Do post-translational modifications control cell differentiation?

Nancy Kendrick, Kendrick Labs, Inc., Madison, WI 53713 nancy@kendricklabs.com

Abstract

The four human tissue types (epithelial, muscle, nervous and connective) are comprised of about 200 different cell types all with the same genome but radically different morphologies [1]. While Alberts et al. describe seven control points of cell differentiation in the textbook "Molecular Biology of the Cell" [2] (Table 2), the details remain unclear. Here we propose that proteins from at least 15 human gene families (over 2100 genes) control cell differentiation *via post-translational modifications* that regulate timing, abundance, and function of target proteins.

Introduction

We know that the Central Dogma, "DNA makes RNA makes Protein," is the basis for all life on earth [3], and also that DNA mutations are the root cause of cancer [4]. However, as noted by Roehrl et al., "Proteins are the true machines of life. Protein enzymes carry out virtually all complex chemical transformations in living organisms such as nucleic acid synthesis and replication, post-translational modifications (PTM), carbohydrate and lipid metabolism, hormone biosynthesis proteolysis, and many more" [5]. PTMs such as tyrosine phosphorylation are responsible in turn for animal multicellularity [6].

mRNA ≠ protein: Considerable evidence has accrued showing expressed protein levels are mostly unrelated to mRNA levels. Cao et al. reported that for a proteogenomic characterization of 140 pancreatic cancer samples, the median correlation (Pearson's correlation coefficient) between mRNA and protein levels was 0.35 [7]. Stetson et al. reported that the average mRNA/protein agreement for 27 snap-frozen glioblastoma tumors was 0.22 [8]. Even for normal tissues the agreement between mRNA and protein expression is poor. Wang et al. reported that for a proteome/ transcriptome analysis of 13,413 protein coding genes in 29 healthy human tissues, the median correlation was 0.35. Some proteins could not be detected for highly expressed mRNAs, and some mRNAs could not be detected for highly expressed proteins [9].

The discrepancy between mRNA and protein levels is likely the consequence of gradual metazoan evolution over ~700 million years [10]. The four human tissue types: epithelial, muscle, nervous, and connective, are made up of about 200 distinct cell types [11], all with the same nuclear genome [12] but different expressed proteins. Thus, ***"When and where a protein is synthesized in an organism is as important as the protein's function"*** [13, 14].

What part do PTMs play?

Genome analysis has revealed the existence of multiple gene families whose expressed proteins bring about post-translational modifications (PTMs) of other proteins as shown in Table 1. The family members, scattered throughout the genome, are likely expressed at different times and in different tissues during embryonic development. For example, about 5% of human genes (588 proteases and 377 E3 ubiquitin ligases) code for proteins that break down other proteins. This speaks to the importance of protein turnover in mammalian tissues.

Other post-translational modifications (PTMs) are key to protein interactions. Phosphorylation is the most common PTM [15]. Interestingly, while there are many more serine/threonine kinases (428) than phosphatases (30),the reverse is true for tyrosine: there are more tyrosine *phosphatases* (107) than kinases (90). Since trans-phosphorylation of key tyrosine residues on RTKs creates binding sites that control the actions of many cytoplasmic signaling proteins [4], dephosphorylation of pTyr-RTKs is an important regulatory event. Malfunctions of tyrosine phosphatases are likely oncogenic, as reviewed by Zhao et al. [16].

Human protein families that perform key primary functions (RBP, TF, GPCR, IF, Claudins), or activate/inactivate other proteins via PTMs, shaded)	Number of family members
RNA-binding proteins (RBP) [17]	2,961
Transcription factors (TF) [18]	1,639
G protein-coupled receptors (GPCR) [19]	826
Intermediate filaments [20] (IF, cell shape)	70
Claudins (tight junctions) [21]	23
Proteases [22]	588
Serine/threonine kinases [23]	**428**
Serine/threonine phosphatases [24]	**30**
E3 ubiquitin ligases [25]	377
Glycosyltransferases [26]	244
Tyrosine phosphatases [27]	**107**
Tyrosine kinases (58 RTK & 32 non-R-TK [28]	**90**
Histone acetylation modulator proteins [29]	73
2-OG-dependent dioxygenases [30]	70
Histone lysine methyltransferases [31]	50
DHHC3 Palmitoyltransferases [32]	23
Protein Disulfide isomerases	21
Glutathione S-transferases [33]	16
Sulfotransferases [34]	13
Aldehyde dehydrogenases	12
NSD protein methyltransferases [35]	3
EGLN prolyl hydroxylase [36]	3
Total	7667 (40%)
Total shaded	2148 (11%)
Total protein-coding genes in human genome [37]	19,202

Table 1. Gene families whose expressed proteins regulate cell differentiation by controlling activity, expression, and/or PTMs (shaded rows) of other proteins. TFs control which genes are transcribed into mRNA. RBPs regulate the translation of mRNA transcripts into proteins. GPCRs transduce extracellular signals into physiological effects via G proteins. IFs control cell shape and migration. The next fifteen rows (shaded) show protein families with 2115 members that post-translationally modify other proteins. *Note that this table was compiled from hand-collected literature references and is incomplete*. It shows only 40% of protein-coding genes.

The fact that the serine-threonine kinase/phosphatase ratio (428/30) is so high suggests that this post-translational modification (PTM) is common and mostly irreversible. This hypothesis is supported by the fact that serine/threonine phosphorylation is the most abundant PTM of the human proteome. About 13,000 (68%) of the ~19,000 proteins in the human genome are phosphorylated on either serine or threonine residues [38].

In the textbook "Molecular Biology of the Cell" (7th edition, p 401), Alberts et al. discuss seven control points for pathways that lead to cellular differentiation [2] Fig. 7-8. The fact that the PTM protein families in Table 1 match well to control points described by Alberts et al. (Table 2) suggests that PTM timing is key to cell differentiation.

Control Point	Cell Process from Molecular Bio textbook [2]	Controlling Protein Family	# of Family Members	Ref
1	DNA transcription	Transcription Factors	1,639	[18]
2	RNA-processing including alternative splicing	RNA Binding Proteins	2,961	[17]
3	mRNA transport/localization, from nucleus to cytosol			
4	mRNA translation into protein			
5	mRNA degradation			
6	protein degradation	Proteases	588	[22]
		E3 ubiquitin ligases	377	[25]
7	protein activity control	G protein-coupled receptors	826	[19]
		serine/threonine kinases	428	[23]
		serine/threonine phosphatases	30	[24]
		tyrosine kinases	**90**	[28]
		tyrosine phosphatases	**107**	[27]
		glycosyl transferases	244	[26]
		Histone acetylation	73	[29]
		2-OG-dependent dioxygenases	70	[30]
		Histone lysine methyltransferases	50	[31]
		Glutathione S-transferases	16	[33]
		Sulfotransferases	13	[34]
		NSD methyltransferases	3	[35]
		EGLN prolyl hydroxylase	3	[36]
	Cell shape & migration	Intermediate filaments	70	[20]
		Total PTMs (control points 6-7)	2,988	
		Total protein-coding genes in the human genome	19,202	[37]

Table 2. Correlation of cell processes described in the textbook by Alberts et. al. [2] by protein PTM families comprising 40% of the human genome. A huge family of 1639 transcription factors bind DNA to stimulate or inhibit gene transcription (control point 1). An even larger family of 2961 RNA binding proteins regulates mRNA alternative splicing (control points 2-5). PTMs, including proteases (588 members) and ubiquitin ligases (377 members) regulate protein degradation (control point 6). A myriad (428) of kinases regulate serine/threonine phosphorylation while 90 tyrosine kinases and 107 tyrosine phosphatases, often cancer drivers, cell growth.

Discussion and Conclusions

The root cause of cancer is DNA alterations that lead to uncontrollable growth [39]. However, linking specific DNA "driver" mutations to drug targets is often difficult because, as noted, mRNA transcript abundances correlate poorly with protein abundances for both healthy [9] and cancerous [8, 40] human tissues.

At last count, 40 cancer drugs are available that target receptor tyrosine kinase proteins [41]. Regardless of NGS mutation status and drug availability, however, RTK proteins must be expressed, dimerized by the presence of the corresponding growth factor (e.g., epidermal growth factor) and specific tyrosines trans-phosphorylated to allow binding of signaling proteins containing Src Homology 2 (SH2) and phosphotyrosine-binding (PTB) domains [4]. Mutated RTK proteins lacking tyrosine phosphorylation will be inactive, and the matching cancer drug will fail. Thus, correlation of **pTyr-RTK proteoforms** in cancer biopsies with NGS mutation patterns would likely assist in finding *actionable* genome biomarkers that predict drug response.

References

1. Ovalle, W.K. and P.C. Nahirney, *Netter's essential histology : with correlated histopathology*. Third edition. ed. 2021, Philadelphia, PA: Elsevier. xiii, 552 pages.
2. Alberts, B., *Molecular biology of the cell*. Seventh edition. ed. 2022, New York: W. W. Norton & Company. pages cm.
3. He, C., *Special Issue on Regulating the Central Dogma.* Biochemistry, 2019. **58**(5): p. 295-296.
4. Weinberg, R.A., et al., *The biology of cancer*. Third edition, International student edition. ed. 2023, New York, N.Y: W.W. Norton & Company. xxi, 865, 8, 48, 40 pages : illustrations (chiefly colour).
5. Roehrl, M.H., V.B. Roehrl, and J.Y. Wang, *Proteome-based pathology: the next frontier in precision medicine.* Expert Rev Precis Med Drug Dev, 2021. **6**(1): p. 1-4.
6. Tong, K., Y. Wang, and Z. Su, *Phosphotyrosine signalling and the origin of animal multicellularity.* Proc Biol Sci, 2017. **284**(1860).
7. Cao, L., et al., *Proteogenomic characterization of pancreatic ductal adenocarcinoma.* Cell, 2021. **184**(19): p. 5031-5052 e26.
8. Stetson, L.C., et al., *Proteins inform survival-based differences in patients with glioblastoma.* Neurooncol Adv, 2020. **2**(1): p. vdaa039.
9. Wang, D., et al., *A deep proteome and transcriptome abundance atlas of 29 healthy human tissues.* Mol Syst Biol, 2019. **15**(2): p. e8503.
10. Knoll, A.H. and M.A. Nowak, *The timetable of evolution.* Sci Adv, 2017. **3**(5): p. e1603076.
11. Ovalle, W.K. and P.C. Nahirney, *Netter's essential histology: : with correlated histopathology*. 3rd. ed. 2020, Philadelphia: Elsevier. pages cm.
12. Gabaldon, T., *Origin and Early Evolution of the Eukaryotic Cell.* Annu Rev Microbiol, 2021. **75**: p. 631-647.
13. Fay, J.C. and P.J. Wittkopp, *Evaluating the role of natural selection in the evolution of gene regulation.* Heredity (Edinb), 2008. **100**(2): p. 191-9.
14. Brunet, T. and N. King, *The Origin of Animal Multicellularity and Cell Differentiation.* Dev Cell, 2017. **43**(2): p. 124-140.
15. Pang, K., et al., *Role of protein phosphorylation in cell signaling, disease, and the intervention therapy.* MedComm (2020), 2022. **3**(4): p. e175.
16. Zhao, M., et al., *Protein tyrosine phosphatases: emerging role in cancer therapy resistance.* Cancer Commun (Lond), 2024. **44**(6): p. 637-653.

17. Liao, J.Y., et al., *EuRBPDB: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic RNA binding proteins (RBPs).* Nucleic Acids Res, 2020. **48**(D1): p. D307-D313.
18. Lambert, S.A., et al., *The Human Transcription Factors.* Cell, 2018. **172**(4): p. 650-665.
19. Yang, D., et al., *G protein-coupled receptors: structure- and function-based drug discovery.* Signal Transduct Target Ther, 2021. **6**(1): p. 7.
20. Szeverenyi, I., et al., *The Human Intermediate Filament Database: comprehensive information on a gene family involved in many human diseases.* Hum Mutat, 2008. **29**(3): p. 351-60.
21. Lal-Nag, M. and P.J. Morin, *The claudins.* Genome Biol, 2009. **10**(8): p. 235.
22. Kappelhoff, R., et al., *Overview of transcriptomic analysis of all human proteases, non-proteolytic homologs and inhibitors: Organ, tissue and ovarian cancer cell line expression profiling of the human protease degradome by the CLIP-CHIP DNA microarray.* Biochim Biophys Acta Mol Cell Res, 2017. **1864**(11 Pt B): p. 2210-2219.
23. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.
24. Shi, Y., *Serine/threonine phosphatases: mechanism through structure.* Cell, 2009. **139**(3): p. 468-84.
25. Medvar, B., et al., *Comprehensive database of human E3 ubiquitin ligases: application to aquaporin-2 regulation.* Physiol Genomics, 2016. **48**(7): p. 502-12.
26. Joud, M., M. Moller, and M.L. Olsson, *Identification of human glycosyltransferase genes expressed in erythroid cells predicts potential carbohydrate blood group loci.* Scientific reports, 2018. **8**(1): p. 6040.
27. Alonso, A., et al., *Protein Tyrosine Phosphatases in the Human Genome.* Cell, 2004. **117**(6): p. 699-711.
28. Robinson, D.R., Y.M. Wu, and S.F. Lin, *The protein tyrosine kinase family of the human genome.* Oncogene, 2000. **19**(49): p. 5548-57.
29. Hu, Z., et al., *Genomic characterization of genes encoding histone acetylation modulator proteins identifies therapeutic targets for cancer treatment.* Nat Commun, 2019. **10**(1): p. 733.
30. Zurlo, G., et al., *New Insights into Protein Hydroxylation and Its Important Role in Human Diseases.* Biochim Biophys Acta, 2016. **1866**(2): p. 208-220.
31. Bennett, R.L., et al., *The Role of Nuclear Receptor-Binding SET Domain Family Histone Lysine Methyltransferases in Cancer.* Cold Spring Harb Perspect Med, 2017. **7**(6).
32. Gottlieb, C.D., S. Zhang, and M.E. Linder, *The Cysteine-rich Domain of the DHHC3 Palmitoyltransferase Is Palmitoylated and Contains Tightly Bound Zinc.* J Biol Chem, 2015. **290**(49): p. 29259-69.
33. Nebert, D.W. and V. Vasiliou, *Analysis of the glutathione S-transferase (GST) gene family.* Hum Genomics, 2004. **1**(6): p. 460-4.
34. Gamage, N., et al., *Human sulfotransferases and their role in chemical metabolism.* Toxicol Sci, 2006. **90**(1): p. 5-22.
35. Murali, M. and V. Saloura, *Understanding the Roles of the NSD Protein Methyltransferases in Head and Neck Squamous Cell Carcinoma.* Genes (Basel), 2022. **13**(11).
36. Strocchi, S., et al., *The multifaceted role of EGLN family prolyl hydroxylases in cancer: going beyond HIF regulation.* Oncogene, 2022. **41**(29): p. 3665-3679.
37. Amaral, P., et al., *The status of the human gene catalogue.* Nature, 2023. **622**(7981): p. 41-47.
38. Vlastaridis, P., et al., *Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes.* Gigascience, 2017. **6**(2): p. 1-11.
39. Martinez-Jimenez, F., et al., *A compendium of mutational cancer driver genes.* Nat Rev Cancer, 2020. **20**(10): p. 555-572.

40. Paik, P.K. and C.M. Rudin, *Missing the mark in FGFR1-amplified squamous cell cancer of the lung.* Cancer, 2016. **122**(19): p. 2938-40.
41. Roskoski, R., Jr., *Properties of FDA-approved small molecule protein kinase inhibitors: A 2023 update.* Pharmacol Res, 2023. **187**: p. 106552.