Exploring epigenetic marks by analysis of non-covalent interactions

Judith Millán[1], Alberto Lesarri[2], José A. Fernández[3], and Rodrigo Martínez[1,]*

[1] Departamento de Química, Facultad de Ciencia y Tecnología, Universidad de La Rioja, Madre de Dios, 53, Logroño, 26006 Spain.

[2] Departamento de Química Física y Química Inorgánica - IU CINQUIMA, Universidad de Valladolid, Valladolid 47011 Spain

[3] Departamento de Química Física, Facultad de Ciencia y Tecnología, Universidad del País Vasco-UPV/EHU, Barrio Sarriena s/n, Leioa, 48940 Spain.

ABSTRACT

Epigenetic marks are modest chemical modifications on DNA and histone proteins that regulate the activation or silencing of genes, through modulation of the intermolecular interactions between the DNA strands and the protein machinery. The process is complex and not always well understood. One of the systems studied in greater detail is the epigenetic mark on H3K9: the lysine 9 of the histone 3. The degree of methylation or acetylation of this histone is linked to silencing or activation of the corresponding gene, but it is not clear which effect each mark has in gene expression. We shed light on this particular methylation process using density functional theory calculations (DFT) First, we built a model consisting of a DNA double strand containing three base pairs and a sequence of three amino acids of the histone's tail. Then, we computed the modulation introduced in the intermolecular interactions by each epigenetic modification: from mono to tri-methylation and acetylation. The calculations show that while acetylation and tri-methylation results in a reduction of the DNA-peptide interaction; non-, mono-, and di-methylation increase the intermolecular interactions. Such observations compare well with the findings reported in the literature, highlighting the correlation between the balance of intermolecular forces and biological properties and advancing quantum mechanical studies of large biochemical systems at molecular level through the use of DFT methods.

* rodrigo.martinez@unirioja.es

INTRODUCTION

Chromatin is a supramolecular structure formed by stacked disk-shaped macromolecules known as nucleosomes, which constitutes a way to stabilize and store the genetic code in the cellular nucleus.[1] Each nucleosome is composed of a section of the DNA strand wrapped around a set of histones, held together by non-covalent interactions. The structure of the nucleosome can be divided into two parts: the central part or core, and an inter-nucleosome region that links adjacent cores, also known as the linker region[1]. The core is formed by two units of four different proteins (histones): H2A, H2B, H3, and H4. These are relatively small proteins that form a central domain with a well-defined structure, as can be seen in Figure 1. The structure of the histones is conserved in eukaryotic cells.[2] This means that, independently of the eukaryotic organism, 146 DNA base pairs warp 1.7 times around the octamer formed by the histones (see Figure 1). Conversely, the linker region presents variations even among different types of cells of the same organism. Thus, the total DNA length in the nucleosome varies between 160 and 240 base pairs.[1]

Gene expression requires of previous DNA liberation by a complex set of proteins.[3] So, in essence, gene expression or silencing (interruption or suppression of the expression of a gene) depends on a subtle balance of DNA-protein interactions, which may, in turn, be controlled by methylation, acetylation and other chemical modifications, known as epigenetic marks: simple but fundamental chemical modifications on DNA and histones that are key for gene promoting or silencing.[4] The first epigenetic modifications discovered, and probably the most popular ones, are cytosine and adenine methylation[5] on the DNA strands. However, further experiments demonstrated that epigenetic marks may also be found in the histones.[6]

Histone tails contain a large abundance of lysine (Lys or K) and arginine (Arg or R) and are the target of numerous post-translational modifications that modulate the histone-DNA interaction, promoting or silencing the gene. [7] Thus, acetylation of a specific Lys on the N-terminal side of the histone H3 (that ties to a specific part of the linker region) plays a fundamental role in the formation of euchromatin: an unfolded chromatin domain where DNA is available for transcription.[4] Histone acetylation neutralizes the lysine's positive charge, reducing the DNA-histone binding[7, 8] and hyperacetylation of histones favors euchromatin formation, signaling transcriptionally active regions in this way. Nevertheless,

euchromatin formation is not exclusively related to transcription, since it can be involved in other processes, such as DNA repair.[9]

Mono-, di-, or tri-methylation of lysine side chains in histones can be associated with either transcriptional activation or silencing, depending on the specific lysine residue modified and the degree of methylation.[9] These processes promote the formation of facultative heterochromatin (a compact part of DNA, but involved in gene transcription) or constitutive heterochromatin (a condensed form of DNA that acts in the gene silencing process).[4]

Methylation and demethylation reactions are part of a reversible equilibrium catalyzed by lysine methyltransferases and demethylases.[10] [11] [12] [13] [14] [15] The enzymes involved can work in a distributive manner, where a pre-existing mono- or di-methylated state must be present, or in a progressive way, where a conversion of an unmodified substrate to a tri-methylation state would occur.[16] Moreover, the substrates for methyltransferases or demethylases can be either free histones or assembled chromatin.[16] Thus, all of these agents (enzymes, free or condensed histones, various methylation states, etc.) participate in a dynamic process[17] that can result in gene activation or silencing. Multiple steps of this complex process are still unresolved.[16]

In this work we have focused on the lysine 9 of histone 3 (H3K9), since it plays a double duty: while its acetylation seems to signal gene activation (H3K9ace), tri-methylation results in gene silencing. The substitution of a hydrogen atom by an acetyl moiety, with the subsequent cancellation of the positive charge in lysine's ε-amino group, surely causes a conformational change that reinforces the interaction with other macromolecules. Charge cancellation results in a weaker histone-DNA interaction, enabling the interaction with other proteins, and the recruitment of proteins necessary for the next step in transcription.[18], [19], [20], [21]

On the other hand, methylation of H3K9 seems to be by far a more complex process: different methylations states on H3K9 act promoting or silencing the gene. Mono-methylated lysine (H3K9me1) has been related to gene activation,[22] whereas di- and tri-methylated lysine (H3K9me2 and H3K9me3, respectively) seem to be involved in gene repression,[7] since they are specifically recognized by the chromodomain of the heterochromatin protein 1 (HP1),[7] a non-histone protein with versatile functions.[23] However, the specific role of each mark is

not clear, since there are also data about correlation between gene activation and the presence of H3K9me2.[7]

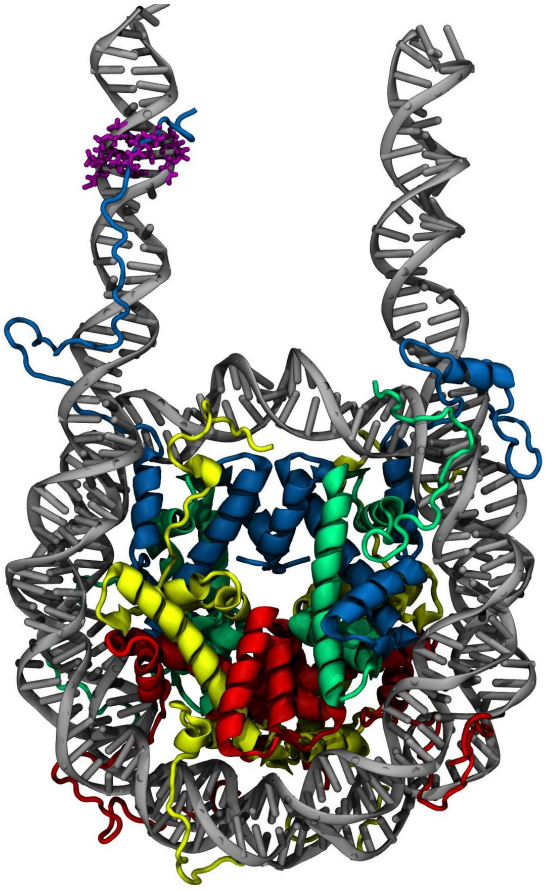

Figure 1: Human nucleosome PDB ID 1kx5.[24] H3 histones in blue, H4 in green, H2A in yellow, H2B in red, and DNA in gray. The system studied is represented in purple.

To shed light on this process, we take a reductionist approach, building a model composed by a DNA segment containing three base-pairs and three amino acids in the N-terminal side of H3. The system includes H3K9, H3R8 (previous arginine in histone H3), recognized as a linker with the DNA minor groove in a human nucleosome,[25] and the H3A7 amino acid (previous alanine to H3R8). This polypeptide (H3A7-H3R8-H3K9, note the linker amino acid in the middle) has been demonstrated to interact with the three pairs of bases CAG-GTC, using molecular dynamics simulations of a human nucleosome[25] (see purple structure in Figure 1). Thus, these three base pairs complete the system studied. Using density functional theory (DFT) calculations, we explored the changes in structure, non-covalent interactions (in particular the intermolecular hydrogen bonds) and interaction energy that each epigenetic

modification (acetylation, mono-, di-, and tri-methylation) introduces in the system. Our results indicate a clear correlation between DNA-peptide interaction and activation/repression of the gene, highlighting the importance of the investigation of intermolecular interactions through accurate quantum mechanical methods and exploiting DFT to advance to large biological systems.

COMPUTATIONAL METHODS

The initial geometry of the three amino acids and the three base pairs was extracted from a snapshot of the trajectory calculated in a 1 microsecond molecular dynamics simulation of a human nucleosome.[25] For this geometry, we verified with Visual Molecular Dynamics (VMD)[26] that the hydrogen atom of the guanidine ion's N-H group of H3R8 presented a distance of 1.95 Å with the N3 of the nearest adenosine nucleobase. This specific interatomic interaction has been identified in silico,[27] verifying that the initial geometry is biologically feasible. We added $-CH_3$ moieties and H atoms with the Tmolex v4.4 program[28] to the initial structure to equilibrate the unbalanced peptide bonds and DNA phosphates, respectively. Compensation of the negative charge on the phosphate groups by adding methyl groups is important to avoid unrealistic electrostatic attraction with positively charged amino acid lateral chains. The resulted structure CH_3-H3A7-H3R8-H3K9-CH_3/CAG-CTG (252 atoms), denoted as H3K9, is depicted in Figure 2.

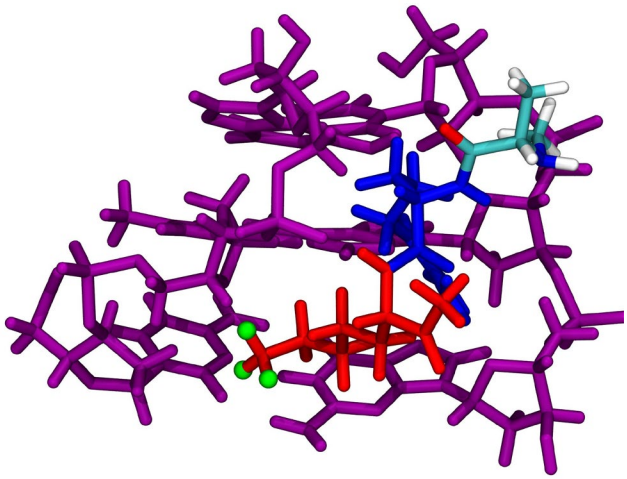

Figure 2: The non-methylated H3K9 system consisting of the CH_3-H3A7-H3R8-H3K9-CH_3/CAG-CTG sequence. DNA bases in purple, H3K7 colored according to the atom type, H3R8 in blue, H3K9 in red. The methylation points on H3K9 are represented as green dots.

DFT theory provides a wide variety of computationally effective methods that permit an accurate characterization of large systems. [29] Here, the potential energy of the system was minimized at the BLYP-D3(BJ)/def2-TZVPP level, with implicit consideration of water using the COSMO approach,[30] as implemented in TURBOMOLE 7.2.[31] The empirical dispersion correction D3(BJ)[32] [33] was used, since the BLYP functional does not include the dispersion effects which are important for non-covalent interactions.[29] To reduce the computational time, the parallel version of RIDFT[34] TURBOMOLE's module and the RI-JK approach[35] were used.

Once the minimum energy geometry of the H3K9 system was obtained, the different epigenetic marks were inserted with Tmolex v4.4, by modifying the ε-amino group of H3K9 amino acid to get H3K9ace, H3H9me1$_a$, H3K9me2$_a$, and H3K9me3 systems. Moreover, the H3K9me1$_a$ and H3K9me2$_a$ fragments were subsequently rotated to consider different orientations of $–(NH_2CH_3)^+$ and $–(NH(CH_3)_2)^+$ moieties (H3K9me1$_b$, H3K9me1$_c$, H3K9me2$_b$, and H3K9me2$_c$).

Finally, the 9 structures representing acetylated (H3K9ace) non-methylated (H3K9), the three rotamers of mono-methylated (H3K9me1$_a$, H3K9me1$_b$, and H3K9me1$_c$), the three rotamers of di-methylated (H3K9me2$_a$, H3K9me2$_b$, and H3K9me2$_c$), and tri-methylated (H3K9me3) systems were optimized at the same level of theory than H3K9.

Once all of the optimized structures were obtained, the root-mean-square deviations (RMSD) for the atomic positions were calculated to find correlations for the different epigenetic marks. These calculations were performed considering only the atoms present in all of the structures (i.e. not considering the atoms of the epigenetic modification). Moreover, the interaction energy between the DNA and polypeptide fragments in the optimized structures was calculated by subtracting the energy of the fragments with the geometry they have in the complex from the energy of the complex. If the geometry of the optimized fragments were used instead, the binding energy would be obtained. However, the relevant data here is the interaction strength between the fragments, while the separation of the peptide from the DNA strand and re-optimization of its structure would result in an important structural rearrangement that would not enable computation of the true interaction energy.

Calculation of the interaction energy usually requires to take into account the basis set superposition error (BSSE).[36] In the case of solvated systems, as the one studied here, the BSSE is overestimated due to effects related to solvation, resulting in energy corrections greater than even the interaction values, which is clearly wrong.[37] For this reason, we used the BSSE error calculated in gas phase but using the solvated geometry to correct the solvated interaction energies. The BSSE calculation is automated in TURBOMOLE following the counterpoise procedure described in [36], where the energy of each monomer is calculated with the dimer's complete basis sets but considering the electron and nuclear charges of the other monomer as zero (ghost basis sets). Thus, the mutual overlapping of the basis sets of the monomers in the dimer structure is compensated when the interaction energy is calculated.

For H3K9me1 and H3K9me2 complexes, where three rotamers are involved, the interaction energy was calculated as a weighted average of the interaction energies of the corresponding rotamers. The weight for each structure was calculated by the Boltzmann distribution at 298 K (the results at 36 ºC, the typical human body temperature, would differ only by 0.04 kJ·mol^{-1}):

$$w_i = \frac{e^{-E_i/k_B \cdot T}}{\sum_{j=i}^{3} e^{-E_j/k_B \cdot T}} \qquad [1]$$

Next, the interaction energies were also calculated at four DFT levels to check the consistency: B3LYP-D3(BJ)/def2-TZVPP, M06-l/def2-TZVPP, M06/def2-TZVPP, and M06-2x/def2-TZVPP. The structure used was that resulting from the optimization at BLYP-D3(BJ)/def2-TZVPP level. In this way, the optimization steps at the more computationally demanding levels are avoided. In the Results and Discussion section the validity of this procedure is confirmed.

With these calculations different DFT approaches, all available in TURBOMOLE, are tested. The selection of functionals covers representative GGA (BLYP), hybrids (B3LYP), meta-GGA (M06-l) and meta-hybrid (M06, and M06-2x) functionals.

Finally, the wave functions of the systems optimized at BLYP-D3(BJ)/def2-TZVPP level, also generated by TURBOMOLE, were used for the analysis of the non-covalent interactions

(NCI) between the investigated polypeptide and the DNA strand using topological methods based on the electron density.

According to the NCI approach,[38], [39] hydrogen bonding, π-π stacking and van der Waals interactions are characterized by a low or very low electron density between the interacting atoms or fragments and they can be characterized by the reduced density gradient:

$$s(\rho) = \frac{|\nabla\rho|}{2\cdot\left(3\cdot\pi^2\right)^{1/3}\cdot\rho^{4/3}} \qquad [2]$$

where ρ is the electron density, $\nabla\rho$ its gradient, and $s(\rho)$ the reduced density gradient. This last property shows sharp peaks when the electron density is low. Moreover, the sign of the second eigenvalue of ρ's Hessian allows one to classify the interactions as attractive, weak, or repulsive. All together can be depicted in very useful molecular representations by NCIplot,[38], [39] using the location and character (attractive, weak, or repulsive) of the non-covalent interactions to explain the preferred geometry of the aggregates or their molecular conformations (see e.g. [27] and [40]).

Recently, the Independent Gradient Model,[41], [42], [43] $|\delta\rho^{IGM}|$, has been developed as a complementary tool for the analysis of the non-covalent interactions. The IGM, also starting from the wave function of the system, allows one to target a specific atom-pair interaction in a molecule, either covalent or non-covalent. So, the plots obtained with the IGMPlot[41],[42], [43] software isolate the exclusive non-covalent interactions between two fragments, e.g. DNA and histone in our investigation.

RESULTS AND DISCUSSION

The optimized structure of the H3K9 system together with the results from the NCI and IGM analysis are depicted in Figure 3. The NCI calculation gives full-of-information pictures as can be seen in the left part of the figure. Weak interactions (green surfaces) are ubiquitous, as they take place between aromatic rings (π-π stacking), aliphatic side chains (interactions due to dispersive forces) or between aliphatic groups and aromatic rings (mainly C-H···π interactions). Furthermore, the system is surrounded by a wide green surface that represents

the implicit solvent effect. The red lobules represent strong repulsive zones and are detected at the center of the ring in sugar units and nucleobases.

Although weak interactions are the most abundant, strong stabilizing interactions, as hydrogen bonds, are also detected and represented as blue disks in the NCI plots. These strong interactions stabilize the DNA duplex but they are also important in the formation of the DNA-protein aggregate. For this reason, they deserve special attention.

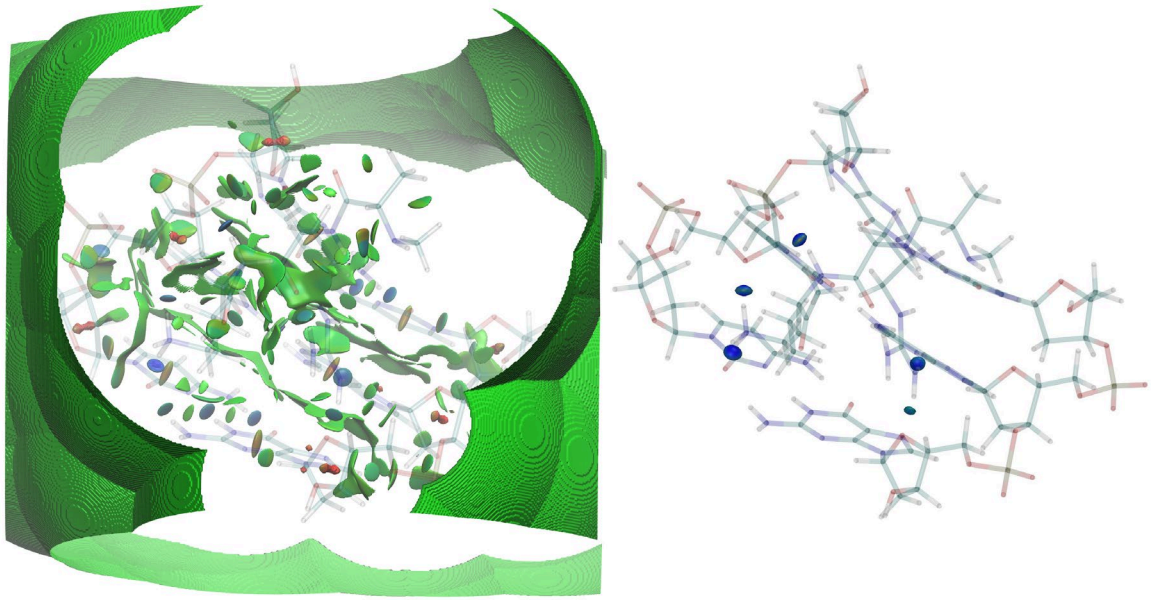

Figure 3: Complete representation of non-covalent interactions in the H3K9 system with the NCIplot (left) and IGMplot (right) methods. Hydrogen bonds are represented as blue disks.

On the other hand, the IGMplot method gives a clear representation of the interactions between DNA and the peptide fragment, since it is possible to differentiate the intramolecular interactions using the wave function of the complex.[42] It is clear from Figure 3 (right) that 5 hydrogen bonds (H-bonds) participate in the stabilization of the DNA-tripeptide aggregate.

Lysine and arginine's tails form 2 H-bonds with the DNA fragment, respectively, whereas the fifth one is formed by an NH group of the alanine's backbone. Detailed pictures of these interactions may be found in Figure S1, while their geometrical properties (distances and angles) are collected in Table S1. The same detailed pictures for the rest of the optimized

structures are collected in Figures S2-S9, and their structural parameters are also shown in Table S1.

In all of the H-bonds detected, an –NH moiety in the peptide acts as proton-donor. However, a larger diversity was found in the acceptor atom, since it can be a nitrogen from guanine or adenine; or an oxygen atom from cytosine, deoxyribose ring, or even from the phosphate fragment. The geometrical characteristics of all of the H-bonds detected (Table S1) allow us to classify them as moderate or weak bonds.[44] The values reported for these systems are in agreement with previous publications on the typical values (1.809-2.042 Å) for hydrogen bonds between amino acids and DNA basis.[27] The guanidinium ion on the arginine tail forms two H-bonds in all cases with the exception of H3K9me1$_b$ system, where a single H-bond is formed (Figure S3, IGM results). The acceptor atom in these two bonds are a nitrogen atom of the adenine base and an oxygen in the deoxyribose ring of guanine. When the acceptor atom is the adenine's N, the H-bond distances are, in general, shorter than the in the N-H ··O bond (~2 Å), showing an N-H···N angle around 170º (165.6º - 171.7º), which is wider than in the N-H···O case (152.1º – 162.5º).

A third interaction in this amino acid can be observed in the NCI analysis (Figures S1-S9). However, it is an intramolecular H-bond and it is not detected in the IGM results.

The nitrogen located at lysine residues participates in the H-bond pattern in different ways depending on the degree of substitution and on the orientation of the rotamer. In H3K9ace (Figure S9 and Table S1), the H-bond with the most unfavorable structural parameters was found (*d*= 1.93 Å, angle= 157.2º). The rest of the systems present bonding distances between 1.88 and 1.68 Å and angles in the 170.3 º - 160.5º interval. In the case of H3K9me2$_b$ and H3K9me2$_c$ rotamers (Figures S5-S6), the acceptor atom in the DNA bases is the N atom from guanine that also acts as proton-donor in the hydrogen bond stablished with its partner base (cytosine), playing a double role: proton-donor and acceptor.

Finally, the -NH group of the alanine backbone and an oxygen located in a DNA's phosphate form a H-bond in all of the systems studied that ranges from 1.66 to 1.75 Å and sometimes, as in the H3K9me1$_b$ and H3K9me2$_b$ cases, it is close to linearity (See Table S1 and Figures S3 and S6). This last hydrogen bond increases the rigidity of the peptide's backbone and

complements the implication of the three amino acids in the formation of a H-bond pattern with the DNA fragment.

The RMSDs among the different aggregates were calculated to study the effect produced in the structure of the peptide and the DNA template by the epigenetic marks. Thus, different RMSD values were obtained considering the whole systems, the DNA strand, and the tripeptide, with and without the hydrogen atoms. Note that the deviations were calculated considering only the common atoms in all the structures, i.e. excluding the different epigenetic modifications. These results are presented in Tables S2-S7.

Finally, the results of the calculated interaction energies for the nine optimized complexes are collected in Table 1.

Table 1: Interaction energy ($kJ \cdot mol^{-1}$) calculated at different DFT levels for the systems studied. In all cases the def2-TZVPP basis set was used.

	BLYP-D3(BJ)	**B3LYP-D3(BJ)**	**M06-l**	**M06**	**M06-2x**
H3K9ace	-180,13	-177,36	-138,45	-126,11	-116,61
H3K9	-202,38	-199,57	-151,84	-142,29	-134,22
H3K9me1a	-206,47	-203,18	-155,37	-146,02	-135,29
H3K9me1b	-183,62	-205,79	-168,69	-158,25	-147,22
H3K9me1c	-210,93	-195,71	-155,44	-149,50	-134,18
H3K9me2a	-193,69	-206,45	-167,90	-164,08	-141,83
H3K9me2b	-205,20	-189,29	-144,00	-135,16	-125,32
H3K9me2c	-167,06	-178,48	-135,06	-132,03	-112,96
H3K9me3	-170,49	-166,18	-128,18	-121,15	-106,75

The present protocol limiting the geometry optimization to BLYP-D3(BJ)/def2-TZVPP level is justified in the considerable computational effort necessary to obtain all of the optimized structures for all the DFT levels analyzed. To validate this procedure the structures of H3K9 and H3K9ace systems were optimized at M06, and M06-2x levels and their interaction energies calculated. The comparison with the interaction energies calculated with the BLYP/def2-TZVPP-D3(BJ) structure show differences below 4.5 $kJ \cdot mol^{-1}$ (see Table S8),

which means that the difference between the values obtained with the two methods are well within the computational error. Also, the maximum RMSD between the structures optimized through the different methods is 0.258 Å (see Table S9) which is also a negligible difference. On the light of these results, we consider that the proposed procedure is accurate for the present purposes and overcomes the necessity of the optimization at the higher DFT levels.

Concerning the BLYP-D3(BJ) results, H3K9, H3K9me1$_a$, H3K9me1$_b$, H3K9me2$_a$, and H3K9me2$_b$ systems present an interaction energy close to -200 kJ·mol^{-1}, whereas the values in H3K9me1$_c$, H3K9me2$_c$, H3K9me3, and H3K9ace systems are between -170 ~ -180 kJ·mol^{-1}, indicating that it is easier to disaggregate these latter systems. A similar conclusion can be extracted from the B3LYP-D3(BJ) results, with the exception of H3K9me1$_c$, that can be included in the group of more stable aggregates.

In the case of Minnesota functionals, i.e. M06-l, M06, and M06-2x, the interaction energy values are smaller, but the above-mentioned grouping is also observed. Thus, in all of these results H3K9, H3K9me1$_a$, H3K9me1$_b$, H3K9me1$_c$, and H3K9me2$_a$ aggregates are more stable than the H3K9ace, H3K9me2$_b$, H3K9me2$_c$, and H3K9me3 ones. This aggrupation can clearly be observed in Figures S10-S14. where the interaction energy values are represented for each structure and method.

No direct correlation between the number or geometry of the intermolecular H-bonds established in the systems and interaction energies has been found. For example, H3K9me1$_b$ is more stable that H3K9me1$_c$ in all of the computational methods tested, but 4 hydrogen bonds were detected in this latter dimer, whereas there are 3 in the former (see Table 1). This observation evidences that the weak but abundant interactions (green disks in NCIplot results: Figures S1-S9) play a fundamental role in the binding of biological aggregates and work together with the hydrogen bonds in the stabilization of the clusters.

Concerning the RMSD values, no correlation between these data and interaction energies was found either. Taking H3K9 as the reference structure (e.g. first row in Table S1), a certain correlation between the RMSD and stability was observed for H3K9 and H3K9me1$_a$ in all DFT results: they present very similar optimized structures, RMSD= 0.112 Å, and interaction energy (less than ~4 kJ·mol^{-1} of energy differences for all theoretical levels). However, this correlation is not preserved for other structures and methods, where the same energy

difference is observed but higher RMSD value is obtained, e.g. H3K9 and H3K9me2$_b$ at BLYP level where the RMSD is 0.476 Å and their interaction energy difference is 2.81 kJ·mol^{-1} or H3K9 and H3K9me2$_b$ at M06-2x level where the RMSD is 0.684 Å and the interaction energy difference is almost null (0.04 kJ·mol^{-1}). The same is applicable to lysine acetylation, which causes a similar distortion on the H3K9 structure than di-methylation, but H3K9ace dimer is less stable than H3K9me2$_a$ and H3K9me2$_b$ in all of the DFT results.

In any case, all of the RMSD values exposed in the tables are smaller than 1 Å, indicating that the epigenetic marks do not induce important geometrical changes on the structure of the aggregate. Nevertheless, some interesting information can be extracted from the graphical representation of the atom-by-atom RMSD.

Representation of the changes in the peptide (Figure S15) highlights that the modifications mainly affect to the lysine and arginine side chains (atoms in red and white), whereas the polypeptide's backbone remains almost unchanged (atoms in blue), surely due to de above-mentioned strong hydrogen bond formed by the tripeptide's side chain.

In the case of DNA strands (Figure S16), the position of the cytosine at the low-left part of the representations presents the largest deviation in comparison with the H3K9 system. This was somehow expected, since this base is close to the lysine modifications. Nevertheless, other DNA zones also present deviations due to the different modifications, contrasting with the observed rigidity of the peptide side chain (see Figure S9).

The fact that the effect of the epigenetic marks in the interaction energy may be divided into two groups can be observed in Figure 4, where the weighted average (Eq. [1]) of the corresponding rotamers is represented.

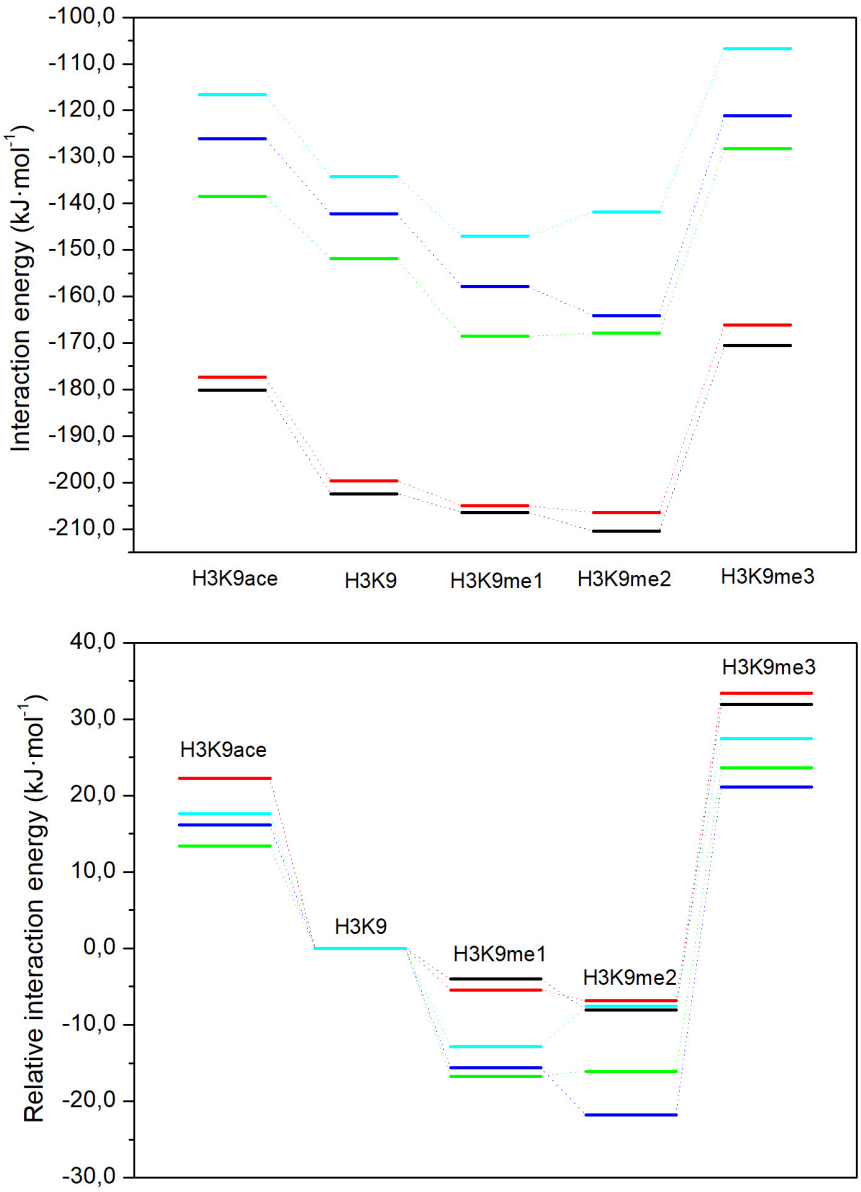

Figure 4. Interaction energies of all of the studied systems in absolute values (upper panel) and relative to H3K9 system (lower panel) at BLYP-D3(BJ) (black), B3LYP-D3(BJ) (red), M06-l (green), M06 (blue), and M06-2x (cyan) levels. The data for mono- and di-methylated systems are Boltzmann averaged

Despite the gap between results from the Minnesota functionals and those from BLYP-D3(BJ) and B3LYP-D3(BJ), there is general trend: it is easier to disaggregate H3K9ace or H3K9me3 systems than H3K9. On the other hand, mono and dimethylation lead to an averaged interaction energy similar to the unmodified system in the case of BLYP and

B3LYP methods. At M06-l and M06 mono and demethylated aggregates are more stable than the unmodified one. Finally, at M06-2x level, the monomethylated aggregate is the most stable and the dimethylated one presents an interaction energy similar to the H3K9 system. This behavior is more evident when the energy of the unmodified H3K9 system is taken as reference, as in the representation in the lower panel of Figure 4.

We propose that the lower stability of H3K9ace and H3K9me3 aggregates could be related to gene activation and suppression because this situation would favor the histone-DNA separation, a key step in the recruitment process of other agents (proteins, enzymes, etc.) by the histone tails. Such separation would trigger the gene activation process in the case of the acetylation[18-21] or silencing in the case of tri-methylation.[7, 16] Thus, we postulate that these recognition processes are energetically favored by low interaction energies (in absolute values) and of course, by geometric, steric, and charge complementarity aspects with the respective partners.

Conversely, H3K9, H3K9me1, and H3K9me2 would be intermediates between gene activation and repression, since H3K9 is the substrate of acetyltransferases[7] but also of methyltransferases, as H3K9me1 and H3K9me2 are.[16, 17]

In this scheme, the environmental conditions inside the nucleus (for example, it has been observed that oxidative stress alters histone modification and DNA methylation[45]), the presence of other agents such as noncoding RNAs[12] or transcription factors[16] would regulate the enzymatic processes through the different intermediates to insert the H3K9ace mark and activate the gene or the H3K9me3 mark and silence it.

This idea of methylated intermediates agrees with several studies where the relative abundance of lysine in different methylation states varies with mutations on demethylase enzymes.[46] Also, this hypothesis of the intermediate would also justify the (apparently) inconsistent data on the detection of H3K9me2 mark in the biological media both in gene silencing and in activation.[7]

Nevertheless, as mentioned in the Introduction, there are evidences that show that H3K9me2 modification also binds to proteins related with gene silencing.[7] This affinity also fits with the interaction energy values of H3K9me2$_b$ and H3K9me2$_c$ rotamers (Figure 6a). Thus,

H3K9me2$_c$ is energetically available for the recruitment, according to the results from all of the DFT levels tested. H3K9me2$_b$ is also available, with the exception of the calculation at BLYP level. That only these rotamers show this "instability" and, once the energies are averaged, the dimethylated system is "stable" would explain why HP1 presents affinity to H3K9me2 but lower than the affinity reported for H3K9me3.[17]

CONCLUSIONS

Acetylation and methylation are epigenetic modifications that promote the activation or silencing of genes. Their effect is related to the formation of eucromatin or heterochromatin domains in the chromosome in a process where a multitude of different agents and factors are involved. The epigenetic marks on histones, simple modifications mainly located in their tails, seem to control this process enabling the union with other proteins or enzymes and promoting other actions.

In this work we have studied the effect of different epigenetic modifications on the aggregate formed by a tripeptide of sequence H3A7-H3R8-H3K9 and its corresponding three pair of DNA bases, CAG-GTC, extracted from a real nucleosome. We have calculated and compared the non-covalent interactions of the DFT-optimized geometry considering water as implicit solvent of the systems, where H3K9 is unmodified, mono-, di-, tri-methylated, and acetylated. We analyzed the hydrogen bond network that produces the stabilization of these aggregates, calculated the RMSD induced by the epigenetic modifications, and estimated the interaction energy between the DNA and tripeptide fragments.

Differences on the hydrogen bond networks and RMSD values did not show a clear correlation neither with the interaction energy nor with the observed biological effect. Conversely, the most clarifying variations have been found in the averaged interaction energies.

H3K9ace and H3K9me3, key modifications in protein recruitment, showed interaction energy values lower than the rest of the marks, highlighting that their disaggregation is easier. Thus, in the context that acetylation, methylation, and reverse processes take part in a dynamical equilibrium.[16] We propose that H3K9, H3K9me1, and H3K9me2 are

intermediates between the species H3K9ace and H3K9me3 that actually trigger the transcription or repression of a certain gen.

It is interesting that using a reductionist approach a clear correlation was found between intermolecular interactions and biological effects. Other factors are surely involved, such as methylation/acetylation of additional amino acids. For example, other lysine amino acids present similar behavior,[7] But we hope that this proposed mechanism can open a future field of investigation where computational chemistry can help solving the epigenetic puzzle. In this sense, this work highlights the importance of using quantum mechanical tools, in particular DFT, to advance in the analysis of molecular mechanism behind the highly complex process of molecular recognition and duplication.

ACKNOWLEDGMENTS

The authors thank the MINECO-FEDER (PGC2018-098561-B) and Basque Government (IIT62-19) for the financial support. We also thank to Drs. E. Hénon and J.C. Boisson from the University of Reims for the help provided to apply IGMplot program to the studied systems. This work used the Scientific Computing Service of the UPV/EHU (IZO-SGI SGIker (UPV/EHU) and Beronia cluster (Universidad de La Rioja), which is supported by FEDER-MINECO grant number UNLR-094E-2C-225.

REFERENCES

[1] R. K. McGinty, S. Tan, *Chemical Reviews* **2015**, *115*, 2255-2273.

[2] G. Felsenfeld, M. Groudine, *Nature* **2003**, *421*, 448-453.

[3] H. Boeger, J. Griesenbeck, J. S. Strattan, R. D. Kornberg, *Molecular Cell* **2003**, *11*, 1587-1598.

[4] S. I. S. Grewal, S. T. Jia, *Nature Reviews Genetics* **2007**, *8*, 35-46.

[5] R. D. Hotchkiss, *Journal of Biological Chemistry* **1948**, *175*, 315-332.

[6] K. Murray, *Biochemistry* **1964**, *3*, 10-+.

[7] L. Vanzan, A. Sklias, Z. Herceg, R. Murr, in *Handbook of Epigenetics (Second Edition)*, "Second Edition" ed. (Ed.: T. O. Tollefsbol), Academic Press, **2017**, pp. 25 - 46.

[8] R. T. Simpson, *Cell* **1978**, *13*, 691-699.

[9] J. V. Tjeertes, K. M. Miller, S. P. Jackson, *Embo Journal* **2009**, *28*, 1878-1889.

[10] H. L. Schubert, R. M. Blumenthal, X. D. Cheng, *Trends in Biochemical Sciences* **2003**, *28*, 329-335.

[11] Y. Tsukada, J. Fang, H. Erdjument-Bromage, M. E. Warren, C. H. Borchers, P. Tempst, Y. Zhang, *Nature* **2006**, *439*, 811-816.

[12] J. Y. Cui, Z. D. Fu, J. Dempse, in *Toxicoepigenetics* (Eds.: S. D. McCullough, D. C. Dolinoy), Academic Press, **2019**, pp. 31 - 84.

[13] L. Morera, M. Lubbert, M. Jung, *Clinical Epigenetics* **2016**, *8*.

[14] J. Li, S. Zhu, X. X. Ke, H. Cui, *Biomed Rep* **2016**, *4*, 293-299.

[15] D. Han, M. X. Huang, T. Wang, Z. P. Li, Y. Y. Chen, C. Liu, Z. J. Lei, X. Y. Chu, *Cell Death & Disease* **2019**, *10*.

[16] T. Jenuwein, *Febs Journal* **2006**, *273*, 3121-3135.

[17] Y. Shi, J. R. Whetstine, *Molecular Cell* **2007**, *25*, 1-14.

[18] K. Karmodiya, A. R. Krebs, M. Oulad-Abdelghani, H. Kimura, L. Tora, *Bmc Genomics* **2012**, *13*.

[19] L. A. Gates, J. J. Shi, A. D. Rohira, Q. Feng, B. K. Zhu, M. T. Bedford, C. A. Sagum, S. Y. Jung, J. Qin, M. J. Tsai, S. Y. Tsai, W. Li, C. E. Foulds, B. W. O'Malley, *Journal of Biological Chemistry* **2017**, *292*, 14456-14472.

[20] A. H. Tencer, K. L. Cox, L. Di, J. B. Bridgers, J. Lyu, X. D. Wang, J. K. Sims, T. M. Weaver, H. F. Allen, Y. Zhang, J. Gatchalian, M. A. Darcy, M. D. Gibson, J. Ikebe, W. Li, P. A. Wade, J. J. Hayes, B. D. Strahl, H. Kono, M. G. Poirier, C. A. Musselman, T. G. Kutateladze, *Cell Reports* **2017**, *21*, 455-466.

[21] K. Struhl, *Genes Dev* **1998**, *12*, 599-606.

[22] A. Barski, S. Cuddapah, K. R. Cui, T. Y. Roh, D. E. Schones, Z. B. Wang, G. Wei, I. Chepelev, K. J. Zhao, *Cell* **2007**, *129*, 823-837.

[23] J. C. Eissenberg, S. C. R. Elgin, *Trends in Genetics* **2014**, *30*, 103-110.

[24] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, T. J. Richmond, *Journal of Molecular Biology* **2002**, *319*, 1097-1113.

[25] A. K. Shaytan, G. A. Armeev, A. Goncearenco, V. B. Zhurkin, D. Landsman, A. R. Panchenko, *Journal of Molecular Biology* **2016**, *428*, 221-237.

[26] W. Humphrey, A. Dalke, K. Schulten, *Journal of Molecular Graphics & Modelling* **1996**, *14*, 33-38.

[27] J. Gonzalez, I. Banos, I. Leon, J. Contreras-Garcia, E. J. Cocinero, A. Lesarri, J. A. Fernandez, J. Millan, *Journal of Chemical Theory and Computation* **2016**, *12*, 523-534.

[28] C. Steffen, K. Thomas, U. Huniar, A. Hellweg, O. Rubner, A. Schroer, *Journal of Computational Chemistry* **2010**, *31*, 2967-2970.

[29] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. Grimme, *Physical Chemistry Chemical Physics* **2017**, *19*, 32184-32215.

[30] A. Klamt, G. Schuurmann, *Journal of the Chemical Society-Perkin Transactions 2* **1993**, 799-805.

[31] S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, A. Hellweg, B. Helmich-Paris, C. Holzer, U. Huniar, M. Kaupp, A. M. Khah, S. K. Khani, T. Müller, F. Mack, B. D. Nguyen, S. M. Parker, E. Perlt, D. Rappoport, K. Reiter, S. Roy, M. Rückert, G. Schmitz, M. Sierka, E. Tapavicza, D. P. Tew, C. v. Wüllen, V. K. Voora, F. Weigend, A. Wodyński, J. M. Yu, *The Journal of Chemical Physics* **2020**, *152*, 184107.

[32] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *Journal of Chemical Physics* **2010**, *132*.

[33] S. Grimme, S. Ehrlich, L. Goerigk, *Journal of Computational Chemistry* **2011**, *32*, 1456-1465.

[34] M. Von Arnim, R. Ahlrichs, *Journal of Computational Chemistry* **1998**, *19*, 1746-1757.

[35] F. Weigend, *Physical Chemistry Chemical Physics* **2002**, *4*, 4285-4291.

[36] S. F. Boys, F. Bernardi, *Molecular Physics* **2002**, *100*, 65-73.

[37] K. Riley, J. Vondrasek, P. Hobza, *Physical Chemistry Chemical Physics* **2007**, *9*, 5555-5560.

[38] E. R. Johnson, S. Keinan, P. Mori-Sanchez, J. Contreras-Garcia, A. J. Cohen, W. T. Yang, *Journal of the American Chemical Society* **2010**, *132*, 6498-6506.

[39] J. Contreras-Garcia, E. R. Johnson, S. Keinan, R. Chaudret, J. P. Piquemal, D. N. Beratan, W. T. Yang, *Journal of Chemical Theory and Computation* **2011**, *7*, 625-632.

[40] C. Perez, I. Leon, A. Lesarri, B. H. Pate, R. Martinez, J. Millan, J. A. Fernandez, *Angewandte Chemie-International Edition* **2018**, *57*, 15112-15116.

[41] C. Lefebvre, G. Rubez, H. Khartabil, J. C. Boisson, J. Contreras-Garcia, E. Henon, *Physical Chemistry Chemical Physics* **2017**, *19*, 17928-17936.

[42] C. Lefebvre, H. Khartabil, J. C. Boisson, J. Contreras-Garcia, J. P. Piquemal, E. Henon, *Chemphyschem* **2018**, *19*, 724-735.

[43] M. Ponce-Vargas, C. Lefebvre, J. Boisson, E. Henon, *Journal of Chemical Information and Modeling* **2020**, *60*, 268-278.

[44] G. A. Jeffrey, W. Saenger, *Hydrogen bonding in biological structures*, Springer-Verlag, **1991**.

[45] Y. M. Niu, T. L. DesMarais, Z. H. Tong, Y. X. Yao, M. Costa, *Free Radical Biology and Medicine* **2015**, *82*, 22-28.

[46] B. D. Fodor, S. Kubicek, M. Yonezawa, R. J. O'Sullivan, R. Sengupta, L. Perez-Burgos, S. Opravil, K. Mechtler, G. Schotta, T. Jenuwein, *Genes & Development* **2006**, *20*, 1557-1562.