New Code for DNA Nucleotide Sequences

Charles D. Schaper, Ph.D.
Transfer Devices, Inc.
Union City, CA 94587 USA
text: 510-378-8260

Corresponding author: Charles D. Schaper, Ph.D.
cschaper@transferdevices.com

March 18, 2020
Abstract

DNA nucleotides consist of the complementary base pairs of Adenine-Thymine (A-T) and Cytosine-Guanine (C-G) that encode as a sequence for genes, and encode for an upstream initiation site that enables transcription. Recently, this lab has shown that steroid hormones are structurally symmetric with each of the four DNA nucleotide pairs and through an ionic binding process may enable gene transcription. Here, a new code is developed for DNA nucleotide sequences that relates to the initiation site for gene transcription. The structural code consists of the orientation of steroid molecules in binding to DNA nucleotides and the class of steroid molecules that form an intermolecular hydrogen bond with an available functional group of Thymine. This later class thereby describes a steroid hormone-DNA nucleotide-ion complex with three hydrogen bonds for A-T and T-A, which thereby matches the three internal hydrogen bonds associated with C-G and G-C. The code consists of two binary vectors to characterize the four configurations of DNA nucleotides and is shown to be consistent with known regulatory elements of DNA sequences associated with gene transcription, including the TATA box and the E-Box, along with other promoters. In addition, the code, which is bijective, is applied to analyze the DNA sequence associated with SARS-CoV-2 to identify regions with relevant structural characteristics.
1 Introduction

DNA nucleotide sequences are comprised of the base pairs A-T (Adenine-Thymine), T-A, G-C (Guanine-Cytosine), and C-G, connected by an oxygen-carbon sugar group that is configured to a chain of phosphor-carbon-oxygen elements, arranged as double strands by multiple internal hydrogen bonds, two for A-T and T-A and three for G-C and C-G. The sequence of the nucleotide base pairs define both the chemical composition of genetic transcription elements, as well as the upstream segments of DNA that initiate genetic expression.

In order to enable the transcription of genes, the current hypothesis is that proteins interact and bind with specific sequences of DNA nucleotides nearby and upstream of the gene to induce the association and enabling of RNA polymerase [1, 2]. There remains much that is unknown about the basic mechanism of transcription initiation under this protein-DNA hypothesis and thus the mechanism of protein interaction with DNA to initiate transcription is an active area of research in which in vitro experimentation and imaging results have been reported of transcription [3, 4, 5, 6, 7, 8, 9]. Thus, in analyzing the consensus base pairs that frequently are upstream of the genes, a number of important regulatory sites have been identified, which include the TATA box [10, 11], E-box [12, 13], and other co-factors and transcriptional elements [14, 15, 16, 17]. Moreover, in addition to the chemical composition, the current hypothesis involve various dynamics and overall shape of the DNA molecule, such as helix and loops, to participate and account for the transcriptional process [18, 19], thereby proposing a rather complex mechanism for genetic transcription.

Nonetheless, recently I have shown that steroid hormones precisely and perfectly coincide structurally with DNA nucleotide pairs, and thus have developed a streamlined mechanism for the initiation of DNA transcription [20, 21] that comprises the direct interaction and binding of steroid hormones to DNA nucleotides for RNA polymerase activity and genetic expression. The results show an unequivocal relation in structure and form, and a theory is thereby developed for the functional capability of steroid hormones and steroid molecules to enable DNA transcription by directly binding to DNA nucleotides. Here, this relationship is examined further to develop a unique code and function for DNA nucleotide sequences, which is addressed towards the initiation of gene transcription process, and compared with known regulatory elements of gene transcription, thereby providing additional confirming evidence of the mechanism. While there is a transcriptional code relating DNA nucleotide sequences to amino acids, the new code developed here relates to the DNA nucleotide sequences that encode for the initiation of the transcription process of such DNA nucleotide sequences that ultimately result in proteins.
2 Results

The first results associating the steroid molecules and steroid hormones with DNA Nucleotides were presented in my earlier work [20], which describes the binding of the steroid hormones to DNA nucleotides, and in [21], which illustrates the structural relations in further detail. In this section, results are presented that analyze the orientation of the steroid-DNA interaction, as to which relative side of the DNA molecule that the binding occurs; and the structural class of steroid hormone that binds to each DNA nucleotide. The code is then used to analyze several regulatory elements that have been indicated in the literature as to have correlation with gene transcription segments of DNA.

2.1 Coding Nucleotide Sequences Based on Steroid-DNA Association

To develop the new code for DNA nucleotide sequences based on its interaction with steroid hormones, the molecules of Figure 1(a) will be utilized. The cortisol molecule, which is a steroid hormone, has a functional oxygen group in the middle of the molecule, in addition to the oxygen-based functional groups on each end. Aldosterone would be a molecule related to cortisol, as it also has a functional group in the center of the molecule. On the other hand, the testosterone molecule, which is also a steroid hormone, is selected because the functional groups are arranged only at the ends of the molecule and it is without a functional group in the middle of the molecule. Progesterone and estrogen would be related molecules. In addition, in Figure 1(b), the DNA nucleotides of the base pairs A-T, T-A, C-G, and G-C are presented. It is noted that the A-T and T-A pairings have two hydrogen bonds, whereas C-G and G-C have three hydrogen bonds. It is also noted that Thymine has an available functional group that is not hydrogen bonded to Adenine. The sugar molecule provides a coupling point, and the phosphate group provide a continuous chain linkage to the base pairs.

As originally shown in [20] and in [21], the A-T and T-A pairing can be associated with the cortisol steroid hormone as shown in Figure 2(a). To get a match to T-A, the cortisol molecule had to be flipped both horizontally and vertically, as indicated in Figure 2(b). The match is perfect in terms of producing three hydrogen bonds through an intermolecular link of cortisol to thymine. Thus, the reason for A-T and T-A to only have two internal hydrogen bonds was to enable an intermolecular binding with steroid hormones similar in structure to cortisol. The testosterone molecule, since it does not contain a functional group in the middle, would not be able to achieve a full set of hydrogen bonds through an intermolecular link with thymine, as does cortisol.

However, since G-C and C-G already have three hydrogen bonds, testosterone is matched to the structure as indicated in Figure 3(a) for G-C. As was the case for A-T and T-A, in order to get the result to work on the reverse C-G, the steroid hormone, testosterone in this case, had to be flipped both horizontally and vertically. The result is indicated in Figure 3(b). It is noted that by flipping the molecule twice, the orientation of the molecule
Figure 1: These structures will be used in developing the new code, and comprise (a) steroid hormones cortisol (upper) and testosterone (lower). Note that the cortisol molecule contains a hydroxyl group in the middle of the molecule, whereas testosterone does not. In (b) the DNA nucleotide pairs are indicated for A-T, C-G, T-A, and G-C. Note that the A-T and T-A pairings have only two internal hydrogen bonds, whereas C-G and G-C have three.

Figure 2: This illustration depicts the A-T and G-C nucleotide base pairs and the binding of the (a) cortisol and (b) testosterone steroid hormones. This binding occurs on the “front” side of the DNA molecule, which will be important in developing the code.

relative to the DNA nucleotide pair, will be the same on the back as it is on the front. It is also noted that the cortisol steroid hormone can not orient itself to the C-G and G-C pairs, because of steric hinderance of the middle section oxygen group. Thus, a full set of three hydrogen bonds results for each of the four steroid hormone and DNA Nucleotide complex.

Having established the positional arrangement of the steroid hormone to DNA nucleotide, when taken together, the orientation and class of steroid hormone correspondence form a new code for DNA as shown in Table 1. For each base pair, there is a unique code of...
For the T-A and C-G pairings, it is necessary to twice flip the steroid hormone, that is in a horizontal and vertical flip, in order to align the steroid hormone to the pairing, which will occur on the “back” of the molecule. Because of the molecule flipped twice, the orientation of the steroid hormone to the nucleotides on the back-side is equivalent to that on the front-side.

<table>
<thead>
<tr>
<th>Base Pair</th>
<th>Steroid Orientation</th>
<th>Steroid Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>A−T</td>
<td>( \mathcal{F} )</td>
<td>( \mathcal{H} )</td>
</tr>
<tr>
<td>T−A</td>
<td>( \mathcal{B} )</td>
<td>( \mathcal{H} )</td>
</tr>
<tr>
<td>C−G</td>
<td>( \mathcal{B} )</td>
<td>( \mathcal{S} )</td>
</tr>
<tr>
<td>G−C</td>
<td>( \mathcal{F} )</td>
<td>( \mathcal{S} )</td>
</tr>
</tbody>
</table>

Table 1: New contribution to the DNA Nucleotide Base Pair Codes indicating the location of the steroid structure, \( \mathcal{F} \) is front side and \( \mathcal{B} \) is back side, and the corresponding class of structurally compatible steroid hormones. \( \mathcal{H} \) signifies the class of cortisol-like molecules, which has an oxygen-based functional group positioned to interact with the available functional group of Thymine, and \( \mathcal{S} \) signifies the class of testosterone-like molecules, which do not have a functional group positioned that can interact with the available functional group of Thymine. Note that each code set is unique to A, C, T and G.

2.2 Binary Representation

Each of the nucleotide pairs has its own unique representation, and a mathematical Base 4 nucleotide pairing is structured mathematically as a decomposition into two Base 2 vector representations of class and orientation. In Table 2, the digital representation is presented,
which can transform the alphabetic code into a consecutive 2-digit code, in which element
represents a structural relation. It can also be expressed as an individual vector sequence
of binary values that indicate the structural arrangement of either the class or orientation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A − T</td>
<td>HF</td>
<td>00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T − A</td>
<td>HB</td>
<td>01</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G − C</td>
<td>SF</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C − G</td>
<td>SB</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Representation of the new code for DNA nucleotide sequences in formats with continuous
2-bit letter or digital code, or presented as a vector code separated as orientation or class
in which by definition 0 = $H$, 1 = $S$, 0 = $F$, 1 = $B$. The vector codes are useful in
searching for patterns, as it is possible to first search for the orientation to establish basic
stability, and then check for the type of hormone binding class to check for the type of
steroid hormone involved.

Thus, for example, in the representation, with $F=0$, $B=1$, $H=0$, $S=1$, a sequence
ATCG, which is represented as HF HB SF SB would be 00 01 10 11, or in vector format as
0011 and 0101. This has practical implementation in a search for a binding sequence from
a long DNA read segment. To pick-up the binding sequence, a certain match would be
needed on one vector, and if that vector is successful, then the corresponding vector would
be checked. Note that it is not necessary to check both vectors, which therefore makes the
search more efficient.

2.3 E-Box

The new code is applied to evaluate transcription regulatory elements, which have known
DNA nucleotide complementary base pairs. For example, an enhancer box or E-box [14] has
been correlated to gene expression in neurons, muscles, and other tissues. The DNA base
sequence for the E-box is generally considered to be CACGTG. In Table 3, the conversion
of the DNA Nucleotide sequence is mapped to the orientation and class of steroid hormone
using Table 2. It is noted that at the 2 and 5 positions, the steroid hormone class is the
same, $H$ and $H$ respectively, but the steroid hormone orientation is different, $F$ and $B$
respectively. As will be indicated in a later, this structural relation can be used to match
with only potential transcription elements to determine an arrangement common to each.

2.4 DNA-Binding Segments

Several segments of DNA have been proposed to have correlation with downstream gene
transcription elements. In this subsection, the transcription factor SMAD2β is evaluated,
in which the DNA base pair is indicated in Table 4 of which there has been proposed
Table 3: For the E-box coding, the hormone orientation, $\mathcal{F}$ is front and $\mathcal{B}$ is back, and corresponding class of structurally compatible steroid class, $\mathcal{H}$ or $\mathcal{S}$. The symmetry is noted on class, and the alternation is noted on orientation. For the 2 and 5 positions, favorable conditions are present to induce steroid hormone binding.

<table>
<thead>
<tr>
<th>DNA Base E-box</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steroid Hormone Orientation</td>
<td>$\mathcal{B}$</td>
<td>$\mathcal{F}$</td>
<td>$\mathcal{B}$</td>
<td>$\mathcal{F}$</td>
<td>$\mathcal{B}$</td>
<td>$\mathcal{F}$</td>
</tr>
<tr>
<td>Steroid Hormone Class</td>
<td>$\mathcal{S}$</td>
<td>$\mathcal{H}$</td>
<td>$\mathcal{S}$</td>
<td>$\mathcal{S}$</td>
<td>$\mathcal{H}$</td>
<td>$\mathcal{S}$</td>
</tr>
</tbody>
</table>

association of protein-DNA interaction in order to enable gene transcription. However, by evaluating the steroid hormone orientation and steroid class, it indicates a possible explanation for gene transcription through the binding of the steroid hormone at the 4, 7 and 9 positions, which would be sufficient to induce gene expression. At these positions, the vector code for class is 000 and the vector code for orientation is 100, which may enable sufficient structural stability of the DNA molecule for gene transcription.

Table 4: For a transcription factor SMAD2β that has been reported to have a DNA binding site, the code is expressed to determine the steroid hormone orientation and class to examine patterns, and see if this segment reported for transcription factors can also be explained as a target site for steroid hormones. In this segment, the positions at 4, 7, and 9 would form compatible binding sites to enable gene transcription by steroid hormones.

<table>
<thead>
<tr>
<th>DNA Base</th>
<th>G</th>
<th>T</th>
<th>C</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steroid Orientation</td>
<td>$\mathcal{F}$</td>
<td>$\mathcal{B}$</td>
<td>$\mathcal{B}$</td>
<td>$\mathcal{B}$</td>
<td>$\mathcal{F}$</td>
<td>$\mathcal{B}$</td>
<td>$\mathcal{F}$</td>
<td>$\mathcal{F}$</td>
<td>$\mathcal{B}$</td>
<td></td>
</tr>
<tr>
<td>Steroid Class</td>
<td>$\mathcal{S}$</td>
<td>$\mathcal{H}$</td>
<td>$\mathcal{S}$</td>
<td>$\mathcal{H}$</td>
<td>$\mathcal{S}$</td>
<td>$\mathcal{H}$</td>
<td>$\mathcal{S}$</td>
<td>$\mathcal{H}$</td>
<td>$\mathcal{S}$</td>
<td></td>
</tr>
</tbody>
</table>

2.5 Application to SARS-CoV-2

In this subsection, the new code is applied to a published DNA sequence associated with the coronavirus SARS-CoV-2, which contains a positive sense RNA strand, which is responsible for COVID-19. This analysis is performed to indicate the utility of the new DNA code to associate structure with sequences. In Table 5, the DNA base sequence is indicated corresponding to a 60 nucleotide snippet of SARS-CoV-2, which is developed from a larger RNA sequence of this data the file www.ncbi.nlm.nih.gov/nuccore/MT027063. Table 1 is used to assign the base of the sequence to its steroid orientation and steroid class. This sequence is then analyzed with a rule to see if the base sequences corresponds to a DNA regulatory sequence. For this case, the coronavirus was evaluated to determine if there was a sequence in which the corticosteroid class of steroid hormone, which has hydrogen bonding (designated as $\mathcal{H}$), was correlated with an orientation the same as the E-box. Thus, a six base pair was analyzed for a $\mathcal{BFBFBF}$ arrangement with a corticosteroid class hormone at the 2 and 5 positions. No match was found with this snippet, however, analysis of the full
DNA sequence did locate multiple targeted sequences of 6 nucleotides and 9 nucleotides arranged in a steroid class and steroid orientation consistent with regulatory elements. The supplemental file contains the coded structures for the steroid hormone orientation and steroid class for a full 29,882 DNA Nucleotide sequence associated with SARS-CoV-2.

<table>
<thead>
<tr>
<th>COVID-19</th>
<th>ATTAAGGT'TATACCCCTCCACGGTAACAAACCAACCTTTCCAGTCTCTCTGTAGATCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation</td>
<td>FBBBFFFFFBBFBBFBFBFBBBFBFFFBFFFBFBFFFBFBFFFBFBFBBFBFBFFFBFBFB</td>
</tr>
<tr>
<td>Class</td>
<td>HHHHHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSSHSH</td>
</tr>
<tr>
<td>E Scan</td>
<td>0000000000000000000000000000000000000000000000000000000000000000</td>
</tr>
</tbody>
</table>

Table 5: A snippet of a DNA base pair corresponding to that of the coronavirus SARS-CoV-2 resultant in the COVID-19 disease. The new code for DNA nucleotide base pair sequences is processed to determine the orientation and class of symmetry with steroid hormone. A candidate base pair sequence associated with the E-box is then run through the two vectors of orientation and class to determine if a sequence is available that can produce a response equivalent to the E-box promoter site. In analysis of the first sixty base pairs, the label E scan denotes the result, which did not produce a match to the targeted six base pairs within the snippet. Of the steroid hormone orientation and class results for the full 29,882 DNA Nucleotides of SARS-CoV-2, there are matches of relevant regulatory sequences.

2.6 Application to 7 bp TATA Box

The TATA box configuration of DNA nucleotides is studied in this subsection. The TATA box is commonly associated with the nucleotide sequence that is upstream to the start site of gene expression, and thus it is very important to analyze to confirm the capability of the code to relate to the initiation of transcription. The pairing is commonly associated as TATAWAW, in which W can be A, T, C or G. The corresponding orientation and class are indicated in Table 6. It is noted that this configuration corresponds to the use of the cortisol steroid hormone to achieve a full set of three hydrogen bonds when associated with the DNA nucleotide pair. In addition, it is noted that the A element in between the two W elements can be associated with a steroid hormone, and then the preceding A as a second coupling site, followed by the first T to achieve three binding sites to stabilize the DNA molecule. Thus, the binding is similar to the E-box studied above, since the DNA-steroid hormone code is the same, even if the DNA Nucleotide sequence is different.

The ability of the steroid class and orientation to stabilize the TATA box is examined using molecular models for the base structure of ATATATAT. The binding is indicated on the backside of one nucleotide and on the frontside of another nucleotide. In Figure 4, the 2, 5 and 7 elements are associated with a cortisol steroid hormone to achieve a stable configuration. In comparison to the energy without the steroid hormone binding of 1,719 kJ/mol, there is significant energy stabilization at -16,642 kJ/mol with three cortisol molecules configured with Ca$^{2+}$ coupling points. There is an additional reduction in energy stabilization of approximately 10% further with Mg$^{2+}$ coupling at -18,519 kJ/mol. It was
Table 6: The steroid hormone orientation and class for the TATA box, which is associated with the initiation point for genes. The binding hormone for the TATA box is of the class associated with strong intermolecular binding to Thymine. The orientation of the steroid hormone to achieve stability can have backside orientation at position 1, followed by frontside orientation at 4 and 6 positions.

<table>
<thead>
<tr>
<th>DNA Base</th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>W</th>
<th>A</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steroid Orientation</td>
<td>B</td>
<td>F</td>
<td>B</td>
<td>F</td>
<td>B/F</td>
<td>F</td>
<td>B/F</td>
</tr>
<tr>
<td>Steroid Class</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>S/H</td>
<td>H</td>
<td>S/H</td>
</tr>
</tbody>
</table>

noted in earlier work, that certain configuration are not stable due to ionic repulsion from neighboring ion connection points; however, this configuration for the TATA box is stable.

Figure 4: To evaluate the TATA box, an eight DNA nucleotide is shown with alternating base pairs of adenine and thymine, thus an arrangement ATATATAT, and stabilized with cortisol binding through Ca$^{2+}$. The cortisol steroid hormone is brought into positional contact with the configuration at the 2, 5 and 7 positions. It is noted that there is a hydroxyl group of cortisol facing outwards toward the environment, which is thus available for further interaction. The molecule is in a stable configuration.

In Figure 6, a side view of the TATA box configuration is indicated for the three steroid hormone interaction at the 2, 5 and 7 positions. The 2 bp gap between 2 and 5 and the staggered configuration minimizes ionic repulsion. Based on the previous rules, it is not possible to place two ions in adjacent pairs, in either a linear or staggered configuration. Thus, another possibility for the TATA box would be 2, 4 and 7, and having the ability for arrangement on both sides of the molecule. However, in this configuration the orientation would be different instead of $BFF$, it would be $BBF$.

The electrostatic potential of the steroid hormone to DNA complex for the TATA box is examined in Figure 6. It is noted that on the top, the coupling point, a Ca$^{2+}$ element, induces a partial positive charge facing the external environment. It is noted that for a coupling point of Mg$^{2+}$, the differentiated positive charge provided by the ionic coupling relative to the phosphor-oxygen-carbon backbone is not as strongly apparent as seen for Ca$^{2+}$. In the 2, 5 and 7 position, the hydroxyl group also is facing the external environment. Thus, there are functional groups available for interaction with the RNA polymerase, which will also align the enzyme with the DNA complex to continue with the process of DNA transcription.
Figure 5: To evaluate the TATA box, the van der Waals spheres of the ATATATAT sequence with cortisol binding at the 2, 5 and 7 positions. It is a stable configuration as the calcium ions maintain association, and do not have significant neighboring ionic repulsion through this arrangement, which is a combination of staggered and linear positioning of the ionic linking elements. It is noted that in addition to calcium ions, magnesium ions would also lead to a stable configuration.

Figure 6: The electrostatic potential of a sequence of DNA nucleotides is shown with alternating base pairs of adenine and thymine, thus an arrangement ATATATAT, of which the steroid hormone cortisol is brought into positional contact and binding with the configuration at the 2, 5, and 7 positions. The partial positive charge is noted at the ionic connection points, and together with additional regulatory elements of the cortisol molecule, can be used to attract RNA polymerase to attach and continue the process of gene transcription.

3 Discussion

The inherent structure common to each DNA Nucleotide pairing is a steroid molecule. The same is true of course for steroid hormones, and thus there is a natural association between the two structures. The steroid hormone stabilizes the DNA molecule, thereby inducing gene transcription by strand separation and the association of the steroid hormone
with RNA polymerase. The attraction of the steroid hormone to DNA will be at every possible site along the chain, provided that there is not steric hindrance preventing access. Furthermore, while it is the case that each steroid hormone would be able to attract to DNA, there are sequences which will only result in stabilization sufficient to bond RNA polymerase for strand separation and gene transcription. Moreover, it is the case that the steroid hormone is transported to the DNA molecule by a receptor, which involves a process in which the steroid hormone is transferred to the DNA nucleotides, which will also preclude certain sites of the DNA molecule from having sufficient access to enable such an interaction.

The development of a new code for DNA nucleotide sequences is necessary to characterize the new recent results that describe the direct binding of steroid hormones to DNA nucleotides. The orientation in which the binding occurs and the molecular structure of the steroid hormones in defining the group produced a couple of binary codes that together form a unique code set of orientation and class. This code is structural in nature, and related to the capability to enable gene transcription. Thus, DNA contains two codes: one associated with the gene in its transcription to amino acid sequences, and the second code, as presented in this article, is associated with the regulatory action of expressing such genes.

There are extensions to the new code that can increase its utility, such as with the application of an energy stabilization measure that is a function of the code sequences. For example, in the TATA box configuration the sequence of three steroid hormones arranged in a $BVVFVFV$ configuration in which the $V$ signifies vacant site, led to a stable arrangement of significant energy stabilization in comparison to the completely vacated section. Moreover, the ionic repulsion was kept to a minimum by the staggered configuration. Thus, it is feasible to compute the energy stabilization over a complete sequence, including upstream and downstream enhancement areas to predict or design sequences for the initiation of gene transcription.

Of the structures studied, the codes faithfully indicated a correspondence between the binding sequence and the structural compatibility of the ion coupling agents. The SARS-CoV-2 sequence analysis produced a non-trivial solution which is consistent with the number of proteins associated with the virus. The E-box analysis indicated cortisol interaction, and this sequence is known to have a circadian rhythm, and thus the interaction with cortisol, which also has a circadian rhythm, has a common link. The TATA box indicated stabilization at three sites of seven base pairs, which is consistent with the observed binding domain inducing gene transcription. Further, the interaction of the steroid hormone with functional sites potentially associated with RNA polymerase was indicated, which would also define a directionality and alignment to the transcription direction.

For genetic applications, it seems reasonable to evaluate the potential for these initiation sites, and the surrounding elements, to be evaluated for their capability to induce transformative phenotypes based upon the inheritance of the sequence to promote the initiation and stabilization of gene transcription. In addition to initiating gene expression,
these initiation sites may also be affiliated with the termination of the overall transcription action, as the removal of the steroid hormone or ion coupling agent from the DNA molecule will result in a termination of gene expression due to a loss of stabilization. Thus, the stability of this site is crucial to the magnitude of the genetic expression. Research to apply the new code and extensions for the induction of transcription from DNA nucleotide sequences by endogenous binding of steroid molecules and hormones is underway for medical and genetic studies.

4 Methods

• Molecular Modeling: Molecular models were constructed using the software program Avogadro was used and the force field MMFF94 [22] was utilized. An optimization program available in the software that minimizes the energy associated with the complex was used for determining the bond length and the positioning of the chemical elements.

• Interaction of Steroid Hormone and DNA Nucleotide Sequence: For the TATA box, the molecular modeling software was used to determine the position of the cortisol molecule and the Adenine-Thymine nucleotide structures. Eight alternating pairs of A-T nucleotides were prepared. The optimization routine was utilized to determine the lowest energy state. The initial positioning of the cortisol molecule and the two calcium ions was performed using previously described methods [23].
Bibliography


