Phys. Biol. 19 (2022) 036005

Physical Biology

Prediction and comparative analysis of CTCF binding sites based on a first principle approach

Nestor Norio Oiwa¹,², Kunhe Li², Claudette E Cordeiro³ and Dieter W Heermann², ², *

¹ Department of Basic Science, Universidade Federal Fluminense, Rua Doutor Silvio Henrique Braune 22, Centro, 28625-630 Nova Friburgo, Brazil
² Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, D-69120 Heidelberg, Germany
³ Department of Physics, Universidade Federal Fluminense, Avenida Atlântica s/n, Gragoatá, 24210-346 Niterói, Brazil

* Author to whom any correspondence should be addressed.
E-mail: heermann@tphys.uni-heidelberg.de

Keywords: CTCF binding sites, Cys2His2 zinc finger, extended ladder model

Supplementary material for this article is available online

Abstract
We calculated the patterns for the CCCTC transcription factor (CTCF) binding sites across many genomes on a first principle approach. The validation of the first principle method was done on the human as well as on the mouse genome. The predicted human CTCF binding sites are consistent with the consensus sequence, ChIP-seq data for the K562 cell, nucleosome positions for IMR90 cell as well as the CTCF binding sites in the mouse HOXA gene. The analysis of Homo sapiens, Mus musculus, Sus scrofa, Capra hircus and Drosophila melanogaster whole genomes shows: binding sites are organized in cluster-like groups, where two consecutive sites obey a power-law with coefficient ranging from 0.3292 ± 0.0068 to 0.5409 ± 0.0064; the distance between these groups varies from 18.08 ± 0.52 kbp to 42.1 ± 2.0 kbp. The genome of Aedes aegypti does not show a power law, but 19.9% of binding sites are 144 ± 4 and 287 ± 5 bp distant of each other. We run negative tests, confirming the under-representation of CTCF binding sites in Caenorhabditis elegans, Plasmodium falciparum and Arabidopsis thaliana complete genomes.

1. Introduction
In mammals the primary insulator is the nucleotide sequence CCCTC-binding factor (CTCF), a protein with 10 Cys2His2 and one C2HC zinc finger and the major eukaryotic DNA-protein binding motifs [1–4] (cf figure 1(b)). These transcription factors are characterized by 3 to 29 zinc finger (ZF) units [5, 6], each composed by one zinc ion linking two cysteines at the end of two β-sheets and two histeines in the C-terminal of one α helix [7, 8]. Chromatin immunoprecipitation assays with DNA microarray indicate at least 13 804 actives binding sites [2] and Xie et al [3] reports a minimum of 15 000 binding sites for CTCF, using chromatin immunoprecipitation assay with massively parallel DNA sequencing (ChIP-seq). Chen et al [4] estimates 326 840 possible sites along the human genome, combining the data from 38 cell lines. Despite CTCF relevance, the quality is poor in 20% to 30% of the available data due to limitations of the experimental apparatus and the algorithms for localizing binding site [2, 4, 9]. Same mistakes are made, adding false binding sites and making impossible in see the structure of the CTCF distribution. In this paper we present a new method to finding CTCF binding sites based on the interaction of the zinc finger and the electronic cloud of the nucleotide π-orbital of the double DNA (dDNA). This is a first principle approach method, because we compute the local electron density of states using electron–nucleotide interactions along the genome [10] (referee 1: item 5). This quantum mechanic charge transport description of the nucleotide, typical in semiconductor physics, adds a new layer of information beyond traditional four letter nucleotide
genomics. In this way, we overcome the limitations of previous works, unveiling a power law along CTCF binding sites in many complete genomes.

The workflow and organization of the article is illustrated in figure 1(a). First of all, we collect 23 experimentally detected CTCF-DNA binding site (see supplementary material S1 (https://stacks.iop.org/PB/19/036005/mmedia)). Then, we study the electronic cloud of the nucleotide $\pi$-orbital using [10]. This analysis extends the usual nucleotide alignment based on hydrogen bonds, adding information about the electronic behavior in CTCF binding sites as ground state, highest occupied orbital (HOMO) and lowest unoccupied orbital (LUMO) (see S2). Once we establish a pattern based on our electronic nucleotide alignment, we apply it over a complete genome in multiple genomes (see S3). We validate our putative CTCF binding sites with the consensus sequence [2, 11–13], ubiquitous ChIP-seq K562 data [4, 14], MNase-seq of IMR90 cell with improved nucleosome positioning (iNPS) [15–17] and the cluster HOXA [18]. After corroboration of our putative CTCFbs, we study the distribution of CTCFbs over the complete human, mouse, pig, goat, fruit fly and Aedes aegypti (mosquito) genomes. We use the complete Caenorhabditis elegans, Plasmodium falciparum and Arabidopsis thaliana genomes as negative controls. We report cluster-like structures for the CTCF distribution in multiple species. Finally, we discuss the limitations of our method as well as ChIP-seq data.

2. Method

2.1. CTCF samples

In order to establish an electronic nucleotide pattern, we consider 23 experimentally confirmed CTCF binding sites, figure 1(b). Detailed descriptions about these CTCFbs are in supplementary material S1.

The nucleotide sequences in figure 1(b) are fasta or gbk files extracted from the GenBank reference map [19]. We do not use the original sequences from the articles, because the literature only publishes the binding site nucleotides. This is insufficient for $\pi$-orbitals. We are not restricted just to the nucleotides of the consensus CTCF motif. The electronic nucleotide description of nucleotide $\pi$-orbitals considers the effects of the surround of the core 20-mers. Electrons can easily hop for 16.8 (AT rich sequences) or 25 Å (CG rich) [20], which comprehend at least 5 to 8 bp of the surrounding nucleotides over the core 20-mer binding site. Results with transcription factor specificity protein 1 (SP1) and early growth response protein 1 (EGR1) [10] show the existence of HOMO and LUMO surrounding binding site. Similar phenomena happen for CTCF as we will report in this work, although the biological function of HOMO and LUMO is unknown yet. We can easily find the selected binding sites in the GenBank reference map with the same SP1, EGR1, initiator element (Inr), Goldberg-Hogness box (TATA box) and other expected genomics features. All selected binding sites must be experimentally confirmed for multiple methods.

2.2. Nucleotide alignment using local electronic density of states

The starting point in our method is the quantum description of the nucleotide $\pi$-orbitals along the genome, considering three terms in the Hamiltonian (equation (1) of the supplementary material S2): electron–electron, electron–nucleotide displacement field and electron–nucleotide interactions. The first term is just the free electron along the base pairs. The electron–nucleotide displacement field is the $\pi$-orbital response with its own nucleotide. The electron–nucleotide interaction between two base pairs are represented by the Morse potential and anharmonic spring. This technique combines DNA melting [21] with the extended ladder model [22, 23]. When we diagonalize the proposed Hamiltonian in eigenvalues and eigenvectors, the nucleotide $\pi$-orbitals along DNA is described as local density of states (LDOS) of the ground state, holes (nucleotides in the valence band without free electron) and highest occupied orbital (HOMO) along with lowest unoccupied orbitals (LUMO), beyond the usual four letters nucleotide alignments [10]. The computation of LDOS is detailed in supplementary material S2.

In the context of charge transport, the valence band is the energy levels of the electrons between the ground state and HOMO. The conduction band are the energy levels of the electrons beyond LUMO. Since we have one free electron per nucleotide in the extended ladder model, the valence band will be completely filled and the conduction band will be empty. The ground state electrons are the least mobile, while the HOMO electrons are the most movable ones and they may hop from HOMO to LUMO. In this work, the nucleotides with ground state electrons, marked in yellow in figure 1(c), are actually the nucleotides with at least 10% of probability of finding the degenerated ground state electrons. The difference between HOMO and LUMO is absent in conductors, while electric insulators present wide gaps. The gap in the extended ladder model [10] gives a semi-conductor characteristic for the double helix.

The common tools for four letters nucleotide alignment are useful for early alignments of the CTCF samples mentioned in the previous section [24, 25], but these drafts should be reevaluated since they do not consider the electronic features pointed in this article. So, we perform a second nucleotide alignment, considering simultaneously adenine (A), citosine (C), guanine (G), timine (T), ground states (yellow), HOMO (orange) and LUMO (red bordered boxes), figure 1(c). This second alignment in figure 1(c) is made manually.
Figure 1. (a) Our workflow. (b) Sketch of the molecular structure of the CCCTC-binding factor (CTCF) with 11 zinc fingers (ZF). Panel (c) shows the electronic nucleotide alignment for CTCF. We indicate the reading direction in reverse and complementary strands with r and c in the parentheses, respectively. Nucleotides with at least 10% of probability in finding the degenerated ground state electrons are in yellow. The highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) nucleotides with at least 10% of probability in localizing one electron are respectively indicated by orange and red bordered boxes. The eleven CTCF zinc finger positions are marked by the succession of blue and white boxes. The CTCF-DNA binding patterns are indicated by the letters CBA. We indicate respectively the binding positions A and C of ZF2 and ZF6 using dashed boxes. B is about the CTCF-DNA binding of ZF4 and ZF5. We tag respectively by $^\ast$BA and CB$^\ast$, when the CTCF miss ZF6 or ZF2 bindings. We remark that there is no nucleotide position zero, in line with [26]. (d) The consensus sequence is the simple majority (number of alignment nucleotides $\geq 12$). The black bordered boxes indicate the guanines that appear in all studied CTCF binding sites. The four modulus in [13] are indicated in light and dark green. (e) Is number of nucleotide occurrence per site. (f) Is the nucleotide motifs along the columns with the number of repetitions in parentheses.
We cannot ignore the symmetries of the genetic code, since CTCF read dDNA in four directions in function of complementary and reflection symmetries. So, the charge patterns of the tips of ZF and the LDOS of DNA chains must be evaluated in the direct or positive strand and direct reading (from 5’ to 3’), in the direct strand and reverse reading (from 3’ to 5’), complementary or negative strand with direct reading and complementary strand with reverse reading.

2.3. Pattern identification

We divide the prediction technique in two parts. In the first part of the technique, we scan the contiguous sequences (contigs), looking for the electronic distribution patterns as a very specific ground state positions in guanines and absence of HOMO and LUMO around CTCFbs. These electronic patterns, figure 1(c), are described further in the text (section 3.1, consensus sequence). Then, we consider the number of nucleotide occurrence and the motifs in figures 1(e) and (f). Since the length and the number of the binding sites is small in figure 1(c), we do not use any algorithm for motifs detection and classification. We arrange the nucleotides manually. Indeed, there are only four and three observed motifs in the ZF4 and ZF5 triplets, figure 1(f). The number of motifs is reduced in the middle of the binding site, but large in the flanking region. So, we divide the nucleotides in two sets: $S_{\text{core}}$ and $S_{\text{flank}}$.

In the core of the CTCFbs, we define the geometric average probability $P_{\text{core}}(S_{\text{core}}) = \prod_{k} P(S_k)^{1/7}$, where $S_{\text{core}} = \cup_i S_i$ and $k = \{ZF2, ZF4, ZF5\}$, and $P(S_i)$ is the probability of occurrence of the motif $S_i$, figure 1(f). We have a cubic root in $P_{\text{core}}$, because we are analyzing the patterns of 3 zinc fingers. After extensive tests localizing the listed figure 1(c) in GenBank flat files, we conclude that a minimum of 9.0% for $P_{\text{core}}$ is required for a valid DNA-CTCF binding.

In the region flanking the core, we define a probability $P_{\text{flank}}(S_{\text{flank}}) = \frac{1}{2} \left[ \prod_{k} P(S_k) \right]^{1/7} + \frac{1}{4} \left[ \prod_{k} P(S_k) \right]^{1/15}$, where $S_{\text{flank}} = \cup_i S_i$ for $k = \{ZF2a, ZF2b, ZF6, ZF7, ZF8a, ZF8b, ZF9\}$, $P(S_i)$ is the probability of occurrence of the motif $S_i$, and $P(S_i)$ is the probability of the nucleotide occurrence $S_i$ in the position $i, i = -11, -10, \ldots, 13$. The first term $\left[ \prod_{k} P(S_k) \right]^{1/7}$ in $P_{\text{flank}}$ guarantees the detection of nucleotide sequences listed in figure 1(f), and we have 7th root in the expression since we are considering seven elements in $S_i$. However, there are considerable variation in $S_{\text{flank}}$, comparing with $P_{\text{core}}$. If we restrict the motifs just in figure 1(f), we will miss valid CTCFbs. So, we introduce $\left[ \prod_{k} P(S_k) \right]^{1/15}$ in $P_{\text{flank}}$. We decompose the flanking sequence in their 15 nucleotides, $S_i = \{a, t, c, g\}$. Then, we estimate the geometric average probability associated with the occurrence of each particular nucleotide $S_i$ along the binding site, figure 1(e). Our tests show that the probability of a valid CTCFbs $P_{\text{flank}}$ should be bigger than 6.5%.

We illustrate the procedure in the supplementary material S4.

3. Validation

3.1. Consensus sequence

The most striking feature of the alignment of 23 CTCFbs in figure 1(c) is the guanine at the positions 2 and 5, marked with a black box in figure 1(d). Actually, guanines at the position 2 and 5 coincide with the middle nucleotide of the triplet of the ZF4 and ZF5 and the amino acid of tip of these ZF tips are base. So, the positive charged tips of ZF4 and ZF5 bind with the ground state electrons of guanines in position 2 and 5; a similar mechanism is described in [5, 7, 8, 27, 28]. Coarse-grained Monte Carlo simulations confirm this finding [29, 30]. Further, Kim et al. [2] increases the specificity of their CTCF binding site prediction using these same nucleotides in positions 2 and 5 as well as -4 and 7 (positions 6, 11, 14 and 16 in their article). There is always adsorption of the zinc fingers 4 and 5 by the DNA.

We do not observe HOMO between -5 to -2 and 2 to 9, and there is no LUMO between -4 to 6. We never observe over-position between ground state and HOMO or LUMO electrons. The core of CTCF-DNA binding sites is a region without mobile electrons and CTCF anchors their zinc fingers in the most stable electrons, i.e. ground state electrons.

Since the electronic alignment considers the charges in the tips of the zinc-finger [10], the eleven ZFs in CTCF reveal more details about the protein-DNA attachment. There are five ZF with well-defined charge motifs: ZF2, ZF4, ZF5, ZF6 and ZF9. The finger tip is acid (negative) for ZF2 and ZF6 as well as base (positive) for ZF4, ZF5, and ZF9. Electrons in the nucleotides will bind the positive tips, and holes in negative ones. We will ignore ZF9, because it is neither in the core binding site nor fundamental for CTCF-DNA binding [29]. ZF4 and ZF5 always bind with the dDNA [29]. We do not find any particular property for ZF3. Thus, we will focus on the binding sites for ZF2 and ZF6 (respectively A and C in figure 1(b)) and ZF4 and ZF5 marked as B in figure 1(c). Instead of three nucleotides of the triplet, we consider five nucleotides in A and C, blue box in figure 1(c), because the CTCF is a flexible molecule and the finger may displace back and forward along the double helix. The site B is the triplets under ZF4 and ZF5. CTCF sometimes misses the binding sites A or C, but it always binds in B. CTCF-DNA binding is successful only if we do not miss A and C sites simultaneously, figure 1(c).

The consensus sequence in figure 1(d) is just the simple majority (number of alignment nucleotides ≥ 12). We avoid the Schneider and Stephens logo, and we use neither the Shannon information content,
Gibbs binding free energy nor position weight matrix for the calculus of the specific-binding free energy [31–34], because we have neither a clear boundary of the major groove and reverse reading (from 3′ to 5′), since it is easier for the CTCF to insert into the major groove. We can see in figure 1(c) that we have 21 samples in the major groove and the matching between predicted and ChIP-seq K562 data increases: 34% of binding sites will have a perfect matching, with Q2 = 401 bp, Q3 = 2238 bp and a maximum discrepancy of 60 676. However, we have only 22% of matching, Q2 = 621 bp, Q3 = 2348 and a maximum of 73 246 bp difference for CTCF binding in the minor dDNA groove. Here, we linked the minor groove with the direct reading in the complementary strand and reverse reading in the direct strand. In figure 1(c) BRCA1 CTCF2 and MEG3 CTCFα are in the direct strand and reverse reading, associated with the minor groove. The absence of major and minor groove distinction in our method is obvious when we see the chromosomal average proportions of each reading direction: direct strand and direct reading is 29 ± 1% of the predicted CTCFbs; direct strand and reverse reading has 21 ± 1%; complementary strand and direct reading values 20.8 ± 0.7%; and complementary strand and reverse reading is 29.4 ± 0.7%. The number of direct strand and reverse reading as well as complementary strand and direct reading could be overestimated.

3.2. CTCF and ChIP-seq K562 data

Once we identify the electronic nucleotide pattern and establish a criteria for CTCF binding sites, we localize all human CTCFbs along the assembly hg38, table 1. We find 335 088 binding sites. This number is remarkable close to the total cumulative number of 326 840 CTCF binding sites identified by Chen et al using data from 38 human cell lines [4].

We compare our predicted CTCFbs to the ChIP-seq K562 ubiquitous binding sites. The 8771 ubiquitous CTCFbs from 5 ENCODE K562 files are described in supplementary material S5. We have 29.8 ± 3.8% of perfect match between our method against experimental data. The median Q2 of the distances between predicted and observed binding sites shows us that 50% of the putative are just at a 473 bp distant from the expected one and 75% of them (third quartile, Q3) are at the maximum 2352 bp. Beyond Q3, we have some huge discrepancies reaching 73 250 bp. As we lay out in the discussion, the discrepancy of the last quartile (25% of data) between our putative CTCF binding sites and those detected by ChIP-seq comes from the limitations of the chromatin immunoprecipitation technique.

We can improve the matching in light of the helical geometry of the dDNA. When we observe the three-dimensional structure of dDNA, there are two possible grooves where the zinc finger will insert into the dDNA to read the π-orbital. The major groove is 22 Å large, while the minor groove has only 11 Å [39]. We expect more CTCFbs in the direct strand and direct reading (from 5′ to 3′) and in complementary strand and reverse reading (from 3′ to 5′), because we have neither a clear boundary of the major groove and reverse reading (from 3′ to 5′), since it is easier for the CTCF to insert into the major groove.
Phys. Biol. 19 (2022) 036005

N N Oiwa et al

UCSC genome browser [42] and the nucleosome positions are predictions using [43]. Since our analysis is done on a genome-wide scale, the relatively remote nucleosomes are considered to be more fluctuated and hard to recognized in the figure; on the other hand, this also demonstrates the prominence of the observed peaks. Each nucleosome in our work may spread for few kbps. Interestingly, the faint signal between the gene A4 and A5, not reported by [18], is positioned in one putative CTCFbs group, detected by our program. This faint signal is a cluster with seven CTCFbs, asterisk in figure 2(b). The CTCFbs organized in cluster-like groups [4] as C6/7 with two binding sites, C7/9 with 4, C10/11 with two, the CTCFbs before and after A13 with two and four respectively, figure 2(b). The CTCFbs assays by [18] have a precision around 1 kbp and are unable to find one particular 20 bp long CTCFbs binding sites. In the case of the CTCFbs after A13 gene, the CTCFbs cluster stretches for 3625 bp. The motif based methods in CTCF assays do not consider the local repeats as alternative sites. They choose one of possible sites that may spread for few kbps. Interestingly, the faint signal between the gene A4 and A5, not reported by [18], is positioned in one putative CTCFbs group, detected by our program. This faint signal is a cluster with seven CTCFbs, asterisk in figure 2(b). We also observe, respectively, a mismatch of 2981 and 2795 bp between our results and [18] for C5/6 and C6/7. The source for this displacement is the considered CTCF consensus motifs. In order to test the robustness of our electronic alignment, we do not include their CTCF motifs (supplemental material in [18]) in our 23 samples, figure 1(b).

In the previous section, we show that we get better results considering only CTCFbs in the major dDNA groove. However, the CTCFbs in the minor DNA cannot be completely neglected. Since we have many CTCFbs in figure 2(b). For example, the CTCFbs of C5/6, C6/7 and C10/11 are all in the minor dDNA groove.

The mismatches between our results and [18] around C5/6 and C6/7 give us an idea about the inaccuracy in the positioning $\lambda_{\text{ctcf}}$ of our method. The estimate of misplacement is around 3 kbp. However,

Table 1. $L$ is the genome length, $n_{\text{ctcf}}$ is the predicted CTCF binding sites (CTCFbs) and $\langle l_{\text{ctcf}} \rangle$ is the chromosomal average of CTCFbs density. The probability distribution $P(\Delta)$ of the difference $\Delta$ between two consecutive CTCFbs obeys a scaling law $\alpha$ from 11 bp $\leq \Delta \leq$ 2000 bp to 16 bp $\leq \Delta \leq$ 17000 bp, depending of the considered genome. $P(\Delta)$ follows an exponential decay with typical length $\lambda$, when we consider the fitting regions from $2000 \ leq \Delta \ leq$ 78 kbp to 9.7 kbp $\leq \Delta \leq$ 99 kbp.

| Species   | $n_{\text{ctcf}}$ | $\langle l_{\text{ctcf}} \rangle$ (kbp) | $\alpha$ | $\lambda$ (kbp) |
|-----------|------------------|---------------------------------|--------|----------------|
| Human all | 2814809546        | 331668                          | 8.8±3.1| 0.511±0.014    | 19.28±0.24                  |
| Human cent | 76305151          | 1892                            | 38±17  | 9.207          | No structure                |
| Human var | 140590877         | 1528                            | 30±14  | 9.207          | No structure                |
| Mouse all | 2541456020        | 277027                          | 9.4±1.9| 0.392±0.0068   | 19.79±0.31                  |
| Mouse chrom Y | 824248315      | 2512                            | 32.74  | Detailed in the text |

| Genome       | $n_{\text{ctcf}}$ | $\langle l_{\text{ctcf}} \rangle$ (kbp) | $\alpha$ | No structure |
|--------------|------------------|---------------------------------|--------|--------------|
| Pig          | 2389924585        | 316919                          | 7.9±2.7| 0.484±0.013  | 22.44±0.37                  |
| Goat         | 2462599355        | 264286                          | 9.7±3.2| 0.5409±0.0064| 24.24±0.35                  |
| Fruit fly    | 128506876         | 8962                            | 14.4±2.2| 0.454±0.012  | 18.08±0.52                  |
| Fruit fly chrom 4 | 1200662         | 20                              | 60.03  | No structure |
| A. aegypti   | 1195030408        | 39777                           | 30.043 | 144±4±bp, 287±5 bp |
| C. elegans   | 100272607         | 2086                            | 48.2±7.7| No structure |
| P. falciparum| 23264338          | 46                              | 530±340| No structure |
| A. thaliana  | 116129212         | 1595                            | 72.7±3.6| No structure |

$^a$Heterochromatin were excluded.
$^b$No standard deviations due to the reduced amount of data.
$^c$Chromosome Y is excluded.
$^d$Well-defined $\Delta$ CTCFbs distances.
Figure 2. (a) The black line shows the average chromosomal density $\rho(i_{\text{nucl}} - i_{\text{ctcf}})$ of nucleosomes per nucleotide around the predicted CTCF binding site (CTCFbs) $i_{\text{ctcf}}$, excluding the nucleosome in the CTCF position. The exact neighbor nucleosome positions are indicated by numbers above the peaks with the error in the parenthesis. (b) Shows the HOXA genes (black line), the predicted (white diamond) and predicted CTCFbs which are experimentally confirmed (red triangle) from the mouse HOXA gene cluster. The red triangles in C5/6, C6/7, C7/9, C10/11, before and after C13 are the CTCFbs reported in [18]. The asterisk indicates the faint response for CTCF in [18], not reported by the authors. (c) and (d) Are probabilities $P(\Delta)$ in finding the next consecutive putative CTCF binding site in percentage against the distance $\Delta$ in base pairs for human (black) and fruit fly (red). The dark dashed and the red dotted dashed lines indicate the power-law for human and D. melanogaster, respectively. (d) The exponential fitting for human and fly in semi-log scale are also pointed by dark dashed and red dotted dashed lines. (e) $P(\Delta)$ of mouse chromosome Y with multiple typical $\Delta_1 - \Delta_{16}$ distances. (f) Aedes aegypti $P(\Delta)$ with the characteristic 144 ± 4 bp ($\Delta_1$) and 287 ± 5 bp ($\Delta_2$) distances.
the most evident feature in figure 2(b) is the coalescence of the CTCFbs, reported by [4] as clusters of binding sites. However, the concept of cluster demand a Gaussian among CTCFbs distribution and we do not observe such structure. Since our electronic alignment is not limited by poor quality data [9, 45] or absence of the expected 20-mer consensus motif [4], we make more accurate analysis.

4. Results

Instead of a cluster organization for CTCFbs suggested by [4, 12], we implement another evaluation, detecting a power law in $P(\Delta)$, table 1 and figure 2(c), indicating organized structure for CTCFbs. Here, $\Delta$ is the distance of two consecutive CTCFbs and $P(\Delta)$ is the probability of finding the next binding site. In humans we adjust $\alpha$ in $P(\Delta) \approx \Delta^\alpha$, considering two or three orders of magnitude. The human euchromatic regions have $\alpha = 0.511 \pm 0.014$, fitting within the interval 14 bp $\leq \Delta \leq 2400$ bp. The region with a power law in $P(\Delta)$ covers 39.98% of the euchromatic binding sites. Furthermore, the chromosomal average $\alpha$ of mouse values 0.3292 $\pm$ 0.0068, covering 43.93% of binding sites, and it is fitted in the interval 20 bp $\leq \Delta \leq 4.1$ kbp.

For the region beyond polynomial fitting, $P(\Delta)$ decays exponentially, $P(\Delta) \approx e^{-\Delta/\lambda}$. The characteristic length $\lambda$ for humans values $\lambda = 19.28 \pm 0.24$ kbp and 15 kbp $\leq \Delta \leq 99$ kbp is the exponential adjustment region, comprising 16.36% of binding sites. In the case of the mouse, the genomic $\lambda$ values 18.06 $\pm$ 0.29 kbp with 11 kbp $\leq \Delta \leq 89$ kbp, containing 23.77% of CTCFbs.

A similar feature is described for the human K562 CTCF binding sites distribution [4]. However, we cannot compare the power law CTCFbs distribution for the entire genome directly with their cluster analysis [4], since the power law has not a characteristic length by definition. Thus we use a cluster analysis, assuming those CTCFbs to be nearest neighbors that are within 3058 bp and hence in one particular cluster. We choose 3058 bp because this is the median for the complete genome $\Delta$ as well as this is close to the upper limit of the power fitting, table 1. Thus, 63.68% of our cluster-like structures can be classified as singletons (isolated CTCFbs), while [4] reports 38.94%. The groups with 2, 3, 4, 5, 6 and more than 6 CTCFbs values respectively 18.09%, 7.08%, 3.45%, 2.11%, 1.26% and 4.26% while [4] indicate 25.09%, 14.60%, 8.79%, 5.22%, 3.10% and 4.26% in their cluster map. Although we have more singletons in our results, we have the same percentage for cluster-like structures with more than 6 CTCFbs reported by [4].

We do not restrict our analysis just to human and mouse. We confirm the existence of cluster-like structures in pig and goat, where we find 316 919 and 264 286 CTCFbs respectively. Both average chromosomal CTCFbs densities ($l_{ctf}$) are compatible with the human and mouse, but direct comparison should be avoided because we exclude the heterochromatin in the human genome. $\alpha$ values are 0.484 $\pm$ 0.013 and 0.5409 $\pm$ 0.0064 for pig and goat respectively. They contain 44.27% (pig) and 41.31% (goat) of the binding sites. Both species have the same regions for $\alpha$ fitting: 14 bp $\leq \Delta \leq 2000$ bp, but the domains for $\lambda$ adjustments are different: 14 kbp $\leq \Delta \leq 99$ kbp for pig, covering 13.75% of binding sites; and 18 kbp $\leq \Delta \leq 99$ kbp in the case of goat, composing 13.31% of CTCFbs.

$P(\Delta)$ is not limited just to polynomial and exponential fittings. We have many CTCFbs that are 13 bp apart from each other as well. 5.67% of human euchromatin, 6.39% of mouse without chromosome Y, 7.58% of pig and 7.19% of goat binding sites are in the region $0 < \Delta \leq 13$ bp, and $P(\Delta)$ distributions are not uniform. We observe few binding sites with $\Delta = 2, 5$ or 7 bp and the height of $P(\Delta)$ is species dependent. By the way, 0.23%, 0.40%, 0.49% and 0.46% of binding sites are $\Delta = 0$ distance respectively in human, mouse, pig and goat, i.e. the CTCF has multiple binding modes in these sites as mentioned previously.

We also apply our method to the fruit fly and localize 8962 binding sites. Although the genome size is just 5% of mammals, $P(\Delta)$ of Drosophila melanogaster resembles mammal with a well-defined power law $\alpha = 0.454 \pm 0.012$ and exponential decay $\lambda = 18.08 \pm 0.52$ kbp. The polynomial and exponential fittings are along 14 bp $\leq \Delta \leq 14000$ bp and 15 kbp $\leq \Delta \leq 78$ kbp, covering 62.73% and 29.83%. 5.87% of the binding sites are 13 bp or less distant each other and 0.12% has $\Delta = 0$.

We study the genome of A. aegypti and identify 39 777 binding sites. We do not find a power law, but 16.33% and 3.57% of binding sites are respectively 144 $\pm$ 5 bp ($\Delta_1$) and 287 $\pm$ 5 bp ($\Delta_2$) at a distance of each other, figure 2(e). We remark that we need 146 bp to wrap one nucleosome. 0.37% of sites has multiple binding modes, $P(0 < \Delta \leq 13)$ is unlike the other genomes, since 1.26% of CTCFbs are just at one bp distance of each other. When we consider a region of 4.5 kbp $\leq \Delta \leq 99$ kbp for the exponential fitting, we have $\lambda = 42.1 \pm 2.0$ kbp. The exponential fitting contains 61.47% of CTCFbs.

This odd behavior can be observed in mouse chromosome Y too, where we find 2512 binding sites. The low density of $\langle l_{ctf} \rangle = 32742$ bp per predicted CTCFbs hides a surprise. This $\langle l_{ctf} \rangle$ is just 9% higher than A. aegypti, and there is neither a power law nor an exponential decay. The number of binding sites in the region where $0 < \Delta \leq 13$ bp is minimal, is just 0.6%. We do not observe multiple CTCF binding modes, $P(\Delta = 0) = 0$. Although mouse chromosome Y lacks a power law and an exponential decay, 35.51% of binding sites presents well-defined $\Delta$ distances: 0.96%, 1.47%, 0.92%, 2.15%, 6.21%, 0.80%, 1.83%, 1.15%, 5.02%, 2.03%, 2.87%, 1.15%, 2.83%,
1.15%, 2.03% and 2.95% of the binding sites are 208 ± 5 (Δ₁), 780 ± 7 (Δ₂), 931 ± 1 (Δ₃), 1539 ± 8 (Δ₄), 1927 ± 2 (Δ₅), 4736 ± 4 (Δ₆), 6949 ± 11 (Δ₇), 7161 ± 22 (Δ₈), 7587 ± 12 (Δ₉), 17523 ± 16 (Δ₁₀), 20862 ± 21 (Δ₁₁), 23551 ± 27 (Δ₁₂), 28153 ± 29 (Δ₁₃), 28460 ± 21 (Δ₁₄), 31452 ± 25 (Δ₁₅) and 40565 ± 91 bp distance of each other (Δ₁₆) in figure 2(d), respectively.

We test our method for *Plasmodium falciparum* (low unicellular eukaryote) and *Arabidopsis thaliana* (plant), where CTCF is absent [46]. In the case of *P. falciparum*, table 1, the number of CTCFbs spotted by our method is so small that we cannot even build *P*(Δ). As a matter of fact, there are only 3 ± 3 CTCF binding sites per chromosome. We have better statistic for *A. thaliana*, table 1, where we detected 1595 CTCFbs. The expected binding sites in the region 0 < Δ ≤ 13 bp is represented by 7.4% of CTCFbs and they are at 6 ± 4 bp distance of each other. We do not report multiple binding modes for these species, *P*(Δ = 0) = 0 and there is neither a polynomial nor an exponential decay for *P*(Δ). These binding sites detected by our method are false positives. They are born from the *P*(S) statistics in *P*_rank from pattern identification and other limitations outlined along this manuscript.

*Caenorhabditis elegans* is another interesting specimen. Although this worm lost its CTCF gene along the evolution [47], we encounter 2086 binding sites, possible remains of its segmented body past [47]. In the region 0 < Δ ≤ 13 bp, we have 5.27% of the binding sites and there are two sites with multiple binding modes. These values are compatible with mammalian genomes. But we do neither find a power law nor an exponential decay in its CTCFbs distribution. The density of CTCFbs in *C. elegans* is 48.2 ± 7.7 kbp. This ⟨Δctf⟩ is not far from human centromeric domains (38 ± 7 kbp per CTCFbs, table 1). Here we have *P*(0 < Δ ≤ 13 bp) = 3.3% and 0.2% of sites present multiple binding mode, but we do neither find a power law nor an exponential decay.

One may argue the absence of a power law and exponential decay in *P*(Δ) is due to the low density of CTCFbs ⟨Δctf⟩ in the human centromeric domain or in mouse chromosome Y. However, we have an unusual concentration of binding sites in the human noncentromic and nontelomeric heterochromatin regions (gvar). These domains are: the entire 3q11.2 and 19q12; the initial part of 9q12, 19p12 and Yq12; final part of 1q12, 13p11.2, 16q11.2 and 22p11.2. They have 14 Mbp of the length, represent 44.7% of all heterochromatic CTCFbs and 0.3% of the sites has multiple binding modes. Nevertheless, similarities with euchromatic segments end at this point. We do not observe the *P*(0 < Δ ≤ 13 bp) distribution of the mammal genomes, but 7.5% of CTCFbs are 10 ± 2 bp distant each other. We do neither observe a power law nor an exponential decay in *P*(Δ) too.

Finally, we report just 20 binding sites in the chromosome 4 of fruit fly. But, this number is too small for conclusive results.

5. Discussion

The molecular basis for the four letters alignment is the hydrogen bonds of the nucleotides. The adaptation of the Peyrad–Bishop model of the DNA melting for the transcription factor binding [21] also considers the hydrogen bonds as responsible for the electronic pattern along the genome. Although the Peyrad–Bishop explains successfully the separation of the base pair under the temperature variation in polymerase chain reaction, transcription factors, as EGR1, SP1 and CTCF, do not open the double Helix in their search for binding sites. They scan the dDNA, inserting zinc fingers into the major and minor grooves of DNA and probing for π-orbital electronic patterns [7]. So, the Peyrad–Bishop cannot be applied directly for the search of the transcription factor binding site. However, the nucleotide π-orbitals have successfully been described by the extended ladder model, which interprets dDNA as semiconductor-like material [22, 23]. When we apply the extended ladder to transcription factor binding DNA sequences, patterns as in figure 1(c) appear. Again we emphasize that this semiconductor-like description is in situ condition dependent.

The method presented in this work is solvent dependent. The electronic nucleotide alignment using the extended ladder model considers the dDNA in atmosphere, low vacuum or Tris-HCl buffers [10, 20, 48, 49]. There is no consensus about the electronic transport properties of dDNA, since the experimental frameworks change the electronic properties of DNA [49]. Ethylenediaminetetraacetic acid (EDTA) or 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) buffers may induce an electric insulator effect [50]. However, at room temperature and under tris(hydroxymethyl)-aminomethane and hydrochloride salt (Tris-HCl), a traditional physiological buffer with pH = 7.382 at 37 °C [51], dDNA has a semiconductor like behavior [10, 48, 49]. This buffer may emulate the living HeLa cytosol and nucleos conditions, i.e. an aqueous solution with pH around 7.35 [52]. Under this circumstance, we may adopt the charge transport formalism to the nucleotide analysis.

The charge transport formalism adds a new layer over the nucleotide alignment. We are not restricted just to four letter pattern. As in the three ZFs of EGR1 and SP1 [10], CTCF also anchors ZFs in the nucleotides with the most stable electrons, i.e. the ground state and the lowest occupied π-orbitals. These nucleotides are at the core of the consensus motif. Numerical simulation in [29, 30] also show that the central ZFs are the most relevant nucleotide for the CTCF binding. Although the CTCF binding sites localization is not temperature dependent,
the CTCF-DNA interaction is affected by the thermal fluctuation [29]. Coarse-grained Monte Carlo of multi-Cys-His \(_2\) (mC \(_2\)H \(_2\)) zinc finger proteins as EGR1, TATAZf, transcription factor IIIA (TFIII A) and CTCF shows rotation-coupled sliding, asymmetrical roles of zinc fingers and nucleotide dependency. Furthermore, simulated mC \(_2\)H \(_2\) binds just with its central zinc fingers as we observe in CTCF.

When we examine our CTCFs with those in [4], we observe many mismatches. One source for the predict and experimental ChiP-seq CTCFs differences is the sample number for nucleotide and electronic pattern in figure 1 and for the statistics of \(P_{\text{core}}\) and \(P_{\text{flank}}\) in section pattern identification. The 23 samples do not cover all possibilities, although they catch the most common features. Actually [3, 4], also mention these additional motifs beyond the 20-mer consensus motif, positions \(-9\) to \(11\) in figure 1(d). Moreover, we are not considering homologous CTCFs [6]. The samples in figure 1(c) belong to mouse and human only, and we are defining one common CTCF pattern for them. Although we may expect a general mechanism from a common arrangement, we may foresee species depend variations in the electronic pattern.

We also introduce noise when we consider the second criteria \(S_{\text{flank}}\) for \(S_{\text{flank}}\) in pattern identification, based on the nucleotide occurrence, figure 1(e). This term plays a similar role as the background frequency correction in DNA sequence motifs. Although, this approach adds flexibility, it introduces systematic error in site prediction: the method will consider some false motifs.

CTCF can bind to dDNA in multiple ways as in shown in figure 1(c), but we combined all binding possibilities in one simple binding pattern. Indeed the literature about CTCF motifs does not consider multiple CTCF binding possibilities. However, experimental results [53], numerical simulations [29] and careful charge analysis of the tips of the zinc fingers show many viable binding arrangements. Unfortunately, the sample number in this work is too small for each individual binding configuration. Thus, we joint all, following the literature [4, 13].

The process for positioning the CTCF binding sites in the K562 uses hg38, which is a consensus sequence of nine healthy males [19], while K562 is a tumoral cell from a woman [54]. So, we are using sequences of one person to find the position in the consensus of nine others individuals. Most of sequences will be placed in the correct spot, but we expect discrepancies between these data.

Despite all limitations and criticisms about our method and the ChIP-seq technique, we have 29.8 ± 3.9% of perfect matching and 20.2% of near matching \((||ctf - k|| < 474 bp, median, Q2), 25% with intermediate misplacing (474 ≤ ||ctf - k|| < 2376 bp, third quartile, Q3) and 25% of mismatching bigger than 2374 bp. Surprisingly [9], reports similar result: 55% of successful identification, around 25% with intermediate quality and 20% with poor quality. [9] attributes the poor quality data to the low depth reading in ChiP-seq assays. [4] also reports nearly 30% of CTCFbs without the characteristic 20-mer consensus motif in ChiP-seq data and [2] reports the 20-mer motif in just over 75% of experimentally identified CTCFbs. Moreover, using limited quality data from ENCODE and only five samples of K562 ubiquitous CTCF binding sites do not help us in the evaluation of the electronic nucleotide alignments. Nevertheless, extensive tested and analyzed genome using huge ENCODE data by independent peer as [2, 4, 40] are rare. Otherwise, we may estimate the amount of misleading binding sites captured by our method from the \(P. falciparum\) and \(A. thaliana\), table 1.

There is no CTCF gene for protozoan and plants [46]. So, these binding sites are false positives generated by \(P(S)\) statistics in \(P_{\text{flank}}\) in pattern identification. Since we have around one CTCFbs in 9 kbp for mammals (human, mouse, pig and goat), we estimate from 2% to 13% of false positives in our technique considering \(P. falciparum\) and \(A. thaliana\) as negative controls. \(C. elegans\) is not a good negative test. Although this worm lost its CTCF genes along its evolution [46], this organism still hold CTCFbs.

There are three regions for the probability distribution \(P(\Delta)\) of the distance \(\Delta\) of two consecutive CTCF binding sites in human, mouse, pig, goat and fruit fly. In the first region, the binding sites appear in tandem and they are very close to each other, \(0 < \Delta \leq 13\) bp. The second region starts at \(11\) bp ∼ \(20\) bp and extends in between \(2\) kbp to \(17\) kbp. These are the domains for the power law fitting. The third domain ranges from \(2\) kbp ∼ \(15\) kbp to \(62\) kbp ∼ \(99\) kbp, when we have an exponential decay in \(P(\Delta).\) Beyond \(100\) kbp, we have visible structures in optical microscope as the high packed chromatin, coordinated by scaffold proteins in mitotic cells. But, this very large scale organization is not a topic in this paper.

In the \(0 < \Delta \leq 13\) bp domain, the number of binding sites represents 5.67% to 7.58% of the total. Further, there are always binding sites with multiple reading modes: 0.12% < \(P(\Delta = 0)\) < 0.49%. Here, we have multiple binding modes due to the molecular CTCF shape variations [29, 53], beyond the different dDNA reading modes due to the symmetries of the genomic code. The upper limit of this region is delimited by the size of the CTCF binding site. The binding site from the position \(-11\) to \(13\) in figure 1, resulting in a 24 bp of length, is compatible with the literature, where the length values 11 bp ∼ 60 bp [4, 13, 55, 56]. However, we need just 4 ∼ 5 ZFs for the CTCF-DNA attachment, using just 13 nucleotides. So, it is not surprise that this region end at \(13\) bp.

We have a power law for \(\Delta\) beyond \(13\) bp. This domain ranges from \(11\) bp ∼ \(20\) bp to \(2\) kbp ∼ \(17\) kbp,
covering between 39.98% to 62.73% of binding sites. For these distances, CTCF may interact with dDNA as well as other transcription factors due to the N and C-terminals. In human, they are respectively 150 and 265 long amino-acid sequences with distinct highly acid and basic domains [1, 57]. Further, the electronic nucleotide alignment in figure 1(b) shows consistently the presence of LUMOs and HOMOs around a binding site, reinforcing such a possibility. Although the SysZNF database provide insights about the molecular structures of the head and end of homologous CTCFs [6], detailed studies about N and C terminals interaction with DNA are rare and vague, despite experimental results [53].

The CTCF alone is not able to explain the power law. *Aedes aegypti* genome gives us a cue about the CTCF organization in these regions. The characteristic distances of $144 \pm 4$ bp and $287 \pm 5$ bp in $P(\Delta)$, table 1 and figure 2(e), reflect the action of the nucleosomes in chromatin. We need $147$ bp to wrap one nucleosome core. Moreover, the mouse chromosome Y has a recognizable $208 \pm 5$ bp distance in $P(\Delta)$, indicating a nucleosome wrapping by 147 bp with linker of 61 bp long. Indeed the mouse chromosome Y distinct distances $780 \pm 7$ bp, $931 \pm 1$ bp, $1539 \pm$ and $1927 \pm 2$ bp, figure 2(d), can be also interpreted as a chromatin with respectively 4, 5, 8 and 10 nucleosomes attached in the dDNA with two CTCF in the extremities. The CTCFs of these complexes may connect each other creating small DNA-loops. In the case of $\Delta$ ranging from $4736 \pm 4$ bp to $40565 \pm 91$ bp, figure 2(d), we have from 25 to 219 nucleosomes between the binding sites. The presence of nucleosomes around CTCF binding sites is confirmed by [4, 40] as well as in figure 2(a).

The interaction of CTCFs and nucleosomes result in a solenoidal, zig-zag ribbon or other irregular chromatin structures with a polynomial decay in $P(\Delta)$. The distribution of CTCFbs will have a cluster-like appearance, figure 2(e), troubling ChIP-seq procedures [9, 45]. Binding sites in tandem will bring ambiguities in motif alignments used in the ChIP-seq protocol too.

The distance between these cluster-like CTCFbs groups can be examined by the behavior of $P(\Delta, \lambda)$, when $\Delta$ ranges from 2 kbp $\sim 15$ kbp to 62 kbp $\sim 99$ kbp. $P(\Delta)$ becomes exponential, because the probability in finding the next CTCFbs after $\Delta$ nucleotides is $p(1 - p)^\Delta$, where $p$ is the probability of occurrence of the CTCFbs. We can approximate this expression as $p e^{n \Delta}$, since $p \ll 1$. So, we expect an exponential decay in the case of random distribution of CTCFbs. Calling $p = 1 / \lambda$, we observe an exponential behavior for $P(\Delta)$, when $\Delta$ is bigger than 2 kbp $\sim 15$ kbp.

We may illustrate the power law and the exponential decay of $P(\Delta)$ in the mouse HOXA gene cluster (cf figure 2(b)). The distance between CTCFbs inside of a cluster-like group never exceed 3058 bp and obeys a power law with $\alpha = 0.3292 \pm 0.0068$ in mouse. Nonetheless, we have a distance around 17 kbp between A4 (‘) and C5/6 as well as before A13 and after A13, and $\lambda = 19,79 \pm 0.31$ kbp in table 1.

The number of binding sites is not small in the exponential distances, ranging from 13.31% to 29.83%. In the case of *A. aegypti*, we have 61.47%. The chromatin folding process in these distances cannot be explained just with CTCF and nucleosomes. Multiple different chromosome folding for these $\Delta$ distances is mediated by non-histone proteins as cohesins, Ying and Yang 1 (YY1) and others [13, 57].

Moreover, CTCF may skip many binding sites [13]. Monte Carlo simulations show that the depletion of histones along the chromatin has influence over the folding process [58]. This is illustrated in the putative cluster-like binding sites of the genes A10, A11 and A13, indicated by diamonds in figure 2(b), where the binding sites were overlooked by CTCF. The number of binding sites localized by ChIP-seq is usually a fraction of the expected ones, with a chromosomal average of just one in 42 $\pm 12$ human ubiquitous euchromatic CTCF binding sites in the KS62 cells.

Finally, we are working with incomplete data. So, direct comparison between species must be done carefully. Major efforts from the community must be done seeking for less fragmented complete sequences. When the number of contigs are large and the size is small, most of them are too short for computing distances between binding sites and the segment number is excessive for handling them individually. The procedures described in this article are not automated yet. So, the manipulation of thousands of contigs is not viable. Furthermore, the many gaps will add noise in the probability distribution $P(\Delta)$ of the distance $\Delta$ between two consecutive binding sites. In fact, most of genomes deposited in GenBank are excessively fragmented, even those organized in chromosomes. However, new sequences deposited in GenBank overcome such limitations. The recently reviewed genomes of pig and goat have few gaps (see material), opening new perspectives to unveil the chromosomal organization in the coming years.

6. Conclusions

The CCCTC transcription factor binding sites (CTCFbs) have a characteristic $\pi$-orbital nucleotide motif. Mobile electrons are absent in the core of CTCF binding regions, i.e. we do neither observe highest occupied molecular orbitals (HOMO) nor lowest unoccupied molecular orbitals (LUMO) between ZF3 to ZF5. The CTCF may miss ZF2 or ZF6 binding with DNA. But, it cannot miss both simultaneously. There are at least three different ways to CTCF attach to the DNA. Our nucleotide alignment match with those reported in the literature.

We report 335 088 predicted CTCFbs in the whole human genome, using the electronic nucleotide
alignment. When we compare our results with the ubiquitous K562 chromatin immunoprecipitation with massively parallel DNA sequencing data (ChiP-seq), we have 29.8 ± 3.8% of matching. And, 75% of mismatches are with less than 2352 bp distance between the measured one and the predicted from our method. These 2 kbp discrepancies are expected because we use reduced number of experimental sequences for the search of our electronic pattern and the limitations of the extended ladder model. However, larger mismatches (>2 kbp) are due to ChiP-seq assay: insufficient depth of reading, the absence of the 20-mer consensus motif in the ChiP-seq data or even position of multiple CTCF motifs, each one related with one possible binding pattern.

When we combine our predicted CTCFbs and nucleosome positions, we localize 15 nucleosomes flanking CTCFbs as expected. Furthermore, the distribution of nucleosomes around CTCF reveal asymmetry, reflecting the N and C-terminous molecular differences.

We also confirm the experimental results with our theoretical study, detecting all CTCFbs in the mouse HOXA cluster.

We have studied the genomes of Mus musculus (mouse), Sus scrofa (pig), Capra hircus (goat), Drosophila melanogaster (fruit fly) and Aedes aegypti (mosquito) finding 277 027, 316 919, 264 286, 8982 and 39 777 CTCF binding sites respectively. We also analyzed Caenorhabditis elegans, Plasmodium falciparum and Arabidopsis thaliana as negative controls. Since C. elegans, protozoans and plants have no CTCF gene, there are few binding sites as expected.

The CTCFs distribution along whole genomes of studied mammals and insects, totaling 11.77 billion nucleotides, may be described as follows: for distances between 11 bp ~ 20 bp and 2 kbp ~ 17 kbp, CTCFsbs compose cluster-like groups, where the interval Δ between two consecutive binding sites obeys a power law with a coefficient α varying from 0.3292 ± 0.0068 (mouse) to 0.5409 ± 0.0064 (goat). There is no power law for the Aedes genome, but 19.9% of binding sites are at 144 ± 4 and 287 ± 5 bp distance of each other. These cluster-like CTCFbs groups are separated with a typical distance between 18.08 ± 0.52 kbp (fruit fly) to 42.1 ± 2.0 kbp (Aedes).

Acknowledgments

The authors wish to thank Lei Liu and Sujeet Kumar Mishra for the discussions about zinc fingers and CTCF. This work is supported by Conselho Nacional de Desenvolvimento Tecnológico e Científico (CNPq), Process Number 248589/2013, Brazil. We acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme Open Access Publishing, by the Baden-Württemberg Ministry of Science, Research and the Arts and by Heidelberg University. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through Grant INST 35/1134-1 FUGG. Kunhe Li would like to acknowledge funding by the Chinese Scholarship Council (CSC). This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy EXC 2181/1—390900948 (the Heidelberg STRUCTURES Excellence Cluster).

Data availability statement

All CTCF binding sites computed for this work are freely available in Heidelberg Open Research Data (HeiDATA): https://doi.org/10.11588/data/RDISCE.

ORCID iDs

Dieter W Heermann https://orcid.org/0000-0002-3148-8382

References

[1] Klenova E M, Nicolas R H, Paterson H F, Carne A F, , Heath C M, Goodwin G H, Neiman P E and Lobanenkov V V 1993 CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-my gene, is an 11-Zn-finger protein differentially expressed in multiple forms Mol. Cell. Biol. 13 7612–24
[2] Kim T H, Abdullahazar K Z, Smith A D, Ching K A, Loukinov D I, Green R D, Zhang M Q, Lobanenkov V V and Ren B 2007 Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome Cell 128 1231–45
[3] Xie X, Mikkelsen T S, Ginrike A, Lindblad-Toh K, Kellis M and Lander E S 2007 Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites Proc. Natl Acad. Sci. USA 104 7145–50
[4] Chen H, Tian Y, Shu W, Bo X and Wang S 2012 Comprehensive identification and annotation of cell-type-specific and ubiquitous CTCF-binding sites in the human genome PLoS One 7 e41374
[5] Iuchi S 2001 Three classes of C2H2 zinc finger proteins Cell. Mol. Life Sci. 58 625–35
[6] Ding G, Lorenz P, Kreutzer M, Li Y and Thiesen H-J 2009 SysZNF: the C2H2 zinc finger gene database Nucleic Acids Res. 37 D267–73
[7] Wolfe S A, Nekludova L and Pabo C 1999 DNA recognition by Cys3His4 zinc finger proteins Annu. Rev. Biophys. Biomol. Struct. 3 183–212
[8] Klug A 2010 The discovery of zinc fingers and their applications in gene regulation and genome manipulation Annu. Rev. Biochem. 79 213–31
[9] Marinov G K, Kundaje A, Park P J and Wold B J 2014 Large-scale quality analysis of published ChiP-seq data Genes, Genomes, Genet. 4 209–23
[10] Oiwa N N, Cordeiro C E and Heermann D W 2016 The electronic behavior of zinc-finger protein binding sites in the context of the DNA extended ladder model Front. Phys. 4 13
[11] Bell A C and Felsenfeld G 2000 Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene Nature 405 482
[12] Essien K, Vigneau S, Apreleva S, Singh L N, Bartolomei M S and Hannenhalli S 2009 CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcripomic features Genome Biol. 10 R131

[13] Ong C-T and Corces V G 2014 CTCF: an architectural protein bridging genome topology and function Nat. Rev. Genet. 15 234–46

[14] The ENCODE Project Consortium 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project Nature 447 799–816

[15] Edgar R, Domrachev M and Lash A E 2002 Gene expression omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 30 207–10

[16] Chen W, Liu Y, Zhu S, Green C D, Wei G and Han J-D J 2014 Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data Nat. Commun. 5 4909

[17] Zhao Y et al 2019 NucMap: a database of genome-wide nucleosome positioning map across species Nucleic Acids Res. 47 D1163–9

[18] Narendra V, Bucha P P, An D, Raviram R, Shok J A, Mazzoni E O and Reineberg D 2013 CTCE establishes discrete functional chromatin domains at the Hox clusters during differentiation Science 347 1017–21

[19] Benson D A, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman D J, Ostell J and Sayers E W 2013 GenBank Nucleic Acids Res. 41 D336–42

[20] Yoo K-H, Ha D H, Lee J-O, Park J W, Kim J, Kim J J, Lee H-Y, Kawai T and Choi H Y 2001 Electrical conduction through poly(dA)–poly(dT) and poly(dG)–poly(dC) DNA molecules Phys. Rev. Lett. 87 198102

[21] Zhu J-X, Rasmussen K O, Balatsky A V and Bishop A R 2007 Local electronic structure in the Peyer–Bishop–Holstein model J. Phys.: Condens. Matter 19 136203

[22] Senthilkumar K, Grozema F C, Guerra C F, Bickelhaupt F M, Lewis F D, Berlin Y A, Ratner M A and Siebelles I D A 2005 Absolute rates of hole transfer in DNA J. Am. Chem. Soc. 127 14984–903

[23] Mehrez H and Antranikian M P 2005 Interbase electronic coupling for transport through DNA Phys. Rev. B 71 115405

[24] Morgan M, Anders S, Lawrence M, Abouyou P, Pages H and Gentleman R 2009 ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data Bioinformatics 25 2607–8

[25] Pagès H, Abouyou P, Gentleman R and DeRoy S 2015 Biostrings: string objects representing biological sequences, and matching algorithms R Package Version 2.3.4.1

[26] Dres D, Ambrosini G, Périer R C and Bucher P 2013 EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era Nucleic Acids Res. 41 D157

[27] Miller J, McLachlan A D and Klug A 1985 Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes EMBO J. 4 1669–14

[28] Nolte J T, Conlin R M, Harrison S C and Brown R S 1998 Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex Proc. Natl Acad. Sci. USA 95 2938–43

[29] Liu L and Heermann D W 2015 The interaction of DNA with multi-Cys/His zinc finger proteins J. Phys.: Condens. Matter. 27 064103

[30] Liu L, Wade W C and Heermann D W 2015 A multiscale approach to simulating the conformational properties of unbound multi-CySH zinc fingers proteins Proteins 83 1604–15

[31] Schneider T D, Stormo G D, Gold L and Ehrenfeucht A 1986 Information content of binding sites on nucleotide sequences J. Mol. Biol. 188 415–31

[32] Schneider T D and Stephens R M 1990 Sequence logos: a new way to display consensus sequences Nucleic Acids Res. 18 6079–100

[33] Bailey T L and Elkan C 1994 Fitting a mixture model by expectation maximization to discover motifs in biopolymers Proc. 2nd Int. Conf. on Intelligent Systems for Molecular Biology pp 28–36

[34] D’haeseleer P 2006 What are DNA sequences motifs? Nat. Biotechnol. 24 423–5

[35] Setubal J and Meidanis J 1997 Introduction to Computational Molecular Biology (Boston: PWS Publishing)

[36] Mount D W 2004 Bioinformatics Sequence and Genome Analysis 2nd edn (New York: Cold Spring Harbor Laboratory Press)

[37] Zhong J, Wasson T and Hartemink A J 2014 Learning protein–DNA interaction landscapes by integrating experimental data through computational models Bioinformatics 30 2698–74

[38] Gerstein M B et al 2007 What is a gene, post-ENCODE? History and updated definition Genome Res. 17 669–81

[39] Wing R, Drew H, Takano T, Broka C, Tanaka S, Ikukura K and Dickerson R E 1980 Crystal structure analysis of a complete turn of B-DNA Nature 287 755–8

[40] Fu Y, Sinha M, Peterson C L and Weng Z 2008 The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome PLoS Genet. 4 e1000134

[41] Frankish A et al 2019 GENCODE reference annotation for the human and mouse genomes Nucleic Acids Res. 47 D766–73

[42] Kent W J, Sugnet C W, Furey T S, Roskin K M, Pringle T H, Zahler A M and Haussler D 2002 The human genome browser at UCSC Genome Res. 12 996–1006

[43] Segal E, Fondufe-Mittendorf Y, Chen L, Thalström A, Field Y, Moore K, Wang J-F Z and Widom J 2006 A genomic code for nucleosome positioning Nature 442 772–8

[44] Clarkson C T, Deeks E A, Samarista R, Mamanyusupova H, Zhurkin V B and Tel V B 2019 CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length Nucleic Acids Res. 47 11181–96

[45] Park P J 2009 ChiP-seq: advantages and challenges of a maturing technology Nat. Rev. Genet. 10 669–80

[46] Heger P, Marin B, Bartkuhn M, Schierenberg E and Wiehe T 2012 The chromatin insulator CTCF and the emergence of metazoan diversity Proc. Natl Acad. Sci. USA 109 17507–12

[47] Heger P, Marin B and Schierenberg E 2009 Loss of the insulator protein CTFC during nematode evolution BMC Mol. Biol. 10 84

[48] Cai L, Takata H and Kawai T 2000 Self-assembled DNA networks and their electrical conductivity Appl. Phys. Lett. 77 3105

[49] Taniuchi M and Kawai T 2006 DNA electronics Physica E 33 1–12

[50] de Pablo J P, Moreno-Herrero F, Cochrero J, Gómez Herrero J, Herrero P, Baró A M, Ordejón P, Soler J M and Artacho E 2000 Absence of de-conductivity in λ-DNA Phys. Rev. Lett. 85 4992

[51] Durst R A and Staples B R 1972 Tris/Tris-HCl: a standard buffer for use in the physiologic pH range Clin. Chem. 18 206–8

[52] Llopis J, McCaffrey J M, Miyawaki A, Farquhar M G and Tsien R Y 1998 Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins Proc. Natl Acad. Sci. USA 95 6803–8

[53] Filippova G N, Fagerlie S, Klenova E M, Myers C, Dehner Y, Goodwin G, Neiman P E, Collins S J and Lobanenkov V K 1996 An exceptionally conserved transcriptional repressor, CTFC, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes Mol. Cell. Biol. 16 2802–13

[54] Loizzi C and Loizzo B 1975 Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome Science 188 415–31

[55] Ohlsbø R, Renkawitz R and Lobanenkov V 2001 CTFC is a uniquely versatile transcription regulator linked to epigenetics and disease Trends Genet. 17 520–7
[56] Phillips J E and Corces V G 2009 CTCF: master weaver of the genome Cell 137 1194–211
[57] Zlatanova J and Caiafa P 2009 CTCF and its protein partners: divide and rule? J. Cell Sci. 122 1275–84
[58] Tark-Dame M, Jerabek H, Manders E M M, Heermann D W and van Driel R 2014 Depletion of the chromatin looping proteins CTCF and cohesin causes chromatin compaction: insight into chromatin folding by polymer modelling PLoS Comput. Biol. 10 e1003877