# INTRODUCTION

# Bioinformatics:
## Mystery, Astrology or Service Technology?

**Frank Eisenhaber***

## Abstract

Mathematical interpretation and integration of experimental data for the goal of biological theory development has had little, if no, impact on previous progress in life sciences compared with the sophistication of experimental approaches themselves. The genesis of recent spectacular breakthroughs in molecular biology that led to the discovery of the enzymatic function of several nonmetabolic enzymes illustrates that this relationship is beginning to change.

The development of high-throughput technologies, for example of complete genome sequencing, leads to large amounts of quantified data on biological systems without direct link to biological function that require formalized and complex mathematical approaches for their interpretation. The research success in life sciences depends increasingly on the ability of researchers in experimental and theoretical biology to jointly focus on important questions. Currently, theoretical methods have best chances to contribute to new biological insight independently of experiments in the area of genome text interpretation and especially for gene function prediction. Experimental studies can help progress in the development of theoretical methods by providing verified, sufficiently large and variable sequence datasets for the exploration of sequence-function relationships.

## Introduction

To caricature, the typical research process in life sciences consists of periodic repetitions of weeks/months of bench work by a PhD or postdoctoral student followed by an hour of looking at the results by the lab head after which the coworker again disappears into the cold room or behind the microscope with new directives. Generations of life scientists have been educated that the most important goal consists in producing "hard", quantitative experimental data describing biological structures and processes. Pure theoretical efforts directed at biological data analysis are believed to add little more than intellectual speculation or a colorful illustration in the form of a graph or an alignment. The biological theory itself has remained logically simple and with little or no mathematics or formal structure. Typically, all creativity has been directed into sophistication and rationalization of experimental procedures and techniques for the wet lab. This type of life science has successfully produced breakthroughs and will, apparently, continue to stay the major source of new biological insight in the near future.

*Frank Eisenhaber—Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna, Republic Austria. Email: Frank.Eisenhaber@imp.univie.ac.at

*Discovering Biomolecular Mechanisms with Computational Biology*,
edited by Frank Eisenhaber. 

This situation is especially astonishing for people that come from more formalized sciences such as physics where a typical experiment is preceded by months of calculations and computer simulations. Such research is necessary in this area to derive the most interesting research targets and to check the consistency of new hypotheses with existing knowledge. There had been several waves of efforts to inject mathematics into life sciences, for example statistics (beginning with Mendel's ratios), kinetics (of enzymes and ligand binding, of transport systems, in population dynamics) or 3D biomolecular structure modeling (together with quantum chemistry, QSAR studies and molecular dynamics; especially in context with the hypothesis of DNA double-helical structure). Although each of these waves have enriched life sciences in aspects, neither one has become a continuous source of qualitatively new biological knowledge or has made biology a truly theoretical, a quantitative and predictive science.

Beginning with the 1960s, yet another stream of efforts focused on the esoteric topic of analysis of text strings representing the monomer sequences of proteins and nucleic acids and of the evolution of these strings after multiple single-point mutations.[1,2] Thanks to these pioneering efforts, theoretical concepts, computational methods and sequence databases have been established that allow the prediction of function for experimentally uncharacterized genes from their sequence, most importantly, together with the quantification of the prediction error (prediction reliability) in probabilistic terms.[3] The impact of this development is perceived in different ways by various parts of the generally wet lab-focused life science community depending on personal background and experience: as *mystery*, *astrology* or *service technology*. None of these three ways is a really appropriate assessment for the recent step in the difficult development of life sciences towards a formalized theory of living systems as the discussion below will attempt to show.

## Mystery

Sometimes, success stories are sensed euphorically as a *mystery* by those scientists that receive a tremendous boost in their experimental work from a function prediction. At the background of general weakness of theory in life sciences, it is indeed perceived as a bolt from the blue by the experimental life science research community that a number of recent scientific breakthroughs in biology have originated from theoretical studies for gene function prediction. Several instances of discoveries of enzyme activities for a number nonmetabolic proteins, typically without any previous hint or suspicion from experimental findings, are remarkable evidence for the growing predictive power of theoretical biology.

Important science-organizational and cognitive aspects of this process towards new biological knowledge can be illuminated by viewing some recent examples of enzymatic function assignment to nonmetabolic enzymes. Three stories with considerable biological impact, namely

1. the discovery of the molecular function of Fringe in Notch signaling,
2. the determination of the protease domain of separin triggering the transition from metaphase to anaphase during the cell cycle, and
3. the understanding of heterochromatin formation as initiated by the histone methyltransferase activity of the Su(var)3-9 homologues, are described in brief in Boxes 1, 2 and 3.

There are more of such nontrivial findings, and it is not possible to give a complete list here. For example, a C-terminal domain in yeast protein dot1p was assigned to the SAM sequence family with suggested methyltransferase activity. The loss-of-function phenotype of the dot1 gene (disruption of telomere silencing) implied a possible role in the posttranslational modification of histones.[4] Indeed, a biochemical assay was able to show that dot1p does methylate histone H3 at Lys79.[5] As the authors acknowledge, the previously published theoretical report was critical for their decision to launch the experimental test.

In another case, the yeast protein eco1p was found critical for the establishment of cohesion between sister chromatids,[6] but the biological experiments did not give any hint with respect to a possible molecular function of eco1p. Sequence analysis studies pointed to an