

Preface

Since the pioneering discovery of Watson and Crick, the scientific community has come to accept the general principle that biological information is stored in an encrypted code within the base sequence of DNA. As in other forms of classical cryptography, gene transcription and protein translation are therefore subject to rules of linguistics, since they convert information from one linear sequence (nucleic acids) into another (amino acids). In a similar, albeit more complex way, proteins can also be described by a set of linguistic 'rules' governing semantics, grammar, and syntax. However, the cryptographic complexity of proteins encodes not only sequence, but somehow explicitly specifies folding, structure, and biological function as well. How, then, can one learn to read this 'language of proteins'?

One of the most powerful approaches to 'cracking the protein code' has involved sequence comparisons between and within species, a task now greatly simplified by the ever-expanding genome sequencing efforts. Early efforts in this area led to a realization that proteins can be grouped into families and superfamilies, in which individual family members perform related, though not identical, tasks. The second remarkable insight was that unrelated proteins often share significant portions of sequence similarity and that these regions of similar sequence function as independently folded modules or domains capable of performing a specific biochemical reaction or event. Thus, a large number of functionally diverse proteins can be thought of as molecules built by combining a limited number of structurally-stable folded domains. Within these multidomain proteins, each domain is capable of autonomous function but also interacts with the other domains to allow for higher levels of regulation and control.

In this light, the human proteome becomes the equivalent of a dictionary, with the individual proteins corresponding to single entries and the modular domains akin to the words that constitute each entry. The task that lies before us, learning how to read the dictionary and decipher the rules of 'protein linguistics', then becomes a bit simpler – we can start by learning to read the words – that is, recognizing the domains and understanding their functions. Once we accomplish this, we can extend our understanding beyond simple word (domain) recognition, to understanding the biological meaning conveyed in each sentence. This book is an attempt to codify the 'words' of the signaling language spoken by proteins in living organisms: the modular protein-interaction domains, nodules, and Lego blocks described in the Prologue by Sir Tom Blundell or the sockets as Harel Weinstein defines them in his inspiring Epilogue. The compilation of articles in

this book provides only a glimpse of the current knowledge of modular protein domains – as evidenced by the fact that we can recognize some ‘protein sentences’ as conveying large amounts of information while others seem to have no identifiable words at all. It is our sincere hope that this book will stimulate others to extend our understanding of protein modules by identifying and characterizing new domains and by discovering the grammatical rules by which domains interconnect to convey a broader biological meaning by the proteins that contain them.

Leonardo da Vinci told us: “Realize that everything connects with everything else”. Indeed it does, in any integrated and efficient system that responds to the environment and has an intrinsic strategy to be flexible, grow, and adapt. Whether the system is biological, human-engineered, or a combination of both (see the Epilogue), it usually has a complex ‘wiring’ that interconnects its multiple primary elements to make the system an integrated whole. Modular protein-interaction domains are these primary integrating elements.

The representative repertoire of protein- and lipid-interacting domains are described here by experts in the field, including bench scientists, who originally defined them or significantly contributed to their characterization, and bio-informaticians, structural biologists, and protein chemists, who were instrumental in the rapid progress of the protein domain field. The Prologue and Epilogue in the book reflect this ‘perfect marriage’.

As editors we thank all the contributors very much. When we approached the potential authors with invitations, several of them remarked that they do not write chapters for books, they publish original research reports only, but in the same breath they said that they would gladly do it. Although we still do not know what exactly changed their minds, we appreciate their decisions and efforts. The exciting ‘intellectual covers’ of the book, the Prologue and Epilogue, were written after most of the chapters were completed, when we sent Sir Tom Blundell and Harel Weinstein a large package of partially edited manuscripts (that at least temporarily overflowed their e-mail inboxes) to comment upon. We thank them for their insightful and clearly inspiring comments. We see their thoughts as trend-setting vectors of current and future research on modular protein-interaction domains. The editorial process reminded us that in constructing a book about our field, as in primary laboratory research, we are never experts, just permanent students. Thank you, we learned a lot. Finally, we owe a large debt of gratitude to Frank Weinreich and the entire Wiley-VCH team for invaluable assistance.

To our readers: Enjoy!

August 2004

Gianni Cesareni
Mario Gimona
Marius Sudol
Michael Yaffe