Essentials of Genomics and Bioinformatics

Edited by C. W. Sensen

WILEY-VCH

This Page Intentionally Left Blank

Essentials of Genomics and Bioinformatics

Edited by C. W. Sensen

This Page Intentionally Left Blank

Essentials of Genomics and Bioinformatics

Edited by C. W. Sensen

WILEY-VCH

Edited by:
Prof. Christoph W. Sensen
University of Calgary
Faculty of Medicine
Department of Biochemistry and
Molecular Biology
3330 Hospital Drive N.W.
Calgary, Alberta
Canada

This book was carefully produced. Nevertheless, authors, editor and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No. applied for

A catalogue record for this book is available from the British Library

Die Deutsche Bibliothek – Cataloguing-in-Publication Data:
A catalogue record for this book is available from Die Deutsche Bibliothek
ISBN 3-527-30541-6

Printed on acid-free paper.

Composition and Printing: Zechner Datenservice und Druck, Speyer.
Bookbinding: Grossbuchbinderei J. Schäffer GmbH + Co. KG, Grünstadt.

Printed in the Federal Republic of Germany

Preface

Genomics has become the new paradigm in biological and medical research. Initially being defined as the large-scale characterization of genomes (mainly through mapping and DNA sequencing), the field has become increasingly diverse, now covering the characterization of genomes on the sequence level, as well as the large-scale study of gene expression and the protein complement of organisms. Due to the rapid development of the field, it is quite difficult even for the specialist to stay informed about the wealth of new approaches that is being developed every year. This applies even more to the interested general public, thus this book tries to provide comprehensive introductory level information about the methods used in Genomics research, the model organisms studied, the bioinformatics approaches, which are used to analyze genomic data, and the ethical implications of genome research.

The first section of the book is dedicated to model organisms, including humans and the study of their genomes. Several chapters in this section deal with the characterization and analysis of the human genome, and the potential of Genomics for the cure of diseases. The next two sections of the book deal with DNA and protein technologies, including chapters on genomic mapping, DNA sequencing, DNA high-density arrays and the analysis of the proteome through Mass Spectrometry and Capillary Electrophoresis. The following section contains six chapters on the bioinformatics aspects of Genomics. While many books now deal with bioinformatics in general, these chapters are different as they exclusively deal with the analysis of complete genomes, a topic rarely covered elsewhere. The book concludes with a chapter about the ethical implications of genome research and an outlook to the future of Genomics.

This book started out as part of the Second Edition of *Biotechnology*. The success of Volume 5b *Genomics and Bioinformatics* encouraged the series editors, the publisher and myself to pursue a "spin off" from the series, which resulted in the current revised concise softcover edition. I would like to thank H.-J. Rehm, G. Reed, A. Pühler, P. Stadler and WILEY-VCH for the permission to let *Biotechnology* Volume 5b be used as the basis for this book. Special thanks go to Karin Dembowsky, who has now patiently guided the creation of this book in an extremely competent manner for over two years. Without her constant encouragement and support, it would have been impossible to complete this book. Finally, we are grateful for the suggestions provided by Stephanie E. Minnema (The University of Calgary, Faculty of Medicine) after critically reading *Biotechnology* Volume 5b *Genomics and Bioinformatics*.

Calgary, January 2002 C. W. Sensen

This Page Intentionally Left Blank

Contents

Ethical, Legal and Social Issues

Color Figures *see*
http://www.wiley-vch.de/books/info/
3-527-30541-6

Contributors

Dr. Lothar Altschmied
AG Expressionskartierung
Institut für Pflanzengenetik und
Kulturpflanzenforschung (IPK)
Corrensstraße 3
D-06466 Gatersleben
Germany
Chapter 1

Dr. Rolf Apweiler
EMBL Outstation
The European Bioinformatics Institute
Hinxton Hall, Hinxton
Cambridge, CB10 1SD
United Kingdom
Chapter 11

Gary D. Bader
Samuel Lunenfeld Research Institute
Mount Sinai Hospital
600 University Avenue
Toronto, Ontario, M5G 1X5
Canada
Chapter 16

Dr. Roland Brousseau
National Research Council of Canada
6100 Royalmount Avenue
Montreal, Quebec, H4P 2R2
Canada
Chapter 8

Dr. Detlev Buttgereit
Zell- und Entwicklungsbiologie
FB Biologie
Universität Marburg
Karl-von-Frisch-Straße
D-35032 Marburg
Germany
Chapter 1

Dr. Miroslaw Cygler
Biotechnology Research Institute
6100 Royalmont Avenue
Montreal, Quebec, H4P 2R2
Canada
Chapter 14

Prof. Dr. Antoine Danchin
Hong Kong University,
Pasteur Research Centre
Dexter HC Man Building
8, Sassoon Road, Pokfulam
Hong Kong
and
Genetics of Bacterial Genomes,
CNRS URA 2171
Institut Pasteur
28, rue du Docteur Roux
75724 Paris Cedex 15
France
Chapter 1

Dr. Daniel B. Davison
Director of Bioinformatics
Bristol Myers Squibb
Bioinformatics-853
5 Research Parkway
Wallingford, CT 06492-7660
USA
Chapter 4

Dr. Graham Dellaire
MRC – Human Genetics Unit
Crewe Road
Edinburgh, EH4 2XU
United Kingdom
Chapter 3

Dr. Norman Dovichi
Department of Chemistry
University of Washington
Seattle, WA 98195-1700
USA
Chapter 10

Dr. Thure Etzold
Lion Bioscience
Compass House
80–82 Newmarket Road
Cambridge, CB5 8DZ
United Kingdom
Chapter 11

Prof. Dr. Horst Feldmann
Adolf-Butenandt-Institut
Schillerstraße 44
D-80336 München
Germany
Chapter 1

Dr. Daniel Figeys
MDS Proteomics Inc.
251 Attwell Drive
Toronto, Ontario, M9W 7H4
Canada
Chapter 9

Dr. Terry Gaasterland
The Rockefeller University
1230 York Avenue
New York, NY 10021-6399
USA
Chapter 15

Paul Gordon
National Research Council of Canada
1411 Oxford Street
Halifax, Nova Scotia, B3H 3Z1
Canada
Chapter 15

Dr. Dwayne Hegedus
AAFC, Saskatoon Research Center
Molecular Genetics Section
107 Science Place
Saskatoon, Saskatchewan, S7N 0X2
Canada
Chapter 5

Dr. Christopher Hogue
Samuel Lunenfeld Research Institute
Mount Sinai Hospital
600 University Avenue
Toronto, Ontario, M5G 1X5
Canada
Chapter 16

Dr. Shen Hu
Department of Chemistry
University of Washington
Seattle, WA 98195-1700
USA
Chapter 10

Dr. Patrick G. Humphrey
Advanced Research and Development
Li-Cor, Inc.
4308 Progressive Avenue
P.O. Box 4000
Lincoln, NE 68504
USA
Chapter 7

Dr. Doris Jording
Lehrstuhl für Genetik
Universität Bielefeld
D-33594 Bielefeld
Germany
Chapter 1

Dr. Jörn Kalinowski
Lehrstuhl für Genetik
Universität Bielefeld
D-33594 Bielefeld
Germany
Chapter 1

Dr. Hans-Peter Klenk
EPIDAUROS Biotechnologie AG
D-82347 Bernried
Germany
Chapter 1

Prof. Bartha Maria Knoppers
Université de Montréal
Faculté de Droit
3101 chemin de la Tour
C.P. 6128, succursale A
Montreal, Quebec, H3C 3J7
Canada
Chapter 17

Prof. Dr. Manfred Kröger
Institut für Mikro- und Molekularbiologie
Universität Gießen
Heinrich-Buff-Ring 26–32
D-35392 Gießen
Germany
Chapter 1

Dr. Sergej N. Krylov
Department of Chemistry
York University
Toronto, Ontario, M3J 1P3
Canada
Chapter 10

Dr. Rodrigo Lopez
EMBL Outstation
The European Bioinformatics Institute
Hinxton Hall, Hinxton
Cambridge, CB10 1SD
United Kingdom
Chapter 11

Dr. Derek Lydiate
AAFC, Saskatoon Research Center
Molecular Genetics Section
107 Science Place
Saskatoon, Saskatchewan, S7N 0X2
Canada
Chapter 5

Dr. Alan Matte
Biotechnology Research Institute
6100 Royalmont Avenue
Montreal, Quebec, H4P 2R2
Canada
Chapter 14

Dr. David Michels
Department of Chemistry
University of Washington
Seattle, WA 98195-1700
USA
Chapter 10

Dr. Lyle R. Middendorf
Advanced Research and Development
Li-Cor, Inc.
4308 Progressive Avenue
P.O. Box 4000
Lincoln, NE 68504
USA
Chapter 7

Dr. Narasimhachari Narayanan
Advanced Research and Development
Li-Cor, Inc.
4308 Progressive Avenue
P.O. Box 4000
Lincoln, NE 68504
USA
Chapter 7

Dr. Isobel A. P. Parkin
AAFC, Saskatoon Research Center
Molecular Genetics Section
107 Science Place
Saskatoon, Saskatchewan, S7N 0X2
Canada
Chapter 5

Prof. Dr. Alfred Pühler
Lehrstuhl für Genetik
Universität Bielefeld
D-33594 Bielefeld
Germany
Chapter 1

Dr. Chandra S. Ramanathan
Bristol Myers Squibb
Bioinformatics-853
5 Research Parkway
Wallingford, CT 06492-7660
USA
Chapter 4

Prof. Dr. Renate Renkawitz-Pohl
Zell- und Entwicklungsbiologie
FB Biologie
Universität Marburg
Karl-von-Frisch-Straße
D-35032 Marburg
Germany
Chapter 1

Peter Rice
Lion Bioscience
Compass House
80–82 Newmarket Road
Cambridge CB5 8DZ
United Kingdom
Chapter 12

Dr. Stephen J. Robinson
AAFC, Saskatoon Research Center
Molecular Genetics Section
107 Science Place
Saskatoon, Saskatchewan, S7N 0X2
Canada
Chapter 5

Dr. Stephen C. Roemer
Advanced Research and Development
Li-Cor, Inc.
4308 Progressive Avenue
P.O. Box 4000
Lincoln, NE 68504
USA
Chapter 7

Dr. Kevin Rozwadowski
AAFC, Saskatoon Research Center
Molecular Genetics Section
107 Science Place
Saskatoon, Saskatchewan, S7N 0X2
Canada
Chapter 5

Dr. Stephen W. Scherer
The Hospital for Sick Children
HSC Research Institute
555 University Avenue
Toronto, Ontario, M5G 1X8
Canada
Chapter 2

Dr. Joseph D. Schrag
Biotechnology Research Institute
6100 Royalmont Avenue
Montreal, Quebec, H4P 2R2
Canada
Chapter 14

Dr. Christoph W. Sensen
University of Calgary
Faculty of Medicine
Department of Biochemistry and Molecular Biology
Calgary, Alberta, T2N 4N1
Canada
Chapters 15 and 18

Dr. Andrew G. Sharpe
AAFC, Saskatoon Research Center
Molecular Genetics Section
107 Science Place
Saskatoon, Saskatchewan, S7N 0X2
Canada
Chapter 5

Dr. Daniel C. Tessier
National Research Council of Canada
6100 Royalmount Avenue
Montreal, Quebec, H4P 2R2
Canada
Chapter 8

Dr. David Y. Thomas
McGill University
Faculty of Medicine
Department of Biochemistry
McIntyre Medical Sciences Building
Room #801
3655 Promenade Sir William Osler
Montreal, Quebec, H3G 1Y6
Canada
Chapter 8

Dr. Theerayut Toojinda
DNA Fingerprinting Unit
Biotechnology Center
Kasetsart University
Kampangsaen Campus
Nakorn Pathom 73140
Thailand
Chapter 6

Dr. Somvong Tragoonrung
DNA Fingerprinting Unit
Biotechnology Center
Kasetsart University
Kampangsaen Campus
Nakorn Pathom 73140
Thailand
Chapter 6

Prof. Lap Chee Tsui
The Hospital for Sick Children
HSC Research Institute
555 University Avenue
Toronto, Ontario, M5G 1X8
Canada
Chapter 2

Dr. Apichart Vanavichit
DNA Fingerprinting Unit
Biotechnology Center
Kasetsart University
Kampangsaen Campus
Nakorn Pathom 73140
Thailand
Chapter 6

Dr. David Wishart
University of Alberta
Faculty of Pharmacy and Pharmaceutical Sciences
Dentistry/Pharmacy Centre 2123
Edmonton, Alberta, T6G 2N8
Canada
Chapter 13

Dr. Evgeni M. Zdobnov
EMBL Outstation
The European Bioinformatics Institute
Hinxton Hall, Hinxton
Cambridge, CB10 1SD
United Kingdom
Chapter 11

Dr. Zheru Zhang
Analytical R & D
Pharmaceutical Research Institute
Bristol – Myers Squibb Company
Syracuse, NY 13221-4755
USA
Chapter 10

Introduction

CHRISTOPH W. SENSEN

Calgary, Canada

Genomics has revolutionized biological and medical research and development over the last fifteen years. The speed and magnitude by which Genomics has outgrown the disciplines from which it originally developed have taken many by surprise. The rapid development of the field has left much of the early history of Genomics behind and many key events have not been recorded properly.

It may be even all but forgotten how the term "genomics" originated. According to the first editorial of *Genomics* (1987, **1**, 1–2), the term was coined by T. H. RODERICK from the Jackson Laboratories in Bar Harbor, MN, some time around 1987 in discussions with editors VICTOR A. MCKUSICK and FRANK H. RUDDLE, who were looking for suggestions to name their new journal.

There is no all-encompassing definition for genomics, the word is used with many meanings. At the time when MCKUSICK and RUDDLE wrote their editorial, they understood genomics to be mapping and sequencing to analyze the structure and organization of genomes. When the Genomics journal was founded, only three years had passed since the invention of automated DNA sequencers, which dominated the first phase of the development of genomics as a science. Thus a definition such as MCKUSICK and RUDDLE's of the word genomics can be understood in the context of that time.

Today, genomics is very often subdivided into "structural genomics", which deals with the determination of the complete sequence of genomes (DNA sequencing), or the complete set of proteins (proteome) in an organism (proteomics), and "functional genomics", which studies the functioning of genes and metabolic pathways (metabolomics) or the gene expression patterns in an organism (chip technologies). To complicate matters, X-ray crystallographers have adopted the term structural genomics to refer to protein 3-D structure determination.

For the purpose of this book, genomics has the broader meaning of "genome research", including bioinformatics and other studies of the genome and proteome to understand the blueprint and function of organisms. Many of the technologies that are part of today's genomics toolkit were developed previously and then automated in an attempt to apply them in large-scale, high-throughput environments. Some people, including the late Canadian Nobel laureate MICHAEL SMITH (UBC), have claimed that they were doing genomics all along, which is true to a certain degree when using a broad definition of genomics, considering its strong roots in molecular biology, biophysics, and biochemistry. With such a definition, we may say that genomics really started when WATSON and CRICK discovered the structure of DNA.

Without a doubt, the introduction of computers into molecular biology laboratories was one of the key factors in the development of genomics. Laboratory automation led to the production of large amounts of data, and the need to analyze, combine, and understand these resulted in the development of "bioinformatics", a new discipline at the interface of several traditional disciplines. Bioinformatics is the glue that integrates all the diverse aspects of genomics. Of similar importance to the development of the field is the development of laser-based technologies. The use of laser-based systems, which can be coupled to computerized detection systems, has replaced most of the radioactive techniques in genomics laboratories, allowing the complete automation of many types of experiments.

Considering the rapid pace of development, it is quite difficult to organize a book that reflects all aspects of genomics. Chapters about model organisms are followed by overviews of the key technologies. Because of the importance to the field, several chapters are dedicated to bioinformatics. Genomics is a science with huge impact on society, thus ethical and legal issues that need to be dealt with arise daily. One of the book chapters is devoted to ethical and legal aspects of genome research. The book closes with an outlook to future developments in genomics.

With the completion of the human genome, the true tasks for genomics are only starting to emerge. We are far from understanding how organisms with small genomes function, let alone how the human genome is organized. As more and more scientific disciplines get "genomicized", the field will undergo continual transformation, thus a book like this one can only capture a flavor and a moment in time. This may be frustrating to some, but this book is intended to summarize the essence of the first fifteen years of research and development in genomics, during which the cornerstone for a very exiting future was laid.

Calgary, January 2002 Christoph W. Sensen

Application Domains

This Page Intentionally Left Blank

1 Genome Projects of Model Organisms

ALFRED PÜHLER,
DORIS JORDING,
JÖRN KALINOWSKI
Bielefeld, Germany

DETLEV BUTTGEREIT
RENATE RENKAWITZ-POHL
Marburg, Germany

LOTHAR ALTSCHMIED
Gatersleben, Germany

ANTOINE E. DANCHIN
Hong Kong

HORST FELDMANN
Munich, Germany

HANS-PETER KLENK
Bernried, Germany

MANFRED KRÖGER
Giessen, Germany

1 Introduction

Genome research allows the establishment of the complete genetic information of organisms. The first complete genome sequences established were those of prokaryotic and eukaryotic microorganisms followed by plants and animals (see, e.g., the TIGR web page at *http://www.tigr.org/*). The organisms selected for genome research were mostly those which already played an important role in scientific analysis, thus they can be considered model organisms. In general, organisms are defined as model organisms when a large amount of scientific knowledge has been accumulated in the past. For this chapter on genome projects of model organisms, a number of experts in genome research have been asked to give an overview on specific genome projects and to report on the respective organism from their specific point of view. The organisms selected include prokaryotic and eukaryotic microorganisms as well as plants and animals.

We have chosen the prokaryotes *Escherichia coli*, *Bacillus subtilis*, and *Archaeoglobus fulgidus* as representative model organisms. The *E. coli* genome project is described by M. KRÖGER (Giessen, Germany). He gives a historical outline about the intensive research on microbiology and genetics of this organism, which cumulated in the *E. coli* genome project. Many of the technological tools presently available have been developed during the course of the *E. coli* genome project. *E. coli* is without doubt the best analyzed microorganism of all. The knowledge of the complete sequence of *E. coli* has confirmed its reputation to represent the leading model organism of gram-negative eubacteria.

A. DANCHIN (Hong Kong) reports on the genome project of the environmentally and biotechnologically relevant gram-positive eubacterium *B. subtilis*. The contribution focuses on the results and analysis of the sequencing effort and gives a number of examples for specific and sometimes unexpected findings of this project. Special emphasis is given to genomic data which support the understanding of general features such as translation, specific traits relevant for living in its general habitat or its usefulness for industrial processes.

A. fulgidus is the subject of the contribution by H.-P. KLENK (Bernried, Germany). This genome project was started before the genetic properties of the organism had been extensively studied. However, its unique lifestyle as a hyperthermophilic and sulfate-reducing organism makes it a model for a large number of environmentally important microorganisms and species with a high biotechnological potential. The structure and results of the genome project are described in the contribution.

The yeast *Saccharomyces cerevisiae* has been selected as a representative of eukaryotic microorganisms. The yeast project is presented by H. FELDMANN (Munich, Germany). *S. cerevisiae* has a long tradition in biotechnology as well as a long-term research history as the eukaryotic model organism *per se*. It was the first eukaryote to be completely sequenced and has led the way to sequencing other eukaryotic genomes. The wealth of the yeast sequence information as a useful reference for plant, animal or human sequence comparisons is outlined in this contribution.

Among plants, the small crucifere *Arabidopsis thaliana* was identified as the classical model plant due to simple cultivation and short generation time. Its genome was originally considered to be the smallest in the plant kingdom and was, therefore, selected for the first plant genome project which is described here by L. ALTSCHMIED (Gatersleben, Germany). The sequence of *A. thaliana* helped to identify that part of the genetic information unique to plants. In the meantime, other plant genome sequencing projects were started, many of which focus on specific problems of crop cultivation and nutrition.

The roundworm *Caenorhabditis elegans* and the fruitfly *Drosophila melanogaster* have been selected as animal models due to their specific model character for higher animals and also for humans. The genome project of *C. elegans* is summarized by D. JORDING (Bielefeld, Germany). The contribution describes how the worm – in spite of its simple appearance – became an interesting model organism for features such as neuronal growth, apoptosis, or signaling pathways. This genome project has also provided a number of bioinformatic tools which are widely used for other genome projects.

The genome project concerning the fruitly *D. melanogaster* is described by D. BUTTGEREIT and R. RENKAWITZ-POHL (Marburg, Germany). *D. melanogaster* being the best analyzed multicellular organism up to now, is capable of serving as a model system for features such as the development of limbs, the nervous system, circadian rythms, and even for complex human diseases. The contribution gives examples for the genetic homology and similarities between *Drosophila* and humans and outlines perspectives for studying features of human diseases using the fly as a model.

2 Genome Projects of Selected Prokaryotic Model Organisms

2.1 The Gram-Negative Enterobacterium *Escherichia coli*

2.1.1 The Organism

The development of the most recent field of molecular genetics is directly connected to one of the best described model organisms, the eubacterium *Escherichia coli*. There is no textbook in biochemistry, genetics or microbiology without extensive sections describing the numerous basic observations, which have been noted first in *E. coli* cells or the respective bacteriophages, or using *E. coli* enzymes as a tool, respectively. Consequently, several monographs solely regarding *E. coli* have been published. Although it seems to be impossible to name or count the number of scientists involved in the characterization of *E. coli*, Tab. 1 is an attempt to name some of the most relevant people in chronological order.

The scientific career of *E. coli* (Fig. 1) started in 1885, when the German paediatrician T. ESCHERICH described the first strain from the faeces of newborn babies. As late as 1958, this is internationally recognized by using his name to classify the respective group of bacterial strains. In 1921 the very first report on virus formation was published for *E. coli*. Today we call the respective observation "lysis by bacteriophages". In 1935 these bacteriophages became the most powerful tool in defining the characters of individual genes. Because of their small size, they turned out to be ideal tools for statistical calculations performed by the former theoretical physicist M. DELBRÜCK. His very intensive and successful work has attracted many others to this area of research. In addition, DELBRÜCK's extraordinary capability to catalyze the exchange of ideas and methods yielded the legendary Cold Spring Harbor Phage course. Everybody interested in basic genetics once attended the famous summer course or at least came to the respective annual phage meeting. This course, which was an optimal combination of joy and work, became an ideal source to spread practical methods. For many decades it was the most important exchange forum for results and ideas, as well as strains and mutants. Soon, the so-called phage family was formed, which interacted almost like one big laboratory, e.g., results were communicated by preprints preferentially. Finally, 15 Nobel prize laureates have their roots in this summer school (see Tab. 1).

The substrain *E. coli* K12 was first used by E. TATUM as a prototrophic strain. It was chosen more or less by chance from the strain collection of Stanford Medical School. Since it

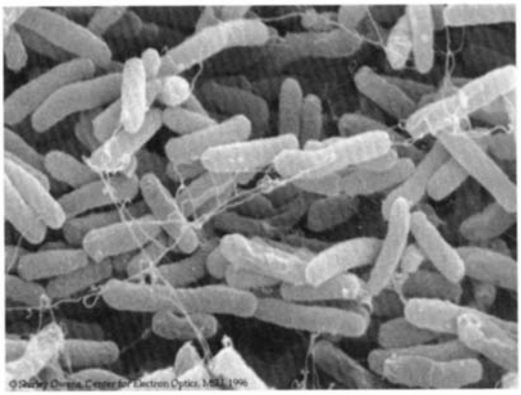

Fig. 1. Scanning electron micrograph (SEM) of *Escherichia coli* cells (image courtesy of: SHIRLEY OWENS, Center for Electron Optics, MSU; found at *http://commtechlab.msu.edu/sites/dlc-me/zoo/zah0700.html#top#top*).

Tab. 1. Chronology of the Most Important Primary Detections and Method Applications with *E. coli*

1885	"bacterium coli commune" by T. Escherich
1921	Lysogeny and prophages by d'Herelle
1939	Growth kinetics for a bacteriophage by M. Delbrück (Nobel prize 1969)
1943	Statistical interpretation of phage growth curve (game theorie) by S. Luria (Nobel prize 1969)
1946	Konjugation by E. Tatum and J. Lederberg (Nobel prize 1958) Repair of UV damage by A. Kelner and R. Dulbecco (Nobel prize for tumor virology)
1952	DNA as the carrier of genetic information, proven via radioisotopes by M. Chase and A. Hershey (Nobel prize 1969)
1953	Phage immunity as the first example for gene regulation by A. Lwoff (Nobel prize 1965) Transduction of *gal* genes (first isolated gene) by E. and J. Lederberg Host-controlled modification of phage DNA by G. Bertani and J. J. Weigle
1956	DNA polymerase I by A. Kornberg (Nobel prize 1959) Polynucleotide phosphorylase (RNA synthesis) by M. Grunberg-Manago and S. Ochoa (Nobel prize 1959)
1958	Semiconservative duplication of DNA by M. Meselson and F. Stahl
1959	Operon theory and induced fit by F. Jacob and J. Monod (Nobel prize 1965)
1962	Restriction enzymes by W. Arber (Nobel prize 1978)
1963	Physical genetic map with 99 genes by A. L. Taylor and M. S. Thoman Strain collection by B. Bachmann
1967	DNA ligase by several groups at the same time
1972	DNA hybrids by P. Lobban and D. Kaiser
1973	Recombinant DNA from *E. coli* and SV40 by P. Berg (Nobel prize 1980) Patent on genetic engineering by H. Boyer and S. Cohen
1974	Sequencing techniques using *lac* operator by W. Gilbert and *E. coli* polymerase by F. Sanger (Nobel prize 1980)
1975	Promoter sequence by H. Schaller Attenuation by C. Yanowsky General ribosome structure by H. G. Wittmann
1977	Rat insulin expressed in *E. coli* by H. Goodmann Synthetic gene expressed by K. Itakura and H. Boyer
1978	Site directed mutagenesis by M. Smith (Nobel prize 1993)
1984	Polymerase chain reaction by K. B. Mullis (Nobel prize 1993)
1987	Restriction map of the complete genome by Y. Kohara and K. Isono
1989	Organism specific sequence data base by M. Kröger
1995	Total sequence of *Haemophilus influenzae* using an *E. coli* comparison
1996	Systematic sequence finished by a Japanese consortium under leadership of H. Mori
1997	Systematic sequence finished by F. Blattner
1999	Three dimensional structure of ribosome by four groups at the same time

was especially easy to cultivate and since it is, as an inhabitant of our gut, a nontoxic organism by definition, the strain became very popular. Because of the already acquired vast knowledge and because of its lack to form fimbriae, in 1975 *E. coli* K12 was chosen as the only organism to allow early cloning experiments in the famous Asilomar Conference on biosafety (Berg et al., 1975). No wonder that almost each of the following basic observations in life sciences was either done with or within *E. coli*. However, what started as the "phage family", dramatically split into hundreds of individual groups working in tough competion. As one of the most important outcomes, sequencing of *E. coli* was performed more than once. Because of the separate efforts, the genome finished only as number seven (BLATTNER et al., 1997; YAMAMOTO et al., 1997; KRÖGER and WAHL, 1998). However, the amount of knowledge acquired is certainly second to none, and the way how this knowledge has been acquired is interesting, both for the history of sequencing methods and bioinformatics, and for the influence of national and individual pride.

The work on *E. coli* is not finished with the completion of the DNA sequence. Data will be continously acquired to fully characterize the genome in terms of genetic function and protein structures (THOMAS, 1999). This is very important, since a number of toxic *E. coli* strains is known. Thus research on *E. coli* has turned from basic science into applied medical research. Consequently, the human toxic strain O157 has been also completely sequenced, again more than once (PERNA et al., 2001 and unpublished data).

2.1.2 Characterization of the Genome and Early Sequencing Efforts

With its history in mind and realizing the impact of the respective data, it is obvious that an ever growing number of colleagues worldwide worked with or on *E. coli*. Consequently, there was an early need for the organization of data. This led to the first physical genetic map of any living organism comprising 99 genes, published by TAYLOR and THOMAN (1964). This map was improved and refined for several decades by BACHMANN (1983), and M. BERLYN (see Neidhard, 1996). These researchers still maintain a very useful collection of strains and mutants at Yale University. A number of 1,027 loci had been mapped in 1983 (BACHMANN, 1983), these were used as a basis for the very first sequence database specific to a single organism (KRÖGER and WAHL, 1998). As shown in Fig. 2 of KRÖGER and WAHL (1998), sequencing of *E. coli* started as early as 1967, with one of the first ever characterized tRNA sequences. Immediately after DNA sequencing had been established, numerous laboratories started to determine sequences of their personal interest.

2.1.3 Structure of the Genome Project

In 1987 the group of K. ISONO published a very informative and incredibly exact restriction map of the entire genome (KOHARA et al., 1987). With the help of K. RUDD, it was possible to locate sequences quite precisely (see NEIDHARD, 1996; ROBERTS, 2000). But only very few saw any advantage in closing the sometimes very small gaps, thus a worldwide joint sequencing approach could not be established. Two groups, one in Kobe, Japan (YAMAMOTO et al., 1997) and one in Madison, WI (BLATTNER et al., 1997) started systematic genome sequencing in parallel. In addition, another laboratory at Harvard University used *E. coli* as a target to develop a new sequencing technology. Several meetings, organized especially on *E. coli,* did not result in any unified systematic approach, thus many genes have been sequenced two or three times. Specific databases have been maintained to bring some order into the increasing chaos. However, even this type of tool has been developed several times in parallel (KRÖGER and WAHL, 1998; Roberts, 2000). Whenever a new contiguous sequence was published, about 75% had already previously been submitted to the international databases by other laboratories. The progress of data acquisition followed a classical e-curve, as shown in Fig. 2 of KRÖGER and WAHL (1998). Thus in 1992 it was possible to predict the completeness of the sequence for 1997, without knowing about the enormous technical innovations in between (KRÖGER and WAHL, 1998).

Both the Japanese consortium and the group of F. BLATTNER started early; people say they started too early. They subcloned the DNA first, and they used manual sequencing and older informatic systems. Sequencing has been performed semiautomatically, and many students were employed to read and control the X-ray films. When the first genome sequence of *Haemophilus influenzae* appeared in 1995, the science foundations wanted to discontinue the support of the *E. coli* projects, which received their grant support mainly because of the model character of the sequencing techniques developed.

Three facts and the truly international protest, convinced the juries to continue financial support: First, in contrast to other completely sequenced organisms, *E. coli* is an autonomously living organism. Second, when the first complete very small genome sequence was released, even the longest contiguous sequence for *E. coli* was already longer. Third, the other laboratories could only finish their sequences because the *E. coli* sequences were already publicly available. Consequently, the two competing main laboratories were allowed to purchase several of the meanwhile developed sequencing machines and use the shutgun approach to complete their efforts. Finally, they finished almost at the same time: H. Mori and his colleagues included already published sequences from other laboratories into their sequence data and sent them to the international databases on December 28, 1996 (Yamamoto et al., 1997) while F. Blattner turned in an entirely new sequence on January 16, 1997 (Blattner et al., 1997). They added last changes and additions as late as October 1998. Very sadly, at the end *E. coli* had been sequenced almost three times (Kröger and Wahl, 1998). However, most people nowadays forget about all other sources and refer to the Blattner sequence.

2.1.4 Results of the Genome Project

When the sequences finally were finished, most of the features of the genome had already been known. Consequently, people did no longer celebrate the *E. coli* sequence as a major breakthrough. At that time, everybody knew that the genome is almost completely covered with genes, although less than half of them have been genetically characterized. Tab. 2 illustrates this and shows the differences

Tab. 2. Some Statistical Features of the *E. coli* Genome

Total size	4,639,221 bp[a]	acc. to Regulon[d]	acc. to Blattner[e]
Transcription units	proven	528	
	predicted	2,328	
Genes	total found	4,408	4,403
	regulatory	85	
	essential	200	
	nonessential[b]	2,363	1,897
	unknown[c]	1,761	2,376
	tRNAS	84	84
	rRNA	29	29
Promoters	proven	624	
	predicted	4,643	
Sites (?)		469	
Regulatory interactions	found	642	
	predicted	275	
Terminators	found	96	
RBSs		98	
Gene products	regulatory proteins	85	
	RNAs	115	115
	other peptides	4,190	4,201

[a] Additional 63 bp compared to the original sequence.

[b] Genes with known or predicted function.

[c] Yet no other data available than the existence of an open reading frame with a start sequence and more than 100 codons.

[d] Data from *http://tula.cifn.unam.mx8850/regulondb/regulon_doc/summary.*

[e] Data from *http://www.genome.wisc.edu.*

in counting. Because of this high density of genes, F. BLATTNER and coworkers defined "grey holes" whenever they found more than 2 kb of noncoding region (BLATTNER et al., 1997). It turned out that the termination of replication is almost exactly opposite to the origin of replication. No special differences have been found for either direction of replication. Approximately 40 formerly described genetic features could not yet be located or supported by the sequence (KRÖGER and WAHL, 1998; NEIDHARD, 1996). On the other hand, we have several examples for multiple functions encoded by the same gene. It turned out that the multifunctional genes are mostly involved in gene expression and used as general control factor. M. RILEY determined the number of gene duplications, which is also not unexpectedly low when neglecting the ribosomal operons (see ROBERTS, 2000).

Everybody is convinced that the real work is starting only now. A number of strain differences may be the cause of deviations between the different sequences available. Thus the number of genes as well as nucleotides differ slightly (see Tab. 2). Everybody would like to know the function of each of the open reading frames (THOMAS, 1999), but nobody has received the grant money to work on this important problem. Seemingly, the other model organisms are of more public interest, thus, it may well be that research on other organisms will now help to understand *E. coli*, just the way how *E. coli* provided information to understand them. In contrast to yeast, it is very hard to produce knock-out mutants. Thus, we may have the same situation in the postgenomic era as we had before the genome was finished. A number of laboratories will continue to work with *E. coli*, they will constantly characterize one or the other open reading frame, but there will not be any mutual effort (THOMAS, 1999).

2.1.5 Follow up Research in the Postgenomic Era

Today, it seems to be more attractive to work with toxic *E. coli* strains like O157, rather than with *E. coli* K12. This strain has recently been completely sequenced, the data are available via Internet. The comparison between toxic and nontoxic strains will certainly help to understand the toxic mechanisms. On the other hand, it turned out that it was correct to use *E. coli* K12 as the most intensively used strain for biological safety regulations (BERG et al., 1975). No additional features turned out to change this.

Surprisingly, the colleagues from mathematics or informatics are those who showed most interest in the bacterial sequences. They did all kind of statistical analyses and tried to find some evolutionary roots. Here another fear of the public is already formulated: People are afraid of the attempts to reconstruct the first living cell. So there are at least some attempts to find the minimal set of genes for the most basic needs of a cell. We have to ask again the very old question: Do we really want to "play God"? If we want, *E. coli* indeed could serve as an important milestone.

2.2 The Gram-Positive Spore Forming *Bacillus subtilis*

2.2.1 The Organism

Self-taught ideas have a long life: articles about *Bacillus subtilis* (Fig. 2) almost invariably begin with words such as: "*B. subtilis*, a soil bacterium ...", nobody taking the elementary care to check whether this is based on experimental observations. *Bacillus subtilis*, first identified in 1885, is named *ko so kin* in japanese and *laseczka sienna* in polish,"hay bacterium", and this refers to the real biotope of the organism, the surface of grass or low lying plants (SEKOWSKA, 1999). Interestingly, it required its genome to be sequenced to conquer again its right biotope. Of course, plant leaves fall on the soil surface, and one must find *B. subtilis* there, but its normal niche is the surface of leaves, the phylloplane. Hence, if one wishes to use this bacterium in industrial processes, to engineer its genome, or simply to understand the functions coded by its genes, it is of fundamental importance to understand where it normally thrives, and what the environmental parameters are controlling its life

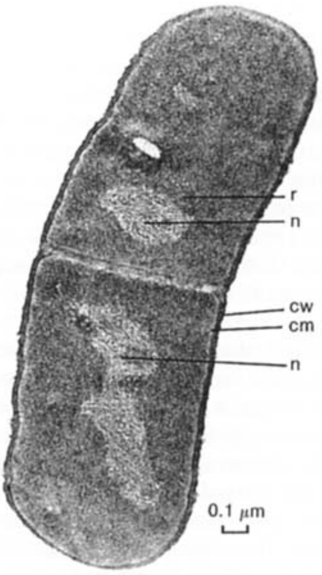

Fig. 2. Electron micrograph of a thin section of *Bacillus subtilis*. The dividing cell is surrounded by a relatively dense wall (CW), enclosing the cell membrane (cm). Within the cell, the nucleoplasm (n) is distinguishable by its fibrillar structure from the cytoplasm, densely filled with 70S ribosomes (r).

cycle and the corresponding gene expression. Among other important ancillary functions, *B. subtilis* has thus to explore, colonize, and exploit the local resources, while at the same time it must maintain itself, dealing with congeners and with other organisms: Understanding *B. subtilis* requires understanding the general properties of its normal habitat (Sekowska, 1999).

2.2.2 A Lesson from Genome Analysis: The *Bacillus subtilis* Biotope

The genome of *B. subtilis* (strain 168), sequenced by a team of European and Japanese laboratories, is about 4,214,780 bp long. Of the more than 4,100 protein-coding genes, 53% are represented once. One quarter of the genome corresponds to several gene families which have been probably expanded by gene duplication. The largest family contains 77 known and putative ATP-binding cassette (ABC) permeases indicating that, despite its large metabolism gene number, *B. subtilis* has to extract a variety of compounds from its environment (Kunst et al., 1997). In general, the permeating substrates are unchanged during permeation. Group transfer, where substrates are modified during transport, however, plays an important role in *B. subtilis*. Its genome codes for a variety of phosphoenolpyruvate-dependent systems (PTS) which transport carbohydrates and regulate general metabolism as a function of the nature of the supplied carbon source. A functionally-related catabolite repression control, mediated by a unique system (not cyclic AMP), exists in this organism (Saier, 1998). Remarkably, apart from the expected presence of glucose-mediated regulation, it appears that carbon sources related to sucrose play a major role, via a very complicated set of highly regulated pathways, indicating that this carbon supply is often encountered by the bacteria. In the same way, *B. subtilis* can grow on many of the carbohydrates synthesized by grass-related plants.

In addition to carbon, oxygen, nitrogen, hydrogen, sulfur, and phosphorus are the core atoms of life. Some knowledge about other metabolisms in *B. subtilis* has accumulated, but significantly less than in its *E. coli* counterpart. However, knowledge of its genome sequence is rapidly changing the situation, making *B. subtilis* a model of similar general use as *E. coli*. A frameshift mutation is present in an essential gene for surfactin synthesis in strain 168, but it has been found that including a small amount of a detergent into plates allowed these bacteria to swarm and glide extremely efficiently (C.-K. Wun and A. Sekowska, unpublished observations). The first lesson of the genome text analysis is thus that *B. subtilis* must be tightly associated with the plant kingdom, with grasses in particular (Kunst et al., 1997). This should be considered a priority when devising growth media for this bacterium, in particular in industrial processes.

Another aspect of the *B. subtilis* life cycle consistent with a plant-associated life, is that it can grow over a wide range of different temperatures, up to 54–55 °C – an interesting feature for large-scale industrial processes. This indicates that its biosynthetic machinery comprises control elements and molecular chaperones that permit this versatility. Gene duplication may permit adaptation to high temperature, with isozymes having low and high temperature optima. Because the ecological niche of *B. subtilis* is linked to the plant kingdom it is subjected to rapid alternating drying and wetting. Accordingly, this organism is very resistant to osmotic stress, and can grow well in media containing 1 M NaCl. Also, the high levels of oxygen concentration reached during daytime are met with protective systems: *B. subtilis* appears to have as many as six catalase genes, both of the heme-containing type (*katA*, *katB*, and *katX* in spores) and of the manganese-containing type (*ydbD*, *yjqC*, and *cotJC* in spores).

The obvious conclusion of these observations is that the normal *B. subtilis* niche is the surface of leaves (Arias et al., 1999). This is consistent with the old observation that *B. subtilis* makes up the major population of bacteria of rotting hay. Furthermore, consistent with the extreme variety of conditions prevailing on plants, *B. subtilis* is an endospore-forming bacterium, making spores highly resistant to the lethal effects of heat, drying, many chemicals, and radiation.

2.2.3 General Organization of the Genome: A First Law of Genomics

Analysis for repeated sequences in the *B. subtilis* genome discovered an unexpected feature: strain 168 does not contain insertion sequences. A strict constraint on the spatial distribution of repeats longer than 25 bp was found in the genome, in contrast to the situation in *E. coli*. The correlation of the spatial distribution of repeats and the absence of insertion sequences in the genome suggests that mechanisms aiming at their avoidance and/or elimination have been developed (Rocha et al., 1999a). This observation is particularly relevant for biotechnological processes where one has multiplied the copy number of genes in order to improve production. Although there is generally no predictable link between the structure and function of biological objects (Danchin, 1999), the pressure of natural selection has adapted together gene and gene products. Biases in features of predictably unbiased processes is evidence for prior selective pressure. In the case of *B. subtilis* one observes a strong bias in the polarity of transcription with respect to replication: 70% of the genes are transcribed in the direction of the replication fork movement (Kunst et al., 1997). Global analysis of oligonucleotides in the genome demonstrated that there is a significant bias not only in the base or codon composition of one DNA strand with respect to the other, but, quite surprisingly, there is a strong bias at the level of the amino acid content of the proteins. The proteins coded by the leading strand are valine-rich, and those coded by the lagging strand are threonine + isoleucine-rich. This first law of genomics seems to extend to many bacterial genomes (Rocha et al., 1999b). It must result from a strong selection pressure of a yet unknown nature, demonstrating that, contrary to an opinion frequently held, genomes are not, at a global scale, plastic structures.

2.2.4 Translation: Codon Usage and the Organization of the Cell's Cytoplasm

Exploiting the redundancy of the genetic code, coding sequences exhibit highly variable biases of codon usage. The genes of *B. subtilis* are split into three classes on the basis of their codon usage bias. One class comprises the bulk of proteins, another is made up of genes that are expressed at a high level during exponential growth, and a third class, with A + T-rich codons, corresponds to portions of the genome that have been horizontally exchanged (Kunst et al., 1997).

When mRNA threads are emerging from DNA they become engaged by the lattice of ribosomes, and ratchet from one ribosome to the next, like a thread in a wiredrawing machine (Danchin et al., 2000). In this process, nascent proteins are synthesized on each ribo-