chain termination method, essentially as described by Sanger et al (235,236).

Five  $\mu$ l of single-stranded M13 recombinant DNA was mixed with 5  $\mu$ l of 20 mM tris-HCl pH 8.3, 20 mM MgCl<sub>2</sub>, 1 mM DTT containing a universal sequencing primer at 1 ng/ $\mu$ l. Initially the primer provided by BRL was used, but this was found to be unreliable - one batch produced overlapping sequencing ladders, and 'stuttering' was commonly observed at the bottoms of the gels. Most work was done with the New England Biolabs primer, which anneals further from the insert than does the BRL product; no problems were encountered with this reagent. The mixture was heated to 95°C and cooled to room temperature over 30 - 40 minutes.

Each sequencing reaction was carried out by mixing 2  $\mu$ l of the template-primer with 2  $\mu$ l of a solution containing non-cognate dNTPs (other than dGTP) at 100  $\mu$ M, dGTP at 1.7  $\mu$ M,  $\alpha$ -(<sup>32</sup>P)dGTP at 0.3  $\mu$ Ci/ $\mu$ l, the cognate dNTP at 4.5  $\mu$ M, the cognate dideoxy-NTP at 20 - 100  $\mu$ M (exact concentrations were determined empirically), in 2.5 mM tris-HCl pH 8, 0.45 mM EDTA. In the G reaction, the concentration of dGTP was 1.7  $\mu$ M. After a 15 minute incubation,

2  $\mu$ l of 0.5 mM dGTP were added, as a chase, and the reactions were incubated for a further 15 minutes. At first, reactions were performed at room temperature, but extensive premature termination of the G reaction was observed in some 20 % of the clones. Subsequent reactions were performed at 37°C, at which this problem did not occur.

Reactions were terminated by addition of 4  $\mu$ l of 90 % formamide, 10 mM EDTA, 0.01 % each of bromophenol blue and xylene cyanol. and were heated to 95°C for 2 minutes. Gels were loaded with 3 - 5  $\mu$ l per lane, and were run as described in section 2C6. For short runs, buffer gradient gels were used, and for long runs, simple 1 x buffer gels were used.

## 2I Maxam and Gilbert sequencing

### 2I1 Preparation of DNA fragments

Fragments labelled at one end with ( $^{32}\text{P}$ ) were prepared by three approaches. For labelling protocols, see section 2B7; for purification from agarose gels see section 2C6. For each set of reactions, 50 - 500 ng of DNA was used, giving exposures to X-ray film of 15 - 60 hours.

In the first approach, linear DNA fragments were produced by cleavage with restriction enzymes, and all termini in the mixture were labelled (or a fragment was gel-purified if appropriate) either with polynucleotide kinase or with the Klenow fragment of DNA polymerase I from E coli. The resulting labelled ends were separated by cleavage with a second restriction enzyme and were purified from agarose gels.

The second approach consisted of preparing linear DNA fragments which could be labelled with the Klenow enzyme in such a way that only one end was able to be labelled. For example, a fragment terminating with BamHI and HindIII sites can only label at the BamHI end with Klenow enzyme and  $\alpha$ -( $^{32}\text{P}$ )dGTP, while a fragment with TaqI and BamHI ends can be labelled at either end selectively: at the former with  $\alpha$ -( $^{32}\text{P}$ )dCTP and at the latter with  $\alpha$ -( $^{32}\text{P}$ )dGTP.

Finally, a number of shorter fragments were prepared as single-stranded DNA: a double-stranded fragment labelled at both ends was denatured by heating to 95°C in 100  $\mu\text{l}$  of 90 % formamide, 10 mM EDTA, 0.01 % each of bromophenol blue and xylene cyanol, and cooled on iced water.

The sample was applied to a large slot in a 40 cm long, 1 mm thick, 10 % polyacrylamide, 0.26 % bisacrylamide gel and run at 40 V/cm until the xylene cyanol was two-thirds of the way down the gel. The gel was exposed wet to X-ray film for 1 - 2 hours, and radioactive bands were cut out. These were placed in a slot cut into a 1 % neutral agarose gel, sealed in with molten agarose, and eluted as described in section 2C6. This method of recovery was used in preference to direct elution from acrylamide gels because of speed and proven efficacy.

#### 2I2 Modification and cleavage reactions

These were performed essentially as described by Maxam & Gilbert (237), though with slight modifications. The reaction times described below were for fragments greater than 200 bases in length; shorter fragments were incubated for 2 - 5 times as long.

The purified end-labelled fragments were dissolved in 40  $\mu$ l of water containing 5  $\mu$ g of carrier sheared salmon sperm DNA (size 20 - 30 kbp). Ten  $\mu$ l of this were used for two-base reactions, and 5  $\mu$ l for single-base reactions.

#### G reaction

Five  $\mu$ l of DNA were mixed with 200  $\mu$ l of 50 mM sodium cacodylate pH 8, 10 mM  $MgCl_2$ , 1 mM EDTA and 1  $\mu$ l of DMS was added. The reaction was not constructed on ice as recommended, as this was found unnecessary. Incubation was for 5 minutes at room temperature, after which the reaction was stopped by the addition of 50  $\mu$ l of 1.5 M NaAc pH 7, 1 M 2-mercaptoethanol.

#### A + G reaction

Ten  $\mu$ l of DNA were mixed with 25  $\mu$ l of 98 % formic acid and incubated at room temperature for 3 minutes. The reaction was stopped

by addition of 200  $\mu$ l of 300 mM NaAc pH 6.

#### C reaction

Five  $\mu$ l of DNA were mixed with 10  $\mu$ l of 5 M NaCl and 30  $\mu$ l of hydrazine. After 7 minutes at room temperature, 200  $\mu$ l of 1 M acetic acid was added to terminate the reaction.

#### C + T reaction

This was done as the C reaction, except that 10  $\mu$ l of DNA were used, and 10  $\mu$ l of water were used in place of the NaCl.

#### A + C reaction

Ten  $\mu$ l of DNA were mixed with 100  $\mu$ l of 1.2 M NaOH, 20 mM EDTA, and were incubated at 90°C for 10 minutes. The reaction was stopped by the addition of 150  $\mu$ l of acetic acid.

#### Preparation of products for cleavage

Five  $\mu$ l of 1  $\mu$ g/ $\mu$ l E coli tRNA were added to each terminated sequencing reaction, and products were precipitated with 3 volumes of ethanol at -80°C for 10 minutes. Modified DNA was recovered by centrifugation in an Eppendorf centrifuge in the cold for 10 minutes. After washing in 70 % ethanol, the DNA was redissolved in 200  $\mu$ l of 300 mM NaAc pH 6 and reprecipitated with 3 volumes of ethanol. After pelleting, the DNA was again washed with 70 % ethanol. Finally, the samples were dried briefly in vacuo.

#### Cleavage reactions

All samples were redissolved in 100  $\mu$ l of freshly diluted 1 M piperidine and heated to 90°C for 30 minutes. Piperidine was

removed by lyophilisation, first of the reaction mixture, then from 200  $\mu$ l of water, and finally from 20  $\mu$ l of water.

The resulting products were dissolved in 5 - 15  $\mu$ l of 90 % formamide, 10 mM EDTA, 0.01 % each of bromophenol blue and xylene cyanol, and sequencing gels were run as described in section 2C5.

Buffer gradient gels were not used for Maxam & Gilbert sequencing, as the excessive spacing of bands in the lower part of the gel observed with dideoxy sequencing products was not found: presumably the fragments move more slowly when produced chemically, probably because of the much greater amounts of DNA and carrier RNA present.

## 2J Primer extension

### 2J1 Cloning of the primer

Primer extension was performed in order to estimate the length of sequence at the 5' end of the mRNA not represented in the clones sequenced. A primer was prepared from one of the lectin clones (pRCL6), and subcloned into pUC8 - this approach was used because, in order to produce a primer fragment from the clone directly, difficult fragment separations would be required.

Fig 2J-1 shows a partial restriction map of clone pRCL6. The 99 bp BamHI - AluI fragment (fragment A) was cloned into pUC8 as follows: first, the BamHI-BglII fragment (fragment B) was cleaved from clone pUC617 (a pUC8 subclone of pRCL6), and gel purified. This fragment was cleaved with AluI to generate the desired fragment, along with several others, including a 110 bp AluI-BglII fragment (fragment C). In order to prevent the cloning of this fragment, the mixture was digested with HhaI, which produces an incompatible terminus. In the resulting mixture, the only fragment with one blunt end and one end compatible with a BamHI site is the 99 bp target fragment, fragment A. pUC8 DNA was cleaved with BamHI and SmaI, dephosphorylated and gel purified. The two DNAs were mixed and ligated overnight, and E coli DH1 cells were transformed with the products.

Plasmids were prepared from 10 of the colonies obtained, and were cleaved with EcoRI and BamHI to release any inserts. The cleaved samples were labelled with Klenow enzyme and  $\alpha$ -(<sup>32</sup>P)dGTP to label at the BamHI site, and run on a neutral agarose gel. One of the 10 plasmids (designated pPRIM8) contained an insert of the

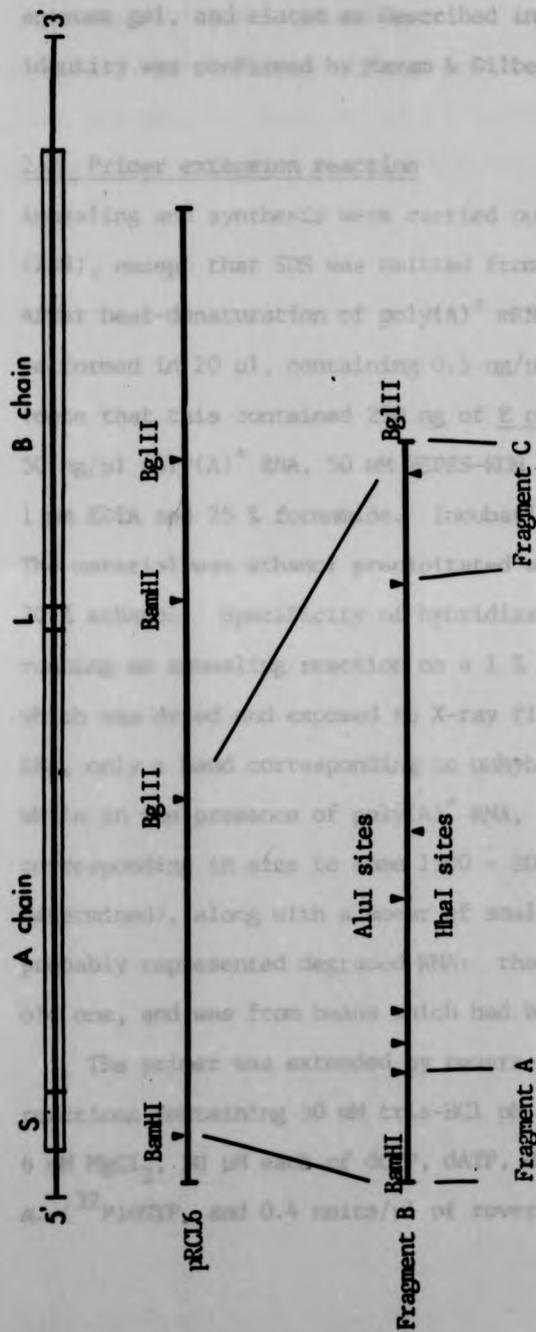


Fig 2J-1: Partial restriction map of pRCL6 to show subcloned primer fragment

expected size: this was purified from a DNA sequencing gel (the band was cut out from this gel, placed in a slot in a neutral agarose gel, and eluted as described in section 2C6), and its identity was confirmed by Maxam & Gilbert sequencing.

#### 2J2 Primer extension reaction

Annealing and synthesis were carried out after the method of Hall et al (238), except that SDS was omitted from the annealing stage. After heat-denaturation of poly(A)<sup>+</sup> mRNA, hybridisation was performed in 20  $\mu$ l, containing 0.5 ng/ $\mu$ l of end-labelled primer (note that this contained 250 ng of E coli tRNA coprecipitant), 50 ng/ $\mu$ l poly(A)<sup>+</sup> RNA, 50 mM HEPES-KOH pH 7.6, 450 mM NaCl, 1 mM EDTA and 25 % formamide. Incubation was at 43°C for 4 hours. The material was ethanol precipitated and thoroughly washed with 70 % ethanol. Specificity of hybridisation was ascertained by running an annealing reaction on a 1 % neutral agarose gel, which was dried and exposed to X-ray film. In the absence of poly(A)<sup>+</sup> RNA, only a band corresponding to unhybridised primer was visible, while in the presence of poly(A)<sup>+</sup> RNA, an extra band appeared, corresponding in size to some 1500 - 2000 bases (exact size not determined), along with a smear of smaller material. This latter probably represented degraded RNA: the preparation used was an old one, and was from beans which had been stored for a long period.

The primer was extended by reverse transcription in 10  $\mu$ l reactions containing 50 mM tris-HCl pH 8.3, 60 mM NaCl, 20 mM DTT, 6 mM MgCl<sub>2</sub>, 50  $\mu$ M each of dCTP, dATP, dTTP, 4  $\mu$ M dGTP, 1  $\mu$ Cl/ $\mu$ l  $\alpha$ -(<sup>32</sup>P)dGTP, and 0.4 units/ $\mu$ l of reverse transcriptase.

The mixture was incubated at 37°C for 30 minutes, and was then lyophilised. After dissolving in 5 µl of sequencing gel loading buffer, the products were denatured by heating to 95°C for 2 minutes, and were run on a sequencing gel, along with either dideoxy or Maxam and Gilbert sequence ladders.

## 2K Nucleic acid transfer and hybridisation

### 2K1 Southern blots (ref 239)

DNA samples were run on neutral agarose gels as described. Gels were soaked for 30 minutes in 10 volumes of 400 mM NaOH, 800 mM NaCl to denature the DNA. The alkali was neutralised by soaking in 500 mM tris-HCl pH 7.6, 1.5 M NaCl. The gel was then placed on a support covered with 3MM paper running down into a tank containing 1 litre of 10 x SSC, to act as wicks, and a sheet of nitrocellulose (S & S type BA85/1), wetted with 10 x SSC, was placed on the gel. The edges were covered with Parafilm sealing tape, and 4 sheets of 3MM paper and a 6 inch pile of tissues were placed on top. After overnight transfer the stack was dismantled, and the filter was dried and baked in vacuo for 2 hours at 80°C.

Blots to be probed with oligonucleotides were prehybridised and hybridised and washed as described in section 2F1.

Other probes were labelled by nick translation, by the method of Rigby et al (240). Blots were prehybridised in 3 x SSC, 50 % formamide, 50 mM HEPES-NaOH pH 7, 5 x Denhardt's solution, 10 µg/ml sheared denatured salmon sperm DNA and 500 µg/ml total yeast RNA. The filters were prehybridised for 4 hours at 43°C, and 10<sup>6</sup> cpm per ml of probe was added, and hybridisation was for 16 hours at 43°C. Washes were: 3 times for 40 minutes each in 2 x SSC at room temperature, 10 minutes at 55°C in 0.2 x SSC, and a rinse in 0.1 x SSC at room temperature.

### 2K2 Northern blots

RNA samples on formamide agarose gels were prepared as described in section 2C3. RNA was transferred to nitrocellulose as above; prehybridisation, hybridisation and wash conditions were as above except that poly(U) was added to 20 µg/ml, and the stringent wash was at 65°C in 0.1 x SSC.

### CHAPTER 3: RESULTS AND DISCUSSION

#### 3A Isolation and characterisation of mRNA

Castor beans at developmental stages D - E (81) were harvested and total RNA was extracted. After extensive washing with 3 M NaAc pH 5.5, the material was run on an oligo(dT)-cellulose column to purify the poly(A) - containing fraction. Typically, 150 - 200 mg of total RNA was obtained, of which 500 - 600  $\mu$ g was polyadenylated (0.25 - 0.40 %). This compares with a proportion of 0.5 - 1.0 % obtained when castor bean mRNA was characterised previously (241).

The presence of mRNA species directing the synthesis of the lectin precursors was shown by analysis of the products of in vitro translation: Fig 3A-1 shows the time-course of incorporation of radioactivity from (<sup>35</sup>S)methionine into acid-insoluble products, and fig 3A-2 shows the SDS-polyacrylamide gel analysis of these products. In the total products of translation, a considerable number of proteins are synthesised, with a wide range of molecular weights. The immunoprecipitated sample shows two bands, one at an apparent  $M_r$  of 60 kDal and the other with an apparent  $M_r$  of 34 kDal. These correspond respectively with the lectin precursor identified by Butterworth and Lord (45) and the castor bean albumin precursor previously thought to be the A chain precursor (44).

This poly(A)<sup>+</sup> RNA was fractionated on the basis of size by sucrose density gradient centrifugation. Thirty fractions were collected, and the RNA in each was precipitated with ethanol, and a sample was translated in a rabbit reticulocyte system.

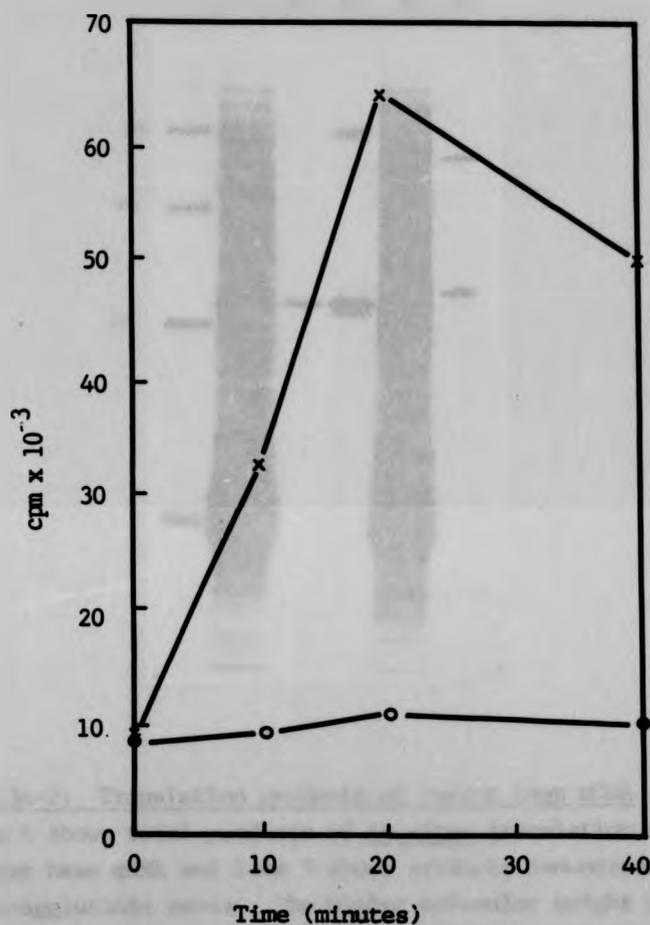


Fig 3A-1: Time course for polypeptide synthesis in reticulocyte lysate supplemented with 40  $\mu\text{g/ml}$  poly(A)<sup>+</sup> RNA from maturing castor beans. 1  $\mu\text{l}$  samples were removed from a 25  $\mu\text{l}$  assay and hot TCA - insoluble material was determined by scintillation counting.

o = No added RNA  
x = Poly(A)<sup>+</sup> mRNA added

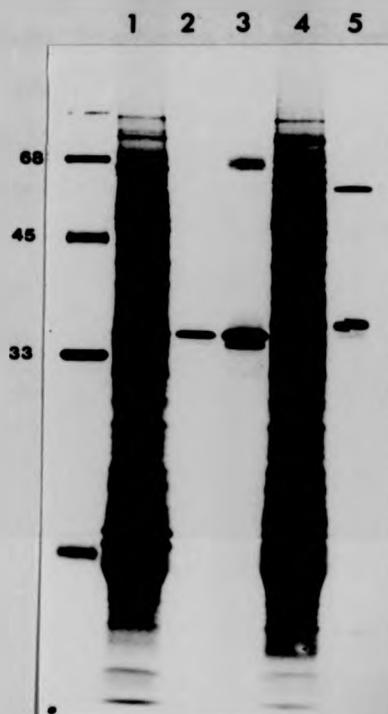


Fig 3A-2: Translation products of castor bean mRNA

Lane 4 shows total products of in vitro translation of developing castor bean mRNA and lane 5 shows products immunoreactive with anti-agglutinin serum. The higher molecular weight products are the lectin precursor, and the lower molecular weight bands are the albumin precursor.

Lane 1 shows total products obtained in the presence of dog pancreas membranes; the reaction was fractionated by centrifugation and immunoreactive products are visible in lanes 2 (supernatant) and 3 (pellet). Compartmentalisation of the lectin precursor is evident, along with an increase in its apparent molecular weight as compared with the product obtained in the absence of membranes.

Markers as fig 3A-3.

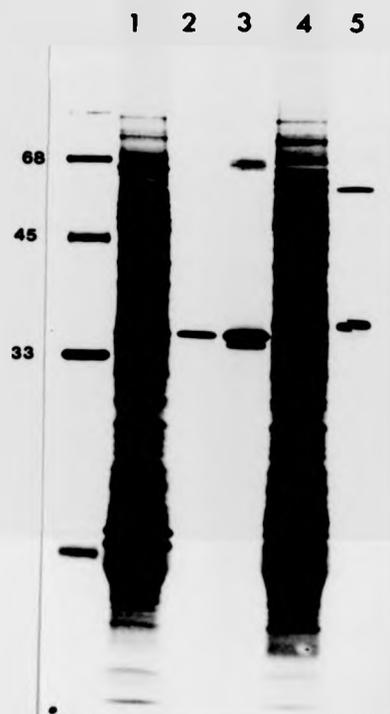


Fig 3A-2: Translation products of castor bean mRNA

Lane 4 shows total products of *in vitro* translation of developing castor bean mRNA and lane 5 shows products immunoreactive with anti-agglutinin serum. The higher molecular weight products are the lectin precursor, and the lower molecular weight bands are the albumin precursor.

Lane 1 shows total products obtained in the presence of dog pancreas membranes; the reaction was fractionated by centrifugation and immunoreactive products are visible in lanes 2 (supernatant) and 3 (pellet). Compartmentalisation of the lectin precursor is evident, along with an increase in its apparent molecular weight as compared with the product obtained in the absence of membranes.

Markers as fig 3A-3.

Fig A3-3 shows the total products of translation of each fraction, and fig A3-4 shows the products immunoprecipitated with anti-RCA serum. The lectin precursor mRNA is predominantly present in fractions 25 - 27, while the albumin precursor is in fractions 16 - 19. The appearance of smaller species towards the bottom of the gradient is probably a result of RNA aggregation, even though the material was heat-denatured prior to loading on the gradient. Translational activity was fairly constant over the gradient, each fraction stimulating the activity of a lysate by a factor of 7 - 12 (data not shown).

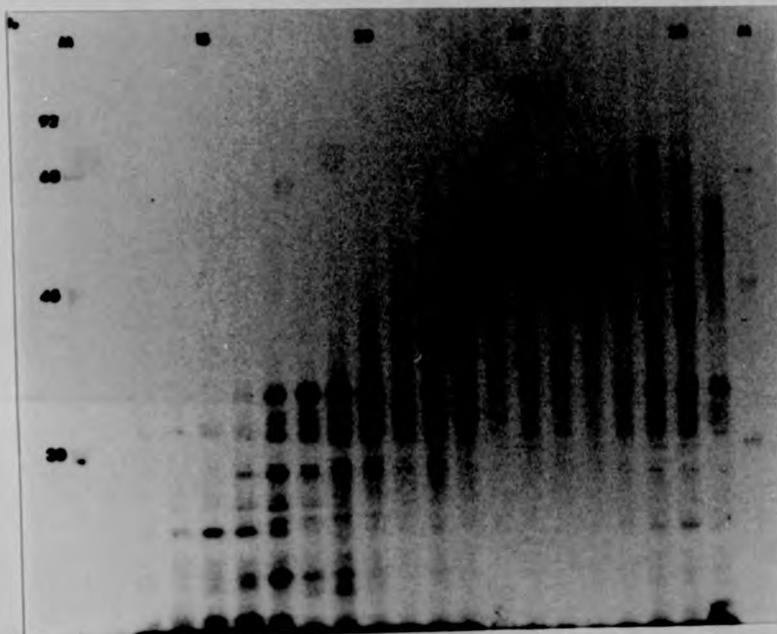


Fig 3A-3: Sucrose density gradient fractionation of mRNA. RNA was recovered from fractions and translated in rabbit reticulocyte lysates in 25  $\mu$ l assays. 3  $\mu$ l of each assay were analysed by SDS-PAGE.

MW markers: carbonic anhydrase (30 kDal), ovalbumin (45 kDal), BSA (68 kDal) and phosphorylase b (92 kDal).

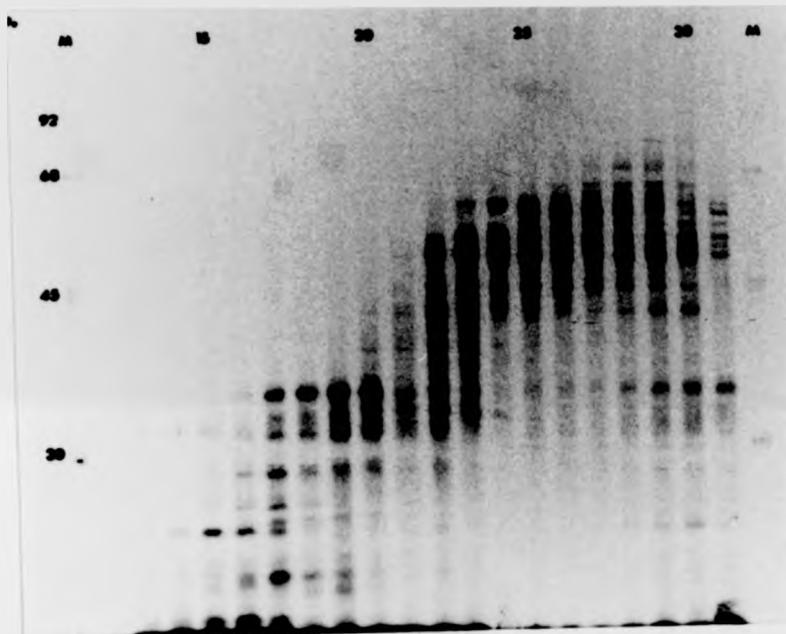


Fig 3A-3: Sucrose density gradient fractionation of mRNA. RNA was recovered from fractions and translated in rabbit reticulocyte lysates in 25  $\mu$ l assays. 3  $\mu$ l of each assay were analysed by SDS-PAGE.

MW markers: carbonic anhydrase (30 kDal), ovalbumin (45 kDal), BSA (68 kDal) and phosphorylase b (92 kDal).

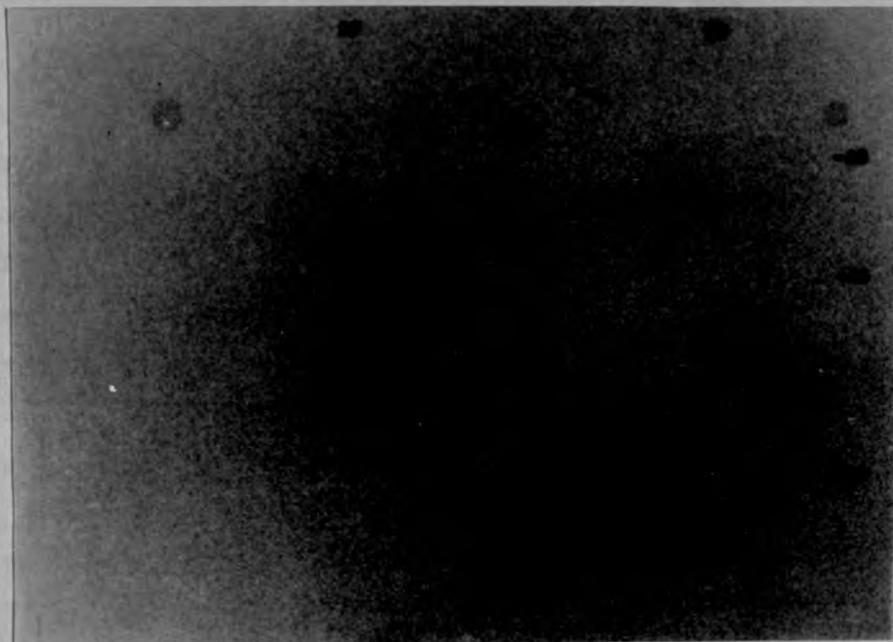


Fig 3A-4: Sucrose density gradient fractionation of mRNA. Aliquots of the translation products shown in fig 3A-3 were immunoprecipitated with anti-RCA serum and analysed by SDS-PAGE.

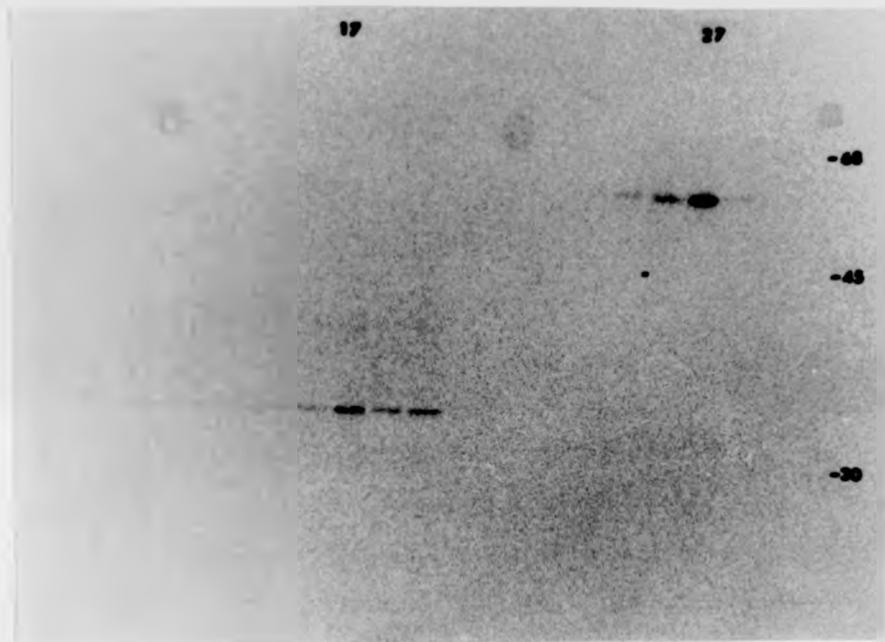


Fig 3A-4: Sucrose density gradient fractionation of mRNA. Aliquots of the translation products shown in fig 3A-3 were immunoprecipitated with anti-RCA serum and analysed by SDS-PAGE.

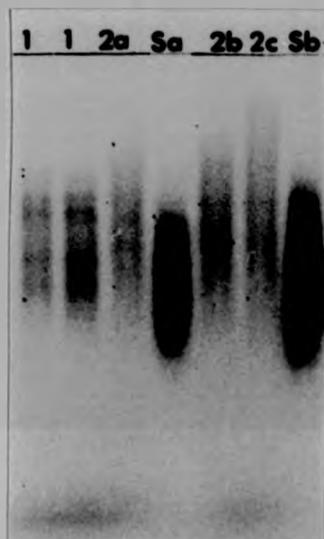
### 3B cDNA synthesis

First strand synthesis reactions as described in section 2E1 generally worked well, producing cDNAs of the expected size range: typically 1400 - 2000 bases from RNA fractions 25 - 27.

Early attempts at second strand synthesis were less successful, as were those for the  $S_1$  nuclease reaction. Incubation of the second strand synthesis at 15°C, using DNA polymerase I produced length increments over the first strands of only 30 - 40 %, with enzymes provided by several colleagues. The use of reverse transcriptase for this reaction did not result in any improvement - nor did an attempt using the cDNA synthesis method of Land et al (242), in which homopolymer tails are added to the first strand reaction products, the second strand being primed with the complementary oligo(dN) (results not shown).

A fresh supply of DNA polymerase I from Boehringer produced products of the desired length: different temperatures were tried, 30°C producing the longest second strands. However, it is known that at higher temperatures, many of the longer molecules are probably artefactual (243), so a compromise temperature of 20°C was subsequently used. Fig 3B-1 shows products of cDNA synthesis and  $S_1$  cleavage reactions, with the second strand reactions at 15, 20 and 30 °C. Size markers are not shown here, but the second strands are approximately double the length of the first strands.

Fig 3B-2 shows the cDNA synthesis and  $S_1$  cleavage products which were used for the following cloning experiments.



**Fig 3B-1: cDNA synthesis.**

**Effect of temperature on second strand synthesis with DNA pol I.**  
Lanes 1 are two different first strand reactions. Lanes 2 are second strand reactions and lanes S are products of  $S_1$  cleavage. Lanes a are at  $15^{\circ}\text{C}$ , b at  $20^{\circ}\text{C}$  and c at  $30^{\circ}\text{C}$ .

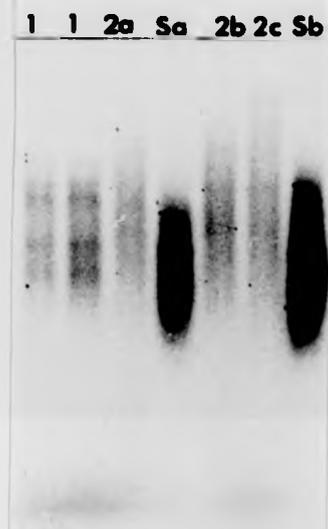


Fig 3B-1: cDNA synthesis.

Effect of temperature on second strand synthesis with DNA pol I.  
Lanes 1 are two different first strand reactions. Lanes 2 are  
second strand reactions and lanes S are products of  $S_1$  cleavage.  
Lanes a are at  $15^{\circ}\text{C}$ , b at  $20^{\circ}\text{C}$  and c at  $30^{\circ}\text{C}$ .

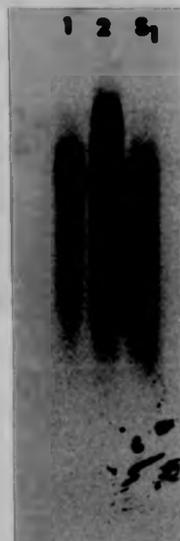


Fig 3B-2: Synthesis of cDNA used for cloning. Aliquots of the first strand synthesis with reverse transcriptase (lane 1), second strand synthesis with DNA pol I at 20°C (lane 2) and S<sub>1</sub> cleavage reactions were analysed on a 1 % alkaline agarose gel.



Fig 3B-2: Synthesis of cDNA used for cloning. Aliquots of the first strand synthesis with reverse transcriptase (lane 1), second strand synthesis with DNA pol I at 20°C (lane 2) and S<sub>1</sub> cleavage reactions were analysed on a 1 % alkaline agarose gel.

Double stranded cDNA cleaved with  $S_1$  nuclease was tailed with dCTP and terminal transferase. The reaction was assayed by passing the products through a small column of Biogel P60; an aliquot of the excluded fraction was TCA precipitated and its radioactivity counted, and another aliquot was run on a 1% neutral agarose gel and exposed to X-ray film. Such a gel is shown in fig 3B-3: lane 1 indicates that the untailed cDNA was labelled (during its synthesis), and that the tailed cDNA in lane 2 contains far more radioactivity.

The resulting tailed cDNA was annealed with PstI-cut, dG-tailed pBR322 and transformed into E coli DH1 cells (see next section). Recombinants were identified by resistance to tetracycline and sensitivity to ampicillin, and small-scale plasmid preparations were carried out: only four of the 68 recombinants obtained were detectably larger than the vector plasmid, as judged by agarose gel electrophoresis. For this reason, the tailed cDNA was size fractionated on a sucrose density gradient. A mixture of restriction fragments of pBR322 were run on a parallel gradient as markers, and the cDNA gradient was loaded with 2.25  $\mu$ g of material. 21 fractions were collected, and pooled to form three final fractions:

(a) Small cDNA	600 - 1000 bp	25 % of input label
(b) Medium cDNA	1000 - 2200 bp	13 % of input label
(c) Large cDNA	greater than 2200 bp	3.5 % of input label

Results of transformation with these fractions are presented in the next section.

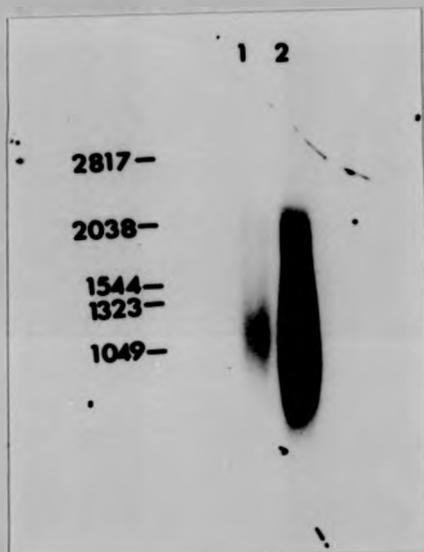


Fig 3B-3: Tailing of cDNA.

Double-stranded cDNA labelled during synthesis (lane 1) was tailed with  $\alpha$  - ( $^{32}\text{P}$ )dCTP and terminal transferase, and an aliquot was analysed on a 1 % neutral agarose gel (lane 2). Size markers are restriction fragments of pBR322.

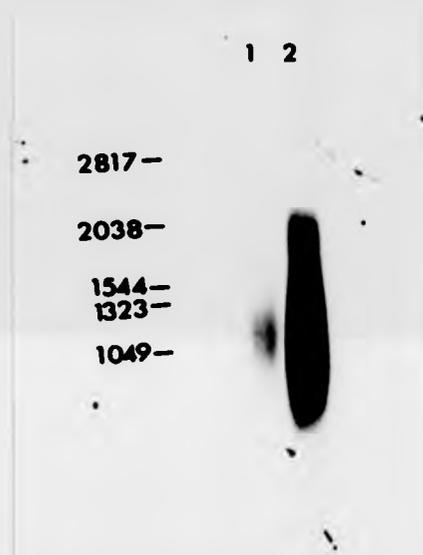


Fig 3B-3: Tailing of cDNA.

Double-stranded cDNA labelled during synthesis (lane 1) was tailed with  $\alpha$  -( $^{32}$ P)dCTP and terminal transferase, and an aliquot was analysed on a 1 % neutral agarose gel (lane 2). Size markers are restriction fragments of pBR322.

### 3C Construction of clones

Attempts to tail PstI - cut pBR322 with dG were of dubious success. On no occasion did the incorporated counts indicate that more than 1,- 2 bases had been added per 3' terminus. The products of one reaction, in which some labelling was apparent, were digested with HaeIII. This should produce two bands, one of 145 bp plus tail and one of 122 bp plus tail. Only one clear band was produced, corresponding to the larger of these, along with a number of faint bands of larger size. It is possible that significant amounts of tailing were taking place at single-stranded breaks in the DNA. To avoid further problems, commercially prepared PstI-cut, dG-tailed pBR322 was obtained from BRL.

Before recombinants were constructed, a transformation procedure was characterised, which gave good and consistent results. Initially, a protocol in use at EMBL was tried, whereby competent E coli cells were prepared by two incubations on ice in 100 mM CaCl<sub>2</sub> pH 8, followed by a heat-shock. This was tried with E coli X1776 (244), HB101 (245) and DH1 (202). Although a wide variety of conditions were tried, the transformation efficiency never exceeded some 2000 transformants per µg of supercoiled pBR322 DNA. Omitting the pH adjustment of the CaCl<sub>2</sub> improved the efficiency to around  $5 \times 10^4$  per µg, but this was still considered inadequate. In all cases, cell viability was checked by taking control cells through the procedure and plating dilutions on antibiotic-free medium. Typically, there were some  $0.2 - 2.0 \times 10^6$  viable cells per µl.

The procedure described in Methods section 2B3 (RbCl/MnCl<sub>2</sub> method) was then tried, and has consistently given some 10<sup>7</sup> - 10<sup>8</sup> transformants per µg of supercoiled pBR322 which is satisfactory.

The fractionated, tailed cDNA was then annealed with the vector as described and transformed using the above protocol. A total of 1587 recombinants were obtained, of which 150 were examined on agarose gels to check for the presence of inserts:

<u>cDNA fraction</u>	<u>No of recombinants</u>	<u>% with large inserts</u>
Total (section 3B)	68	6
Large (this section)	52	40
Medium ( " " )	672	72
<u>Small ( " " )</u>	<u>795</u>	<u>55</u>

In this context, "large insert" means that the supercoiled recombinant plasmids were detectably larger than the vector. This is not an accurate measurement, but is adequate for the present purpose.

Thus a library of 1587 recombinant plasmids was obtained. All of these were subjected to screening for the presence of lectin - encoding insert sequences.

### 3D Screening of the cDNA library

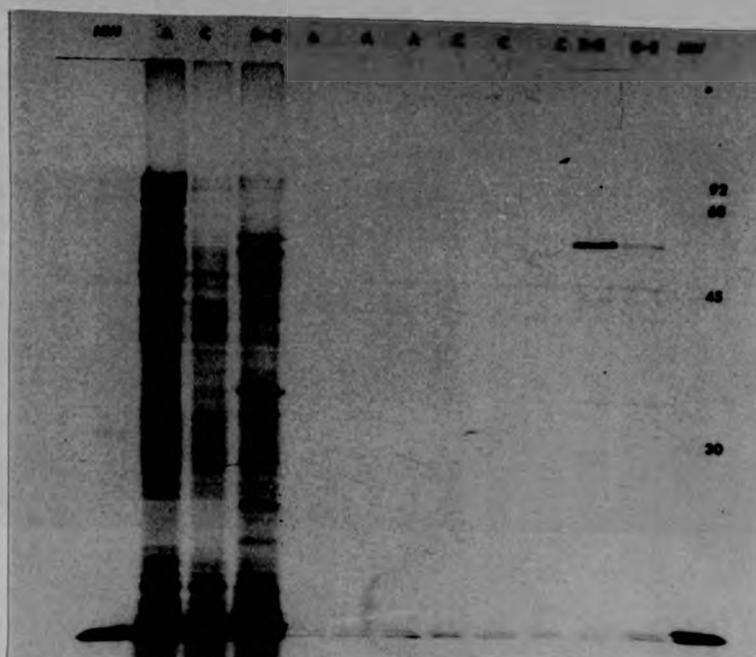
The first screening exploited the developmental regulation of lectin mRNAs: these species are translatable during stages D-F, but are absent during stages A-C (81). Polyadenylated mRNA was extracted from beans at stages A-C ("early") and from stages D-F ("late"). Aliquots of various preparations of these mRNAs were translated in rabbit reticulocyte lysates and the products were immunoprecipitated and analysed on SDS-polyacrylamide gels. This is shown in fig 3D-1, from which it is clear that the late mRNAs do indeed contain lectin mRNAs which are not present in the early mRNAs.

These mRNAs were reverse transcribed to produce single-stranded cDNA probes, using oligomeric calf thymus DNA as primer.

The library was grown up on nitrocellulose filters and the colonies were lysed. The filters were hybridised first with the "late" probe, then washed and rehybridised with the "early" probe. This order was used to avoid false positives which might have arisen if the "early" probe had been used first, by incomplete removal by washing of non-lectin cDNAs bound to clones of non-lectin mRNAs. Colonies of cells containing pBR322 did not hybridise with either probe.

348 colonies were identified as developmentally regulated. These were replated onto duplicate filters, one set being probed with each cDNA preparation. Fig 3D-2 shows an example: most of the colonies hybridise with the "late" probe, though some (arrowed) hybridise with both probes. A few, none of which are shown on this filter, hybridised only with the "early" probe.

At this time, an oligonucleotide probe was made available by Celltech Ltd. This was designed from the published sequences of the ricin A and B chains: for details of design and determination of



**Fig 3D-1: Translation of early and late castor bean mRNAs**

Poly(A)<sup>+</sup> mRNA was extracted from castor beans at the developmental stages A, C and D+E (indicated above lanes) and translated in rabbit reticulocyte lysates. The first three lanes after the molecular weight markers show total products, while the remaining lanes show products immunoreactive with anti-agglutinin serum, each lane being from a different batch of mRNA. Developmental stages are defined in ref 81.

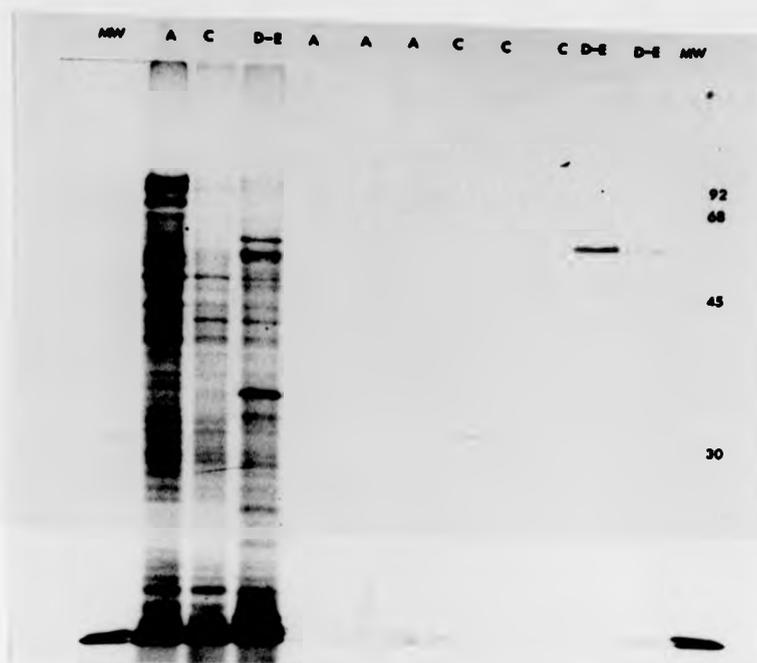


Fig 3D-1: Translation of early and late castor bean mRNAs

Poly(A)<sup>+</sup> mRNA was extracted from castor beans at the developmental stages A, C and D+E (indicated above lanes) and translated in rabbit reticulocyte lysates. The first three lanes after the molecular weight markers show total products, while the remaining lanes show products immunoreactive with anti-agglutinin serum, each lane being from a different batch of mRNA. Developmental stages are defined in ref 81.

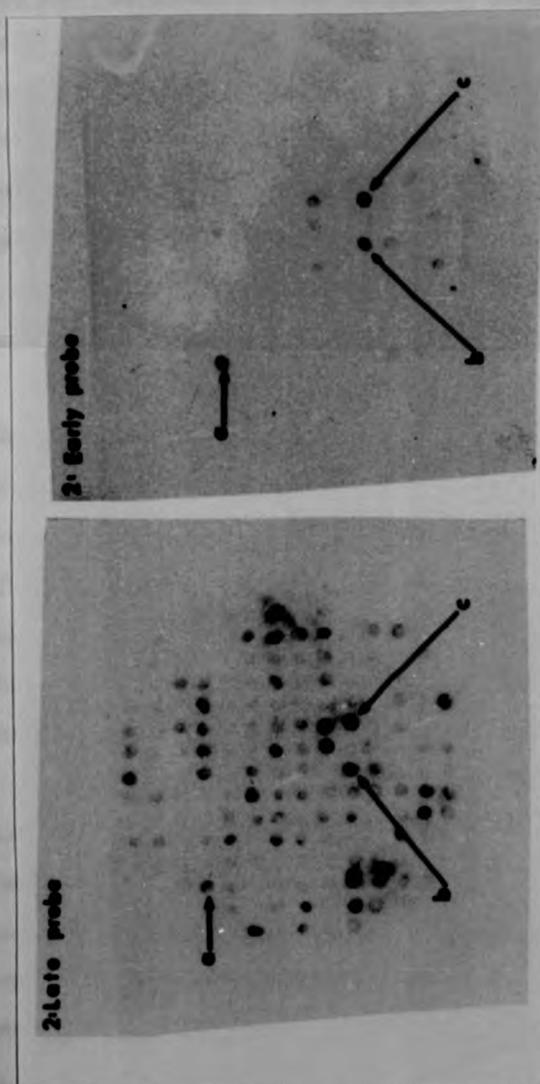


Fig 3D-2: Rescreening of cDNA library with early and late probes. Three colonies which hybridise with both probes are arrowed.

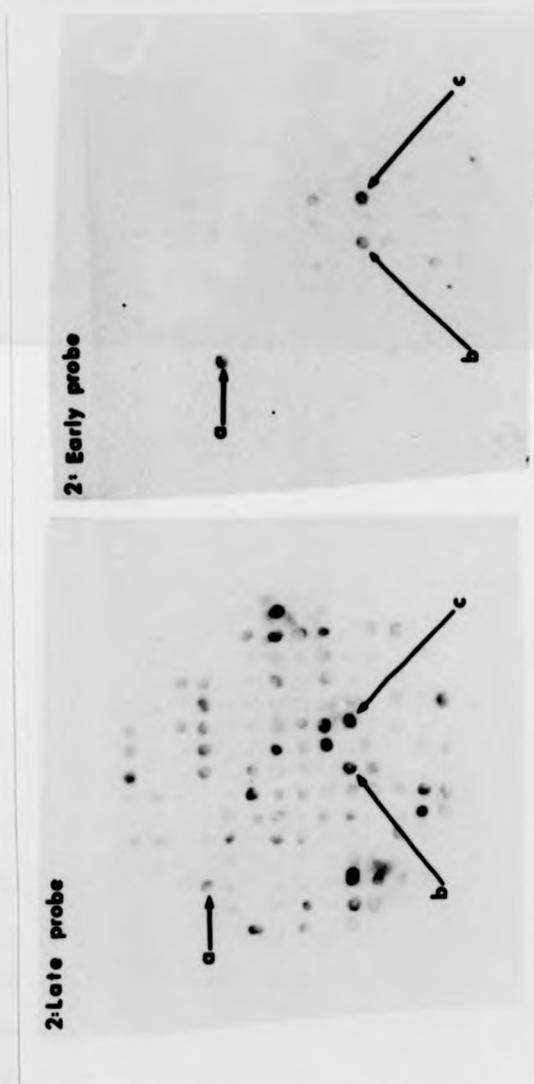


Fig 3D-2: Rescreening of cDNA library with early and late probes.  
Three colonies which hybridise with both probes are arrowed.

hybridisation and wash conditions see section 2F2. The probe was 5' - end-labelled and used as described in Methods.

The initial screening of the whole library with this probe identified 85 hybridising colonies, of which 23 produced only faint images on the autoradiograph. These were designated "possibles." All the positives, all the possibles, four negatives and cells containing pBR322 were rescreened: all but four of the positives hybridised again, and 13 of the possibles became positive. All the negatives, and the pBR322 - containing cells, remained negative. Thus, 71 clones were finally identified. These were named pRCLx, where RCL stands for Ricinus communis lectin, and x represents the number of the colony designated for this experiment.

Fig 3D-3 shows the autoradiograph of the rescreening experiment: the sequence of the probe gives a range of predicted melting points of 52 - 60 °C (see section 2F2). The actual sequence to which the probe hybridised was later found to give a melting point of 56°C: it would be expected that all the oligomer would have just washed off at this temperature, but it clearly has not. The rule is empirical and, evidently, not entirely accurate, given that the conditions were accurately maintained.

Comparison of these results with those of the differential screening reveals that 44 colonies were positive by both methods. The differential screening thus missed 27 colonies, while the oligomer missed 304: these may be non-lectin developmentally regulated sequences, or lectin-specific clones which lack the oligomer hybridising site. The 27 missed on the first screening could well be lectin clones containing the oligomer site, but not long enough to hybridise with the cDNA probe.



**Fig 3D-3: Rescreening with the oligomer.**  
Positive colonies and negative controls were replated in triplicate and rescreened, stringent washes being at the temperatures indicated.

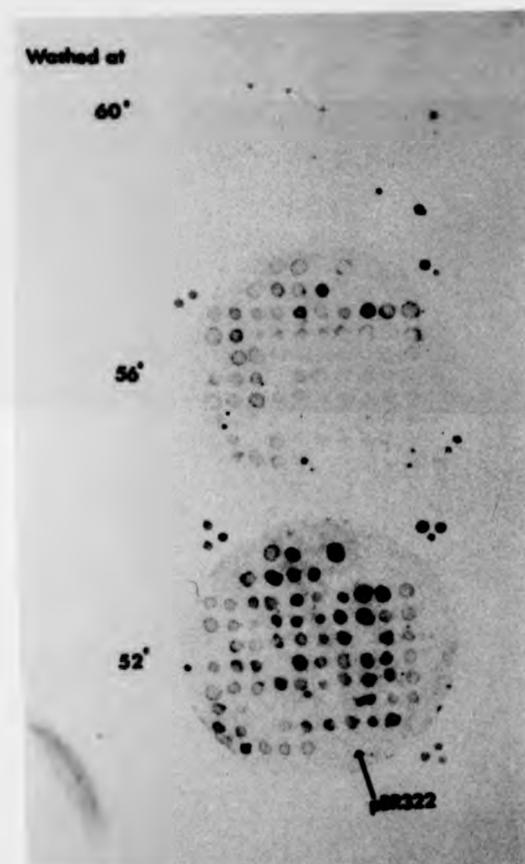


Fig 3D-3: Rescreening with the oligomer.  
Positive colonies and negative controls were replated in triplicate  
and rescreened, stringent washes being at the temperatures indicated.



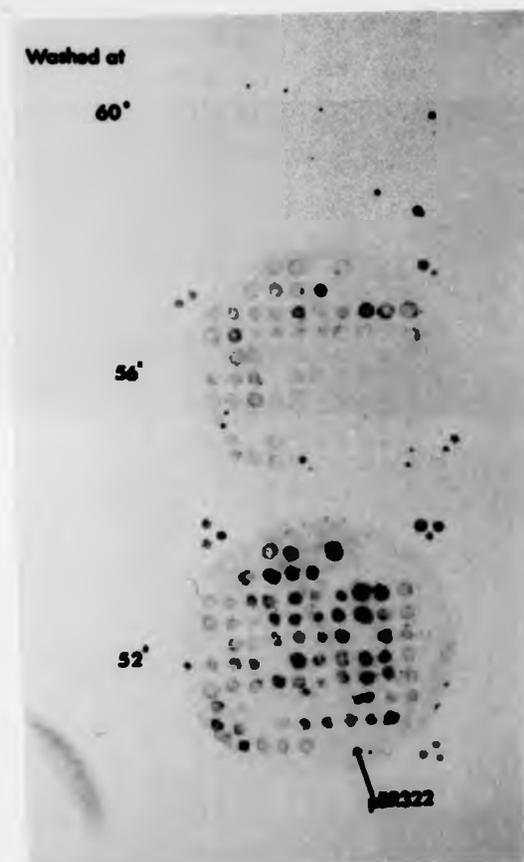


Fig 3D-3: Rescreening with the oligomer.  
Positive colonies and negative controls were replated in triplicate  
and rescreened, stringent washes being at the temperatures indicated.

Small-scale plasmid preparations were performed on all the colonies which hybridised with the oligomer, and their sizes were crudely estimated by comparison with those of supercoiled pBR322 and pAT153. The eight largest plasmids were then rescreened by translation of hybridisation-selected mRNA. The first step of this procedure involves linearisation of the plasmids with EcoRI. Fig 3D-4 shows the products of digestion: plasmids pRCL 6, 17, 58 and 61 do not contain EcoRI sites within the insert, whereas pRCL 15, 52, 57 and 59 do. This was the first indication that two classes of cDNA clones had been obtained, and these were later shown to correspond to ricin and agglutinin clones.

The products of translation of the hybridisation-selected mRNAs are shown in fig 3D-5, for plasmids pRCL 6, 15, 17, 58 and 59. The other three were analysed on a separate gel (not shown). All these plasmids select a mRNA encoding a protein of apparent  $M_r$  57 kDal which is immunoreactive with anti-RCA serum. That no other proteins were synthesised by the selected mRNAs is shown in the lane containing non-immunoprecipitated products of the mRNA selected by clone pRCL15.

Thus, 8 plasmids with fairly large inserts were shown to contain lectin - specific sequences.



Fig 3D-4: EcoRI digests of plasmids prior to hybridisation with RNA.  
Lanes 1 - 8 are plasmids pRCL 6,15,17,52,57,58,59 and 61, cut with  
EcoRI.

Lane 9 is pBR322 cut with EcoRI.

Lane 10 is undigested pBR322.

Samples analysed on 1 % neutral agarose gel in TAE buffer system.

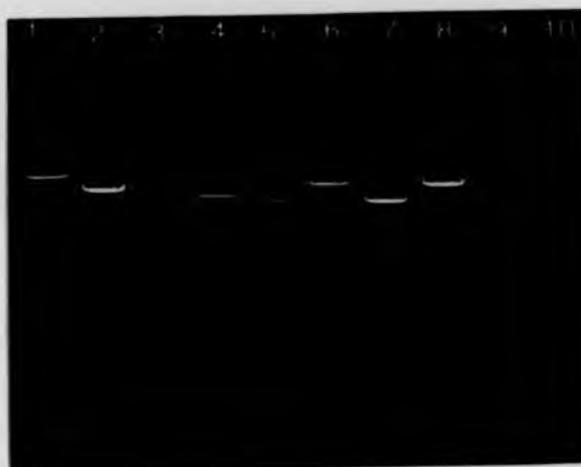


Fig 3D-4: EcoRI digests of plasmids prior to hybridisation with RNA.  
Lanes 1 - 8 are plasmids pRCL 6,15,17,52,57,58,59 and 61, cut with  
EcoRI.

Lane 9 is pBR322 cut with EcoRI.

Lane 10 is undigested pBR322.

Samples analysed on 1 % neutral agarose gel in TAE buffer system.

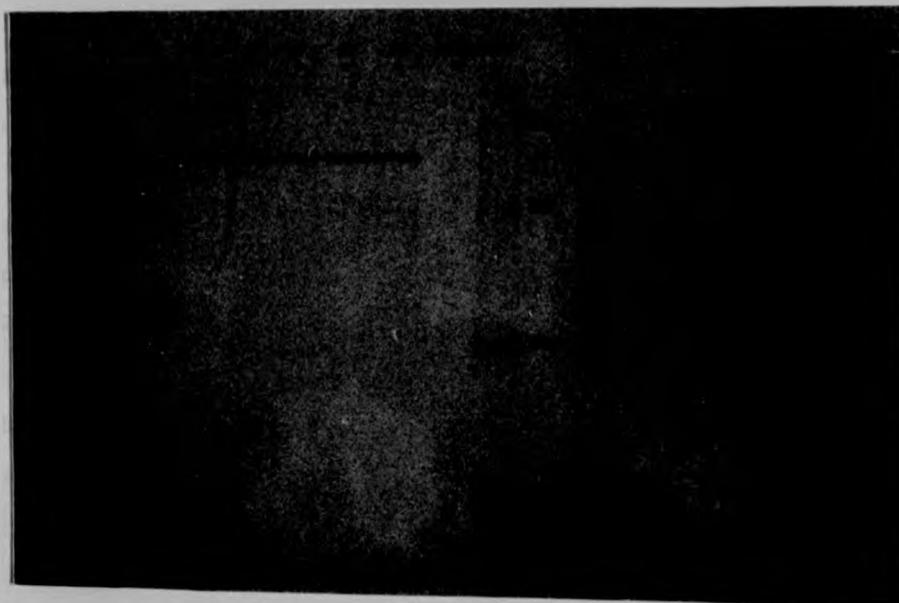
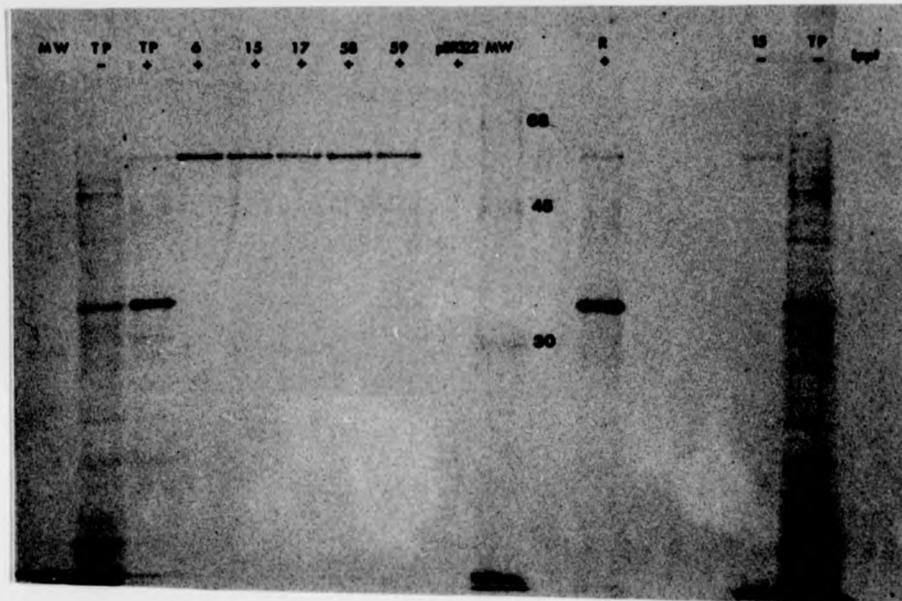


Fig 3D-5: Translation of hybridisation-selected mRNA

mRNAs were selected by plasmids pRCL 6, 15, 17, 58, 59 and pBR322 (indicated above lanes), translated in rabbit reticulocyte lysates and analysed by SDS/PAGE. Lanes marked + are immunoprecipitated with anti-agglutinin serum, while those marked - are run without immunoprecipitation.

Lanes TP show products of translation of mRNA prior to hybridisation with plasmids, and lane R shows products encoded by mRNA recovered from the hybridisation reaction.

Molecular weight markers are as fig 3A-3.



**Fig 3D-5: Translation of hybridisation-selected mRNA**

mRNAs were selected by plasmids pRCL 6, 15, 17, 58, 59 and pBR322 (indicated above lanes), translated in rabbit reticulocyte lysates and analysed by SDS/PAGE. Lanes marked + are immunoprecipitated with anti-agglutinin serum, while those marked - are run without immunoprecipitation.

Lanes TP show products of translation of mRNA prior to hybridisation with plasmids, and lane R shows products encoded by mRNA recovered from the hybridisation reaction.

Molecular weight markers are as fig 3A-3.

### 3E Characterisation of clones

The eight clones screened by translation of hybridisation-selected mRNA were cleaved with PstI and the sizes of the inserts were estimated on a 1 % neutral agarose gel, along with AluI and HinfI digests of pBR322 DNA. The gel is shown in fig 3E-1, along with the results, and the sizes of four of the clones, which were later determined by sequencing. (The relative absence of vector DNA is because the samples analysed were aliquots of purified inserts.)

Since the original intention was to sequence in M13, no detailed restriction mapping was undertaken. However, the presence of two restriction classes identified by EcoRI digestion (see fig 3D-4) led to further investigation of restriction site heterogeneity.

First, the eight clones were digested with BamHI. The largest, pRCL52, produced four bands (fig 3E-2), indicating that this insert has three sites (the fourth one is in the vector, pBR322). The bands were designated A - D, in decreasing order of size; the same fragments from the other clones are referred to by the same letters, even if fewer or different bands are produced.

DNA from the BamHI digests was transferred from the gel to nitrocellulose and probed with the oligodeoxynucleotide used for screening the library. Fig 3E-3 shows the autoradiograph, which shows that in all plasmids which produce BamHI fragment D, this band hybridises. It follows that fragment BamHI-D contains B chain sequences; its size is consistent with its containing most of the B chain: back-translation of the amino acid sequence implies possible sites at amino acids 5 - 7 (one in two possibility) and at amino acids 236 - 238 (one in eight possibility). This was later confirmed by sequencing.

Fig 3E-2 also shows that pRCL6,15 and 52 contain BamHI fragment C, which was later shown to contain the whole of the A chain sequence, and to precede fragment D in the clones. The BamHI and sequence



**Fig 3E-1: Insert sizes**

Plasmids containing lectin-specific inserts were digested with PstI and the inserts (purified as part of another experiment) were run on a 1 % neutral agarose gel in TAE buffer system.

Lanes 1 - 8 are pRCL 6,15,17,52,57,58,59 and 61 respectively.

Lane A is pBR322 cut with AluI and lane B is pBR322 cut with HinfI.

<u>Plasmid</u>	<u>Size from gel (bp)</u>	<u>Size from sequencing (bp)</u>
6	1650	1624
15	1650	
17	1350	1287
52	1800	1688
57	1200	1171
58	1000	
59	1300	
61	1050	

NB Sequencing sizes assume GC tails of 10 bp at each end;  
pRCL17 has a poly(A) tail of 27 residues.

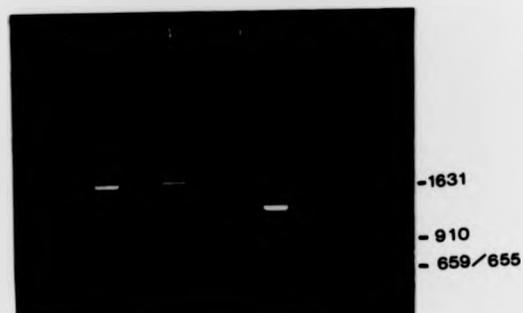


Fig 3E-1: Insert sizes

Plasmids containing lectin-specific inserts were digested with PstI and the inserts (purified as part of another experiment) were run on a 1 % neutral agarose gel in TAE buffere system.

Lanes 1 - 8 are pRCL 6,15,17,52,57,58,59 and 61 respectively.

Lane A is pBR322 cut with AluI and lane B is pBR322 cut with HinfI.

<u>Plasmid</u>	<u>Size from gel (bp)</u>	<u>Size from sequencing (bp)</u>
6	1650	1624
15	1650	
17	1350	1287
52	1800	1688
57	1200	1171
58	1000	
59	1300	
61	1050	

NB Sequencing sizes assume GC tails of 10 bp at each end;  
pRCL17 has a poly(A) tail of 27 residues.

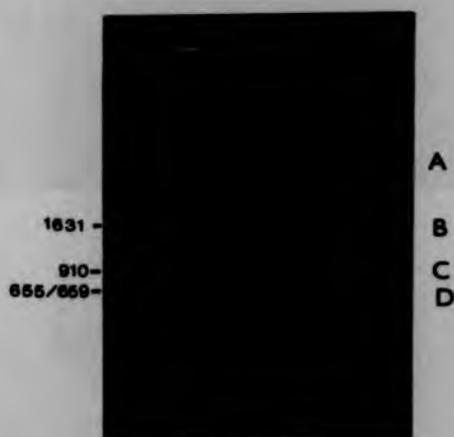


Fig 3E-2: BamHI digests

The eight plasmids which were screened by translation of hybridisation-selected mRNA were digested with BamHI and the products were run on a 1 % neutral agarose gel in TAE buffer system.

Lanes 1-8 are pRCL 6,15,17,52,57,58,59 and 61 respectively.

Lane A is pBR322 cut with HinfI.

Lane B is pBR322 cut with AluI.

Bands A - D are defined in the text (p108).



Fig 3E-2: BamHI digests

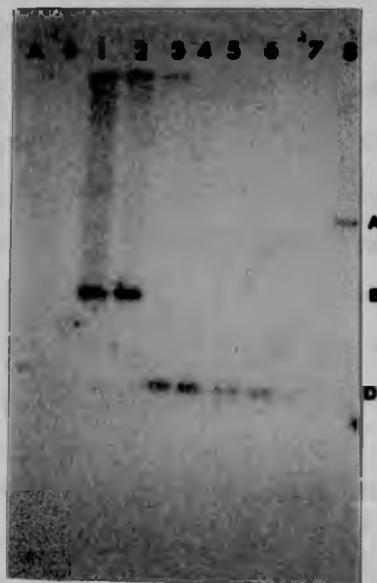
The eight plasmids which were screened by translation of hybridisation-selected mRNA were digested with BamHI and the products were run on a 1 % neutral agarose gel in TAE buffer system.

Lanes 1-8 are pRCL 6,15,17,52,57,58,59 and 61 respectively.

Lane A is pBR322 cut with HinfI.

Lane B is pBR322 cut with AluI.

Bands A - D are defined in the text (p108).



**Fig 3E-3: Southern blot of BamHI digests**

The gel shown in fig 3E-2 was blotted onto nitrocellulose and probed with the 20-mer oligodeoxynucleotide which hybridises with sequences encoding residues 214-220 of the ricin B chain.

Lanes 1-8 are pRCL 6,15,17,52,57,58,59 and 61 respectively.

Lanes A and B are pRR322 digested with HinfI and AluI respectively.

For definition of bands A,B and D see text (p108).

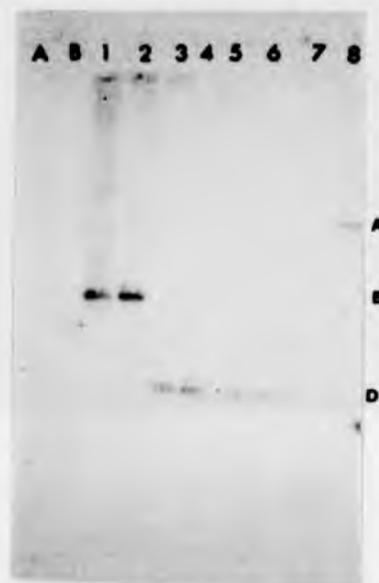


Fig 3E-3: Southern blot of BamHI digests

The gel shown in fig 3E-2 was blotted onto nitrocellulose and probed with the 20-mer oligodeoxynucleotide which hybridises with sequences encoding residues 214-220 of the ricin B chain.

Lanes 1-8 are pRCL 6,15,17,52,57,58,59 and 61 respectively.

Lanes A and B are pBR322 digested with HinfI and AluI respectively.

For definition of bands A,B and D see text (p108).

data, along with Sau96I digests of insert preparations, were later used to determine the extents of all eight clones: see fig 3E-4. The downstream Sau96I site is present in all the clones, as is the central BamHI site. The ends of the clones were estimated from the sizes of the Sau96I fragments, along with knowledge of the presence or absence of the downstream and upstream BamHI sites.

Sequencing of clones pRCL6,17,52 and 57 implied the existence of another heterogeneity: the EcoRI<sup>-</sup> clones contained a BglII site in BamHI fragment D, which was absent in the EcoRI<sup>+</sup> clones. In order to investigate this in the other clones, all were cut with BamHI and PstI: those clones lacking a 'natural' BamHI-D fragment thus gain an analogous BamHI-PstI fragment. Half of each digest was then cleaved with BglII, and all fragments in all the reactions were end-labelled, and analysed on agarose gels. The autoradiographs are shown in fig 3E-5: in each EcoRI<sup>+</sup> clone the BamHI fragment D lacks a BglII site, while in each EcoRI<sup>-</sup> clone, the site is present.

Heterogeneity with respect to TaqI sites exists within BamHI fragments C and D. In fragment C, clone pRCL6 contains two TaqI sites, and clone pRCL52 contains one (which is at a different position from both of those in pRCL6). The only other clone having a BamHI fragment C is pRCL15, which is in the same EcoRI class as pRCL52. pRCL15 was shown to contain only the TaqI site present in pRCL52. TaqI site heterogeneity in BamHI fragment D was not further investigated.

Thus, the eight clones can be divided into two restriction classes in terms of EcoRI, BglII and TaqI sites. No variation with respect to BamHI or Sau96I was detected, and no clone had any sites for HindIII, SmaI, AvaI or HindII (data not shown).

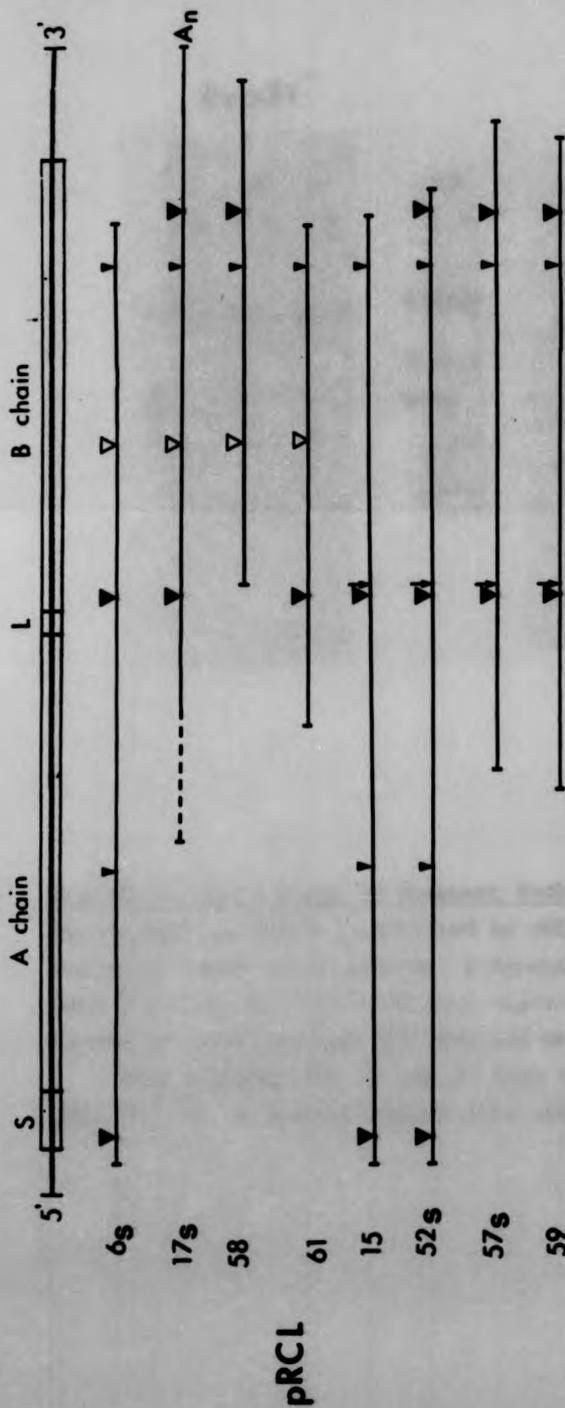


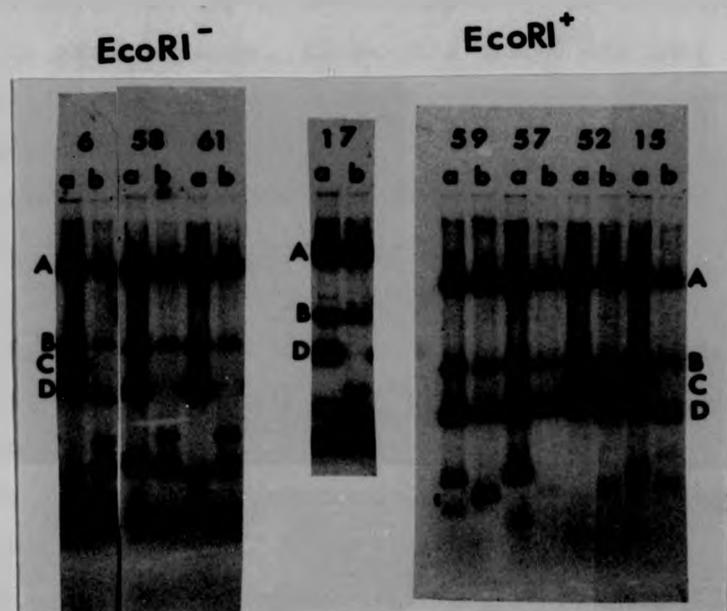
Fig 3E-4: Extents of lectin clones

Extents of clones were mapped using Sau96I sites (▼) and BamHI sites (▽), and by sequencing (clones subscripted 's').

Also shown are the heterogeneous EcoRI sites (↓) and BglII sites (▽) (the BglII site in the A chains is not shown.)

Dotted line in pRCL17 indicates the B chain fragment within this clone - see section 3G1.

An indicates the presence of a poly(A) tail.



**Fig 3E-5: BglII sites in fragment BamHI-D**

Recombinant plasmids (identified by pRCL no) were cleaved with BamHI and PstI (lanes a) to generate fragments analogous to BamHI-D (see text for details). Half of each reaction was then digested with BglII (lanes b), end-labelled and analysed on a 1 % neutral agarose gel.

Clones pRCL 6, 58, 61 and 17 lack an EcoRI site, while clones pRCL 59, 57, 52 and 15 contain this site.

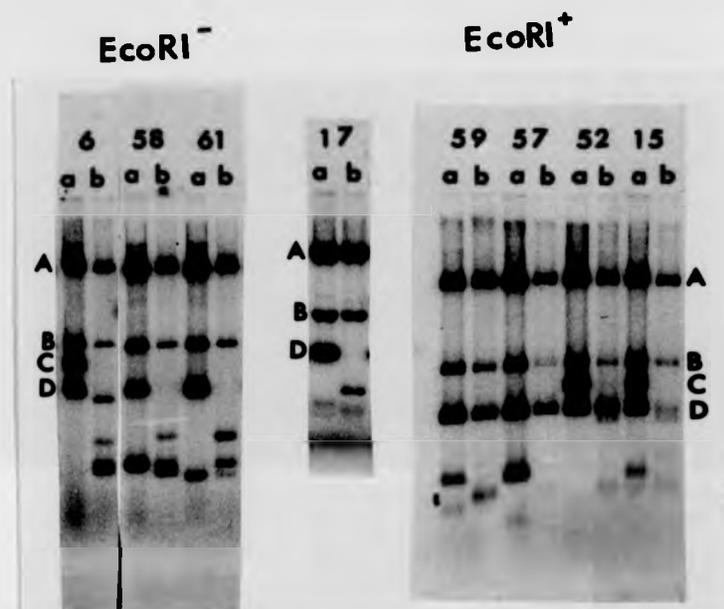


Fig 3E-5: BglII sites in fragment BamHI-D

Recombinant plasmids (identified by pRCL no) were cleaved with BamHI and PstI (lanes a) to generate fragments analogous to BamHI-D (see text for details). Half of each reaction was then digested with BglII (lanes b), end-labelled and analysed on a 1% neutral agarose gel.

Clones pRCL6, 58, 61 and 17 lack an EcoRI site, while clones pRCL 59, 57, 52 and 15 contain this site.

### 3F Subcloning into pUC8

The inserts of clones pRCL6 and pRCL52 were subcloned into pUC8, in order to facilitate the sequencing of their ends by Maxam and Gilbert sequencing, and for the construction of clones containing complete coding sequences. Clones pRCL17 and 57 were later shown to overlap with clones pRCL6 and 52 respectively, and were then also subcloned into pUC8.

The inserts were excised with PstI and gel purified, and ligated into suitably prepared vector, preparations of which were first shown to give very low backgrounds of transformation. Some 30 - 50 colonies were obtained from the ligated material, compared to 1 - 2 from the vector. Small-scale plasmid extractions were performed: fig 3F-1 shows eight such plasmids from the cloning of the pRCL52 insert. All eight are of the expected size. The identities of the inserts were confirmed by limited restriction mapping, with BamHI, BglIII and EcoRI, the orientations of the inserts being determined from the BamHI digests. Fig 3F-2 shows the BamHI digests of recombinants in pUC8 containing the insert of pRCL6: Clones pUC611, 616 and 618 are in orientation A (see fig 3F-3), while pUC613, 614, 615 and 617 are in orientation B. Plasmid pUC612 failed to cut, as it comigrates with uncut pUC613 - this was not investigated further.

The orientations of the inserts in the pRCL52 subclones were determined by EcoRI digestion: pUC8 has a single EcoRI site, near the BamHI site shown in fig 3F-3, while the insert has a single site cutting it into fragments of approximately 1050 bp and 630 bp. Clones pUC521, 522, 523, 524 and 527 were in orientation A, while pUC525, 526 and 528 were in orientation B.

### 3F Subcloning into pUC8

The inserts of clones pRCL6 and pRCL52 were subcloned into pUC8, in order to facilitate the sequencing of their ends by Maxam and Gilbert sequencing, and for the construction of clones containing complete coding sequences. Clones pRCL17 and 57 were later shown to overlap with clones pRCL6 and 52 respectively, and were then also subcloned into pUC8.

The inserts were excised with PstI and gel purified, and ligated into suitably prepared vector, preparations of which were first shown to give very low backgrounds of transformation. Some 30 - 50 colonies were obtained from the ligated material, compared to 1 - 2 from the vector. Small-scale plasmid extractions were performed: fig 3F-1 shows eight such plasmids from the cloning of the pRCL52 insert. All eight are of the expected size. The identities of the inserts were confirmed by limited restriction mapping, with BamHI, BglII and EcoRI, the orientations of the inserts being determined from the BamHI digests. Fig 3F-2 shows the BamHI digests of recombinants in pUC8 containing the insert of pRCL6: Clones pUC611, 616 and 618 are in orientation A (see fig 3F-3), while pUC613, 614, 615 and 617 are in orientation B. Plasmid pUC612 failed to cut, as it comigrates with uncut pUC613 - this was not investigated further.

The orientations of the inserts in the pRCL52 subclones were determined by EcoRI digestion: pUC8 has a single EcoRI site, near the BamHI site shown in fig 3F-3, while the insert has a single site cutting it into fragments of approximately 1050 bp and 630 bp. Clones pUC521, 522, 523, 524 and 527 were in orientation A, while pUC525, 526 and 528 were in orientation B.

A B C 1 2 3 4 5 6 7 8 C B A

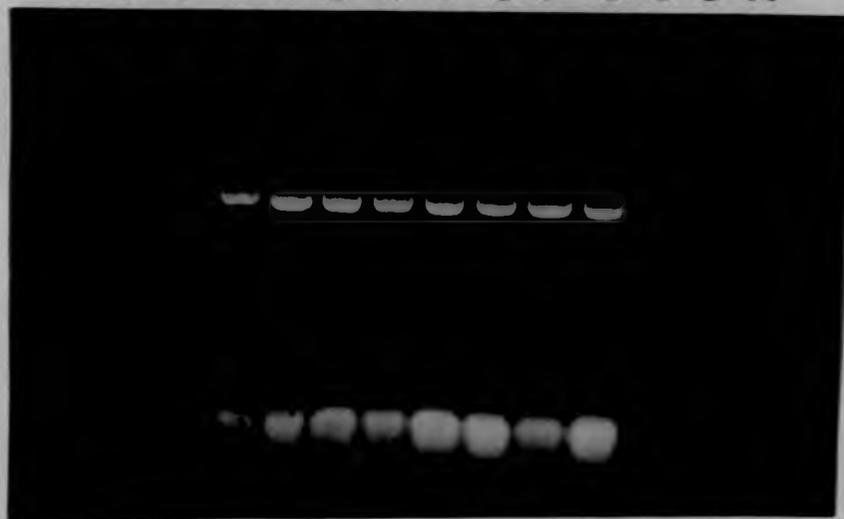


Fig 3F-1: Subcloning into pUC8.

The insert of pRCL52 was cloned into pUC8. Plasmids were extracted from eight colonies and analysed on a 1 % neutral agarose gel.

Lanes 1 - 8 are plasmids pUC521 - 528.

Lane A is pUC8, B is pBR322 and C is pAT153.

Samples analysed on 0.8 % agarose gel in TAE buffer system.



Fig 3F-1: Subcloning into pUC8.

The insert of pRCL52 was cloned into pUC8. Plasmids were extracted from eight colonies and analysed on a 1 % neutral agarose gel.

Lanes 1 - 8 are plasmids pUC521 - 528.

Lane A is pUC8, B is pBR322 and C is pAT153.

Samples analysed on 0.8 % agarose gel in TAE buffer system.

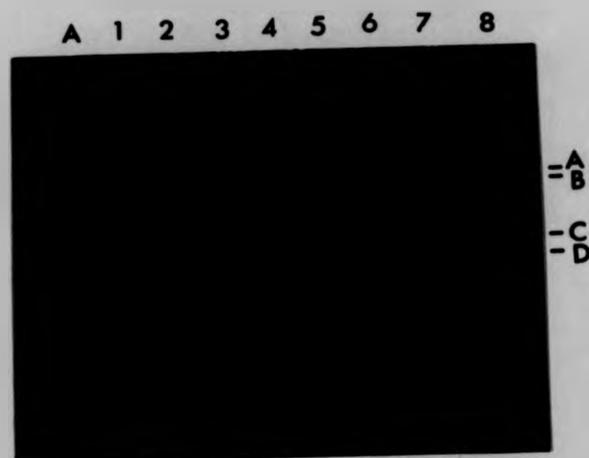


Fig 3F-2: Subcloning in pUC8.

The insert of pRCL6 was cloned in pUC8. Eight of the clones were digested with BamHI and run on a 1 % neutral agarose gel.

Lanes 1 - 8 are clones pUC611 - 618 cut with BamHI.

Lane A is undigested pUC613.

Fragments A - D are explained in fig 3F-3.

Samples were analysed on a 1 % agarose gel in TAE buffer system.

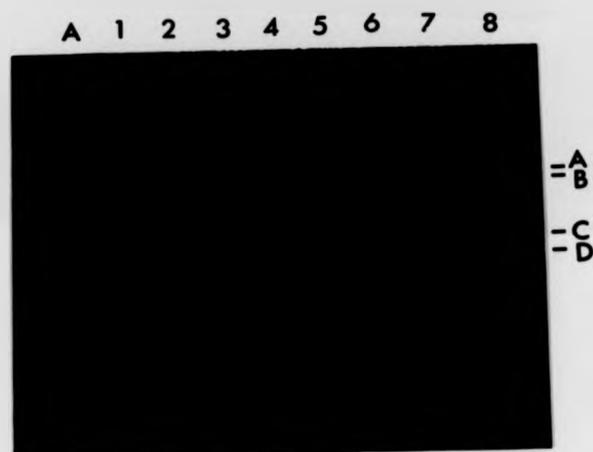


Fig 3F-2: Subcloning in pUC8.

The insert of pRCL6 was cloned in pUC8. Eight of the clones were digested with BamHI and run on a 1 % neutral agarose gel.

Lanes 1 - 8 are clones pUC611 - 618 cut with BamHI.

Lane A is undigested pUC613.

Fragments A - D are explained in fig 3F-3.

Samplese were analysed on a 1 % agarose gel in TAE buffer system.

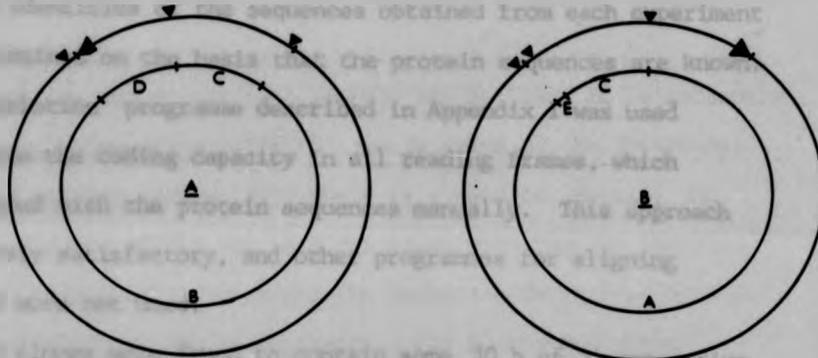


Fig 3F-3: Orientations of pUC8 subclones of pRCL6 insert.

Above: Location of BamHI sites in pRCL insert, numbered from beginning of insert, and assuming G-tails of 10 bases. Arrowhead indicates 5' to 3' direction.

Below: Orientations in pUC8 with respect to the BamHI site of the vector. Boundaries of inserts are indicated, and correspond to PstI sites. Fragments A,B,C and D are the bands visible in fig 3F-2; band E is too small to see on the gel.

### 3G Sequencing: introduction

#### 3G1 Summary of results

Clones pRCL6 and pRCL52 were initially selected for sequencing - that is, the largest member of each of the two restriction classes. Sequencing was at first carried out by the dideoxy method, after subcloning restriction digests of the inserts into M13 mp8; the remaining gaps were filled by Maxam and Gilbert sequencing.

The identities of the sequences obtained from each experiment were determined on the basis that the protein sequences are known: the 'translation' programme described in Appendix 1 was used to generate the coding capacity in all reading frames, which were aligned with the protein sequences manually. This approach was entirely satisfactory, and other programmes for aligning sequences were not used.

Both clones were found to contain some 30 b of 5' non-coding region, followed by a putative signal sequence of 24 amino acids. This was followed by sequences corresponding to the mature A chain, which was in turn followed by B chain sequences. A 12-residue linker was found between the two chains. Neither clone ran as far as the C-terminus of the B chain, clone pRCL52 being somewhat longer than pRCL6.

It was suspected, from restriction data (see section 3E) that clones pRCL17 and 57 overlap with clones pRCL6 and 57 respectively, in such a way that the additional clones should contain sequences downstream of those in the clones originally sequenced. These two were therefore sequenced, entirely by the Maxam & Gilbert method: clone pRCL17 contained the whole 3' - non-coding region, including the poly(A) tail, whereas

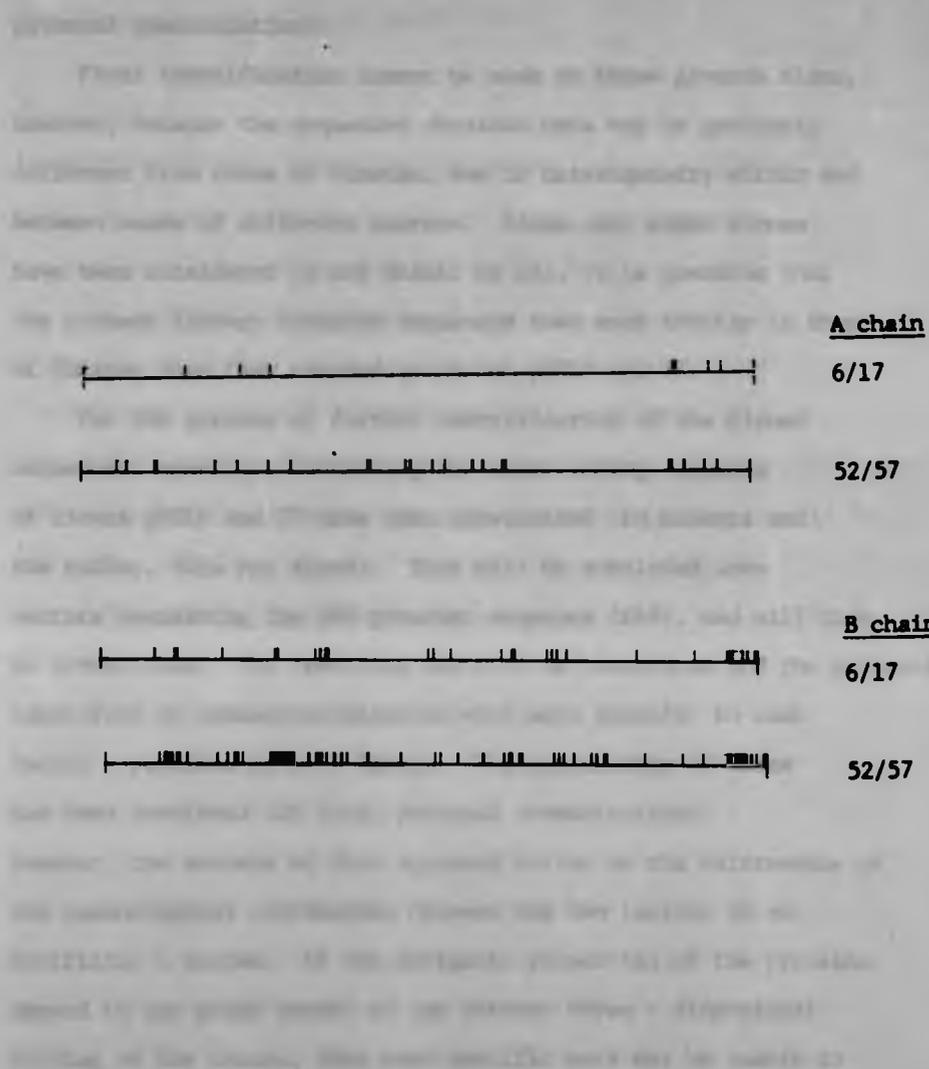
pRCL57 terminated within the 3' - non-coding region. Throughout the overlapping parts, no differences existed within the two pairs of clones.

However, pRCL17 contains an extra fragment: the greater part of its insert consists of sequences encoding the end of the A chain, the linker, the whole B chain and the 3' - non-coding region. Preceding this, and in the other strand, there is a small fragment encoding part of the B chain. As will be discussed at the end of this section, this is probably a cloning artefact.

Since the two castor bean lectins are believed to be extremely similar, both immunologically (3) and in terms of the N-terminal sequences of both chains (31), there is a need to identify the coding sequences of the two sets of clones with the two lectins. This was approached by comparing the sequences determined here with those published for the ricin chains. Fig 3G-1 shows plots indicating the locations of differences between the sequences deduced from the cDNA clones and the published sequences. It is clearly apparent that the sequence of clones pRCL6 and 17 is far more similar to the published amino acid sequences than is that encoded by clones pRCL52 and 57. The similarity of the clone pRCL6 and 17 sequence to the published one may well be much greater than indicated here, as will be shown later.

The two sequences were tentatively identified as ricin (clones pRCL6 and 17) and RCA (clones pRCL52 and 57) on this basis.

It is of note that the cDNA data indicate that the two B chains are of very similar length - although EndoH digestion of ricin and agglutinin B chains indicates that the latter is longer (45), more recent evidence obtained in the presence of tunicamycin indicates that they are indeed of similar length (JM Lord,



**Fig 3G-1**

Differences between the deduced amino acid sequences and those determined by Funatsu's group. Each vertical line represents an amino acid difference, deletion or insertion.

personal communication).

Final identification cannot be made on these grounds alone, however, because the sequences obtained here may be genuinely different from those of Funatsu, due to heterogeneity within and between beans of different sources. Since only eight clones have been considered in any detail at all, it is possible that the present library contains sequences even more similar to those of Funatsu than that encoded by clones pRCL6 and 17.

For the purpose of further identification of the cloned sequences, subclones containing the whole coding sequence of clones pRCL6 and 17 have been constructed (LM Roberts and the author, data not shown). This will be subcloned into vectors containing the SP6 promoter sequence (246), and will then be transcribed. The resulting RNA will be translated and the products identified by immunoprecipitation with sera specific to each lectin - provided by Dr P Thorpe. The specificity of these has been confirmed (JM Lord, personal communication).

However, the success of this approach relies on the maintenance of the immunological distinction between the two lectins in so artificial a system. If the antigenic properties of the proteins depend to any great extent on the correct three - dimensional folding of the chains, then even specific sera may be unable to distinguish between them.

Another approach relies on the differences in the functions of the subunits of the two lectins. If the cDNA sequences can be expressed in such a way as to produce toxic A chains or sugar-binding B chains, then quantitative assay of A chain function, or determination of sugar specificity of B chains should assist in final identification. Towards this end, the clones containing the whole ricin coding sequence will be subcloned into,

the yeast expression vector pMA91 (247).

Clone pRCL17

As mentioned above, this clone appears to contain a cloning artefact. A map is shown in fig 3G-2, and the sequence surrounding the junction is shown in fig 3G-3.

The origin of this anomaly is unclear. The three G residues (encoding the first glycine residue of the additional fragment) might suggest an origin during annealing: this hypothesis requires that the end of one molecule failed to tail, such that these G residues were left overhanging, subsequently to hybridise with the C tails on another molecule. However, this model implies that the G residues would be at the 3' end of a DNA strand; they are, in fact, at the 5' end.

The E coli strain used was DH1, which is recA<sup>-</sup> (202). However, a scenario for a recombination event can nonetheless be constructed, based on the findings of Conley & Saunders (248), who report the transformation of E coli with linearised pBR322 DNA into a variety of recombinational backgrounds, including recA<sup>-</sup>. They find that, although the efficiency of transformation is low with linear molecules, the transformants which do arise generally contain deletions, implying a recombination event. It is suggested that most transformants in these experiments arose through recombination of homologous sequences, rather than by annealing of sticky restriction ends. In support of this, they transformed cells with linear dimers of pBR322, and obtained perfect monomeric plasmid DNA from transformants.

It is possible that some similar event has occurred in clone pRCL17: if a single cell were transformed with two

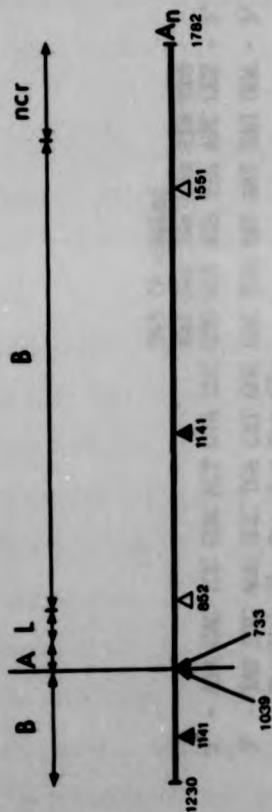


Fig 3G-2: Map of clone pRCLL17.

Upper line: Coding regions (A=A chain; B=B chain; L=linker; ncr=3' non-coding region).

Lower line: Numbers correspond to nucleotides in the preproinsulin cDNA sequence.

Restriction sites: Δ = BamHI; ▲ = BglII.

The upper line (5' to 3') gives the determined sequence.

Below this is the complementary strand.

Numbers indicate positions in the preproinsulin sequence.

245 (A chain)  
Val Ser Ile Leu Ile Pro  
5' - ATA GAC TCC CCG ACT GTA CCC GTG AGT ATA TTA ATC CCT - 3'  
3' - TAT CTG AGG GCC TGA CAT GGG CAC TCA TAT AAT TAG CGA - 5'  
Tyr Val Gly Pro Ser Tyr Gly  
347 (B chain)

Fig 3G-3: Sequence of the 'junction' in pRCL17.  
The upper line (5' to 3') gives the determined sequence.  
Below this is the complementary strand.  
Numbers indicate positions in the preprorin sequence.

recombinant molecules, each containing ricin sequences, then a recombination - deletion event may have occurred, parts of one or both of the original molecules being deleted in the process. However, this should imply some homology surrounding the recombination site, but a search of this region, comparing regions upstream and downstream of both termini at the junction, revealed no detectable homology. The search was conducted by the generation of a dot matrix (249) using a simple computer programme (results not shown).

#### 3G2 M13 cloning and dideoxy sequencing

The inserts of clones pRCL6 and 52 were initially sequenced in M13 subclones, generated from Sau3AI and AluI digests of their inserts. In each case, libraries of 300 - 400 M13 recombinants were obtained. Sequencing of many of these, along with screening of the rest by T-tracking, showed that not all of the target sequence had been represented in the libraries. Fig 3G-4 shows the extent of the Sau3AI and AluI libraries for both clones, as well as a HaeIII - AluI double digest library subsequently made from pRCL6. The sites for these enzymes are also shown, as determined from the completed sequences.

In each case, there are gaps - these are in similar regions of each clone. Although the large region lacking AluI sites might be understood not to have cloned very efficiently in the presence of smaller fragments, there are several Sau3AI sites in this region, which would be expected to have cloned. Non-random distribution of M13 clones has previously been reported (235), and it may be relevant that Drouin (250) found that certain MboI fragments were difficult to clone in pBR322.

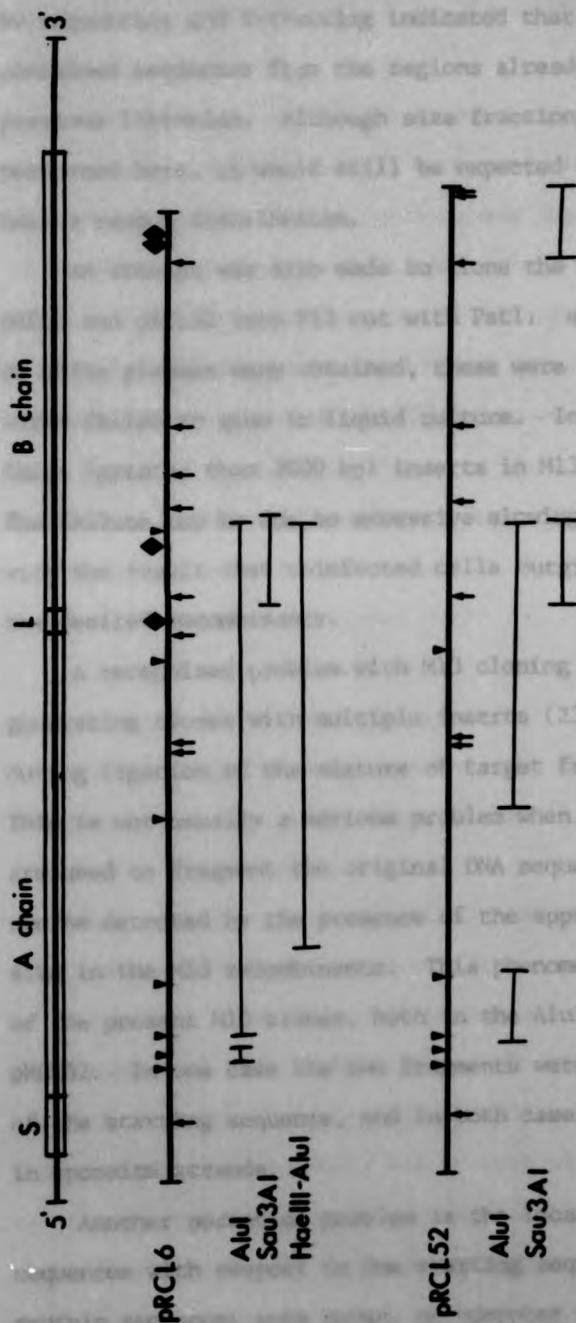


Fig 3C-4: Extents of M13 libraries from pRCL6 and pRCL52. Libraries were made using the enzymes indicated at left; the sites for these enzymes are shown in the clones: ▽ = AluI; ↑ = Sau3AI; ◆ = HaeIII.

A further clone bank of about 100 clones was constructed by a DNase I cleavage method (234). Limited analysis of this bank by sequencing and T-tracking indicated that most of its members contained sequences from the regions already represented in the previous libraries. Although size fractionation was not performed here, it would still be expected that the clones should have a random distribution.

An attempt was also made to clone the whole inserts of pRCL6 and pRCL52 into M13 cut with PstI: although large numbers of white plaques were obtained, these were very small, and the virus failed to grow in liquid culture. Instability of large (greater than 2000 bp) inserts in M13 has been reported (235). The failure may be due to excessive slowing of virus replication, with the result that uninfected cells outgrow those containing the desired recombinants.

A recognised problem with M13 cloning is the risk of generating clones with multiple inserts (235), this occurring during ligation of the mixture of target fragments with the vector. This is not usually a serious problem when restriction digests are used to fragment the original DNA sequence, as its occurrence can be detected by the presence of the appropriate restriction site in the M13 recombinants. This phenomenon was found in two of the present M13 clones, both in the AluI library derived from pRCL52. In one case the two fragments were from adjacent regions of the starting sequence, and in both cases the fragments were in opposite strands.

Another potential problem is the location of the determined sequences with respect to the starting sequence: since the protein sequences were known, no computer programmes for aligning

these sequences were used. No problems of this type were found during this work, but it is suspected that if pRCL17 had been sequenced in M13, some confusion would have arisen from the duplication within this clone.

Some of the difficulties associated with these techniques were not encountered here, such as band compression (251). However, a fairly frequent problem was that of non-specific termination of synthesis of the new DNA strand. Regions of sequence which were difficult to determine because of this were obtained by sequencing M13 clones containing the same fragment in the opposite orientation, and by Maxam & Gilbert sequencing.

Because of the incompleteness of the M13 results, the remaining sequence was determined by the Maxam & Gilbert chemical method, as described below.

### 3G3 Maxam & Gilbert sequencing

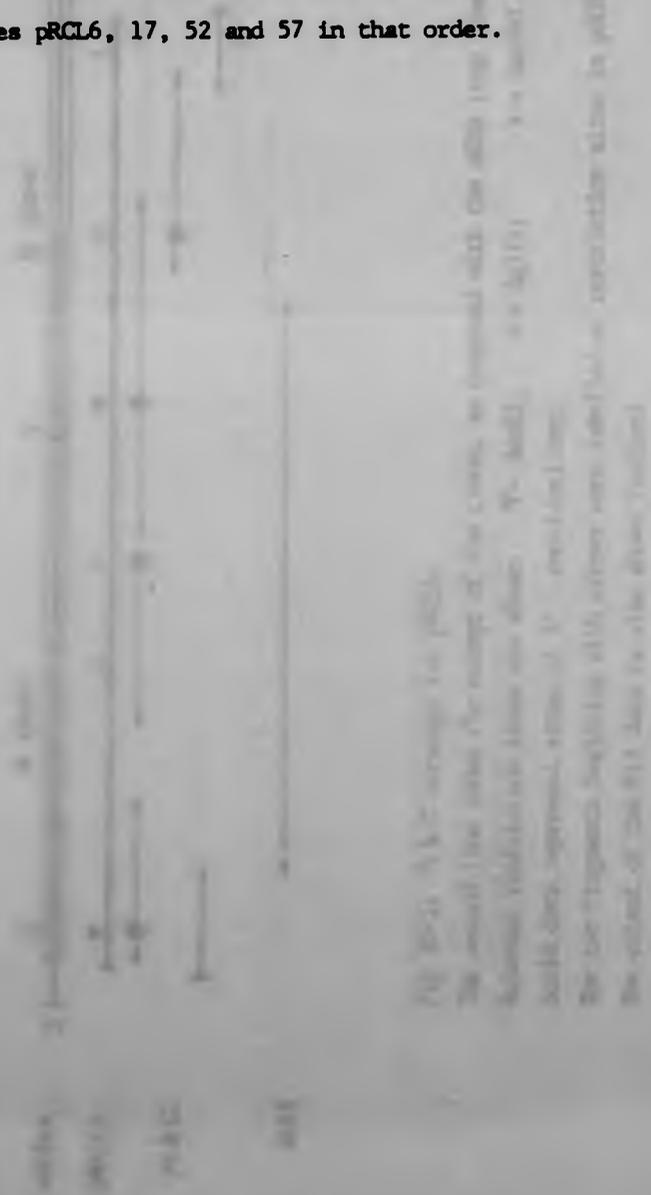
Very few problems were encountered with this technique.

Occasionally, gels would smear, or bands would appear diffuse, these problems probably arising from poor fragment preparation. In one or two cases, non-specific bands were observed, but these did not reappear on repetition of the experiment.

The most important difficulties occurred when sequencing from *Ava*I sites and from *Hind*III sites. In the former case, the site in pUC8 which was labelled has two adjacent C residues. In the simple labelling protocol, one or both of these may be labelled - giving overlapping sequence ladders. This was easily overcome by following the labelling reaction with a chase of dCTP at much higher concentration than that of the labelling reaction. *Hind*III sites mostly produced unreadable ladders, with badly

overlapping sequences, though some such experiments worked well. Even if this enzyme can cut at either of the A residues in its recognition sequence (A/AGCTT), this should have no effect, as the reaction was performed in the presence of 100  $\mu$ M cold dATP, and radioactive dGTP - only one position should label. The reason for this problem thus remains unknown.

The fragments sequenced are shown in figs 3G5-3G8, for clones pRCL6, 17, 52 and 57 in that order.



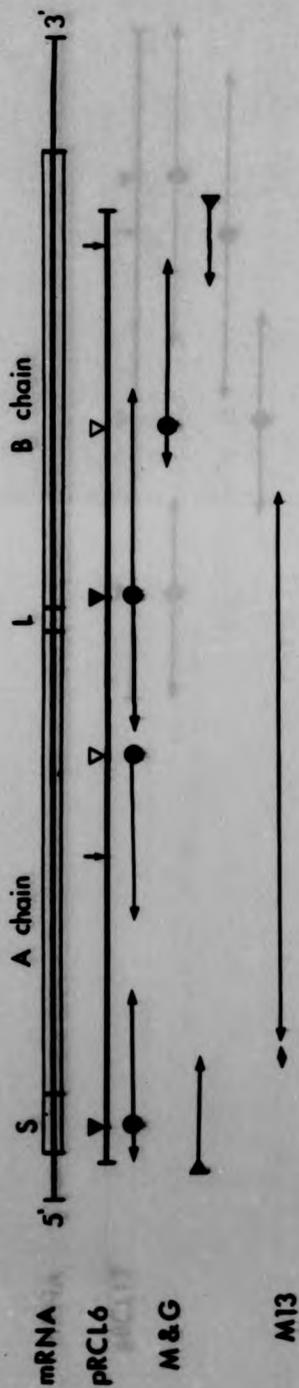


Fig 3G-5: M & G strategy for pRCL6.

The second line shows the extent of the clone, as compared with the mRNA (top line).

Relevant restriction sites are shown: ▽ = BamHI; ▽ = BglII; ↓ = Sau96I.

Solid dots represent sites of 3' - end-labelling.

The two fragments beginning with arrows were labelled at restriction sites in pUC8 subclones.

The extent of the M13 data is also shown (bottom).

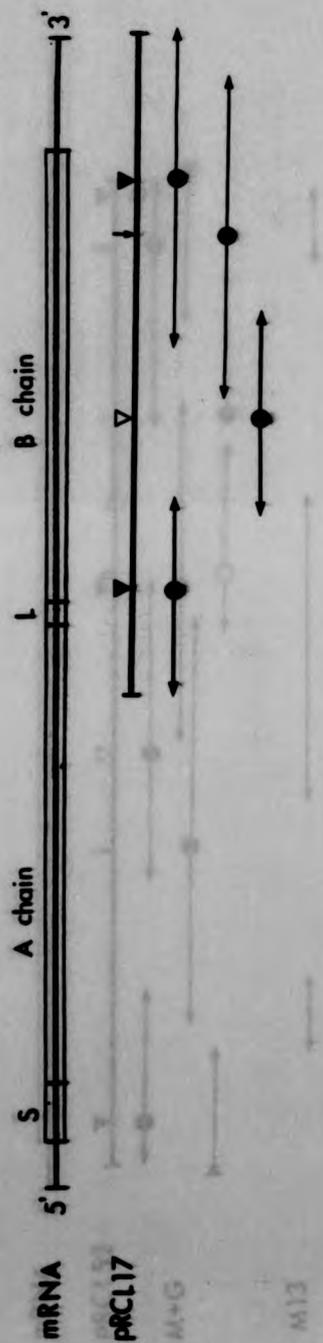


Fig 3G-6: M & G strategy for pRCL17.  
Only the main coding part of the clone is shown.  
For explanation of symbols see fig 3G-5.

For explanation of symbols see fig 3G-5.  
Additional symbols: Tail sites are indicated by ⊕, and EcoRI sites by ▽.  
Fragments from open circles were labelled at the 5' end.

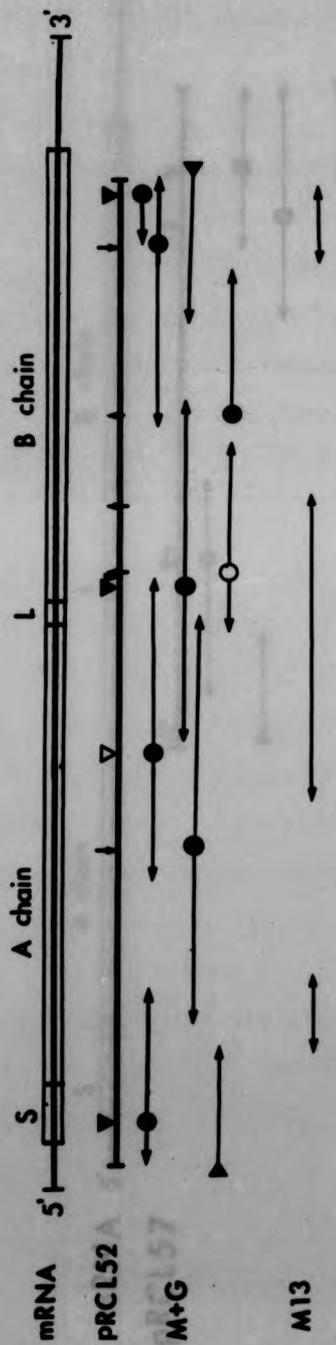


Fig 3G-7: H & G strategy for pRCL52.  
For explanation of symbols see fig 3G-5.  
Additional symbols: TaqI sites are indicated by ▼, and EcoRI sites by ●.  
Fragments from open circles were labelled at the 5' end.

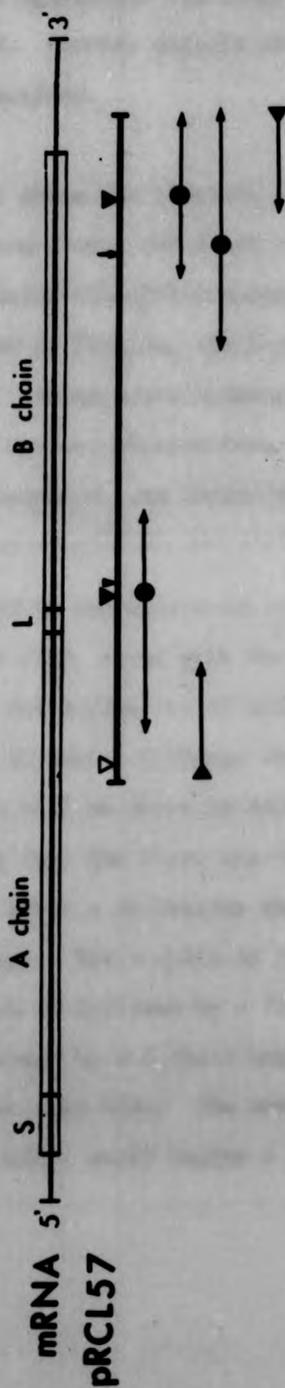


Fig 3G-8: M & G strategy for pRCL57.  
For explanation of symbols see fig 3G-7.

### 3H The preproricin sequence

In this and the following sections, the cDNA sequences encoding ricin and the agglutinin (in their prepro- forms) are presented and described. Further details will be discussed in subsequent sections.

Fig 3H-1 shows the complete preproricin cDNA sequence, as determined from clones pRCL6 and 17. The deduced amino acid sequence is below the cDNA sequence, and when this differs from those obtained by Funatsu, the latter's data are placed below the deduced data. Amino acids missing from Funatsu's data are indicated by dashes, whereas those present in Funatsu, but absent in the deduced sequence, are indicated by asterisks. Potential sites for N-glycosylation are enclosed in boxes.

Clone pRCL6 contains bases -102 to 1512, and pRCL17 contains bases 733 to 1782, along with the poly(A) tail (but see also section 3G1 for discussion of pRCL17). The 5' non-coding region consists of 30 bases, although the mRNA probably contains another 80 bases, as will be shown in section 3L. This is followed by an ATG codon (not the first one) which opens up a reading frame containing, after a 24-residue signal sequence, the whole ricin A and B chains. The A chain is 267 residues long (Funatu's sequence was 265), and is followed by a linker of 12 amino acids. This is in turn followed by a B chain sequence of 262 residues (Funatsu's sequence contained 260). The coding sequence is terminated by a TGA stop codon, which begins a 3' non-coding region of 156 bases.

3'-AAAGGKXKAG GAATACATAT TGTATATGG ATG TAT CCA GTG CCA ACA TGG CTT TGT TTT CCA TCC ACC TCA GGG TGG TCT TTC ACA TTA CAG  
Met Tyr Ala Val Ala Thr Trp Leu Cys Phe Gly Ser Thr Ser Gly Trp Ser Phe Thr Leu Glu

GAT AAC AAC ATA TTC CCG AAA CAA TAC CCA ATT ATA AAC TTT ACC ACA GGG GGT GGC ACT GTG CAA ACC TAC ACA AAC TTT ATC AGA GCT  
Asp Asn Asn Ile Phe Pro Lys Lys Tyr Pro Ile Ile Asn Phe Thr Thr Ala Gly Ala Thr Val Gln Ser Tyr Thr Asn Phe Ile Arg Ala

GTT GCG GGT GGT TTA ACA ACT GGA GCT GAT GTG AGA CAT GAT ATA CCA GTG TTG CCA AAC ACA GTT GGT TGG GCT ATA AAC CAA GGG TTT  
Val Arg Gly Arg Leu Thr Thr Gly Ala Asp Val Arg His Asp Ile Pro Val Leu Pro Asn Arg Val Gly Leu Pro Ile Asn Gln Arg Phe

ATT TTA GTT GAA CTC TCA AAT CAT CCA CAG CTT TCT GGT ACA TTA CCA CCG GAT GTC ACC AAT CCA TAT GGG GTC GGC TAC GGT GCT GGA  
Ile Leu Val Glu Leu Ser Asn His Ala Glu Leu Ser Val Thr Leu Ala Leu Asp Val Thr Asn Ala Tyr Val Val Gly Tyr Arg Ala Gly

AAT ACC CCA TAT TTC TTT CAT CCG GAC AAT CAG GAA GAT CCA GAA CCA ATC ACT CAT GTT TTC ACT GAT GTT CAA AAT CCA TAT ACA TTC  
Asn Ser Ala Tyr Phe Phe His Pro Asp Asn Gln Glu Asp Ala Glu Ala Ile Thr His Leu Phe Thr Asp Val Gln Asn Arg Tyr Thr Phe

GCC TTT GGT GGT AAT TAT GAT AGA CTT GAA CAA CTT GCT GGT AAT CCG AGA GAA AAT ATC GAG TTG GGA AAT GGT CCA CTA CAG GAG GCT  
Gly Phe Gly Asn Tyr Asp Arg Leu Glu Phe Leu Ala Gly Asn Leu Arg Glu Asn Ile Glu Leu Gly Asn Gly Pro Leu Glu Glu Ala

ATC TCA GCG CTT TAT TAT TAC AGT ACT GGT GGC ACT CAG CTT CCA ACT CCG GCT GGT TCC TTT ATA ATT TCC ATC CAA ATC ATT TCA GAA  
Ser Leu Val Glu Leu Tyr Tyr Tyr Ser Ile Gly Gly Thr Gln Leu Pro Thr Leu Ala Arg Ser Phe Ile Ile Cys Ile Gln Met Ile Ser Glu

CCA CCA ACA TTC CAA TAT ATT GAG CCA GAA ATG GCG ACC AGA ATT ACC TAC AAC CCG ACA TCT CCA CCA GAT GCT ACC GTA ATT ACA CTT  
Ala Ala Arg Phe Gln Tyr Ile Glu Gly Glu Met Arg Thr Arg Ile Arg Tyr Asn Arg Arg Ser Ala Pro Asp Pro Ser Val Ile Thr Leu

GAG AAT AGT TGG GGG ACA CTT TCC ACT CCA ATT CAA GAG TCT AAC CAA CCA GCG TTT GCT AGT CCA ATT CAA CCG CAA CCA GCT AAT GCT  
Glu Asn Ser Trp Gly Arg Leu Ser Thr Ala Ile Gln Glu Asp Ser Asn Gln Gly Ala Phe Ala Ser Pro Ile Gln Leu Gln Arg Arg Asn Gly  
Asp

TGC AAA TTC ACT GTC TAC GAT GTC ACT ATA TTA ATC GCT ATC ATA GCT CTC ATG GTG TAT AGA TCC CCA GCT CCA CCA TCC TCA CAG TTT  
Ser Lys Phe Ser Val Tyr Asp Val Ser Ile Leu Ile Pro Ile Ile Ala Leu Met Val Tyr Arg Cys Ala Pro Pro Pro Ser Ser Gln Phe

TCT TGT CTT ATA ACC CCA GTG GGA CCA AAT TTT AAT GCT GAT GGT TGT ATG GAT GCT GAG GCG ATA GTG GGT ATC GGA GGT CCA AAT GGT  
Leu Leu Leu Ile Arg Pro Val Val Pro Asn Phe Asn Ala Asp Val Cys Met Asp Pro Glu Pro Ile Val Arg Ile Val Gly Gln Arg Asn Gly

CTA TGT GGT GAT GGT ACC GAT GGA ACA TTC CAC AAC GCA AAC CCA ADA CCG TTG TGG CCA TCC AAC TCT AAT ACA GAT CCA AAT CAG CTC  
Leu Cys Val Asp Val Arg Asp Gly Arg Phe His Asn Gly Asn Ala Ile Gln Leu Trp Pro Cys Lys Asp Ser Asn Thr Asp Ala Asn Gln Leu

TGG ACT TGG AAA AEA GAC AAT ACT ATT CCA TCT AAT GGA AAG TGT TTA ACT ACT TAC GGG TAC AGT GGG GGA GTC TAT GTG ATG ATC TAT  
Trp Thr Leu Lys Arg Asp Asn Thr Ile Arg Ser Asn Gly Lys Cys Leu Thr Thr Tyr Gly Tyr Ser Pro Ser

GAT TCC AAT ACT GGT CCA ACT GAT ACC GCG GCG TGG CAA AEA TGG GHE AAT GGA ACC ATC ADA AAT GCG AGA TCT AGT CTA GTT TTA CCA  
Asp Cys Asn Thr Ala Ala Thr Asp Ala Thr Arg Trp Gln Ile Trp Asp Asn Gly Thr Ile Ile Asn Pro Arg Ser Ser Leu Val Leu Ala

GCG ACA TCA GCG AAC AGT GGT ACC ACA CTT ACC GTG CAA ACC AAC ATT TAT GCG GGT AGT CAA GGT TGG CTT GCT ACT AAT AAT ACA CAA  
Ala Thr Ser Gly Asn Ser Gly Thr Thr Leu Thr Val Gln Thr Asn Ile Tyr Ala Val Ser Gln Gly Trp Leu Pro Thr Asn Asn Thr His

GCT TTT GGT ACA ACC ATT GGT GGG CTA TAT GGT CCG TGC TGG CAA CCA AAT AGT GGA CAA GTA TCC ATA CAG GAC TGT ACC AGT GAA AAU  
Pro Phe Val Thr Thr Ile Val Gly Leu Tyr Gly Leu Cys Leu Gln Ala Asn Ser Gly Gln Val Trp Ile Glu Asp Cys Ser Ser Glu Lys

GCT GAA CAA CAG TGG GCT CTT TAT CCA GAT GGT TCA ATA GGT GCT CAG CAA AAC CCA GAT AAT TCC CTT ACA AGT GAT TGT AAT ATA GLE  
Ala Glu Gln Gln Trp Ala Leu Tyr Ala Asp Gly Ser Ile Arg Phe Gln Gln Asn Arg Asp Asn Cys Leu Thr Ser Asp Ser Asn Ile Arg

GAA ACA GTT GGT AAG ATC CTC TCT TGT GDU OCT CCA TCC TCT GCG CAA CCA TGG ATG TTC AAG AAT GAT GGA ACC ATT TTA AAT TTG TAT  
Glu Thr Val Val Lys Ile Leu Ser Cys Gly Pro Ala Ser Ser Gly Gln Arg Trp Met Phe Lys Asn Asp Gly Thr Ile Leu Asn Leu Tyr

ACT GGA TGG GTG TTA GAT GTG AGG CCA TGG GAT GEE AGE CTT AAA CAA ATC ATT CTT TAC GGT CTC CAT GGT GAC CCA AAC CAA ATA TGG  
Ser Gly Leu Val Leu Asp Val Arg Ala Ser Asp Pro Ser Leu Lys Gln Ile Ile Leu Tyr Pro Leu His Gly Asp Pro Asn Gln Ile Trp  
Gly His

TTA CCA TTA TTT TGA TAGACGATT ACTCTCTCC AGCTGTATG TCCGCCGATG AATATGATG OCTTAATGA AAGGACATT GTAATTTTG TAATGAAAG  
Leu Pro Leu Phe Asp

CACACGAGT TATTCAGTC CAGTATCTA TAGAGACCA ACTATTCTCT TGTGATCTT AATTT - Poly(A) - C-tail

Fig 3H-1: The preproricin cDNA sequence

1'-AAAGXXXXG GAATAGTAT TGTATATGG ATG TAT GCA GTG GCA ACA TGG CTT TGT TTT GCA TCC AGC TCA GGG TGG TCT TTC ACA TTA GAG  
 Met Tyr Ala Val Ala Thr Trp Leu Cys Phe Gly Ser Thr Ser Gly Trp Ser Phe Thr Leu Glu  
 GAT AAC AAC ATA TTC CCG AAA CAA TAC CCA ATT ATA AAC TTT ACC ACA GGG GGT GGC ACT GTG CAA AGC TAC ACA AAC TTT ATC AGA GCT  
 Asp Asn Asn Ile Phe Pro Lys Gln Tyr Pro Ile Ile Asn Phe Thr Thr Ala Gly Ala Thr Val Gln Ser Tyr Thr Asn Phe Ile Arg Ala  
 GTT CCG GGT GGT TTA ACA ACT GCA GCT GAT GTG ACA GAT GAT ATA GCA GTG TGG GCA AAG ACA GTT GGT TRG GCT ATA AAC CAA GGG TTT  
 Val Arg Gly Arg Leu Thr Thr Gly Ala Asp Val Arg His Asp Ile Pro Val Leu Pro Asn Arg Val Gly Leu Pro Ile Asn Gln Arg Phe  
 ATT TTA GTT GAA CTC TCA AAT CAT GCA GAG CTT TCT GTT ACA TTA GCG CAG GAT GTC AAT GCA TAT GGG GTC GCG TAC GGT GCT GCA  
 Ile Leu Val Glu Leu Ser Asn His Ala Glu Leu Ser Val Thr Leu Ala Leu Asp Val Thr Asn Ala Tyr Val Val Gly Tyr Arg Ala Gly  
 AAT AAC GCA TAT TTC TTT CAT GCT GAC AAT CAG CAA GAT GCA GAA GCA ATC ACT CAT CTT TTC ACT GAT GTT CAA AAT GCA TAT ACA TTC  
 Asn Ser Ala Tyr Phe Phe His Pro Asp Asn Gln Glu Asp Ala Glu Ala Ile Thr His Leu Phe Thr Asp Val Gln Asn Arg Tyr Thr Phe  
 GCG TTT GGT GGT AAT TAT GAT AGA CTT GAA CAA CTT GCT GGT AAT CAG ACA GAA AAT ATC GAG TTC GCA AAT GGT CCA CTA GAG GAG GCT  
 Ala Phe Gly Gly Asn Tyr Asp Arg Leu Glu Gln Leu Ala Gly Asn Leu Arg Glu Asn Ile Glu Leu Gly Asn Gly Pro Leu Glu Glu Ala  
 ATC TCA GGG CTT TAT TAT TAG AGT ACT GGT GGC ACT CAG CTT GCA ACT GTG GCT GGT TCC TTT ATA ATT TCC ATC CAA ATG ATT TCA GAA  
 Ile Ser Ala Leu Tyr Tyr Tyr Ser Thr Gly Gly Thr Gln Leu Pro Thr Leu Ala Arg Ser Phe Ile Ile Cys Ile Gln Met Ile Ser Glu  
 GCA GCA AGA TTC CAA TAT ATT GAG GCA GAA ATG GCG ACC AGA ATT ACC TAC AAC GCG AGA TCT GCA CCA GAT GCT AGC GTA ATT ACA GTT  
 Ala Ala Arg Phe Gln Tyr Ile Glu Gly Glu Met Arg Thr Arg Ile Arg Tyr Asn Arg Arg Ser Ala Pro Asp Pro Ser Val Ile Thr Leu  
 GAG AAT AGT TGG GGG AGA CTT TCC ACT GCA ATT CAA GAG TCT AAC CAA GCA GCG TTT GCT AGT CCA ATT CAA CTG CAA AGA GGT AAT GGT  
 Glu Asn Ser Trp Gly Arg Leu Ser Thr Ala Ile Gln Glu Ser Asn Gln Gly Ala Phe Ala Ser Pro Ile Gln Leu Gln Arg Arg Asn Gly  
 TCC AAA TTC ACT GTG TAC GAT GTG AGT ATA TTA ATC GCT ATC ATA GCT CTC ATG GTG TAT AGA TCC CCA CCT CCA CCA TGG TCA CAG TTT  
 Ser Leu Leu Ile Arg Pro Val Val Pro Asn Phe Asn Ala Asp Val Cys Met Asp Pro Glu Pro Ile Val Arg Ile Val Gly Arg Asn Gly  
 TCT TRG CTT ATA AGC CCA GTG GCA CCA AAT TTT AAT GCT GAT GTT TGT ARG GAT GCT GAG GCG ATA GTG GGT ATC GCA GGT CCA AAT GCT  
 Leu Cys Val Asp Val Arg Asp Gly Arg Phe His Asn Gly Asn Ala Ile Gln Leu Trp Pro Cys Lys Ser Asn Thr Asp Ala Asn Gln Leu  
 TGG ACT TRG AAA AGA GAC AAT ACT ATT CCA TCT AAT GCA AAG TGT TTA ACT ACT TAC GGG TAC AGT GGG GCA GTC TAT GTG ATG ATC TAT  
 Trp Thr Leu Lys Arg Asp Asn Thr Ile Arg Ser Asn Gly Lys Cys Leu Thr Thr Tyr Gly Tyr Ser Pro Gly Val Tyr Val Met Ile Tyr  
 GAT TCC AAT ACT GCT GCA ACT GAT GCG ACC GCG TGG CAA AEA TGG GAT AAT GCA ACC ATC ATA AAT GCG AGA TCT AGT GTA GTT TTA CCA  
 Asp Cys Asn Thr Ala Ala Thr Asp Ala Thr Arg Trp Gln Ile Trp Asn Asn Gly Thr Ile Ile Asn Pro Arg Ser Ser Leu Val Leu Ala  
 GCG ACA TCA GGG AAC AGT GGT ACC ACA CTT ACC GTG CAA ACC AAC ATT TAT GCG GTT AAT CAA GGT TGG CTT GCT ACT AAT AAT ACA CAA  
 Ala Thr Ser Gly Asn Ser Gly Thr Thr Leu Thr Val Gln Thr Asn Ile Tyr Ala Val Ser Gln Gly Trp Leu Pro Thr Asn Asn Thr Gln  
 OCT TTT GTT ACA ACC ATT GTT GGG GCA TAT GGT CTG TCC TRG CAA GCA AAT AGT GCA GAA GTA TGG ATA GAG GAG TGT ACC AGT GAA AAG  
 Pro Phe Val Thr Thr Ile Val Gly Leu Tyr Gly Leu Cys Leu Gln Ala Asn Ser Gly Gln Val Trp Ile Glu Asp Cys Ser Ser Glu Lys  
 OCT GAA CAA CAG TGG GCT CTT TAT CCA GAT GGT TCA ATA GGT OCT CAG CAA AAC CCA GAT AAT TGG CTT ACA AGT GAT TCT AAT ATA GAG  
 Ala Glu Gln Gln Trp Ala Leu Tyr Ala Asp Gly Ser Ile Arg Pro Gln Gln Asn Arg Asp Asn Cys Leu Thr Ser Asp Ser Asn Ile Arg  
 GAA ACA GTT GTT AAG ATC CTC TCT TGT GCG OCT CCA TCC TCT GCG CAA GCA TGG ATG TTC AAG AAT GAT GCA ACC ATG TTA AAT TRG TAT  
 Glu Thr Val Val Lys Ile Leu Ser Cys Gly Pro Ala Ser Ser Gly Gln Arg Trp Met Phe Lys Asn Asp Gly Thr Ile Leu Asn Leu Tyr  
 AAT GCA TRG GTG TTA GAT GTG AGG CCA TGG GAT GGG AGC CTT AAA CAA ATC ATT CTT TAC GCT CTC GAT GGT GAC CCA AAC CAA ATA TGG  
 Ser Gly Leu Val Leu Asp Val Arg Ser Asp Pro Ser Leu Lys Gln Ile Ile Leu Tyr Pro Leu His Gly Asp Pro Asn Gln Ile Trp  
 TTA CCA TTA TTT TGA TAGACAGATT ACTCTCTCC ACTGTGTATG TCGTCCATG AAGATAGATG GCTTAATTA AAGGACATT GTAATTTT TAACTGAAG  
 Leu Pro Leu Phe  
 GACACCGAT TATTGAGTC CAGTATCTAA TAGAGACCA ACTATTCTCT TGTGATTCI AATTTT - Poly(A) - C-tail

Fig 3H-1: The preproricin cDNA sequence

### 3I The preproagglutinin sequence

The preproagglutinin sequence is shown in fig 3I-1, compared with Funatsu's ricin A and B chain sequences. The format is as fig 3H-1.

The overall structure of the sequence is the same as that of the prororicin sequence, but the A chain is three bases shorter, corresponding to deletion of ricin's alanine at position 130. An additional potential N-glycosylation site is located at residues 357 - 359, in the B chain, and the A chain contains two cysteine residues, at positions 84 and 156, which are absent in the ricin sequence.

Both sequences begin at the 5' end at the same nucleotide: this may reflect homology extending towards the 5' terminus, resulting in identical  $S_1$  nuclease cleavage points for both sequences.

The preproagglutinin sequence contains only 69 bases of 3' non-coding region, and these are identical to their counterparts in the prororicin sequence.

There are numerous differences in the coding regions, which will be discussed in detail in following sections.

Clone pRCL52 contains bases -102 to 1566, and clone pRCL57 contains bases 538 to 1689, though not all of this clone has been sequenced. No differences were found in the overlapping regions.

5'-AACGGGGGGA GAATACTAT TGTATATATU ATU TAT GUG GUG OCA ACA TGG CTT TGT TTT GUA TOC ACC TUA GUG TGU TCT TTC ACA TTA (24)  
Met Tyr Ala Val Ala Thr Trp Leu Cys Phe Gly Ser Thr Ser Gly Trp Ser Phe Thr Leu Glu

GAT AAC AAC ATA TTC CCG AAA GAA TAC OCA ATT ATA AAC TTT ACC ACA OCA GAT GGC ACT GUG GAA AGC TAC ACA AAC TTT ATC AIA GUT  
Asp Asn Asn Ile Phe Pro Lys Gln Tyr Pro Ile Ile Asn Phe Thr Thr Ala Asp Ala Thr Val Glu Ser Tyr Thr Asn Phe Ile Arg Ala

GUG CCG AGT CAT TTA ACA ACT GCA OCT GAT GGG ACA CAT GAA GAA ATA OCA GUG TGC OCA AAC ADA GTT GGT TGU OLT ATA ACC CAA GUG TTT  
Val Arg Ser His Leu Thr Thr Gly Ala Asp Val Arg His Glu Ile Pro Val Leu Pro Asn Arg Val Glu Ser Tyr Thr Asn Phe Ile Arg Phe

ATT TTA GTT GAA CTC TCA AAT CAT GCA GAG CTT TCT GTT ACA TTA OCA CIG GAT GTC ACC AAT OCA TAT GTG GTC GGC TUC CCG GCT GGA  
Ile Leu Val Glu Leu Ser Gln Asn His Ala Glu Leu Ser Val Thr Leu Ala Leu Asp Val Thr Asn Ala Tyr Val Val Gly Cys Arg Ala Gly

AAT ACC CCG TAT TTC TTT CAT OCT GAC AAT CAA GAA GAT OCA GAA ACA ATC ACT CAT CTT TTC ACC GAT GTT CAA AAT TCA TTT ACA TTC  
Asn Ser Ala Tyr Phe Phe His Pro Asp Asn Gln Glu Asp Ala Glu Ala Ile Thr His Leu Phe Thr Asp Val Gln Asn Ser Phe Thr Phe

GCC TTT GGT GGT AAT TAT GAT AGA CTT GAA CAA CTT OCA GGT CIG AGA GAA AAT ATT CAG TTG GCA ACT GGT OCA TTA GAG CAC GCT ATC  
Ala Phe Gly Gly Asn Tyr Asp Arg Leu Glu Gln Leu Gly Leu Arg Glu Asn Ile Glu Leu Gly Thr Asn Gly Pro Leu Glu Asp Ala Ile

TCA CCG CTT TAT TAT TAT AGT AGT TGT GGC ACT CAG ATT OCA ACT CIG OCT GGT TOC TTT ATG GTT TOC ATC CAA ATG ATT TCA GAA OCA  
Ser Ala Leu Tyr Tyr Tyr Ser Thr Cys Thr Gly Gln Ile Pro Thr Leu Ala Arg Ser Phe Met Val Ile Ile Ile Ile Ile Ile Ile Ile Ile Ile

OCA ACA TTC CAG TAC ATT CAG GCA GAA ATG CCG ACC AGA ATT AGG TAG AAC CCG AGA TCT OCA CCA GAT OCT ACC GTA ATT ACA CTT GAG  
Ala Arg Phe Gln Tyr Ile Glu Gly Glu Met Arg Thr Arg Ile Arg Tyr Asn Arg Arg Ser Ala Pro Asp Pro Ser Val Ile Thr Leu Glu

AAT AGT TGG CCG AGA CTT TOC ACT UCA ATT CAA GAG TCT AAC CAA GGA GGC TTT GCT AGT OCA ATT CAA CTG CAA ACA GAT AAC GGT TUG  
Asn Ser Trp Gly Arg Leu Ser Thr Ala Ile Gln Glu Ser Asn Gln Gly Ala Phe Ala Ser Pro Ile Gln Leu Gln Arg Arg Asn Gly Ser

AAA TTC AAT GUG TAC GAT GUG AGT ATA TTA ATC OCT ATC ATA OCT CTC ATG GUG TAT AGA TOC GCA OCT OCA GUG TUG TCA CAG TTT TCT  
Lys Phe Asn Val Tyr Asp Val Ser Ile Leu Ile Pro Ile Ile Ala Leu Met Val Tyr Arg Cys Ala Pro Pro Pro Ser Ser Gln Phe Ser

TTG CTT ATA AGC OCA GUG GUG OCA AAT TTT AAT GCT GAT GGT TGT ATG GAT OCT GAG CCG ATA GUG GGT ATC GTA GGT GGA AAT GGT CTA  
Leu Leu Ile Arg Pro Val Val Pro Asn Phe Asn Ala Asp Val Cys Met Asp Pro Glu Pro Ile Val Arg Ile Val Gly Arg Asn Gly Leu

TGT GTT GAT GTT ACA GGT GAA GAA TTC TTC GAT GCA AAC OCA ATA CAA TIG TGG OCA TGC AAA TCT AAT ACA GAT TGG AAT CAG TTA TGG  
Cys Val Asp Val Thr Gly Glu Glu Phe Phe Asp Gly Asn Pro Ile Gln Leu Trp Pro Cys Lys Ser Asn Thr Asp Trp Asn Gln Leu Trp

ACT TIG ACA AAA GAT ACC ACT ATT GCA TCT AAT GGC AGG TGT TIG ACC ATT TGC AAG TGC AGT OCA AGA CAG CAG GTC GTC ATA TAT AAT  
Thr Leu Arg Lys Asp Ser Thr Ile Arg Ser Asn Gly Lys Cys Leu Thr Ile Ser Lys Ser Ser Pro Arg Gln Gln Val Val Ile Tyr Val

TOC AGT ACC OCT ACA GTT GGT CCG ACC GGT TGG CAA ATA TGG GAC AAT OCA ACC ATC ATA AAT CCG OCA TCT GGT GTA GGT TTG OCA CCG  
Cys Ser Thr Ala Thr Val Gly Ala Thr Arg Trp Gln Ile Trp Asp Asn Arg Thr Ile Ile Asn Pro Arg Ser Gly Leu Val Leu Ala Ala

ACA TCA CCG AAC AGT GGT ACC AAA CTT ACA GUG CAA ACC AAG ATT TAT GGC GGT AGT CAA GGT TGC CTT GLT ACT AAT AAT ACA CAA CTT  
Thr Ser Gly Asn Ser Gly Thr Thr Lys Leu Thr Val Gln Thr Asn Ile Tyr Ala Val Ser Gln Gly Trp Leu Pro Thr Asn Asn Thr Gln Pro

TTT GUG ACA ACC ATT GTT CCG GTA TAT GGC AAG TGC TTG CAA OCA AAT AGT GGA AAA GTA TUG TTA CAG CAC TGT ACC AGT GAA AAG GCT  
Phe Val Thr Thr Ile Val Gly Leu Tyr Gly Met Cys Leu Gln Ala Asn Ser Gly Lys Val Trp Leu Glu Asp Cys Thr Ser Ser Glu Lys Ala

GAA CAA CAA TGG OCT CTT TAT OCA GAT GGT TCA ATA GGT OCT CAG CAA AAC CCG GAT AAT TUC CTT ACA ACT GAT OCT AAT ATA AAA (24)  
Glu Gln Gln Trp Ala Leu Tyr Ala Asp Gly Ser Ile Arg Pro Gln Gln Asn Arg Asp Asn Cys Leu Thr Thr Asp Ala Asn Ile Lys Gly

ACA GTT GTC AAG ATC CTC TCT TGT GXC OCT GCA TOC TCT GGC CAA GCA TGG ATG TTC AAG AAT GAT GUA ACC ATT TTA AAT TRT TAT AAT  
Thr Val Val Lys Ile Leu Ser Cys Gly Pro Ala Ser Ser Gly Gln Arg Trp Met Phe Lys Asn Asp Gly Thr Ile Leu Asn Leu Tyr Asn Ser

OCA TTG GUG TTA GAT GUG AGC OCA TGG GAT CCG ACC CTT AAA CAA ATC ATT GTT CAC OCT TTC CAT GUA AAC CTA AAC CAA ATA TRT TTA  
Gly Leu Val Leu Asp Val Arg Ser Asp Pro Ser Leu Lys Gln Ile Ile Val His Pro Phe His Gly Asn Leu Asn Gln Ile Trp Leu

OCA TTA TTT TGA TAGACGATT ACTCTTTTC AUTOTATATU TCTTCCACT AATATAGATG OCTAATAA AAGGGA  
Pro Leu Phe Pro

Fig 3I-1: The preproagglutinin cDNA sequence

### 3J Analysis of protein sequences

#### 3J1 Comparison with protein-determined sequences

The molecular weights of the deduced sequences are shown in fig 3J-1, along with those calculated from Funatsu's data. Values for the ricin chains are close to those of Funatsu; the RCA chains are of similar sizes to those of ricin.

The comparison of the two deduced sequences with the protein-determined data already presented (fig 3G-1) includes all differences, regardless of cause. As will be shown, many of these may result from errors in the protein sequence data, though it must be borne in mind that seed proteins do tend to be heterogeneous, and that some of the differences which might be explicable on these grounds may in fact be genuine differences. It must also be remembered that cDNA synthesis does not always produce absolutely faithful copies of an mRNA sequence (252,253), so that some of the differences may be errors in the cDNA.

Fig 3J-2 lists all the differences between the sequence determined by Funatsu and the deduced ricin sequence, and classifies them into those explicable by transposition of residues, those due to misidentification or missing of tryptophan, and those involving confusion with amidated residues. Those lacking comments are assumed to be inexplicable by any obvious problem in protein sequencing. These remaining differences are plotted in fig 3J-3, which shows that the two sequences assume a much greater degree of similarity than is at first apparent.

Funatsu's sequence data were obtained from A and B chains isolated either by performic acid oxidation of native ricin, or by reduction with 2-mercaptoethanol (51). In the case of the A chain, much of the sequence was determined solely on oxidised chains (33,34), while the B chain was determined both on

**Fig 3J-1: Amino acid compositions and molecular weights**

<u>Amino acid</u>	<u>Ricin A</u>	<u>Ricin B</u>	<u>RCA A</u>	<u>RCA B</u>	<u>Signal</u>	<u>Linker</u>
Ala	24	14	23	12	2	-
Cys	2	9	4	9	1	-
Asp	9	17	10	14	1	-
Glu	14	5	15	6	1	-
Phe	14	4	15	6	2	1
Gly	17	20	15	20	2	-
His	4	2	5	2	-	-
Ile	22	18	22	18	-	1
Lys	2	7	2	11	-	-
Leu	23	24	21	23	2	2
Met	3	3	4	3	1	-
Asn	17	21	15	21	2	2
Pro	15	13	15	13	-	2
Gln	14	15	13	16	-	-
Arg	21	14	19	13	-	1
Ser	19	20	21	19	3	1
Thr	17	21	18	22	3	-
Val	15	17	16	19	1	2
Trp	1	9	1	10	2	-
<u>Tyr</u>	<u>14</u>	<u>9</u>	<u>12</u>	<u>5</u>	<u>1</u>	<u>-</u>
<b>Total</b>	<b>267</b>	<b>262</b>	<b>266</b>	<b>262</b>	<b>24</b>	<b>12</b>
<b>MW (Dal)</b>	<b>29399</b>	<b>28531</b>	<b>29261</b>	<b>28584</b>	<b>2670</b>	<b>1309</b>
<b>Funatsu</b>	<b>29172</b>	<b>27944</b>				

MW of preproricin is 61855 Dal.

MW of preproagglutinin is 61779 Dal.

Fig 3J-2: Amino acid differences

List of amino acid differences between the ricin sequences according to Funatsu's group and that deduced from the cDNA. Numbering is from fig 3H-1; a + following a number indicates position of a residue in Funatsu's sequence which is absent in the cDNA-derived sequence.

Position	cDNA	Funatsu	Comments
A 41	Asp	Glu	Alternative acid
63	Ser	Gln	
75	Asp	Ser	
235	Arg	---	Misinterpretation of Arg-Arg as Arg
236	Asn	Asp	Acid-amide
249	Ile	Leu )	Transposed Leu plus missed Ile
254	Leu	---	
B 301	Asp	Asn	Acid-amide
308	His	Asn )	Transposition
309	Asn	His )	
328	Trp	---	Missed Trp
349	Ser	Pro )	Transposition
350	Pro	Ser )	
365	Asp	Thr )	Transposition
367	Thr	Asp )	
369	Trp	---	Missed Trp
370	Gln	Glu	Acid-amide
410	Trp	Pro )	Complex transposition
412	Pro	Phe )	
419	Phe	Trp )	
439	Trp	Val	Misidentification of Trp?
443	Cys	Ser )	Transposition
444	Ser	Cys )	
457	Asp	Ser )	Transposition plus acid-amide
459	Ser	Asn )	
461	Arg	Asn )	Transposition
465	Asn	Arg )	
493	Gln	Glu	Acid-amide

Continued....

Fig 3J-2 (continued)

<u>Position</u>	<u>cDNA</u>	<u>Funatsu</u>	<u>Comments</u>
B 516	Arg	Ala	
529+	---	Trp	Incorrect position of Trp-537
530	His	Gly	} Transposition
531	Gly	His	
535+	---	Leu	Incorrect position of Leu-540
537	Trp	---	Correct position of Trp (529+)
540	Leu	---	Correct position of Leu (535+)

For further details and comments see text.



oxidised material and on reduced, carboxymethylated chains (35,36). Performic acid treatment of proteins is not entirely specific for modification of cysteine residues - it is well-known that tryptophan residues are largely degraded to kynurenine and other products (254); tyrosine and histidine residues may also be modified and degraded (254). Serine and threonine residues may undergo acyl shifts - that is, the peptide bond is converted into an ester linkage - though this reaction is reversible and its significance to sequencing studies is unclear (254). Methionine is also modified, almost quantitatively, to the sulphone, but this derivative presents no problems. Funatsu (34) comments that methionine was indeed converted to the sulphone (or the sulphoxide), that cysteine residues were partially converted to cysteic acid, but that no modification of tyrosyl or histidyl residues was observed. In the case of the A chain, tryptophan residues were checked by cleavage with N-chlorosuccinimide, followed by N-terminal analysis of the product mixture - this confirmed that the A chain contains only one tryptophan. No such analysis is reported for the B chain, possibly because it contains several tryptophan residues and the products may have been complex. This is unfortunate, as several of the possible errors in the B chain concern tryptophan residues.

Reduction and carboxymethylation of the chains may also result in sequencing problems, in that reports suggest that some 2 % of lysine residues may be carboxymethylated, and that up to 10 % of methionine residues may be degraded (254). However, the low proportions are unlikely to have a very great effect, and appear not to have caused Funatsu any difficulty.

The differences between the amino acid sequences determined by Funatsu's group and those deduced from the cDNA will now be discussed in more detail.

The C-terminus of the A chain was only sequenced in one fragment by Funatsu (33,34): a peptic peptide of 18 residues, designated fragment T-23. Part of this fragment was also examined by determination of the amino acid composition of a small (6 residue, according to Funatsu) peptic fragment derived from a large cyanogen bromide fragment (designated CBII-P2):

Funatsu (33,34): S K F S V Y D V S I L L P I I A - M V Y R C A  
cDNA: S K F S V Y D V S I L I P I I A L M V Y R C A

T-23  
CBII-P2

The composition of CBII-P2 is given as Pro Ala Ile<sub>2</sub> Leu Hse (Hse = homoserine, product of cyanogen bromide action on methionine). It appears that the leucine in fragment T-23 has been misplaced, and that the failure to detect the extra isoleucine in the composition analysis of CBII-P2 has resulted in this not being corrected. It is noteworthy that both apparent errors are contained within both fragments.

The next region to be discussed is that which starts at residue 355 - following this residue there is a transposition, an omitted tryptophan and a possible asparagine-aspartic acid confusion. The fragments used by Funatsu are shown below, along with the deduced sequence, and the sequence obtained by Li's group (50):

Funatsu (35,36): M I Y D C N T A A T T A D R - E I W N N

Li (50): M I Y D C N T A A T D A T R W E I W

cDNA: M I Y D C N T A A T D A T R W E I W D N

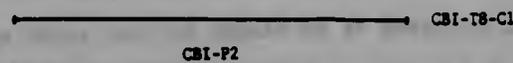


Of the first fragment, CBI, only the first 5 residues were sequenced. Its subfragment CBI-T4 was completely sequenced, though the fragment P11 was not sequenced (its composition is given as Asp<sub>2</sub> Thr<sub>2</sub> Ala<sub>2</sub> Ile Tyr, ie 8 amino acids, though Funatsu draws it as 9 residues in length). However, the region containing the transposition was sequenced twice (in CBI-T4 and in CBI-P3). Similarly, the region with the missing tryptophan and the acid-amide confusion was also sequenced twice (in CBI-T1 and in CBI-P3). Note that the transposition was never split over two fragments. Li's group also sequenced the cyanogen bromide fragment corresponding to Funatsu's CBI and obtained a sequence identical to that from the cDNA sequence (50), which supports the possibility that Funatsu is in error in this region.

The 'complex transposition' indicated in fig 3J-2 occurs at residues 410, 412 and 419. This region was sequenced twice by Funatsu:

Funatsu (36): T N I Y A V S Q G P L F T N N T Q P W V T T I V G

cDNA: T N I Y A V S Q G W L P T N N T Q P F V T T I V G





As can be seen from fig 3J-2, three differences between the Funatsu and the cDNA-deduced sequences remain, two in the A chain sequence and one in the B chain sequence. A similar analysis has not been performed for the putative agglutinin sequence, because this differs much more from the Funatsu sequences. It must once again be emphasised that some of the apparent errors in Funatsu's data may not be errors, but may reflect heterogeneity between different castor bean varieties. The beans used in this study are not of defined lineage, and may possibly contain other mRNAs encoding proteins more closely related to those sequenced by Funatsu's group than the ones which were cloned here.

The deduced ricin A chain sequence contains one potential N-glycosylation site absent from Funatsu's sequence, caused by the presence of an asparagine residue at position 236, where Funatsu's sequence has aspartic acid. This is in accord with recent results, which indicate that the ricin A chain does carry two oligosaccharide moieties (B Foxwell, personal communication).

3J2 Comparison of deduced amino acid sequences

The two deduced amino acid sequences are presented in fig 3J-4, differences between them being indicated by vertical lines.

The signal and linker peptides have identical sequences; the A chains are identical at all but 18 positions, and are thus 93.3 % homologous, while the B chains differ at 41 positions and are 84.4 % homologous.

The differences between the sequences are listed in fig 3J-5, along with their codons. 45 changes are due to single base substitutions, 10 involve alteration of two bases within a codon, while three changes result from complete replacement of codons. One codon present in the preproricin sequence is absent in the preproagglutinin. The reversal of lysine and arginine at positions 330 and 331 may be a cDNA synthesis artefact, since it involves displacement of a G in a run of A's (the agglutinin has AGA-AAA and ricin has AAA-AGA). Clearly, several changes could be explained as cDNA synthesis errors, which are known to occur (252). However, the number of changes is far too high for many to be due to such errors - the rate of misreading by reverse transcriptase has been estimated as one in 600 bases (253). Confirmation of changes can only be achieved by sequencing more clones for both sequences.

Of the total of 59 amino acid differences, 22 might be expected not to alter the properties of the protein significantly, as in these cases the amino acids in both proteins are of similar characteristics (that is, they are members of the same groups as defined by von Heijne (263) - eg hydrophobic, aromatic etc). These changes are marked with asterisks in fig 3J-5. A further four changes

```

(MYAV ATMLCGGSTIS GASFTLENN)IFPKQYPIIN FTTAGATVQS YNFIRAVRG RLITGADVRH DIPVLPNRVG LPINQRFILV
(MYAV ATMLCGGSTIS GASFTLENN)IFPKQYPIIN FTTADATVES YNFIRAVRS HLTTCADVRH EIPVLPNRVG LPISQRFILV

EISNHAELSV TLALDVNAY WGRAGNSA YFFHPNQED AEATHLFTD VQRVYFAFG GNYRLEQLA GNRENIELG NEPLEEAISA
EISNHAELSV TLALDVNAY WGRAGNSA YFFHPNQED AEATHLFTD VNSFTFAFG GNYRLEQL- GELRENIELG TGPLDAAISA

LYYSYGGIQ IPTLARSFII CIQMISEAAR FQYIEGEMRT RIRYRNSAP DPSVITLNS WERLSTAIQE SNOCAFASPI QLQRRNGSKF
LYYSYGGIQ IPTLARSFNW CIQMISEAAR FQYIEGEMRT RIRYRNSAP DPSVITLNS WERLSTAIQE SNOCAFASPI QLQRRNGSKF

SVYDVSILIP IIALMVRCA PFPSSQF(SL LIRPWFNFN)ADWCIDPEPI VRIVERGELC VDVRDERFHN GNALQLWPKC SNTDANQLMT
NVDVSILIP IIALMVRCA PFPSSQF(SL LIRPWFNFN)ADWCIDPEPI VRIVERGELC VDVTGEEFFD GNPIQLWPKC SNTDANQLMT

LRDNTIRSN GKCLITIGYS PGVWMIYDC NTAATDTRW QIMDNGTIIN PRSSLVLAAT SENSCTILTV QNIYAVSQG WLPTNNTQPF
LRDNTIRSN GKCLITIKSS PRQWVYINC STATVGTATRW QIMDNTIIN PRSELVLAAT SENSCTKLV QNIYAVSQG WLPTNNTQPF

VITIVGLXEL CLOANSQWV IEDCSSEKAE QQALYADGS IRPQQRUNC LITSNSIRET WKILSOEPA SSGQRAMFKN DGTILNLYSG
VITIVGLXCH CLOANSKAW LEDCTSEKAE QQALYADGS IRPQQRUNC LITDANKGT WKILSOEPA SSGQRAMFKN DGTILNLYNG

LVLVRRSDP SLKQIILYPL HEDNQIMLP LF
LVLVRRSDP SLKQIIVHPF HENLNQIMLP LF

```

Fig 3J-4: The preproricin amino acid sequences is above that of the preproagglutinin; the signal sequence and linker peptide are enclosed in brackets. Vertical lines indicate differences.

Continued....

Fig 3J-5: Amino acid differences

Amino acid differences between the two deduced sequences are listed, along with their codons. Numbering is according to the preproagglutinin sequence. One asterisk indicates that both residues shown are of similar properties (see text), while two indicate an amidated residue change. Residue 129+ is an extra amino acid in the preproargin sequence which follows residue 129 of the preproagglutinin sequence.

Position	Aggl	Ricin	Aggl	Ricin
(A) 15	Asp	Gly	GAT	GGT
19	Glu **	Gln	GAA	CAA
30	Ser	Gly	AGT	GGT
31	His	Arg	CAT	CGT
41	Glu *	Asp	GAA	GAT
54	Ser	Asn	AGC	AAC
84	Cys	Tyr	TGC	TAC
114	Ser	Arg	TCA	CGA
115	Phe *	Tyr	TTT	TAT
129+	---	Ala	---	GCT
131	Gly	Asn	GGT	AAT
140	Thr	Asn	ACT	AAT
145	Asp *	Glu	GAC	GAG
156	Cys	Gly	TGT	GGT
160	Ile *	Leu	ATT	CTT
168	Met *	Ile	ATG	ATA
169	Val *	Ile	GTT	ATT
240	Asn	Ser	AAT	AGT
(B) 302	Thr	Arg	ACA	AGG
303	Gly	Asp	GGT	GAT
304	Glu	Gly	GAA	GGA
305	Glu *	Arg	GAA	AGA
307	Phe *	His	TTC	CAC
308	Asp **	Asn	GAT	AAC
311	Pro	Ala	CCA	GCA
323	Trp	Ala	TGG	GCA
330	Arg *	Lys	AGA	AAA
331	Lys *	Arg	AAA	AGA
333	Ser	Asn	AGC	AAT

Continued....

Fig 3J - 5 (continued)

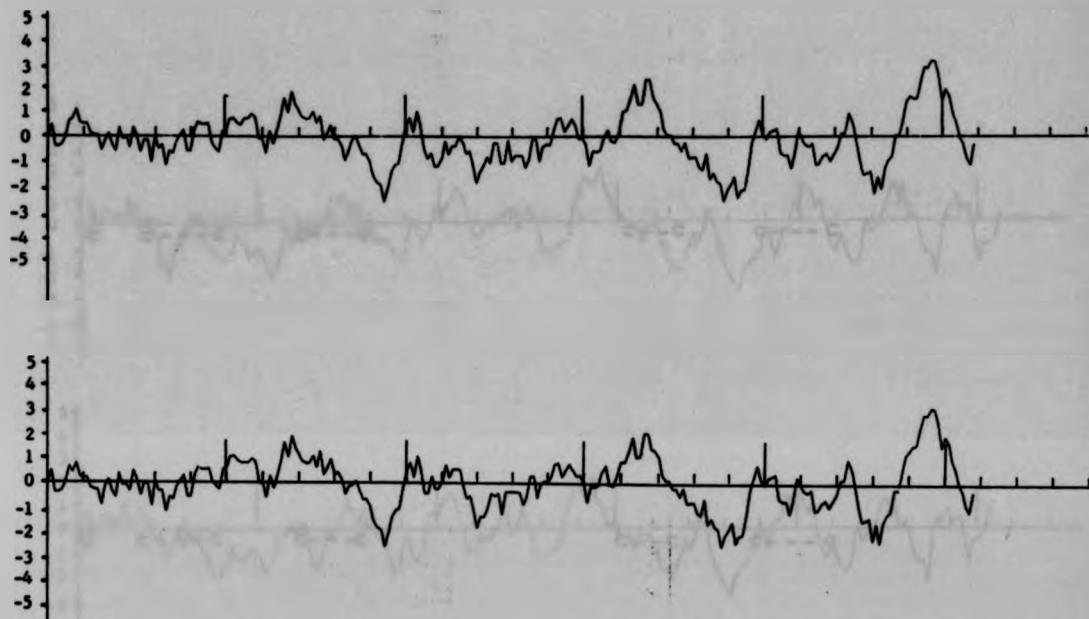
Position	Aggl	Ricin	Aggl	Ricin
(B) 344	Ile	Thr	ATT	ACT
345	Ser	Tyr	TCC	TAC
346	Lys	Gly	AAG	GGG
347	Ser *	Tyr	TCC	TAC
350	Arg	Gly	AGA	GGA
351	Gln	Val	CAG	GTC
352	Gln	Tyr	CAG	TAT
354	Val *	Met	GTG	ATG
357	Asn **	Asp	AAT	GAT
359	Ser	Asn	AGT	AAT
362	Thr *	Ala	ACA	GCA
363	Val	Thr	GTT	ACT
364	Gly	Asp	GGT	GAT
374	Arg	Gly	CGA	GGA
382	Gly	Ser	GGT	AGT
395	Lys	Thr	AAA	ACA
428	Met *	Leu	ATG	CTG
436	Lys	Gln	AAA	CAA
439	Leu *	Ile	TTA	ATA
443	Thr *	Ser	ACC	AGC
471	Thr *	Ser	ACT	AGT
473	Ala *	Ser	GCT	TCT
476	Lys *	Arg	AAA	CGG
477	Gly	Glu	GGA	GAA
507	Asn	Ser	AAT	AGT
525	Val *	Leu	GTT	CIT
526	His *	Tyr	CAC	TAC
528	Phe *	Leu	TTC	CTC
531	Asn **	Asp	AAC	GAC
532	Leu	Pro	GTA	CCA

involve amidated residues, and are indicated in fig 3J-5 by double asterisks.

The effects of the differences in sequence were examined by generating hydropathy plots of the A and B chains, using the programme described in Appendix 1. Hydropathy values determined by Kyte and Doolittle (255) were used: these are averaged over a number of residues (the span), and plotted against position in the protein. Kyte and Doolittle noted that spans of 7 - 11 residues gave the most meaningful results in terms of signal-to-noise ratio, and best correlation with structural features determined by X-ray diffraction. That is, hydrophilic regions tend to be located on the external surface of the molecule, while hydrophobic regions are buried in the interior.

The resulting plots are shown in figs 3J-6 (A chains) and 3J-7 (B chains); the latter also shows the positions of the disulphide bridges. The cysteine residues, and consequently, the disulphide bridges, are conserved. The bridges tend to occur at the limits of hydrophobic regions - these may thus represent polypeptide loops associating together in the inside of the molecule, a feature which may be required to construct the sugar-binding sites by bringing together different parts of the chains at the surface.

Fig 3J-8, a difference plot for hydropathy, shows that the region of greatest difference coincides with the second hydrophobic loop; the agglutinin has a less hydrophobic loop than does ricin. The agglutinin loop may be less deeply buried than that of ricin, with the possible result that different residues are exposed at the surface - causing different sugar specificity. Differently exposed residues may also participate differently in intermolecular interactions, allowing the agglutinin to form its dimers of



**Fig 3J-6: Hydropathy plots of A chains.**

**The ricin A chain hydropathy plot is shown above, and the agglutinin below. Both were determined with span = 9 residues.**

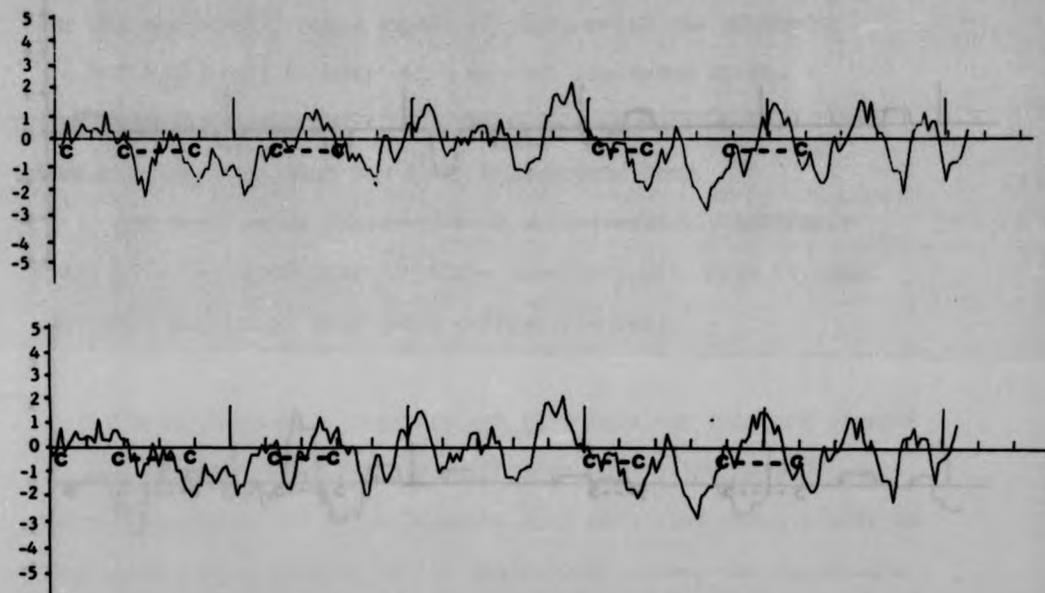


Fig 3J-7: Hydropathy plots of B chains.

The ricin B chain hydropathy plot is shown above, and the agglutinin below. Both were determined with span = 9 residues.

Positive peaks indicate that ricin is more hydrophobic than the agglutinin at the region of the peak, and vice versa.

The A chains are shown above, and the B chains below.

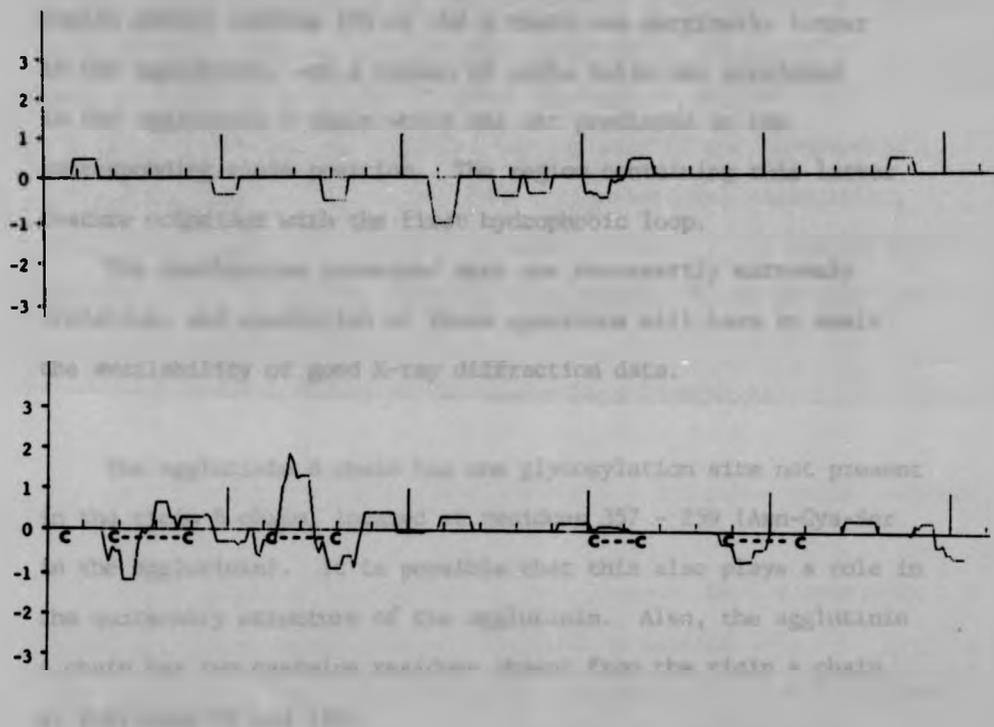


Fig 3J-8: Hydrophathy difference plots

Hydrophathies over a span of 9 residues were determined simultaneously for both sequences, and the difference was plotted. Positive peaks indicate that ricin is more hydrophobic than the agglutinin at the region of the peak, and vice versa. The A chains are shown above, and the B chains below.

ricin-like species.

Secondary structures were predicted by the method of Garnier et al, as described in Appendix 1. No gross differences emerged between ricin and the agglutinin, though an alpha-helical region around residue 100 of the A chain was marginally longer in the agglutinin, and a region of alpha helix was predicted in the agglutinin B chain which was not predicted in the corresponding ricin position. The region containing this latter feature coincides with the first hydrophobic loop.

The conclusions presented here are necessarily extremely tentative, and resolution of these questions will have to await the availability of good X-ray diffraction data.

The agglutinin B chain has one glycosylation site not present in the ricin B chain, located at residues 357 - 259 (Asn-Cys-Ser in the agglutinin). It is possible that this also plays a role in the quaternary structure of the agglutinin. Also, the agglutinin A chain has two cysteine residues absent from the ricin A chain, at positions 84 and 156.

### 3J3 The signal and linker peptides

The cDNA sequences predict the presence of 24 amino acids preceding the N-terminus of the mature lectin A chain, these being the same for both proteins. These are presumed to represent signal sequences involved in the cotranslational segregation of the nascent polypeptides, and have been observed in all protein body proteins thus far investigated (eg refs 87,95,256-262).

It is unlikely that the signals play any role in the topogenesis of protein body proteins other than cotranslational segregation, since they are cleaved cotranslationally, though it is conceivable that they may direct the proteins to a receptor which carries them to their ultimate destination.

The signal sequence of the castor bean lectins is:

M Y A V A T W L C F G S T S G W S F T L E D N N - mature A chain

The hydrophobic properties of this sequence were examined by producing a hydropathy plot as described in section 3J2. Fig 3J-9 shows the results obtained with spans of 1 and 5 residues. The former shows hydropathies for individual amino acids; the span of 5 was used because the wider spans previously used are inappropriate for so short a sequence.

The lower plot (span = 5) indicates that the first 10 residues are in general hydrophobic, though the upper plot (span = 1) reveals the presence of hydrophilic residues among them. This is not entirely characteristic of signal sequences - they usually have a hydrophobic core, rather than a hydrophobic beginning, and usually terminate with a small uncharged residue (263).

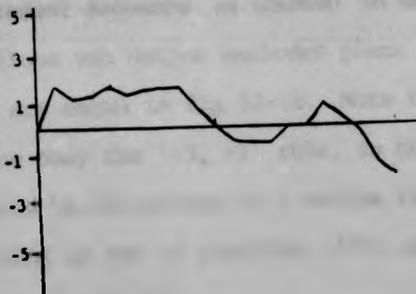
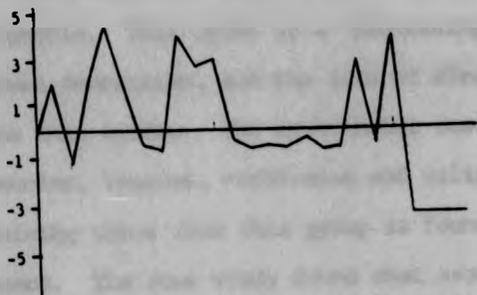


Fig 3J-9: Signal sequence hydropathy plots.

Plots were constructed as described in the text, with span = 1 (above) and span = 5 (below).

Von Heijne examined 78 eukaryotic signal sequences (no plant signals were included) and formulated a set of rules for predicting the cleavage site (263): starting from the N-terminus, a search is carried out for a group of four residues, three of which are hydrophobic. This opens up a 'processing window' a number of residues downstream, and the site of cleavage is predicted in detail within this window. The hydrophobic residues are phenylalanine, isoleucine, leucine, methionine and valine: no quadruplet containing three from this group is found within the present signal sequence. The same study found that aspartate residues are never present in position -3 (with respect to the cleavage site), and in only one case was an asparagine or a glutamine found at position -1. The present sequence is unusual in having both of these features.

Since von Heijne excluded plant signal sequences, a small group are shown in fig 3J-10. Note that all except the castor bean lectins obey the '-3, -1' rule, in that the -1 position is always occupied by an alanine or a serine residue, and no forbidden residues are found at the -3 position (270) and that proline does not occur between -3 and -1.

Von Heijne suggests (270) that selection of residues in this region is designed to offer a single site to signal peptidase, other potential sites being blocked by the presence of forbidden residues. The unusual nature of the castor bean cleavage site may perhaps be explicable by an unusual signal peptidase: the -1 position is occupied by asparagine, which may be associated with precursor processing sites (see later in this section).

Although one study (271) suggested that only entirely non-polar signals can interact with the receptor, several signals do contain polar residues, in the hydrophobic core (263). The hydrophobic region of the castor bean lectin signal may thus represent a more extreme case of this observation.

<u>Protein</u>	<u>Cleavage site</u>	<u>Reference</u>
Thaumatin II	TLSRA/ATFE	264
Zein clone A30	SAASA/TIFP	265
Zein p22.1	SATNA/FIIP	276
Phaseolus lectin	SHANS/ATET	266
Phaseolin	SASFA/TSLR	258
Pea legumin	GGCFA/LREQ	267
Pea vicilin	VCVSS/RSDQ	268
Pea seed lectin	FKVNS/TETT	262
Napin	LLTNA/STYR	95
Wheat gliadin	TATTA/VRFP	269
Barley amylase	GALAS/GHQV	275
Soybean seed lectin	SKANS/AETV	262
<u>Castor bean lectins</u>	<u>LEDNN/IFPK</u>	<u>Present work</u>

**Fig 3J-10: Plant signal sequence cleavage sites**

Cleavage sites shown were determined by comparison with mature proteins, or are most likely sites - by a variety of criteria, for which see references cited.

Between the C terminus of the A chain sequence and the N terminus of the B chain, the cDNA precursor sequence reveals a 12 amino acid linker peptide, which is the same for both lectins. This is shown below, along with the termini of the chains:

A chain - P P P S S Q F / S L L I R P V V P N F N / A D V C M - B chain

Linker peptides from a few other proteins have also been sequenced: that of napin (95) is 19 or 20 residues long, depending on heterogeneity, while that of pea seed lectin is 6 residues (94), though identification in this case is not final. By analogy, lentil lectin, which is homologous to pea seed lectin, may have a tetrapeptide linker, while conA, which is otherwise homologous to these two lectins, is not cleaved - and is divergent at the processing site (94). An analogous situation has been found to apply to pea vicilin: the 47 kDal variant is processed by a single endoproteolytic cleavage, while the 50 kDal species is not (268). The proteins are homologous except at this region, in which a number of cDNA base changes occur, and two small deletions are quite close. Pea legumin also appears to be processed by a single cleavage (267), thus not having a linker. Soybean glycinin has a 4 - residue linker (272). The sequences surrounding these processing sites are shown in fig 3J-11.

The common factor which emerges from these is the asparagine residue at the C-terminus of the linkers. This may have some bearing on the asparagine residue at the end of the castor bean lectin signal sequence, if the signal peptidase is related to the processing enzyme.

<u>Protein</u>	<u>Processing sites</u>	<u>Ref</u>
Napin	...SCGG / PMW....DMENPQG / PQQ...	95
Pea seed lectin	...TYPN / SLEEN / VISYTL...	94
Favlin	...LYPN / ..... / LTGYTL	See 94
Lentil lectin	....QNG / ..... / VISYT	See 94
Pea vicilin	...CKEN / DKEEEQ....	268
Pea legumin	...QCIN / GLEETV...	267
Soybean glycinin	...RQSK / RSRN / GIDET....	272
Castor bean lectins	...SSQF / SILL...PNEF / ADVCM....	This work

Fig 3J-11: Plant heterodimer precursor processing sites

The exact site of cleavage at the N-terminus of the pea seed lectin is unclear.

The sites for favin and lentil lectin are possible sites, estimated by analogy with pea seed lectin.

The mammalian protein insulin is also synthesised as a heterodimer precursor, and has a linker (or connecting peptide) of 33 residues, accounting for more than one third of the precursor (96). Both ends of the linker have dibasic sites, which is not observed in the plant linkers. Haptoglobin is also synthesised as a precursor, and in this case has a single arginine residue as its linker, this not being part of a dibasic site (274).

The enzymes responsible for processing heterodimer precursors have as yet received little attention. However, the castor bean lectin processing enzyme has been located within the protein bodies, and has been shown to cleave the lectin precursors to products of the expected sizes, with a pH optimum of 4: this is consistent with processing taking place in protein bodies (S Harley and JM Lord, personal communication).

### 3K Analysis of nucleotide sequences

#### 3K1 Comparison of preproricin and preproagglutinin sequences

Both cDNA sequences are shown in fig 3K-1; the agglutinin sequence is above the ricin sequence. The agglutinin deduced amino acid sequence is presented above its cDNA sequence, while only those residues which differ are shown for ricin, below the sequence.

Throughout the overlapping part of the sequences there are 114 base differences, giving identity at 93.25 % of positions. The B chains diverge more than do the A chains, having homologies of 90.5 % and 95.4 % respectively. It is curious that the protein sequences appear to be less conserved than the cDNA sequences, the B and A chains having homologies of 84.4 % and 93.3 % respectively. One would expect the DNA sequences to diverge more, since many base changes do not affect the coding properties. However, of 96 codons containing nucleotide alterations, only 37 fail to alter the amino acid encoded. Similarly, it is unexpected that the 5' and 3' non-coding regions are identical - in the absence of selective pressure acting upon them, they should change more than the coding sequence.

The absence of one codon in the agglutinin A chain as compared to the ricin A chain might possibly be a cloning artefact, though similar deletions have previously been observed, for example, in cDNA sequences encoding pea vicilin (268), in which a single codon deletion and a three codon deletion occur close together.

The majority of base changes in the present sequences are transitions (64.5 %); 35.2 % of these are silent, whereas 28.2 % of the transversions are silent. This might be expected since doublet codons have either purines or pyrimidines in the third position. Nonetheless, the high frequency of A-G transitions (43.6 % of all base changes) is curious.

Fig 3K-1: Preproagglutinin and preproricin sequences

Figure appears on the next page.

The upper nucleotide sequence is that of the preproagglutinin cDNA, with its deduced amino acid sequence above it. The lower nucleotide sequence is the preproricin cDNA; only when its amino acids differ from those of preproagglutinin are they shown, below the sequence.

The first amino acid in each row is numbered at left, and the last nucleotide in each row is numbered at right (numbers correspond to positions in the preproagglutinin sequence).

The ends of various parts of the sequences are indicated by arrows; S = signal sequence, A = A chain, L = linker peptide, and B = B chain.



3K2 Base composition, dinucleotide frequencies and codon usage

The base composition of the preproricin sequence, along with those of several other plant sequences, was determined with a simple computer programme. The results are tabulated in fig 3K-2, in which the preproricin data are enclosed in brackets.

Note that the data for general plant sequences are uncorrected: a small number of errors will have no significant effect on a sample of this size.

The non-coding regions of plant sequences appear to be A+T rich, and the preproricin sequence accords with this conclusion. However, plant sequences in general have almost no preference for A+T or for G+C in the coding region, whereas the preproricin sequence is somewhat A+T rich. Although the significance of base composition is unclear, it is of note that data derived from a catalogue of animal sequences indicates that these have an A+T composition of 45.2 % in the coding regions, lower than the plant sequences.

The dinucleotide frequencies for the preproricin sequence are shown in fig 3K-3. The nucleotide pair CG is grossly under-represented, as has been noted for eukaryotes in general (288,289) and in plants (281,290,291). There are no entirely satisfactory explanations for selective dinucleotide usage, though it has been suggested that it may be associated with constraints on chromatin structure (see ref 289 for a discussion of dinucleotides).

Codon usage in the preproricin sequence is shown in fig 3K-4. As is common in eukaryotic sequences (277), codons involving the dinucleotide CG are relatively infrequent, as would be expected from the rarity of this pair. Thus, preproricin uses 18  $AC_G^A$  codons for arginine, and an equal number of  $CGN$  codons.

<u>Base</u>	<u>5' NCR</u>	<u>CR</u>	<u>3' NCR</u>	<u>Total</u>
A	195 (12)	4945 (494)	987 (53)	6127 (559)
T	127 ( 7)	4232 (486)	1105 (51)	5464 (544)
C	149 ( 3)	4702 (347)	501 (24)	5352 (374)
G	66 ( 8)	4164 (368)	607 (31)	4837 (407)
A+T (%)	60.0 (63.3)	50.9 (57.8)	65.4 (65.4)	53.2 (58.5)
G+C (%)	40.0 (36.7)	49.1 (42.2)	34.6 (34.6)	46.8 (41.5)
No of seq:	13	21	22	22

Fig 3K-2: Base composition

Data were collected from the number of sequences indicated; data for the preprolactin cDNA sequence are presented in brackets.

Several of the published sequences lacked 5' non-coding regions.

References: 95, 262, 264-269, 276, 278-287.

	A	T	C	G
A	140	152	99	102
T	118	136	93	138
C	134	98	75	40
G	101	99	80	88

Fig 3K3: Dinucleotide frequencies

The first base is in the column at left, and the second base is indicated at the top. The analysis is of the preproricin sequence, including non-coding regions.

	U		C		A		G	
U	U	12 Phe	U	11 Ser	U	15 Tyr	U	6 Cys
	C	9 Phe	C	5 Ser	C	9 Tyr	C	6 Cys
	A	11 Leu	A	7 Ser	A	0 ***	A	0 ***
	G	9 Leu	G	2 Ser	G	0 ***	G	12 Trp
C	U	16 Leu	U	11 Pro	U	5 His	U	6 Arg
	C	5 Leu	C	3 Pro	C	1 His	C	3 Arg
	A	4 Leu	A	14 Pro	A	22 Gln	A	6 Arg
	G	5 Leu	G	2 Pro	G	7 Gln	G	3 Arg
A	U	14 Ile	U	14 Thr	U	28 Asn	U	14 Ser
	C	12 Ile	C	9 Thr	C	14 Asn	C	5 Ser
	A	16 Ile	A	15 Thr	A	4 Lys	A	14 Arg
	G	7 Met	G	2 Thr	G	5 Lys	G	4 Arg
G	U	14 Val	U	12 Ala	U	23 Asp	U	16 Gly
	C	3 Val	C	6 Ala	C	4 Asp	C	3 Gly
	A	4 Val	A	19 Ala	A	10 Glu	A	14 Gly
	G	14 Val	G	3 Ala	G	10 Glu	G	6 Gly

Fig 3K-4: Codon usage.

First base is at left, second at top, and third in centre.

Data are for preproricin sequence, excluding the stop codon.

There are 153 amino acids in the sequence which can have codons ending with CG, but only 12 such codons are used. This accounts for 30 CG dinucleotides altogether, leaving the remaining 10 to be distributed between codons rather than within codons.

Grantham (277) points out that mammalian mRNAs contain several times as much degenerate G than A, with the result that the sequences prefer the codons CUC and CUG for leucine, GCC for alanine, AAG for lysine and CAG for glutamine. The preproricin sequence contains more than twice as much degenerate A as G (T being the most frequent), the result being that these codon preferences are not observed here. This has previously been noted for a number of plant mRNAs (291). The avoidance by plant mRNAs of the UUA leucine codon (291) is not observed here.

In spite of the mammalian preference for degenerate G, Grantham (277) notes that the more abundant mRNAs tend to prefer C and U in the third position. The preproricin mRNA produces an abundant protein, and has 57.5 % of the degenerate third bases being C or U, and may thus fit the hypothesis.

### 3K3 The 5' non-coding region

Thirty bases of 5' non-coding region are represented in both the preproricin and preproagglutinin sequences. These are identical in both sequences, and both begin at exactly the same nucleotide. This probably indicates continuation of the homology to the very 5' terminus of the mRNAs, since the hairpin loops of the cDNA produced by the first strand synthesis apparently end in the same position (222).

Primer extension (see section 3L) indicates that some 80 bases are missing from the clones; only one band was seen, implying that the two mRNAs are of the same length (assuming a similar abundance for both).

In the following discussion bases are numbered from the first base of the codon for the first residue of the mature A chain.

The first AUG codon occurs at position -76, and is followed by a short reading frame of 23 amino acids (including the initiator methionine), which ends at the termination codon TAA at -7. A second AUG is found at position -72, which opens up a reading frame which goes on to include the whole coding sequence of the A and B chain precursor. Fig 3K-5 shows a Maxam & Gilbert sequencing gel from which this part of the sequence was determined. One of the pUC8 subclones of pRCL6 was cleaved with *Ava*I and *Hind*III and the insert was purified on an agarose gel. The *Ava*I end was selectively labelled and the fragment was sequenced. The region of the gel showing the two AUG codons is marked, as is the part containing the beginning of the mature A chain sequence. It is clear from this gel that it is indeed the second AUG which opens the correct reading frame. It is remotely possible that this region became altered during subcloning into pUC8, but the same sequences were obtained from both of the original clones (pRCL6 and pRCL52) by labelling at the *Bam*HI site located at position -42, that is, within the signal sequence. It is also possible that these



formed during cDNA synthesis: it is known that the 5' ends of cDNAs are less faithfully synthesised than are the other regions (285). However, it is unlikely that the same artefacts would be obtained in two different clones, which were annealed and transformed at different times.

Since initiation at a second AUG is unusual, the evidence favouring initiation at the first will be briefly reviewed, and a small survey of plant mRNAs will be presented.

The currently favoured model of translation initiation is the scanning model proposed by Kozak (292), in which the 40 S ribosomal subunit is believed to bind to the 5' end of a mRNA, and then to migrate downstream until it encounters an AUG codon. At this point, migration stops, presumably because of an interaction between the initiator codon and the met-tRNA<sub>1</sub> which is already bound to the 40 S subunit. Peptide bond formation then ensues, and continues until a stop codon is reached. The model was inspired by the observation that the vast majority of mRNAs then sequenced (mostly from animal sources) had no AUG codons in the 5' untranslated region. The few exceptions known at that time were explained away with a variety of arguments. For example, one upstream AUG was immediately adjacent to a 5' cap structure, while another was involved in a stable secondary structural formation and, it was suggested, unavailable to the 40 S subunit. Presumably in this case the 40 S subunit was supposed to migrate past the bottom of the stem-and-loop structure.

Under certain abnormal conditions (for example in the presence of edeine, or in low Mg<sup>2+</sup> concentrations), it was found that the 40 S subunit could migrate right through an initiator AUG, and on this basis the model was modified (293). It was now proposed that

if the upstream AUGs were 'weak' initiators, then the 40 S subunit was able to ignore them, perhaps initiating with decreased efficiency at subsequent initiator codons. In support of this, one set of experiments involving the introduction of extra AUGs in the 5' non-coding region of a Herpes simplex thymidine kinase gene, cloned in a retrovirus vector, indicated that the more upstream AUGs there were, then the less efficiently was translation initiated at the correct codon (294). Similar experiments in a different system (295) found that if the upstream AUG was an efficient one, then reinitiation at a subsequent AUG would only occur if the reading frame already started were terminated before the ribosome arrived at the correct reading frame.

Sequencing of three procollagen genes (296) indicated that in each one there were two or more upstream AUGs - all encoded short reading frames of the type described above. However, the functional AUG was in each mRNA involved in a stem-and-loop structure, apparently invalidating the supposition that AUG codons could be 'hidden' from the 40 S subunit by secondary structure. This work also suggested that the function of the short peptides initiated by the upstream AUGs was in the regulation of translation: it is known that N-terminal oligopeptides of the gene product (preprocollagen) inhibit translation of the mRNA species concerned (296). Also noteworthy is hamster 3-hydroxy-3-methylglutaryl-coA reductase cDNA, which contains an AUG codon 160 b upstream of the correct initiation site. Its reading frame overlaps the correct frame, and it has a 'better' initiation environment than the correct AUG (297).

The 'quality' of an AUG has been investigated by searching for similarities in sequences flanking initiator AUG codons in many mRNAs. The initial study (298) found a consensus around the AUG of  
A  
GXXAUGG(X - any nucleotide). The significance of this result was

supported by the finding that the presence of a purine in positions -3 and +4 enhanced the binding of AUG-containing oligonucleotides to wheatgerm ribosomes. More rigorously, a variety of preproinsulin cDNA - containing plasmids were constructed with different nucleotides in the -3 and +4 positions (299). At -3, the base A gave the most efficient initiation, while G was better than C or T. At the +4 position, G was better than A.

A later survey (300) of 211 mRNAs (of which only 8 were of plant origin) enlarged the consensus sequence to  $CC^A_GCAUUG$ . Only six of the sequences had a pyrimidine at -3, and  $1\frac{1}{2}$  of these were from plants (the ' $\frac{1}{2}$ ' is a result of a sequence which Kozak treats as one sequence, but is in fact two mRNAs, one of which has A, while the other has C). As far as the first four bases of this consensus are concerned, more than half the mRNAs had three or four nucleotides in common with the consensus, and only 10 were perfect matches.

On the basis of 20 sequences of plant origin, Messing (301) suggested that the plant initiation environment was  $C^GAAAXXAUGG$ , where X is any base. Note that of the 16 sequences for which this region is given completely, 8 are zein variants and 5 are soybean leghaemoglobin variants.

A number of plant translation initiation sequences are given in fig 3K-6 and in fig 3K-7 these are used to determine another consensus sequence:  $ACAA^A_C AUG G$ . The validity of including several closely related sequences is dubious - ie the zeins and the soybean leghaemoglobins - since these are likely to be conserved for reasons other than singularity of function. In fig 3K-8 the sequences are relisted, omitting all members of these two families, and two sequences whose functional AUGs are

Fig 3K-6: Plant 5' non-coding regions, AUG environments and upstream AUGs  
For explanation see notes following table.

Source	5' NCR	Environment	Comments	Ref
Rapeseed napin	Incomplete	ACAAGA AUG G	Only 9 b of NCR	95
Phaseolin	Complete	UCUACU AUG A	No AUGs	258
Soybean seed lectin	Complete	AAAGCA AUG A	No AUGs	262
Maize zeins: Z22	Complete	ACACCA AUG A	No AUG's	276
ZE19	Complete	CCAACA AUG A	No AUGs	302
pML1	Complete	ACAACA AUG A	No AUGs	55
Z4, ZG7, 124, 31	Complete	CCAA <sup>U</sup> <sub>A</sub> AUG G	ZG7 and 124 have 1 AUG	303
A30	Incomplete	CUCCUU AUG C	1 AUG in 30 b	265
Pea vicilin	Incomplete	CGGUUA AUG U	Only 8 b of NCR	268
Thaumatin	Incomplete	ACAUCA AUG G	No AUGs in 31 b	264
Soybean actin	Complete	UAAAAG AUG G	?Start site; 2-3 AUGs	281
Barley pre-amylase	Incomplete	CGCCCC AUG G	No AUGs in 96 b	283
Phaseolus lectin	Incomplete	AUGAUC AUG G	See note (1) below	266
Soybean pre-SSU	Complete	AAGAAA AUG G	No AUGs	304
Pea legumin	Complete	CUGUUC AUG G	No AUGs	267
Wheat gliadin	Complete	UCCACC AUG A	No AUGs	269

Continued next page....

Fig 3K-6 (continued)

Source	5' NCR	Environment	Comments	Ref
Octopine T-DNA transcript 7	Complete	CACAUC AUG A	No AUGs	287
Soybean Lb	Complete	AGAAAU AUG G	No AUGs	280
Lba	Complete	AGAAAU AUG G	No AUGs	278
Lbc	Complete	AGAAAU AUG G	?AUGs ?Start site	278
Lbc2	Complete	AGAAAU AUG G	No AUGs	282
Lemna gibba SSU	Incomplete	AGAUAU AUG AUG	See note (2) below	285
Parsley chalcone synthase	Incomplete	AAAAAA AUG G	Only 15 b of NCR	284

For explanation of table, and for notes, see next page....

Abbreviations

SSU = small subunit of ribulose biphosphate carboxylase/oxygenase

Lb = leghaemoglobin

NCR = non-coding region

Fig 3K-6: Explanation and notes

In the column labelled 5' NCR, complete indicates that the data are derived from genomic clones; incomplete implies cDNA clones, in which case AUGs not shown in reports could exist upstream (as also applies to the sequences for the castor bean lectins).

If the 5' non-coding region is very short, then its length is found in the comments column. The presence of any upstream AUGs is also noted here.

Note 1

The AUG codon for Phaseolus lectin is predicted on the basis of opening the correct reading frame and having the best surrounding sequence. The upstream AUG shown in the table has a less favourable environment, and there are two more AUGs out of phase:

AUGAAUGCAUGAUGGCU  
1 2 3 4

AUG no 4 is proposed to be the functional initiator codon.

Note 2

This is the first AUG in the cDNA clones. However, it is suspected that the sequence at the 5' end of the cDNA is in error, because of cloning artefacts (285). Genomic clones indicate that there is another AUG 12 b upstream, which may be the actual initiator. The N-terminal sequence of the protein is unknown.

Fig 3K-7: Plant initiation environments: all sequences

	-6	-5	-4	-3	-2	-1	AUG	+4	
A	13	5	15	15	8	11		8	
U	3	3	1	4	6	7		1	
C	8	10	5	3	9	5		1	
G	0	6	3	2	1	1		14	

Con:    A        C        A        A        C  
         A        A        A        A        A        G

Data from fig 3K-6.

The Zein group comprising clones Z4, ZG7, ZG124 and ZG31 has been counted twice, once with U at -2 and once with C (See fig 3K3-2).

Con = consensus sequence.

Fig 3K-8: Plant initiation environments: Selected sequences

Source	-6	-5	-4	-3	-2	-1	+4	Ref
Napin	A	C	A	C	G	A	G	95
Phaseolin	U	C	U	A	C	U	A	258
Soybean lectin	A	A	A	G	C	A	A	262
Pea vicilin	C	C	G	U	U	A	U	268
Thamatin	A	C	A	U	C	A	G	264
Soybean actin	U	A	A	A	A	G	G	281
Barley amylase	C	G	C	G	C	C	G	283
Soybean SSU	A	A	G	A	A	A	G	304
Pea legumin	C	U	C	U	U	C	G	267
Wheat gliadin	U	C	C	A	C	C	A	269
Parsley CS	A	A	A	A	A	A	G	284
Octopine T-DNA Tr7	C	A	C	A	U	C	A	287
<b>Totals</b>	<b>A:</b>	5	5	5	6	3	6	4
	<b>U:</b>	3	1	1	3	3	1	1
	<b>C:</b>	4	5	4	1	5	4	0
	<b>G:</b>	0	1	2	2	1	1	7
<b>Con:</b>	A	A	A	A	C	A	G	
	C	C	C			C		
	U							

Abbreviations:

SSU = small subunit of ribulose biphosphate carboxylase/oxygenase

CS = chalcone synthase

Tr7 = Transcript 7

con = consensus sequence

uncertain. The consensus above becomes less strong: all that is apparent is a preference for A or C in positions -1 to -6, and a complete absence of G from position -6.

The validity of consensus sequences derived in this way is limited, first by the small number of sequences examined, and also by the dissimilarities between the different genes and their requirements for regulation. That is, if the sequences surrounding the AUG are important, then it is more meaningful to compare only those sequences whose modes of regulation are known to be similar. Consensus sequences derived here, and by other authors, probably include sequences which are so different that they should not directly be compared without some other organising factor.

The castor bean lectin initiator AUG environments are shown below along with various consensus sequences:

First castor bean lectin AUG:	UGUAAU AUG G
Second " " " " ;	AUAUGG AUG U
Kozak (300):	CC <sup>A</sup> CC AUG G
Messing (301):	C <sup>G</sup> AAXX AUG G
Here:	AAA <sup>G</sup> ACA AUG G

The first castor bean lectin AUG is a better fit with the Kozak and Messing consensus sequences than is the second; the two are equal when compared with the present consensus. Again, the inadequacy of consensus sequences must be emphasised: although the first AUG may have a 'better' environment, castor bean ribosomes may not agree! That is, deviation from the consensus sequences may imply that castor bean ribosomes have a different specificity than others,

and the details of the difference may reflect some special feature of the regulation of translation of the particular mRNA species examined here.

The position of the initiator AUG may be of greater significance than the sequence surrounding it: although most eukaryotic mRNAs initiate at the first such codon, not all do. For example, fig 3K3-2 shows that three zein mRNAs have upstream AUGs, the soybean actin mRNA probably has two, Phaseolus vulgaris lectin has two or three, and one of the soybean leghaemoglobin sequences may have one (the translation start site is uncertain). The importance of these mRNAs, and of the present sequences, to the scanning model of translation initiation is unclear. It is conceivable that upstream AUGs are themselves involved in regulation of translation. Thus, a ribosome might migrate along a mRNA, stop at an upstream AUG codon, and then wait for some signal to continue, such a signal perhaps being a developmentally regulated gene product.

### 3K4 The 3' non-coding region

The preproricin cDNA sequence contains 159 bases of 3' non-coding sequence, including the termination codon, as determined from clone pRCL17. The preproagglutinin sequence has only 69 3' non-coding bases, and these are identical to their counterparts in the preproricin sequence.

The stop codon is UGA, and is immediately followed by another stop codon, UAG. A survey of termination codons and their environments (305) found that such tandem stop codons are unusual in eukaryotic mRNAs, and may even be selected against (the only two cases found in 74 sequences examined were from related viruses).

Few plant sequences were included in this report, so a number are listed in fig 3K-9. This shows that of the 23 sequences presented, three have tandem stop codons: a wheat gamma-gliadin (21) has UGA-UAA, the pea vicilin 50 kDal precursor has UAA-UGA (268) and a plastocyanin sequence has UAA-UAG (337). The importance of tandem stops is obscure, as in eukaryotes termination at a single such codon is apparently extremely efficient (305). In the 'all eukaryotes' sample, the stop codon UAA is used more frequently than UAG and UGA (which are used roughly equally) (305); in the plant mRNAs, no obvious trend is seen (UAA, UAG and UGA occur 9, 7 and 7 times respectively). The sample is of course very small - accumulation of more plant sequences may indicate the existence of the other eukaryote trend, or a plant preference may become apparent.

Fig 3K-10 shows the frequencies of occurrence of each base in the codons before and after the termination codon; also shown are similar results from the survey already mentioned (305), when significant deviations from randomness were observed.

Fig 3K-9: Plant termination environments

<u>Source</u>	<u>Environment</u>	<u>Ref</u>
Lemna gibba SSU	ACC UAA GCU	285
Soybean seed lectin	AUC UAA AUG	262
Soybean Lb's (two of these)	GCA UAA UUA	278
Thaumatin	GAG UAA GAG	264
Soybean actin	UUC UAA CUU	281
Pea vicilin (50 kDal)	UUU UAA UGA *	268
Octopine T-DNA Tr7	AGC UAA GCU	287
S pratensis plastocyanin	AAC UAA UAG *	337
Pea chlorophyll a/b BP	AAA UAA ACA	338
Wheat gliadin	AAC UGA GAA	269
Wheat gamma-gliadin	UAC UGA UAA *	286
Barley alpha-amylase	AGC UGA AGU	283
Parsley chalcone synthase	CAC UGA AGU	284
Soybean vicilin (Gmcp53.58)	AAA UGA CAA	306
Phaseolin	UAC UGA AUA	258
Castor bean lectins	UUU UGA UAG *	This work
Pea legumin	GCU UAG AUU	307
Napin	UAC UAG AUU	95
Pea seed lectin	GCA UAG UUU	260
Zeins	UUU UAG AUU	265,276,303
Phaseolus lectin	CUC UAG ACU	266
Maize Adhl	AAC UAG AUU	279
<u>Soybean Lb's (two of these)</u>	<u>UUU UAG GAU</u>	<u>280,282</u>

Notes

Several zeins are counted together, as all in this report had the same termination environment.

Sequences marked with an asterisk have tandem termination codons.

Lb = leghaemoglobin; SSU = small subunit of ribulose biphosphate carboxylase/oxygenase.

S pratensis is Silene pratensis; BP = binding protein

**Fig 3K-10: Distribution of bases flanking plant stop codons**

Stop codon	Base	-3	-2	-1	+1	+2	+3
UAA	A	5	3	2	2	2	3
	U	2	3	1	3	3	3
	C	0	2	5	1	3	0
	G	2	1	1	3	1	3
UGA	A	3	5	1	3	4	4
	U	3	1	1	2	1	2
	C	1	0	5	1	0	0
	G	0	1	0	1	2	1
UAG	A	1	2	1	5	1	0
	U	3	3	3	1	5	7
	C	1	2	3	0	1	0
	G	2	0	0	1	0	0
ALL	A	9	10	4	10	7	7
	U	8	7	5	6	9	12
	C	2	4	13	2	4	0
	G	4	2	1	5	3	4

**Significant results from ref 305 (present results in brackets)**

All values in percentages.

Stop codon	-3	-2	-1	+1	+2	+3
UAA	G+C=69(22)			G+A=82(65) all stops		
UGA	A=90(71)			C+U=67(55) all stops		
UAG	C=0(14)		C=68(43)			

**Above** The number of sequences (out of 23) having specified bases in positions before and after the stop codons are tabulated by stop codon.

**Below** Significant results for specified positions from ref 305 are reproduced, with present results for those positions in brackets.

Results obtained here at those positions are given in brackets. Although the sample of plant sequences is small, they do appear to prefer the preceding codon to be of the form  $\overset{A}{U}AC$ , and the subsequent codon to be  $A\overset{A}{U}$ . Further sequences may or may not bear this out.

The majority of 3' non-coding regions of eukaryotic mRNAs contain the hexanucleotide sequence AAUAAA, usually 10 - 40 bases from the poly(A) addition site. Polyadenylation occurs in the nucleus, and probably involves site-specific cleavage of the pre-mRNA followed by addition of A residues to the exposed 3' terminus (308). The cleavage site may be defined by hybridisation of the pre-mRNA with snRNA U4, through the AAUAAA signal and other sequences (309). Although most animal mRNAs contain only one AAUAAA sequence, examples are known with multiple signals (297, 310-312).

Plant mRNAs are distinctly different in that multiple polyadenylation signals appear to be the rule, rather than the exception (291,301). Poly(A) addition rarely follows the first AAUAAA signal, though in one case, an octopine T-DNA gene (transcript 7, ref 287), polyadenylation does occur after the first signal in the pre-mRNA, not after the second signal.

The castor bean mRNAs are quite typical of plant mRNAs with respect to polyadenylation signals, with one 59 bases from the last coding base, and another 38 bases from the polyadenylation site. The first is a typical AAUAAA sequence, and the second is a variant, AAUAAG: this pattern has previously been noted (301).

Berget (309) analysed 3' non-coding sequences in more detail, and found that most animal genes contain variations on the consensus sequence CAYUG (where Y = a pyrimidine), either between the polyadenylation signal and the polyadenylation site, or a short distance downstream of the polyadenylation site. In many cases, a 4 out of 5, or only a 3 out of 5 fit was obtained. She suggested that the AAUAAA signal provides primary recognition for pre-mRNA cleavage, while the CAYUG signal directs the precise cleavage point. Benoist et al (313) found that several animal sequences contain variations of the model sequence UUUUCACUGC (note that this contains CAYUG), the polyadenylation site often immediately following this. Lycett et al (291) noted that pea legumin mRNA contains AUUUCAGUGC, a variation on the Benoist sequence, the polyadenylation site following after a further two bases.

Since arguments such as these may explain the selection of the appropriate polyadenylation signal in plant mRNAs, a search of a number of such sequences was undertaken. The computer programme RE SITES (See Appendix 1) was used to find sequences which matched, or gave a 4 out of 5 fit with, CAYUG. Potential polyadenylation sequences were located by visual examination of published data.

The sequences examined were from references 95,258,262, 264-6,269,276,278,283-7,303 and 306-7.

First, the frequencies of the numbers of AAUAAA sequences are tabulated:

<u>No of signals</u>	<u>No of sequences</u>
0	0
1	3
2	8
3	4
4	4
5	0
6	1

Thus, the multiplicity of polyadenylation signals (AAUAAA and variants) is amply confirmed.

Although three sequences contain only one such signal, all of these may be special cases. First, the barley alpha-amylase mRNA (283) contains an additional GAUUC upstream of the functional signal - this may be a variant of AAUAAA. Second, the octopine T-DNA transcript 7 gene (287) contains a second AAUAAA signal downstream of the cleavage site - the significance of this is unclear, and this appears to be a completely non-functional signal. Finally, the zein clone pZ22.3 (276) may conceivably contain another signal downstream: genomic clones were not given in this case, but another report (303), in which cDNA and genomic sequences were compared, indicated that zein genes can use either of two signals. That is, although the second signal may be the dominant one, earlier ones can also result in polyadenylation.

Of the 20 sequences examined, 8 contained perfect matches or 4 out of 5 matches with the CAYUG sequence, between the AAUAAA signal and the polyadenylation site. If this sequence is as important in plants as it appears to be in animals, it may well be that in the remaining 12 sequences, its location is downstream of the polyadenylation site. Examination of the polyadenylation signals (that is, AAUAAA signals) in the 8 mRNAs mentioned reveals

that the last signal in 5 cases is a variant, rather than a perfect match: of the remaining 12 sequences, only three have AAUAAA variants, as opposed to perfect matches.

The distance between the polyadenylation signal and the CAYUG variant in the 8 mRNAs is variable, ranging from 0 to 20 bases. In two cases, pea legumin (307) and the castor bean lectins (present work) the CAYUG sequence can be extended to variants of the UUUUCACUGC sequence of Benoist et al (313):

Legumin: AAUAAAAGGUAAAAUUUCAGUGCUC - poly(A)  
CB Lectins: AAUAAGAGCACAACUAUUGUCUUGUGCAUUCUAAAAUUU - poly(A)

In the twelve sequences lacking the CAYUG variants after the polyadenylation signal, there is no consistent presence or absence of CAYUG variants before the signal, and there is no consistent relationship with upstream AAUAAA signals.

A similar search of six genomic sequences (262,267,269,278,287) revealed that four of these contain AAUAAA sequences downstream of the polyadenylation site. Three of these lack CAYUG variants between the functional AAUAAA signal and the polyadenylation site, and of these three, two had CAYUG variants just downstream of the polyadenylation site, by 15 and 30 bases. In only one case of downstream AAUAAAs was a CAYUG variant close, this being the pea legumin gene (267), which has a CAYUG variant some 20 bases beyond a double AAUAAA.

Note that the pea seed lectin mRNA (260) has an unusually large distance between its polyadenylation signal and its polyadenylation site, 57 bases, and the signal is immediately followed by a CAYUG variant.

Although this sample is severely limited by size, it appears that the mechanism of polyadenylation site selection may well be

similar to that in animals, involving hybridisation to a snRNA species, as most of the plant sequences examined here have features common to many other eukaryotic mRNAs.

In summary, the castor bean lectin mRNAs have typical plant 3' non-coding regions, though the tandem stop codons may turn out to be unusual, though certainly not unique.

### 3L Primer extension

A primer was cloned which corresponded to the region of pRCL6 encoding part of the signal sequence and the beginning of the A chain sequence, as described in section 2J1. The fragment was excised from its plasmid, end-labelled and purified from a gel. After heat denaturation it was annealed with total poly(A)<sup>+</sup> mRNA from maturing castor beans and extended with reverse transcriptase, the products of the reaction being analysed on a 6 % acrylamide, 7 M urea sequencing gel.

Fig 3L-1 shows such a gel, with dideoxy sequencing products as size markers: the number of bases between the unextended primer and the extended product is about 143, indicating that the mRNA has this number of bases 5' to the end of the primer. The end of the primer is 61 bases from the 5' terminus of the cDNA sequences present in the clones, indicating that the mRNA has approximately 80 bases which are not represented in the clones. The number is approximate because extension may be terminated 2 - 3 bases before a 5' cap structure (238).

The castor bean lectin mRNAs thus contain some 195 bases of 5' non-coding sequence. It was originally intended to attempt to sequence this part of the mRNA by inhibition of the primer extension reaction with dideoxynucleotides, but the amount of mRNA available was unfortunately inadequate.

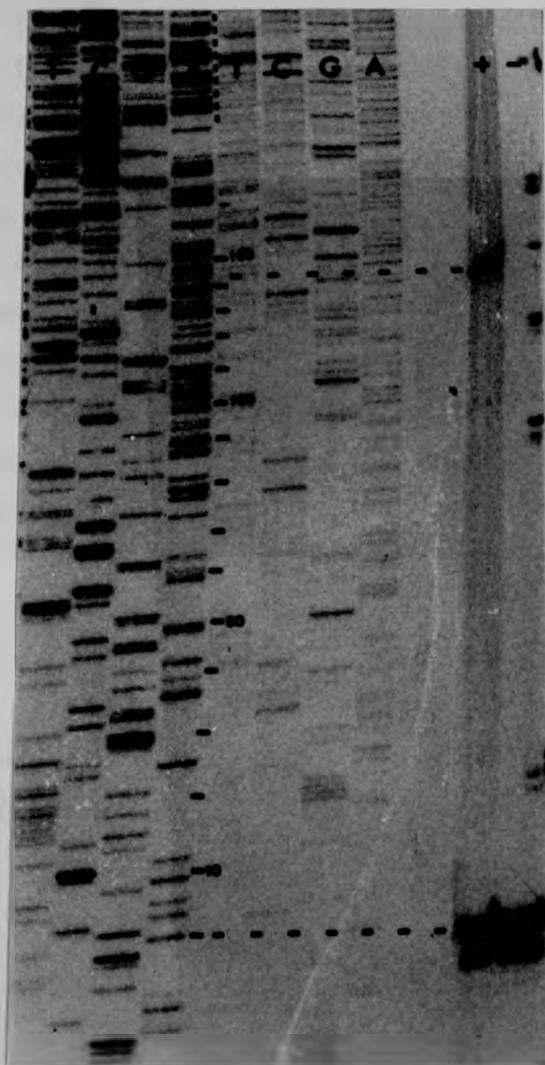


Fig 3L-1: Primer extension.

Bands corresponding to unextended primer and to extension product are marked (dashed lines). Track + contains mRNA, track - is a control. Marks at right are spilled over from adjacent track. Dideoxy sequencing ladders act as size markers.

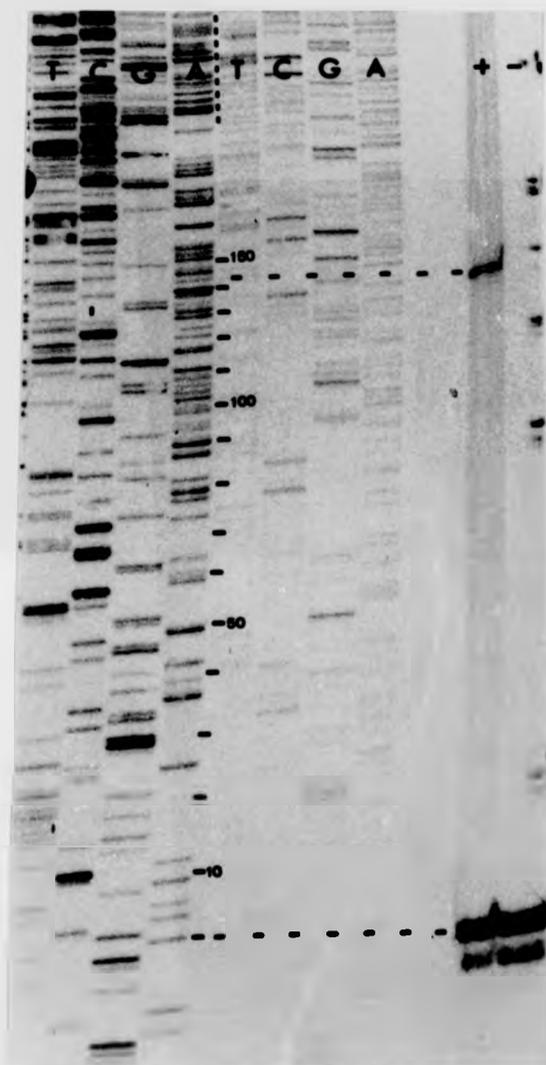


Fig 3L-1: Primer extension.

Bands corresponding to unextended primer and to extension product are marked (dashed lines). Track + contains mRNA, track - is a control. Marks at right are spilled over from adjacent track. Dideoxy sequencing ladders act as size markers.

### 3M Northern blotting

Total poly(A)<sup>+</sup> RNA from maturing castor beans was run on a formaldehyde agarose gel and transferred to nitrocellulose, and hybridised with nick-translated pRCL6. The resulting autoradiograph is shown in fig 3M-1; identical results were obtained with nick-translated insert from the same plasmid, though in this case the markers were not visible (markers are digests of pBR322 DNA, the vector for pRCL6).

Lane 1 contains the mRNA, and three bands are visible, with sizes estimated at 3000, 2000 and 1700 bases. The largest and smallest bands comigrated with bands visible on the ethidium bromide stained gel (not shown) - these correspond to ribosomal RNA species, indicating that the probe has hybridised non-specifically, and that the washing conditions were not sufficiently stringent.

It is of course possible that the remaining band, of 2000 bases, is also an artefact, but its similarity of size with estimates from the sequences, and from the primer extension experiment, suggest that it may indeed represent castor bean mRNA. Unfortunately the experiment could not be repeated due to lack of castor beans.

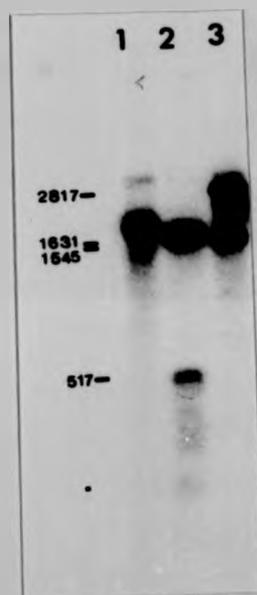


Fig 3M-1: Northern blot of castor bean mRNA

Lane 1: 10  $\mu$ g of castor bean poly(A)<sup>+</sup> mRNA was run on a 1 % formamide-agarose gel, transferred to nitrocellulose and hybridised with nick-translated pRCL6. This also hybridises with pER322, allowing the markers to be visible.

Markers are digests of pER322: lane 2 with HinfI and lane 3 with PstI and PvuII.



Fig 3M-1: Northern blot of castor bean mRNA

Lane 1: 10  $\mu$ g of castor bean poly(A)<sup>+</sup> mRNA was run on a 1 % formamide-agarose gel, transferred to nitrocellulose and hybridised with nick-translated pRCL6. This also hybridises with pBR322, allowing the markers to be visible.

Markers are digests of pBR322: lane 2 with HinfI and lane 3 with PstI and PvuII.

3N Oocyte injections

It was of interest to determine whether the lectin cDNA-containing plasmids could direct the synthesis in Xenopus oocytes of the proteins which they encode. The eight plasmids screened by translation of hybridisation-selected mRNA were injected into oocytes and labelled in the presence of (<sup>35</sup>S)methionine (experiment performed by Professor H Woodland). Expression of a number of sequences in this system has previously been reported (336).

Fig 3N-1 shows that although a large number of proteins were made, there were no polypeptides which reacted with anti-RCA serum. The failure to obtain expression was not further investigated.

In Oocytes  
It was of  
plasmids  
proteins  
translating  
oocytes  
(experiment)  
number of  
Fig 3  
made, there  
The failure

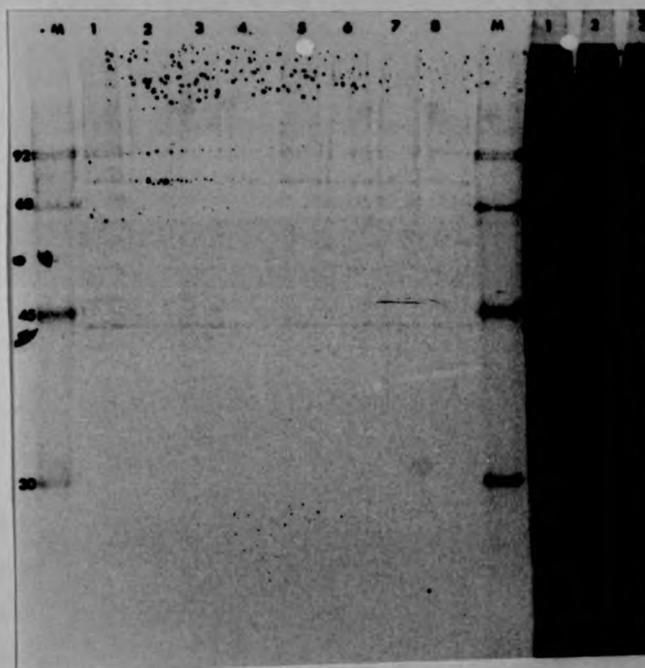


Fig 3N-1: Oocyte injection.

Plasmids were injected into oocytes and incubated with  $(^{35}\text{S})$ methionine. Left side tracks are immunoprecipitated with anti-RCA serum; right side tracks are total protein synthesis. MW markers as fig 3A-3.

Lanes 1 - 8 are, respectively, pRCL 6,15,17,



Fig 3N-1: Oocyte injection.

Plasmids were injected into oocytes and incubated with (<sup>35</sup>S)methionine. Left side tracks are immunoprecipitated with anti-RCA serum; right side tracks are total protein synthesis. MW markers as fig 3A-3.

Lanes 1 - 8 are, respectively, pRCL 6,15,17,18,19,20,21,22.

presence of  
with

d 61.

CHAPTER 4: Final Discussion

At the outset of this work it was desired to determine the nature of the precursor-product relationships of the castor bean lectins. This arose from previous work which indicated that the then putative B chain precursor was far larger than the authentic B chains (44), the requirement for this extra size being unclear. At this time little work had been reported on plant storage protein and lectin precursors, those for legumin (93) and canavalin (314) having been published in 1980: in each of these cases it was shown that the mature heterodimers were assembled post-translationally from single-chain precursors containing both subunits.

During the following three years it has become clear that this class of precursor-product relationship is common in plant storage proteins and lectins; Butterworth and Lord (45) in 1981 obtained evidence from immunological and proteolytic cleavage approaches that castor bean lectins are also synthesised in this manner. Thus, it was shown that the large 'B chain precursor' almost certainly contained polypeptides corresponding to both the A and the B chains of the mature proteins. The present studies were designed to elucidate the structure of this precursor by cDNA cloning and sequencing - an approach already used to confirm these precursors in plants (eg 95, 307).

The results obtained here confirm the protein-based evidence previously reported, and also show that the castor bean lectin precursor has a signal sequence and a linker between the two chains, which must be removed post-translationally, very likely in the protein bodies. In addition, two different sequences were obtained, which have been tentatively identified as ricin and Ricinus communis agglutinin. Although the full significance of

presence of  
with

nd 61.

the differences between the primary structures of these two proteins is somewhat unclear, it is hoped that collaborations with groups using more sophisticated computer modelling techniques (eg ref 335) will help to elucidate this aspect, especially when good X-ray diffraction structures become available.

The existence of two proteins of similar sequences, but which are different in their quaternary structures and detailed functions, is intriguing. It might be supposed that the differences have arisen by genetic drift (315) and by lack of selection against them, but the existence of similar situations in other plants, such as Abrus precatorius (3), may imply evolutionary importance for two related, but different lectins. It is of note that some toxins similar to ricin, such as viscumin, also tend to form dimers (57), dimer formation in this case being dependent on concentration. Perhaps the agglutinins in Ricinus and Abrus have evolved to stabilise such dimeric forms.

In a similar vein, it is interesting that many plants contain toxins which enzymatically inhibit ribosomes, while only in certain species are they associated with lectin chains (37), and that whenever a lectin chain is present, it is specific for terminal galactosyl residues. This may imply that only galactose-bearing receptors can mediate the uptake of the toxins, though this argument suggests the supposition that the molecules evolved for the specific purpose of being toxic. It is also possible that the A chains have some important but as yet unknown function in the cells which synthesise them (for example in the regulation of protein synthesis), and that some plants may have improved the interaction of the A chain - like species with a receptor by linking them with galactose-binding moieties - namely lectin

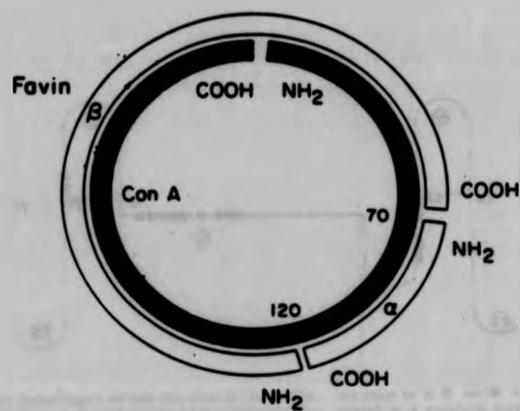
chains, already present and subserving some other purpose.

The existence of gene families in plants encoding variant forms of storage proteins is a widely documented feature (eg 286,316-319), and there are probably over 100 zein genes (55). Such gene families have not been examined in the case of the castor bean lectins - extensive analysis of the present cDNA library with this aim in mind has not been performed. Southern blot analysis of genomic DNA from castor bean tissues is currently being undertaken in this laboratory (JM Lord, personal communication). However, the finding reported here that of eight large lectin-specific clones, all appear to fall into one of two restriction site classes is a preliminary indication that only the two are present: that is, a ricin gene and an agglutinin gene. It is of note that analysis of lectin precursors synthesised in the presence of tunicamycin reveals two bands on SDS/PAGE, very close together. If a wide variety of genes were transcribed and translated, then more bands, or a smear, would be expected (JM Lord, personal communication). Genomic cloning has also been undertaken, to determine the structural features of the lectin genes, and to investigate the mechanisms by which they arose. It will be of interest to locate any short repeats around the genes, which might have been involved in duplication events. Also of interest from the genomic sequences will be the detection of any relationship between exons and functional domains. Thus, if the genes for the two chains arose separately, and then became linked by recombination, one might expect to find introns between their coding sequences. Similarly, an intron might be present between the two halves of the B chain coding sequence. Such relationships between exons and structural and functional domains have been suggested in animal

systems by Go (320,321), though Rashin (322) argued against them, suggesting that structural domains are "stable protein fragments found in biochemical experiments" - that is, artefacts.

Curious evolutionary events have certainly occurred in other plant lectins, notably within the conA homology group. There is considerable homology among lectins from Phaseolus vulgaris, Vicia faba, pea, jackbean and lentil (see ref 323). A detailed comparison of favin and conA revealed a relationship defined by a circular permutation, such that the termini of the two proteins were homologous to the central regions of the other (325). This relationship is indicated in fig 4-1, and a discussion of how it may have arisen is to be found in reference (325).

Complex events may also have occurred in the evolution of the castor bean lectins. Villafranca and Robertus (72) compared the two halves of the ricin B chain, and found that 26 amino acids were conserved - see fig 4-2, which is reproduced from reference 72. A similar analysis of the present ricin sequence is shown in fig 4-3, along with the cDNA sequence. In this amino acid sequence 41 residues are conserved between the two halves of the B chain, or 28.6 % (note that Villafranca and Robertus did not box two of the conserved residues in their sequence; also I have introduced three extra gaps to maximise homology). The great amount of divergence within the cDNA sequences encoding the non-conserved residues (26.3 % homology), and the surprisingly large amount of divergence in the codons for the conserved amino acids (71.7 % homology; differences especially notable in leucine codons), implies that the conserved residues may be important functionally. That is, the DNA sequence appears to have changed nearly as much as it can without changing its coding properties.



**Fig 4-1: Circular homology of conA and favin**

Schematic drawing of the alignment of favin alpha and beta chains (open bar) with conA (solid bar) showing the circular permutation that gave maximum homology between the two sequences.

Reproduced from reference 325.

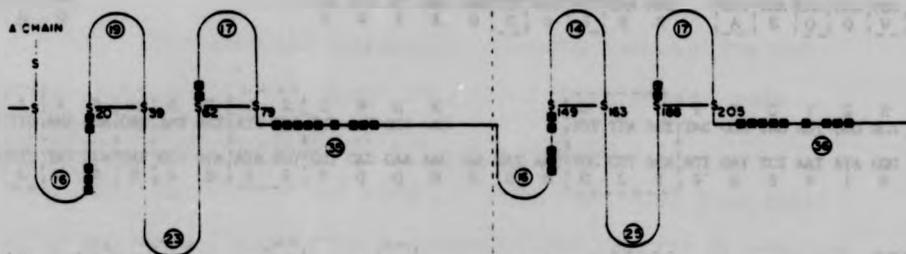


FIG. 2. Structural homologies within the ricin B chain. The linear sequence of the B chain is represented as a line with disulfide bridges forming loops. The dashed vertical line divides the chain into an NH<sub>2</sub>-terminal half (1-132) and a COOH-terminal half (133-260). The circled numbers described the number of residues within each loop. Regions of strong amino acid homology are marked off along

the chain by ☒, ☙, and ☚. A stretch of a given pattern in one half is related to a matching sequence pattern in the second half. The pattern of disulfide bridges, loop size, and sequence homologies suggest the two halves of the B chain are strongly related in a translational sense. Cys 4 of the B chain forms the disulfide bond to the A chain as indicated on the left-hand side of the figure.

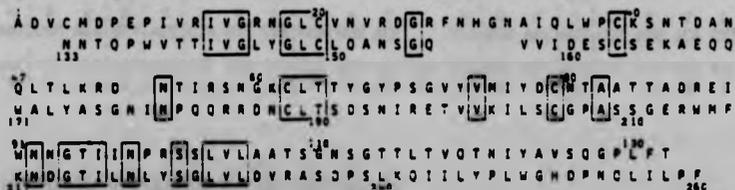


FIG. 3. Amino acid sequence comparisons of the two halves of ricin B chain. The sequence of the NH<sub>2</sub>-terminal half (residues 1-132) is compared with the COOH-terminal half (133-260), aligning the 2 pairs of disulfide bridges in each half. Identical residues are enclosed in boxes. Amino acid single-letter symbols are according to Dayhoff (28).

Fig 4-2: Structural homologies within the ricin B chain.

Reproduced from Villafranca & Robertus (ref 72).

1  
A D V C M D P E P I V R I V G R N G L C V D V R  
GCT GAT GTT TGT ATG GAT OCT GAG CCC ATA GTG GGT ATC GTA GGT OCA AAT GGT CTA TGT GTT GAT GTT AGG  
AAT AAT ACA CAA OCT TTT GTT ACA ACC AIT GTT GGT CTA TAT GGT CTG TCG ITG CAA GCA AAT  
N N T Q P F V T T I V G L Y G L C L Q A N

D G R F H N G N A I Q L W P C K S N T D A N Q L W  
GAT GGA AGA TTC CAG AAC GGA AAC GCA ATA CAG ITG TGG OCA TGG AAG TCT AAT ACA GAT OCA AAT CAG CTC TGG  
AGT OCA CAA GTA TGG ATA GAG GAC TGT ACC AGT GAA AAG OCT GAA CAA CAG TGG  
S C Q V W I E D C S S E K A E Q Q W

50  
T L K R D N T I R S N G K C L T T Y G Y S P  
ACT TTG AAA AGA GAG AAT ACT AIT OCA TCT AAT OCA AAG TGT TTA ACT TAC GGG TAC AGT CCG OGA  
OCT CIT TAT OCA GAT GGT TCA ATA OCT OCT CAG CAA AAC OGA GAT AAT TGC CTT ACA AGT GAT TCT AAT ATA CCG  
A L Y A D G S I R P Q Q N R D N C L T S D S N I R

G V Y V M I Y D C N I A A T D A T R W Q I W D N G  
OGA GTC TAT GTG ATG ATC TAT GAT TGC AAT ACT GCT GCA ACT GAT GCC ACC GGC TGG CAA ATA TGG GAT AAT GCA  
CAA ACA GTT GTT AAG ATC CTC TCT TGT GGC CCT GCA TCC TCT GGC CAA OGA TGG ATG TTC AAG AAT GAT GGA  
E T V V K I L S C G P A S S G Q R W M F K N D G

100  
T I I N P R S S L V L A A T S G N S G T T L T V  
ACC ATC ATA AAT CCC AGA TCT AGT CTA GTT TTA OCA GGC ACA TCA GGG AAC AGT GGT ACC ACA CTT ACC GTG  
ACC AIT TTA AAT TTG TAT AGT GCA TTG GTG TTA GAT GTG ACC OGA TGG GAT CCG ACC CTT AAA CAA ATC ATT CTT  
T I L N L Y S G L V L D V R R S D P S L K Q I I L

Q T N I Y A V S Q G W L P T  
CAA ACC AAC ATT TAT GCC GTT AGT CAA GGT TGG CTT OCT ACT  
TAC CCT CTC CAT GGT GAC OCA AAC CAA ATA TGG TTA CCA TTA TTT  
Y P L H G D P N Q I W L P L F

Fig 4-3

Comparison of the two halves of the ricin B chain, and their cDNA sequences. Upper line is the N-terminal 134 residues, lower line is residues 135 to 262. Identical amino acids are enclosed in boxes.

That the changes in the sequences encoding the two halves of the B chain are subject to the constraint that the general properties are not too greatly altered is indicated by the conservation of the hydropathy profile: fig 4-4 shows the hydropathy plot of the ricin B chain, with the two halves one above the other. It is clear that the profile is not greatly altered, even though the amino acids are grossly different.

A similar analysis of the agglutinin B chain is shown in fig 4-5. This also has 41 residues conserved between the two halves of the B chain, though only 37 of these are the same residues as are conserved in the ricin B chain.

Note that overall, the two B chains differ at 15.65 % of their amino acids: it would thus be expected that, of the 82 residues of the ricin B chain which are involved in the intrachain homology, 12.8 of these would be different in the agglutinin B chain. The observed number is 6 - which may support the suggestion that these are indeed important residues. The differing sugar specificities of the ricin and agglutinin B chains may be reflected in these differences, though the other differences may in fact be responsible. Similarly, which of the differences between the two B chains are responsible for the different quaternary structures is unclear.

That the intrachain homology formed much earlier than the duplication of the whole precursor is indicated by the much greater conservation of nucleotides between the two B chains than between the two halves of each B chain. Of the 74 residues participating in the ricin intrachain homology which are identical in the agglutinin B chain, all but 9 have identical codons: much less mutation has occurred between the two types of B chain than between the two halves of each type.

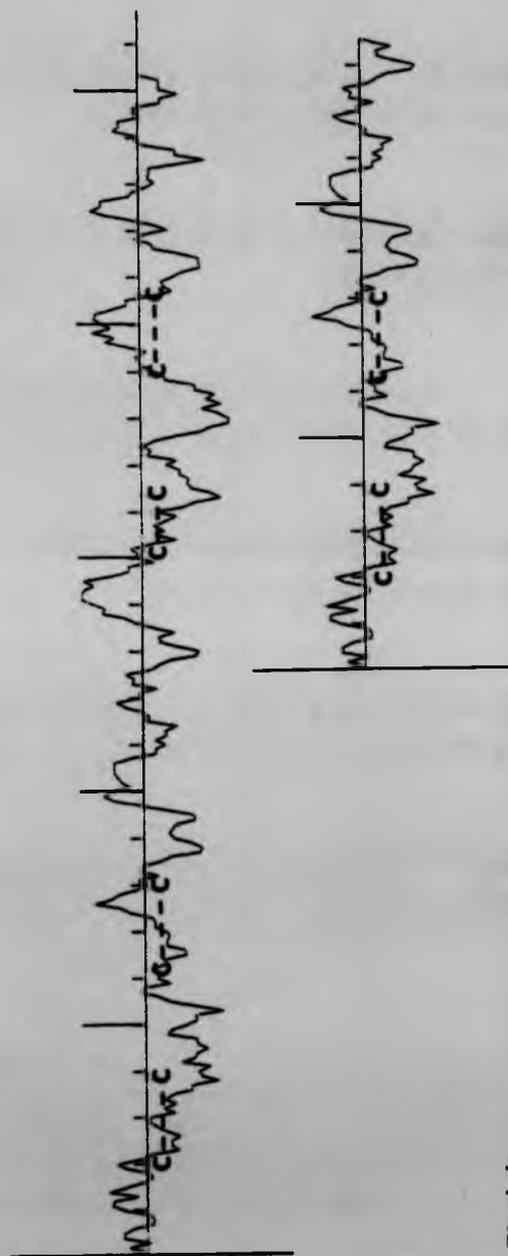


Fig 4-4

Hydropathy plots of deduced ricin B chain sequences.

Above: Whole B chain.

Below: Residues 1 - 134, aligned to emphasise similarity of peaks.

Positions of disulphide bridges are indicated by dotted lines joining Cys residues.

A D V C M D P E P I V R I V G R N G L C V D V T  
 GCT GAT GTT TGT ATG GAT OCT CAG OCC ATA GTG GGT ATC GTA GGT GGA AAT GGT CIA TGT GTT GAT GTT ACA  
 AAT AAT ACA CAA OCT TTT GTG ACA ACC ATT GTT GGG CIA TAT GGC ATG TGC TTG CAA GCA AAT  
 N N T Q P F V T T I V G L Y G H C L Q A N

G E E F F D G N P I Q L W P C K S N T D W N Q L  
 GGT GAA GAA TTC TTC GAT GGA AAC CCA ATA CAA TTG TGG CCT TGC AAA TCT AAT ACA GAT TGG AAT CAG TTA  
 AGT GCA AAA GTA TGG TTA CAG GAC TCT ACC AGT GAA AAG GCT CAA CAA CAA  
 S G K V W L E D C T S E K A E Q Q

W T L R K D S T I R S N G K C L T I S K S S  
 TGG ACT TTG ACA AAA GAT AGT ACT ATT GCA TCT AAT GGC AAG TGT TTG ACC AAT TCC AAG TCC AGT  
 TGG OCT GTT TAT GCA GAT GGT TCA ATA GGT OCT CAG CAA AAC GGC GAT AAT TGC CTT ACA ACT GAT GCT AAT ATA  
 W A L Y A D G S I R P Q Q N R D N C L T T D A N I

P R Q Q V V I Y N C S T A T V G A T R W Q I W D  
 CCA AGA CAG CAG GTG GTG ATA TAT AAT TGC AGT ACC GGT ACA GTT GGT GGC ACC GGT TGG CAA ATA TGG GAC  
 AAA GGA ACA GTT GTC AAG ATC CTC TCT TGT GGC OCT GCA TCC TCT GGC CAA GCA TGG ATG TTC AAG AAT  
 K G T V V K I L S C G P A S S G Q R W H F K N

N R T I I N P R S G L V L A A T S G N S G T K L  
 AAT GGA ACC ATC ATA AAT GGC AGA TCT GGT CIA GTT TTG GCA GGC ACA TCA GGC AAC AGT GGT ACC AAA CTT  
 GAT GGA ACC AAT TTA AAT TTG TAT AAT GGA TTG GTG TTA GAT GTG AGG GCA TGG GAT GGC ACC CTT AAA CAA ATC  
 D G T I L N L Y N G L V L D V R R S D P S L K Q I

T V Q T N I Y A V S Q C W L P T  
 ACA GTG CAA ACC AAC ATT TAT GCC GTT AGT CAA GGT TGG CTT OCT ACT  
 ATT GTT CAC OCT TTC CAT GGA AAC CIA AAC CAA ATA TGG TTA CCA TTA TTT  
 I V H P F H G N L N Q I W L P L F

Fig 4-5

Comparison of the two halves of the agglutinin B chain, and their cDNA sequences. Upper line is the N-terminal 134 residues, and the lower line the remaining residues, 135 to 262. Identical amino acids are enclosed in boxes.

Thus the results obtained here strengthen the argument of Villafranca and Robertus. A possible series of events in the evolution of the castor bean lectins is as follows: first, there would be a primitive A chain gene and a primitive 'half-B' chain gene. The 'half-B' gene would then duplicate to form the complete B chain gene. At some stage, probably after the formation of the whole B chain gene, the A and B chain genes would form a single expressing unit, which later would become duplicated. Subsequent mutations, under whatever influence of selection applied, would generate the two lectin genes present today.

It is of interest that, during the sequencing of the C-terminus of the A chain and the linker sequence, a gel reading error showed that, given a deletion of a single base at the C-terminus of the linker, a passable signal sequence is produced. This is 22 amino acids long, and has a signal-like hydropathy plot, shown in fig 4-6. In this region there are two base changes between ricin and the agglutinin, one changing an isoleucine to a valine, and one changing a threonine to an alanine. The signal-like sequence overlaps the end of the A chain and the linker peptide. It is conceivable that the B chain gene, complete with signal peptide, became attached to the end of the A chain gene, in such a way as to delete the stop codon of the latter. Insertion of a base by mutation could then have restored the reading frame, leading to the expression of the heterodimer as a single-chain precursor. In time, the appropriate enzymatic processing mechanism would evolve to produce the correct product from this precursor. This is entirely speculative, and its elucidation may be achieved by analysis of genomic clones: if there is an intron in this region, then such a scenario is unlikely to

Met Ala Thr Ser Thr Ile Val Thr Val Phe Ala Tyr Lys Ala Ser Gly Thr Lys Phe Asn -  
Ricin: ATG CGC ACC TCC ACC ATC GTC ACA GTC TTT GCT TAT AAG GCC AGT GGT ACC AAA TTT AAT -  
RCA: G Val G Ala

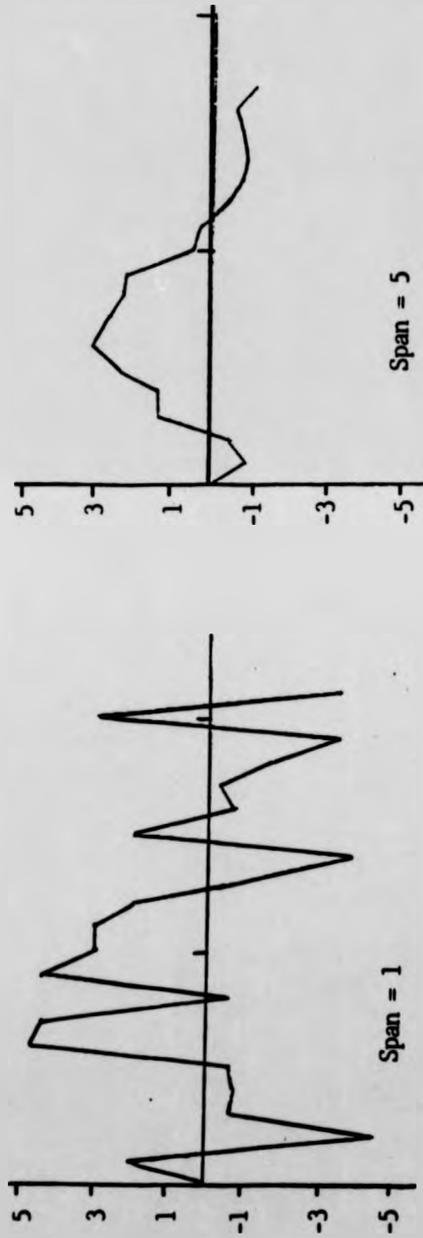


Fig 4-6: The signal-like sequence of the ricin B chain.

Above: Nucleotide and amino acid sequences; the base omitted is a T in the phenylalanine codon. The agglutinin counterpart, where it differs, is shown beneath.

Below: Hydropathy plots with spans of 1 (left) and 5 (right) for the ricin signal-like sequence.

be correct.

#### Future developments

The definitive identification of the two sequences presented here needs to be determined: the tentative identification is based only on the degree of similarity to the published protein sequences for the ricin chains; the accuracy of these may be questionable. Approaches to solving this problem have already been discussed, in section 3G1.

The most important advances related to the use of the clones have been in the immunotoxin field. Major problems have been encountered using native toxins for this purpose, and it is probable that the best toxic agent will be the whole ricin molecule, but lacking its sugar-binding function, and its carbohydrate. Work is in progress on the construction of clones which will express the sequences in yeast, using the vector pMA91 (247; provided by A and S Kingsman). Once expression of functional proteins is obtained, and adequate screening methods for alteration of function have been devised, a variety of mutagenesis techniques can be applied - libraries of mutagenised clones will be screened for loss of the galactose-binding site. As well as providing clones suitable for the development of immunotoxins, however, characterisation of mutated sequences will help to elucidate the location and mechanism of the galactose-binding site, and the second function of the B chain in transmembrane penetration of the A chain.

As far as obtaining the modified protein in unglycosylated form is concerned, yeast mutants have been constructed which are deficient in N-linked glycosylation of proteins (324). If the appropriate product can be obtained from such cells, and if it can be correctly linked to sufficiently specific antibodies, then the

remaining problem will be that of scale of production.

Other possibilities for exploiting the clones include their use as probes for genomic lectin sequences - not only from castor beans, but also to determine any homologies between these and related proteins in other species - and further mutagenesis experiments may be of value in understanding the functions of the A chains. Mutants of the A chains with increased toxicity may even be found - these would be even better than non-mutant A chains in immunotoxins!

### Conclusions

The work described here reports the determination of a new plant cDNA sequence, to be added to a small, but now rapidly growing group of such sequences. The castor bean lectin sequences contain features characteristic of comparable sequences - that is, they encode a single chain precursor which is post-translationally cleaved to form a disulphide-linked heterodimeric protein.

The precursor has a transiently associated signal sequence, and a short linker peptide is removed during processing of the translation product. Also, the 3' non-coding region is characteristic of plants, in that it contains multiple polyadenylation signals, and is of characteristic length, as is the 5' non-coding region.

Although there are unusual features - the tandem stop codon, and the upstream initiator codon - these have been observed before in plant systems.

The problems of definitively identifying the two sequences have been discussed, as have the possible uses of the clones, with special emphasis on the immunotoxin field.

The construction and characterisation of these clones is but the first of many steps, in a number of directions, towards the

achievement of a variety of ultimate goals. Every answer  
raises many questions....

Appendix 1: Computer Programmes

Sequences were analysed on an Apple II Europlus computer with 48 k RAM. Two such systems were used: one was equipped with a disk drive and a printer, and the other used tape storage and lacked a printer. Since the disk operating system uses some 10,000 bytes of RAM, this memory space was used for programming and data storage in the diskless system. This has the consequence that the restriction enzyme site searching programme is unoperable on the disk system, as presented here. Where programmes were run on both systems, the disk version is presented, the other differing in the memory areas used for sequence storage, and in instructions relating to disk control. For a description of the hardware and the programming languages used, see references 326-330.

The programming languages available on the Apple are the Applesoft dialect of BASIC, and 6502 machine code. The latter was used alone for the restriction sites programme, because of its speed, and in combination with Applesoft as simple subroutines. These subroutines are stored in the BASIC segment of programmes, and written into memory using READ...DATA instructions, and accessed by CALL commands. Routines and data are loaded into defined memory areas using POKE commands, and retrieved with the PEEK instruction.

There are three major advantages in using defined memory locations to store data: (i) the length of the sequence is limited only by the amount of memory available (in Applesoft the maximum character string length is some 255 characters); (ii) machine code programmes can directly access the data, allowing fast-running programmes; and (iii) data can be recoded into a format less arbitrary than letters of the alphabet, cutting down on 'irrelevant' processing time. Thus, each nucleotide or combination of nucleotides was recoded as a number from 1 to 15: A,T,C and G were represented as

1,2,4 and 8 respectively, and uncertainties, such as 'A or T' were generated additively. Each base is consequently represented as a single bit in each byte: bits set to 1 contain information as to which base(s) is located at the position represented by the byte. Simple logical comparisons can be performed on this data. For example, the AND function is able to detect similarities: if two bases being compared are identical, the product of AND is  $\emptyset$ , whereas if they are different, the product is NOT  $\emptyset$ ; the zero flag in the processor status register being adjusted accordingly. Subsequent instructions can cause the programme to jump, branch or continue, in response to the zero flag, and thus to the result of the comparison.

Protein sequences were not re-encoded, except inasmuch as their amino acids were represented by ASCII codes. This does not allow comparisons such as are described above for nucleotides, but does allow long sequences to be stored and rapidly retrieved.

Programmes on the disk system were organised in a menu-driven format. The menu programme itself is not presented here, as it is trivially simple. Some of the programmes end with the statement 'RUN MENU', which allows a subsequent programme to be loaded for analysis of data already entered. Loss of data in response to the RUN command is prevented by prior execution of a HIMEM: instruction.

Each programme will be introduced, briefly annotated, and listed in full. References 326-330 will assist in the interpretation of instruction usage and syntax, and further information will be provided on request.

A number of other programmes were also used, but are not described here, either because they are very simple or because the results are not presented. These include programmes to determine base composition, dinucleotide frequencies, codon usage and for plotting dot matrix homology plots.

SEQUENCE LOADER: NOTES & COMMENTS

The programme asks for input in the form of base or instruction codes. The former are letters representing bases or combinations of bases, as detailed below. Instruction codes are numbers, and allow clear screen presentation of sequence entered thus far, or corrections. Corrections can be changes, additions or deletions of bases. Alternatively, all data entered can be deleted and entering restarted.

Base and instruction codes

<u>Base code</u>	<u>Meaning</u>	<u>Instr code</u>	<u>Meaning</u>
A	A	1	Print sequence & End
T	T	2	Print & continue
C	C	3	Change base(s)
G	G	4	Add base(s)
B	A or T	5	Delete base(s)
D	A or C	6	Start again
E	A or G	7	Print part of sequence
F	T or C		
H	T or G		
I	C or G		
J	A or T or C		
K	A or T or G		
L	A or C or G		
M	T or C or G		
N	A or T or C or G		

Programme functions are now described by line number.

<u>Line</u>	<u>Comments</u>
3	Protect memory above byte 25000. NB Sequence data start at byte 25003, with length at 25001 (lsb) and 25002 (msb).
20 - 50	Assign base code letters to C\$
80	Initialise base number
90 - 160	Gets and interprets each keypress. Line 100 detects a letter code & skips line 110, which directs the programme to the line specified by the number entry. GOSUB 900 is to a subroutine which converts the letter code to the number code; line 140 stores the result in memory.

- 200 - 258 GOSUB 1000 prints the end of the sequence on screen; lines 220 - 240 determine and store the length of the sequence. GOTO 610 is to the routine which handles hard copy and disk storage.
- 260 - 300 Prints last part of sequence on screen for checking and returns to get another base or instruction code.
- 310 - 370 Line 330 gets range for base changes; D is the no of the first base to change, and E is the no of bases to change. Line 331 allows exit if routine called by error. Lines 335-365 enter the new data, interpret it (ie GOSUB900) and store it in place of the old bases. Line 370 returns to "print DNA and continue".
- 380 - 440 Line 410 gets the range of bases to be added. Lines 420 - 440 move following bases up in memory, and lines 450 - 485 get the new bases and place them in position.
- 500 - 570 Line 510 gets the range: in this case, parameters are entered as strings, to allow input of "E" to mean "delete to end". Line 527 reduces the length parameter to D, if range is "E". Line 525 converts the string characters to numerical values. Lines 530 - 550 move bases beyond the deletion down, and line 560 modifies the length parameter.
- 599 - 600 On "scrap entry" instruction, line 600 executes RLN.
- 610 - 800 If hard copy is required, subroutine at 1300 provides it. If disk storage is required, line 790 handles this. Line 800 then returns to the MENU programme.
- 900 - 940 Compares the entered base code to all its possible values and assigns a number to the variable A. If no assignation is made, lines 932 - 935 announe an input error, and request a replacement base.
- 1000 - 1070 Prints last 100 - 120 bases of sequence on screen. in groups of 10 bases, 3 such groups per line. The number of the last base in the line is displayed at right.
- 1100 -1210 Prints specified region of sequence on screen; format as above.
- 1300 - 1370 As 1000 - 1070, except that the whole sequence is printed.

```
1      REM ***SEQUENCE LOADER***
3      HIMEM: 25000
5      DIM C$(15)
10     HOME
20     FOR T = 1 TO 15 : READ C$(T) : NEXT
50     DATA A,T,B,C,D,F,J,G,E,H,K,I,L,M,N
60     PRINT "SEQUENCE LOADER" : PRINT
70     PRINT "ENTER BASE OR INSTRUCTION CODE" : PRINT
80     N = 1
90     GET AS
100    IF VAL(AS) = 0 THEN 120
110    ON VAL(AS) GOTO 200,260,310,380,500,600,1100
120    GOSUB 900
130    PRINT AS;
140    POKE 25002 + N, A
150    N = N + 1
160    GOTO 90

200    REM ***PRINT DNA AND END***
210    GOSUB 1000
220    A = INT((N - 1) / 256)
230    B = N - 1 - 256 * A
240    POKE 25001,B : POKE 25002,A
250    PRINT : PRINT N - 1; " BASES"
258    GOTO 610

260    REM ***PRINT DNA AND CONTINUE***
270    GOSUB 1000
280    PRINT : PRINT : PRINT N - 1; " BASES"
290    PRINT : PRINT "ENTER BASE OR INSTRUCTION CODE"
295    PRINT : PRINT
300    GOTO 90

310    REM ***CHANGE BASES***
325    PRINT : PRINT : PRINT : PRINT
330    INPUT "CHANGE BASES....ENTER RANGE "; D,E
331    IF E = 0 THEN 260
333    PRINT : PRINT : PRINT "ENTER NEW BASES" : PRINT : PRINT
335    FOR T = 1 TO E
340    GET AS
345    PRINT AS;
350    GOSUB 900
360    POKE 25001 + D + T, A
365    NEXT T
370    GOTO 260

380    REM ***ADD BASES***
400    PRINT : PRINT
410    INPUT "ADD BASES....ENTER RANGE "; D,E
415    IF E = 0 THEN 260
420    FOR T = N - 1 TO D + 1 STEP -1
430    POKE 25002 + T + E, PEEK (25002 + T)
440    NEXT
450    PRINT : PRINT : PRINT "ENTER NEW BASES" : PRINT : PRINT
453    N = N + E
455    FOR T = 1 TO E
460    GET AS : PRINT AS;
465    GOSUB 900
480    POKE 25002 + D + T, A
```

```
485 NEXT
490 GOTO 260
500 REM ***DELETE BASES***
505 PRINT : PRINT : PRINT
510 INPUT "DELETE BASES....ENTER RANGE "; D$,E$
520 IF E$ = "0" THEN 260
525 D = VAL(D$) : E = VAL(E$)
527 IF E$ = "E" THEN N = D : GOTO 260
530 FOR T = D + E TO N - 1
540 POKE 25001 + T - E, PEEK(25002 + T)
550 NEXT
560 N = N - E
570 GOTO 260

599 REM ***RESTART***
600 RUN

610 REM ***PRINT HARD COPY***
620 PRINT : PRINT "PRINT SEQUENCE?" : GET A$
640 IF A$ # "Y" THEN 730
650 HOME
660 INPUT "NAME OF SEQUENCE? "; A$
670 PR#1
680 PRINT A$ : PRINT : PRINT
690 GOSUB 1300
700 PRINT : PRINT
710 PR#0
730 PRINT : PRINT
735 PRINT "STORE ON DISK? " : GET B$
737 IF B$ "Y" THEN PRINT CHR$(4)
740 IF B$ # "Y" THEN PRINT : PRINT CHR$(4); "RUN MENU"
750 IF LEN(A$) > 1 THEN 790
760 PRINT : INPUT "NAME OF SEQUENCE? "; A$
770 PRINT
790 PRINT CHR$(4); "BSAVE";A$; ",A25001,L";N+2
800 PRINT : PRINT CHR$(4);"RUN MENU"

900 REM ***BASE CODE TO NUMBER CODE***
910 A = (A$ = "A") + 2 * (A$ = "T") + 4 * (A$ = "C") + 8 * (A$ = "G");
IF A # 0 THEN 940
920 A = 3 * (A$ = "B") + 5 * (A$ = "D") + 9 * (A$ = "E") +
6 * (A$ = "F") + 10 * (A$ = "H") + 12 * (A$ = "I");
IF A$ # 0 THEN 940
930 A = 7 * (A$ = "J") + 11 * (A$ = "K") + 13 * (A$ = "L") +
15 * (A$ = "N")
IF A # 0 THEN 940
931 CALL 64477
932 PRINT : PRINT "INPUT ERROR...TRY AGAIN ...BASE CODES ONLY"
933 GET A$
934 GOTO 910
935 RETURN

1000 REM ***PRINT DNA SUBROUTINE***
1010 HOME
1020 S = 0 : U = 0
1022 IF N - 1 < 130 THEN R = 1 : GOTO 1030
```

```
485 NEXT
490 GOTO 260
500 REM ***DELETE BASES***
505 PRINT : PRINT : PRINT
510 INPUT "DELETE BASES....ENTER RANGE "; D$,E$
520 IF E$ = "" THEN 260
525 D = VAL(D$) : E = VAL(E$)
527 IF E$ = "E" THEN N = D : GOTO 260
530 FOR T = D + E TO N - 1
540 POKE 25001 + T - E, PEEK(25002 + T)
550 NEXT
560 N = N - E
570 GOTO 260

599 REM ***RESTART***
600 RUN

610 REM ***PRINT HARD COPY***
620 PRINT : PRINT "PRINT SEQUENCE?" : GET A$
640 IF A$ # "Y" THEN 730
650 HOME
660 INPUT "NAME OF SEQUENCE? "; A$
670 PR#1
680 PRINT A$ : PRINT : PRINT
690 GOSUB 1300
700 PRINT : PRINT
710 PR#0
730 PRINT : PRINT
735 PRINT "STORE ON DISK? " : GET B$
737 IF B$ "Y" THEN PRINT CHR$(4)
740 IF B$ # "Y" THEN PRINT : PRINT CHR$(4); "RUN MENU"
750 IF LEN(A$) > 1 THEN 790
760 PRINT : INPUT "NAME OF SEQUENCE? "; A$
770 PRINT
790 PRINT CHR$(4); "ESAVE";A$; ",A25001,L";N+2
800 PRINT : PRINT CHR$(4);"RUN MENU"

900 REM ***BASE CODE TO NUMBER CODE***
910 A = (A$ = "A") + 2 * (A$ = "T") + 4 * (A$ = "C") + 8 * (A$ = "G");
IF A # 0 THEN 940
920 A = 3 * (A$ = "B") + 5 * (A$ = "D") + 9 * (A$ = "E") +
6 * (A$ = "F") + 10 * (A$ = "H") + 12 * (A$ = "I");
IF A$ # 0 THEN 940
930 A = 7 * (A$ = "J") + 11 * (A$ = "K") + 13 * (A$ = "L") +
15 * (A$ = "N")
931 IF A # 0 THEN 940
932 CALL 64477
933 PRINT : PRINT "INPUT ERROR...TRY AGAIN ...BASE CODES ONLY"
934 GET A$
935 GOTO 910
940 RETURN

1000 REM ***PRINT DNA SUBROUTINE***
1010 HOME
1020 S = 0 : U = 0
1022 IF N - 1 < 130 THEN R = 1 : GOTO 1030
```

```
1024 R = 10 * INT((N - 1) / 10) - 109
1030 FOR T = R TO N - 1
1040 PRINT C$(PEEK(25002 + T));
1050 S = S + 1 : IF S = 10 THEN PRINT " " : S = 0 : U = U + 1 :
IF U = 3 THEN PRINT " " : T : PRINT : U = 0
1060 NEXT
1070 RETURN

1100 REM ***PRINT PART OF SEQUENCE***
1110 PRINT : PRINT : PRINT
1120 INPUT "ENTER PRINT RANGE "; D,E
1130 HOME
1140 S = 0 : U = 0
1150 FOR T = D TO E
1160 PRINT C$(PEEK(25002 + T));
1170 S = S + 1 : IF S = 10 THEN PRINT " " : S = 0 : U = U + 1 :
IF U = 3 THEN PRINT " " : T : PRINT : U = 0
1180 NEXT
1190 PRINT : PRINT : PRINT
1200 PRINT "ENTER INSTRUCTION CODE "; GET A
1210 ON A GOTO 200,260,310,380,500,600,1100

1300 REM ***PRINT HARD COPY SUBROUTINE***
1310 HOME
1320 S = 0 : U = 0
1330 FOR T = 1 TO N - 1
1340 PRINT C$(PEEK(25002 + T));
1350 S = S + 1 : IF S = 10 THEN PRINT " " : S = 0 : U = U + 1 :
IF U = 3 THEN PRINT " " : T : PRINT : U = 0
1360 NEXT
1370 RETURN
```

TRANSLATION: NOTES & COMMENTS

The DNA sequence entered by SEQUENCE LOADER, or from disk, is used to generate amino acid sequences from all six possible reading frames. The first three frames start at the first, second and third bases, using the sequence as entered, assuming that it is 5' to 3'. Note that the DNA sequence must be absolutely certain.

One-letter amino acid codes are stored in a three-dimensional array, indexed by numbers 1 - 4. Since the sequence is stored in terms of numbers 1,2,4 & 8, a machine code routine carries out the appropriate conversion - for the complementary frames, a second machine code routine converts 1,2,4 & 8 into 2,1,8 & 4 respectively. These routines are stored in READ...DATA statement in the BASIC programme. Operation requires only the entering of the sequence name.

<u>Line number</u>	<u>Comments</u>
20 - 80	Store one-letter codes in 3-dimensional array.
100 - 140	Store machine code subroutine on p3 of memory,
142	Switch on printer (later switched off by PR#0)
150	Read sequence length
160 - 190	Read reading frame parameters; M\$ = title of frame; CA = adrs of machine code subroutine; D = direction of translation
200	Reading frame parameters: S = first base in specified frame
201 - 206	Reading frame parameters: E = last base
210 & 390	Control loop calling each frame in turn
300 & 370	Control loop reading through DNA sequence
310 - 350	Determine indexes for array containing AA codes
360	Print appropriate one-letter AA code
400	Return to MENU programme

```
5 REM ***TRANSLATION***
7 HIMEM: 20000
10 HOME
20 DIM W$(4,4,4)
30 FOR X = 1 TO 4
40 FOR Y = 1 TO 4
50 FOR Z = 1 TO 4
60 READ W$(X,Y,Z)
70 NEXT Z,Y,X
80 DATA K,N,N,K,I,I,I,M,T,T,T,T,R,S,S,R,*,Y,Y,*,L,F,F,L,S,S,
S,S,*,C,C,W,Q,H,H,Q,L,L,L,L,P,P,P,P,R,R,R,R,E,D,D,E,
V,V,V,V,A,A,A,A,G,G,G,G
100 FOR N = 768 TO 823
110 READ A
120 POKE N,A
130 NEXT
140 DATA 165,8,201,1,208,1,96,201,2,208,1,96,201,4,208,5,169,3,
133,8,96,169,4,133,8,96,165,8,201,4,208,1,96,201,1,208,
5,169,2,133,8,96,201,2,208,5,169,1,133,8,96,169,3,133,8,96
141 INPUT "NAME OF SEQUENCE? "; A$
142 PR#1
143 PRINT A$
144 PRINT : PRINT
150 L = PEEK(25001) + 256 * PEEK(25002)
160 FOR N = 1 TO 6
170 READ M$(N), CA(N), D(N)
180 NEXT
190 DATA RF1,768,1,RF2,768,1,RF3,768,1,CRF1,794,-1,CRF2,794,-1,
CRF3,794,-1
200 S(1) = 1 : S(2) = 2 : S(3) = 3 : S(4) = L : S(5) = L - 1 :
S(6) = L - 2
201 E(1) = 3 * INT(L / 3) - 2
202 E(2) = 3 * INT((L - 1) / 3) - 1
203 E(3) = 3 * INT((L - 2) / 3)
204 E(4) = L - 3 * INT(L / 3) + 2
205 E(5) = L - 3 * INT((L - 1) / 3) + 1
206 E(6) = L - 3 * INT((L - 2) / 3)
210 FOR P = 1 TO 6
220 PRINT M$(P)
300 FOR Z = S(P) TO E(P) STEP 3 * D(P)
310 FOR N = 0 TO 2
320 POKE 8, PEEK(25002 + D(P) * N + Z)
330 CALL CA(P)
340 W(N) = PEEK(8)
350 NEXT N
360 PRINT W$(W(0),W(1),W(2));
370 NEXT Z
380 PRINT : PRINT
390 NEXT P
392 PR#0
395 PRINT : PRINT "PRESS ANY KEY TO RETURN TO MENU " : GET A$
397 PRINT
400 PRINT CHR$(4); "RUN MENU"
```

### RE SITES

This programme is written entirely in machine code, and comprises four parts: (1) the main control loop, located at \$7D01; (2) NFILE, the file of restriction enzyme names, at \$7531, which also contains the addresses of the restriction enzyme recognition sequences, and the lengths of the sequences; (3) the list of recognition sequences, at \$7F01; and (4) a number of subroutines which start at \$8100. These subroutines are: (1) ADD1, which increments a two-byte number on the zero page, at \$8100; (2) PRNFILE, at \$8120, which displays the name of the current restriction enzyme on screen; (3) RDADRS, at \$8170, which reads the address of the current enzyme's recognition sequence, and places it in a location accessible by the main programme, and also reads its length.

The main programme starts by initialising a number of variables, and by printing the first message ("RE SITES") in the NFILE on screen. The programme then enters the main control loop by calling the next entry in NFILE. If this is \$FF, with which NFILE must end, then the programme prints the message "END" on screen, and ends. Otherwise, the next NFILE entry is displayed on screen, and the address and length of the current enzyme are read. The first base of the RE sequence is then compared with the first base of the DNA sequence, the variable measuring the latter being incremented. If these two bases are different, a test is performed to determine whether the whole DNA sequence has been examined; if so, the programme reinitialises parameters relating to it, and returns to get the next enzyme sequence. If the end of the DNA sequence has not been reached, the next base is compared with the first RE sequence base.

When the first RE sequence base is the same as a base in the DNA sequence, the position of the latter is stored and a variable (the "+ counter") is incremented. The next base in the RE sequence is then compared with the next in the DNA sequence, and a test is performed to determine whether a complete site has been found. If not, then the next 2 bases are compared, and so on until either a complete recognition sequence is found, or until a mismatch is found. In the latter case, the programme returns to the main loop and looks at the next base in the DNA sequence. If a complete RE site is found, its location in the DNA sequence is printed

and the programme returns to examine the next base in the DNA sequence.

These events are summarised in the flow chart presented on the next page. Programme control variables are located on the zero page, and are listed below:

<u>Address (\$)</u>	<u>Name</u>
6,7	NFILE current address
8,9	RE-FILE current address
18	RE-FILE current length minus one
19	ENDFLAG
1A,1B	DNA sequence current address
1C,1D	Current base number (hexadecimal)
1E,1F	Last base number
4A,4B	RE-FILE address working location
4D	+ counter
CE,CF	DNA sequence address working location
<u>EB,EC,ED</u>	<u>Current base number (bcd)</u>

The programme is used as follows: it is loaded from tape by entering the system monitor with CALL-151, and loaded with the instruction 7531.8190R. (Note that a DNA sequence must be loaded with sequence loader - both programmes can reside in memory at the same time, so the sequence can be loaded after loading RE SITES, as long as RE SITES is not run without a sequence in position.) RE SITES is then run with the instruction 7D01G. After each enzyme search the computer beeps; the next search is initiated by pressing any key.

The programme is listed on following pages in assembler format, though JSR and JMP destination addresses have been replaced with mnemonics.

and the programme returns to examine the next base in the DNA sequence.

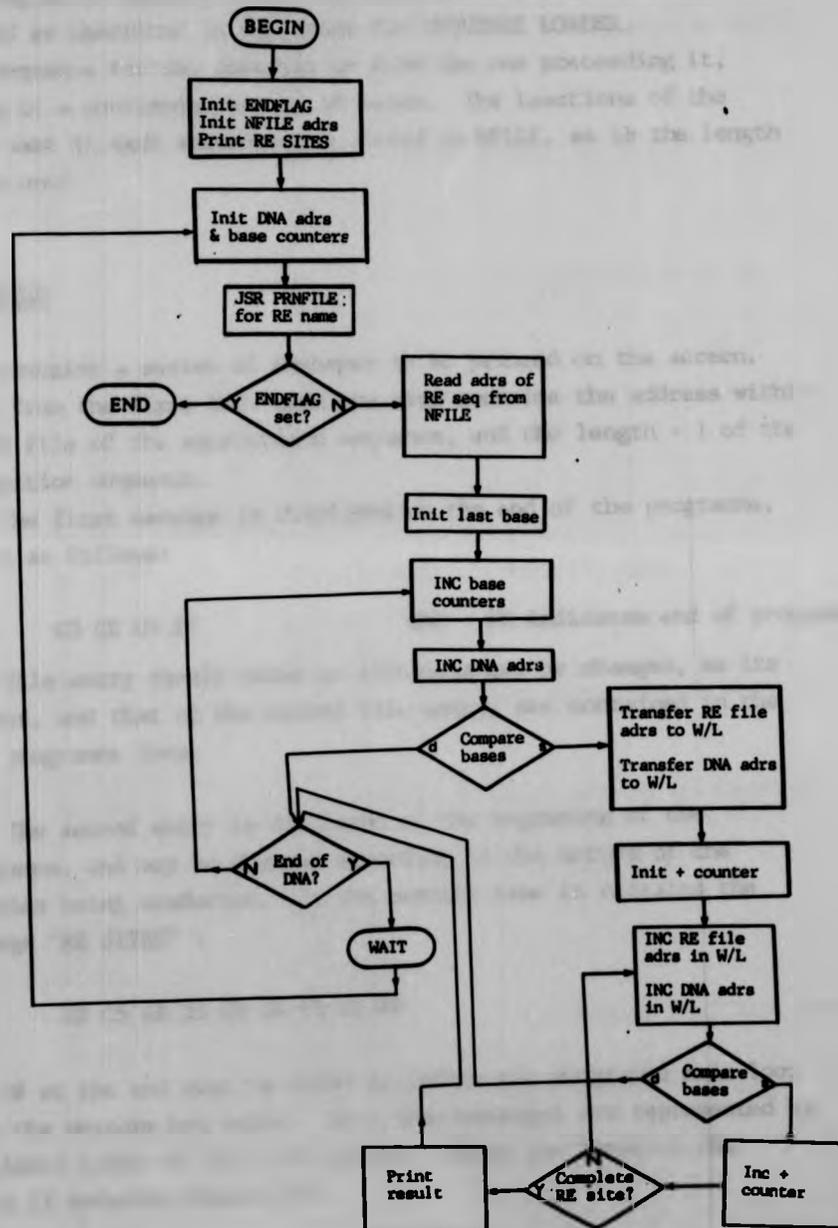
These events are summarised in the flow chart presented on the next page. Programme control variables are located on the zero page, and are listed below:

<u>Adress (\$)</u>	<u>Name</u>
6,7	NFILE current address
8,9	RE-FILE current address
18	RE-FILE current length minus one
19	ENDFLAG
1A,1B	DNA sequence current address
1C,1D	Current base number (hexadecimal)
1E,1F	Last base number
4A,4B	RE-FILE address working location
4D	+ counter
CE,CF	DNA sequence address working location
<u>EB,EC,ED</u>	<u>Current base number (bcd)</u>

The programme is used as follows: it is loaded from tape by entering the system monitor with CALL-151, and loaded with the instruction 7531.8190R. (Note that a DNA sequence must be loaded with sequence loader - both programmes can reside in memory at the same time, so the sequence can be loaded after loading RE SITES, as long as RE SITES is not run without a sequence in position.) RE SITES is then run with the instruction 7D01G. After each enzyme search the computer beeps; the next search is initiated by pressing any key.

The programme is listed on following pages in assembler format, though JSR and JMP destination addresses have been replaced with mnemonics.

RE SITES: Flow chart



### The RE file

This begins at address \$7F01, and contains the RE sequences, encoded as described in the notes for SEQUENCE LOADER. Each sequence follows directly on from the one preceding it, giving it a continuous series of bases. The locations of the first base in each sequence are stored in NFILE, as is the length (minus one).

### THE NFILE

This contains a series of messages to be printed on the screen. Apart from the first two, each one also contains the address within the RE file of the appropriate sequence, and the length - 1 of the recognition sequence.

The first message is displayed at the end of the programme, and is as follows:

```
7531   C5 CE C4 FF           END   FF indicates end of programme
```

This file entry should under no circumstances be changed, as its address, and that of the second file entry, are contained in the main programme loop.

The second entry is displayed at the beginning of the programme, and may be changed according to the nature of the searches being conducted. In the present case it contains the message "RE SITES" :

```
7535   D2 C5 A0 D3 C9 D4 C5 D3 00
```

The 00 at the end must be added to inform the programme main loop that the message has ended. Note that messages are represented by the ASCII codes of their characters. These are found in the Apple II Reference Manual p15.

Subsequent entries contain the RE name in ASCII codes, followed by 00 to indicate end of entry, followed by the address of the sequence (msb first), then by the length of the sequence - 1. Entries follow continuously, and the last one is followed by \$FF, which indicates end of NFILE.

RE SITES: LISTING

7D01	20 58 FC	JSR CLRSCRN	
7D04	A9 00	LDA #00	ENDFLAG Init
7D06	85 19	STA \$19	
7D08	A9 35	LDA #35	Init NFILE adrs to \$7535
7D0A	85 06	STA \$06	
7D0C	A9 75	LDA #75	
7D0E	85 07	STA \$07	
7D10	20 20 81	JSR PRNFILE	Print "RE SITES"
7D13	20 8E FD	JSR CROUT	
7D16	20 8E FD	JSR CROUT	
7D19	A9 42	LDA #42	Init DNA sequence adrs to \$9C42
7D1B	85 1A	STA \$1A	
7D1D	A9 9C	LDA #9C	
7D1F	85 1B	STA \$1B	
7D21	A9 00	LDA #00	Init both base counters to \$00
7D23	85 1C	STA \$1C	
7D25	85 1D	STA \$1D	
7D27	85 EB	STA \$EB	
7D29	85 EC	STA \$EC	
7D2B	85 ED	STA \$ED	
7D2D	20 20 81	JSR PRNFILE	Print RE name
7D30	A5 19	LDA \$19	ENDFLAG test
7D32	C9 01	CMP #01	
7D34	D0 01	BNE \$7D37	Branch if clear
7D36	60	RTS	End programme if set
7D37	20 8E FD	JSR CROUT	
7D3A	20 8E FD	JSR CROUT	
7D3D	20 70 81	JSR RDADRS	Read adrs of RE sequence from NFILE
7D40	38	SEC	Init last base
7D41	AD 41 9C	LDA \$9C41	
7D44	E5 18	SBC \$18	
7D46	85 1E	STA \$1E	
7D48	E9 00	SBC #00	
7D4B	85 1F	STA \$1F	
7D4F	A2 1C	LDX #1C	Inc hexadecimal base counter
7D51	20 00 81	JSR ADD1	
7D54	F8	SED	Inc bcd base counter
7D55	18	CLC	
7D56	A2 FF	LDX #FF	
7D58	E8	INX	
7D59	E0 03	CPX #03	
7D5B	F0 0A	BEQ \$7D67	

7D5D	B5 EB	LDA \$EB,X	
7D5F	69 01	ADC #\$01	
7D61	95 EB	STA \$EB,X	
7D63	C9 00	CMP #\$00	
7D65	F0 F1	BEQ \$7D58	
7D67	D8	CLD	
7D68	A2 1A	LDX \$1A	Inc current DNA address
7D6A	20 00 81	JSR ADD1	
7D6D	A0 00	LDY #\$00	LDA current DNA base
7D6F	B1 1A	LDA (\$1A),Y	
7D71	31 08	AND (\$08),Y	Compare with RE base
7D73	D0 27	BNE \$7D9C	Branch if bases same
7D75	A5 1C	LDA \$1C	End of DNA sequence?
7D77	C5 1E	CMP \$1E	No...jump to \$7D4F
7D79	F0 03	BEQ \$7D7E	Yes..continue
7D7B	4C 4F 7D	JMP \$7D4F	
7D7E	A5 1D	LDA \$1D	
7D80	C5 1F	CMP \$1F	
7D82	F0 03	BEQ \$7D87	
7D84	4C 4F 7D	JMP \$7D4F	
7D87	20 DD FB	JSR BELL1	Beep the speaker
7D8A	20 0C FD	JSR RDKEY	Get keypress
7D8D	20 58 FC	JSR CLRSCRN	
7D90	4C 19 7D	JMP \$7D19	Jump to next enzyme
7D93	EA x 9	NOP codes	Space to allow programme modification
Jump to here on bases same:			
7D9C	A5 08	LDA \$08	Transfer RE-FILE adrs to working location
7D9E	85 4A	STA \$4A	
7DA0	A5 09	LDA \$09	
7DA2	85 4B	STA \$4B	
7DA4	A5 1A	LDA \$1A	Transfer DNA sequence address to working location
7DA6	85 CE	STA \$CE	
7DA8	A5 1B	LDA \$1B	
7DAA	85 CF	STA \$CF	
7DAC	A9 00	LDA #\$00	Init + counter
7DAE	85 4D	STA \$4D	
7DB0	A2 4A	LDK \$4A	Inc RE-File adrs in working location
7DB2	20 00 81	JSR ADD1	
7DB5	A2 CE	LDK \$CE	Inc DNA adrs in working location
7DB7	20 00 81	JSR ADD1	

7DBA	A0 00	LDY #000	LDA DNA base
7DBC	B1 CE	LDA (\$CE),Y	
7DBE	31 4A	AND (\$4A),Y	Compare with RE base
7DC0	D0 03	BNE \$7DC5	Branch if same
7DC2	4C 75 7D	JMP \$7D75	Return if different
7DC5	E6 4D	INC \$4D	Inc + counter
7DC7	A5 4D	LDA \$4D	Test for complete RE site
7DC9	C5 18	CMP \$18	
7DCB	D0 E3	BNE \$7DB0	Next pair of incomplete
7DCD	EA x 10	NOP codes	Space for modification
7DD7	A5 ED	LDA \$ED	Print result on screen
7DD9	20 DA FD	JSR PRBYTE	
7DDC	A5 EC	LDA \$EC	
7DDE	20 DA FD	JSR PRBYTE	
7DE1	A5 EB	LDA \$EB	
7DE3	20 DA FD	JSR PRBYTE	
7DE6	20 48 F9	JSR PRBLNK	
7DE9	4C 75 7D	JMB \$7D75	Return to ENDTEST

Subroutine: ADD1

Increments a 2-byte number on the zero page; needs the address of the first of these (the lsb) in the X register on entry.

8100	8A	TXA	
8101	8D 08 81	STA \$8108	lsb into INC instruction
8104	8D 0A 81	STA \$810A	lsb into LDA instruction
8107	E6 00	INC \$00	Inc lsb
8109	A5 00	LDA \$00	Transfer to accumulator
810B	C9 00	CMP #000	Inc msb?
810D	F0 01	BEQ \$8110	Yes...go and do it
810F	60	RTS	No....return
8110	E8	INC X	
8111	8A	TXA	
8112	8D 16 81	STA \$8116	msb into INC instruction
8115	E6 00	INC \$00	Inc msb
8117	60	RTS	Return

Subroutine: PRNFILE

Prints contents of current entry in NFILE. Each byte is tested for \$FF, which indicates no more RE name files, and then for \$00, which indicates end of current entry. NFILE current address is incremented for every byte read.

8120	A0 00	LDY #000	Load byte from NFILE
8122	B1 06	LDA (\$06),Y	
8124	C9 FF	CMP #0FF	End of programme?

8126	F0 17	BEQ \$813F	Yes...goto endroutine
8128	B1 06	LDA (\$06),Y	Reload byte from NFILE
812A	C9 00	CMP #00	End of current entry?
812C	D0 06	BNE \$8134	No...goto print character
812E	A2 06	LDX #06	Inc NFILE address
8130	20 00 81	JSR ADD1	
8133	60	Return	
8134	20 F0 FD	JSR COUT1	Print character
8137	A2 06	LDX #06	Inc NFILE address
8139	20 00 81	JSR ADD1	
813C	4C 22 81	JMP \$8122	Get next byte from NFILE
813F	A5 19	LDA \$19	ENDFLAG set?
8141	C9 00	CMP #00	
8143	D0 0F	BNE \$8154	Yes...Goto return
8145	A9 01	LDA #01	Set ENDFLAG
8147	85 19	STA \$19	
8149	A9 31	LDA #31	Set NFILE adrs for 'END'
814B	85 06	STA \$06	message
814D	A9 75	LDA #75	
814F	85 07	STA \$07	
8151	4C 20 81	JMP \$8120	Print 'END' from NFILE
8154	60	RTS	Return

Subroutine: RDADRS

This reads the address of the first byte of the current enzyme in the NFILE, storing it in \$08, \$09. It then reads the (length - 1) of the recognition sequence and stores it in \$18.

8170	A0 00	LDY #00	lsb goes into \$08
8172	B1 06	LDA (\$06),Y	
8174	85 08	STA \$08	
8176	A2 06	LDX #06	Inc NFILE address
8178	20 00 81	JSR ADD1	
817B	B1 06	LDA (\$06),Y	msb goes into \$09
817D	85 09	STA \$09	
817F	A2 06	LDA #06	Inc NFILE address
8181	20 00 81	JSR ADD1	
8184	B1 06	LDA (\$06),Y	Length - 1 goes into \$18
8186	85 18	STA \$18	
8188	A2 06	LDX #06	Inc NFILE address
818A	20 00 81	JSR ADD1	
818D	60	RTS	Return

Protein Hydropathy: Notes & Comments

Hydropathy values for individual amino acids were obtained from reference 255. These are averaged over a specified number of residues (the span) and the result is plotted against the position in the polypeptide chain of the first residue in the span.

<u>Line</u>	<u>Comments</u>
15	Reserve memory above 20000
20-40	Read hydropathy values into variable A, indexed by ASCII codes for single letter amino acid codes.
50 -100	Enter sequence name, and sequence, and poke latter into memory locations starting at 20001.
110	Adjust length variable.
120	Print sequence on screen for checking.
140	END programme to allow for error correction. This is done by calculating position of error (ie position in sequence +20000), and POKEing the ASC function of the character into this location, in immediate mode.
200	Determines horizontal distance plotted per residue.
210	Switch on printer/plotter and adjust origin of graph.
220-310	Draw axes with markers. GET lines (285,305) stop the programme to allow adjustment of paper position in plotter.
320	Enter span variable.
340-400	Control loop for calculations.
350-380	Reset hydropathy variable, then accumulate hydropathies over the set span.
390	Calculate average hydropathy over the span and plot it.
410-460	Print name of sequence and span, then ask if another span, or another sequence, is required.

```
10 REM ***HYDROPATHY***
15 HIMEM: 20000
20 DIM A(89)
30 FOR N = 65 TO 89 : READ A(N) : NEXT
40 DATA 1.8,0,2.5,-3.5,-3.5,2.8,-.4,-3.2,4.5,0,-3.9,3.8,
1.9,-3.5,0,-1.6,-3.5,-4.5,-.8,-.7,0,4.2,-.9,0,-1.3
50 HOME
51 PRINT "ENTER SEQUENCE NAME " : INPUT NS : PRINT : PRINT
55 PRINT "ENTER SEQUENCE " : PRINT : PRINT : PRINT
60 Q = 1
70 GET AS : PRINT AS; : IF AS = "Z" THEN 110
80 POKE 20000 + Q, ASC(AS)
90 Q = Q + 1
100 GOTO 70
110 HOME : Q = Q - 1
120 FOR A = 1 TO Q : PRINT CHR$(PEEK(20000 + A)); : NEXT
130 PRINT : PRINT
140 PRINT "CORRECT ANY ERRORS" : PRINT : PRINT "THEN TYPE GOTO
200 HOME : INPUT "PLOT POINTS PER RESIDUE? "; Y
205 PR#1
210 PRINT CHR$(18); "I"; "M0,-999"; "I"
220 C = 799 - C * Q
230 PRINT "L0"; M200,"C"; "D200,999,0,999,350,999"
250 PRINT "M200,999"
270 FOR T = 999 TO C STEP - 50 * Y : PRINT "M200,"T : PRINT
"D250,"T : NEXT : PRINT "M200,999"
280 GET CS
290 PR#1
300 FOR T = 999 TO C STEP - 10 * Y : PRINT "M200,"T : PRINT
"D210,"T : NEXT : PRINT "M200,999"
305 GET CS
310 PR#0
320 HOME : PRINT "ENTER SPAN " : INPUT S
330 PR#1
340 FOR N = 1 TO 1 + Q - S
350 H = 0
360 FOR P = 0 TO S - 1
370 H = H + A(PEEK(20000 + N + P))
380 NEXT P
390 PRINT "D'200 + 30 * H / S, 999 - Y * N
400 NEXT N
410 PRINT "M0,950"; "Q1"
420 PRINT : PRINT "P"; NS; " SPAN = "; S
430 PRINT "M0,"C - 100
440 PR#0
450 HOME : PRINT "ANOTHER PLOT? " : GET XS : IF XS = "Y" THEN 200
460 PRINT : PRINT "ANOTHER SEQ? " : GET XS : IF XS = "Y" THEN RUN
```

HYDROPATHY DIFFERENCE: NOTES & COMMENTS

Values for hydrophathies of individual amino acids are as in HYDROPATHY. This programme functions in the same way as the previous one, but stores two sequences, no 1 at byte 30000 and no 2 at byte 32000. As the programme steps through the sequences, the average hydrophathy value is calculated for both sequences simultaneously, and the difference is plotted. Since the operation of the programme is the same as that of HYDROPATHY, no further details are given here.

```
10 REM ***HYDROPATHY DIFFERENCE***
15 HIMEM: 30000
20 DIM A(89) : FOR N = 65 TO 89 : READ A(N) : NEXT
30 DATA 1.8,0,2.5,-3.5,-3.5,2.8,-.4,-3.2,4.5,0,-3.9,-3.8,
1.9,-3.5,0,-1.6,-3.5,-4.5,-.8,-.7,0,4.2,-.9,0,-1.3
100 FOR N = 0 TO 1 : HOME
110 PRINT "ENTER NAME OF SEQ "; N + 1 : INPUT A$(N + 1)
120 PRINT : PRINT "ENTER SEQ "; N + 1 : PRINT
130 IF N = 0 THEN PRINT "SEQ 2 IS SUBTRACTED FROM SEQ 1" : PRINT
140 B = 1
150 GET A$ : PRINT A$ ; : IF A$ = "Z" THEN 200
160 POKE 30000 + B + 2000 * N, ASC(A$)
180 B = B + 1
190 GOTO 150
200 PRINT : PRINT "CORRECT ANY ERRORS" : PRINT
210 IF N = 0 THEN PRINT "THEN TYPE GOTO 250" : END
220 PRINT "THEN TYPE GOTO 260" : END
250 NEXT N
260 HOME : PRINT "ENTER SPAN "; S
265 PR#1
270 PRINT CHR$(18); "I" : C = 4 * B : PRINT "M0,-999"; "I"
300 PRINT "M200," 999 - C; "D200, 999"; "M0,999"
310 PRINT "D400,999"; "M200,999"
360 FOR T = 999 TO 999 - C STEP -200 : PRINT "M200,"T : PRINT
"D250,"T : NEXT : PRINT "M200,999"
370 GET D$
380 PR#1
390 FOR T = 999 TO 999 - C STEP -40 : PRINT "M200,"T : PRINT
"D210,"T : NEXT : PRINT "M200,999"
395 GET D$
397 PR#1
400 FOR N = 1 TO B - S
410 H(1) = 0 : H(2) = 0
420 FOR P = 0 TO S - 1
430 H(1) = H(1) + A(PEEK(30000 + N + P))
460 H(2) = H(2) + A(PEEK(32000 + N + P))
450 NEXT P
460 D = (H(1) - H(2)) / S
470 PRINT "D"200 + 50 * D", "999 - 4 * N
480 NEXT N
490 PRINT "M10,980"; "Q1"; "HYDROPATHY DIFFERENCE: "; A$(1);
" - "; A$(2); " SPAN = "; S
500 PRINT "M0,"900 - C
510 PR#0
```

Secondary structure prediction programme

The prediction method of Garnier et al (331) was used, since a comparison of various methods (332) found that it is methodologically simpler, more computer-adaptable and more objective than others, including that of Chou and Fasman (333).

The sequences were first analysed with both decision constants set to zero, as suggested by Thornton & Taylor (334). This indicated that the A chains are possible beta/alpha proteins (as categorised by Thornton & Taylor), while the B chains are all-beta proteins, having only about 5 % of alpha structure. Again, on the suggestion of Thornton & Taylor, the A chains, as possible beta/alpha proteins, were reanalysed with the alpha decision constant set to 100. This reduced its predicted alpha content to 9.2 %, putting it into the all-beta category.

The programme starts by setting up a three-dimensional array of 4 x 25 x 17 elements. The first dimension refers to the conformational states alpha, extended, turn and coil, that is to tables 1 to 4 in ref 331. The second dimension represents the one-letter amino acid codes (the unused ones are filled with zero's). The third dimension refers to the relative position of the amino acid to the one whose conformational state is being predicted. The programme then loads sequence data as ASCII codes, and runs through the sequence evaluating the informational parameters S(1 - 4) for each amino acid residue. For each, the largest is selected, and indicates which conformational state is most likely for that residue.

```
5 HIMEM: 30000
10 REM ***GARNIER PROT PREDICTION***
20 HOME
30 DIM A(4,25,17)
40 FOR N = 1 TO 4
50 FOR P = 1 TO 25
60 FOR Q = 1 TO 17
70 READ A(N,P,Q)
80 NEXT Q,P,N
85
to DATA statements
880
1000 HOME
1010 PRINT "ENTER AA SEQUENCE" : PRINT : PRINT
1015 L = 1
1020 GET AS : PRINT AS;
1025 IF AS = "Z" THEN L = L - 1 : GOTO 1100
1030 POKE 30000 + L, ASC(AS)
1040 L = L + 1
1050 GOTO 1020
1100 REM ***DECISION CONSTANTS***
1110 HOME
1120 PRINT "DECISION CONSTANTS" : PRINT : PRINT
1130 INPUT "DC-ALPHA: "; DA : PRINT : PRINT
1140 INPUT "DC-EXTEN: "; DE
1180 REM ***CALCULATIONS***
1190 HOME : K = 0
1200 FOR N = 9 TO L - 8
1210 S(1) = 0 : S(2) = 0 : S(3) = 0 : S(4) = 0
1220 FOR M = -8 TO 8
1230 FOR P = 1 TO 4
1240 S(P) = S(P) + A(P,PEEK(30000 + N + M) - 64, M + 9)
1243 NEXT P
1245 NEXT M
1250 S(1) = S(1) - DA
1260 S(2) = S(2) - DE
1290 X = S(1) : Y = 1
1300 IF S(2) > X THEN X = S(2) : Y = 2
1310 IF S(3) > X THEN X = S(3) : Y = 3
1320 IF S(4) > X THEN X = S(4) : Y = 4
1330 IF K = 0 THEN PRINT Y : K = 1 : GOTO 1340
1335 PRINT Y; " "; : K = K + 1 : IF K = 21 THEN K = 1 : PRINT
1337 IF N = 149 THEN CALL 64477 : GET X$: HOME
1340 NEXT N
1360 INPUT "NEW DC'S? "; X$
1370 IF X$ = "Y" THEN 1100
1380 PRINT : PRINT : INPUT "ANOTHER SEQUENCE? "; X$
1390 IF X$ = "Y" THEN 1000
```

Appendix 2: Restriction sites and maps

Restriction sites were located in the two lectin sequences using the programme described in Appendix 1. Recognition sequences were obtained from a chart supplied by BRL (1983 edition). Only one enzyme per recognition sequence was examined. In the table below, numbers refer to those given in figs 3H-1 and 3I-1 for preproricin and preproagglutinin respectively; the absence of a number indicates the absence of a recognition sequence.

Enzyme	Preproricin	Preproagglutinin
AatI		
AatII		
AccI		
AflII		
AflIII		
AhaIII	1509	1506
AluI	58,78,105,200,479,756,977	58,78,105,200,753
ApaI		
AsuII		479
AvaI		
AvaII	424	421
AvaIII	233	233
AvrII		
BalI	946,1473	943,1470
BamHI	-42,852,1551	-42,849,1548
EbaI		
EbvI	532,1084,1159	240,529,1156
BclI		
BglI		
BglII	589,1141	586
BinI	-42,852,1446,1551	-42,598,849,1443,1548
EaePI		
BstEII	1591,1610	1607
BstXI		
BvuI		

Continued....

Enzyme	Preproricin	Preproagglutinin
CfrI	946,1473	943,1470
Clal	338	
DdeI	-34,397,444,476,857,1385	-34,394,441,473,854,1382
EcoPI	-26,889,1282	-26,391,886,1144
EcoP15	1082	38,1051
EcoRI		913
EcoRII	219	
EcoRV		
Fnu4HI	532,1084,1159	248,251,529,1156
FokI	-74,222,280,513,909, 1466,1484	-74,222,280,327,510, 1463,1481
GdlII		
HaeII	447	444
HaeIII	815,947,1459,1474	812,944,1456,1471
HgaI		
HgiAI		
HgiCI	822,1180	1177
HgiDI		
HgiEII	34,1601	1598
HhaI	266,448,564,776	84,254,266,445,773
HincII		
HindIII		
HinI	348,1006	348,478,1003
HinfI	539,658,917,1052,1417	477,536,655,969
HpaI		
HpaII	-99, 585,1048	-99,582
HphI	227,1591	227,907,1059,1575
KpnI	822,1180	1177
MboI	-41,590,601,853,1010,1065, 1142,1446,1547,1552	-41,587,598,850,1007,1139, 1443,1544,1549
MboII	295,914	295,910,917
MluI		
MnlI	-95,-35,-12,433,436,553, 781,1321,1384,1449,1469, 1542,1582	-95,-35,-12,389,430,550,778, 1318,1381,1446,1466,1539
NotI	563,775	83,772

Continued....

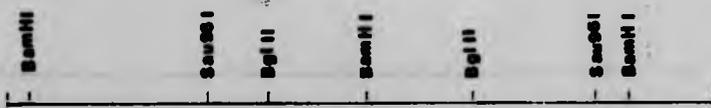
Enzyme	Preporicin	Preproagglutinin
MstII	-35	-35
NaeI		
NciI	-99,1048	-99
NcoI	1587	1584
NdeI		
NruI		
NspBII		
Nso7524 I		
Nsp7524 II		
PstI		
PvuI	1546	1543
PvuII		
RsaI	464,579,726,823,1041,1181	543,576,723,1076,1178,1325
SalI		
Sau96I	424,1459	421,1456
ScaI	463	
ScrFI	-99,219,1048	-99
SfaNI	298,512,773,967,1093,1465	43,298,509,770,1414,1462
SmaI		
SnaI		
SphI		
SstI		
SstII		
StuI		
TaqI	339,410,882,1008	480,879,920,1005,1119
ThaI	85	253
TthIII I		
TthIII II	9,137,843,1288,1294	9,137,694,840,1021,1285,1291
XbaI		
XhoI		
XhoII	-42,589,852,1141,1551	-42,586,849,1188,1548
XmaIII		
XmaI		



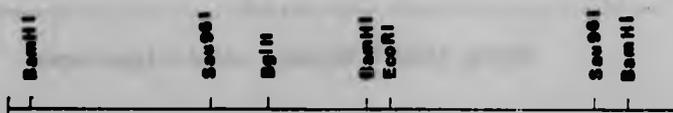
Preproricin: Unique sites



Preproagglutinin: Unique sites



Preproricin: BamHI, BglII & Sau96I



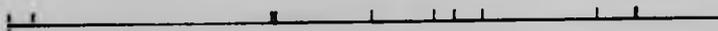
Preproagglutinin: BamHI, BglII, EcoRI & Sau96I



Preproricin: AluI sites



Preproagglutinin: AluI sites



Preproricin: Sau3AI (MboI) sites



Preproagglutinin: Sau3AI (MboI) sites

REFERENCES

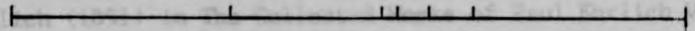
1 BE Tally & H Macvern (1976) *Plant Physiol* 50:719-6

2 RJ 5'  3'

3 G Olsson & A Pihl (1976) In *The Specificity and Action of Animal, Bacterial and Plant Toxins, Receptors & Recognition Series B* Vol 1

4 JW Lin **Preproricin: TaqI sites** & TG Berg (1970) *Nature* 227:292-3

5 G Fokstad, G Kvalheim, & Gavel, J Lofberg, S Axial, H Host & A Pihl (1984) *Cancer Res* 44:563-5

6 P Ehrlich  Vol 2 pp21-30.

7 Pergamo **Preproagglutinin: TaqI sites**

8 GA Salant (1974) *Toxicology* 1:77-102

9 B Knight (1979) *Brit Med J* 1:330-1

10 M. Craig, DJ Alderice, AM Corwin, SM Jacobs & CI Kavel (1962) US Patent no 3,060,355

11 The Times 22 May 1986 p15

12 M Larrick (1983) *Nature* 301:404

13 K Coleman (1983) *Nature* 301:404

14 I Greenfield, M Bjorn, H Jona, V Pava, G Suck, RJ Gillier & DA Tapian (1983) *Nature* 301:404-7

15 D Leung, M Coleman & M Murphy (1983) *Nature* 301:404-7

16 D Leung, M Coleman & M Murphy (1983) *Nature* 301:404-7

17 M Casarok, F Delgortcha, S Chomizer, AM Pissach, M Murphy, P Saper & P Thielis (1983) *Nature* 301:404-7

18 B Viana, C Saper-Ullman, H Gavel & J Lofberg (1983) *Brit J Biochem* 133:171-6

REFERENCES

- 1 RE Tully & H Beevers (1976) Plant Physiol 58:710-6
- 2 RJ youle & AHC Huang (1976) Plant Physiol 58:703-9
- 3 S Olsnes & A Pihl (1976) in The Specificity and Action of Animal, Bacterial and Plant Toxins. Receptors & Recognition Series B Vol 1 pp 129-173. P Cuatrecasas (Ed); Chapman & Hall
- 4 JW Lin, KY Tserng, CC Chen, LT Lin & TC Tung (1970) nature 227:292-3
- 5 O Fodstad, G Kvalheim, A Godal, J Lotsberg, S Aamdal, H Host & A Pihl (1984) Cancer Res 44:862-5
- 6 P Ehrlich (1891) in The Collected Works of Paul Ehrlich Vol 2 pp21-30. Pergamon Press
- 7 GA Balint (1974) Toxicology 2:77-102
- 8 B Knight (1979) Brit Med J 1:350-1
- 9 HL Craig, OH Alderks, AH Corwin, SH Dieke & CL Karel (1962) US Patent no 3,060,165
- 10 The Times 22 May 1984 p15
- 11 JW Larrick (1983) Nature 301:651
- 12 K Coleman (1983) Nature 302:649
- 13 L Greenfield, MJ Bjorn, G Horn, D Fong, GA Buck, RJ Collier & DA Kaplan (1983) Proc Natl Acad Sci 80:6853-7
- 14 D Leong, KD Coleman & JR Murphy (1983) Science 220:515-8
- 15 D Leong, KD Coleman & JR Murphy (1983) J Biol Chem 258:15016-20
- 16 M Kaczorek, F Dalpeyroux, N Chenciner, RE Streeck, JR Murphy, P Boquet & P Tiollais (1983) Science 221:855-8
- 17 R Vlasak, C Unger-Ullmann, G Kreil & A-M Frischauf (1983) Eur J Biochem 135:123-6

- 18 ML Gennaro & PJ Greenaway (1983) *Nucleic Acids Res* 11:3855-61
- 19 H Lockman & JB Kaper (1983) *J Biol Chem* 258:13722-6
- 20 EK Spicer & JA Noble (1982) *J Biol Chem* 257:5716-21
- 21 T Yamamoto, T Tamura & T Yokota (1984) *J Biol Chem* 259:5037-44
- 22 J Mellor, MJ Dobson, NA Roberts, MF Tuite, JS Entage, S White, PA Lowe, T Patel, AJ Kingsman & SM Kingsman (1983) *Gene* 24:1-14
- 23 European Patent Office (1982) European Patent Application no 0,060,129
- 24 K Eiklid, S Olsnes & A Pihl (1980) *Exptl Cell Res* 126:321-6
- 25 *Encyclopaedia Britannica* (1980) *Micropaedia* Vol II p627  
HK Airey Shaw (1975) *Kew Bulletin Additional Series IV*. HMSO  
HK Airey Shaw (1980) *Kew Bulletin Additional Series VIII*. HMSO
- 26 Jonah 4:6 *The Jerusalem Bible* (1968) Darton Longman & Todd
- 27 M Funatsu, K Hara, M Ishiguro, G Funatsu & R Kishikawa (1973) *Proc Japan Acad* 49:829-34
- 28 DB Cawley & LL Houston (1979) *Biochem Biophys Acta* 581:51-62
- 29 GL Nicolson, J Blaustein & ME Etzler (1974) *Biochemistry* 13:196-204
- 30 NK Gonatas, SU Kim, A Stieber & S Avrameas (1977) *J Cell Biol* 73:1-13
- 31 DB Cawley, ML Hedblom & LL Houston (1978) *Arch Biochem Biophys* 190:744-55
- 32 M Kimura, G Funatsu & M Funatsu (1977) *Agric Biol Chem* 41:1733-6
- 33 G Funatsu, S Yoshitake & M Funatsu (1978) *Agric Biol Chem* 42:501-3
- 34 S Yoshitake, G Funatsu & M Funatsu (1978) *Agric Biol Chem* 42:1267-74
- 35 G Funatsu, M Kimura & M Funatsu (1979) *Agric Biol Chem* 43:2221-4
- 36 M Kimura & G Funatsu (1981) *Agric Biol Chem* 45:265-75
- 37 L Barbieri & F Stirpe (1982) *Cancer Surveys* 1:489-520
- 38 L Barbieri, AI Falasca & F Stirpe (1984) *FEBS Lett* 171:277-9
- 39 B McKeever & R Sarma (1982) *J Biol Chem* 257:6923-5
- 40 D FitzGerald, RE Morris & CB Saeling (1980) *Cell* 21:867-73

- 18 ML Gennaro & PJ Greenaway (1983) *Nucleic Acids Res* 11:3855-61
- 19 H Lockman & JB Kaper (1983) *J Biol Chem* 258:13722-6
- 20 EK Spicer & JA Noble (1982) *J Biol Chem* 257:5716-21
- 21 T Yamamoto, T Tamura & T Yokota (1984) *J Biol Chem* 259:5037-44
- 22 J Mellor, MJ Dobson, NA Roberts, MF Tuite, JS Entage, S White, PA Lowe, T Patel, AJ Kingsman & SM Kingsman (1983) *Gene* 24:1-14
- 23 European Patent Office (1982) *European Patent Application* no 0,060,129
- 24 K Eiklid, S Olsnes & A Pihl (1980) *Exptl Cell Res* 126:321-6
- 25 *Encyclopaedia Britannica* (1980) *Micropaedia* Vol II p627  
HK Airey Shaw (1975) *Kew Bulletin Additional Series* IV. HMSO  
HK Airey Shaw (1980) *Kew Bulletin Additional Series* VIII. HMSO
- 26 Jonah 4:6 *The Jerusalem Bible* (1968) Darton Longman & Todd
- 27 M Funatsu, K Hara, M Ishiguro, G Funatsu & R Kishikawa (1973) *Proc Japan Acad* 49:829-34
- 28 DB Cawley & LL Houston (1979) *Biochem Biophys Acta* 581:51-62
- 29 GL Nicolson, J Blaustein & ME Etzler (1974) *Biochemistry* 13:196-204
- 30 NK Gonatas, SU Kim, A Stieber & S Avrameas (1977) *J Cell Biol* 73:1-13
- 31 DB Cawley, ML Hedblom & LL Houston (1978) *Arch Biochem Biophys* 190:744-55
- 32 M Kimura, G Funatsu & M Funatsu (1977) *Agric Biol Chem* 41:1733-6
- 33 G Funatsu, S Yoshitake & M Funatsu (1978) *Agric Biol Chem* 42:501-3
- 34 S Yoshitake, G Funatsu & M Funatsu (1978) *Agric Biol Chem* 42:1267-74
- 35 G Funatsu, M Kimura & M Funatsu (1979) *Agric Biol Chem* 43:2221-4
- 36 M Kimura & G Funatsu (1981) *Agric Biol Chem* 45:265-75
- 37 L Barbieri & F Stirpe (1982) *Cancer Surveys* 1:489-520
- 38 L Barbieri, AI Falasca & F Stirpe (1984) *FEBS Lett* 171:277-9
- 39 B McKeever & R Sarma (1982) *J Biol Chem* 257:6923-5
- 40 D FitzGerald, RE Morris & CB Saelinger (1980) *Cell* 21:867-73

- 41 RD Sekura, F Fish, CR Manclark, B Meade & Y Zhang (1983)  
J Biol Chem 258:14647-51
- 42 J Dwyer & V Bloomfield (1982) Biochemistry 21:3227-31
- 43 R Montesano, J Roth, A Robert & L Orci (1982) Nature 296:651-3
- 44 LM Roberts & JM Lord (1981) Eur J Biochem 119:31-41
- 45 AG Butterworth & JM Lord (1983) Eur J Biochem 137:57-65
- 46 S Olsnes & A Pihl (1982) in Molecular Action of Toxins and  
Viruses (Cohen & van Heyningen Eds) pp 51-105. Elsevier.
- 47 JM Lord, FI Lamb & LM Roberts (1984) Oxford Surv Plant Mol  
Cell Biol 1:85-101
- 48 D Vazquez (1979). Inhibitors of Protein Biosynthesis.  
Vol 30 of Molecular Biology, Biochemistry and Biophysics  
( Ed Kleinzeller). Springer-Verlag.
- 49 S Olsnes, E Saltvedt & A Pihl (1974) J Biol Chem 249:803-10
- 50 H Bull, SSL Li, E Fowler & TTS Lin (1980) Int J Peptide  
Protein Res 16:208-18
- 51 G Funatsu & M Funatsun (1977) Agric Biol Chem 41:1211-5
- 52 G Funatsu, T Mise, H Matsuda & M Funatsu (1978) Agric Biol Chem  
42:851-9
- 53 CH Wei & C Koh (1978) J Mol Biol 123:707-11
- 54 LG Gurtler & J Horstmann (1973) Biochim Biophys Acta 295:582-94
- 55 P Langridge & G Feix (1983) Cell 34:1015-22
- 56 A Surolia, BK Bachhawat & SK Podder (1978) Ind J Biochem Biophys  
15:248p50
- 57 G Lutsch, F Noll, P Ziska, A Kindt & H Franz (1984)  
FEBS Lett 170:335-8
- 58 S Olsnes, E Saltvedt & A Pihl (1974) J Biol Chem 249:803-10
- 59 E Saltvedt (1976) Biochim Biophys Acta 451:536-48
- 60 G Funatsu, S Ueno & M Funatsu (1977) Agric Biol Chem 41:1737-43
- 61 S Olsnes, K Refsnes, TB Christensen & A Pihl (1975)  
Biochim Biophys Acta 405:1-10
- 62 JJ Marchalonis & JK Weltman (1971) Comp Biochem Physiol 38B:609-25

- 63 S Olsnes & E Saltvedt (1975) *J Immunol* 114:1743-8
- 64 RD Marshall (1974) *Biochem Soc Symp* 40:17-26
- 65 SC Hubbard & RJ Ivatt (1981) *Ann Rev Biochem* 50:555-83
- 66 V Gross, TA Tran-Thi, K Vosbeck & PC Heinrich (1983)  
*J Biol Chem* 258:4032-6
- 67 H Ishihara, N Takahashi, S Oguri & S Tejima (1979)  
*J Biol Chem* 254:10715-9
- 68 A Vitale, TG Warner & MJ Chrispeels (1984) *Planta* 160:256-63
- 69 A Masaki & IJ Goldstein (1977) *J Biol Chem* 252:6995-9
- 70 A Vitale, A Ceriotti, R Bollini & MJ Chrispeels (1984) *Eur J Biochem* 141:97-104
- 71 DA Lappi, W Kapmeyer, JM Beglau & NO Kaplan (1978)  
*Proc Natl Acad Sci (USA)* 75:1096-1100
- 72 JE Villafrance & JD Robertus (1981) *J Biol Chem* 256:554-6
- 73 G Funatsu & M Funatsu (1964) *J Biochem (Tokyo)* 55:587
- 74 CH Wei (1973) *J Biol Chem* 248:3745-7
- 75 K Shimazaki, EF Walborg, G Nerl & B Jirgensons (1975)  
*Arch Biochem Biophys* 169:731-6
- 76 K Shelley & A McPherson (1980) *Arch Biochem Biophys* 202:431-41
- 77 S Olsnes, AM Pappenheimer & RA Meren (1973) *J Immunol* 113:842-7
- 78 SSL Li (1980) *Experientia* 36:524-7
- 79 MA Schuler, JJ Doyle & RN Beachy (1983) *Plant Mol Biol* 2:119-27
- 80 DJ Gifford, JS Greenwood & JD Bewley (1982) *Plant Physiol* 69:1471-8
- 81 LM Roberts & JM Lord (1981) *Planta* 152:420-7
- 82 LM Roberts & JM Lord (1979) *J Exptl Bot* 30:739-49
- 83 RB Goldberg, G Hoschek & LO Vodkin (1983) *Cell* 33:465-75
- 84 IM Evans, JA Gatehouse, RRD Croy & D Boulter (1984) *Planta* 160:559-68
- 85 RJ Youle & AHC Huang (1978) *Plant Physiol* 61:1040-2
- 86 FS Sharief & SSL Li (1982) *J Biol Chem* 257:4753-9
- 87 BA Larkins & WJ Hurkman (1978) *Plant Physiol* 62:256-63

- 88 B Burr & FA Burr (1976) Proc Natl Acad Sci (USA) 73:515-9
- 89 WJ Hurkman & L Beevers (1982) Plant Physiol 69:1414-7
- 90 CF Phelps (1980) in The Enzymology of Post-Translational  
Modification of Proteins (Eds Freedman & Hawkins) Academic  
Press pp 136-168
- 91 L Beevers (1982) in Nucleic Acids & Proteins in Plants Vol I  
(Eds D Boulter & B Parthier). Encyclopedia of Plant Physiology  
New Series Vol 14A pp169-188. Springer-Verlag.
- 92 MJ Chrispeels (1983) Planta 158:140-51
- 93 RRD Croy, JA Gatehouse, IM Evans & D Boulter (1980) Planta 148:49-56
- 94 TJV Higgins, PM Chandler, G Zurawski, SC Button & D Spencer  
(1983) J Biol Chem 258:9544-9
- 95 ML Crouch, KM Tenbarger, AE Simon & R Ferl (1983) J Mol Appl Genet  
2:273-83
- 96 R Chance, RM Ellis & WW Bromer (1968) Science 161:165-7
- 97 J Finidori, Y Laperche, R Haguenaer-Tsapis, R Barouki,  
G Guellaen & J Hanoune (1984) J Biol Chem 259:4687-90
- 98 F Stirpe, S Olsnes & A Pihl (1980) J Biol Chem 255:6947-53
- 99 L Barbieri, E Lorenzoni & F Stirpe (1979) Biochem J 182:633-5
- 100 H Lis & N Sharon (1981) in The Biochemistry of Plants: A  
Comprehensive Treatise (Ed A Marcus) Vol 6 pp372-449
- 101 M Funatsu, K Hara, M Ishiguro, G Funatsu, H Otsuka & M Ide (1973)  
Proc Japan Acad 49:835-9
- 102 DJ Janzen (1983) in Physiological Plant Ecology Vol III (Ed O Lange)  
Encyclopedia of Plant Physiology New Series Vol 12C pp 625-656.  
Springer-Verlag.
- 103 RJ Youle & AHC Huang (1978) Plant Physiol 61:13-16
- 104 S Olsnes, R Heiberg & A Pihl (1973) Mol Biol Rep 1:15-20
- 105 LJG Haas-Kohn, AAJ Lugnier, O Tiboni, O Ciferri & G Dirheimer  
(1980) Biochem Biophys Res Commun 97:962-7

- 106 SM Harley & H Beevers (1982) Proc Natl Acad Sci (USA) 79:5935-8
- 107 B Parisi & O Ciferri (1966) Biochemistry 5:1638-45
- B Parisi, G Milanese, JL van etten, A Perani & O Ciferri (1967)  
J Mol Biol 28:295-309
- 108 MG Battelli, E Lorenzoni & F Stirpe (1984) J Exptl Bot 35:882-9
- 109 RA Owens, G Bruening & RJ Shepherd (1973) Virology 56:390-3
- 110 AAJ Lugnier, H Kuntz & G Dirheimer (1979) FEBS Letters 66:202-5
- 111 M Greco, L Montanaro, F Novello, C Saccone, S Sperti  
& F Stirpe (1974) Biochem J 142:695-7
- 112 OG Wilde, S Boguslawski & LL Houston (1979) Biochem Biophys  
Res Commun 91:1082-8
- 113 S Olsnes, C Fernandez-Puentes, L Carrasco & D Vazquez (1975)  
Eur J Biochem 60:281-8
- 114 O Fodstad & S Olsnes (1977) Eur J Biochem 74:209-15
- 115 S Benson, S Olsnes & A Pihl (1975) Eur J Biochem 59:573-80
- 116 S Sperti, L Montanaro, A Mattioli & F Stirpe (1973) Biochem J  
136:813-5
- 117 K Onosaki, H Hayatsu & T Ukita (1975) Biochim Biophys Acta 407:99-107
- 118 J Skorve, KA Abraham, S Olsnes & A Pihl (1977) Eur J Biochem 79:559-64
- 119 S Olsnes & A Pihl (1972) Nature 238:459-61
- 120 S Olsnes & A Pihl (1972) FEBS Letters 20:327-9
- 121 L Carrasco, C Fernandez-Puentes & D Vazquez (1975) Eur J Biochem  
54:499-503
- 122 L Montanaro, S Sperti & F Stirpe (1973) Biochem J 136:677-83
- 123 CH Wei & C Koh (1978) J Mol Biol 123:707-11
- 124 K Hara, M Ishiguro, G Funatsu & M Funatsu (1974) Agric Biol Chem 38:65-70
- 125 L Carrasco, C Fernandez-Puentes & D Vazquez (1975) Eur J Biochem  
54:499-503

- 126 RD Nolan, H Grasmuk & J Drews (1976) Eur J Biochem 64:69-75
- 127 AO Fuller & E Fowler (1980) J Cell Biol 87:273a
- 128 O Fodstad, S Olsnes & A Pihl (1976) Br J Cancer 34:418-25
- 129 TL Rodes III & JD Irvin (1981) Biochim Biophys Acta 652:160-7
- 130 LL Houston (1978) Biochem Biophys Res Commun 85:131-9
- 131 M Zamboni, G Battelli, L Montanaro & S Sperti (1981)  
Biochem J 194:1015-7
- 132 G Sacco, K Drickamer & IG Wool (1983) J Biol Chem 258:5811-8
- 133 WK Roberts & TS Stewart (1979) Biochemistry 18:2615-21
- 134 G Funatsu, S Yoshitake & M Funatsu (1977) Agric Biol Chem 41:1225-31
- 135 C Zentz, JP Frenoy & H Bourrillon (1977) FEBS Letters 81:23-27
- 136 JU Baenziger & D Fiete (1979) J Biol Chem 254:9795-9
- 137 SK Podder (1974) Eur J Biochem 44:151-60
- 138 MI Khan & A Surolia (1982) Eur J Biochem 126:495-500
- 139 RJ Youle, GJ Murray & DM Neville Jr (1981) Cell 23:551-9
- 140 L Simeral, W Kapmeyer, WP MacConnell & NO Kaplan (1980) J Biol Chem  
255:11098-11101
- 141 MS Herrman & WD Behnke (1980) Biochim Biophys Acta 621:43-52
- 142 I Matsumoto, A Jimbo, Y Mizuno, N Seno & RW Jeanloz (1983)  
J Biol Chem 258:2886-91
- 143 S Olsnes & E Saltvedt (1975) J Immunol 114:1743-8
- 144 K Refsnes, S Olsnes & A Pihl (1974) J Biol Chem 249:3557-62
- 145 GL Nicolson (1974) Nature 251:628-30
- 146 GL Nicolson, M Lacorbiere & TR Hunter (1975) Cancer Res 35:144-55
- 147 S Olsnes, K Refsnes & A Pihl (1974) Nature 249:627-31
- 148 LL Houston (1982) J Biol Chem 257:1532-9
- 149 A Surolia, BK Bachhawat & SK Podder (1975) Nature 257:802-4
- 150 A Surolia & BK Bachhawat (1978) Biochem Biophys Res Commun 83:779-85
- 151 G Kayser, E Goormaghtigh, M Vandenbranden & JM Ruysschaert (1981)  
FEBS Letters 127:207-10

- 152 S van Heyningen (1982) in Molecular Action of Toxins and Viruses, pp 169-190 (Eds Cohen & van Heyningen). Elsevier.
- 153 NK Gonatas, SU Kim, A Steiber & S Avrameas (1977) J Cell Biol 73:1-13
- 154 NK Gonatas, A Steiber, SU Kim, DI Graham & S Avrameas (1975) Exptl Cell Res 94:426-31
- 155 J Gonatas, A Stieber, S Olsnes & NK Gonatas (1980) J Cell Biol 87:579-88
- 156 GL Nicolson, JR Smith & R Hyman (1978) J Cell Biol 78:565-76
- 157 K Sandvig & S Olsnes (1982) J Biol Chem 257:7504-13
- 158 K Sandvig, S Olsnes & A Pihl (1978) Eur J Biochem 82:13-23
- 159 E Mekada, T Uchida & Y Okada (1981) J Biol Chem 256:1225-8
- 160 S Olsnes & K Sandvig (1981) in Receptor-Mediated Binding and Internalisation of Toxins and Hormones (Eds JL Middlebrook & LD Kohn) pp81-94, Academic Press.
- 161 FR Maxfield, J Schlessinger, Y Schachter, I Pastan & MC Willingham (1978) Cell 14:805-810
- 162 P Boquet, MS Silverman, AM Pappenheimer Jr & WB Vernon (1976) Proc Natl Acad Sci (USA) 73:4449-53
- 163 P Bacha, JR Murphy & S Reichlin (1983) J Biol Chem 258:1565-70
- 164 P Boquet & AM Pappenheimer (1976) J Biol Chem 251:5770-8
- 165 BL Kagan, A Finkelstein & M Colombini (1981) Proc Natl Acad Sci (USA) 78:4950-4
- 166 K Sandvig & S Olsnes (1981) J Biol Chem 256:9068-76
- 167 WK Adair & S Kornfeld (1974) J Biol Chem 249:4696-704
- 168 B Ishada, DB Cawley, K Reue & BJ Wisnieski (1983) J Biol Chem 258:5933-7

- 169 T Uchida, E Mekada & Y Okada (1980) *J Biol Chem* 255:6687-93
- 170 DG Gilliland, J Mannhalter & RJ Collier (1981) in  
Receptor-Mediated Binding and Internalisation of Toxins  
and Hormones (Eds JL Middlebrook & LD Kohn) pp311-327.  
Academic Press.
- 171 PE Thorpe & WCJ Ross (1982) *Immunol Rev* 62:119-157
- 172 RT Dean, W Jessup & CR Roberts (1984) *Biochem J* 217:27-40
- 173 P Matile (1982) in *Nucleic Acids & Proteins in Plants Vol I*  
(Eds D Boulter & B Parthier) *Encyclopedia of Plant Physiology*  
*New Series Vol 14A*, pp169-188. Springer-Verlag.
- 174 S Olsnes & E Saltvedt (1975) *J Immunol* 114:1743-8
- 175 K Sandvig & S Olsnes (1982) *J Biol Chem* 257:7495-503
- 176 PNT Urwin & PD Ennis (1984) *Nature* 307:609-13
- 177 MW Wooten & RW Wrenn (1984) *FEBS Letters* 171:183-6
- 178 M Yokoyama, F Nishiyama, N Kawai & H Hirano (1980)  
*Exptl Cell Res* 125:47-53
- 179 MJ Capaldi, MJ Dunn, CA Sewry & V Dobowitz (1984) *J Neuro Sci*  
63:129-42
- 180 B Batard, H Debray, JP Kerckaert & G Biserte (1977)  
*FEBS Letters* 80:35-40
- 181 M Neukirch, V Moennig & B Liess (1981) *Arch Virol* 69:287-90
- 182 D Tsao & YS Kim (1981) *J Biol Chem* 256:4947-50
- 183 T Oelmann & J Forbes (1981) *Arch Biochem Biophys* 209:362-70
- 184 T Irimura & GL Nicolson (1984) *Cancer Res* 44:791-8
- 185 M Shiba, T Ohiwa & AJP Klein-Szanto (1984) *J Natl Cancer Inst*  
72:43-51
- 186 RA Roth, BA Maddux & KY Wong (1981) *J Biol Chem* 256:5350-4
- 187 R Arnon & M Sela (1982) *Immunol Rev* 62:5-27

- 188 RA De Weger, HFJ Dullens & W Den Otter (1982) *Immunol Rev* 62:29-45
- 189 V Raso (1982) *Immunol Rev* 62:93-117
- 190 G Svet-Moldavsky & V Hamburg (1964) *nature* 202:303-4
- 191 FL Moolten, EM Schreiber & SH Zajdel (1982) *Immunol Rev* 62:47-73
- 192 FL Moolten & SR Cooperbrand (1970) *Science* 169:68-70
- 193 G Kohler & C Milstein (1975) *Nature* 256:495-7
- 194 PE Thorpe, F Stirpe, JAG Bremner Jr, ANF Brown & SI Detre (1982)  
Imperial Cancer Research Fund Scientific Report for 1982  
pp 83-86 and p 242
- 195 HE Blythman, P Casellas, O Gros, P Gros, FK Jansen, F Paolucci,  
B Pau & H Vidal (1981) *Nature* 290:145-6
- 196 DB Cawley, HR Herschman, DG Gilliland & RJ Collier (1980) *Cell*  
22:563-70
- 197 KA Krolick, C Villemez, P Isakson, JW Uhr & ES Vitetta (1980)  
*Proc Natl Acad Sci (USA)* 77:5419-23
- 198 DP McIntosh, DC Edwards, AJ Cumber, GD Parnell, CJ Dean, WCJ Ross,  
& JA Forrester (1983) *FEBS Letters* 164:17-20
- 199 ES Vitetta, W Cushley & JW Uhr (1983) *Proc Natl Acad Sci (USA)*  
80:6332-5
- 200 V Raso & T Griffin (1981) *Cancer Res* 41:2073-8
- 201 EF Neufeld & G Ashwell (1980) in *The Biochemistry of Glycoproteins  
and Proteoglycans* (Ed WJ Lennarz) Ch 6 pp 241-266. Plenum Press.
- 202 D Hanahan (1983) *J Mol Biol* 166:557-80
- 203 HG Birnboim & J Doly (1979) *Nucleic Acids Res* 7:1513-23
- 204 D Clewell (1972) *J Bact* 110:667-72
- 205 *Molecular Cloning: A Laboratory Manual* (Ed T Maniatis).  
Cold Spring Harbour 1982.
- 206 M Bazaral & DR Helinski (1968) *J Mol Biol* 36:185-94
- 207 LM Roberts (1981) PhD Thesis, University of Bradford.
- 208 RJ Mans & GD Novelli (1961) *Arch Biochem Biophys* 94:48-53

- 209 G Bray (1960) Anal Biochem 1:279-285
- 210 MW McDonnell, MN Simon, & FW Studier (1977) J Mol Biol 110:119-46
- 211 H Lehrach, D Diamond, JM Wozney & H Boedtker (1977)  
Biochemistry 16:4743-51
- 212 U Laemmli (1970) Nature 227:680-5
- 213 WM Bonner & RA Laskey (1974) Eur J Biochem 46:83-88
- 214 J Chamberlain (1979) Anal Biochem 98:132-5
- 215 F Sanger & AR Coulson (1975) J Mol Biol 94:441-8
- 216 MD Biggin, TJ Gibson & GF Hong (1983) Proc Natl Acad Sci (USA)  
80:3963-5
- 217 W Ansorge & L de Maeyer (1980) J Chromatog 202:45-53
- 218 L Bowden-Bonnett & JM Lord (1979) Plant Physiol 63:769-73
- 219 B Dobberstein, H Garoff & G Warren (1970) Cell 17:759-69
- 220 G Buell, MP Wockens, F Payvar & RT Schimke (1978) J Biol Chem  
253:2471-82
- 221 E Retzel, MS Collett & AJ Faras (1980) Biochemistry 19:513-8
- 222 A Efstratiadis, FC Kafatos, AM Maxam & T Maniatis (1976)  
Cell 7:279-88
- 223 MP Wickens, GN Buell & RT Schimke (1978) J Biol Chem 253:2483-95
- 224 G Deng & R Wu (1981) Nucleic Acids Res 9:4173-88
- 225 F Bolivar, RL Rodriguez, PJ Greene, MC Betlach, HL Heyneker,  
HW Boyer, JH Crosa & S Falkow (1977) Gene 2:95-113
- 226 M Grunstein & DS Hogness (1975) Proc Natl Acad Sci (USA)  
72:3961-5
- 227 D Kau & J Myers (1976) Proc Natl Acad Sci (USA) 73:2191-5

- 228 PR Shank, JG Cohen, HE Varmus, KR Yamamoto & GM Ringold (1978)  
Proc Natl Acad Sci (USA) 75:2112-6
- 229 SV Suggs, T Hiroshi, T Miyake, EH Kawashima, MJ Johnson,  
K Itakura & RB Wallace (1981) in Developmental Biology using  
Purified Genes; ICN-ICLA Symposium of Molecular and Cellular  
Biology (Eds DD Brown & CF Fox). Academic Press; Vol 23 pp683-93
- 230 RP Ricciardi, JS Miller & BE Roberts (1979) Proc Natl Acad Sci (USA)  
76:4927-31
- 231 J Vieira & J Messing (1982) Gene 19:259-68
- 232 J Messing & J Vieira (1982) Gene 19:269-76
- 233 J Messing (1979) Recombinant DNA Technical Bulletin 2:43-48
- 234 S Anderson (1981) Nucleic Acids Res 9:3015-27
- 235 F Sanger, AR Coulson, BG Barrell, AJH Smith & BA Roe (1980)  
J Mol Biol 143:161-78
- 236 F Sanger, S Nicklen & AR Coulson (1977) Proc Natl Acad Sci (USA)  
74:5463-7
- 237 AM Maxam & W Gilbert (1980) Meth Enzymol 65:499-560
- 238 L Hall, JE Laird, JC Pascall & RK Craig (1984) Eur J Biochem  
138:585-9
- 239 E Southern (1979) Meth Enzymol 68:152-76
- 240 FWJ Rigby, M Dieckmann, C Rhodes & P Berg (1977) J Mol Biol 113:237-51
- 241 LM Roberts (1981) PhD Thesis, University of Bradford
- 242 H Land, M Grez, H Hauser, W Lindermaier, & G Schutz (1981)  
Nucleic Acids Res 9:2251-66
- 243 MP Wickens, GN Buell & RT Schimke (1978) J Biol Chem 253:2483-95
- 244 MV Norgard, K Keem & JJ Monahan (1978) Gene 3:279-92
- 245 H Boyer & Rouland-Dussoix (1969) J Mol Biol 41:459-72
- 246 MR Green, T Maniatis & DA Melton (1983) Cell 32:681-94
- 247 J Mellor, MJ Dobson, NA Roberts, MF Tuite, JS Emtage, S White,  
PA Lowe, T Patel, AJ Kingsman & SM Kingsman (1983) Gene 24:1-14

- 248 EC Conley & JR Saunders (1984) *Mol Gen Genet* 194:211-8
- 249 J Williams, TC Elleman, IB Kingston, AG Wilkins & KA Kuhn  
(1982) *Eur J Biochem* 122:297-303
- 250 J Drouin (1980) *J Mol Biol* 140:15-34
- 251 F Sanger, AR Coulson, GF Hong, DF Hill & GB Petersen (1982)  
*J Mol Biol* 162:729-73
- 252 NM Gough, EA Webb, S Cory & JM Adams (1980) *Biochemistry* 19:2702-10
- 253 N Battula & LA Loeb (1975) *J Biol Chem* 250:4405-9
- 254 J Leggett Bailey (1967) in *Techniques in Protein Chemistry*. Elsevier.  
LR Croft (1980) in *Handbook of Protein Sequence Analysis*. Wiley.
- 255 J Kyte & RF Doolittle (1982) *J Mol Biol* 157:105-32
- 256 L Dure III, C Chlan & GA Galau (1983) in *Structure & Function  
of Plant Genomes* (Eds O Ciferri & L Dure III). Plenum.  
NATO/ASI Advanced Series A Vol 63 pp 113-121
- 257 AC Brinegar & DM Peterson (1982) *Plant Physiol* 70:1767-9
- 258 JL Slightom, SM Sun & TC Hall (1983) *Proc Natl Acad Sci (USA)*  
80:1897-901
- 259 JJ Hemperley, KE Mostov & BA Cunningham (1982) *J Biol Chem*  
257:7903-9
- 260 TJV Higgins, PM Chandler, D Spencer, MJ Chrispeels  
& G Zurawski (1983) in *Structure & Function of Plant Genomes*.  
NATO ASI Advanced Series A Vol 63 pp 93-99. Plenum Press.
- 261 N Ereken-Tuner, JD Richter & NC Nielsen (1982) *J Biol Chem*  
257:4016-8
- 262 LO Vodkin, PR Rhodes & RB Goldberg (1983) *Cell* 34:1023-31
- 263 G von Heijne (1983) *Eur J Biochem* 133:17-21
- 264 L Edens, L Heslings, R Klok, AM Ledeboer, J Maat, MY Toonen,  
C Visser & CT Verrips (1982) *Gene* 18:1-12

- 265 D Geraghty, MA Peifer, I Rubinstein & J Messing (1981)  
Nucleic Acids Res 9:5163-74
- 266 LM Hoffman, Y Ma & RF Barker (1982) Nucleic Acids Res 10:7819-28
- 267 GW Lycett, RRD Croy, AH Shirsat & D Boulter (1984)  
Nucleic Acids Res 12:4493-506
- 268 GW Lycett, AJ Delauney, JA Gatehouse, J Gilroy, RRD Croy  
& D Boulter (1983) Nucleic Acids Res 11:2367-80
- 269 JA Rafalski, K Scheets, M Metzler, DM Peterson, C Hedgcoth  
& DG Soll (1984) EMBO J 3:1409-15
- 270 G von Heijne (1984) J Mol Biol 173:243-51
- 271 AV Finkelstein, P Bendzko & TA Rapoport (1983) FEBS Lett 161:176-9
- 272 NC Nielsen (1984) Phil Trans Roy Soc Lond B304:287-96
- 273 D Boulter (1984) Phil Trans Roy Soc Lond B304:323-32
- 274 A van der Straten, A Herzog, P Jacobs, T Cabezon & A Bollen  
(1983) EMBO J 2:1003-7
- 275 JC Rogers (1983) J Biol Chem 258:8169-74
- 276 MD Marks & BA Larkins (1982) J Biol Chem 257:9976-83
- 277 R Grantham, C Gautier, M Gouy, M Jacobzone & R Mercier  
(1981) Nucleic Acids Res 9:r43-r74
- 278 JJ Hyldig-Nielsen, EO Jensen, K Paludan, O Wiborg, R Garrett,  
P Jorgensen & KA Marcker (1982) Nucleic Acids Res 10:689-701
- 279 WL Gerlach, AJ Pryor, ES Dennis, RJ Ferl, MM Sachs & WJ Peacock  
(1983) Proc Natl Acad Sci (USA) 79:2981-5
- 280 N Brisson & DPS Verma (1982) Proc Natl Acad Sci (USA) 79:4055-9
- 281 DM Shah, RC Hightower & RB Meagher (1982) Proc Natl Acad Sci (USA)  
79:1022-6
- 282 O Wiborg, JJ Hyldig-Nielsen, EO Jensen, K Paludan & RA Marcker  
(1982) Nucleic Acids Res 10:3487-94
- 283 JC Rogers & C Milliman (1983) J Biol Chem 258:8169-74
- 284 U Reimold, M Kroger, F Kreuzaler & K Hahlbrock (1983) EMBO J  
2:1801-5

- 285 WJ Stieckema, CF Wimpee & EM Tobin (1983) *Nucleic Acids Res* 11:8051-61
- 286 D Bartels & RD Thompson (1983) *Nucleic Acids Res* 11:2961-77
- 287 P Dhaese, H De Greve, J Gielen, J Seurinck, M Van Montagu & J Schell (1983) *EMBO J* 2:419-26
- 288 R Nussinov (1981) *J Mol Biol* 149:125-131
- 289 R Nussinov (1981) *J Biol Chem* 256:8458-62
- 290 SM Sun, JL Slightom & TC Hall (1981) *Nature* 289:37-41
- 291 GW Lycett, AY Delauney & RRD Croy (1983) *FEBS Letters* 153:43-46
- 292 M Kozak (1978) *Cell* 15:1109-23
- 293 M Kozak (1980) *Cell* 22:7-8
- 294 PK Bandyopadhyay & HM Temin (1984) *Mol Cell Biol* 4:743-8
- 295 CC Liu, CC Simonsen & AD Levinson (1984) *nature* 309:82-85
- 296 Y Yamada, M Mudryj & B de Crombrughe (1983) *J Biol Chem* 258:14914-9
- 297 DJ Chin, G Gil, DW Russell, L Liscum, KL Luskey, SK Basu, H Okoyama, P Berg, JL Goldstein & MS Brown (1984) *Nature* 308:613-7
- 298 M Kozak (1981) *Nucleic Acids Res* 9:5233-52
- 299 M Kozak (1984) *Nature* 308:241-6
- 300 M Kozak (1984) *Nucleic Acids Res* 12:857-72
- 301 J Messing (1983) in *Genetic Engineering of Plants: An Agricultural Perspective* (Ed T Kosuge). Plenum. Pp 211-227
- 302 A Spena, A Viotti & V Pirrotta (1983) *J Mol Biol* 169:799-811
- 303 G Heidecker & J Messing (1983) *Nucleic Acids Res* 11:4891-906
- 304 SL Berry-Lowe, TD McKnight, DM Shah & RB Meagher (1982) *J Mol Appl Genet* 1:483
- 305 J Kohli & H Grosjean (1981) *Mol Gen Genet* 182:430-9
- 306 RN Beachy, JJ Doyle, BF Ladin & MA Schuler (1983) in *Structure & Function of Plant Genomes*. (Eds O Ciferri & L Dure III) NATO ASI Series A Vol 62 pp 101-112. Plenum, NY.

- 307 RRD Croy, GW Lycett, JA Gatehouse, JN Yarwood & D Boulter (1982)  
Nature 295:76-79
- 308 E Hofer & JE Darnell (1981) Cell 23:585-93
- 309 SM Berget (1984) Nature 309:179-82
- 310 M Tosi, RA Young, O Hagenbuchle & U Schibler (1981)  
Nucleic Acids Res 9:2313-23
- 311 Y Furutani, Y Morimoto, S Shibahara, M Noda, H Takahashi,  
T Hirose, M Asai, S Inuyama, H Hayashida, T Miyata & S Numa (1983)  
Nature 301:537-40
- 312 H Kakidani, Y Furtani, H Takahashi, M Noda, Y Morimoto,  
T Hirose, M Asai, S Inayama, S Nakanishi & S Numa (1982)  
Nature 298:245-9
- 313 C Benoist, K O'Hare, R Breathnach & P Chambon (1980)  
Nucleic Acids Res 8:127-42
- 314 A McPherson (1980) J Biol Chem 255:10472-80
- 315 M Kimura (1968) Nature 217:624-6
- 316 G Walburg & BA Larkins (1983) Plant Physiol 72:161-5
- 317 GA Galau, CA Chlan & L Dure III (1983) Plant Mol Biol 2:189-206
- 318 BJ Mifflin, S Rahman, M Kreis, BG Forde, L Blanco & PR Shewry  
(1983) in Structure & Function of Plant Genomes (Eds O Ciferri  
& L Dure III). NATO ASI Series A Vol 63 pp 85-92. Plenum, NY.
- 319 H Yamagata, T Sugimoto, K Tanaka & Z Kasai (1982)  
Plant Physiol 70:1094-1100
- 320 M Go (1983) Proc Natl Acad Sci (USA) 80:1964-8
- 321 M Go (1981) Nature 291:90-92
- 322 A Rashin (1981) Nature 291:85-87
- 323 R Casey & C Domoney (1984) Phil Trans Roy Soc Lond B304:349-58
- 324 TC Huffaker & FW Robbins (1983) Proc Natl Acad Sci (USA)  
80:7466-70
- 325 BA Cunningham, JJ Hamperley, TP Hopp & GM Edelman (1979)  
Proc Natl Acad Sci (USA) 76:3218-3222

- 326 Apple II Reference Manual (1979) Apple Computers Inc  
Part no A2L0001A
- 327 Applesoft BASIC Programming Reference Manual (1978)  
Apple Computer Inc Part no A2L0006
- 328 Apple DOS Manual (1980) Apple Computer Inc Part no A2L0036
- 329 Apple Interface Manual (1982) Apple Computer Inc Part no A2L0045
- 330 Radio Shack CGP-115 Printer-Plotter Manual
- 331 J Garnier, DJ Osgudthorpe & B Robson (1978) J Mol Biol 120:97-120
- 332 W Kabsch & C Sander (1983) FEBS Letters 155:179-82
- 333 PY Chou & GD Fasman (1974) Biochemistry 13:222-45
- 334 WR Taylor & JM Thornton (1984) J Mol Biol 173:487-514
- 335 FA Quioco & NK Vyas (1984) Nature 310:381-6
- 336 A Colman (1984) in Transcription & Translation: A Practival  
Approach (Eds BD Hames & SJ Higgins) Chapter 2. IRL Press.
- 337 S Smeekens (1984) in Molecular Form & Function of the Plant  
Genome p A12 (Programme & Abstracts of the NATO ASI & FEBS  
Advanced Course, Renesse, Holland, July 1984)
- 338 AR Cashmore (1984) Proc Natl Acad Sci (USA) 81:2960-4
- 339 TC Huffaker & PW Robbins (1983) Proc Natl Acad Sci (USA) 80:7466-70