

Organic codes and their identification: Is the histone code a true organic code?

by

Stefan Kühn

*Thesis presented in partial fulfilment of the requirements
for the degree of Master of Science (Biochemistry) in the
Faculty of Science at Stellenbosch University*



Department of Biochemistry
University of Stellenbosch
Private Bag X1, 7602 Matieland, South Africa

Supervisor: Prof. J.-H.S. Hofmeyr (supervisor)

March 2014

Acknowledgements

- Prof. Hofmeyr, your boundless enthusiasm for life, the universe, and everything else has been a source of inspiration for me.
- Marcello Barbieri for kickstarting code biology and his fiery conviction.
- The NRF of South Africa for funding.
- My family for unquestioning support.
- The office — Chris, Jaléne, it's been a pleasure.
- Meghan and Sarah, for keeping the faith.

Meinen Eltern, meiner Oma, und Bienchen — ohne euch wäre dieses Werk
nie entstanden

Contents

Declaration	i
Contents	iv
List of Figures	vi
List of Tables	viii
Summary	ix
Opsomming	xi
1 Introduction	1
2 Code biology	5
2.1 On codes	9
2.2 Evolution by natural conventions	12
2.3 Information	14
3 Some organic codes	18
3.1 The genetic code	21
3.2 The metabolic code	23
3.3 The signal transduction code	25

3.4	The sugar code	27
3.5	The splicing code	28
3.6	The ubiquitin code	30
3.7	The compartment code	31
3.8	The regulatory code	32
3.9	The <i>Hox</i> code	34
4	The histone code	36
4.1	What are histones and the ‘histone code’?	37
4.2	The function of post-translational histone modifications	38
4.3	The histone post-translational modification zoo	39
	Acetylation	41
	Methylation	45
	Ubiquitylation	50
4.4	Binding domains: The adaptors of the histone code	52
	Acetyl-recognising domains	52
	Methyl-recognising domains	53
	Ubiquitin-recognising domains	54
4.5	How does it all fit together? Is the histone code an organic code?	54
4.6	Criticisms of the histone code model	59
5	The Görlich-Dittrich algorithm for identifying ‘molecular codes’: A critique	63
6	Discussion	69
	Bibliography	74

List of Figures

2.1	Mappings f and g between a set of nucleotide triplets and a set of amino acids	7
4.1	The structure of a nucleosome.	37
4.2	Major sites of post-translational modifications of histones H2A, H2B, H3 and H4. Symbols: A denotes acetylation, M methylation and U ubiquitylation. Sites on the polypeptide chains are numbered and identified with the one-letter abbreviations of their amino acids.	40
4.3	A: In the absence of acetyl groups on lysine 9 and 14 on histone 3, the double bromodomains of TAFII250 are unable to bind to H3K9 and H3K14 and as a result, TAFII250 does not phosphorylate TAFIIF, which in turn does not lead to transcriptional initiation. B: once H3K9 and H3K14 have been acetylated (red circles), the double bromodomains are now able to recognise and bind the H3K9ac and H3K14ac, which allows TAFII250 to phosphorylate TAFIIF and thus permit transcription to proceed. . .	58
5.1	A binary molecular code according to Görlich and Dittrich [59]. The set, $S = \{A, B\}$, is mapped to the set $M = \{C, D\}$ by the contexts, $C = \{E, G\}$ or $C' = \{F, H\}$	64

5.2 The mapping of the reaction network mapping D-Glucose and D-Talose onto D-Mannose and D-Galactose	68
--	----

List of Tables

3.1	The various proposed organic codes, the independent worlds that they link, and their adaptors.	19
3.2	The mRNA/amino acid translation scheme	22
4.1	The major histone <i>acetylations</i> and <i>ubiquitylations</i> , the binding domains that specifically recognise them, and their corresponding cellular effects.	42
4.2	The major histone <i>methylications</i> , the binding domains that specifically recognise them, and their corresponding cellular effects.	46

Summary

Codes are ubiquitous in culture—and, by implication, in nature. Code biology is the study of these codes. However, the term ‘code’ has assumed a variety of meanings, sowing confusion and cynicism. The first aim of this study is therefore to define what an *organic code* is. Following from this, I establish a set of criteria that a putative code has to conform to in order to be recognised as a true code. I then offer an information theoretical perspective on how organic codes present a viable method of dealing with biological information, as a logical extension thereof.

Once this framework has been established, I proceed to review several of the current organic codes in an attempt to demonstrate how the definition of and criteria for identifying an organic code may be used to separate the wheat from the chaff. I then introduce the ‘regulatory code’ in an effort to demonstrate how the code biological framework may be applied to novel codes to test their suitability as organic codes and whether they warrant further investigation.

Despite the prevalence of codes in the biological world, only a few have been *definitely* established as organic codes. I therefore turn to the main aim of this study which is to cement the status of the histone code as a *true* organic code in the sense of the genetic or signal transduction codes. I provide a full review and analysis of the major histone post-translational

modifications, their biological effects, and which protein domains are responsible for the translation between these two phenomena. Subsequently I show how these elements can be reliably mapped onto the theoretical framework of code biology.

Lastly I discuss the validity of an algorithm-based approach to identifying organic codes developed by Görlich and Dittrich. Unfortunately, the current state of this algorithm and the operationalised definition of an organic code is such that the process of identifying codes, without the necessary investigation by a scientist with a biochemical background, is currently not viable.

This study therefore demonstrates the utility of code biology as a theoretical framework that provides a synthesis between molecular biology and information theory. It cements the status of the histone code as a true organic code, and criticises the Görlich and Dittrich's method for finding codes by an algorithm based on reaction networks and contingency criteria.

Opsomming

Kodes is alomteenwoordig in kultuur—en by implikasie ook in die natuur. Kodebiologie is die studie van hierdie kodes. Tog het die term ‘kode’ ’n verskeidenheid van betekenis en interpretasies wat heelwat verwarring veroorsaak. Die eerste doel van hierdie studie is dus om te bepaal wat ’n *organiese kode* is en ’n stel kriteria te formuleer wat ’n vermeende kode aan moet voldoen om as ’n ware kode erken te word. Ek ontwikkel dan ’n inligtings-teoretiese perspektief op hoe organiese kodes ’n manier bied om biologiese inligting te hanteer as ’n logiese uitbreiding daarvan.

Met hierdie raamwerk as agtergrond gee ek ’n oorsig van ’n aantal van die huidige organiese kodes in ’n poging om aan te toon hoe die definisie van en kriteria vir ’n organiese kode gebruik kan word om die koring van die kaf te skei. Ek stel die ‘regulering kode’ voor in ’n poging om te wys hoe die kode-biologiese raamwerk op nuwe kodes toegepas kan word om hul geskiktheid as organiese kodes te toets en of dit die moeite werd is om hulle verder te ondersoek.

Ten spyte daarvan dat kodes algemeen in die biologiese wêreld voorkom, is relatief min van hulle onomwonde bevestig as organiese kodes. Die hoofdoel van hierdie studie is om vas te stel of die histoonkode ’n *ware* organiese kode is in die sin van die genetiese of seintransduksie kodes. Ek verskaf ’n volledige oorsig en ontleding van die belangrikste histoon post-translasionele

modifikasies, hul biologiese effekte, en watter proteïendomeine verantwoordelik vir die vertaling tussen hierdie twee verskynsels. Ek wys dan hoe hierdie elemente perfek inpas in die teoretiese raamwerk van kodebiologie.

Laastens bespreek ek die geldigheid van 'n algoritme-gebaseerde benadering tot die identifisering van organiese kodes wat deur Görlich en Dittrich ontwikkel is. Dit blyk dat hierdie algoritme en die geoperasionaliseerde definisie van 'n organiese kode sodanig is dat die proses van die identifisering van kodes sonder die nodige ondersoek deur 'n wetenskaplike met 'n biochemiese agtergrond tans nie haalbaar is nie.

Hierdie studie bevestig dus die nut van kodebiologie as 'n teoretiese raamwerk vir 'n sintese tussen molekulêre biologie en inligtingsteorie, bevestig die status van die histoonkode as 'n ware organiese kode, en kritiseer Görlich en Dittrich se poging om organiese kodes te identifiseer met 'n algoritme wat gebaseer is op reaksienetwerke en 'n kontingensie kriterium.

Chapter 1

Introduction

Code biology, the study of all codes of life, holds that the 4 billion-year history of life on earth saw the appearance of more than just the genetic code at the beginning and the various cultural codes at the end [14].

The concept of codes in biology is by no means new; the mRNA-tRNA-amino acid translation code, known as the genetic code, was the first code to be discovered and elucidated in the early 1960s [35, 114, 153]. After a hiatus of a decade the use of the term ‘code’ reared its head again in the 1970s in the context of a ‘metabolic code’ [161] and an ‘epigenetic code’ [44]. The code concept only really started to gain momentum in the early days of the new millennium, when Turner [166] proposed an ‘epigenetic code’, Strahl and Allis [156] the ‘histone code’, and Gabius [51] the ‘sugar code’. Recently, amongst others, there has been talk of a ‘cytoskeleton code’ [58] and a ‘ubiquitin code’ [83]. Despite these uses, the ‘codes’ they refer to lacked a general framework that defines what a biological code is and which components it should have in order to be classified as such; it was not at all clear whether these proposed codes were really true biological codes and whether they forced us to view life differently. The question remained

whether we should not just view such codes as the majority of biologists do the genetic code: as oddities, ‘frozen accidents’. Such a unifying framework was provided by Barbieri with his general concept of an *organic code* [8], one that arose from his earlier work on semantic biology [6] and which now forms the basis for the new research field of code biology [13] which has already recognised a number of other biological codes. Code biology recognises that biological codes are ubiquitous and absolutely essential for life, that coding in fact provides for a mechanism of evolution by natural conventions [7] that is distinct from the copying mechanism that underlies evolution by natural selection. The establishment of new organic codes introduce absolute novelties into the evolutionary process and are associated with major evolutionary transitions and increases in biocomplexity; natural selection, on the other hand, only provides for relative novelties [11].

In the following text I aim to (1) establish a set of criteria against which future codes may be tested to ascertain their veracity, (2) provide an overview of some of those biological codes currently thought to exist, as well as test them against the criteria set forth in 1, and (4). test, in depth, whether the ‘histone code’ conforms to the precepts of an organic code.

Chapter 2 will deal with the question, “What is code biology?”. Here I shall provide a detailed overview of code biology, focusing on concepts such as *organic signs*, *organic meanings*, and *adaptors*. I shall also provide a clear definition of what a code is and contrast this with the somewhat haphazard usage it has suffered to date. Then I will provide a list of criteria, or questions that should be answered when considering whether a putative organic code is indeed a *bona fide* organic code. Furthermore, I will attempt to provide a brief overview of the use and importance of the concept of ‘information’ in biology. The conclusion to this chapter shall deal with the

concept of ‘evolution by natural conventions’ as an extension to current thinking on evolutionary theory.

Chapter 3 will provide a brief, but thorough summary of several putative organic codes. Herein I shall also demonstrate how the previously mentioned criteria can be put to good use in identifying *bona fide* organic codes. The codes I shall be dealing with are as follows:

Genetic code: As the oldest and unanimously recognised biological code, the mapping of mRNA codons to amino acids, known as the genetic code, is a ‘safe’ test case to explore and test the criteria against.

Metabolic code: This code, proposed in 1975 by Tomkins [161], considers the association between certain ‘indicator’ molecules (putatively termed ‘symbols’) and unique metabolic states which they are a symptom of.

Signal transduction code: The associations between the various 1st and 2nd messengers are the subject of the signal transduction code, after the genetic code probably the most important code for life on earth.

Sugar code: The associations between various mono/oligosaccharides and the biological effects specified by them.

Splicing code: The system of signs that governs the *correct* splicing of an mRNA transcript at a given time and place.

Ubiquitin code: The mapping of ubiquitin ‘tags’ to unique biological effects in the context of post-translational protein modification.

Compartment code: This code details the process of recognition and translation whereby a protein is assigned the correct cellular compartment.

***Hox* code:** The idea that in the timing and distribution of *Hox* gene expression there lies a code. However, whether this is a code according to the definition and precepts of an organic code that I provide in Chapter 2 remains to be seen.

Regulatory code: A speculative code governing the associations between allosteric effector molecules and their effects on enzymes. To date, no work exists on the regulatory code; I explore a possibility of such a code as well as the form it could take.

Chapter 4 is the body of work representing the histone code. I begin with an introduction to the basic biochemistry of histones and then proceed to the possible functions of the histone code as it pertains to the role it plays in eukaryotic life. I then provide a detailed overview of the *major* histone post-translational modifications and the unique biological effects which they specify. Following this I spend some time identifying the adaptor molecules in the histone code, as well as the effector proteins they form part of. I then test the precepts of the histone code against the criteria I have previously defined.

Chapter 5 considers the efforts of Görlich and Dittrich [59] at designing an algorithm capable of identifying what they call ‘molecular codes’. I provide a brief overview of the methods they use as well as an analysis of the veracity of their results and the feasibility of trying to identify codes algorithmically.

Chapter 6 offers a summary and discussion of the foregoing work, with a final section on possible avenues of investigation that future work shall bring.

Chapter 2

Code biology

Our social life is inextricably linked with codes. From the codes governing the various languages, religious doctrines, judicial systems, to the rules of games and, in modernity, those of programming languages, codes are ubiquitous in culture. Further, codes are necessary in culture: without these codes and many more, society as we know it simply would not exist. For this reason codes were long thought to affirm the nature/culture divide that has characterised scientific inquiry of the 20th century. The discovery of the genetic code in the 1960s threatened to upend this long-standing convention. For the first time, codes had become a part of the natural world. However, the science of the time needed to be reducible and the concept of a ‘code’ was therefore reduced to a metaphor - a ‘protective belt’ had enveloped it and robbed it of much of its potential [12].

It was soon pointed out that the presence of the genetic code implied that the cell is a physical system controlled by symbols [121]. Simultaneously, Thomas Sebeok argued that if man has roots in nature, so too must culture have roots in nature [12]. Thus began the inquiry in earnest into biosemiotics.

Barbieri [11] provides a preliminary definition of a semiotic system as a system consisting of two independent worlds, signs and meanings, that are connected by the conventional rules of a code. The introduction of ‘signs’ and ‘meanings’ to the molecular world invited the unwelcome guest of ‘interpretation’, for in order to divine meaning from a sign, one would need interpretation, and if interpretation is implied, does this not imply an interpreter - a mind? Indeed it would, if we were dealing with the cultural codes, where subjectivity is a factor. However, on the molecular scale there is no need for interpretation. All that is required is for some ‘thing’ to link these two ‘worlds’ of sign and meaning. This thing (henceforth adaptor) would be required to do little more than to instantiate the correct *sign* \rightarrow *meaning* mapping.

Such a mapping often takes the form of Fig. 2.1. This details a typical mapping as it occurs in the genetic code. As one can see, it is possible for more than one sign to map to the correct meaning (given by the functions f and g), however it is rare that a single adaptor molecule is able to link more than a single binary code pair. Such a ‘many-to-one’ mapping is called a degenerate code. It is possible that such degeneracy became part of biological codes in an effort to increase the robustness of the code. Furthermore, biological codes, as opposed to some cultural ones, do not allow for bidirectional mapping; a biological code is strictly a one-way mapping from sign to meaning. This does not, however, preclude the meanings from acting as signs in another code.

An organic code is therefore a molecular system for translating an *organic sign* into its *biological meaning*. In the genetic code, which has been shown to be a true organic code [8], the organic signs are triplet sequences of three nucleotides in mRNA which has been transcribed from DNA and

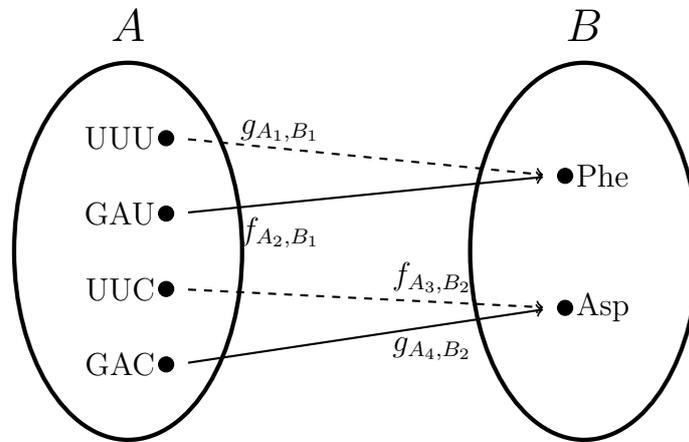


Figure 2.1: Mappings f and g between a set of nucleotide triplets and a set of amino acids

subsequently processed into a mature form. The 64 possible triplet sequences are called *codons*. These codons are recognised by complementary nucleotide triplets, called *anticodons*, on tRNA molecules that have been charged with amino acids. Each codon/anticodon pair corresponds to a particular amino acid according to a convention called the genetic code; the amino acid is therefore the biological meaning of the codon sign. Since more than one codon/anticodon pair can be associated with a particular amino acid the genetic code is a degenerate code. A sequence of mRNA codons is translated into a corresponding sequence of amino acids in a polypeptide in a process called translation, which is catalysed by a ribosome. On a higher level a particular mRNA nucleotide sequence can be regarded as the organic sign that is decoded into its biological meaning, here a specific polypeptide. It should be remembered that the *worlds* of nucleotide sequences on the one hand and amino acid sequences on the other are completely *independent* of each other. The *set of rules* of the genetic code that associate codons with amino acids are *conventional* in nature since the specificity of this correspondence is not dictated by the laws of chemistry but have been fixed in

the course of an evolutionary process. There are no deterministic reasons for the rules of the genetic code; in this sense they are arbitrary, but once fixed they remain frozen.

Prior to the discovery of the genetic code the concept of a code in molecular biology was already put forth by Schrödinger [139]. In this scenario the chromosomes were thought to contain a ‘code-script’ that orchestrates the endeavour of genetic translation; they were simultaneously a container for the description of the organism, including themselves, as well as the implementers of this code [139]. In tandem with the discovery of the genetic code came John von Neumann’s theory of *self-replicating automata* [170]. Herein he suggested that any self-replicating automaton would first need to possess a description of itself, which would function as a template for self-replication. Such an internally asserted description of structure and function is according to Barbieri [11] what makes life an act of “artifact-making” and provides biological systems with closure, instead of invoking the need for an externally imposed description. Secondly, such a description would need to be symbolic in nature [170]. The importance of symbols and signs as information carriers was further stressed by Pattee [120]. Signs that act as information carriers in turn act as *constraints* upon dynamic processes; they restrict the number of *allowed* physical interactions from the pool of *possible* physical interactions. Moreover, information (and by implication signs) only has meaning in events where the outcome *could* be otherwise, they provide a necessary distinction between events with multiple outcomes [120]. In other words, they make the *arbitrariness* of codes possible.

2.1 On codes

The term ‘code’ has seen much use in biological studies since the 1960s, however, rarely with a formal definition in tow. The most common use of the term appears to be in conjunction with state-dependent ‘snapshots’ of metabolic states. The *Hox code* [69] for example is used to describe a ‘readout table’ detailing which combination of *Hox* genes are active in which tissues at which time. The *metabolic code* on the other hand claims that certain key metabolites are symbols for particular metabolic states, much like a red light at a traffic light would designate ‘stop’, however, no mention is made of the driver or adaptor that is able to link the symbol with the state.

For the purpose of this thesis, I shall employ a definition, slightly adapted from Barbieri *et al.* [16] and Brier and Joslyn [27]:

An *organic* code is a mapping that describes the associations between two discrete organic ‘worlds’: one, a set of biomolecules that act as *organic signs* and, two, a set of biomolecules or biological effects that act as *organic meanings*. The link between these two worlds is created by an adaptor molecule that is able to recognise an organic sign on the one end, and mediate the organic meaning on the other. These associations are *arbitrary* in the sense that they exist independent of physical or chemical necessity and are therefore purely due to *natural convention*.

Therefore, in order to correctly identify a putative code as a *bona fide* organic code, one needs to:

1. Demonstrate that the code links two independent worlds, namely that

of organic signs to their biological meanings. The organic signs will be biomolecules, but their biological meanings need not necessarily be; instead of molecules they can, for example, be biological effects such as activation or repression of gene transcription, which is relevant in, for example, the case of histone modifications. Independence implies that *in the absence of the code there is no deterministic relationship* between an organic sign and its biological meaning. The relationship between organic sign and its meaning is therefore a *natural convention*.

2. Identify the set of *adaptor molecules* that instantiate the rules of the putative organic code. On the one hand, such an adaptor must specifically recognise the organic sign molecule and, on the other hand, translate this sign into its biological meaning, either directly or indirectly. The charged tRNA in the genetic code is an example of indirect translation: uncharged tRNA on its own can only recognise a codon; it needs another agent, a specific aminoacyl-tRNA synthetase, to create the translation to an amino acid. The signal-transduction code [8] is an example of direct translation, where the adaptor, here a protein complex spanning the cell membrane, both recognises the external organic sign (first messenger) and mediates the production of the internal second messenger, the biological meaning of the first messenger.
3. Show that the set of rules that implement the code is conventional in nature in that it can be experimentally altered and still act as a code, albeit now with different rules. Alternatively, it may be that nature has provided alternative implementations of the code in question, such as, for example, the 20 known versions of the genetic code [78, 117].

However, unlike the first two identification criteria, this *contingency* criterion is neither necessary nor sufficient, but provides verification of the conventional nature of the organic code in question. This point will be taken up in Chapter 5 in the discussion of a proposed algorithm for discovering molecular codes.

The signature component of any organic code is the *adaptor* molecule that links the world of organic signs to the world of its biological meanings. In the genetic code this role is played by the charged tRNAs. One could say the genetic code is realised in these adaptors. However, the ‘writers’ of the genetic code are the aminoacyl-tRNA synthetases that charge tRNAs with their correct amino acids. All of these components of the genetic code are produced by the cell itself; the cell is therefore what Barbieri [9] calls the *codemaker*.

An adaptor molecule should therefore exhibit the following properties:

- An adaptor molecule must be an independent third-party to the organic sign/meaning-system. Much like an enzyme is able to catalyse a reaction without itself being altered significantly by the reaction, an adaptor molecule needs to remain independent of any chemical processes that occur during translation—it should therefore not change the meaning of the sign during the process of translation. Imagine the chaos were a tRNA molecule to decide, willy-nilly, to which amino acid it would translate a codon.
- The adaptor molecule has a dual function: on the hand it must recognise the organic sign and on the other it must produce or mediate the biological meaning, either a biomolecule or a biological effect. In those codes that we have so far verified, the organic sign is a partic-

2.2. Evolution by natural conventions

ular biomolecule or part of a biomolecule. For example, the tRNA molecule has a specific RNA sequence, the anticodon, which specifically recognises and binds to the corresponding codon on a mature mRNA transcript. The recognition site for the biological meaning however, does not always bind a biomolecule. Since a significant portion of the organic codes tend to follow a molecule \rightarrow effect trajectory, the recognition site for the sign is often attached to an effector protein of sorts. This is especially prominent in the sugar code (Chapter 3) and the histone code (Chapter 4).

Code biology views the cell as a ‘*codepoietic*’ system; one which is able to create and conserve its own codes [14]. Often these codes are not expressly defined in the DNA of a cell, however the fact remains that cells are able to implement the rules of these codes nonetheless. The genetic code, as expansive as it is, does not code for every chemical or physical interaction between the various components of a cell. It is not a director of events as originally thought. For example, while the genetic code would specify the identity of a particular amino acid in a particular position of a particular polypeptide sequence, whether or not this amino acid will be subject to post-translational modification or not, is not under the purview of the genetic code.

2.2 Evolution by natural conventions

A defining element of code biology is evolution by natural conventions [7], which is not meant to replace or invalidate evolution by natural selection, but rather provide an extension thereof.

2.2. Evolution by natural conventions

However, before I can fully delve into the details of evolution by natural conventions, I need to highlight the differences between the two molecular mechanisms that underlie natural selection and natural conventions—namely copying and coding.

Copying concerns the replication of information with high fidelity. In the biological context, copying operates on individual molecules (eg., DNA) and errors or variation in these molecules are able to change the information contained therein, but not the meaning. We can therefore say that copying, the process that underlies evolution by natural selection, introduces *relative* novelties by modifying existing entities.

Coding on the other hand involves a collective set of rules for *translating* information. Changes to these rules, or the introduction of new rules, alter the meaning of the information they pertain to. These changes—and the resulting effect on the meaning of information—are what underlie the evolution by natural conventions and therefore we can say that this process produces *absolute* novelties [12].

Natural selection is a mechanism based on copying (DNA replication and DNA transcription to RNA). However, copying is not a process with 100% fidelity; in DNA replication, for example, for every one million bases copied at least one will be copied incorrectly. What this means is that a unique, but relative change in the current message (DNA) is introduced, which results in a variation in form or function of an existing structure (RNA or protein) [11]. If this variation is beneficial to an organism in a given environment, the chances for that organism surviving increases; ultimately that variation is propagated until the point where it becomes detrimental to an organism in a given (albeit different) environment.

Absolute novelties must have been part of the evolutionary process *at*

least once, however, it is more likely that during the course of evolutionary history, absolute novelties, i.e., new biological codes, arose several times. By linking molecular worlds that were not related before, each new code opens up a set of new possibilities for the organism to explore. This could offer an explanation for the major evolutionary transitions and sudden increases in biocomplexity not yet fully explained by the modern synthesis. The number of codes an organism is able to use could be seen as a measure of biocomplexity—more complex organisms are able to employ more codes.

Nucleotides and amino acids for example, necessarily pre-date the genetic code, but the *mapping* of nucleotide sequences to amino acid sequences is the start of a 4 billion-year story which still has not reached its conclusion. The absolute novelty here is the mapping and it has, undeniably, resulted in a sudden increase in biocomplexity [7, 12, 13]. The appearance and ‘settling in’ of such mappings, or codes, is what we call the evolution by natural convention.

2.3 Information

The concomitant discoveries of the genetic code and protein translation suggested that the DNA molecule carried information and that this information could be translated to give rise to new structures. This revelation quickly became the ‘central dogma’ of modern biology [145], as counter-intuitive as that seems (dogmas usually being anathema to science). Regardless, this discovery did necessitate a conceptual framework for the management of information in biology.

Barbieri asserts that information is a *new observable* that can not be measured, in the physical sense, other than by naming it—the sequence

or structure of the information you are dealing with [13]. Barbieri asserts that information is the result of “a template-dependent copying process” [10], which is undoubtedly true. But I believe biological information can also be produced in other ways: protein post-translational modifications—processes that undoubtedly alter the information present in a protein—are not the result of template-dependent copying, but they are *iterable*, that is to say that they can be repeated *ad infinitum* given suitable materials and conditions. Similarly when one considers the sugar code, the saccharides are not produced according to a template, however they are able to inform the lectins of specific functions that are in turn performed. Template-dependent copying should therefore, in my opinion, be regarded as a special case of information production rather than being the rule when considering biological information.

To further talk about biological information we need to approach the topic from two angles. Firstly, Shannon [144], considered the *meaning* of information “irrelevant to the engineering problem”. Rather, as an engineer, his main concern was the reliable transfer of information from source to receiver. Since a great deal of biological systems are concerned with communication, one consideration of information is the sound arrival of the exact message (or a close approximation thereof) that has been fabricated at one end, at another distant point [19]. A relevant biological example would be the vertical transfer of genetic information (hereditary) from one generation to the next. Herein it is important that the ‘message’ (in this case genetic information of the progenitor) arrives at the receiver (the next generation) in a manner resembling the original message as exactly as possible. However, virtually all channels of communication are unreliable and, inevitably, the message shall suffer decay [144]. In order to combat this,

messages are encoded with redundant bits [20]. This is a form of encoding where the message proper is peppered with nonsense bits, short sequences that have no value. Therefore, if decay occurs, it is less likely to affect a bit of the original message, preserving the original content. Again, an analogue presents itself in the biological world in the form of introns and non-coding DNA. I therefore propose that these sequences are conserved within the genome precisely to increase the robustness thereof, making it less susceptible to deleterious mutations.

Ultimately, the sound transfer of information is a concept that deals with the *copying* of information since this does not deal with the actual *meaning* of information. In other words, DNA replication and transcription, the processes of copying a strand of DNA into DNA and RNA respectively, deal with just such an issue.

The second consideration of biological information concerns the *meaning* thereof. Once a message has been properly encoded and sent, the next logical step would be, upon reception, for this message to be decoded by removing the redundant bits and translating it. The processes of mRNA editing (splicing) and protein translation come to mind as analogues of these processes.

The following would therefore be logical necessities for the decoding of information:

- A description of the original message in terms of a specific set of signs that are independent of the translated message insofar that the latter does not affect the content of the former.
- A schematic, or code, detailing the translation of the sent information into a form that is usable by whichever system received the message.

- An adaptor, able to link the signs to their designated meanings without having any impact upon the information carried by either sign or meaning. In other words, a ‘blind’ adaptor.

Organic codes are therefore superbly suited to the task of translating information into meaning. Firstly, they are mappings from one set of (organic) signs to another, independent set of (biological) meanings, secondly, codes are used to decode structural or sequence information to other, meaningful information, and lastly these codes do not depend on the individual features of the information [4]. Information however, only becomes meaning when it is translated according to the rules of the appropriate code. For example, the genetic code is nonsense when translated into the English language, but when it is translated into a polypeptide sequence it makes biological sense in the context of the cell. Codes therefore, are *necessary* for the meaningful translation of biological information and for the correct function of the various biological system under their purview.

Chapter 3

Some organic codes

In this chapter I will review some of those biological systems thought to be codes. Since the advent of biological codes with the genetic code, many biological systems have (sometimes falsely) been called codes. In the following discussion I shall adhere to the definition of a code set forth in Chapter 2, because often a ‘code’ is not a code as defined there. I will therefore distinguish between those codes I believe are self-evident, those that warrant further investigation, and those that do not conform to the precepts of an organic code.

Table 3.1 provides a cursory overview of the organic codes as they are presently known.

Of the known organic codes there are several that conform to an organic code *prima facie*; these include the genetic, signal transduction, splicing, sugar, and regulatory codes. These codes all nominally possess the required two worlds, specialised adaptor molecules, and the arbitrariness which defines an organic code. Although these codes appear on solid ground, more detail is required on the exact functioning of these codes in order to properly cement their status as *bona fide* organic codes. Another possible code that

Table 3.1: The various proposed organic codes, the independent worlds that they link, and their adaptors.

Code	World 1	Adaptor	World 2	References
Genetic code	mRNA codons	charged tRNAs	amino acids	[35, 114, 153]
Splicing code	intron/exon boundaries	spliceosome proteins	properly joined exons	[49]
Sequence codes	DNA sequences	protein receptor/effector complex	transcriptional behaviour	[8, 163]
Signal transduction code	1 st messengers	transmembrane receptors	2 nd messengers	
Sugar code	saccharides	lectins	biological effects	[51]
Compartmental code	protein signals	endoplasmic reticulum/Golgi apparatus	cellular location	[8]
Ubiquitin code	ubiquitin 'tags'	ubiquitin binding domains	biological effects	[83]
Histone code	post-translational histone modifications	protein domains	biological effects	[156]
Regulatory code	allosteric effectors	allosteric binding sites	enzymatic activation or inhibition	
Metabolic code	molecular symbols	the scientist	metabolic states	[161]
<i>Hox</i> code	<i>Hox</i> transcription states	the scientist	developmental stages	[69]
Tubulin code	microtubule modifications	maps, +TIPs, motor proteins	cellular trafficking, mitosis, assembly of cellular structures, i.e., cilia	[169]
Cytoskeletal code	microtubules	anchoring molecules	cellular structures	[8]
Apoptosis code	protein modifications		cell death	[8, 18, 50]
Nuclear signalling code	phosphoinositides	nuclear receptors	transcriptional regulation	[97]
Adhesion code	cadherins	receptor site on homotypic cadherins	specific cell-cell adhesion	[130]
Quorum sensing code	autoinducers	bacterial receptor proteins	gene transcription	

would be easy to cast in the code biological framework would be that of quorum sensing in bacteria. Quorum sensing involves two (or more) different species of bacteria that are able to send, receive, and properly respond to chemical messages; these responses range from alterations in the virulence of a species to the suppression or incitement of growth.

The largest category is that of the *possible* organic codes; this is the set of proposed codes that have not been properly verified yet or where doubts as to their plausibility as an organic code exist. Several examples of such possible codes are currently available: the ubiquitin code, compartmental code, cytoskeletal code, and adhesion code to name but a few. Although these systems to conform nominally to the precepts of an organic code, the question remains whether they necessitate their own code or whether they could be assimilated in the larger project of constructing a protein post-translational modification code. It may perhaps be simpler to construct these codes as individual entities first and then integrate them into a larger whole as this would simplify our understanding of these codes immensely; one would be able to deal with a particular system without necessitating the comprehensive knowledge of the entire protein post-translational modification system.

As I've mentioned in Chapter 2, there are instances where the term 'code' has been used to describe something akin to a fingerprint rather than an organic code. The metabolic and *Hox* codes are, as I will discuss in sections 3.2 and 3.9, precisely such instances.

A recent paper by Stergachis *et al.* [155] has generated much furore in the media as a 'second' genetic code. Upon closer inspection, however, it appears that much of this hype is misplaced as the idea that certain sequences in the genome are able to affect the binding of transcription

factors is not new [163], and indeed has seen use in other codes as well (eg., the splicing code, see section 3.5). Thus it would be more appropriate to term this the *transcription factor code*. The paper itself, however, is less concerned with the codification of these elements than with the impact they may have had on the evolution of proteins. This once again highlights the confusion that may arise when the term ‘code’ is used *ad hoc* to describe a particular biological system.

3.1 The genetic code

The first universally recognised organic code was the genetic code. Discovered and codified in the 1960s by Crick *et al.* [35], Nirenberg *et al.* [114] and Söll *et al.* [153], the mRNA/amino acid translation scheme revolutionised molecular biology. However, the concept of coding at the molecular level was quickly dismissed as it went against the deterministic bent of molecular biology at the time. The concept of an organic code therefore was dismissed as a mere metaphor [15]. Nevertheless, it would be useful to test this primal organic code within the framework provided by code biology since it appears *prima facie* to fulfil the criteria for an organic code.

The genetic code describes the association of one of the 64 triplet codons formed by the four N-bases of mRNA with either one of 20 amino acids or with one of three ‘stop’ signals as detailed in Table 3.2. The universality of the code is near absolute, however in certain organisms the associations between codon and amino acid are different, owing to the degeneracy of the genetic code. A degenerate code does not describe a one-to-one mapping (as one would find with the Morse code), rather it appears that a level of redundancy has evolved that allows several similar signs to code for one

Table 3.2: The mRNA/amino acid translation scheme

Nucleotides		U	C	A	G
U	U	Phe	Phe	Leu	Leu
	C	Ser	Ser	Ser	Ser
	A	Tyr	Tyr	Stop	Stop
	G	Cys	Cys	Stop	Trp
C	U	Leu	Leu	Leu	Leu
	C	Pro	Pro	Pro	Pro
	A	His	His	Gln	Gln
	G	Arg	Arg	Arg	Arg
A	U	Ile	Ile	Ile	Met
	C	Thr	Thr	Thr	Thr
	A	Asn	Asn	Lys	Lys
	G	Ser	Ser	Arg	Arg
G	U	Val	Val	Val	Val
	C	Ala	Ala	Ala	Ala
	A	Asp	Asp	Glu	Glu
	G	Gly	Gly	Gly	Gly

meaning. In the genetic code this is exemplified by the six codons that code for the single amino acid, leucine.

The translation from codon to amino acid is enabled by a correctly charged tRNA molecule. In one of its unpaired loops (the so-called anticodon loop) this RNA possesses a specific triplet sequence, the anticodon, capable of pairing with a specific codon on a mature mRNA molecule. At the 3'-end is a sequence to which the amino acid corresponding to the anticodon is ligated by the amino-acyl tRNA synthetase specific for that tRNA/amino acid combination.

The malleability of the genetic code has been amply demonstrated by the artificial creation of quadruplet and quintuplet codons, of ribosomes and tRNA molecules that recognise and decode quadruplet codons, the incorporation of unnatural amino acids, and the creation of a 65th codon [3, 23, 66, 112, 173]. This malleability is however not restricted to the labo-

ratory. Nature herself has demonstrated, with at least 20 known variations¹, that the genetic code is an arbitrary association of mRNA codons and amino acids. Mitochondrial genetic codes detail different mRNA → amino acid mappings when compared to nuclear genetic codes, and the bacterial genus *Mycoplasma* is known to employ a genetic code that differs to that used by, for example, humans [77].

In conclusion, the genetic code establishes a conventional relationship between two independent worlds, that of mRNA codons and that of amino acids. These worlds would not be linked to one another were it not for the properly charged tRNA molecules that act as adaptor molecules. Therefore, the genetic code can be considered a *bona fide* organic code.

3.2 The metabolic code

The *metabolic code* was proposed by Tomkins [161], who explored the possibility that particular organic molecules (specifically cyclic AMP, guanosine-pentaphosphate, and hormones) could act as ‘symbols’ denoting unique metabolic states. For example, in *Escherichia coli*, the presence of cAMP was thought to symbolise carbon starvation since the production of this particular metabolite is increased dramatically during periods of carbon starvation. Similarly, in mammals, cAMP production is up-regulated as a result of increased glucagon and epinephrine production during periods of starvation. The metabolic code thus constitutes a ‘fingerprint’ of cellular activity, which gives us an idea of what occurs within cellular metabolism at a given time. Further, Tomkins [161] theorised that each symbol has under its purview a set of biological processes and molecules, called its ‘domain’. He

¹<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>

3.2. The metabolic code

thought that, although each symbol has its own, unique domain, processes and molecules can be shared amongst domains and therefore amongst symbols. With the advent of metazoan life there appeared progressively larger and more complicated forms. Hormones were thought to have evolved as more stable symbols, since cAMP and ppGpp were in a continuous state of flux. Hormones therefore, were ‘encoded’ with a specific message, secreted into the organism and, at their specific receptors, ‘decoded’ into very specific biological effects. In reality these effects took the forms of cAMP or ppGpp (or, in a more modern context, secondary messenger molecules) and thus each hormone carried with it a symbol-message which, once decoded, indicated the particular metabolic state of a cell.

If the metabolic code were a true organic code it would be unique in the sense that it is a mapping from larger phenomena, such as metabolic states or biological effects, to biomolecules. This is because the symbol molecules appear as a *result of* the foregoing biological phenomena. These symbol molecules are thus indicators of a particular metabolic state. However, for the metabolic code to be considered an organic code according to the definition laid out in the foregoing chapter, it would require an adaptor molecule. Upon inspection, it seems doubtful that an adaptor molecule is a useful concept for the metabolic code. The only beneficiary of the metabolic code would be the scientist, who upon measuring the levels of a particular metabolite would then be able to deduce certain aspects of the cellular metabolism—the cell is already “aware” of its metabolic state since it is producing the metabolite in question. It does appear however that certain aspects of the metabolic code could be subsumed by the signal-transduction code (which is dealt with in the next section), particularly those dealing with the communication of states such as glucose starvation

(cAMP), amino acid shortages (ppGpp), or satiety or fear (hormones).

While it does appear to link two worlds, (metabolic state and biomolecule) the question must be asked: what is the adaptor? What is the codemaker? Concerning the metabolic code, this would be the scientist observing the cellular system that is undergoing a particular form of stress-response. A mind is necessary to *interpret* these molecular symbols. These two points: a non-molecular agent and interpretation, disqualify the metabolic code from being an organic code. In conclusion, the metabolic code appears to fall into the category of a molecular fingerprint, not an organic code.

3.3 The signal transduction code

A logical step for a subject dealing with the nature of signs, meanings, and codes would be to take a look at signal transduction. Signal transduction provided the cell with the means to react to the external environment [9], it was thus a ‘sensing’ mechanism, which allowed to the cell to respond to various environmental stimuli; chemotaxis comes to mind as an example hereof. In this process a micro-organism detects the concentration of metabolites in its environment and move towards (nutrients) or away (toxins) from them.

Signal transduction involves the sensing of an extracellular stimulus by a set of highly specialised receptor proteins that in turn translate this extracellular event into the production of an intracellular messenger molecule. Most often the extracellular signal takes the form of a specific metabolite, such as an ionic species (Ca^{2+} , for example) or a small biomolecule or a hormone; the exceptions are some neural cells where the extracellular signal is an electrical impulse. Whether neural signals warrant their own code or whether they can be incorporated into the signal transduction code proper

3.3. The signal transduction code

is still a matter of some uncertainty. The difference between synaptic signal transmission and signal transduction must be stressed; the transmission of neural signals is a process of sequential changes in polarity as an electrical signal is channelled along neurons. The transduction of this signal occurs when it reaches a synapse and results in the release of specific neural transmitters (acetylcholine for example) that cross the synaptic gap and in turn are able to effect a de- or hyper-polarisation. Signal transduction therefore involves the relay of a message by an intermediary in a form different to that of the received message.

Each type of extracellular signal is recognised and bound by a specific transmembrane receptor protein. These receptors are in turn bound to specific proteins or protein complexes that are able to synthesise a specific second messenger (where the initial extracellular signal is the first messenger). In eukaryotic cells, these second messengers are any one of the following four: diacylglycerol (DAG), inositol triphosphate (IP₃), ionic calcium (Ca²⁺), and cyclic adenosine monophosphate (cAMP).

The association between first and second messenger is entirely arbitrary since there is no chemical necessity for a particular first messenger to specify a particular second messenger. This association has become ‘locked-in’ over the millennia.

Signal transduction makes a very clear case for an organic code. Two worlds, first messengers that act as organic signs and second messengers that act as biological meaning, are linked to one another by an adaptor molecule—the transmembrane receptor protein.

3.4 The sugar code

With the introduction of ‘information’ as a biological concept, certain groups began to explore the possibility that information transfer outside of the genetic code was possible. The sugar code presented such a possibility [51].

Post-translational protein modification undeniably expands the range of functions of any protein [54] (this will be explored in depth in Chapter 4 in the context of the histone code). Protein glycosylation, the addition of a carbohydrate molecule to a protein, and the recognition of these glycosylated proteins by specific protein molecules, called lectins, forms the basis of the *sugar code* [51–54].

Protein glycosylation is estimated to occur in >70% of proteins across all organisms [51, 111] and easily outstrips the genetic code in terms of sheer complexity, with over 1000 unique *N*-glycan structures already catalogued by the CarbBank database [52]. The position of these glycan structures as well as their length and modification status (e.g., O-acetylation, sulfation) are able to confer new ‘meaning’ upon the glycans since the altered structure necessitates a different lectin to bind to it, which in turn results in a function different to that specified by the prior modification status [51]. These qualities of glycans are all highly malleable and occur in a state of high-turnover, hinting that the sugar code may be responsible for transient metabolic regulation.

Glycoproteins assume a wide variety of functions such as cell-adhesion, receptor-targeting, and growth control, each of which appears to be controlled by a specific sugar/lectin pair, where the lectin appears to act as both receptor and effector [51, 111].

The adaptors for the sugar code are therefore thought to be the lectins, a class of proteins that possess no catalytic activity on carbohydrates [52].

Further, lectins possess a high degree of selectivity for the various carbohydrates [54], making them ideal candidates for possible adaptor molecules. Moreover, it appears that the sugar code can be altered experimentally with the introduction of biomimetic glycoclusters, strengthening the suspicion that lectins are able to act as molecular adaptors [111].

In conclusion it appears that the sugar code can be viewed as a potential organic code; it contains the necessary two worlds as well as a possible adaptor that is able to link these two worlds to one another. The sugar code is an example where a world of biomolecules is linked to a world of biological effects, rather than a different set of biomolecules.

3.5 The splicing code

A typical gene consists of various coding and non-coding elements, exons and introns respectively. While exons are relatively short, 100 to 300 bp, an intron can assume a length of up to 100 kpb. Were a cell to translate all the introns and exons present in a gene it would be presented with a cumbersome, and wasteful, task indeed. *Splicing* is the process whereby introns and exons are separated from one another and the exons are in turn the joined together in the order that they occur in DNA to form an mRNA transcript. When the order in which exons are joined is shuffled the process is called *alternative splicing*, which allows for the creation of a much larger, diverse set of proteins than specified by genes alone. For example, the *Drosophila* cell-surface protein, *Dscam* has, due to alternative splicing, more than 38,000 isoforms [136]. In humans, 95% of multi-exon genes are consistently spliced in a variety of ways depending on cell and tissue type and mutations in the splicing mechanism accounts for some

15–50% of genetic diseases [5].

Each intron contains 5' and 3' splicing sites, as well as a branch point sequence, that are recognised several times during spliceosome assembly by a variety of proteins: the U1 and U6 snRNPs (small nuclear ribonucleic particles) and SF1/mBBP and U2 snRNP respectively [174]. These sequence features are present in each and every intron. However, the cell is then presented with another problem in the form of pseudo-exons, DNA sequences that lie in between introns and possess similarity to exons, but translate into nonsense. Indeed, the abundance of pseudo splice sites, which give rise to pseudo exons, has the capacity to outnumber the real exons [42].

The splicing machinery is able to differentiate the real exons from the pseudo exons; however, since real exons contain key sequence features that define them, known as exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) and their intronic counterparts (ISEs and ISSs) [49, 101, 123]. The splicing enhancers tend to recruit members of the SR protein family whereas the splicing silencers recruit from the diverse hnRNP class of proteins [174].

The splicing code, therefore, would have to be an association between 'real' exons and a mature mRNA transcript. The adaptors of the splicing code would therefore lie within those proteins that recognise, firstly, the 5' and 3' sites and, secondly, the ESSs, ESEs, ISSs, and ISEs. However, the evidence, while not conclusive, suggests that the splicing code deserves more attention at the very least. A further dimension of the splicing code is that the mRNA transcripts vary in terms of their exon composition depending on the cell or tissue type they originate from [5]. This hints that there may be another code, one of tissue-dependent splicing that may be worth a look.

3.6 The ubiquitin code

Ubiquitin is a small protein of ca. 76 amino acid residues found in almost all types of eukaryotic tissues, hence the name. One of the major post-translational modifications involves the addition of a ubiquitin molecule (ubiquitylation) to a protein, most commonly at a lysine residue. However, ubiquitylation is not limited to the addition of a single ubiquitin molecule or the formation of linear chains. Multimono-ubiquitylation and branched or unbranched ubiquitin chains are all possible. Ubiquitylation involves the ubiquitin-activating enzymes (E1s), ubiquitin-conjugating enzymes (E2s), and the ubiquitin ligase enzymes (E3s), which ultimately catalyse the addition of a ubiquitin molecule to the target protein [83].

Ubiquitin has been implicated in a variety of functions, mainly protein degradation [64, 83] and its role has expanded considerably since its discovery. Ubiquitylated proteins are intricately linked to processes such as transcriptional regulation [74], cell-cycle control [177], and membrane transport [65]. The appointment of each of these functions depends on the length of the ubiquitin chain and the degree of branching that the ubiquitylation forms [65, 83].

The execution of these functions is achieved by a variety of proteins, but only a limited number of ubiquitin-binding motifs exist. These are specialised protein structural domains that recognise and bind, with high specificity, particular ubiquitylated proteins. Currently ca. 20 families of ubiquitin-binding domains have been recognised [72], but that number is sure to expand. These binding domains are usually bound to a particular *effector* protein that is able to execute the function specified by the unique ubiquitin tag; these two domains, binding and effector (or catalytic) are separate from one another in terms of their position on a protein. This is

exemplified by the histone deacetylase, HDAC6, where the ubiquitin binding domain, the *zinc finger*, is responsible for the recognition of a ubiquitin ‘tag’ on a protein (in this case a histone) and found at the C-terminus of the protein, but is otherwise separate from the catalytic domain (the effector protein/deacetylase), which is found toward the N-terminus [67, 71].

The ubiquitin system does appear to fit the criteria for an organic code; two independent worlds (biomolecule and biological effect) that are linked by an adaptor molecule, in this case any one of the various ubiquitin-binding domains. Further evidence would be needed, in particular whether the binding domains are interchangeable and thus whether one is able to ‘re-write’ the ubiquitin code.

3.7 The compartment code

Eukaryotic cells, with all the various membranes and compartments they possess, need a process that enables them to correctly assign each protein to its compartment, be it the cell membrane, the nuclear membrane, the mitochondria, etc. The cell is able to accomplish this in two stages. First, after a protein has been synthesised, it may contain a leader or signal peptide. These short amino acid sequences determine whether the protein is destined for the endoplasmic reticulum or, if they are absent, the cytosol [8]. Once the protein has reached the cytosol, its journey is at an end. If, however, the protein has been sent to the endoplasmic reticulum, it then enters the second stage. The endoplasmic reticulum packages the protein into a vesicle that is to be sent to the Golgi apparatus. Once there, the protein is, depending on the leader peptide, packaged into vesicles destined either for intra- or extracellular transport, or, if a specific destination signal

is absent, the default destination is the plasma membrane [8].

The system of cellular compartmentalisation is thus subject to codified behaviour. The presence, nature of, as well as the absence of these peptide signals are analogous to organic signs in that they specify, without a deterministic link, the cellular location of a protein. This location is in turn analogous to the biological meaning of an organic code. Lastly, no organic code would be one without an adaptor molecule. In this case I believe that it may exist in two stages, firstly a recognition site on the endoplasmic reticulum that is able to ferry the nascent protein on its way (should it contain a leader peptide). Secondly, a recognition site on the Golgi apparatus that is able to bind the leader peptide and then shuffle the protein toward the intra or extracellular environments it is destined for.

3.8 The regulatory code

An allosteric molecule is a small bio-molecule that is able to regulate the activity of a protein by enhancing or diminishing the affinity of the protein for its substrate or the activity (k_{cat}) of the enzyme [61]. It achieves this by binding to a specialised ‘allosteric’ site on a protein. Allosteric modulation is different to ‘classical’ reversible enzyme regulation since the allosteric molecule does not, unlike traditional agonists or antagonists, bind to the active site of a protein [140]. In fact, the allosteric site and active site of such a protein are suitably spatially separated from one another for us to assume them to be independent [179]. Further, there is no apparent need for an allosteric modulator to be chemically similar to the endogenous ligand of a protein in order to affect the function of said protein [88]. Allosteric regulation is also subject to cooperativity: subsequent binding of the same

3.8. The regulatory code

modulator to other subunits of the multimeric enzyme serves to reinforce the effect brought on by the initial binding of a specific modulator [40].

Allosteric regulation is present a variety of proteins, such as the GPCRs (G-protein coupled receptors) or the 7TM (7-helix transmembrane protein) or hemoglobin [22, 103, 140]. The most common consequence following the binding of an allosteric modulator is a conformational change in the protein, but this is not always the case. Recently there has been a shift away from the dogmatic view of allosteric regulation, namely the structural view, in favour of allosteric communication based on thermodynamic fluctuations [164]. For example, the enzyme DHDPS (dihydrodipicolinate synthase), which is inhibited by lysine (the end-product of the pathway DHDPS is the first step of), shows no conformational change (at least none that is detectable) upon the binding of lysine [88]. This suggests that conformational changes alone do not account for the full story of allosteric regulation.

Another factor that supports the concept of a regulatory code is the mutability of the code. Allosteric sites can be engineered to recognise specific modulators that are not endogenous to a protein without significant disruption of biochemical activity [91, 179]. Thus one is able to re-write the regulatory code, indicating that the association between allosteric modulator and biological effect is arbitrary in nature.

The *regulatory code* would therefore explore the possibility that allosteric modulation is part of a two-world system: allosteric modulator and biological effect, linked by an adaptor molecule, in this case a specific recognition site on a dynamic protein (an effector protein). The state of the field is such that, to date, no thought has been given to a regulatory code. However, I do believe that the evidence warrants a closer look at the specifics of allosteric regulation in order to solidify, or debunk, its status as an organic

code.

3.9 The *Hox* code

Hox genes are those responsible for the correct patterning and segmentation of metazoan cells during cellular differentiation. Incorrect translation of the *Hox* genes results in fatal phenotypes.

The idea that within the expression of the *Hox* genes lies a code was developed by Hunt *et al.* [68, 69] during their investigation of the development of the vertebrate head. The definition of ‘code’ by Hunt *et al.* [69] and Ryan *et al.* [134] as the patterns of combinatorial gene expression rather than a mapping between two independent worlds with an adaptor linking them is nevertheless incorrect.

Although the *Hox* genes are sensitive to certain signals such as retinoic acid [99], this can be explained by the function of other codes, such as the signal transduction code or the histone code.

It appears that most of the current aspects of the *Hox* code can be explained by the presence of other codes. For example the proper translation of the *Hox* genes is the domain of the genetic code, whereas the correct spatio-temporal distribution of gene product as well as the timing of gene translation or repression is explained by the histone code, while the sensitivity to environmental disturbances or chemicals is under the purview of the signal transduction code. However, the possibility does exist that a ‘meta-code’ does exist which allows for the proper synchronisation of the above-mentioned codes, but this is pure speculation for now. In summation, as they stand currently, the precepts of the *Hox* code are insufficient in order for it to qualify as an independent organic code as the crucial element,

a unique adaptor molecule, is missing.

Chapter 4

The histone code

One of the major aims of code biology, besides that of discovering and elucidating new biological codes, is to examine all previously proposed biological codes in the light of Barbieri's framework in order to test whether they truly are organic codes. In order to do this for the histone code, I first provide a detailed overview of the complexities of the post-translational modifications of histones, the subsequent recognition of these modifications by specialised protein domains, and the resulting biological effects. This then makes it possible to tackle the objective of testing the histone code against the criteria that characterise a true organic code. In order for this to be accomplished, it implies that I will be able to identify certain elements of the histone regulatory system as organic signs, organic meanings, and adaptor molecules.

4.1. What are histones and the 'histone code'?

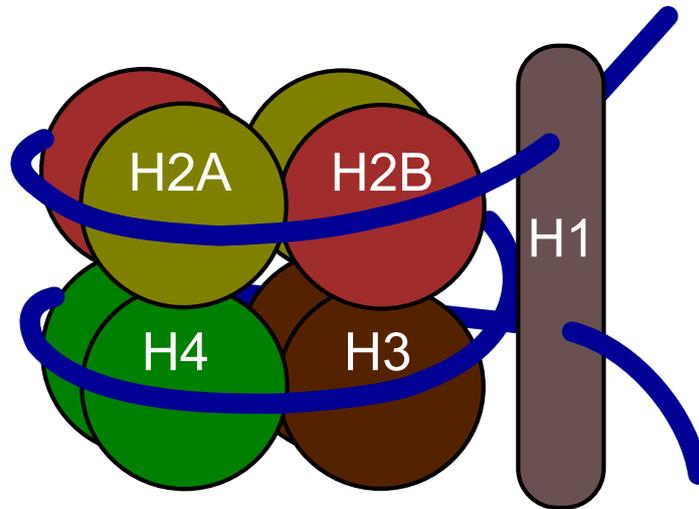


Figure 4.1: The structure of a nucleosome.

4.1 What are histones and the 'histone code'?

Histones are small, basic proteins that complex together to form a core particle around which DNA wraps to form a nucleosome [128]. This core particle consists of two molecules each of four histone types: H2A, H2B, H3, H4 that associate in two H2A-H2B dimers and one H3-H4-H3-H4 tetramer to form an octamer [178]. A ca. 146bp length of DNA, called the core DNA, wraps around this bead of histones in roughly 1.75 turns [128]. Nucleosomes are linked by stretches of DNA called linker DNA. For each nucleosome, a fifth histone type, the linker histone H1, binds to both incoming and outgoing linker DNA and joins nucleosomes to one another in strings of several thousand nucleosomes (see Figure 4.1). These long, 11nm thick fibres, called chromatin, are subject to super-helical winding and torsional forces that arrange them from 11nm to a 30nm thick and ultimately a 600nm thick fibre, called the chromosome.

4.2. The function of post-translational histone modifications

Jutting out from the core particle, and into solution, are the N-terminal tails of the histones, which are rich in basic amino acids such as lysine and arginine. These residues are often subjected to *post-translational modifications* (PTMs) through the addition of small organic molecules [84, 178]. The PTMs identified so far are acetylation, methylation, phosphorylation, ubiquitylation, SUMOylation and ADP-ribosylation [17]. These PTMs provide ‘marks’ that in turn are recognised by and bind to specialised protein domains that locally alter the chromatin structure, causing specific effects such as transcriptional activation or repression [37, 76]. It is this modification-recognition-effect system, which Strahl and Allis [156] dubbed the ‘histone code’, that will be described in detail in the next section.

4.2 The function of post-translational histone modifications

Histones play a crucial role in the development of eukaryotic life. In contrast, prokaryotes, with the exception of the Archaea, have no histones. In Archaea, histones are thought to have a purely structural function since they are involved in the condensation of DNA, but do not possess the various sites for post-translational modification that eukaryotic histones contain. Thus, while the histones of the Archaea maintain genomic integrity, they do not regulate the expression of specific genes as the eukaryotic histones are able to do [122].

What do histone PTMs allow eukaryotes to do that prokaryotes cannot? They constitute a type of epigenetic memory [94] that enables new cells to “remember” what their predecessors were and develop accordingly. This memory also allows certain cells to remember specific previous states.

4.3. The histone post-translational modification zoo

For example, neural cells have an epigenetic memory mechanism coded into the histones that allows them to generate action potentials faster the more they are used; this in turn has a further effect on the histone/chromatin structure, strengthening this memory, and making it even easier to generate subsequent action potentials [94]. The modification of histones and the resulting effects are also responsible for the spatio-temporal regulation of genetic activity during cell differentiation and development. In higher eukaryotes, for example, the *Hox* genes, which are responsible for proper embryonic segmentation and development, are regulated through histone modifications [171, 183]. Regulation through the histone code extends to virtually every gene and is absolutely crucial to the normal functioning of an organism.

Histone PTMs have been of particular importance in recent advances in stem cell research. Without the proper epigenetic programming provided by histone PTMs (in terms of timing, localisation and specificity), the project of inducing stem cells to differentiate into the desired tissue type would be impossible to realise. Those research groups that have realised this are now making the necessary effort to understand and harness the histone PTMs [57].

4.3 The histone post-translational modification zoo

What follows is a discussion of the major histone modifications through acetylation, methylation and ubiquitylation. The list of modifications is not exhaustive; I chose those modifications for which there is enough information about their recognition and effects to enable us to judge them as

4.3. The histone post-translational modification zoo

possible elements of an organic code. For most modifications by phosphorylation, SUMOylation and ADP-ribosylation the required information is to our knowledge not yet available. These modifications and their positions are summarised in Fig. 4.2).

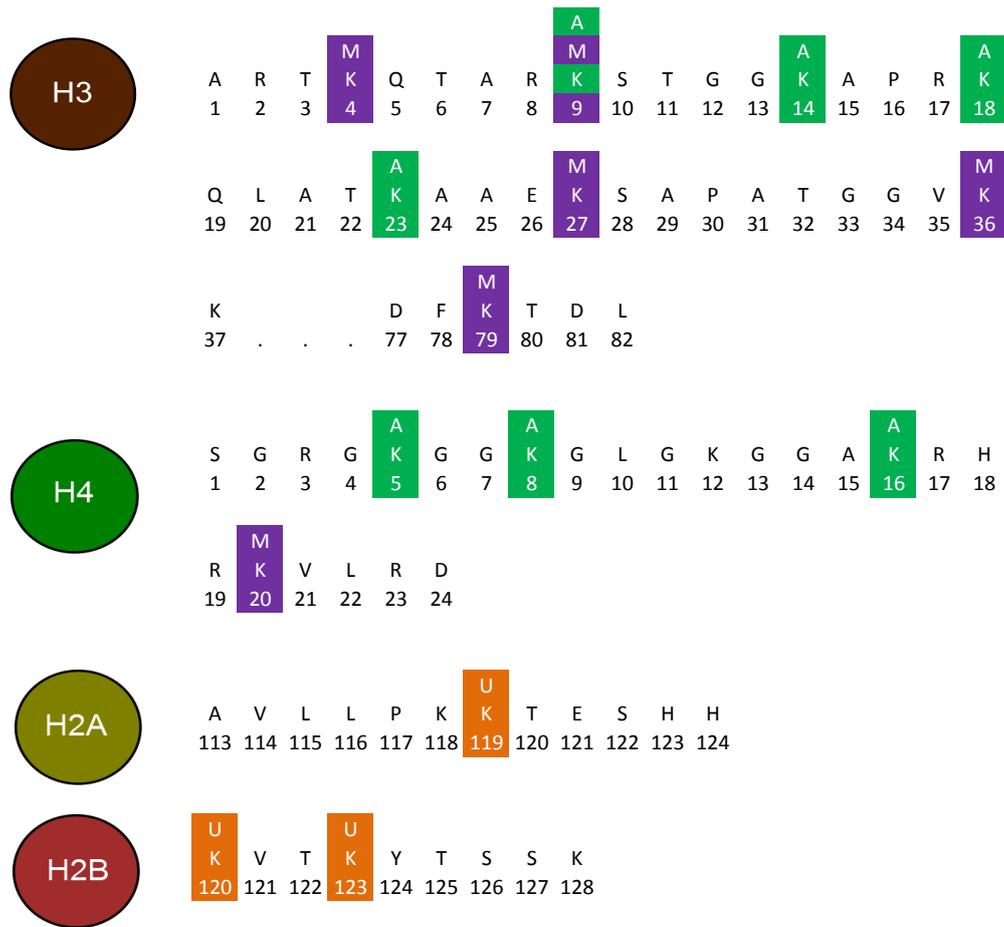


Figure 4.2: Major sites of post-translational modifications of histones H2A, H2B, H3 and H4. Symbols: A denotes acetylation, M methylation and U ubiquitylation. Sites on the polypeptide chains are numbered and identified with the one-letter abbreviations of their amino acids.

Histone post-translational modifications are highly conserved across all eukaryotic life, from *S. cerevisiae* to *H. sapiens* [96]. The best-studied post-

4.3. The histone post-translational modification zoo

translational modifications are the histone acetylations at lysine residues [175]. Originally it was thought that the decrease in positive charge brought about by the acetylation of lysine residues in histones would affect the electrochemical association between histones and DNA, and that this would result in the variety of biological effects so far observed [24, 142]. While partly true, we now know that this is not the entire picture. Several more recent discoveries have cast reasonable doubt on this hypothesis [1, 124]. The discovery of modifications through methylation, ubiquitylation and phosphorylation, none of which introduce a change in charge, provided sufficient evidence that if there is an effect on the electrical charge of histones, it would be ancillary to the main event [124].

In what follows the nomenclature convention $HxKy$ is used, where x identifies the histone in question, and y the lysine position (K is the one-letter abbreviation for the amino acid lysine, the three-letter abbreviation is Lys). Acetylation, methylation, and ubiquitylation are indicated by $HxKyac$, $HxKymet$, and $HxKyq$ respectively.

Acetylation

First identified almost 50 years ago by Allfrey *et al.* [2], histone acetylations by the histone acetyltransferases (HATs) have become the best characterised of histone PTMs [166]. Of the histones, H3 and H4 are most often acetylated, with lysine residues being the common target Rice and Allis [131]. Currently it seems that only lysine residues in histones are acetylated [85]. These acetylated histones are mostly associated with the activation of transcription [162]. The acetylation of histones is reversible through the action of histone deacetylases (HDACs) [28], which keeps them in a state of high turnover. Furthermore, it seems that the functions associated with

4.3. The histone post-translational modification zoo

acetylated histones (mostly restricted to transcriptional activation) are less diverse than those associated with methylated histones. Table 4.1 summarises the major histone acetylations, which are discussed in detail in the following.

Table 4.1: The major histone *acetylations* and *ubiquitylations*, the binding domains that specifically recognise them, and their corresponding cellular effects.

Modification	Binding domain	Biological effect	References
H3K9ac	bromodomain	Transcriptional initiation	[1]
H3K14ac	bromodomain	Transcriptional initiation	[126]
H4K5ac	bromodomain	Transcriptional initiation in embryonic genes	[31, 133]
H4K8ac	bromodomain	Transcriptional initiation and chromatin remodelling	[1]
H4K16ac	bromodomain	Transcriptional initiation (prevents the binding of the ISWI and Sir3 transcriptional silencers)	[56, 86, 87]
H2BK120ac	bromodomain	Provides a binding site for an E3 ubiquitin ligase to ubiquitylate H2AK119	[55]
H2Bk120q	Cps35	Methylation of H3K79, transcriptional activation, transcriptional repression	[75, 118, 181]
H2BK119q	unkown domain	Transcriptional repression	[172]

H3K9: Acetylation of H3K9 to H3K9ac is associated with the initiation and elongation phases of transcription [1, 126, 175]. This mark acts oppositely to methylation of the Lys9 residue, which codes for transcriptional repression. Strašák *et al.* [157] have shown that H3K9ac is also important for nuclear reorganisation of the chromatin, since the inhibition of HDACs

4.3. The histone post-translational modification zoo

caused a spike in acetylation and prevented the binding of heterochromatin protein 1 (HP1) to H3K9met. The Gcn5 HAT enzyme binds to H3K9 prior to initiation of transcription and there is a subsequent peak in acetylated histones associated with the transcriptional start site of active genes [126].

H3K14: Like H3K9ac, H3K14ac is associated with the initiation of transcription and is similarly tied to Gcn5 recruitment at the transcriptional start sites of various genes [126]. This mark also correlates with an increase in transcription rates.

H3K18: Of all the histones tested by Kurdistani *et al.* [87], acetylation of H3K18 increased transcriptional activity to the highest degree. Acetylation of H3K18 seems to prevent that of H4K16 and vice versa, although by which mechanism is not known.

H3K23: H3K23ac is acetylated at Lys23 and with, no methyl group at Lys36, binds to TRIM24, a chromatin and estrogen response modulator. These two sites are recognised by the two binding domains of the PHD-bromo cassette on TRIM24 [165] (the plant-homeo finger and bromo domains of proteins that bind to histones will be discussed in a later section). This TRIM24-H3K23ac complex is associated with the up-regulation of estrogen-related genes associated with cell-proliferation and tumorigenesis. Aberrant TRIM24 expression (and the resulting disruption in H3K23ac patterns) have adverse consequences on breast cancer survival rate [165]. Dimethylated and trimethylated H3K4met are other marks that counter activation by H3K23ac by preventing the binding of TRIM24 and thus suppressing activation of transcription [127, 165].

4.3. The histone post-translational modification zoo

H4K5: H4K5ac plays a crucial role in the earliest stages of embryonic development [31, 146] and is, like most other acetylated histones, strongly correlated with transcriptional activation [60, 133].

H4K8: H4K8ac is associated with SWI/SNF recruitment, which in turn is responsible for ATP-dependent chromatin remodelling [1].

H4K16: Unlike other acetylated histones, the absence as well as the presence of an acetylated Lys16 on H4K16 has a function. Like most acetylated histones, H4K16ac is linked to the up-regulation of transcription [56, 156]. However, removing the acetyl group from H4K16ac allows for the interaction of the Sir3 transcription-silencing protein with H4 [86]. Similarly the ISWI protein, which is part of a nucleosome-remodelling complex associated with silent chromatin, interacts with de-acetylated H4K16 [87]. Further, as mentioned before, it appears that there is an antagonistic relationship between H4K16ac and H3K18ac in that an increase in H4K16ac decreases H3K18ac and vice versa [87]. The precise reason and mechanism for this is not yet known.

H4K16ac also engages in extensive cross-talk with the transcriptionally repressive mark, H3K36met2/3. The latter recruits an H4K16-specific deacetylase which abolishes the transcriptional activation of H4K16ac [21].

H2BK120: Acetylation of H2BK120 is closely related to its ubiquitylation [55]. The addition of ubiquitin to H2BK120 needs the presence of an acetyl group at Lys120 since the inhibition of KAT3A/B, the acetylase that acetylates Lys120, by siRNA prevents ubiquitylation of H2BK120. However, these two marks are inversely correlated in that an increase in H2BK120q is accompanied by a decrease in H2BK120ac. Preliminary ev-

4.3. The histone post-translational modification zoo

idence suggests that the acetylation of H2BK120 acts as a ‘hot’ switch, essentially keeping the position primed for ubiquitylation by an E3-ligase.

Methylation

Methylated histones were originally and exclusively associated with transcriptional repression [28]. However, since the discovery of histone demethylases (HDMTs), their role in regulation has been found to be more diverse than previously thought [131, 159]. This discovery laid to rest the hypothesis that histone methylation is a permanent mark and therefore ‘locks’ the chromatin in its heterochromatin state. Histone methylation also cast doubt on the idea that the variety of structural and genetic activity that was found in chromatin was due to changes in charge-charge interactions since methylations do not alter the ionic properties of a histone [124]. However it is possible that steric hindrance or interactions may still play a role in the biological effects associated with the methylation of a histone. In the following the various functions of methylated histones that have been discovered to date (see Table 4.2) are discussed.

H3K4: Methylated H3K4 is associated with two different functions, depending on the type of protein domain that it binds to. Originally thought to facilitate mostly transcriptional activation and elongation [148], evidence also exists that H3K4met can cause transcriptional repression [147]. Transcriptional elongation is mediated by the chromodomains of transcriptional activator Chd1 that recognise di or tri-methylated H3K4met_{2/3} [127, 148] (chromodomains as adaptors will be discussed in a later section). In addition to possessing two chromodomains, Chd1 also contains a DNA-binding domain and a helicase domain [149]. When bound to H3K4met_{2/3} via its

4.3. The histone post-translational modification zoo

Table 4.2: The major histone *methylations*, the binding domains that specifically recognise them, and their corresponding cellular effects.

Modification	Binding domain	Biological effect	References
H3K4met	various	DNA repair	[82]
H3K4met2/3	chromodomain	Transcriptional initiation	[127, 148]
H3K4met3	PHD Finger	Transcriptional repression	[147]
H3K9met2/3	chromodomain	DNA methylation and transcriptional repression	[26, 152]
H3K27met3	chromodomain	Transcriptional repression, H2BK119 ubiquitylation, and H3K27 transmethylation	[98, 100, 159]
H3K36met3	chromodomain	Transcriptional initiation	[181]
H3K36met2	chromodomain	Histone deacetylation	[29, 92]
H3K79met2/3	tudor	Transcriptional initiation and DNA repair	[75, 113]
H4K20met2	tudor	DNA repair	[25]

chromodomain, Chd1 associates with various transcription elongation factors (Spt5, Pob3, Rtf1) [104, 149]. Evidence also suggests that in higher eukaryotes Chd1 is also responsible for transcriptional regulation and termination [104]. Schneider *et al.* [137] have discovered that the presence of H3K4met decreases toward the 3' end of genes, suggesting that the absence of this modification provides a signal for the termination of transcription.

On the other hand, proteins that bind to H3K4met3 via a PHD finger usually cause active repression of genes [147]. However, even this role is not as clear-cut as it seems, as there are several instances where the PHD finger is part of a protein that is involved in the activation of gene transcription [148]. Here is a clear case of the flexibility often encountered in regulatory mechanisms. Not only is H3K4met associated with activation and repression of transcription, but also with activities such as DNA repair, chromatin remodelling, and sporulation [82, 132]. Even the unmethylated H3K4 has a

4.3. The histone post-translational modification zoo

known function: binding to the PHD domain of the autoimmune regulator (AIRE) resulting in chromatin remodelling and active transcription [116]. In *Drosophila*, trimethylated H3K4met3 prevents binding of H3K27met3 to polycomb group (PcG) proteins and, therefore, repression of transcription caused by PcG proteins [141].

H3K9: To date, di- and tri-methylated H3K9 has been associated predominantly with transcriptional repression [152, 167]. This is due to H3K9met3 acting as a ‘beacon’ to which several DNA-methylases (DNMTs) are recruited [26]. It appears that repression is due to a cooperation between HP1, the protein that recognises the H3K9met3 mark, and the various DNMTs (1, 3a, 3b) [151].

Vakoc *et al.* [168], however, reported that H3K9met2/3 can also be associated with active transcription. It appears that H3K9met2/3 seems to stabilise the open reading frame to permit transcriptional elongation to occur, rather than being associated with the initiation of transcription. Certain genes show increases in H3K9met3 *and* H3K4met3 levels upon initiation of transcription, but, despite this apparent correlation, the evidence that suggests that these methyl marks interact with one another is weak [168].

H3K27: H3K27 is methylated by the SET-containing E(Z) (EZH2 in humans) proteins which form part of the Polycomb group (PcG) of proteins [100]. H3K27met3 is read by the Polycomb repressor-complex 1 (PRC1), which contains a chromodomain. H3K27met3 is mainly associated with the repression of transcription, X chromosome inactivation and genomic imprinting [100]. Furthermore, it appears that H3K27met is an important signal for the localisation of a PRC1-like E3 ligase that ubiquitylates histone 2A at Lys119 since, in the absence of E(Z), H2AK119 ubiquitylation

4.3. The histone post-translational modification zoo

does not occur [100]. However while they are related, the methyl mark is not dependent on the ubiquitin mark, indicating that H2AK119q depends on H3K27met3, and not vice-versa [172].

So far, H3K27met3 has been linked only to transcriptional repression [159], which is due to the interactions between H3K27met3 and various regulatory complexes, and not to alteration of the nucleosomal architecture [150]. This holds for mono-, di-, and tri-methyl marks [100]. It was first thought that these methyl marks cause permanent repression, but the recent discovery of a H3K27-specific demethylase undermines this idea [39].

Another interesting feature of H3K27met3 is that it is a self-perpetuating mark [98] like H3K9. H3K27me3 recruits a histone methyltransferase (HMT) to monomethylated H3K27me2 to its trimethyl form. Of the three types of methyl marks, dimethylated H3K27me2 is the most abundant, being present in roughly 50% of nucleosomes [98, 141]. While H3K27met2 itself is of limited importance in gene repression, evidence suggests that it may not only act as an inactive precursor of H3K27met3, but also prevents methylated H3K27 from being acetylated to H3K27ac[98], a mark associated with active transcription and antagonistic to H3K27met3 [36].

H3K36: H3K36met prevents the methylation of H3K27 and is commonly associated with the activation of transcription. Experimental evidence shows that the repressive mark H3K27met3 rarely co-exists with the activating H3K36met2/3 in the same histone [181]. This study also showed that histones rarely, if ever, exist without one of these modifications, and that nucleosomes therefore do not exist in a ‘blank’ state. It also showed that although a pre-existing H3K36-methyl mark does inhibit the methylation of H3K27 by PRC1, the reverse is not true. This indicates that the

4.3. The histone post-translational modification zoo

unmethylated H3K36 position could play a role in the binding of PRC1 or, alternatively, that the methylated H3K36 somehow prevents the binding of PRC1. The exact mechanism is not known, but it is clear that the Lys36 position in H3 is important since a mutation of this lysine to alanine decreases PRC1 activity considerably [181].

The Eaf3 protein subunit of the deacetylase Rpd3 contains a chromodomain essential to the recognition of H3K36me2 by Rpd3. The absence of this subunit or its chromodomain has been shown to leave histone acetylation levels unchanged [92]. Interestingly, Carrozza *et al.* [29] have identified the H3K36me2 mark in actively transcribed regions of the genome and it has been positively associated with transcription elongation. However, the deacetylation that results from the binding of Rpd3 is linked to transcriptional repression [92]. This mark has also been strongly associated with the recruitment of a HDAC to H4K16ac during active transcription [21] effectively inducing transcriptional silencing. This suggests an intricate web of cross-talk and inter-regulation between the various histone modifications. A possible function of this deacetylation is to prevent spurious transcription from being initiated [92] or to regulate the length of the open reading frame in order to allow alternate transcripts to be produced or simply to signal the end of transcription, allowing the euchromatic area to condense to its heterochromatin state again.

H3K79: Methylated H3K79 is associated mostly with transcriptional activation, however it has also been found to occasionally result in transcriptional repression [154, 158]. The most interesting aspect of H3K79 is the extensive crosstalk with H2BK120q [75]. H2BK120q has been shown to recruit the Dot1 methylase, which is responsible for more than 90% of the

4.3. The histone post-translational modification zoo

H3K79 methylations [75, 105, 108]. Khan and Hampsey [80] have shown that it is indeed H3K79met that is associated with transcriptional activation, since the replacement of Lys79, or the deletion of the dot1 gene, both result in the silencing of a particular genic region. The methylated Lys79 seems to be particularly enriched on the histone variant H3.3, which is prevalent in actively transcribed regions of the genome, but there is no clear understanding as to why this is so [106].

H4K20: Unlike the previously mentioned histone methylations, the methylation of H4K20 is possesses a single function only; trimethylated H4K20 is associated with transcriptional repression [138].

Ubiquitylation

Ubiquitylation of histones involves the addition of ubiquitin, a highly conserved, 76 amino acid protein molecule, to the lysine residues in a histone. It was the ubiquitylation of H2A that first heralded the discovery of ubiquitin and the modification of histones [118, 135]. Unfortunately the importance of both discoveries has been underestimated for some time. Unlike acetylated or methylated histones, the two instances of histone ubiquitylation which have been studied in some detail (see Table 4.1) seem to be recognised by structurally unrelated binding domains, although this is not yet certain. Ubiquitin is ubiquitous (hence the name) and diverse enough in function to perhaps warrant the investigation of a code of its own [83]. Originally it was proposed that histone-ubiquitylation affected transcription via three possible mechanisms: firstly ubiquitin itself, due to its relatively large size, directly affected the chromatin structure and histone/DNA affinity, secondly that ubiquitin acted as a beacon for the recruitment of various

4.3. The histone post-translational modification zoo

regulatory proteins, and thirdly that ubiquitin affected transcription by directly influencing the other histone modifications [182]. Currently the evidence seems to favour the third option, however the first two have not been entirely ruled out.

H2AK119: In most most eukaryotes, the ubiquitylated H2AK119 (H2AK119q) is present in 5–15 % of histones [38]. H2AK119q is overwhelmingly associated with transcriptional repression, either directly or through the indirect mechanism mentioned earlier that involves the recruitment of ubiquitin to transcriptionally silent chromatin (such as a Barr body) that has already been marked by H3K27met3 [38, 172]. Ubiquitin is ligated to H2A by the PRC1-like proteins Ring1A and Ring1B, both of which are crucial for forming and maintaining the H2AK119q mark [38]. These PRC1-like proteins contain a chromodomain that specifically reads the H3K27met3 mark, directly implicating it in the ubiquitylation of H2AK119 [100]. However, Tavares *et al.* [160] have recently shown that while H3K27met3 does code for H2AK119q, it is not essential since PRC2-null mutants, which abolish H3K27met3 entirely, show near normal levels of H2AK119q. The precise mechanism by which H2AK119q is able to facilitate the repression of transcription is not yet known.

H2BK120: H2B is ubiquitylated in higher eukaryotes at Lys120 and in lower eukaryotes at Lys123. Both H2BK120q and H2BK123q have been shown to directly influence the methylation of H3K79 by the Dot1 methylase [75, 105] and the methylation of H3K4 by COMPASS, a Set1 methyltransferase [89, 118]. Both H3K79met and H3K4met are associated with transcriptional regulation, however the majority of H2BK120q-mediated methylation is responsible for transcriptional activation. It is interesting

4.4. Binding domains: The adaptors of the histone code

that the presence of H2BK120q does not affect the methylation of H3K36, which is also linked to active transcription [181]. It also seems that H2K123q is not essential for the monomethylation of histones, since in the absence of H2BK123q, Dot1 and COMPASS still monomethylate H3K4 and H3K79, however what is impeded is their ability to di and trimethylated these positions in the presence of a pre-existing monomethyl mark [143]. Osley [118] has shown that H2BK120q/H2BK123q can also function as a transcription-repressing mark, although the precise mechanism for this is unknown. It is thought that, as with transcriptional activation, this is due to the effect that ubiquitylated histones have on the other histone modifications [43, 182].

4.4 Binding domains: The adaptors of the histone code

Acetyl-recognising domains

Currently, the only protein domain known to be capable of recognising acetyl-lysines is the *bromodomain* [110]. For each acetylated Lys in a particular histone there is a specific bromodomain. Within their ca. 110 amino acid structure, bromodomains contain a conserved hydrophobic pocket of aromatic amino acids that specifically recognises a specific acetyl-lysine [110, 119]. Studies have shown that if one or more of the critical residues in this pocket are mutated, the bromodomain loses its ability to recognise a specific acetyl-lysine [109]. This shows that the bromodomain is absolutely essential for the recognition of acetyl-lysines and that this structural domain fulfils one of criteria for being an adaptor in the histone code, recognising the sign posed by a specific acetyl-lysine.

Methyl-recognising domains

Unlike acetylated histones, the methylated histones recognised, depending on their location and degree of methylation, by a much greater variety of protein domains, such as the Royal family of protein domains [102]. The most common of these are the *chromodomains*, *Tudor*, *plant-Agenet*, *MBT* and *PHD finger domains*, as well as several smaller domains, such as *PWWP* and *JMJ* [30, 81, 107]. Each of these domains discriminate according to the degree of methylation, the position of the lysine and frequently even according to the surrounding residues, although there is evidence that suggests that the latter serve only to strengthen binding: the crucial element remains the modified residue [85].

Evidence is emerging which suggests that the chromodomains can be experimentally exchanged between proteins [47]. In this study the chromodomains of the protein Polycomb (Pc) and of HP1 were interchanged, giving Pc^{HP1} and $HP1^{Pc}$ respectively. As a result, Pc^{HP1} recognised the original target of HP1 and $HP1^{Pc}$ recognised that of Pc (H3K27met and H3K9met respectively).

This suggests that although these domains give each protein a specific identity in terms of being able to recognise a specific modification, they are not 'locked' to a protein. If these findings can be confirmed, and even expanded to include inter-domain exchanges, they would further cement the claim that a protein domain acts as the molecular adaptor for the histone code. We can however say with confidence that those domains involved in the recognition of methyl-lysines are molecular adaptors for the histone code. They ably recognise the methyl-lysines as organic signs and subsequently translate them into their corresponding biological effect.

4.5. How does it all fit together? Is the histone code an organic code?

Ubiquitin-recognising domains

While the number of domains that interact with ubiquitin is large, very few have been found in proteins that specifically interact with histone-bound ubiquitin. One of these is the *zinc-finger (ZnF)*, ubiquitin-specific processing protease (UBP) which is found in the HDAC6 deacetylase [71, 72].

Unfortunately it seems that, currently, no ubiquitin-binding domain has been identified on the Dot1 or COMPASS methylases which bind to H2BK120q. However if one is to be found, it is likely to be found on Cps35 which binds to H2BK120q and then recruits the COMPASS methylase [93]. The domain responsible for the binding of Dot1 to ubiquitin is to the best of our knowledge not known.

4.5 How does it all fit together? Is the histone code an organic code?

As discussed earlier, for the histone modification system to act as an organic code we need to demonstrate that not only does it consist of two independent worlds, here that of histone modifications (which would act as organic signs) and their biological effects (the meaning of the signs), but that there are chemical molecules, called adaptors, that recognise the signs with absolute specificity and translate them into their meanings. Furthermore, it must be possible in principle (and, preferably, by experiment) to alter the rules of the code, i.e., the relationships between organic signs and their meanings, by interchanging those parts of the adaptor molecules that recognise the signs.

4.5. How does it all fit together? Is the histone code an organic code?

From the above discussion of histone post-translational modifications it is clear that each of these modifications is linked to a highly specific biological effect; to our knowledge there are no instances where a particular PTM in a particular organism results in more than one biological effect. These relationships can therefore be regarded as a set of rules between the independent worlds of PTMs (the organic signs) and biological effects (biological meanings). In order for this set of rules to be regarded as code, it is, however, necessary to establish that are molecules that act as the adaptors that translate signs into their meanings. From the details of the histone PTMs it is clear that the role of adaptors is played by the *effector proteins* that consist of a binding domain that specifically recognises the PTM and a domain that acts as a mediator of biological effect associated with that PTM, albeit transcriptional regulation, structural remodelling of chromatin, or even a post-translational modification of another histone.

As explained previously, the relationship between an organic sign and its meaning in an organic code must be arbitrary in the sense that it is not determined by the laws of chemistry or physics (although completely compatible with these laws), but rather has the nature of a convention that arose naturally through evolution. The relationship between a histone PTM and its biological effect fulfils this criterion. The recognition of a specific histone PTM by its corresponding binding domain is analogous to the interaction of a codon on mRNA with its corresponding anticodon on an tRNA. Without the binding domain being part of the effector protein, the effect specified by a certain PTM will not come to be. For example, the mammalian Brd4 protein, a protein involved in transcriptional regulation, contains two bromodomains, and the deletion of even one of these bromodomains abolishes the interaction of Brd4 with acetylated histones [41] and

4.5. How does it all fit together? Is the histone code an organic code?

prevents the biological effect of Brd4. In another study, Flanagan *et al.* [48] showed that the mutation of tryptophan 64 or 67 in the active site of one of the two chromodomains of a CHD1 protein significantly reduced the ability of this protein to bind H3Kmet4.

That the histone code exhibits the required arbitrariness of an organic code has been proven by experimentally altering the coding scheme. As previously mentioned, Fischle *et al.* [47] replaced the binding domain of one effector protein with that of a different effector protein; the modified effector protein now had the binding specificity of the other one. More specifically, they interchanged the chromodomain of the polycomb (PC) protein with the chromodomain of heterochromatin protein 1 (HP1). The hybrid HP1^{PC} now only recognised the H3K27met mark, the target of PC, instead of the original H3K9met. Similarly, the hybrid PC^{HP1} now recognised H3K9met instead of H3K27met.

An example of the histone code in action is provided by the TAFII250 (transcription initiation factor TFIID 250 kDa subunit) protein that orchestrates transcriptional activation. TAFII250 achieves this by binding to the promoter of a gene, thus acting as a scaffold for the assembly of the transcription complex, and positioning RNA polymerase correctly. The primary targets for TAFII250 were shown to lie on either H4 (at the lysine residues K5, K8, K12, and K16) or on H3 (the lysine pair K9 and K14) [37, 73], with the latter pair being the more prominent. Functionally TAFII250 appears to be incredibly diverse, from histone acetylation or ubiquitylation, to phosphorylation of other transcription factors [115, 125, 176]. The underlying theme however, is that in all cases TAFII250 is responsible for the initiation and progression of transcription.

A typical day in the life of TAFII250 begins as follows:

4.5. How does it all fit together? Is the histone code an organic code?

1. H3K9 and H3K14 become acetylated. These newly acetylated residues exist without function until they are bound by TAFII250. Usually, H3K9 and H3K14 become acetylated in response to an environmental stimulus (such as a viral infection [1]) that resulted in the recruitment of a histone acetyltransferase such as Gcn5.
2. TAFII250 binds via both of its bromodomains to H3K9ac and H3K14ac. The bromodomains specifically and discriminately recognise and bind to these acetylated lysines.
3. Once bound, TAFII250 either acetylates upstream histones, ubiquitylates histone H1, or phosphorylates TFIIF, which in turn promotes transcription.

Figure 4.3, depicts how the double bromodomains, in the absence of the requisite acetyllysines, are unable to bind to H3. Only once H3K9 and H3K14 have been acetylated is TAFII250 able to bind H3 and perform its function(s).

In other words, we have the creation of an organic sign, the binding of an adaptor molecule, and the translation of that organic sign into biological meaning.

Transplanting the double bromodomains of TAFII250 onto another protein, Brd2, confers the binding specificity of TAFII250 onto this protein [79], indicating that the double bromodomains are the molecular adaptors in this case, as they confer their binding specificity to other proteins regardless of the original target of those proteins.

The biological meaning of the H3K9 and H3K14 signs, here transcription initiation, would not have come about if a) H3K9 and H3K14 had not been acetylated [1], or b) if the TAFII250 protein were made non-functional

4.5. How does it all fit together? Is the histone code an organic code?

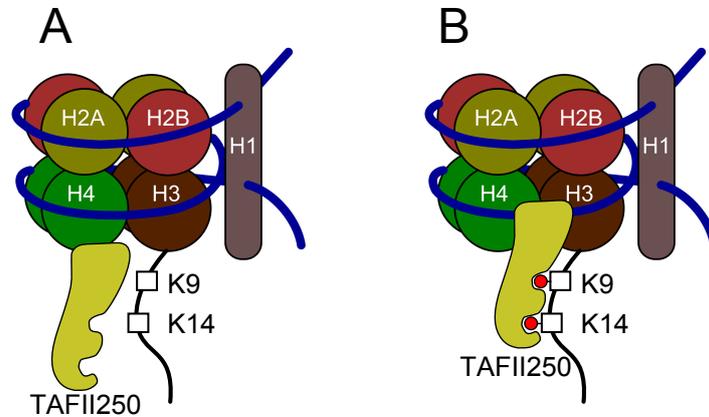


Figure 4.3: A: In the absence of acetyl groups on lysine 9 and 14 on histone 3, the double bromodomains of TAFII250 are unable to bind to H3K9 and H3K14 and as a result, TAFII250 does not phosphorylate TAFIIF, which in turn does not lead to transcriptional initiation. B: once H3K9 and H3K14 have been acetylated (red circles), the double bromodomains are now able to recognise and bind the H3K9ac and H3K14ac, which allows TAFII250 to phosphorylate TAFIIF and thus permit transcription to proceed.

by either silencing the TAF1 gene or removing either one of the bromodomains [62]. This emphasises that both the post-translational modifications (organic sign) *and* the bromodomains (adaptor molecules) are necessary for transcription initiation (biological meaning).

These considerations show that the histone code fulfils all the criteria for an organic code. Although we probably do not yet know the complete histone code, we have, as argued in this paper, more than enough information to be able to recognise the histone code as a *bona fide* organic code.

Whereas the genetic code, which after its discovery came as a “bolt from the blue”, was quickly surrounded by a “protective belt” that emptied it from all its revolutionary potential [11], we hope that we have ensured with this paper that the histone code does not suffer the same fate.

4.6 Criticisms of the histone code model

Recently, the histone code hypothesis has been criticised by Liu *et al.* [95] and Rando [129].

The study by Rando [129] makes the argument that the loss of certain residues (in terms of a K \rightarrow A mutation) differs little from a similar mutation at another residue. For example, it is asserted that the loss of H3K9 acetylation is similar to the loss of H3K18 acetylation. The histone code developed in the preceding sections confirms that, indeed, both H3K9ac and H3K18ac code for transcriptional activation and that loss of acetylation at either residue would negatively influence this process. However, what is not mentioned is that different residues are often modified in response to different stresses as varied as salt-stress responses (in *Arabidopsis thaliana*) or T-cell activation (in mouse tumor cells) [32, 33]. Rando [129] goes on to question whether histone modifications do anything at all, citing that the deletion of the H3K4 methylase Set1 has a minor effect on transcriptional activity. However, what the author fails to mention is that H3K4met0 *also* has an effect. The auto-immune regulator (AIRE), for example, contains a PHD domain that is able to recognise and bind unmethylated H3K4, and elicit a transcriptional response [34]. It is the modification *state* rather than the modification itself that must be regarded as coding for biological effects. Further, Rando [129] seems to argue that the differences between *in vitro* and *in vivo* histone modification patterns show that histone-associated proteins lack the necessary specificity to participate in a code.

Crucially, however, Rando [129] appears to use the term ‘code’ when referring to the pattern of histone modifications and the resulting localisation of the associated proteins. The study therefore does not examine the histone code hypothesis from the point of organic codes. While it is true

4.6. Criticisms of the histone code model

that, nominally, a link between a post-translational histone modification (or patterns thereof) and a localisation event has been assumed, the salient point is that the histone code describes a mapping from modification to biological *effect*. Indeed, the author makes certain points for the histone code as being an organic code. He mentions that the deletion of the Eaf3 chromodomain of the Rpd3S complex has no effect on the localisation of the complex (it binds to the RNA Polymerase II C-terminal domain), but the loss of interaction with H3K36met3 affects the functioning of Rpd3S, which is a histone deacetylase [129]. Thus, while the author mentions the various binding domains, they are never more than vehicles that localise the various protein complexes to their respective targets instead of the mediators between the histone modification and the resulting biological effect—they are therefore not seen as adaptor molecules.

The criticism by Liu *et al.* [95] appears to be focused on the combinatorial nature of the histone code, in particular that more than one modification, or combination of modifications, can result in similar outcomes (in this case, transcriptional activation). The authors' view of the 'code' aspect suggests that this degeneracy is not suggestive of a code.

The authors mention that modifications rarely occur in discrete states, rather they seem to exist in a continuum of modification states [95]. However, when viewed in the context of their assertion that histone modifications are subject to high turnover and that the methodology they employed did not allow for the examination of single nucleosomes, but rather provided a population average of modification states that could not rule out the possibility of discrete states being obscured, we see that a continuum of states is not unexpected, especially when we consider that, as Fischer *et al.* [45], Rando [129], Wang *et al.* [175] have pointed out, there

4.6. Criticisms of the histone code model

is combinatorial complexity in how histone modifications are read to bring about biological effects. The matter is further complicated when we consider that many histone modifications are the result of cross-talk with one another [46, 70, 90, 180]. For example, the ubiquitylation of H2BK120 results in the subsequent methylation of H3K79 and H3K4 by the Dot1 and COMPASS methylases respectively (see section 4.3). This led to Henikoff [63] asking the question whether there is true combinatorial complexity or whether it is cumulative simplicity, since the rapid turnover of histone acetylations (in particular) complicates matters when attempting to tease apart this question. Liu *et al.* [95] conclude that the modification patterns they observed are often the result of rather than the cause of transcription. This is true, to a degree. The histone acetyltransferase Gcn5 has been shown by Pokholok *et al.* [126] to be recruited to and acetylate H3K9 prior to initiation of transcription. It is however important to remember that many histone modifications are implicated in the elongation phases of transcription, particularly H3K9ac [126], H4K5ac [148], and H3K9me2/3 [168]. Further, Henikoff [63] describes a mechanism whereby nucleosomes are excised during transcription and moved further along the chromatin strand before being reinserted, potentially confusing experimental data concerning which modifications or patterns thereof occur when and where. Lastly, Liu *et al.* [95] performed their study on actively dividing yeast. During this process the entirety of the yeast genome would be subject to active transcription, potentially muddying the data they gathered concerning the transcriptional nature of histone modifications. As mentioned previously, histone modifications can often be the result of environmental stimuli—in a state of pervasive transcription, in the absence of disruptive environmental stimuli, it is unclear how much of the more subtle behaviour of histone

4.6. Criticisms of the histone code model

modifications was lost. Finally, when one considers that these conclusions were made in the absence of histone methylation or histone ubiquitylation data, I suggest that the combinatorial complexity of histone modifications is more nuanced than suggested by the authors.

Chapter 5

The Görlich-Dittrich algorithm for identifying ‘molecular codes’: A critique

In a recent paper, Görlich and Dittrich [59] propose an algorithm for identifying codes and classifying the elements along the lines of signs and meanings. In doing so they first operationalise (partially) the concept of a code. While they do, nominally, take their definition of a code from Barbieri as, “a mapping from sign to meaning”, they use contingency as their defining criterion for a code. Although they do mention adaptors in passing, this crucial element of organic codes are not part of their definition of what they call ‘molecular codes’. Instead they use a reaction network-based approach where the concepts of ‘sign’ and ‘meaning’ ultimately lose their value as they become stand-ins for ‘left-hand side’ and ‘right-hand side’ molecules.

The authors state that a reaction network is able to implement a *molecular* code if one set of the molecular species can be mapped onto another set of molecular species. This is exemplified by Fig. 5.1, where the left set,

which contains the elements, A and B, is mapped by one of two ‘contexts’ (E and G, or F and H) that are in reality closer to chemical reactions, to the right set containing the elements C and D [59]. (E,G) and (F,H) can be thought of as pairs of enzymes that exist under different sets of conditions.

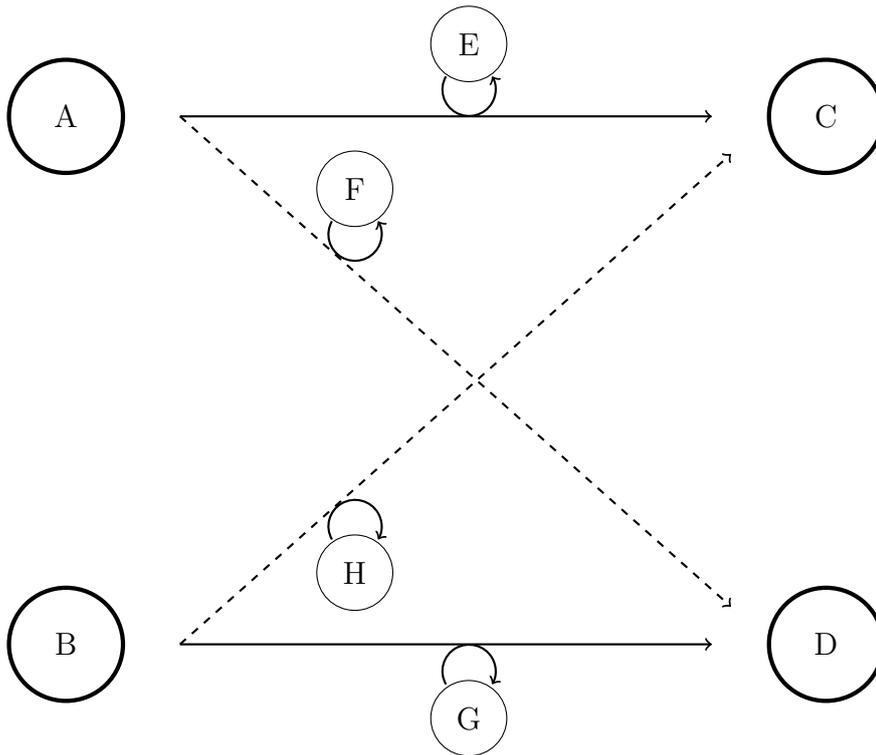


Figure 5.1: A binary molecular code according to Görlich and Dittrich [59]. The set, $S = \{A, B\}$, is mapped to the set $M = \{C, D\}$ by the contexts, $C = \{E, G\}$ or $C' = \{F, H\}$.

Further, a molecular code is contingent—the mappings could be different—and, following from this, alternative mappings *must* exist for a reaction network to implement a code, since the *contingency* of a code implies that another context exists under which the signs are interpreted differently [59]. While it is true that this ‘contingency’ is a quality of most codes, it is not a necessary quality for a code. Imagine that the genetic code was universal

with no variations and therefore no contingency. In terms of the Görlich-Dittrich definition their algorithm would not recognise it as a code. However, for the biochemists who discovered the genetic code (and at that time that was the only known instance) this lack of contingency posed no problem for them to recognise the genetic code as a true organic code. From purely chemical considerations it was clear enough that there is no prior deterministic relationship between nucleotide and amino acid sequences.

Another necessity for a molecular code is the *molecular context*. The definition provided for the molecular context is however highly ambiguous; according to Görlich and Dittrich [59], the molecular context is, “necessary for the reaction to happen”. This definition offers two problems: firstly, an organic code (defined in Section 2.1), is not a transformative chemical reaction, and secondly this definition does not implicate *adaptors* specifically, rather it allows for a slew of other agents to be ‘necessary’. The genetic code for example necessitates the various RNA polymerases, an mRNA molecule, a pool of amino acids, GTP, and a host of other molecules—each of which is necessary for the reaction (here being translation) to happen. This issue is further reinforced by the reaction network approach where the authors imply that a code details a *transformative reaction* rather than an *association*. Reaction networks detail deterministic relationships between related molecular species, whereas organic codes map out non-deterministic relationships between *independent* molecular worlds.

The epimerisation of D-glucose and D-talose as described in Figure 5.2 would satisfy their criteria for a molecular code, but fall short of those set for an organic code. In this reaction network it is clear that one set of molecular species are mapped onto another: D-glucose and D-talose onto D-mannose and D-galactose. This mapping is contingent in the sense that

an alternative molecular context (C-2 or C-4 epimerisation) that is able to alter the mapping and that at no point is it dictated by natural law which of these mappings are preferred. The algorithm devised by the authors would therefore identify the epimerisation of D-glucose and D-talose as a molecular code. This would certainly not convince any biochemist or molecular biologist. A molecular code therefore also refer to a situation where the signs are *indexical* in nature rather than *symbolic*. Indexical signs (D-glucose and D-talose) represent objects (D-mannose and D-galactose) by virtue of a physical link that exists between them (in this case a strong structural similarity). Symbolic signs (a specifically acetylated histone, or a nucleotide codon) represent objects (a particular biological function or an amino acid) by entirely arbitrary links that have no established physical link between them (H3K14ac could just as easily have specified a silent transcriptional state, or UUU could have specified serine instead of phenylalanine).

Further, how the algorithm would be able to identify the ‘signs’ and ‘meanings’ by itself is unknown. It seems likely that, if presented with a reaction network the algorithm would not be possible to identify true organic codes. Instead it would identify a slew of mappings that could by no means be considered a code. The authors have demonstrated this by amalgamating 17 of the known genetic codes into a single reaction network and subjecting this to the algorithm. While the algorithm did manage to identify some 16 molecular codes, it did not identify a single cogent genetic code. Rather it identified multiple instances where more than one codon specified for an amino acid across each of the genetic codes [59]. Another problem that crops up is how the algorithm distinguishes between sign and meaning *if this is not explicitly defined* prior to running the algorithm? If the mappings of a reaction network are not made explicit, but instead the

algorithm is provided exclusively with the signs and meanings (it is therefore blind to which is which), it would not be possible for it to reliably distinguish between these sets. Here it would have been useful to introduce the concept of an adaptor (as with organic codes) as this molecule is the ‘fingerprint’ of an organic code—the sign-recognition site would have immediately identified which of the entities are the signs (as with the anticodon of a tRNA molecule) and the meaning-implementation site would have immediately identified which is the biological meaning (as with the effector site on chromatin-associated proteins). This also highlights a crucial failing of the algorithm, it functions only when presented with a reaction network, the large set of organic codes that deal with molecule \rightarrow effect mappings would remain untouched.

The identification of organic codes by algorithms therefore, seems to be a task that, for the time-being, is not possible. Firstly, an adaptor molecule—the molecule that ultimately associates the world of signs with that of meanings—is not part of the definition. Secondly, unless the associations are made explicit, the algorithm seems unlikely to be able to identify them correctly. Thirdly, the reaction network approach opens up much scope for ambiguity that allows for the identification of molecular codes (such as the example in Fig. 5.2) that clearly do not function as an organic code, robbing the term of much of its impact. Lastly, the algorithm is only able to identify *contingent* molecule \rightarrow molecule mappings, this implies that a large number of codes such as the histone code, the sugar code, the compartment code, the regulatory code and many more would not qualify as molecular codes.

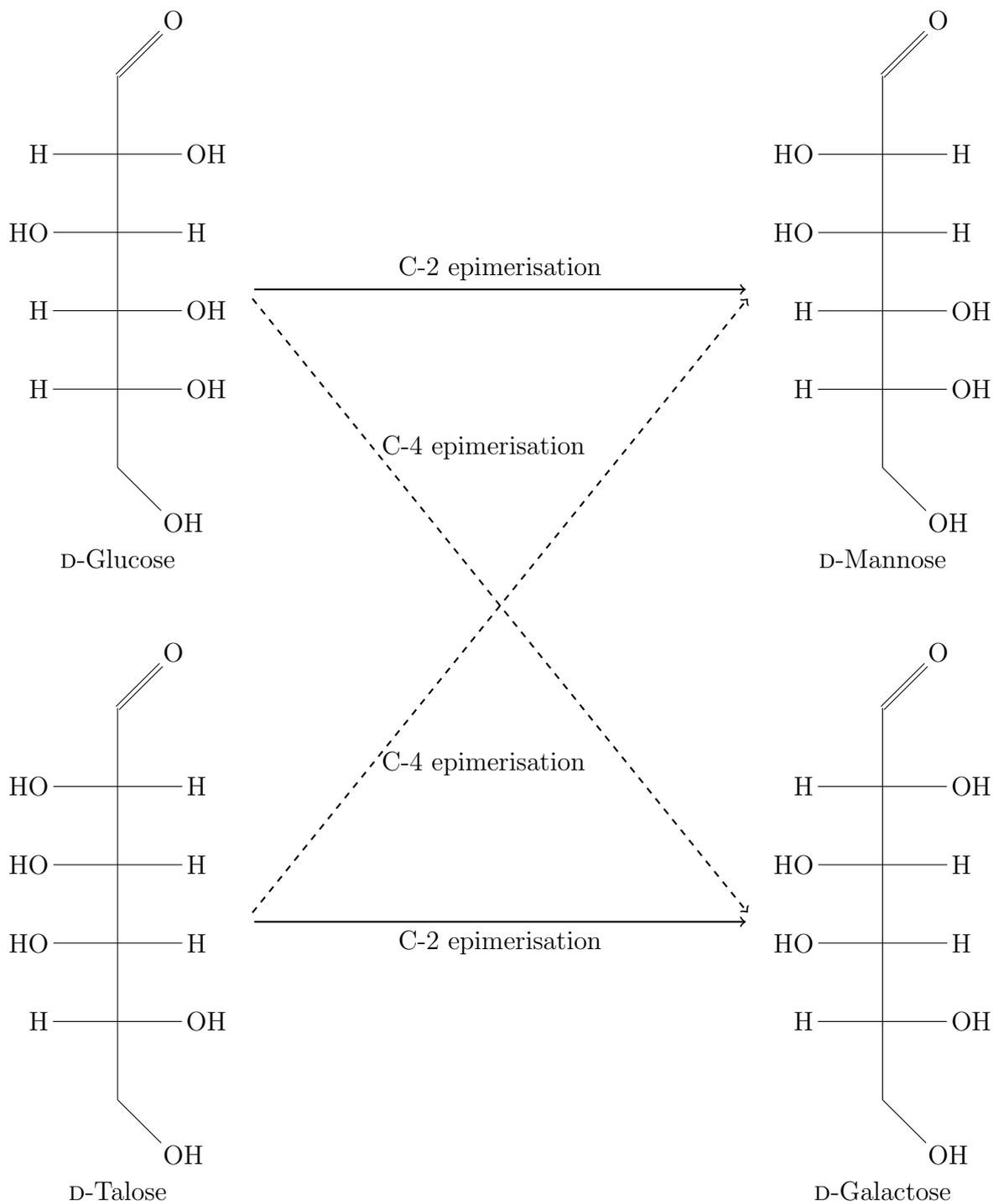


Figure 5.2: The mapping of the reaction network mapping D-Glucose and D-Talose onto D-Mannose and D-Galactose

Chapter 6

Discussion

The use of the term ‘code’ has cropped up time and again in the biological sciences—more so since the discovery of the genetic code. However, what has rarely been made explicit is precisely what is meant by ‘code’. Therefore, one of the first tasks of the research described in this dissertation was to formulate a definition of an *organic code* based on Barbieri’s work on code biology. This definition as well as the ensuing characteristics a putative coding system needs to fulfill form the bedrock of this thesis. Without a unifying framework, the field of biology would be replete with codes that in reality are not codes. In Chapter 3 I touched upon two systems where the classification of a system as a code has been a misnomer: the ‘metabolic code’ and the ‘*Hox* code’ were shown not to be codes according to the criteria that I laid out in Chapter 2. Rather than forming organic codes, these systems are in reality closer to being ‘fingerprints’, which in both instances need a scientist to interpret them. Code biology is not the study of these molecular fingerprints; it is rather the study of how cells themselves are able to implement coding systems *without* the need of an outside agent to provide interpretation. Cells are able to translate organic signs into biolog-

ical meaning in the absence of a mind or the need for interpretation. This distinction between molecular fingerprinting and biological coding has been necessary for some time since since code biology is currently in the process of cataloguing the various biological codes with little emphasis given to their suitability as organic codes.

When we regard a code, it would be remiss to do so without paying due consideration to the role that *information* plays in the function of a code. To that effect I distinguished between two ways in which we deal with information: one concerned with only the reliable transfer thereof and the other with the reliable translation thereof. These two processes underpin much of information theory as it applies to biological systems. First I demonstrated that real parallels exist between biological mechanisms and those proposed by information theorists. The reliable transfer of information finds succour in the biological world as the transmission of DNA from one generation to the next. In particular, I believe that the explanation of sequence redundancy as a feature to improve overall robustness of the sequence itself is an important step toward the synthesis of information theory and biology. Building on the reliable transmission of information, the next, and possibly more important, consideration deals with the reliable translation of this information so that it may be used in one form or another. To wit, the removal of redundancies is of paramount importance since the translation of redundant bits (non-coding DNA) would be a waste. Once the meaningful bits are extricated, they must be translated (much like the strings of 1s and 0s must be converted to English for this to be legible). The parallels I point to here are the splicing of mRNA in order to remove non-coding sequences (introns) and, of course, protein translation—the process that translates mRNA sequences to amino acid sequences. Finally I

demonstrated that the concepts of codes and information are inextricably linked—information makes sense only when viewed through the lense of the appropriate code—and that *organic codes* in particular offer a suitable framework for the integration of these concepts into molecular biology.

Bearing this and the foregoing discussion of the criteria of an organic code in mind, I then turn to an analysis of several systems thought to be codes. The genetic code makes an excellent and obvious choice to test against the criteria for an organic code. As befits the progenitor of organic codes, it passes the test. There are codes that do not pass this test such as the *Hox* code or the metabolic code. The ubiquitin code, while conforming to the criteria, presents a problem in the sense that there is uncertainty whether it is a discrete code or whether it is subsumed by a more inclusive ‘protein post-translational modification’ code. Ultimately I present a novel framework for the identification and classification of a ‘regulatory’ code that describes the association between allosteric effectors and enzymatic behaviours, although not as fully understood as the genetic code (or the histone code), I do believe the preliminary inspection of the associated elements warrants further investigation. Interestingly, what does crystallise from this chapter (Chapter 3) is that a large cluster of organic codes (possibly the majority) are concerned with a molecule \rightarrow effect correspondence rather than a molecule \rightarrow molecule correspondence as with the genetic or signal transduction codes.

What exactly do I mean by ‘further investigation’? Chapter 4, which concerns a thorough investigation into the elements of the histone code, is just such an investigation. In this chapter I explore, thoroughly, the crucial elements of the histone code: the post-translational histone modifications, the slew of associated biological effects, and the specific binding/effector

molecule pairings that allow for the translation between these phenomena. I then draw parallels between these elements and those of organic codes, identifying them as organic sign, biological meaning, and adaptor molecules respectively. I believe that I have, in this chapter, presented sufficient evidence for the identification of the histone code as a *true* organic code. What the foregoing attempts at elucidating a histone code have all lacked was either the unifying framework provided by code biology or the in-depth analysis and synthesis of the existing elements.

Finally I turned to the attempt to identify codes by algorithmic means. While it would be welcome to obviate the necessity for an exhaustive investigation into the elements of a putative code in favour of a speedier, computer-based approach, the theoretical framework that underlies just such an approach (as advocated by Görlich and Dittrich [59], is flawed. Not only does their concept of a ‘molecular code’ not make any mention of an adaptor, the reaction network based approach is one founded on transformative reactions and indexical signs rather than an association of symbolic signs. This opens the door for the misidentification of deterministic reactions (such as the epimerisations in Fig. 5.2) as molecular codes, when in fact these are not codes in the least. The other problem with the algorithmic approach is that, in order for the algorithm to identify a code from a reaction network, it needs to be told which of the elements constitute the signs and meanings, and further, the mappings that exist (or the reactions that convert one to the other) would need to be made explicit in order for the algorithm to work. As a result I believe that the algorithmic identification of organic codes is not yet applicable.

However, that is not to say that, in the years to come, such an approach (with due refinements and a theoretical overhaul) will not be of use. As

such it may help in the classification of the plethora of codes that, to date, have not been catalogued or identified. Of these, I believe the regulatory and the quorum sensing codes should be a priority. Another, and possibly larger task remains that of investigating the current codes as thoroughly as I have done with the histone code (which is by no means complete as it relies heavily on experimental data in order to expand). Subsequently it would be necessary to separate the true codes from those that have been falsely called thus. Further, to strengthen the theoretic framework of code biology, a description of organic coding in terms of category theoretical mappings would be very useful.

The concepts that I have dealt with here are, I believe, essential to a complete understanding of life on earth.

Bibliography

- [1] Agalioti, T., Chen, G. and Thanos, D. [2002] “Deciphering the transcriptional histone acetylation code for a human gene” *Cell* **111**, 381–392.
- [2] Allfrey, V. G., Faulkner, R. and Mirsky, A. E. [1964] “Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis” *Proc. Natl. Acad. Sci. U.S.A.* **51**, 786.
- [3] Anderson, J. C., Wu, N., Santoro, S. W., Lakshman, V., King, D. S. and Schultz, P. G. [2004] “An expanded genetic code with a functional quadruplet codon” *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7566–7571.
- [4] Artmann, S. [2009] “Basic semiosis as code-based control” *Biosemiotics* **2**, 31–38.
- [5] Barash, Y., Calarco, J., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. and Frey, B. [2010] “Deciphering the splicing code” *Nature* **465**, 53–59.
- [6] Barbieri, M. [1985] *The semantic theory of evolution* Harwood Academic Publishers.
- [7] Barbieri, M. [1998] “The organic codes. The basic mechanism of macroevolution.” *Riv. Biol.* **91**, 481.

- [8] Barbieri, M. [2003] *The organic codes: an introduction to semantic biology* Cambridge University Press, Cambridge.
- [9] Barbieri, M. [2005] “Life is “artifact-making”” *J. Biosemiotics* **1**, 81–101.
- [10] Barbieri, M. [2006] “Life and semiosis: The real nature of information and meaning” *Semiotica* **2006**, 233–254.
- [11] Barbieri, M. [2008] “Biosemiotics: a new understanding of life” *Naturwissenschaften* **95**, 577–599.
- [12] Barbieri, M. [2009] “A short history of biosemiotics” *Biosemiotics* **2**, 221–245.
- [13] Barbieri, M. [2012] “Code biology—a new science of life” *Biosemiotics* pp. 1–27.
- [14] Barbieri, M. [2012] “Codepoiesis—the deep logic of life” *Biosemiotics* pp. 1–3.
- [15] Barbieri, M. [2012] “The paradigms of biology” *Biosemiotics* pp. 1–27.
- [16] Barbieri, M., de Beule, J. and Hofmeyr, J.-H. S. [2012] “Code biology: A glossary of terms and concepts” .
URL: <http://www.codebiology.org/glossary.html>
- [17] Bártoová, E., Krejčí, J., Harničarová, A., Galiová, G. and Kozubek, S. [2008] “Histone modifications and nuclear architecture: A review” *J. Histochem. Cytochem.* **56**, 711–721.

- [18] Basañez, G. and Hardwick, J. M. [2008] “Unravelling the bcl-2 apoptosis code with a simple model system” *PLoS Biol.* **6**, e154.
- [19] Battail, G. [2009] “Applying semiotics and information theory to biology: A critical comparison” *Biosemiotics* **2**, 303–320.
- [20] Battail, G. [2012] “Biology needs information theory” *Biosemiotics* pp. 1–27.
- [21] Bell, O., Wirbelauer, C., Hild, M., Scharf, A. N., Schwaiger, M., MacAlpine, D. M., Zilbermann, F., van Leeuwen, F., Bell, S. P., Imhof, A., Garza, D., Peters, A. H. F. M. and Schübeler, D. [2007] “Localized h3k36 methylation states define histone h4k16 acetylation during transcriptional elongation in drosophila” *EMBO J.* **26**, 4974–4984.
- [22] Benesch, R. and Benesch, R. E. [1967] “The effect of organic phosphates from the human erythrocyte on the allosteric properties of hemoglobin” *Biochem. Biophys. Res. Commun.* **26**, 192–167.
- [23] Benner, S. A. [1994] “Expanding the genetic lexicon: incorporating non-standard amino acids into proteins by ribosome-based synthesis” *Trends Biol.* **12**, 158–163.
- [24] Berger, S. [2002] “Histone modifications in transcriptional regulation” *Curr. Opin. Genet. Dev.* **12**, 142–148.
- [25] Botuyan, M. V., Lee, J., Ward, I. M., Kim, J.-E., Thompson, J. R., Chen, J. and Mer, G. [2006] “Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair” *Cell* **127**, 1361–1373.

- [26] Brenner, C. and Fuks, F. [2007] “A methylation rendezvous: reader meets writers” *Dev. Cell* **12**, 843–844.
- [27] Brier, S. and Joslyn, C. [2012] “What does it take to produce interpretation? Informational, Peircean and Code-Semiotic Views on Biosemiotics” *Biosemiotics* pp. 1–17.
- [28] Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y., Allis, C. D. *et al.* [1996] “Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation” *Cell* **84**, 843–852.
- [29] Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K. and Lee, K. K., Shia, W.-J., Anderson, S., Yates, J., Washburn, M. P. and Workman, J. L. [2005] “Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription” *Cell* **123**, 581–592.
- [30] Cavalli, G. and Paro, R. [1998] “Chromo-domain proteins: linking chromatin structure to epigenetic regulation” *Curr. Opin. Cell Biol.* **10**, 354–360.
- [31] Chen, C.-H., Chang, W.-F., Liu, C.-C., Su, H.-Y., Shyue, S.-K., Cheng, W. T., Chen, Y. E., Wu, S.-C., Du, F., Sung, L.-Y. *et al.* [2012] “Spatial and temporal distribution of Oct-4 and acetylated H4K5 in rabbit embryos” *Reprod. Biomed. Online* **24**, 433–442.
- [32] Chen, L.-T., Luo, M., Wang, Y.-Y. and Wu, K. [2010] “Involvement of *Arabidopsis* histone deacetylase HDA6 in ABA and salt stress response” *J. Exp. Bot.* **61**, 3345–3353.

-
- [33] Chen, X., Wang, J., Woltring, D., Gerondakis, S. and Shannon, M. F. [2005] “Histone dynamics on the interleukin-2 gene in response to T-cell activation” *Mol. Cell. Biol.* **25**, 3209–3219.
- [34] Chignola, F., Hetenyi, C., Gaetani, M., Rebane, A., Liiv, I., Maran, U., Mollica, L., Bottomley, M. J., Musco, G. and Peterson, P. [2008] “The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression” *EMBO Rep.* **9**, 370–376.
- [35] Crick, F., Barnett, L., Brenner, S. and Watts-Tobin, R. [1961] “General nature of the genetic code for proteins” *Nature* **192**, 1227–1232.
- [36] Cui, K., Zang, C., Roh, T.-Y., Schones, D. E., Childs, R. W., Peng, W. and Zhao, K. [2009] “Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation” *Cell Stem Cell* **4**, 80–93.
- [37] de la Cruz, X., Lois, S., Sánchez-Molina, S. and Martínez-Balbás, M. [2005] “Do protein motifs read the histone code?” *Bioessays* **27**, 164–175.
- [38] de Napoles, M., Mermoud, J. E., Wakao, R., Tang, Y. A., Endoh, M., Appanah, R., Nesterova, T. B., Silva, J., Otte, A. P., Vidal, M., Koseki, H. and Brockdorff, N. [2004] “Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation” *Dev. Cell* **7**, 663–676.
- [39] De Santa, F., Totaro, M. G., Prosperini, E., Notarbartolo, S., Testa, G. and Natoli, G. [2007] “The histone H3 lysine-27 demethylase Jmjd3 links inflammation to inhibition of polycomb-mediated gene silencing” *Cell* **130**, 1083–1094.

- [40] Denisov, I. G. and Sligar, S. G. [2012] “A novel type of allosteric regulation: Functional cooperativity in monomeric proteins” *Arch. Biochem. Biophys.* **519**, 91–102.
- [41] Dey, A., Chitsaz, F., Abbasi, A., Misteli, T. and Ozato, K. [2003] “The double bromodomain protein Brd4 binds to acetylated chromatin during interphase and mitosis” *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8758–8763.
- [42] Dhir, A., Buratti, E., van Santen, M. A., Lührmann, R. and Baralle, F. E. [2010] “The intronic splicing code: multiple factors involved in ATM pseudoexon definition” *EMBO J.* **29**, 749–760.
- [43] Dover, J., Schneider, J., Tawiah-Boateng, M. A., Wood, A., Dean, K., Johnston, M. and Shilatifard, A. [2002] “Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6” *J. Biol. Chem.* **277**, 28368–28371.
- [44] Elder, D. [1979] “An epigenetic code” *Differentiation* **14**, 119–122.
- [45] Fischer, J., Toedling, J., Krueger, T., Schueler, M., Huber, W. and Sperling, S. [2008] “Combinatorial effects of four histone modifications in transcription and differentiation” *Genomics* **91**, 41–51.
- [46] Fischle, W., Wang, Y., Allis, C. D. *et al.* [2003] “Histone and chromatin cross-talk” *Curr. Opin. Cell Biol.* **15**, 172–183.
- [47] Fischle, W., Wang, Y., Jacobs, S. A., Kim, Y., Allis, C. D. and Khorasanizadeh, S. [2003] “Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains” *Gene. Dev.* **17**, 1870–1881.

- [48] Flanagan, J. F., Mi, L.-Z., Chruszcz, M., Cymborowski, M., Clines, K. L., Kim, Y., Minor, W., Rastinejad, F. and Khorasanizadeh, S. [2005] “Double chromodomains cooperate to recognize the methylated histone H tail” *Nature* **438**, 1181–1185.
- [49] Fu, X. [2004] “Towards a splicing code” *Cell* **119**, 736–738.
- [50] Füllgrabe, J., Hajji, N. and Joseph, B. [2010] “Cracking the death code: Apoptosis-related histone modifications” *Cell Death Diff.* **17**, 1238–1243.
- [51] Gabius, H. [2000] “Biological information transfer beyond the genetic code: the sugar code” *Naturwissenschaften* **87**, 108–121.
- [52] Gabius, H., André, S., Kaltner, H. and Siebert, H. [2002] “The sugar code: functional lectinomics” *Biochim. Biophys. Acta* **1572**, 165–177.
- [53] Gabius, H., Siebert, H., André, S., Jiménez-Barbero, J. and Rüdiger, H. [2004] “Chemical biology of the sugar code” *ChemBiochem* **5**, 740–764.
- [54] Gabius, H.-J., André, S., Jiménez-Barbero, J., Romero, A. and Solís, D. [2011] “From lectin structure to functional glycomics: principles of the sugar code” *Trends Biochem. Sci.* **36**, 298–313.
- [55] Gatta, R., Dolfini, D., Zambelli, F., Imbriano, C., Pavesi, G. and Mantovani, R. [2011] “An acetylation-monoubiquitination switch on lysine 120 of H2B” *Epigenetics* **6**, 630–637.
- [56] Gelbart, M. E., Larschan, E., Peng, S., Park, P. J. and Kuroda, M. I. [2009] “Drosophila MSL complex globally acetylates H4K16 on the

- male X chromosome for dosage compensation” *Nat. Struct. Mol. Biol.* **16**, 825–832.
- [57] Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A. K., Kelley, D. R., Shishkin, A. A., Issner, R., Zhang, X., Coyne, M., Fostel, J. L., Holmes, L., Meldrim, J., Guttman, M., Epstein, C., Park, H., Kohlbacher, O., Rinn, J., Gnirke, A., Lander, E. S., Bernstein, B. E. and Meisner, A. [2013] “Transcriptional and epigenetic dynamics during specification of human embryonic stem cells” *Cell* .
- [58] Gimona, M. [2008] “Protein linguistics and the modular code of the cytoskeleton” in *The Codes of Life* (Barbieri, M., ed.) pp. 189–206 Springer, Berlin.
- [59] Görlich, D. and Dittrich, P. [2013] “Molecular codes in biological and chemical reaction networks” *PloS One* **8**, e54694.
- [60] Grunstein, M. [1997] “Histone acetylation in chromatin structure and transcription” *Nature* **389**, 349–352.
- [61] Gunasekaran, K., Ma, B. and Nussinov, R. [2004] “Is allostery an intrinsic property of all dynamic proteins?” *Proteins: Struct., Funct., Bioinf.* **57**, 433–443.
- [62] Hassan, A. H., Prochasson, P., Neely, K. E., Galasinski, S. C., Chandy, M., Carrozza, M. J. and Workman, J. L. [2002] “Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes” *Cell* **111**, 369–379.

- [63] Henikoff, S. [2005] “Histone modifications: Combinatorial complexity or cumulative simplicity?” *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5308–5309.
- [64] Hershko, A. and Ciechanover, A. [1998] “The ubiquitin system” *Annu. Rev. Biochem.* **67**, 425–479.
- [65] Hicke, L. [2001] “Protein regulation by monoubiquitin” *Nat. Rev. Mol. Cell. Biol.* **2**, 195–201.
- [66] Hohsaka, T., Ashizuka, Y., Murakami, H. and Sisido, M. [2001] “Five-base codons for incorporation of nonnatural amino acids into proteins” *Nucleic Acids Res.* **29**, 3646–3651.
- [67] Hook, S. S., Orian, A., Cowley, S. M. and Eisenman, R. N. [2002] “Histone deacetylase 6 binds polyubiquitin through its zinc finger (PAZ domain) and copurifies with deubiquitinating enzymes” *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13425–13430.
- [68] Hunt, P., Gulisano, M., Cook, M., Sham, M.-H., Faiella, A., Wilkinson, D., Boncinelli, E. and Krumlauf, R. [1991] “A distinct *Hox* code for the branchial region of the vertebrate head” *Nature* **353**, 861–864.
- [69] Hunt, P., Whiting, J., Nonchev, S., Sham, M.-H., Marshall, H., Graham, A., Cook, M., Alleman, R., Rigby, P. W. and Gulisano, M. [1991] “The branchial *Hox* code and its implications for gene regulation, patterning of the nervous system and head evolution” *Development* **2**, 63–77.
- [70] Hunter, T. [2007] “The age of crosstalk: phosphorylation, ubiquitination, and beyond” *Mol. Cell* **28**, 730–738.

- [71] Hurley, J., Lee, S. and Prag, G. [2006] “Ubiquitin-binding domains” *Biochem. J* **399**, 361–372.
- [72] Husnjak, K. and Dikic, I. [2012] “Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions” *Annu. Rev. Biochem.* **81**, 291–322.
- [73] Jacobson, R. H., Ladurner, A. G., King, D. S. and Tjian, R. [2000] “Structure and function of a human TAFII250 double bromodomain module” *Science* **288**, 1422–1425.
- [74] Jason, L. J., Moore, S. C., Lewis, J. D., Lindsey, G. and Ausió, J. [2002] “Histone ubiquitination: A tagging tail unfolds?” *Bioessays* **24**, 166–174.
- [75] Jeltsch, A. and Rathert, P. [2008] “Putting the pieces together: histone H2B ubiquitylation directly stimulates histone H3K79 methylation” *ChemBiochem* **9**, 2193–2195.
- [76] Jenuwein, T. and Allis, C. [2001] “Translating the histone code” *Sci. STKE* **293**, 1074.
- [77] Jukes, T. H. [1985] “A change in the genetic code in *Mycoplasma capricolum*” *J. Mol. Evol.* **22**, 361–362.
- [78] Jukes, T. H. and Osawa, S. [1993] “Evolutionary changes in the genetic code” *Comp. Biochem. Physiol. B* **106**, 489–494.
- [79] Kanno, T., Kanno, Y., Siegel, R. M., Jang, M. K., Lenardo, M. J. and Ozato, K. [2004] “Selective recognition of acetylated histones by bromodomain proteins visualized in living cells” *Mol. Cell* **13**, 33–43.

- [80] Khan, A. U. and Hampsey, M. [2002] “Connecting the DOTs: covalent histone modifications and the formation of silent chromatin” *Trends Genet.* **18**, 387–389.
- [81] Kim, J., Daniel, J., Espejo, A., Lake, A., Krishna, M., Xia, L., Zhang, Y. and Bedford, M. T. [2006] “Tudor, MBT and chromo domains gauge the degree of lysine methylation” *EMBO Rep.* **7**, 397–403.
- [82] Koche, R. P., Smith, Z. D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B. E. and Meissner, A. [2011] “Reprogramming factor expression initiates widespread targeted chromatin remodeling” *Cell Stem Cell* **8**, 96–105.
- [83] Komander, D. and Rape, M. [2012] “The ubiquitin code” *Annu. Rev. Biochem.* **81**, 203–229.
- [84] Kornberg, R. and Lorch, Y. [1999] “Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome” *Cell* **98**, 285–294.
- [85] Kouzarides, T. [2007] “Chromatin modifications and their function” *Cell* **128**, 693.
- [86] Kurdistani, S. K. and Grunstein, M. [2003] “Histone acetylation and deacetylation in yeast” *Nat. Rev. Mol. Cell Biol.* **4**, 276–284.
- [87] Kurdistani, S. K., Tavazoie, S. and Grunstein, M. [2004] “Mapping global histone acetylation patterns to gene expression” *Cell* **117**, 721–733.

- [88] Laskowski, R. A., Gerick, F. and Thornton, J. M. [2009] “The structural basis of allosteric regulation in proteins” *FEBS Lett.* **583**, 1692–1698.
- [89] Latham, J. A., Chosed, R. J., Wang, S. and Dent, S. Y. [2011] “Chromatin signaling to kinetochores: transregulation of Dam1 methylation by histone H2B ubiquitination” *Cell* **146**, 709–719.
- [90] Latham, J. A. and Dent, S. Y. R. [2007] “Cross-regulation of histone modifications” *Nat. Struct. Mol. Biol.* **14**, 1017–1024.
- [91] Lee, J., Natarajan, M., Nahine, V. C., Socolich, M., Vo, T., Russ, W. P., Benkovic, S. J. and Ranganathan, R. [2008] “Surface sites for engineering allosteric control in proteins” *Science* **322**, 438–442.
- [92] Lee, J.-S. and Shilatifard, A. [2007] “A site to remember: H3K36 methylation a mark for histone deacetylation” *Mutation Res. Fundamental Mol. Mech. Mutagen.* **618**, 130–134.
- [93] Lee, J.-S., Shukla, A., Schneider, J., Swanson, S. K., Washburn, M. P., Florens, L., Bhaumik, S. R. and Shilatifard, A. [2007] “Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS” *Cell* **131**, 1084–1096.
- [94] Levenson, J. M. and Sweatt, J. D. [2005] “Epigenetic mechanisms in memory formation” *Nat. Rev. Neurosci.* **6**, 108–118.
- [95] Liu, C. L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S. L., Friedman, N. and Rando, O. J. [2005] “Single-nucleosome mapping of histone modifications in *S. cerevisiae*” *PLoS Biol.* **3**, e328.

- [96] Lo, W., Henry, K., Schwartz, M. and Berger, S. [2003] “Histone modification patterns during gene activation” *Methods Enzymol.* **377**, 130–153.
- [97] Maraldi, N. M. [2008] “A lipid-based code in nuclear signalling” in *The Codes of Life* (Barbieri, M., ed.) pp. 207–221 Springer.
- [98] Margueron, R. and Reinberg, D. [2011] “The Polycomb complex PRX2 and its mark in life” *Nature* **469**, 343–349.
- [99] Marshall, H., Nonchev, S., Sham, M.-H., Muchamore, I., Lumsden, A. and Krumlauf, R. [1992] “Retinoic acid alters hindbrain *Hox* code and induces transformation of rhombomeres 2/3 into 3/4 identity” *Nature* **360**, 737–741.
- [100] Martin, C. and Zhang, Y. [2005] “The diverse functions of histone lysine methylation” *Nat. Rev. Mol. Cell Biol.* **6**, 838–849.
- [101] Matlin, A., Clark, F. and Smith, C. [2005] “Understanding alternative splicing: towards a cellular code” *Nat. Rev. Mol. Cell Biol.* **6**, 386–398.
- [102] Maurer-Stroh, S., Dickens, N. J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F. and Ponting, C. P. [2003] “The tudor domain ‘Royal family’: tudor, plant agenet, chromo, PWWP and MBT domains” *Trends Biochem. Sci.* **28**, 69–74.
- [103] May, L. T., Leach, K., Sexton, P. M. and Christopoulos, A. [2007] “Allosteric modulation of G protein-coupled receptors” *Annu. Rev. Pharmacol. Toxicol.* **47**, 1–51.

- [104] McDaniel, I. E., Lee, J. M., Berger, M. S., Hanagami, C. K. and Armstrong, J. A. [2008] “Investigations of CHD1 function in transcription and development of *Drosophila melanogaster*” *Genetics* **178**, 583–587.
- [105] McGinty, R. K., Kim, J., Chatterjee, C., Roeder, R. G. and Muir, T. W. [2008] “Chemically ubiquitylated histone H2B stimulates hDot1L-mediated intranucleosomal methylation” *Nature* **453**, 812–816.
- [106] McKittrick, E., Gafken, P. R., Ahmad, K. and Henikoff, S. [2004] “Histone H3. 3 is enriched in covalent modifications associated with active chromatin” *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1525–1530.
- [107] Mellor, J. [2006] “It takes a PHD to read the histone code” *Cell* **126**, 22–24.
- [108] Mellor, J. [2009] “Linking the cell cycle to histone modifications: Dot1, G1/S, and cycling K79me²” *Mol. Cell* **35**, 729–730.
- [109] Mujtaba, S., He, Y., Zeng, L., Farooq, A., Carlson, J. E., Ott, M., Verdin, E. and Zhou, M.-M. [2002] “Structural basis of lysine-acetylated HIV-1 Tat recognition by PCAF bromodomain” *Mol. Cell* **9**, 575–586.
- [110] Mujtaba, S., Zeng, L. and Zhou, M. [2007] “Structure and acetyl-lysine recognition of the bromodomain” *Oncogene* **26**, 5521–5527.
- [111] Murphy, P., André, S. and Gabius, H.-J. [2013] “The third dimension of reading the sugar code by lectins: design of glycoclusters with cyclic scaffolds as tools with the aim to define correlations between spatial presentation and activity” *Molecules* **18**, 4026–4053.

- [112] Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. and Chin, J. W. [2010] “Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome” *Nature* **464**, 441–444.
- [113] Nguyen, A. T. and Zhang, Y. [2011] “The diverse functions of Dot1 and H3K79 methylation” *Gene. Dev.* **25**, 1345–1358.
- [114] Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. and O’neal, C. [1965] “RNA codewords and protein synthesis, VII. on the general nature of the RNA code” *Proc. Natl. Acad. Sci. U.S.A.* **53**, 1161.
- [115] O’Brien, T. and Tjian, R. [1998] “Functional analysis of the human $TA_{i,i}$ 250 N-terminal kinase domain” *Mol. Cell* **1**, 905–911.
- [116] Org, T., Chignola, F., Hetenyi, C., Gaetani, M., Rebane, A., Liiv, I., Maran, U., Mollica, L., Bottomley, M. J., Musco, G. and Peterson, P. [2008] “The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression” *EMBO Rep.* **9**, 370–376.
- [117] Osawa, S., Jukes, T. H., Watanabe, K. and Muto, A. [1992] “Recent evidence for evolution of the genetic code.” *Microbiol Rev* **56**, 229–264.
- [118] Osley, M. A. [2004] “H2B ubiquitylation: the end is in sight” *Biochim. Biophys. Acta* **1677**, 74–78.
- [119] Owen, D. J., Ornaghi, P., Yang, J.-C., Lowe, N., Evans, P. R., Ballario, P., Neuhaus, D., Filetici, P. and Travers, A. A. [2000] “The structural basis for the recognition of acetylated histone H4

- by the bromodomain of histone acetyltransferase Gcn5p” *EMBO J.* **19**, 6141–6149.
- [120] Pattee, H. H. [2012] “Epistemic, evolutionary, and physical conditions for biological information” *Biosemiotics* pp. 1–23.
- [121] Pattee, H. H. H. [1967] “The physical basis of coding and reliability in biological evolution” in *Toward a theoretical biology* (Waddington, C. H., ed.) vol. 1 Edinburgh University Press, Edinburgh.
- [122] Pereira, S. L., Grayling, R. A., Lurz, R. and Reeve, J. N. [1997] “Archaeal nucleosomes” *Proc. Natl. Acad. Sci. U.S.A.* **94**, 12633–12637.
- [123] Pertea, M., Mount, S. M. and Salzber, S. L. [2007] “A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*” *BMC Bioinformatics* **8**.
- [124] Peterson, C. L., Laniel, M.-A. *et al.* [2004] “Histones and histone modifications” *Curr. Biol.* **14**, 546–551.
- [125] Pham, A.-D. and Sauer, F. [2000] “Ubiquitin-activating/conjugating activity of TAFII250, a mediator of activation of gene expression in *Drosophila*” *Science* **289**, 2357–2360.
- [126] Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E. *et al.* [2005] “Genome-wide map of nucleosome acetylation and methylation in yeast” *Cell* **122**, 517–527.

- [127] Pray-Grant, M. G., Daniel, J. A., Schieltz, D., Yates, J. R. and Grant, P. A. [2005] “Chd1 chromodomain links histone H3 methylation with SAGA-and SLIK-dependent acetylation” *Nature* **433**, 434–438.
- [128] Ramakrishnan, V. [1997] “Histone structure and the organization of the nucleosome” *Annu. Rev. Biophys. Biomol. Struct.* **26**, 83–112.
- [129] Rando, O. [2012] “Combinatorial complexity in chromatin structure and function: revisiting the histone code” *Curr. Opin. Genet. Dev.* **22**, 148–155.
- [130] Redies, C. and Takeichi, M. [1996] “Cadherins in the developing central nervous system: An adhesive code for segmental and functional subdivisions” *Dev. Biol.* **180**, 413–423.
- [131] Rice, J. C. and Allis, C. D. [2001] “Histone methylation versus histone acetylation: New insights into epigenetic regulation” *Curr. Opin. Cell Biol.* **13**, 263–273.
- [132] Roest, H. P., Baarends, W. M., de Wit, J., van Klaveren, J. W., Wassenaar, E., Hoogerbrugge, J. W., van Cappellen, W. A., Hoesjmakers, J. H. and Grootegoed, J. A. [2004] “The ubiquitin-conjugating DNA repair enzyme HR6A is a maternal factor essential for early embryonic development in mice” *Mol. Cell. Biol.* **24**, 5485–5495.
- [133] Rundlett, S. E., Carmen, A. A., Suka, N., Turner, B. M. and Grunstein, M. [1998] “Transcriptional repression by UME6 involves deacetylation of lysine 5 of histone H4 by RPD3” *Nature* **392**, 831–835.

- [134] Ryan, J. F., Mazza, M. E., Pang, K., Matus, D. Q., Baxevanis, A. D., Martindale, M. Q. and Finnerty, J. R. [2007] “Pre-bilaterian origins of the *Hox* cluster and the *Hox* code: evidence from the sea anemone *Nematostella vectensis*” *PLoS One* **2**, e153.
- [135] Schlesinger, D. H., Goldstein, G. and Niall, H. D. [1975] “Complete amino acid sequence of ubiquitin, an adenylate cyclase stimulating polypeptide probably universal in living cells” *Biochemistry (Mosc.)* **14**, 2214–2218.
- [136] Schmucker, D. and Flanagan, J. G. [2004] “Generation of recognition diversity in the nervous system” *Neuron* **44**, 219–222.
- [137] Schneider, R., Bannister, A. J., Myers, F. A., Thorne, A. W., Crane-Robinson, C. and Kouzarides, T. [2003] “Histone H3 lysine 4 methylation patterns in higher eukaryotic genes” *Nat. Cell Biol.* **6**, 73–77.
- [138] Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D. and Jenuwein, T. [2004] “A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin” *Gene. Dev.* **18**, 1251–1262.
- [139] Schrödinger, E. [1944] *What is life? The physical aspect of the living cell* Cambridge University Press, Cambridge.
- [140] Schwartz, T. W. and Holst, B. [2007] “Allosteric enhancers, allosteric agonists and ago-allosteric modulators: where do they bind and how do they act?” *Trends Pharmacol. Sci.* **28**, 366–373.
- [141] Schwartz, Y. B. and Pirrotta, V. [2008] “Polycomb complexes and epigenetic states” *Curr. Opin. Cell Biol.* **20**, 266 – 273.

- [142] Shahbazian, M. D. and Grunstein, M. [2007] “Functions of site-specific histone acetylation and deacetylation” *Annu. Rev. Biochem.* **76**, 75–100.
- [143] Shahbazian, M. D., Zhang, K. and Grunstein, M. [2005] “Histone H2 ubiquitylation controls processive methylation but not monomethylation by Dot1 and Set1” *Mol. Cell* **19**, 271–277.
- [144] Shannon, C. E. [1948] “A mathematical theory of communication” *ACM SIGMOBILE Mob. Comp. Comm. Rev.* **5**, 3–55.
- [145] Shapiro, J. [2009] “Revisiting the central dogma in the 21st century” *Ann. N. Y. Acad. Sci.* **1178**, 6–28.
- [146] Shi, L., Ai, J., Ouyang, Y., Huang, J., Lei, Z., Wang, Q., Yin, S., Han, Z., Sun, Q. and Chen, D. [2008] “Trichostatin A and nuclear reprogramming of cloned rabbit embryos” *J. Anim. Sci.* **86**, 1106–1113.
- [147] Shi, X., Hong, T., Walter, K. L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Pena, P., Lan, F., Kaadige, M. R. *et al.* [2006] “ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression” *Nature* **442**, 96–99.
- [148] Shilatifard, A. [2008] “Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation” *Curr. Opin. Cell Biol.* **20**, 341–348.
- [149] Simic, R., Lindstrom, D. L., Tran, H. G., Roinick, K. L., Costa, P. J., Johnson, A. D., Hartzog, G. A. and Arndt, K. M. [2003] “Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes” *EMBO J.* **22**, 1846–1856.

- [150] Simon, J. A. and Kingston, R. E. [2009] “Mechanisms of polycomb gene silencing: knowns and unknowns” *Nat. Rev. Mol. Cell Biol.* **10**, 697–708.
- [151] Smallwood, A., Estève, P.-O., Pradhan, S. and Carey, M. [2007] “Functional cooperation between HP1 and DNMT1 mediates gene silencing” *Gene. Dev.* **21**, 1169–1178.
- [152] Snowden, A., Gregory, P., Case, C. and Pabo, C. [2002] “Gene-specific targeting of H3K9 methylation is sufficient for initiating repression in vivo” *Curr. Biol.* **12**, 2159–2166.
- [153] Söll, D., Ohtsuka, E., Jones, D., Lohrmann, R., Hayatsu, H., Nishimura, S. and Khorana, H. [1965] “Studies on polynucleotides, XLIX. stimulation of the binding of aminoacyl-sRNA’s to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids.” *Proc. Natl. Acad. Sci. U.S.A.* **54**, 1378.
- [154] Steger, D. J., Lefterova, M. I., Ying, L., Stonestrom, A. J., Schupp, M., Zhuo, D., Vakoc, A. L., Kim, J.-E., Chen, J., Lazar, M. A., Blobel, G. A. and Vakoc, C. R. [2008] “DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells” *Mol. Cell. Biol.* **28**, 2825–2839.
- [155] Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E. M., Akey, J. M. and Stamatoyannopoulos, J. A. [2013] “Exonic transcription factor binding directs codon choice and affects protein evolution” *Science* **342**, 1367–1372.

- [156] Strahl, B. and Allis, C. [2000] “The language of covalent histone modifications” *Nature* **403**, 41.
- [157] Strašák, L., Bártová, E., Harničarová, A., Galiová, G., Krejčí, J. and Kozubek, S. [2009] “H3K9 acetylation and radial chromatin positioning” *J. Cell. Physiol.* **220**, 91–101.
- [158] Sun, Z.-W. and Allis, C. D. [2002] “Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast” *Nature* **418**, 104–108.
- [159] Swigut, T. and Wysocka, J. [2007] “H3K27 demethylases, at long last” *Cell* **131**, 29–32.
- [160] Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H., Wutz, A., Vidal, M., Elderkin, S. and Brockdorff, N. [2012] “RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3” *Cell* .
- [161] Tomkins, G. [1975] “The metabolic code” *Science* **189**, 760–763.
- [162] Tóth, K. F., Knoch, T. A., Wachsmuth, M., Frank-Stöhr, M., Stöhr, M., Bacher, C. P., Müller, G. and Rippe, K. [2004] “Trichostatin A-induced histone acetylation causes decondensation of interphase chromatin” *J. Cell Sci.* **117**, 4277–4287.
- [163] Trifonov, E. [1989] “The multiple codes of nucleotide sequences” *Bull. Math. Biol.* **51**, 417–432.

- [164] Tsai, C.-J., Del Sol, A. and Nussinov, R. [2008] “Allostery: absence of a change in shape does not imply that allostery is not at play” *J. Mol. Biol.* **378**, 1–11.
- [165] Tsai, W.-W., Wang, Z., Yiu, T. T., Akdemir, K. C., Xia, W., Winter, S., Tsai, C.-Y., Shi, X., Schwarzer, D., Plunkett, W., Aronow, B., Or, G., Fischle, W., Hung, M.-C., Patel, D. J. and Barton, M. C. [2010] “TRIM24 links a non-canonical histone signature to breast cancer” *Nature* **468**, 927–932.
- [166] Turner, B. [2000] “Histone acetylation and an epigenetic code” *Bioessays* **22**, 836–845.
- [167] Turner, B. [2002] “Cellular memory and the histone code” *Cell* **111**, 285–291.
- [168] Vakoc, C. R., Mandat, S. A., Olenchock, B. A. and Blobel, G. A. [2005] “Histone H3 lysine 9 methylation and HP1 γ are associated with transcription elongation through mammalian chromatin” *Mol. Cell* **19**, 381–391.
- [169] Verhey, K. J. and Gaertig, J. [2007] “The tubulin code” *Cell Cycle* **6**, 2152–2160.
- [170] von Neumann, J. [1966] *Theory of self-reproducing automata* (edited by Arthur W. Burks) University of Illinois Press, Urbana, Illinois.
- [171] Wang, G. G., Cai, L., Pasillas, M. P. and Kamps, M. P. [2007] “NUP98–NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis” *Nat. Cell Biol.* **9**, 804–812.

- [172] Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R. S. and Zhang, Y. [2004] “Role of histone H2A ubiquitination in Polycomb silencing” *Nature* **431**, 873–878.
- [173] Wang, K., Schmied, W. and Chin, J. [2012] “Reprogramming the genetic code: From triplet to quadruplet codes” *Angew. Chem., Int. Ed.* **51**, 2288–2297.
- [174] Wang, Z. and Burge, C. [2008] “Splicing regulation: from a parts list of regulatory elements to an integrated splicing code” *RNA* **14**, 802–813.
- [175] Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cudapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q. *et al.* [2008] “Combinatorial patterns of histone acetylations and methylations in the human genome” *Nat. Genet.* **40**, 897–903.
- [176] Wassarman, D. A., Aoyagi, N., Pile, L. A. and Schlag, E. M. [2000] “TAF250 is required for multiple developmental events in *Drosophila*” *Proc. Natl. Acad. Sci. U.S.A.* **97**, 1154–1159.
- [177] Wickliffe, K., Williamson, A., Jin, L. and Rape, M. [2009] “The multiple layers of ubiquitin-dependent cell cycle control” *Chem. Rev.* **109**, 1537–1548.
- [178] Wu, J. and Grunstein, M. [2000] “25 years after the nucleosome model: Chromatin modifications” *Trends Biochem. Sci.* **25**, 619 – 623.
- [179] Yang, J.-S., Seo, S. W., Jang, S., Jung, G. Y. and Kim, S. [2012] “Rational engineering of enzyme allosteric regulation through sequence evolution analysis” *PLoS Comput. Biol.* **8**, e1002612.

- [180] Yang, X.-J. and Seto, E. [2008] “Lysine acetylation: codified crosstalk with other posttranslational modifications” *Mol. Cell* **31**, 449–461.
- [181] Yuan, W., Xu, M., Huang, C., Liu, N., Chen, S. and Zhu, B. [2011] “H3K36 methylation antagonizes PRC2-mediated H3K27 methylation” *J. Biol. Chem.* **286**, 7983–7989.
- [182] Zhang, Y. [2003] “Transcriptional regulation by histone ubiquitination and deubiquitination” *Gene. Dev.* **17**, 2733–2740.
- [183] Zhu, B., Zheng, Y., Pham, A.-D., Mandal, S. S., Erdjument-Bromage, H., Tempst, P. and Reinberg, D. [2005] “Monoubiquitination of human histone H2B: The factors involved and their roles in *HOXs* gene regulation” *Mol. Cell* **20**, 601–611.