Preface

The last decade has witnessed the dawn of a new era of 'silicon-based' biology, opening the door, for the first time, to the possible investigation and comparative analysis of complete genomes. In its broadest sense, genome analysis (the quest to elucidate and characterise the genes and gene products of an organism) is underpinned by a number of pivotal concepts, concerning, principally, the processes of evolution (divergence and convergence), the mechanism of protein folding, and, crucially, the manifestation of protein function.

Today, our use of computers to model such processes is limited by, and must be placed in the context of, the current limits of our understanding of these central themes. At the outset, it is important to recognise that we do not yet fully understand the rules of protein folding; we cannot invariably say that a particular sequence or fold has arisen by divergent or convergent evolution; and we cannot necessarily diagnose protein function, given knowledge only of its sequence, or of its structure, in isolation. Accepting what we *cannot* do with computers plays an essential role in forming an appreciation of what, in fact, we can do. Without this kind of understanding, it is easy to be misled, as spurious arguments are often used to promote perhaps rather overenthusiastic points of view about what particular programs and software packages can achieve.

In the field of bioinformatics, the current research drive is to be able to understand evolutionary relationships in terms of the expression of protein function. Two computational approaches have been brought to bear on the problem, tackling the identification of protein function from the perspectives of sequence analysis and of structure analysis respectively. From the point of view of sequence analysis, we are concerned with the detection of relationships between newly determined sequences and those of known function (usually within a database). This may mean pinpointing functional sites shared by disparate proteins (probably the

result of convergent evolution), or identifying related functions in similar proteins (most commonly the result of divergent evolution).

Put like this, the identification of protein function from sequence sounds straightforward, and, indeed, sequence analysis is usually a fruitful technique; but there are two important caveats. First, in 1998, function cannot be inferred from sequence for about one third of proteins in any of the sequenced genomes, largely because biological characterisation cannot keep pace with the volume of data issuing from the genome projects (large numbers of database sequences thus either carry no annotation beyond the parent gene name, or are simply designated 'hypothetical proteins'). Second, in some instances, closely related sequences, which may be assumed to share a common structure, may not share the same function. The classic example here is lysozyme, an enzyme that catalyses the hydrolysis of bacterial cell-wall polysaccharides, which shares around 50% identity (70% similarity) with α-lactalbumin, a non-catalytic regulatory milk protein whose presence is required for lactose synthase to transfer a galactose molecule from UDPgalactose to D-glucose. In spite of this high level of sequence similarity, and highly similar folds, the functions of the proteins differ: the two key catalytic residues of lysozyme (glutamic and aspartic acid) are not conserved in α-lactalbumin; and the acidic calcium-binding motifs characteristic of α-lactalbumins are shared by only a few lysozymes. The lesson here is not that sequence or structure analysis cannot be used as a route to deducing function, but rather that *neither* technique can be applied infallibly without reference to the underlying biology.

The sequence and structure of a protein are clearly different components of its overall functionality – in mathematical terms, we might think of function as being a convolution of sequence and structure. It is perhaps useful to consider a protein fold as providing a basic scaffold, or architecture, and that this framework may be modulated, or decorated, in different ways with different sequences to confer different functions. By analogy, think of a very simple architecture – let us say an empty room. Let us now place within it a table and a chair. At this stage, we cannot tell exactly what the room is for, except perhaps that it is unlikely to be a bedroom or a bathroom. However, if we place a computer on the table, then we could say that it is more likely to be a study, for example, than a dining room. But this is not the full story. Supposing that the room might be a study has not given us an insight into the environment of the room: the study could be a room within a domestic house, where the computer is used largely, say, for keeping accounts; or it could be an office in a company or college perhaps, where the computer is used for running database searches. We only begin to get a full appreciation of the function of the room when we understand, first, how its basic framework is decorated with particular items of furniture and key accessories, and then how its location provides a context for the precise expression of its function.

At a simplistic level, we can think of proteins in much the same way. To take a trivial example, a basic fold, such as a β-barrel (our room), may

occur in many different protein environments, where it may have different functions, depending on how the framework is modulated by a particular protein sequence. In this case, the folding unit may simply be the consequence of convergence to a stable architecture (a room is just a room). In the case of lysozyme and α-lactalbumin, however, which are the result of divergent evolution, both sequence and structure are similar (our room has essentially the same furniture). Here, the different functions of the proteins can only be resolved at the level of *specific* residues that mediate their different activities (i.e., the room's accessories differ – e.g., the computer is a powerful workstation rather than a PC).

The message here is that Nature has her own complex rules, which we only poorly understand, and which we cannot easily encapsulate within computer programs. No current algorithm can 'do' biology. Programs provide mathematical, and therefore infallible, models of biological systems. To interpret correctly whether sequences or structures are meaningfully similar, whether they have arisen by the processes of divergence or of convergence, whether similar sequences, or similar folds, have the same or different functions: these are challenging problems, even for the experienced biologist. There are no simple solutions, and computers do not give us the answers; rather, given a sea of data, they help to narrow the options down so that we, the users, can begin to draw informed, biologically reasonable conclusions.

The issues relating to sequence and structure analysis are many and varied, and separately they constitute vast disciplines. Although, perhaps ideally, investigation of protein function should draw on insights derived from both sequence and structure, we will see that, at the present time, the amount of available structural information is limited by comparison with the quantity of available sequence data. It is clear that the three-dimensional structure of proteins is more faithfully preserved, or conserved, than is the underlying sequence. Thus, seemingly different sequences may adopt remarkably similar folds; in studying evolutionary relationships, this presents the sequence analyst with some daunting challenges. In light of the tidal wave of genome data now before us, these challenges, far from diminishing, are growing in scale. This book therefore considers the problems of detecting the distant relationships sequestered in these oceans of data solely from the perspective of sequence analysis, and, in this context, explores both the possibilities and the current limitations of biology *in silico*.