1

History for Canonical and Non-canonical Structures of Nucleic Acids

The main points of the learning:

Understand canonical and non-canonical structures of nucleic acids and think of historical scientists in the research field of nucleic acids.

1.1 Introduction

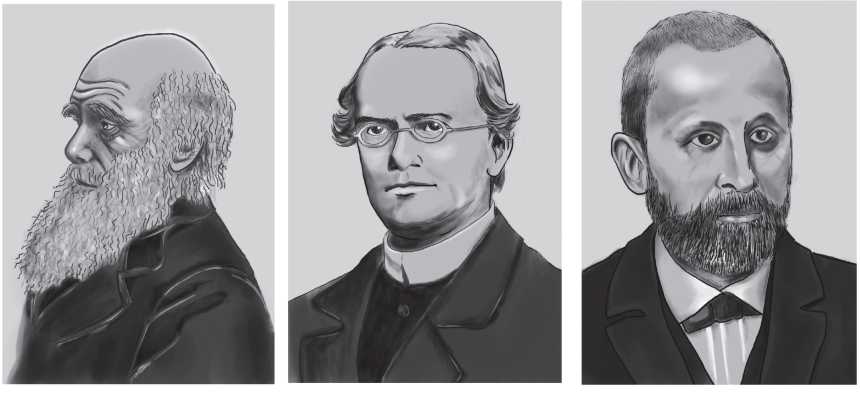
This book is to interpret the non-canonical structures and their stabilities of nucleic acids from the viewpoint of the chemistry and study their biological significances. There is more than 60 years' history after the discovery of the double helix DNA structure by James Dewey Watson and Francis Harry Compton Crick in 1953, and chemical biology of nucleic acids is facing a new aspect today. Through this book, I expect that readers understand how the uncommon structure of nucleic acids became one of the common structures that fascinate us now. In this chapter, I introduce the history of nucleic acid structures and the perspective of research for non-canonical nucleic acid structures (see also Chapter 15).

1.2 History of Duplex

The opening of the history of genetics was mainly done by three researchers. Charles Robert Darwin, who was a scientist of natural science, pioneered genetics. The proposition of genetic concept is indicated in his book *On the Origin of Species* published in 1859. He indicated the theory of biological evolution, which is the basic scientific hypothesis of natural diversity. In other words, he proposed biological evolution, which changed among individuals by adapting to the environment and be passed on to the next generation. However, that was still a primitive idea for the genetic concept. After that, Gregor Johann Mendel, who was a priest in Brno, Czech Republic, confirmed the mechanism of gene evolution by using "factor" inherited from parent to children using pea plant in 1865. This discovery became the concept of genetics. At the almost same time in 1869 as Mendel, Johannes

Chemistry and Biology of Non-Canonical Nucleic Acids, First Edition. Naoki Sugimoto.

Friedrich Miescher, who was a biochemist in Swiss, discovered nucleic acids as a chemical substance of the gene identity. He named it "nuclein" (later, it was named "nucleic acid," which exists acidic substance in nucleus) and made the opportunity to study nucleic acid chemistry. However, it would be doubtful if he realized that nucleic acid is the gene identity. After that, it was needed to take a lot of time to conclude that the gene identity is proved the chemical substance.



Charles Robert Darwin (left), Gregor Johann Mendel (middle), and Johannes Friedrich Miescher (right)

Erwin Rudolf Josef Alexander Schrödinger, who was a great physicist, pioneered to go after the mystery of gene. He published a book titled *What Is Life*? in 1944 [1]. This book invited the study of the gene to many researchers. He mentioned in the book that he believed a gene – or perhaps the whole chromosome fiber – to be an aperiodic solid, although he also mentioned that gene is probably one big protein molecule. After the 1950s, chemistry regarding nucleic acids had been developing. One of the organic chemists was Erwin Chargaff, who was a professor at Colombia University in the United States and born in Austria. He discovered that from the result of paper chromatography targeted to the different types of DNA, the number of guanine units equals the number of cytosine units and the number of adenine units equals the number of thymine units [2]. It is called Chargaff's rules. On the other hand, analysis of the superstructure of nucleic acids was also proceeding. At the beginning of the 1950s, at King's College London, the results of X-ray crystal analysis were accumulated by Maurice Hugh Frederick Wilkins, Rosalind Elsie Franklin, and others. Finally, based on their result, Watson and Crick who worked at Cavendish Laboratory in Cambridge and proposed the model of double helix structure of DNA (Figure 1.1 and see Chapter 2), published as a single-page paper about DNA double helix in *Nature* issued on 25 April 1953 [3]. By discovering DNA double helix structure, Watson, Crick, and Wilkins were awarded the Nobel Prize in Physiology or Medicine in 1962.

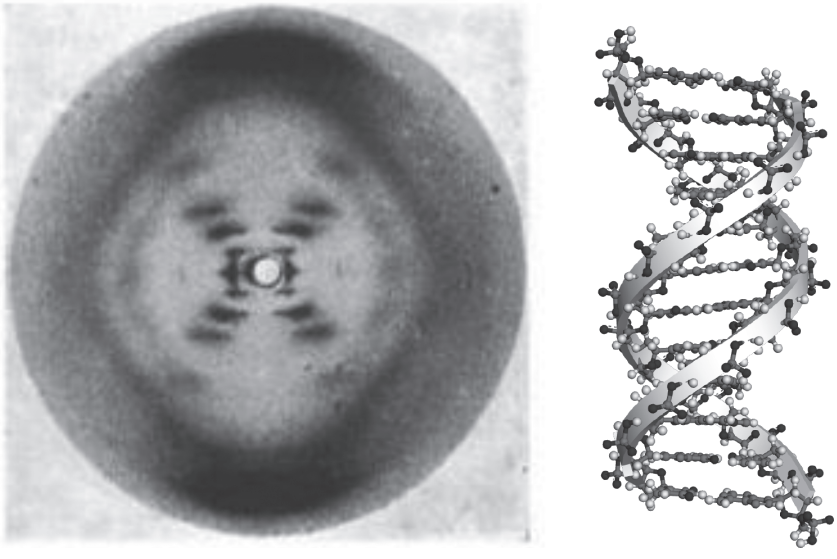
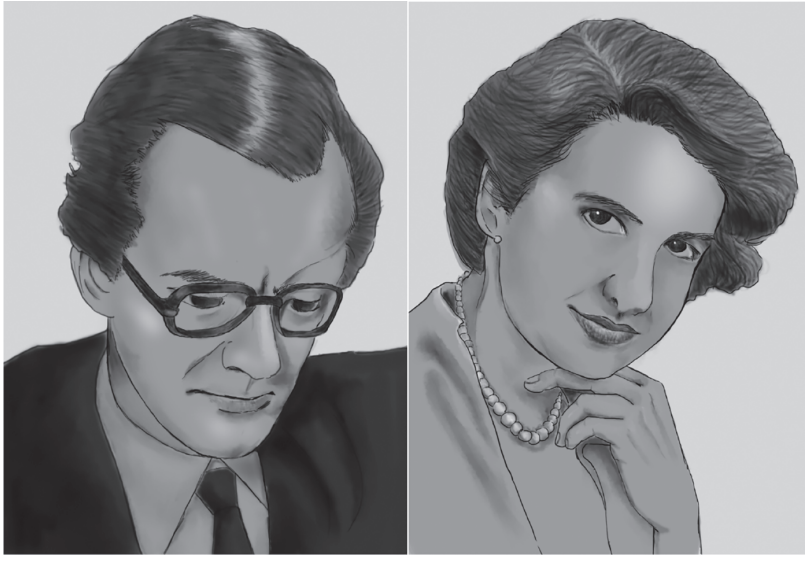
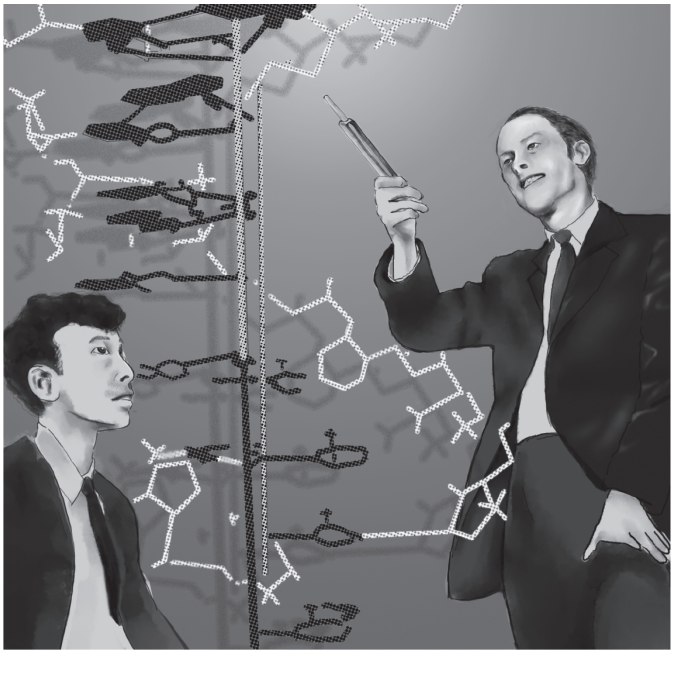


Figure 1.1 The diffraction pattern of the canonical DNA duplex and its chemical structure. Source: Kings College London.

Erwin Rudolf Josef Alexander Schrödinger (left) and Erwin Chargaff (right)



Maurice Hugh Frederick Wilkins (left) and Rosalind Elsie Franklin (right)



James Dewey Watson and Francis Harry Compton Crick

1.3 Non-Watson–Crick Base Pair

Although the discovery of Watson–Crick base pairs is famous, we need to make sure that Watson and Crick initially "proposed" their model. Moreover, Watson and Crick were not the first researchers who proposed the structure of nucleic acids. The physicist Linus Pauling, who earned the Nobel Prize two times in his career, first proposed the helix model of nucleic acids with his associate Robert Corey [4]. However, the structure was fault: it was a triple helix having negatively charged phosphates located at the core of the helix, which could not exist in nature. After the proposal of Watson–Crick base pairs, the race for determination of the helical structure of DNA had been started using purine and pyrimidine monomers. The first such study was reported in 1959, when Karst Hoogsteen – an associate of Robert Corey at Caltech – used single-crystal X-ray analysis to determine the structures of cocrystals containing 9-methyladenine and 1-methylthymine, where methyl groups were used to block hydrogen bonding to nitrogen atoms otherwise bonded to sugar carbons in DNA [5]. However, the structure was NOT Watson–Crick base pair, in which the adenine base was flipped upside down. The different base pair was later named Hoogsteen base pair (Figure 1.2 and see Chapter 2). After the discovery of

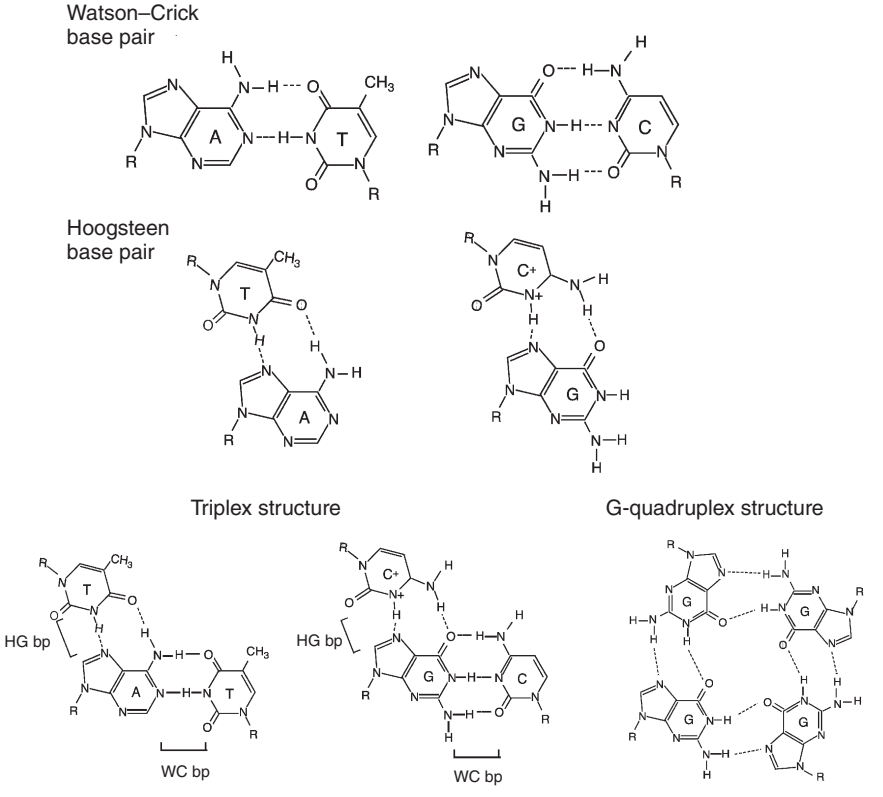


Figure 1.2 Chemical structures of base pairs via Watson–Crick or Hoogsteen types.

Hoogsteen base pairs, many researchers looked for Watson–Crick base pairs. However, only Hoogsteen base pairs were identified. In 1973, Alexander Rich first discovered Watson–Crick base pairs in the cocrystal of the AU and GC dinucleoside phosphate complex [6]. And soon after, Richard E. Dickerson, who took over the Pauling's lab, first solved the single-crystal structure of a DNA dodecamer using heavy atom X-ray crystallography in 1980 [7]. It takes more than 20 years after the discovery of Watson–Crick base pairs. These results suggest that Watson–Crick base pairs tended to stably form under the constraint of the helical structure of nucleic acids, whereas Hoogsteen base pairs form in other structural conditions. Therefore, there are canonical structures composed by Watson–Crick base pairs in the duplex structures. On the other hand, non-canonical structures include non-Watson–Crick base pairs such as Hoogsteen base pairs.

Linus Pauling (left), Robert Corey (middle), and Karst Hoogsteen (right)

Alexander Rich (left) and Richard E. Dickerson (right)

1.4 Nucleic Acid Structures Including Non-Watson–Crick Base Pairs

Behind the extensive efforts to identify the duplex structure of Watson–Crick base pairs, Hoogsteen base pairs were also found in the structure of nucleic acids in the 1960s. Felsenfeld and Rich explained how poly(rU) strands might associate with poly(rA)-poly(rU) duplexes to form triplexes [8]. From the chemical shift of NMR, they identified evidence for triplex formation via protonated G–C^+ Hoogsteen base pairs at cytosine N3 in a poly(dG)-poly(dC) complex with dGMP at low pH [9]. In 1962, it was found that short guanine-rich stretches of DNA could assume unusual structures [10]. The diffraction studies of poly(guanylic acid) gels suggested that if four guanines were close enough together, they could form planar hydrogen-bonded arrangements now called guanine quartets (G-quartets). With a stack of a few G-quartets, a tetraplex structure is formed called as G-quadruplex (see Chapter 2). In the crystal structure, Hoogsteen base pairs of polynucleic acids were first found in tRNA structure [11]. In the structure Watson–Crick base pairs formed the secondary structure of tRNA, whereas Hoogsteen base pairs supported the tertiary structure. Not only Hoogsteen base pairs but also other types of non-Watson–Crick base pairs were found in tRNA structures. The tertiary structure of nucleic acids is important especially for non-coding RNAs, which do not code genetic information. The landmark of research of non-coding RNA is the discovery of ribozyme (ribonucleic acid enzyme) by Thomas Robert Cech in 1982 [12]. Ribozymes catalyze chemical reactions as well as protein enzymes. Later structural studies revealed that there are a lot of non-Watson–Crick base pairs to produce the active core of enzymatic reaction of ribozymes. Therefore, non-canonical Watson–Crick base pairs including Hoogsteen base pairs have been thought of as a tool for the tertiary structure of nucleic acids except for duplexes.

Thomas Robert Cech

With the progress of structural analysis technology in the 1990s, Hoogsteen base pairs are gradually revealed to exist in DNA complexes with low molecular weight compounds and proteins as well as transiently in Watson–Crick-type double helix. Furthermore, another type of tetraplex structures was identified from DNA sequence enriched in cytosine due to the cross intercalations of hemiprotonated cytosine–cytosine ($C–C^+$) base pairs under acidic conditions [13]. This tetraplex is called as i-motif (see Chapter 2). Soon after, the roles of the non-canonical structures have been gaining attention. Especially since the 2000s, research on the G-quadruplex structure formed from Hoogsteen base pairs has made remarkable progress. When a G-quadruplex is formed on DNA or RNA, the reactivity of the protein involved in gene expression is affected (see Chapters 6–8). This means that the central dogma proposed by Crick – that genetic information is determined centrally by the flow of replication, transcription, and translation – is highly controlled by the formation of a G-quadruplex structure. In general, it has been thought that the regulation of gene information expression is due to protein functions. However, the specific structure of Hoogsteen base pairs controls gene expression so that the nucleic acid itself can function like a protein. That is, the roles of nucleic acids might be properly used according to base pairs: Watson Crick base pair = information, non-Watson Crick base pair = function. Many sequences that can have a G-quadruplex structure are distributed in the telomere at the end of the chromosome and the promoter region of the oncogene of the gene. Starting with the 2013 report, there have been many reports on the formation of G-quadruplexes and i-motifs in cells. These reports point out that the oncogene may be activated by the formation (or dissociation) of the G-quadruplex to cause cancer (see Chapter 11). Furthermore, it has been suggested that the phase-separated structure formed by the aggregation of RNAs with G-quadruplexes contributes to neurological diseases such as amyotrophic lateral sclerosis (see Chapter 12).

1.5 Perspective of the Research for Non-canonical Nucleic Acid Structures

As the regulation of gene expression by the specific structure of nucleic acids has been clarified, the next important issue is knowing what specific structures are formed where and when in cells. For example, Hoogsteen base pairs are affected by the molecular environments such as ions, pH, and water activity. Cells are in an environment crowded with molecules, so-called molecular crowding (see Chapters 3 and 4), and the molecular environment changes depending on the cell cycle [14]. For example, the nucleolus causes a change in the molecular density in the nucleus by repeating formation and dissociation according to the cell cycle. This regulates the timing of activation of rRNA transcription in each cell cycle, because the transcription of rRNA specifically occurs in nucleolus. In addition, the environment of mitochondria is particularly crowded (up to $500\,mg\,ml^{-1}$) but heterogeneous due to locally increased proton concentration by the proton gradient required for ATP synthesis. Therefore, it is desirable to develop a technology that can predict

physicochemical property of specific structures due to Hoogsteen base pairs in each characteristic molecular environment [15]. In addition, there is a possibility to make a new approach of drug development that treats diseases by changing the molecular environments of cells, rather than targeting genes and proteins.

1.6 Conclusion and Perspective

According to Pauling's personal communication revealed by the Nobel Foundation's disclosure, he considered Watson and Crick's Nobel award to be premature. In spite of his opinion, the Nobel Foundation decided to award the Prize to Watson and Crick. This might suggest that Watson–Crick base pairs were very common and meaningful at that time but non-Watson–Crick base pairs were artifact and meaningless. Nowadays, non-Watson–Crick base pairs are becoming common and significant as Pauling perhaps predicted. Now, the day when the essence of nucleic acids becomes beyond the concept of Watson and Crick is coming closer.

References

1 Schrödinger, E. (1944). *What Is Life? The Physical Aspect of the Living Cell and Mind*. Cambridge: Cambridge University Press.
2 Tamm, C., Hodes, M., and Chargaff, E. (1952). *J. Biol. Chem.* 195: 49–63.
3 Watson, J.D. and Crick, F.H. (1953). *Nature* 171: 737–738.
4 Pauling, L. and Corey, R.B. (1953). *Nature* 171: 346–346.
5 (a) Hoogsteen, K. (1959). *Acta Crystallogr.* 12: 822–823. (b) Hoogsteen, K.R. (1963). *Acta Crystallogr.* 16: 907–916.
6 (a) Day, R.O., Seeman, N.C., Rosenberg, J.M., and Rich, A. (1973). *Proc. Natl. Acad. Sci. U. S. A.* 70: 849–853. (b) Rosenberg, J.M., Seeman, N.C., Kim, J.J. et al. (1973). *Nature* 243: 150–154.
7 Wing, R., Drew, H., Takano, T. et al. (1980). *Nature* 287: 755–758.
8 Felsenfeld, G. and Rich, A. (1957). *Biochim. Biophys. Acta* 26: 457–468.
9 Kallenbach, N.R., Daniel, W.E. Jr., and Kaminker, M.A. (1976). *Biochemistry* 15: 1218–1224.
10 Gellert, M., Lipsett, M.N., and Davies, D.R. (1962). *Proc. Natl. Acad. Sci. U. S. A.* 48: 2013–2018.
11 Robertus, J.D., Ladner, J.E., Finch, J. et al. (1974). *Nature* 250: 546.
12 (a) Kruger, K., Grabowski, P.J., Zaug, A.J. et al. (1982). *Cell* 31: 147–157. (b) Zaug, A.J., Grabowski, P.J., and Cech, T.R. (1983). *Nature* 301: 578–583.
13 Gehring, K., Leroy, J.L., and Gueron, M. (1993). *Nature* 363: 561–565.
14 Nakano, S., Miyoshi, D., and Sugimoto, N. (2014). *Chem. Rev.* 114: 2733–2758.
15 (a) Takahashi, S. and Sugimoto, N. (2020). *Chem. Soc. Rev.* 49: 8439–8468. (b) Takahashi, S. and Sugimoto, N. (2021). *Acc. Chem. Res.* 54. In press.