BIOTECHNOLOGY INFORMATION '86

R. Wakeford

IRL PRESS
OXFORD · WASHINGTON DC

IRL PRESS
OXFORD · WASHINGTON DC

BIOTECHNOLOGY INFORMATION '86

Proceedings of a Conference
held at Sussex University, UK,
September 22 – 25, 1986

Edited by
R.Wakeford

IRL Press Ltd
PO Box 1
Eynsham
Oxford OX8 1JJ
England

British Library Cataloguing in Publication Data

Biotechnology Information '86: proceedings of a conference
held at Sussex University, UK, September 22–25, 1986.
1. Information storage and retrieval systems—Biotechnology
2. On-line bibliographic searching
I. Wakeford,R.
025'.066606 Z699.5.B5/

ISBN 1 85221 009 5

Printed by Information Printing Ltd, Oxford, England.

Introduction

The dependence of biotechnology on diverse information sources makes it perhaps unique among high technology industries today. The first Biotechnology Information conference, held in September 1986 in Brighton, UK set out to map this variety of material and to examine the many developments that are now taking place. Some fundamental features of biotechnology make for a close linkage with information systems—molecular sequences are a linear code and closely analogous to the digital strings used within computers; commercial products all eventually derive from protein and proteins number in the tens of thousands; the profusion of these potential products leads to a corresponding variety of new companies. All these features threaten information overload and stimulate industry's appetite for more efficient intelligence.

Biotechnology Information'86 brought together an unusually diverse group of speakers and delegates to discuss each link of the information chain; from the capture of raw, experimental data to the evaluation of investment decisions. The composition of the audience showed that even if information is not a discipline in its own right, it is an inescapable feature of everybody's activity.

The first step, the gathering of experimental data was represented by a look at some recent applications of gel electrophoresis. In these proceedings, Robert Silman (St Bartholomew's Hospital Medical School) describes the characterization of cells and microbial strains by a system that requires a sophisticated image detector and computer processing to deal with the data that emerge from the gel. Data are collected from such automated equipment and by hard manual labour in laboratories all over the world but it is of limited use when it is locked in local collections. Their true value is only realised when the data are disseminated and made available from public databanks, be they electronic or paper based. The topic that drew the most attention at this meeting was the management of databanks—the collection, validation, presentation and distribution of factual information. Howard Bilofsky (Bolt, Beranek and Newman Inc.) and Winona Barker (National Biomedical Research Foundation) describe the GenBank and Protein Information Resource molecular sequence collections. In the field of cell culture collections, Micah Krichevsky (National Institutes of Health) relates how the Microbial Strain Data Network will be used to index collections and detailed strain databanks on an international basis. Two of these sources of detailed information – MiCIS, the Microbial Culture Information Service which catalogues the UK national collections, and the Hybridoma Databank which lists international holdings of monoclonal antibodies, are presented by Geraldine Alliston (Laboratory of the Government Chemist) and Alain Bussard (CODATA and the University of Nice).

Banks of data, however detailed and comprehensive are only useful when processed and manipulated. Howard Bilofsky and Jean Michel Claverie (Institut Pasteur) describe how computer packages are being used to extract knowledge of sequence patterns from sequence databanks like GenBank. Chris Rawlings (Imperial Cancer Research Fund) talks about the results achieved so far in transforming one dimensional sequences into three-dimensional protein structures via Intelligent Knowledge Based Systems.

A conference on bio-*technology* (as distinct from biological science) should discuss the practical use of information in the world of industry and commerce and several papers look at these applications. The point of view of the production manager, beset with the increasing burden of regulations is given by Jan Ayres (Wellcome Biotech Ltd). Information as industrial property is reviewed by Peter Elliott (Marks and Clark Ltd) when held in patent documents, and by Ivan Bousfield (National Collection of Industrial and Marine Bacteria Ltd) when embodied in living form in culture collections. For the business manager, information is an economic resource in its own right and Bernard Jones (ICI Ltd) relates his experience of running a commercial intelligence centre within a large multi-national company, while the future of selling published information is examined by Jack Franklin (Asfra Consultants). A major problem for any company is understanding its markets, its competitors and its customers. Tom Clark (European Business Associates) reveals some of the black arts of the market researcher and Sarah Gordon (Kleinwort Grieverson) points out the key data that the City looks for when judging where to place its investments.

The biotechnology industry will only prosper when it operates within a society which is well informed and able to make mature criticisms of change and innovation. Bernard Dixon (contributing editor of the journal *Bio/Technology*) shows how a science journalist works; Edward Yoxen (University of Manchester) and Ray Zilinskas (University of California) discuss how public opinion forms and government decision-making is supported.

At this point the chain of information comes to an end—observations have become data, data have been collected and enhanced to produce information, information contributes to expert criticism and becomes knowledge. We must now just hope that some of this knowledge turns into wisdom!

Biotechnology Information'86 was organized by the European Biotechnology Information Project and IRL Press Ltd. The European Biotechnology Information Project was based within the British Library to provide an information service for industry and research. Having completed a fixed term of support from the Commission of the European Communities, its activities will now continue on a permanent basis as the Biotechnology Information Service at the British Library. IRL Press Ltd are publishers of books, journals, videos and software specializing in the fields of biology, biomedicine and biotechnology.

The organizers of Biotechnology Information'86 acknowledge, with thanks, the financial support of the Commission of the European Communities.

Richard Wakeford

The British Library, Biotechnology Information
Service, 9 Kean Street, London WC2 4AT

LIST OF CONTRIBUTORS

G.V.Alliston
Microbial Culture Information Service, Laboratory
of the Government Chemist, Cornwall House,
Waterloo Road, London SE1 8XY, UK

N.G. Anderson
Proteus Technology, ATCC, 12301 Parklawn
Drive, Rockville, MD 20852, USA

J.J.Ayres
Wellcome Biotech Ltd, Langley Court, Beckenham,
Kent BR3 3BS, UK

W.C.Barker
National Biomedical Research Foundation,
Georgetown University Medical Center, 3900
Reservoir Road, NW, Washington DC 20007, USA

H.Bilofsky
BBN Laboratories Inc., Cambridge, MA 02138,
USA

M.C.Blomquist
National Biomedical Research Foundation,
Georgetown University Medical Center, 3900
Reservoir Road, NW, Washington DC 20007, USA

I.J.Bousfield
The National Collections of Industrial & Marine
Bacteria Ltd, Torry Research Station, PO Box 31,
Aberdeen AB9 8DG, UK

A.Bussard
Hybridoma Data Bank, International Council of
Scientific Unions, Committee on Data for Science
and Technology, 51 Boulevard de Montmorency,
75016 Paris, France

T.J.Clark
European Business Associates, 22 rue Dernier Sol,
L-2543 Luxembourg, Luxembourg

J.-M.Claverie
Computer Science Unit, Institut Pasteur, F-75724
Paris Cedex 15, France

M.D.Dibner
Director, Business Studies in Biotechnology, The
North Carolina Biotechnology Center, PO Box
13547, Research Triangle Park, NC 27709, USA

B.Dixon
81 Falmouth Road, Chelmsford, Essex CM1 5JA,
UK

P.Elliott
Marks and Clerk, 57-60 Lincoln's Inn Fields,
London WC2A 3LS, UK

J.Flensholt
Computer Resources International A/S,
Vesterbrogade 1A, DK-1620, Copenhagen V,
Denmark

J.Franklin
ASFRA, Voorhaven 33, 1135 BL Edam, The
Netherlands

D.G.George

National Biomedical Research Foundation,
Georgetown University Medical Center, 3900
Reservoir Road, NW, Washington DC 20007, USA

S.L.Gordon

Hambrecht and Quist Inc., 277 Park Avenue, New
York, NY 10172, USA

D.T.Holland

Department of Reproductive Physiology, St
Bartholomew's Hospital Medical College, London
EC1A 7BE, UK

R.B.Jones

Imperial Chemical Industries p.l.c., Biological
Products Business, Agricultural Division, PO Box
1, Billingham, Cleveland TS23 1LB, UK

M.I.Krichevsky

Microbial Systematics Section, Epidemiology and
Oral Disease Prevention Program, National
Institute of Dental Research, National Institutes of
Health, Bethesda, MA 20892, USA

C.J.Rawlings

Imperial Cancer Research Fund, PO Box 123,
Lincoln's Inn Fields, London WC2A 3PX, UK

R.E.Silman

Department of Reproductive Physiology, St
Bartholomew's Hospital Medical College, London
EC1A 7BE, UK

E.Yoxen

Department of Science and Technology Policy,
University of Manchester, Manchester M13 9PL,
UK

R.A.Zilinskas

818 Sixth Street 102, Santa Monica, CA 90403,
USA

CONTENTS

NEW TECHNOLOGIES – NEW APPLICATIONS

SEQUENCE DATA BANKS

MICROBIAL AND CELLULAR DATA BANKS

New Technologies – New Applications

R.E.Silman
D.T.Holland

Protein fingerprinting and microbial identification

Department of Reproductive Physiology, St Bartholomew's Hospital Medical College, London EC1A 7BE, UK

ABSTRACT

Bacteria were incubated with a radiolabelled amino acid and their radiolabelled protein products separated by polyacrylamide gel electrophoresis (PAGE). In a first series of experiments the gels were autoradiographed and a wide variety of species studied. Visual inspection showed characteristic band patterns for each species and characteristic differences for subgroups within a species. In a second series of experiments the gels were scanned and the data analysed by computer. A small database was constructed and duplicate 'unknowns' from other gels were correctly matched using pattern recognition software. The study illustrates the potential for a simple, general and computerised system which uses pattern differences for bacterial classification and pattern matching for bacterial identification.

THE ORIGINS OF THE PROJECT

Our interest in microbial identification was entirely fortuitous, resulting from some flawed experiments which we had undertaken to investigate the adrenocorticotrophic (ACTH) peptide hormones of the fetal pituitary gland. Our first experiments were straightforward: (a) the extraction of mRNA from pituitary tissue; (b) the translation of mRNA in vitro using 35-S methionine to provide radiolabelled protein products; (c) the separation of the translated products by polyacrylamide gel electrophoresis (PAGE); and (d) the identification of the translated protein products by autoradiography. The results were as we had anticipated with a single prominent protein band representing the primary translation product of pituitary growth hormone (Fig. 1a). We therefore attempted to enhance the primary translation product of ACTH prior to electrophoresis by using a specific antibody directed against its N-terminal sequence and, to our surprise, the result showed a plurality of bands (Fig. 1b). In further experiments we tried several other antibodies, directed against different portions of the ACTH molecule. In each case, the results gave a plurality of bands and, even more surprising, the banding pattern differed from antibody to antibody. Such results were inexplicable for two reasons: (1) there should have been a single band representing the primary translation product of ACTH; and (2) the different antibodies should have bound to the same primary translation product and, therefore, they should have enhanced the same single band. Our error was uncovered when we undertook a control experiment in which no mRNA was added. Instead of the bands being abolished, each antibody continued to provide the same specific banding pattern as before. From this control experiment it became obvious that we had been observing an artefact. Instead of translating mRNA, we had

Biotechnology Information '86
Edited by R.Wakeford

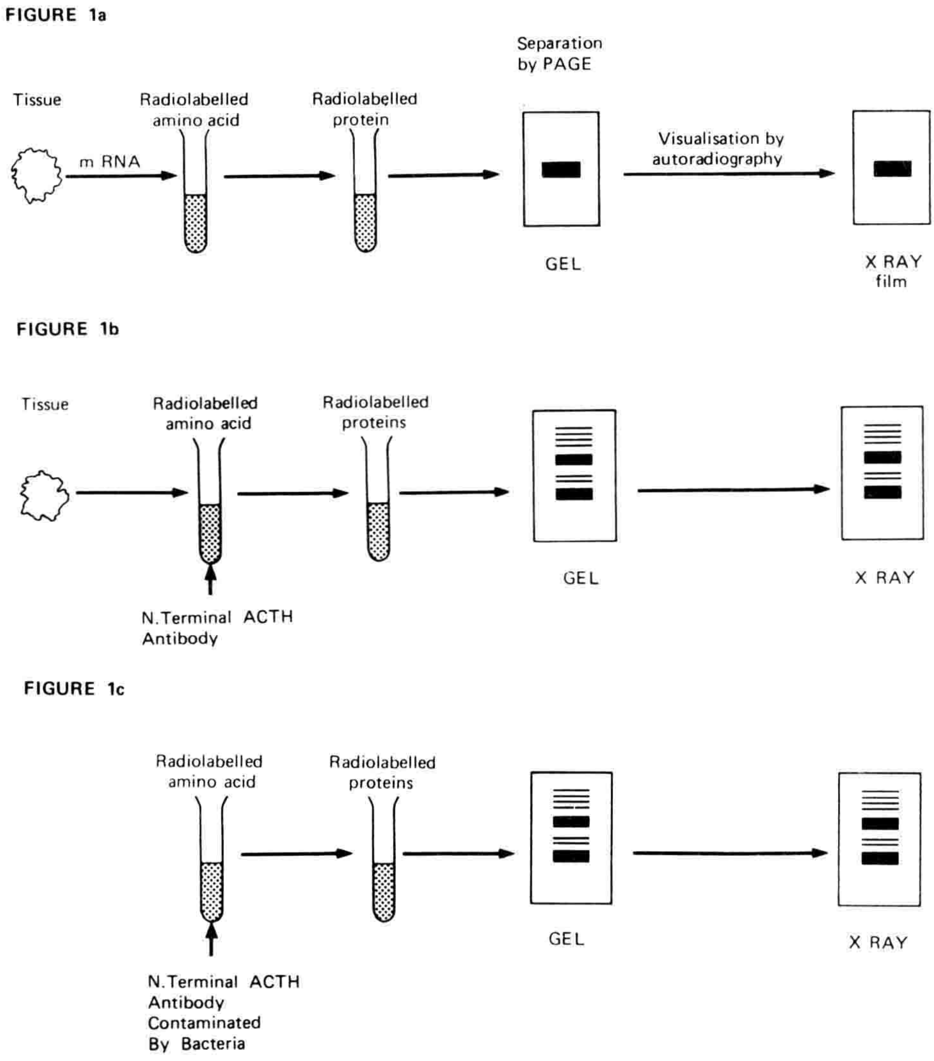

FIGURE 1a

FIGURE 1b

FIGURE 1c

Figure 1.
(a) the incorporation of 35-S methionine into a radiolabelled translation product after incubation with pituitary mRNA. The translation product was separated by PAGE and identified by autoradiography revealing a single band corresponding to growth hormone;
(b) the same as (a) except that the translation product(s) were first bound to an ACTH antibody. Several translation products were separated by PAGE and revealed by autoradiography;
(c) the same as (b) except that no pituitary mRNA was added to the incubation. Several translation products were separated by PAGE and revealed by autoradiography indicating the presence of contaminating microorganisms in the antibody.

been translating extraneous genetic material within our antibody solutions. In other words, the antibodies had become contaminated with unknown microorganisms and the specificity of the banding patterns were due to the specificity of the contaminating microorganisms rather than to the specificity of the antibodies (Fig. 1c). This was confirmed when we sterilised the antibody solutions and found that the banding patterns vanished.

Though these experiments were conducted more than five years ago, they told us something that we should already have known, i.e. that contaminating microorganisms will avidly incorporate a radiolabelled amino acid into the products of their metabolism. Had we been more skilled in molecular biology, we would have been cautious from the outset and would have ensured that the solutions were kept sterile. However, our ignorance had one advantage, it led us to repeat the flawed experiments over many months and, thus, we were able to observe that the various contaminating microorganisms in the different antibody solutions had specific banding patterns which remained constant from experiment to experiment. It was as if they were producing barcodes like one sees in a supermarket. In other words, the banding patterns could be used as 'identifiers' for the contaminating microorganisms.

To test whether this hypothesis was valid, we began working in collaboration with Professor Soad Tabaqchali in the Department of Medical Microbiology at St. Bartholomew's Hospital. A wide variety of bacterial species was cultured, incubated with 35-S methionine, the proteins separated by PAGE and the gels autoradiographed. Using coded samples we were easily able to match 'unknowns' against standards. In trying to identify sub-types within a species, the problem was a little more difficult since the overall pattern of the species was common to all its subtypes. However, small differences between the subtypes were apparent and we were able, for example, to identify twelve different serotypes of E. coli. Finally we applied the method to the typing of isolates of C .difficile and were able to demonstrate the existence of 9 distinct groups within the C. difficile species (1).

From this work we concluded that: (a) microorganisms will incorporate a radiolabelled amino acid into proteins according to their genetic code; (b) the radiolabelled proteins will provide a constant and specific pattern if they are separated by PAGE under constant conditions; and (c) the protein patterns can serve as 'fingerprints' to identify most species and sub-species.

THE DEVELOPMENT OF THE SYSTEM

A company, Automated Microbiology Systems, was created in 1983 to develop these principles into a working system. The Company was set up under the direction of Geoffrey Cross with Edwin Mack as its Director of Research and Development. Previously, Geoffrey Cross had been Chief Executive and Edwin Mack Research Director, of a major international computer company. They therefore focussed their attention, from the outset, on the information technology aspects of the problem.

The banding patterns which are produced by microorganisms are of considerable complexity. This can be seen in figure 2 where patterns from single isolates of Salmonella, E. coli, Proteus, Pseudomonas, Serratia, Klebsiella, and Enterobacter are illustrated. The information which characterises each microorganism can be found by identifying the position of each band in the pattern. However, if this was the sum total of the information, one would probably require two-dimensional electrophoresis to create the resolution needed for distinguishing between subspecies (see, for example, the paper by Dr. Anderson in this volume). In fact there is considerably more to the patterns than simple positional information. As with barcodes, the width of a band contains as much information as its position. More important, the relative intensity of the various bands within a pattern contains as much, if not more, information than their relative positions. In other words, a microorganism may synthesise a specific group of proteins which can be qualitatively assessed by their relative position and width on a gel, however, it will also synthesise these proteins at different rates and this can be quantitatively assessed by the relative intensity of the various bands. This last point is difficult to judge using X-ray film, since the gradations of exposure are linear only within a narrow dynamic range. The problem is, therefore, to harness the complexity of information by providing a computerised system which applies pattern recognition algorithms to the immense number of variables generated by the banding patterns.

It is also necessary to develop a scanning system which can gather this information and thereby replace autoradiography as the means of reading the gels. This is necessary for the reasons outlined above, i.e. the need for a detector with a greater dynamic range than X-ray film. It is also necessary for practical reasons. Exposure times using X-ray film are too long, requiring 24 hours to seven days. The need is for a scanner capable of detecting the same information in 1 or 2 hours and having that information directly stored within a computer ready for data analysis using pattern recognition algorithms. Such a scanner did not exist and it was thanks to the inventive genius of Dr. Brian Pullan that this key element was designed (2).

The final step in creating a working system is the need to provide rapid and reproducible PAGE separation of the proteins. To this end, the company developed a computerised electrophoresis system. The platens were designed to be freon-cooled so as to allow for efficient and uniform dissipation of heat during PAGE separation at high power. However, the most important aspect of PAGE separation, is the quality and reproducibility of the gels. It was thanks to Dr. Philip Bloch that this problem was finally solved and the Company now distributes easy to handle, reproducible, high resolution gels with a shelf-life of several months. A particular advantage of these gels is that they can be dried without cracking. To speed this process the Company also developed a hot-air drier. The patterns displayed in Figure 2 illustrate the quality of these gels.

All the essential elements of the system consisting of (a) consumables: gels and buffers; (b) hardware: electrophoresis unit, air-drier, scanner and computer; and (c) data-acquisition and data-processing software, are now in place. The entire procedure, (a) incorporation of a radiolabelled amino acid; (b) PAGE; (c) air-dry; (d) scan; and (e) computerised data-analysis, can therefore be completed within a 8-hour day. However, although the procedure is standardised, our approach is to avoid, in the short term, the grandiose project whereby a universal database is established in an attempt to solve all possible applications. Instead we have chosen to concentrate on well-defined and limited tasks where alternative methodologies are deficient. Thus, in industry, the system can be used for the quality control of specific microorganisms e.g. yeast in brewing or baking where the database consists of a limited number of specific strains of yeast. In medicine, the system can be used to investigate a spreading hospital-based infection where the problem is

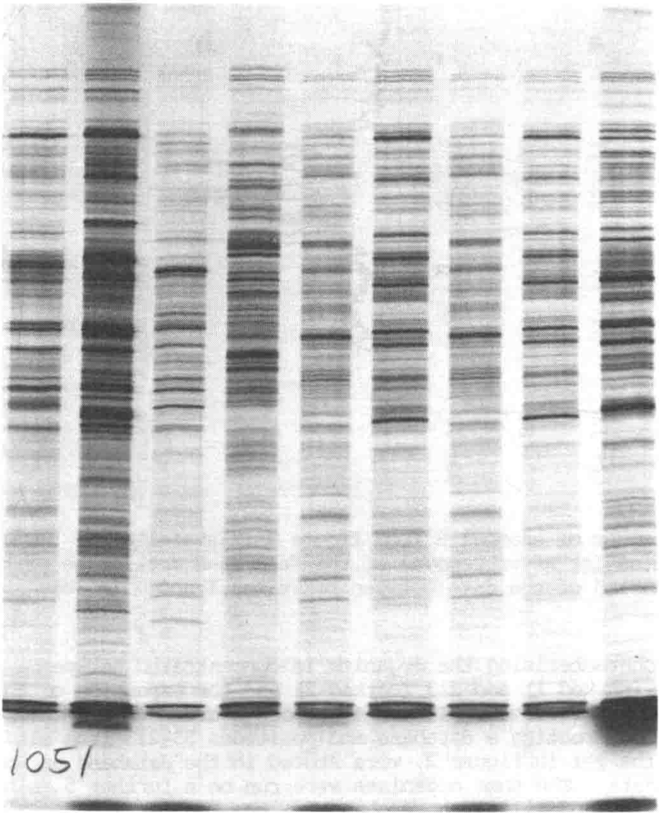

Figure 2. The pattern of radiolabelled proteins produced by microorganisms after incubation with 35-S methionine and separation by PAGE. Channel 1, Salmonella typhimurium; 2, Escherischia coli; 3, Proteus vulgaris; 4, Pseudomonas aeruginosa; 5, Serratia marcescens (also duplicated in channel 7); 6, Klebsiella pneumoniae (also duplicated in channel 8); 9, Enterobacter cloacae.

to determine whether the infection is due to a single type of microorganism and where the database is limited to the actual cases under study. Apart from quality control and epidemiology, the system will also readily lend itself to problems of sub-grouping and taxonomy.

We seek to limit the early use of the system to such specific applications because the software can be used in a variety of ways and is therefore most efficient when tailored to a specific goal. This can be illustrated by the example in Figure 2. The gel was scanned for 1 hour and the channels were extracted and filed. Figure 3a illustrates the full histogram of channel 8 (Klebsiella) with information gathered at all 360 positions. Figure 3b illustrates the same information after excluding positions 312 to 360 which contain the non-specific free 35-S methionine peak. It can be seen from Figure 3b that most of the