Crimson Publishers
Wings to the Research

Opinion

Missing in the Big Data Era: Relative specificity

Xiaoxiao Zhang and Yan Zeng*

Department of Zoology, College of Life Sciences, Nanjing Agricultural University, Nanjing, China

***Corresponding author:** Yan Zeng, Department of Zoology, College of Life Sciences, Nanjing Agricultural University, Nanjing, China

Submission: September 01, 2021
Published: September 21, 2021

Volume 3 - Issue2

How to cite this article: Xiaoxiao Zhang and Yan Zeng*. Missing in the Big Data Era: Relative specificity. J Biotech Biores. 3(2). JBB. 000557. 2021.

Big Data and Relative Specificity

High-throughput methods, single-cell analysis, and machine learning are some of the major technological breakthroughs that have transformed biology and ushered in the new era of big data. It is now possible to predict with a high degree of confidence protein structures based on amino acid sequences [1] and gene functions and disease etiology are frequently discussed through the prism of networks. Genomics techniques, the mainstay and often the starting point of research in the 21st century, generate massive data from which elaborate relationships among genes can be inferred and later verified experimentally. In a typical study, hundreds of hits may be produced, but only a small subset will be selected and then investigated in depth based on criteria such as the levels of expression changes, perceived gene functions, or the pathways involved. This dramatic target narrowing has a strong, pragmatic reason: experimentation with individual genes is still time and effort intensive. But there is another reason: the lack of consideration for relative specificity.

In complex biochemical systems that involve, e.g., the transcription machinery, kinases, RNA-binding proteins, microRNAs (miRNAs), and the ribosome, an enzyme, protein, or RNA has hundreds or thousands of substrates or interacting partners. The relative specificity hypothesis posits that the enzyme, protein, or RNA interacts with and influences its distinct substrates differentially, thereby regulating the underlying biological processes [2]. Although the notion of relative specificity is intuitive, explicit evidence is scarce in the literature, even rarer is the direct demonstration of its physiological significance, as its study necessitates comparing the interactions between an enzyme and its hundreds of substrates as well as the *in vivo* effects [2]. The idea of "quantitative continua" in transcriptional regulation has been proposed [3]. Human Drosha cleaves hundreds of primary miRNA transcripts preferentially, contributing to global differential miRNA expression [4]. The fission yeast CDK phosphorylates its good substrates early to promote DNA replication, and the poor ones later to promote cytokinesis [5]. If it phosphorylates the poor substrates prematurely, cells will divide without DNA replication and subsequently die, demonstrating the functional consequence of relative specificity [5]. Thus, CDK controls cell cycle progression not only by phosphorylating its substrates, but also by selectively phosphorylating them at different times according to relative specificity. Yet despite these examples and others [2,6], relative specificity and its functions remain underappreciated due to historic, conceptual, and practical reasons [2]. The goal using big data is to dissect biological systems in an unprecedentedly comprehensive manner. In most work, however, relative specificity has not been actively considered, and the relevant data either not produced or harnessed. Even when high-throughput techniques yield the information about relative specificity, the aforementioned, standard practice of target narrowing will discount most of such information. Consequently, the resultant conclusions or models lack vital insights and are seriously incomplete. Below, four examples are used to illustrate how pursuing relative specificity might supplement or enhance our knowledge as well as discover novel mechanisms.

What Relative Specificity Can Tell Us

The first example is: what determines differential mRNA expression, which largely depends on gene transcription [7]? This subject has been the most extensively investigated using high-throughput techniques, including by the ENCODE project [8]. Yet it remains impossible to explain or predict, e.g., why gene A produces more mRNA than gene B, at the

genome level. Since genes vary in their amounts of transcription factor binding and DNA and chromatin modifications, i.e., relative specificity, naturally identified by ChIP-seq, one can combine and model these variations systematically, and then correlate to mRNA levels [6-8]. This unifying scheme would address the fundamental question about gene expression.

The second concerns enzymes and substrates. Enzymes routinely have hundreds of cognate substrates, and for a lot of enzymes, many if not most substrates have been identified by genomics and proteomics means. In a handful of enzymes relative specificity has been examined [4-6], but for the vast majority the focus of most research laboratories remains to study their own favorable substrate(s) individually. While this approach yields detailed information on the functions of these substrates, it does not avail the full picture about the function of the enzyme, even when considering the results by all the laboratories collectively. For instance, the traditional strategy cannot unveil the mechanism of the temporal role by CDK in the fission yeast cell cycle [5], or demonstrate that human Drosha, an essential enzyme that cleaves canonical primary miRNAs, further regulates differential miRNA expression [4].

The third relates to gene functions in general and in disease. As examples, how does p53 or c-Myc regulate carcinogenesis, and why does a ubiquitous loss of SMN affect neurons but not other cells [9]? For the first question, a large number of target genes of p53 and c-Myc have been studied over the years, but no single genes have been found to play a dominant role. For the second question, presumably SMN has a yet-to-be-identified target(s) exclusively in motor neurons. In light of relative specificity, it is plausible that p53 or c-Myc mutations induce unbalanced changes in target genes compared to the normal conditions, resulting in cancers, or reduced SMN causes SMN targets' unbalanced changes in neurons that differ from those in other cells. Hence new research avenues to explore: the key is not individual target(s) but the relative levels or changes of many targets.

Lastly, is relative specificity evolutionarily conserved, and what is its impact on evolution? If relative specificity is a general principle, it should be conserved, although there is no evidence currently due to the lack of study. And if it is conserved, relative specificity will constitute a new layer of controlling mechanisms and complexity, for natural selection may act on relative specificity besides the more visible sequence constraints to maintain stability as well as foster evolution in biology.

Conclusion

Relative specificity is a simple yet often overlooked concept, and studying relative specificity will uncover new regulatory mechanisms. Modern technologies and research have generated tremendous amount of data, now further used by artificial intelligence to unravel myriad biological systems. But these efforts have critical pitfalls because in almost all cases relative specificity data are entirely missing or not sufficiently utilized. Fortunately, relative specificity can be studied with current technologies and the right thoughts [2], and its incorporation in the big data era will allow us to understand and model complex biological systems more precisely and completely.

References

1. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, et al. (2021) Highly accurate protein structure prediction for the human proteome. Nature 596(7873): 590-596.
2. Zeng Y (2011) The functional consequences and implications of relative substrate specificity in complex biochemical systems. Front Genet 2: 65.
3. Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. Dev Cell 21(4): 611-626.
4. Feng Y, Zhang X, Song Q, Li T, Zeng Y (2011) Drosha processing controls the specificity and efficiency of global microRNA expression. Biochim Biophy Acta 1809(11-12): 700-707.
5. Swaffer MP, Jones AW, Flynn HR, Snijders AP, Nurse P (2016) CDK substrate phosphorylation and ordering the cell cycle. Cell 167(7): 1750-1761.
6. Zeng Y (2014) Relative specificity: All substrates are not created equal. Genomics Proteomics Bioinformatics 12(1): 1-7.
7. Li JJ, Bickel PJ, Biggin MD (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ 2: e270.
8. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414): 57-74.
9. Nussbacher JK, Tabet R, Yeo GW, Lagier-Tourenne C (2019) Disruption of RNA metabolism in neurological diseases and emerging therapeutic interventions. Neuron 102(2): 294-320.