PART 1

Chapter 1

Introduction

1.1 Overview

In this chapter, we introduce the main perspectives of the book: **bioinformatics** and **computer science**. In Section 1.2, we offer a working definition of the term 'bioinformatics', we discuss where the discipline came from, and consider the impact of the **genome-sequencing** revolution. In Section 1.3, we discuss the origins of computer science, and note the emerging challenges relating to how to manage and describe biological data in ways that are computationally tractable. Having set the scene, we reflect briefly on some of the gaps that now confront computer science and bioinformatics.

By the end of the chapter, you should have an appreciation of how the field of bioinformatics evolved; you should also have gained insights into the extent to which its future progress is linked to the advances in data management and **knowledge representation** that are engaging computer science today.

1.2 Bioinformatics

1.2.1 What is bioinformatics?

Bioinformatics is a term that means different things to different people, with so many possible interpretations – many of them entirely reasonable – that it can sometimes be difficult to know what bioinformatics actually means, and whether it isn't just **computational biology** by another name. One way of making sense of the bioinformatics landscape is to recognise that it has both service and research components. Its service side primarily involves the routine storage, maintenance and retrieval of biological data. While these may seem like rather humdrum tasks for today's technologies, we'll explore why this is far from being true. By contrast, the research side of bioinformatics largely involves analysis of biological data using a variety of tools and techniques, often in combination, to create complex workflows or pipelines, including components ranging from **pattern recognition** and **statistics**, to visualisation and modelling. As we'll see, a particularly important facet of data analysis also concerns the use and development of prediction tools (later, we'll look at some of the ramifications of our heavy reliance on structure- and function-prediction approaches, especially in light of the emergence of high-throughput biology). The union of all of these capabilities into a broad-based,

Bioinformatics challenges at the interface of biology and computer science: Mind the Gap. First Edition.
Teresa K. Attwood, Stephen R. Pettifer and David Thorne. Published 2016 © 2016 by John Wiley and Sons, Ltd.
Companion website: www.wiley.com/go/attwood/bioinformatics

interdisciplinary science, involving both theoretical, practical and *conceptual* tools for the generation, dissemination, representation, analysis and understanding of biological information sets it apart from computational biology (which, as the name suggests, is perhaps more concerned with the development of mathematical tools for modelling and simulating biological systems).

In this book, we broadly explore issues relating to the computational manipulation and conceptual representation of biological sequences and macromolecular structures. We chose this vantage-point for two reasons: first, as outlined in the Preface, this is our 'home territory', and hence we can discuss many of the challenges from first-hand experience; second, this is where the discipline of bioinformatics has its roots, and it's from these origins that many of its successes, failures and opportunities stem.

1.2.2 The provenance of bioinformatics

The origins of bioinformatics, both as a term and as a scientific discipline, are controversial. The term itself was coined by theoretical biologist **Paulien Hogeweg**. In the early 1970s, she established the first research group specialising in bioinformatics, at the University of Utrecht (Hogeweg, 1978; Hogeweg and Hesper, 1978). Back then, with her colleague Ben Hesper, she defined the term to mean 'the study of informatic processes in **biotic systems**' (Hogeweg, 2011). But the term didn't gain popularity in the community for almost another two decades; and, by the time it did, it had taken on a rather different meaning.

In Europe, a turning point seems to have been around 1990, with the organisation of the *Bioinformatics in the 90s* conference (held in Maastricht in 1991), probably the first conference to include this 'new' term – bioinformatics – in its title. Consider that, during the same period, the National Center for Biotechnology Information[1] (NCBI) had been established in the United States of America (USA) (Benson *et al.*, 1990). But this was a centre for ***biotechnology*** information, not a bioinformatics centre, and it was established, at least in part, to provide the nation with a long-term 'biology **informatics** strategy' (Smith, 1990), not a 'bioinformatics' strategy.

With a new label to describe itself, a new scientific discipline evolved from the growing needs of researchers to access and analyse (primarily biomedical) data, which was beginning to accumulate, seemingly quite rapidly, in different parts of the world. This sudden data accumulation was the result of a number of technological advances that were yielding, at that time, unprecedented quantities of ***biological sequence*** information. Hand-in-hand with these developments came the widespread development of the algorithms, and computational tools and resources that were necessary to analyse, manipulate and store the amassing data. Together, these advances created the vibrant new field that we recognise today as bioinformatics.

Looking back, although the full history is convoluted, certain pivotal concepts and milestones stand out in the broadening bioinformatics panorama. In the following pages, we look at two of the major drivers of the evolving story: i) the technological developments that spawned the data deluge and facilitated its world-wide propagation; and ii) the development of **databases** to store the rapidly accumulating data. Before we do so, however, we must first identify a convenient starting point.

[1] http://www.ncbi.nlm.nih.gov

1.2.3 The seeds of bioinformatics

It's useful to think about where and when the seeds of bioinformatics were first sown, as this helps to provide a context for the situation we're in today. But where do we start? We could go all the way back to **Franklin** and **Gosling**'s foundational work towards the elucidation of the structure of **DNA** (Franklin and Gosling, 1953a, b, c), or to **Watson** and **Crick**'s opportunistic interpretation of their data (Watson and Crick, 1953). We could focus on the painstaking work of **Sanger**, who, in 1955, determined the **amino acid** sequence of the first **peptide hormone**. We could consider the ground-breaking work of **Kendrew** *et al*. (1958) and of **Muirhead** and **Perutz** (1963) in determining the first 3-dimensional (3D) structures of **proteins**. Or we could fast-forward to the progenitors of the first databases of macromolecular structures and sequences in the mid-1960s and early 1970s. This was clearly a very fertile era, heralding some of the most significant advances in molecular biology, and leading to the award of a series of **Nobel Prizes** (*e.g.*, 1958, Sanger's prize in chemistry; 1962, Watson, Crick and **Wilkins**' shared prize in physiology or medicine, after Franklin's death; and Perutz and Kendrew's prize in chemistry). These advances, any one of which could provide a suitable stepping-off point, each played an important part in the unfolding story.

Because this book focuses on biological sequences, especially protein sequences, we've chosen Fred Sanger's pioneering work on the peptide hormone **insulin** as our reference. Sanger was the first scientist to elucidate the order of amino acids in the **primary structure** of a protein. This was an immensely difficult puzzle, requiring the use of a range of chemical and enzymatic techniques in a variety of different experiments over many years. Each step of this incremental process was sufficiently new and exciting to warrant a separate publication. Eventually, around ten papers detailed the intricate work that led to the elucidation of the sequences, first of **bovine** insulin (*e.g.*, Sanger, 1945; Sanger and Tuppy, 1951a, b; Sanger and Thompson, 1953a, b; Sanger *et al*., 1955; Ryle *et al*., 1955), and then of **ovine** and **porcine** insulins (Brown *et al*., 1955). This monumental achievement had taken a decade to complete. It seems incredible now that such a small protein sequence (containing only ~50 amino acids – see Figure 1.1) could have resisted so many experimental assaults and withheld its secrets for so long; and more so, that its 3D structure would not be known for a further *14 years* (Adams *et al*., 1969)!

Manual **protein sequencing** was clearly an enormous undertaking, and it was many years before the sequence of the next protein was deduced: this was a small **enzyme** called **ribonuclease**. Work began on ribonuclease in 1955. Following a series of preliminary studies, the first full '**draft sequence**' was published in 1960 (Hirs *et al*., 1960), and a carefully refined final version was published three years later (Smyth *et al*., 1963). Importantly, this eight-year project offered a stepping-stone towards elucidation of the protein's 3D structure: in fact, knowledge of its amino acid sequence provided a vital piece of a 3D **jigsaw puzzle** that was to take a further four years to solve (Wyckoff *et al*., 1967). Viewed in the light of today's high-throughput sequence and structure determinations, these time-scales seem unimaginably slow.

Despite the technical and intellectual hurdles, the potential of amino acid sequences to aid our understanding of the functions, structures and evolutionary histories of proteins was compelling. The crucial role of protein sequences was clear to scientists like **Anfinsen**, arguably the originator of the field of **molecular evolution** (Anfinsen, 1959), and later, **Zuckerkandl** and **Pauling**, who helped build the foundations of molecular paleontology, and introduced evolutionary concepts like the **molecular clock**, based on

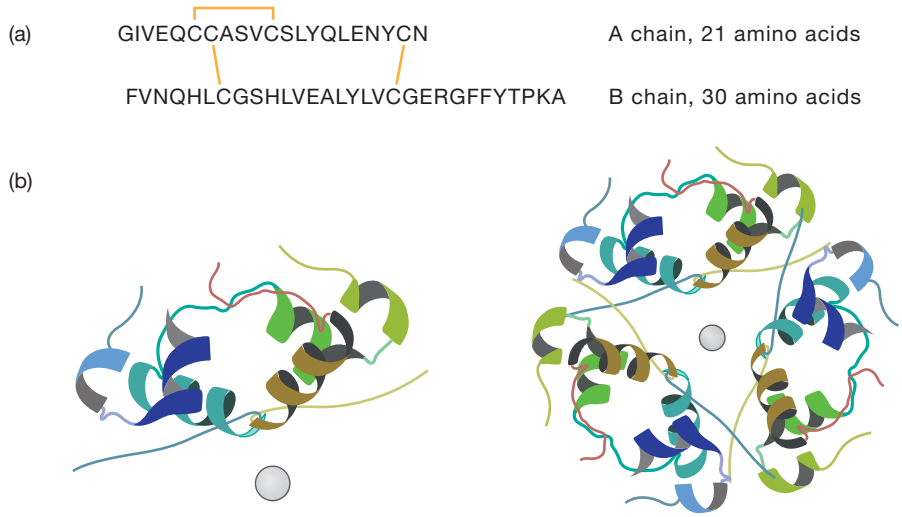

Figure 1.1

Bovine insulin (INS_BOVIN, P01317[i]): (a) the primary structure, showing intra- and interchain disulphide bonds connecting the A chain and B chain; and (b) its zinc-coordinated tertiary structure (PDB: 2INS[ii]), revealing two molecules in the asymmetric unit, and a hexameric biological assembly.

Source: Protein Data Bank.

[i] http://www.uniprot.org/uniprot/P01317
[ii] http://www.rcsb.org/pdb/explore/explore.do?structureId=2ins

rates of change of nucleotide or **polypeptide** sequences (Zuckerkandl, 1987). In their 1965 paper, Zuckerkandl and Pauling asserted that,

> All the potentialities of an individual may be assumed to be inscribed in polypeptide chains that are actually synthesized, or could be synthesized, by the cells under certain circumstances, and in the structures that control the actual and potential rates of this synthesis (Zuckerkandl and Pauling, 1965).

With great prescience, they envisaged biomolecular sequences as 'documents of evolutionary history'; nevertheless, they recognised that extracting their sequestered histories would require much more sequence information than was then available.

As the field of molecular evolution dawned, our hunger to understand more about the functions and structures of biological macromolecules provided the impetus for further sequencing efforts. **Margaret Dayhoff**, who had a keen interest in discerning evolutionary relationships from **sequence alignments**, was among the first scientists to approach this work systematically. In the early 1960s, to facilitate both her own research and the work of others in the field, she began to collect protein sequence information from scientific papers. This growing compendium of sequences was eventually published as a book – the ***Atlas of Protein Sequence and Structure*** (Dayhoff *et al.*, 1965), often simply referred to as the *Atlas*. Interestingly, in a letter she wrote in 1967, she observed,

> There is a tremendous amount of information regarding the evolutionary history and biochemical function implicit in each sequence and *the number of known sequences is growing explosively* [our emphasis]. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it (Dayhoff, 1967; Strasser, 2008).

With the creation of the first *Atlas*, that 'explosive growth' amounted to 65 sequences!

During the next decade, the advent of automated processes overtook time-consuming manual peptide sequencing and dramatically increased the rate of sequence determination. In the meantime, another revolution was taking place, spurred on by the elucidation of the first protein atomic structures using the technique of **X-ray crystallography**: those of **myoglobin** and **haemoglobin** (Kendrew *et al.*, 1958; Muirhead and Perutz, 1963). Building on the sequencing work, this advance set the scene for a new era in which structure determination was to take centre stage in our quest to understand the biophysical mechanisms that underpin biochemical and evolutionary processes. So seductive was this approach that many more structural studies were initiated, and the numbers of deduced structures burgeoned.

In parallel with these developments, advances in sequencing technology in the late 1970s meant that, for the first time, DNA sequences could be determined, and their protein products deduced from them, relatively quickly. Incredibly, where it had taken 8–10 years to determine the sequences of the first small proteins (insulin and ribonuclease), dozens of protein sequences could now be rapidly deduced by translation of sequenced DNA.

The technologies that gave rise to manual peptide-sequencing strategies, then to automated peptide and DNA sequencing, and to protein structure determination at atomic resolution, were thus responsible for producing the first waves of sequence and structural data. Key to handling this expanding information was the recruitment of computers to help systematically analyse and store the accumulating data. Initially, newly determined sequences were published in the literature to make them available to the wider community. In this form, any researcher wishing to exploit the information had first to obtain a copy of the original article, and then to type the sequence(s) into a computer by hand, a process that now seems almost unbelievable.

Eventually, it became clear that collating data into electronic repositories would make it more efficient and easier to store and use the data in future. This realisation led to the birth of the first electronic databases. At this time, the idea that molecular information could be managed using electronic repositories was not only very new, but was also very daunting. Consider that technologies we take for granted today (**email**, the **Internet**, the **World-Wide Web** ('the Web')) hadn't yet emerged; there was therefore no simple way to distribute data from a central database, other than posting **computer tapes** and disks to users on request. This model of data distribution was fraught with difficulties: it was cumbersome and slow; it was also relatively costly, and, alarmingly, led some of the first database pioneers to adopt pricing and/or data-sharing policies that threatened to drive away many of their potential users.

From these tentative awakenings, the first biomolecular databases emerged. From the agonisingly slow trickle of determining a single sequence per decade, to a speed of thousands of sequences per second (a rate that will itself seem inconsequential ten years from now), sequencing technology has revolutionised the pace of data acquisition, and has thrown up new challenges for the field of bioinformatics.

1.3 Computer Science

1.3.1 Origins of computer science

The discipline of computer science is associated with many things, from the design of **silicon chips** and the creation of life-style-enhancing gadgets, to the applications and **operating systems** that have become the mainstay of modern life. It's an enormously

broad subject, touching on topics as diverse as **electronic engineering**, **mathematics**, **aesthetics**, and even human **psychology** and **sociology**. But at its heart, computer science is about one thing: finding mechanisms of representing our universe – both its physical and conceptual nature – such that these can be manipulated and experimented with in ways that are beyond the capacity of the human mind. The process goes something like this: identify an intractable problem in the real world (perhaps one that requires a human to remember too many things at once), or that takes more steps to solve than can be done in a sensible amount of time; devise an abstract representation of that problem and of its constituent parts; and finally, create a device or process that's able to manipulate that representation in an automated way.

This need to manipulate abstract representations in a disciplined and controlled way makes computer science a form of extreme applied mathematics. As we might expect, this anchors its origins much further back in history than bioinformatics. Echoes of the process of abstraction and representation can be seen as far back as 2400 BCE, with the invention of the **abacus**. By drawing lines in the sand, and positioning pebbles in specific patterns, the ancient **Babylonians** were able to represent and manipulate what we would now think of as being positive **integer** numbers – useful for counting concrete physical things such as sheep, slaves and other everyday items of Babylonian life. However, even though it may seem like no enormous conceptual leap to imagine 'half a sheep', the abacus had no way of representing fractional numbers, which did not appear in mathematical systems until several hundred years later, via Egyptian **hieroglyphics**. Neither of these systems coped with negative numbers – what, after all, would it mean to have minus-one **pyramids**? Perhaps more surprisingly, they also lacked a representation for the concept of 'zero': there was no 'thing' that represented 'no thing'. Techniques for capturing and manipulating these more abstract concepts did not appear for at least another two millennia. The pattern of devising increasingly sophisticated and rich representations for abstract concepts, and of building devices that help take the drudge out of manipulating them, has continued to this day. This is the essence of computer science.

Of course, mathematics, and the process of 'computing' answers using mathematical notations, evolved considerably over the centuries, and its history is far too complex and intricate to discuss here; besides, for the most part, its details aren't relevant to the story this book tells. There are, however, a couple of notable exceptions. One was the creation, by Indian mathematician **Brahmagupta** in the 7th century CE, of the idea of an 'algorithm': a formal description of a sequence of steps that, carried out in order, accomplish a specific mathematical task. Another was the realisation, in the 3rd century BCE, by **Pingala**, another Indian mathematician, that numbers could be represented in '**binary notation**' as patterns of *true* or *false* values, *on* or *off* states, or simply by the presence or absence of objects. Centuries later, binary notation became the foundation of computer hardware, with binary numbers being captured, first as the presence or absence of holes in a **punched card**, and later as the presence or absence of a charge in a **valve** or **transistor**. Binary also spawned the idea of '**logic**' in a formal, mathematical sense, which, together with the concept of algorithms, became the cornerstone of modern software methods.

Historians will continue to wrangle over exactly what constitutes the first computing device: whether it was the ancient abacus, '**Napier's Bones**' (c. 1610, a contraption for manipulating **logarithms**), the '**Jacquard Loom**' (c. 1800, which used punched cards to control weaving patterns), **Charles Babbage**'s **Difference and Analytical Engines** (1882 and 1837), or one of the many other machines for automating mathematical calculation, remains a topic of enthusiastic (and often heated) academic discussion. To give

the recognition deserved to the pioneers of mathematics and computing that steered the path from these early crude machines to the **laptops** and **servers** of modern time is well beyond the scope of this book. Instead, we will leap forward in time, past all the technology-related creativity inspired by the **Second World War**, to the early 1960s, and the coining of the term 'computer science' by numerical analyst **George Forsythe**. By this time, the computer had many of the properties associated with contemporary machines: a **Central Processing Unit** (CPU) able both to manipulate numbers represented in binary and to execute a **program** to choreograph such manipulations; some form of **memory** in which to store data and programs; **back-up devices** to record the contents of memory when the device was switched off; connective infrastructure to allow the machine to communicate with devices other than itself; and various forms of input and output (screens, keyboards, *etc*.) to allow users to interact with the machine.

The first computer programs focused on solving mathematical problems (like finding large **prime numbers**) for which no equation existed that could simply be 'solved' by a human with pen and paper, but for which iterative algorithms could be devised that played to the computer's strength of being able to mindlessly and repetitively 'crunch' numbers according to prescribed rules. Over time, programs were written to perform calculations that were related to more pragmatic ends: calculating payrolls and other business tasks, simulating engineering problems, and recording and searching over data about individuals for all manner of purposes. Today, computers perform innumerable jobs, from sending television programmes across **wireless networks**, to hand-held **mobile devices**, to autonomously guiding the trajectory of space probes visiting the outer reaches of our **solar system**.

Of course, this unashamedly selective, whirlwind tour of the history of computer science overlooks a multitude of incredible developments and technological innovations of the past few decades. On the face of it, many of these appear to be merely incremental improvements to existing ideas, resulting in something that is just a bit faster, smaller, bigger, greener, lighter – insert the adjective of your choice here – than before. Many of these improvements, however, are not just the result of 'trying a bit harder': often, development of the next generation of a particular technology has required considerable research and creative effort to overcome or circumvent what were previously considered unbreachable limits of engineering or physics. Other inventions have resulted in turning points in the way we live and work: the first machines 'cheap' enough to have on one's desk, or efficient enough to run off a battery in one's pocket; the early Internet as a connective infrastructure for commerce and research, and the Web as its more friendly front-end, with much broader appeal. More recently, **social networking**, wireless and mobile technology, and the ability to easily create and distribute audio and video, have changed how society interacts in ways that were inconceivable only a few years ago. Many of these technologies are now so integral to modern society that we fail to recognise they exist, much less the effort and thought that led to their creation.

1.3.2 Computer science meets bioinformatics

So, what has happened to allow today's devices to quietly perform such diverse, incredibly complex tasks since the 'primitive' machines of the 1960s? Although faster, cheaper, smaller, less power-hungry, and in almost every way better, more sophisticated and interconnected than their older counterparts, the modern computer has very much the same basic components as its ancestors: its core remains a device for manipulating binary numbers. What has changed is our ability to use patterns of numbers to

represent increasingly complex and sophisticated things. Using a computer to calculate mathematical results is, in some ways, easy – this, after all, is what computers do. Determining which book someone is likely to want to buy next, based on the reading habits of people who have been established as having similar tastes, or – to bring this technological tale back to the life sciences – predicting whether individuals are likely to suffer from particular **diseases**, based on their genetic profiles, are altogether much more complex problems for computers to solve.

Here, then, is a gap: the vast gulf between a computer's ability to manipulate binary numbers, and our desire to use these machines to examine, understand and manipulate concepts – which, ostensibly, have no relationship to binary numbers at all. Bridging this gap, by devising techniques for representing our increasingly sophisticated knowledge in 'computable' form, is a fundamental challenge of modern computer science, one whose solution is poised to transform bioinformatics. Next to the very prominent developments we've just talked about, this is a much quieter revolution, but one that's already shaping how computers deal with data, and particularly **data representation**. In the early days of computing, **flat-file** databases, with their field-based searches, were the norm. In time, however, these were superseded by **relational database** systems that captured some of the structure of the information they modelled; and now we have **graph databases**, which attempt to model the meaning of data through the use of **controlled vocabularies**, allowing 'intelligent' data- and **text-mining** algorithms to scour them for nuggets of knowledge. This trend towards richer, more **semantic** (and thus 'computer friendly') data representation reflects our ongoing quest to make the growing volumes of data accessible to us as knowledge, knowledge that will touch innumerable aspects of our lives, from our relationships with each other and with our planet, to our future medical practice and health.

Albeit very much a thing of the past in the realms of computer science, the flat-file database was the cornerstone of early bioinformatics, playing a pivotal role in housing the gradually accumulating quantities of biomedical information. But modern high-throughput biology changed all this: its data explosion caught many bioinformaticians off-guard, and brought a growing realisation that the technologies underpinning the earliest databases were simply not up to the job. If storing the now-vast quantities of biological data was becoming increasingly demanding, reasoning over the data was becoming more difficult still. Consequently, in the aftershocks that followed, a gulf opened up between what we wanted to be able to do with bioinformatics on the one hand, and what we could actually do with it on the other.

1.4 What did we want to do with bioinformatics?

Because the origins of bioinformatics were rooted in sequence analysis, the earliest analyses aimed to understand what biomolecular sequences could tell us about the functions of **genes** and of their encoded proteins. Ultimately, scientists wanted to discover how amino acid sequences determined 3D **protein folds**, and what their sequences and structures could tell us about their evolutionary histories. Perhaps more importantly, researchers wanted to know how sequence and structure information could be used to elucidate the roles of particular genes and proteins in **pathogenic processes**, and how the assembled data could be used to design better, more efficacious **drugs**.

Later, with the advent of the human **genome-sequencing project**, the goals became increasingly ambitious, and the focus of attention turned more and more towards using bioinformatics to revolutionise molecular medicine. Researchers wanted to identify the

genetic determinants both of rare **syndromes** and of common, pervasive diseases like **cancer**; bioinformatics, it was claimed, would play a major role in the development of new approaches to eradicate such diseases, and would pave the way to **personalised therapies**, where an individual's **genome** could be used to determine which drug regime would offer maximal benefit with the minimum of **side-effects**. Ultimately, the goal was to integrate all molecular and cellular data in such a way as to be able to model **biochemical pathways** and, indeed, whole **cells**, and to understand not just how individual cells work, but how complete assemblies of cells work in whole **organs**, including the brain.

To put some of these aspirations in context, in the run up to the publication of the human genome, huge expectations were placed on what bioinformatics should or would be able to deliver. One commentator predicted that a bioinformatics revolution was afoot from which the next step in man's **evolution** would be

> our acquisition of the power to control the evolution of our own **species** and all others on the planet…to create plants that walk[2], animals that can carry out **photosynthesis** and other unlikely **chimeras**…, [ultimately to] go beyond, in human–computer communication, anything we can remotely conceive of at present (Cantor, 2000).

Other predictions weren't quite so far-fetched. Yet there was a general belief that this new discipline would transform research in fields from **genomics** to **pharmacology**, and would probably 'reverse the long-standing **reductionist** paradigm that has held sway in molecular biology' for more than 50 years;

> In addition, bioinformatics will likely provide the methodology finally to make highly accurate predictions about protein **tertiary structure** based on amino acid sequences and a viable means to design drugs based on computer simulation of the **docking** of small molecules to the predicted protein architecture…New computational methods will likely transform taxonomic and phylogenic [sic] studies as well as our ability to understand and predict the results of complex **signal transduction** cascades and the **kinetics** of intricate **metabolic pathways** (Wallace, 2001).

Such views were not uncommon in this exciting new era. Genomics research was making it possible to investigate biological phenomena on a hitherto-impossible scale, amongst other things, generating masses of **gene expression**, gene and protein sequence, protein structure, and protein–protein interaction data.

> How to handle these data, make sense of them, and render them accessible to biologists working on a wide variety of problems is the challenge facing bioinformatics…The 'post-genomic era' holds phenomenal promise for identifying the mechanistic bases of organismal development, metabolic processes, and disease, and we can confidently predict that bioinformatics research will have a dramatic impact on improving our understanding of such diverse areas as the regulation of gene expression, protein structure determination, comparative evolution, and **drug discovery** (Roos, 2001).

As this book unfolds, we'll touch on some of the predictions that have been made for the bioinformatics revolution, and consider how realistic they are in the context of the challenges that still lie ahead. The reality is that *in silico* approaches won't transport us to Star Trek[3] futures quickly. Indeed, more than a decade ago, a rather prescient

[2] http://en.wikipedia.org/wiki/Triffid
[3] http://en.wikipedia.org/wiki/Star_Trek

Nature Biotechnology editorial suggested that the transformation we could expect genomics and *in silico* tools to have on traditional **empirical** medicine,

> should take about 10 to 15 years (with a following wind) . . . Thus, genomics will not rapidly improve the efficiency of drug development. In fact, it may make it even more complicated (Editorial, 2001).

In the chapters that follow, we will explore some of the complexities. We'll look at the transition from numeric to symbolic algorithms that was necessary to allow bioinformatics to move beyond the computation, say, of sequence comparison scores, to manipulation of concepts or entities, such as '**prion**', '**promoter**', '**helix**', and so on (Attwood and Miller, 2001). We'll touch on many of the emerging techniques that are being used to transform data into knowledge, exploring how **ontologies** can be used to give meaning to raw information, how **semantic integration** is beginning to make it possible to join up disparate data-sets, and how visualisation techniques provide a way of harnessing human intuition in situations where computational techniques fall short. As we navigate the Grand Canyon[4] interface between bioinformatics and computer science, we'll see why, despite the power of computers and progress in **information technology**, bioinformatics is still not as straightforward as it perhaps could or should be. In particular, we'll examine the gap between what we wanted to do with bioinformatics and what we can actually do with it, and why the 'following wind' will need to be very much stronger if we're to make substantial progress even in the next 10–15 years.

Before addressing the technical and philosophical issues that arise when we use computers to try to tackle biological problems, the next chapter will take a brief look at the biological context that provides the foundation for molecular sequence- and structure-based bioinformatics today.

1.5 Summary

This chapter explored the nature and roots of bioinformatics, and the origins of computer science. In particular, we saw that:

1. Bioinformatics has both service and research components;
2. The service side of bioinformatics involves storage, maintenance and retrieval of biological data;
3. The research side of bioinformatics largely involves analysis and conceptual modelling of biological data;
4. The term 'bioinformatics' originally had a different meaning, and pre-dates the discipline we recognise today;
5. The discipline evolved from labour-intensive manual technologies that aimed to deduce molecular sequences and structures, and from largely descriptive manual approaches to catalogue this information;
6. The first such catalogues were books;
7. Manual approaches for deriving molecular sequences were gradually superseded by powerful automatic processes;

[4] http://en.wikipedia.org/wiki/Grand_Canyon

8 Automation of sequencing technologies catalysed both the spread of databases to store, maintain and disseminate the growing quantities of data, and the development of algorithms and programs to analyse them;
9 Automation of DNA sequencing technologies generated data on a scale that was inconceivable 60 years ago, and even now is almost unimaginable;
10 Computer science involves finding ways of representing real-world problems such that they can be manipulated by machines;
11 The scale of modern bio-data production is demanding new computational approaches for data storage and knowledge representation;
12 There is currently a gap between what computers do (manipulate binary numbers) and what we want them to do (examine and manipulate concepts);
13 The impact of bioinformatics on drug discovery and personalised medicine has been slower to emerge than predicted;
14 Bridging the knowledge-representation gap will help to advance bioinformatics in future.

1.6 References

Adams, M.J., Blundell, T.L., Dodson, E.J. *et al.* (1969) Structure of rhombohedral 2 zinc insulin crystals. *Nature*, **224**, 491– 495.

Anfinsen, C. (1959) *The Molecular Basis of Evolution*. John Wiley & Sons, Inc., New York.

Attwood, T.K. and Miller, C. (2001) Which craft is best in bioinformatics? *Computers in Chemistry*, **25**(4), 329–339.

Benson, D., Boguski, M., Lipman, D.J. and Ostell, J. (1990) The National Center for Biotechnology Information. *Genomics*, **6**, 389–391.

Brown, H., Sanger, F. and Kitai, R. (1955), The structure of pig and sheep insulins. *Biochemical Journal,* **60**(4), 556–565.

Cantor, C. (2000) Biotechnology in the 21st century. *Trends in Biotechnology*, **18**, 6–7.

Dayhoff, M.O., Eck, R.V., Chang, M.A. and Sochard, M.R. (eds) (1965) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD, USA.

Dayhoff, M.O. to Berkley, C. (1967) Margaret O. Dayhoff Papers, Archives of the National Biomedical Research Foundation, Washington, DC.

Editorial. (2001) A cold dose of medicine. *Nature Biotechnology*, **19**(3), 181.

Franklin, R.E. and Gosling, R.G. (1953) **a**) The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallographica*, **6**, 673–677; **b**) The structure of sodium thymonucleate fibres. II. The cylindrically symmetrical Patterson function. *Acta Crystallographica*, **6**, 678–685; c) Molecular configuration in sodium thymonucleate. *Nature*, **171**, 740–741.

Hirs, C.H.W., Moore, S. and Stein, W.H. (1960) the sequence of the amino acid residues in performic acid-oxidized ribonuclease. *Journal of Biological Chemistry*, **235**, 633–647.

Hogeweg, P. (1978) Simulating the growth of cellular forms. *Simulation*, **31**, 90–96.

Hogeweg, P. (2011) The roots of bioinformatics in theoretical biology. *PLoS Computational Biology*, **7**(3), e1002021.

Hogeweg, P. and Hesper, B. (1978) Interactive instruction on population interactions. *Computers in Biology and Medicine,* **8**, 319–327.

Kendrew, J.C., Bodo, G., Dintzis, H.M. *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662–666.

Muirhead, H. and Perutz, M. (1963) Structure of hemoglobin. A three-dimensional Fourier synthesis of reduced human hemoglobin at 5.5 Å resolution. *Nature*, **199**, 633–638.

Roos, D.S. (2001) Bioinformatics – Trying to swim in a sea of data. *Science*, **291**, 1260–1261.

Ryle, A.P., Sanger, F., Smith, L.F. and Kitai, R. (1955) The disulphide bonds of insulin. *Biochemical Journal*, **60**(4), 541–556.

Sanger, F. (1945) The free amino groups of insulin. *Biochemical Journal,* **39**, 507–515.
Sanger, F. and Tuppy, H. (1951) **a**) The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.*, **49**, 463–481; **b**) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal,* **49**, 481–490.
Sanger, F. and Thompson, E.O.P. (1953) **a**) The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.*, **53**, 353–366; **b**) The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, **53**, 366–374.
Sanger, F., Thompson, E.O.P. and Kitai, R. (1955) The amide groups of insulin. *Biochemical Journal,* **59**(3), 509–518.
Smith, T.F. (1990) The history of the genetic sequence databases. *Genomics*, **6**, 701–707.
Smyth, D.G., Stein, W.H. and Moore, S. (1963) The sequence of amino acid residues in bovine pancreatic ribonuclease: revisions and confirmations. *Journal of Biological Chemistry*, **238**, 227–234.
Strasser, B. (2008) GenBank – Natural history in the 21st century? *Science*, **322**, 537–538.
Wallace, W. (2001) Bioinformatics: key to 21st century biology. *BioMedNet*, issue **99**, 30 March 2001.
Watson, J.D. and Crick, F.H.C. (1953) Molecular structure of nucleic acids. *Nature*, **171**, 737–738.
Wyckoff, H.W., Hardman, K.D., Allewell, N.M. *et al.* (1967) The structure of ribonuclease-S at 3.5 Å resolution. *Journal of Biological Chemistry*, **242**, 3984–3988.
Zuckerkandl, E. (1987) On the molecular evolutionary clock. *Journal of Molecular Evolution*, **26**, 34–46.
Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, **8**, 357–366.

1.7 Quiz

The following multiple-choice quiz will help you to check how much you've remembered of the origins of bioinformatics and computer science described in this chapter. Be mindful that in this and other quizzes throughout the book – just to keep you on your toes – there may be more than one answer!

1 Who first introduced the term bioinformatics?
- **A** Fred Sanger
- **B** Linus Pauling
- **C** Paulien Hogeweg
- **D** Margaret Dayhoff

2 Who first sequenced a protein?
- **A** Fred Sanger
- **B** Linus Pauling
- **C** Paulien Hogeweg
- **D** Margaret Dayhoff

3 How long did the determination of the sequence of insulin take?
- **A** Five months
- **B** Five years
- **C** Eight years
- **D** Ten years

4 Which was the first enzyme whose amino acid sequence was determined?
- **A** Insulin
- **B** Ribonuclease
- **C** Myoglobin
- **D** Haemoglobin

5 Which was the first protein whose structure was determined?
- **A** Insulin
- **B** Ribonuclease
- **C** Myoglobin
- **D** Haemoglobin

6 Which of the following statements is true?
- **A** The first collection of protein sequences was at the National Center for Biotechnology Information
- **B** The first collection of protein sequences was made by Fred Sanger
- **C** The first collection of protein sequences was the *Atlas of Protein Sequence and Structure*
- **D** None of the above

7 Who was responsible for the invention of the Difference Engine?
- **A** John Napier
- **B** Charles Babbage
- **C** Joseph Marie Jacquard
- **D** George Forsythe

8 What is the smallest number that could be represented using the Babylonian counting scheme?
- **A** One
- **B** Two
- **C** 'Several'
- **D** Zero

9 Binary representation was first conceived by:
- **A** George Boole
- **B** Pingala
- **C** Alan Turing
- **D** John Napier

10 Which of the following statements is true?
- **A** There is a gap between the ability of computers to manipulate concepts and our desire to use them to manipulate binary numbers
- **B** There is a gap between the ability of computers to manipulate primary numbers and our desire to use them to manipulate concepts
- **C** There is a gap between the ability of computers to manipulate binary numbers and our desire to use them to manipulate concepts
- **D** None of the above

1.8 Problems

1 Margaret Dayhoff was one of the pioneers of bioinformatics, actively working in the field in the 1960s, long before the discipline that we know today had even been named. In Section 1.2.3, we described how she was involved in producing the *Atlas of Protein Sequence and Structure*, the first published compendium of protein sequences, one that went on to give life to one of the first protein sequence databases. In addition, she is known for two other pivotal contributions to bioinformatics. What were they?

2 The NCBI, established in 1988, became the new home of the USA's first national nucleotide sequence database in October 1992. What was that database? How many sequences did the database contain when it was first released, and how many were contained in its first release under the auspices of the NCBI?[5] How many sequences does the database contain today?

3 This book is about the gaps we encounter when we explore the interface of bioinformatics and computer science. The nature and types of gap we'll discuss are many and varied: some are subtle and small; others are large and frightening. What is the gap described by Fraser and Dunstan in their 2010 article (*The BMJ*, **342**, 1314–15), and why is it especially disturbing?

[5] Hints: http://www.youtube.com/watch?v=mn98mokVAXI; http://www.ncbi.nlm.nih.gov/genbank/statistics