Preface

The enormous increase in data availability brought about by the genomic projects is paralleled by an equally unprecedented increase in expectations for new medical, pharmacological, environmental and biotechnological discoveries. Whether or not we will be able to meet, at least partially, these expectations it depends on how well we will be able to interpret the data and translate the mono-dimensional information encrypted in the genomes into a detailed understanding of its biological meaning at the phenotypic level.

The major components of living organisms are proteins, linear polymers of amino acids, whose specific sequence is dictated by the genes of the organism. The genes, linear polymers of nucleotides, are directly translated into the linear sequence of amino acids of the encoded protein through a practically universally conserved code. This conceptually simple process is indeed biochemically rather complex, it involves several cellular machineries and is subject to an intricate network of control mechanisms. Even the problem of identifying coding regions in eukaryotic genomes is not completely solved. Far more complex is the identification of the function of the encoded proteins and this will probably be the most challenging problem for the next generations of scientists.

Proteins mediate most of the functions of an organism, and all these functions are, in general, determined by the proteins' three-dimensional structure. Natural proteins spontaneously assume a unique three-dimensional structure. Although this is not a general property of polymers, it is shared, with rare exceptions, by all natural functional proteins. It is achieved by the interplay between molecular evolution and environment-driven selective pressure. Selective pressure acts on function, function requires structure, and therefore protein sequences are selected for being able to fold into a unique structure. This peculiarity of native protein sequences is the reason for their exceptional plasticity and versatility and leads to the exquisite specificity of these macromolecules and to very precise control of their activity through complex networks of interactions. If we could understand the relationship between protein sequence, structure, and function, we could make an effective use of the large body of genetic information available for many organisms, humans included, list all their functions, and study their interplay in shaping life.

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

Protein sequence usually determines protein structure, as was established more than 50 years ago by Christian Anfinsen, the famous Nobel laureate American chemist who was the first to show that a protein can spontaneously refold to its native form and therefore that the information determining the three-dimensional structure of a protein resides in the chemistry of its amino acid sequence. Understanding the underlying rules is, however, a very challenging and as yet an unsolved problem, often referred to as the "holy grail" of biology. The problem has an enormous relevance in many fields, from medicine to biology, from biotechnology to pharmacology and therefore approximate methods for inferring the structure of a protein from its amino acid sequence are flourishing.

Such methods are an essential part of the cultural background and of the toolset of a biologist. There are several structure-prediction tools, easy to use and readily available via internet-based servers, and they are an important contribution of computational biology to the development of the life sciences. All available prediction methods have limitations, however, dictated by the hypotheses on which they rely, and this determines very precisely the field of application of the models produced. Before using any of them it is important to have a clear idea of the biological question for which the model should provide an answer and to verify whether the selected method is sufficiently accurate for the problem at hand. The latter depends upon several factors, and the aim of this book is to provide students and experimental researchers with a guide through the available methods, their underlying assumptions, their limitations, and their expected accuracy in different applications.

Before entering into the heart of the problem, we will discuss the relationship between sequence, structure, and function in proteins, to familiarize the reader with the features of a protein structure (Chapter 1) and with methods and initiatives aimed at evaluating the effectiveness and limitations of prediction methods, and their expected accuracy (Chapter 2).

In Chapter 3 we discuss the extent to which physics can help us compute the structure of a protein on the basis of the knowledge of its chemical structure. The possibility of predicting the structure of a protein from scratch using this approach is, unfortunately, out of our reach at present, and therefore we will explore alternative, knowledge based methods, in subsequent chapters. In Chapter 4 we will survey how a three-dimensional model of a protein can be built on the basis of its evolutionary relationship with a protein of known structure and how to detect the existence of such a relationship. Next, we will discuss how models of proteins can be built taking advantage of the observation that apparently unrelated proteins can share a similar overall structure (Chapter 5) or that, in any case, they share local structure similarities (Chapter 6). Methods which can be used to infer some structural properties of proteins, even if they do not provide us with their complete three-dimensional structure will be surveyed in Chapter 7. These methods are very useful for guiding the construction and assessing the reliability of atomic models obtained with other techniques.

Special treatment must be reserved to membrane proteins. Their peculiar properties, dictated by the environment in which they reside, enable different

methods and strategies to be used for their prediction; these will be discussed in Chapter 8.

Finally, a description of a few successful examples of protein structure modeling will be discussed in the last chapter. The list could be much longer, the literature abounds with scientific reports in which structure prediction is used to guide experiments and to interpret data in the light of a model of the proteins of interest. The examples chosen here partially reflect the personal preference of the author and partially are meant to touch different aspects of the field of protein structure prediction.

The aim of the book is to convince the reader that protein modeling can be extremely useful, if it is performed carefully and with full understanding of the techniques, and that it should be application driven. There is no point in constructing a very approximate model of a protein and later to use it to derive detailed properties about catalytic mechanisms or interaction details; similarly, it is not wise to employ very sophisticated techniques to obtain a model of a protein that, for technical reasons, cannot be used for guiding experiments or for casting light on important biological questions.

Rome, November 2005 *Anna Tramontano*