Determination of hydrogen bonds in Gromacs: new implementation to overcome the limitation

S. V. Gorelov,[*,†,‡] A. I. Titov,[*,†] and A. V. Shvetsov[*,†,‡,¶]

†*Petersburg Nuclear Physics Institute, Gatchina, Russia*

‡*Peter the Great St.Petersburg Polytechnic University, St.Petersburg, Russia*

¶*National Research Centre "Kurchatov Institute", Moscow, Russia*

E-mail: gorelov_sv@pnpi.nrcki.ru; titov_ai@pnpi.nrcki.ru; shvetsov_av@pnpi.nrcki.ru

Abstract

This work describes a new software algorithm associated with determination of hydrogen bonds in biomacromolecules and their environment. The already existing algorithm for determining hydrogen bond networks in Gromacs has a number of impenetrable limitations in the analysis of structures and trajectories. The new implementation of the algorithm for determining hydrogen bond networks in the form of a native Gromacs trajectory analysis module allows to quickly analyze molecular dynamics trajectories without restrictions, thereby overcoming the limitation of original algorithm.

The application of the developed algorithm made it possible to obtain and analyze the networks of hydrogen bonds of the studied defensin-like protein Pentadiplandra brazzeana, as well as to study the lifetime of hydrogen bonds between the residues of this protein and water molecules. The data obtained using the new implementation coincided with the experimental one.

Introduction

A hydrogen bond is a type of attractive (dipole-dipole) interaction between an electronegative atom and a hydrogen atom bonded to another electronegative atom. This bond always involves a hydrogen atom. Such bonds can occur between molecules or within parts of a single molecule.[1] They provide most of the directional interactions that underlie protein folding, protein secondary structure, and molecular recognition, and thus are essential for understanding protein spatial organization and movement. The energy of such bonds is experimentally known, which varies greatly from ≈5–6 kcal/mol for an isolated bond, to ≈0.5–1.5 kcal/mol for proteins in solution. The cores of most protein structures consist of secondary structures such as the α-helices and the β-sheets. This satisfies the hydrogen binding potential between the carbonyl oxygen of the main chain and the nitrogen of the amino group embedded in the hydrophobic core of the protein. The hydrogen bond between a protein and its ligands (which can be biomacromolecules such as proteins, nucleic acids, substrates, effectors, or inhibitors) ensures the directionality and specificity of the interaction, which is a fundamental aspect of molecular recognition. Therefore, the energetics and kinetics of the hydrogen bond must be optimal to allow fast selection and folding kinetics, conferring stability on the protein structure and providing the specificity required for selective macromolecular interactions.[2,3]

An implementation of the algorithm for determining hydrogen bond networks can be used as a Gromacs trajectory analysis module to unambiguously determine the existing hydrogen bonds in a trajectory. Knowledge of existing hydrogen bonds can be used, for example, to find protein residues that actively bind to water, to evaluate protein packaging and stabilization, to search for protein secondary structures, or to evaluate interactions with ligands or at protein-protein interfaces.

Implementation

In the newly developed implementation of the hydrogen bond network determination algorithm, these bonds served as the basis for the analysis, with oxygen or nitrogen atoms serving as acceptors, and oxygen or nitrogen covalently bonded with one or more hydrogen atoms serving as donors. To determine networks of hydrogen bonds in trajectories, a geometric criterion for the existence of a hydrogen bond is used:

$$r \leq r_{HB} = 0,35nm,$$

$$\alpha \leq \alpha_{HB} = 30^{\circ},$$

where r is the distance between the donor and the acceptor, and α is the Hydrogen-Donor-Acceptor angle. The value $r_{HB} = 0.35$ nm corresponds to the first minimum of the radial distribution functions of water in the SPC/E model.[4]

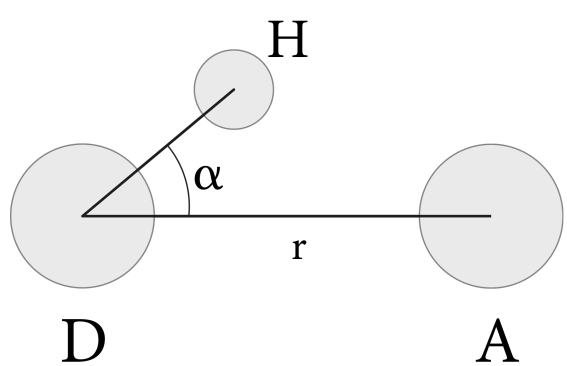


Figure 1: Geometric criterion for the existence of a hydrogen bond between donor D and acceptor A with the participation of the hydrogen atom H.

First of all, the algorithm analyzes the trajectory and topology received at the input, checking their correctness and compatibility with each other. Next, the algorithm asks the user to indicate two groups of atoms from the trajectory - reference and target - in order to analyze the existing hydrogen bonds between them. These two groups of atoms must be either absolutely identical or strictly non-intersecting. Otherwise, the algorithm will report

an error and will not analyze the structure further.

If the topology of the trajectory and the selected groups have successfully passed all the checks, then the algorithm will begin to selection of potential donors and acceptors from topology based on the reference and target groups. All oxygen atoms are always considered as potential hydrogen bond acceptors, while nitrogen atoms can be considered as potential acceptors or ignored in the definition of acceptors (this depends on the parameter passed by the user "-an" or "-noan"; by default, nitrogen atoms are considered potential acceptors, corresponding to the choice of "-an"). Donors are defined in terms of interaction functions obtained from the topology. From the interaction functions of the SETTLE type, hydrogen donors for water molecules are determined, which can be exclusively only oxygen atoms. Of the other interaction functions that can determine the formation of a chemical bond strictly between two arbitrary atoms, oxygen and nitrogen atoms are selected. In addition, for each donor, the hydrogen atoms associated with it are memorized. After all donors and acceptors have been identified, there is a special designation of those acceptors that have also been identified as donors.

If no donor or acceptor was found in any of the groups — the algorithm will issue a warning. If no donor and no acceptor are found in any of the groups, the algorithm will terminate its work prematurely with a special message to the user, since from such a data set, where there are no critically important components (donors and acceptors), hydrogen bond analysis will not be possible.

Next, the existing hydrogen bonds are calculated according to the geometric criterion at each moment of the trajectory. A search for neighboring atoms is carried out using the Neighbor Search algorithm: the reference and target groups are divided into donors and acceptors, and then pairs of atoms are selected (acceptor atoms from the reference group are compared with donor atoms from the target group) located at a distance of no more than 0.35 nm (this value can also be set by the user, but it cannot be less than 0.35 nm). In these pairs, the geometric criterion is checked for the donor, acceptor, and all hydrogen

atoms covalently bonded to the donor. If the criterion was satisfied, then the corresponding pairs with hydrogen are recorded in the storage. After the end of this procedure, the search starts again, but only if the reference group and the target group were not identical to each other. In the new round of the search, hydrogen bonds will be determined already between donors of the reference and acceptors of the target groups.

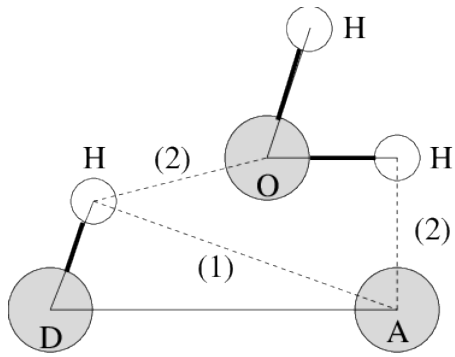


Figure 2: Hydrogen bond calculation. (1) Normal hydrogen bond between two residues. (2) A hydrogen bond bridge across a water molecule.[4]

As a result (by default), a file will be provided with the indices of all atoms of the reference and target groups and the indices of all atoms that are donors and acceptors for these groups. It will also list all hydrogen bonds that exist in the trajectory. These bonds can (at the user's request; parameter "-pf") be specified for each frame separately, or they can (by default; "-nopf") be specified for the entire structure as a whole (in this case, duplicated hydrogen links are not reflected in the output file). It is also possible (parameter "-m") to combine information about hydrogen bonds, which, in the above approximation (for each frame, or for the entire trajectory), having identical donor and acceptor indices, will differ only in the hydrogen index. However, by default (parameter "-nom") pairs of hydrogen bonds will not merge together and all found pairs will be presented in the output.

Molecular Dynamics Analysis Tools

In addition to the index map of hydrogen bond networks, at the output, at the user's request, various statistics of the processed data can be provided. The data obtained from the analysis

of hydrogen bonds on each frame is sent to the data analysis module (tool) integrated into Gromacs. To perform operations, one or more additional modules can be attached to the analysis module. Examples of such operations are averaging, building a histogram, and writing to a file. Modules can perform their data processing as they receive information. Their structure also supports parallel processing of frames: in such cases, information can be accessed in an arbitrary order. In particular, to save data for each frame to a file, as well as for other types of post-processing, pre-sorting by time index is necessary.

List of currently available statistics for output to the output file:

1. Graph of the number of hydrogen bonds versus time.

2. Distance distribution between donor and acceptor for hydrogen bonds.

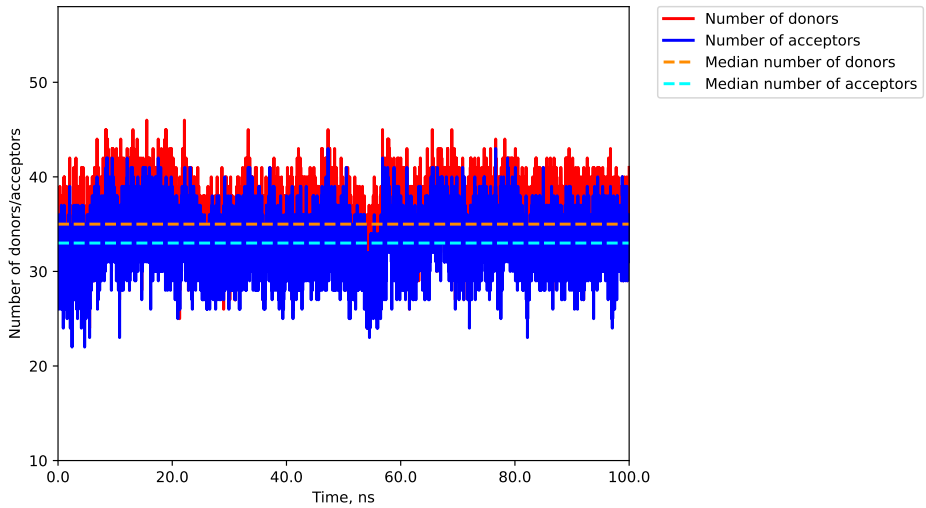
3. Hydrogen bond angle distribution.

4. Graph of the number of donors and acceptors analyzed in each frame.

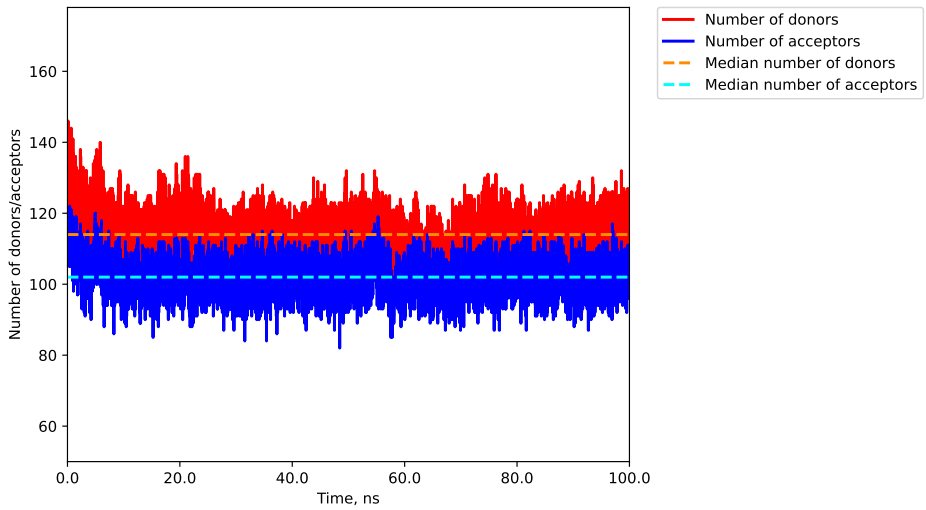
Example applications

As a result, an algorithm for determining networks of hydrogen bonds was developed. The main output of the developed algorithm is an index map, which contains the indices of acceptors, donors, hydrogen atoms associated with donors, and all atoms that participate in the formation of a hydrogen bond. In addition to the index map, the algorithm is also capable of producing optional files that contain data for plot creation of certain key statistics.

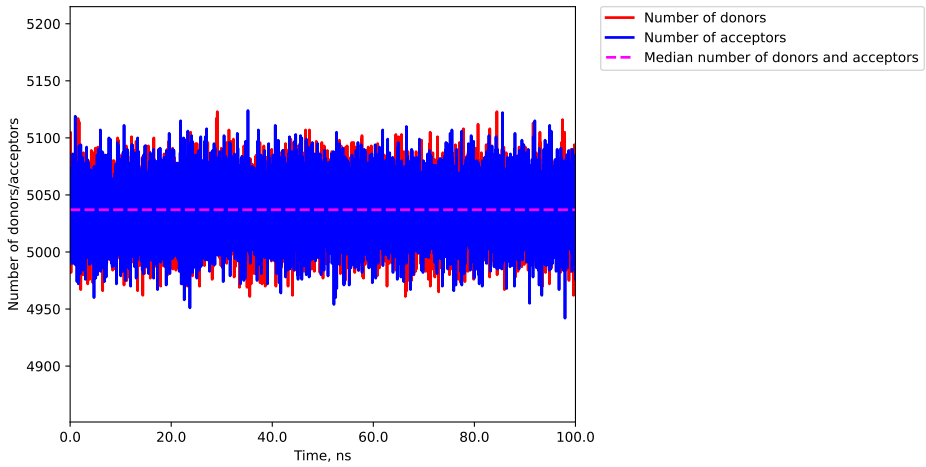
The developed algorithm was applied to the molecular dynamics trajectory of the defensin-like protein of Pentadiplandra brazzeana (Brazzein). The duration of the trajectory is 100 ns.



(a)

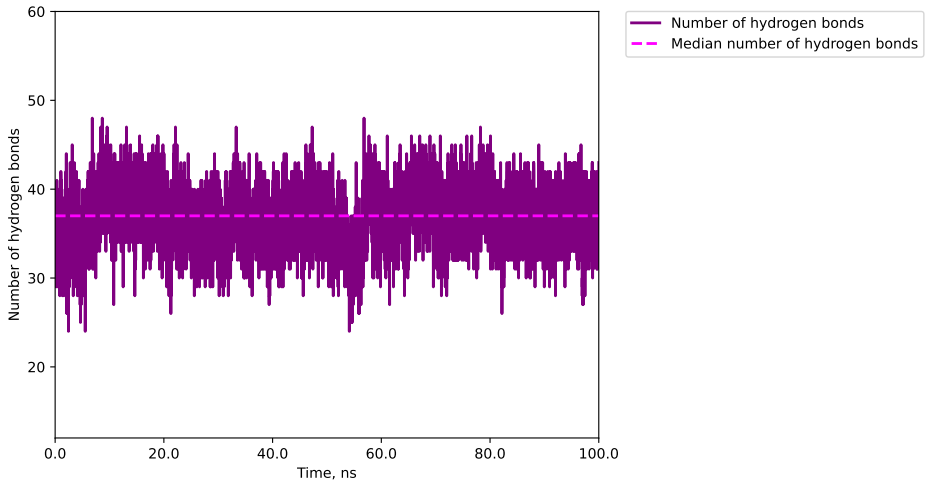


(b)

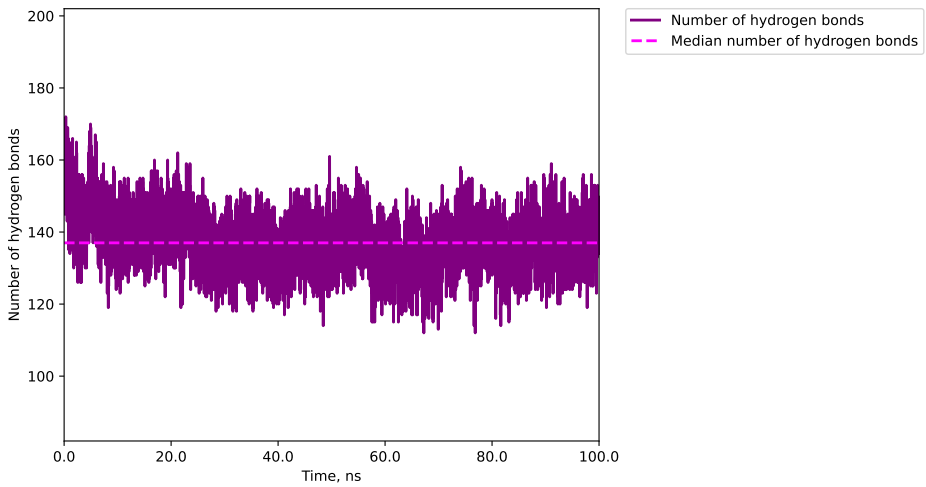


(c)

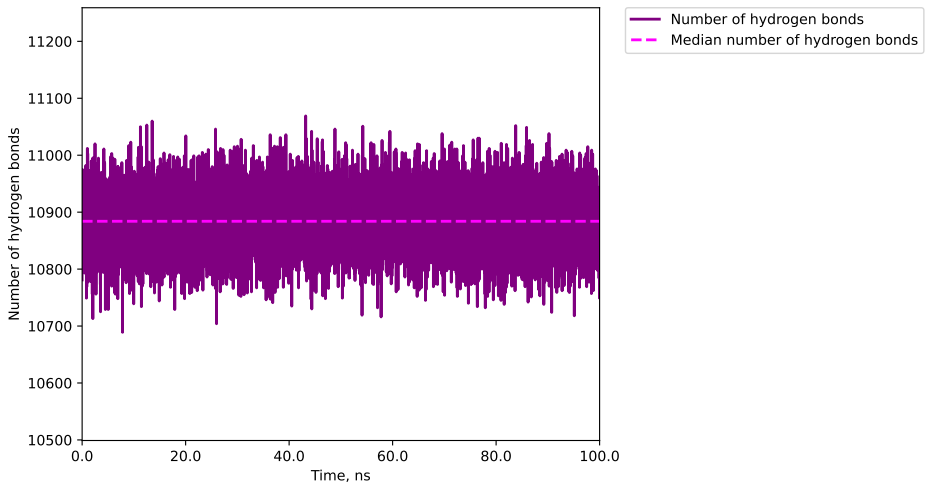
Figure 3: Graphs of the dependence of the number of donors and acceptors of a hydrogen bond between protein atoms (3a), between protein and water atoms (3b) and between water atoms (3c) as a function of time.



(a)

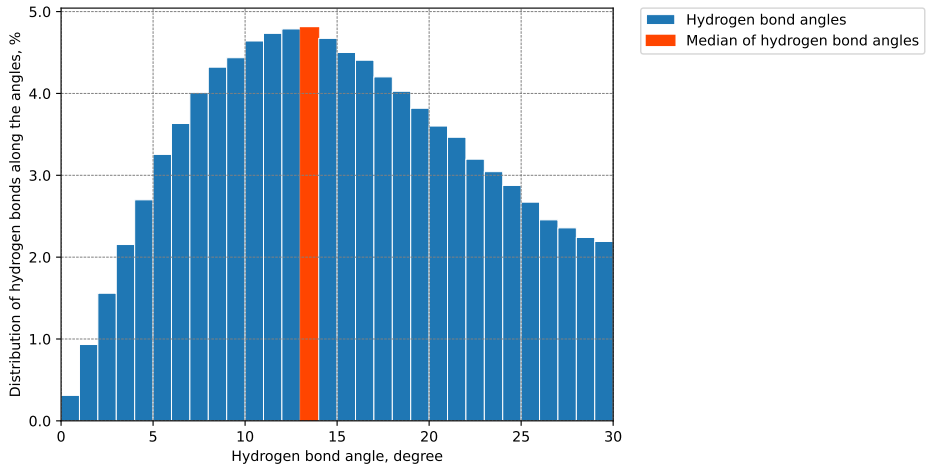


(b)

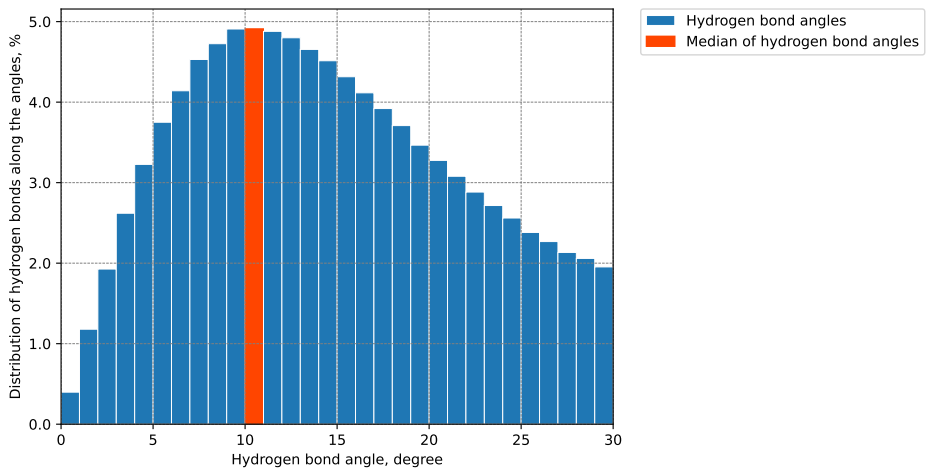


(c)

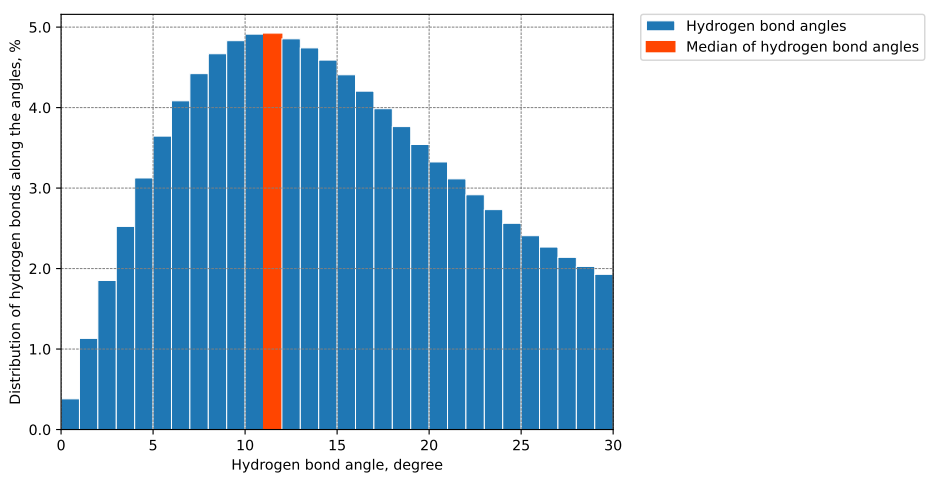
Figure 4: Graphs of the number of hydrogen bonds between protein atoms (4a), between protein and water atoms (4b) and between water atoms (4c) as a function of time.



(a)

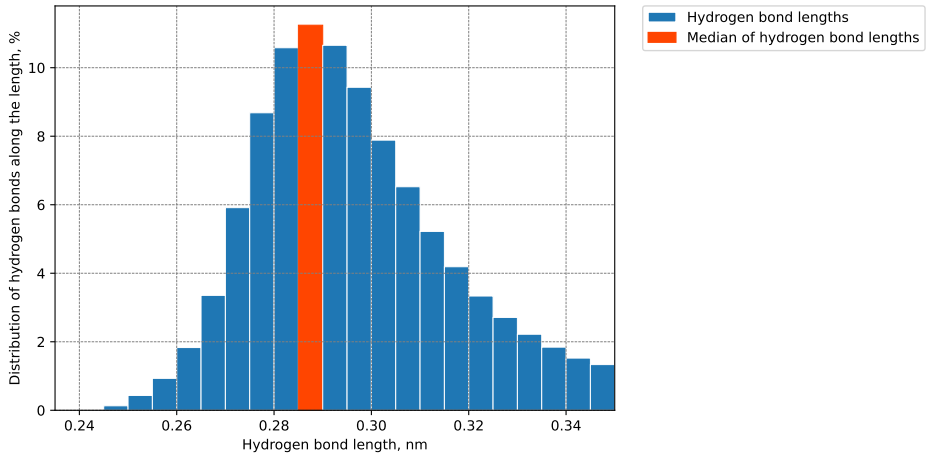


(b)

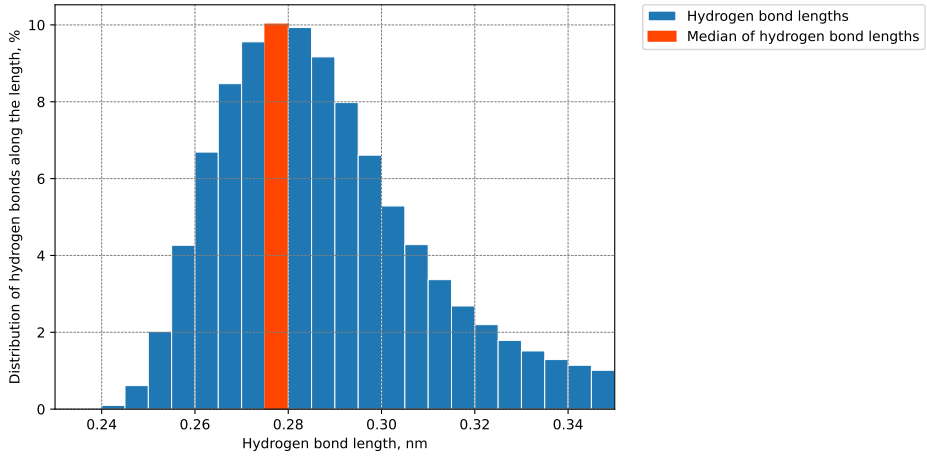


(c)

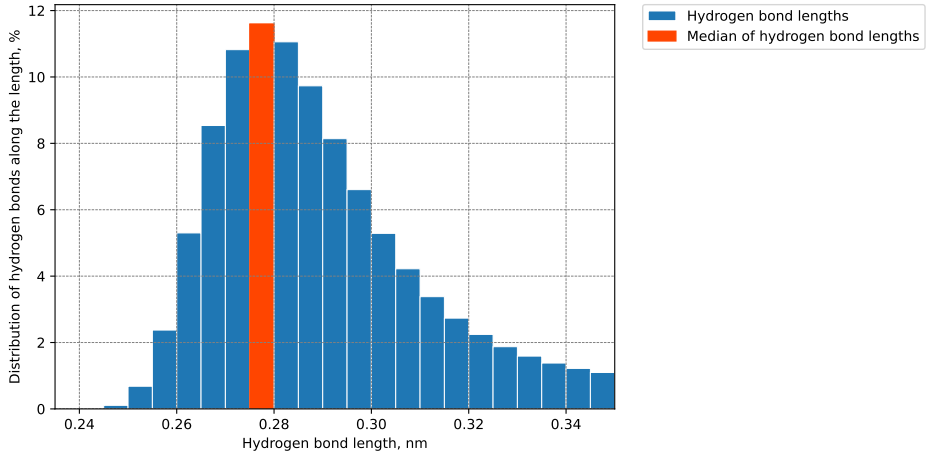
Figure 5: Histograms of distribution of angles of hydrogen bonds between protein atoms (5a), between protein and water atoms (5b) and between water atoms (5c).



(a)



(b)



(c)

Figure 6: Histograms of the distribution of hydrogen bond distances between protein atoms (6a), between protein and water atoms (6b) and between water atoms (6c).

Discussion

From the time dependences of the number of hydrogen bond donors and acceptors (see Fig. 3), it becomes clear that in the trajectoriy of molecular dynamics in pairs protein-protein and, especially, in protein-water pairs, the number of hydrogen bond donors consistently exceeds the number of acceptors. As expected, in water-water pairs, on average, the number of donors is identical to the number of acceptors.

As can be seen from the obtained histograms of the distribution of hydrogen bond distances (see Fig. 6), the most probable values of hydrogen bond lengths are in the range 0 .24-0.35 nm. This is consistent with the theory of hydrogen bond formation, in which hydrogen bonds with a distance between the donor and acceptor of 0.22-0.25 nm are considered strong (mainly covalent), those with a distance of 0.25-0.32 nm are considered moderate (in mostly electrostatic), and those with distances exceeding 0.32 nm are considered weak (electrostatic).[5,6] The presented graphs show that most of the identified hydrogen bonds in the studied trajectory are moderate. Weak and strong hydrogen bonds are also present in a smaller amount.

It is known[7,8] that the distance between hydrogen and a hydrogen bond donor (if the donor is a nitrogen atom N or an oxygen atom O) is ≈1 Å. It is also known that the closer the angle $\angle_{DHA}$ to 180°, the stronger the interatomic interactions inside the hydrogen bond (the bond is considered strong when the angle $\angle_{DHA}$ reaches 170°).[6] From this we can conclude that the smaller the angle $\angle_{HDA}$, the stronger the hydrogen bond will be, and the interaction within the hydrogen bond will be especially strong at $\angle_{HDA} \leq 7.16°$. Weak hydrogen bonds can be considered as bonds with an angle $\angle_{HDA} > 30°$, but, thus, they are eliminated when determining a hydrogen bond by a geometric criterion (see 1) and, therefore, are outside the range of values obtained by the software algorithm in the output file. Based on the obtained histograms of the distribution of hydrogen bond angles (see Fig. 5), one can see that hydrogen bonds of moderate strength are predominantly formed in all trajectoriy.

Using the developed algorithm, it was possible to determine the lifetime of hydrogen

bonds between protein atoms and water atoms for each donor and acceptor in the studied trajectory (see Table 1). The obtained data was compared with the hydrogen bonds found in the crystal structures of the studied proteins obtained experimentally (using the X-ray diffraction method).

Table 1: List of residues involved in the formation of hydrogen bonds as donors of the defensin-like protein of Pentadiplandra brazzeana (Brazzein). Residues that form hydrogen bonds as donors in the crystal structure with identifier 7w8e are highlighted in bold.

Hydrogen bond donor residues	Life time, ps
K3	99260
K15	98840
K6	98770
K27	98300
D2	97750
K30	97300
K5	97100
R43	96850
R43	95600
R33	93590
N20	93440
S14	93250
Q17	92680
K42	92590
Y39	91270
R33	90140
Y24	88900
N23	87750
C22	86040
Y54	83430
Q21	82910
R43	79400
Y39	76640
V7	75780
N10	74400
A19	72250
K3	70610
H31	69870
V13	67870
N20	66650
S14	65960

Table 2: List of residues involved in the formation of hydrogen bonds as acceptors of the defensin-like protein of Pentadiplandra brazzeana (Brazzein). Residues that form hydrogen bonds as acceptors in the crystal structure with identifier 7w8e are highlighted in bold.

Hydrogen bond acceptor residues	Life time, ps	Hydrogen bond acceptor residues	Life time, ps
E9	99480	S14	52230
E9	99380	**Y39**	50170
D29	98760	**K30**	47690
E36	98580	**N23**	45500
Q17	98470	K15	45370
E36	97750	**Y8**	39420
D2	97600	**Y51**	30930
D2	97320	**Y24**	29190
D29	97070	**Q21**	28060
E53	96780	**L45**	15260
E41	96170	**V13**	14220
D29	95960	**N20**	11780
Q21	95950	**H31**	11180
Y54	95920	Q17	11170
Y39	95820	**Q21**	9640
H31	95740	**N44**	9600
E53	95430	**D40**	9430
Q17	95330	**N20**	8450
E41	95190	N10	7270
V7	94890	C4	6700
L28	94360	**N44**	6390
D50	94290	**N23**	6110
C37	94080	**D25**	4920
R43	94060	**N23**	4400
H31	93550	**K6**	3340
E41	92720	Y11	3090
D25	92190	R43	2870
Y54	92020	R43	2640
N10	91380	**R33**	2230
K42	90890	**Q46**	1970
C16	90570	**R33**	1830
N10	90550	P12	1630
Y11	88750	**E36**	1550
N20	88470	**C47**	1300
Y51	88290	C49	1220
K3	87900	**F38**	870
A32	87400	G35	680
C52	86220	R43	540
G35	84640	**K42**	530
L18	84130	Y54	520
D25	82510	D50	500
K5	82450	Y11	360
K27	82280	**D25**, R33	350
D50	81620	E53	330
R33	79180	**S34**	290
Q46	78990	**E9**, Q17, E36	240
N44	78690	**Y39**, E41	230
S14	77730	K15, **F38**	220
D40	76060	**Y24**	210
I48	74850	**Y8**	190
Y8	71770	N10, **N20**	180
D40	71330	**D29**, Q46	160
Y24	70340	**V7**, D40	130
D2	68460	**L18**, A19, N44	120
S34	68040	**L28**, R33, D50	110
Y54	66130	C4, **K5**, P12, S14	90
E53	63130	**K42**, L45	80
E9	57290	C26, **K30**	60
A19	54280	K3, **K5**, V13, K15, C26	50
S14	52230	K3, **K6**, K27, K30, C37, R43	40

It should be noted that some residues in the tables may be repeated, since they can contain several atoms that act as a hydrogen bond donor or acceptor. Therefore, each of these atoms can independently form hydrogen bonds. And since the lifetime of a hydrogen bond was determined specifically for each protein atom individually, the values of the formed hydrogen bonds for them were recorded individually.

All the residues (except for Ala43 residue — it was replaced by Arg43 in the Brazzein protein, examined in the trajectory) that were identified as containing atoms of donors and acceptors of hydrogen bonds from protein crystals also formed hydrogen bonds of varying degrees of stability (mostly stable, i.e. forming hydrogen bonds that exist for more than 2% (2 ns) of the time of the total duration (100 ns) of the trajectory).

From the Brazzein protein trajectory, it was possible to isolate many residues that had hydrogen bonding atoms in the crystal structure and yet also formed a relatively stable hydrogen bond with water through their atoms in a molecular dynamics trajectory. These are the residues: Asp2, Lys3, Cys4, Lys5, Lys6, Val7, Tyr8, Glu9, Asn10, Val13, Cys16, Gln17, Leu18, Asn20, Gln21, Cys22, Asn23, Tyr24, Asp25, Leu28, Asp29, Lys30, His31, Arg33, Ser34, Glu36, Phe38, Tyr39, Asp40, Glu41, Lys42, Asn44, Leu45, Gln46, Cys47, Tyr51, Cys52.

The residues with the longest-lived hydrogen bonds between protein and water were analyzed for the formation of stationary bonds. Of the most stable hydrogen bonds formed by residues in the trajectoriy of molecular dynamics, we selected those that formed a long-term bond (both as a donor and as an acceptor) with the same water molecule.

Table 3: List of Brazzein protein residues that forming stable stationary hydrogen bonds with water molecules. Residues forming hydrogen bonds in the crystal structure with identifier 7w8e are highlighted in bold.

Residues with stationary connections	Longest lifetime, ps	Numer of stationary connections
C22	12870	11
L18	12170	12
K15	7220	7
N23	4940	3
G35	4860	3
E36	2160	1
D50	2140	1
Q21	2010	1

Table 3 shows that the Cys22 and Leu18 residues, which were previously identified as forming hydrogen bonds with water in the crystal structure with identifier 7w8e, in addition to forming stable hydrogen bonds in the molecular dynamics trajectory, also form long-lived stationary bonds with water molecules.

Thus, using the developed algorithm for determining hydrogen bond networks, a conclusion was made not only regarding the general abilities to form hydrogen bonds in individual protein residues, but also regarding the ability to form long-lived stationary bonds with water molecules. The results obtained using the developed algorithm agree with the results obtained experimentally.

Comparison of algorithms' "run time"

The speed of the old and new implementations of hydrogen bond determination algorithms was tested on arbitrary trajectories of different sizes for defensin-like protein of Pentadiplandra brazzeana, Sars-Cov-2 Non-Structural Protein 12 and T7 RNA polymerase. In all cases, only protein and/or water atoms from the molecular dynamics trajectory were analyzed to obtain records of hydrogen bonds.

According to the data presented on figures **??**, **??** and **??** in supporting information section, we can conclude that the developed implementation of the algorithm works faster

than it's counterpart. This result was achieved due to the correct integration of the algorithm into the Gromacs trajectory analysis module.

Conclusions

The re-implementation of the hydrogen bond network determination algorithm in the trajectory of defensin-like protein of Pentadiplandra brazzeana molecular dynamics made it possible to obtain a detailed list of residues that are most actively involved in the formation of hydrogen bonds both as a donor (see Table 1) and as an acceptor (see Table 2) and determine that in the Brazzin protein residues Cys22 and Leu18 are the most active residues that form stationary hydrogen bonds.

Unlike the original algorithm for determining hydrogen bonds nerworks, the new implementation consumes much less memory and works much faster. The data obtained as a result of the work of the new algorithm do not differ from those obtained by the original algorithm. The new implementation is capable of accepting structures of any size as input, without restrictions, and can output information both about the contained hydrogen bonds at each moment of the trajectory, and about hydrogen bonds in the trajectory as a whole. In the new implementation, the shortcomings contained in the previous implementation of this algorithm were eliminated.

Data Availability

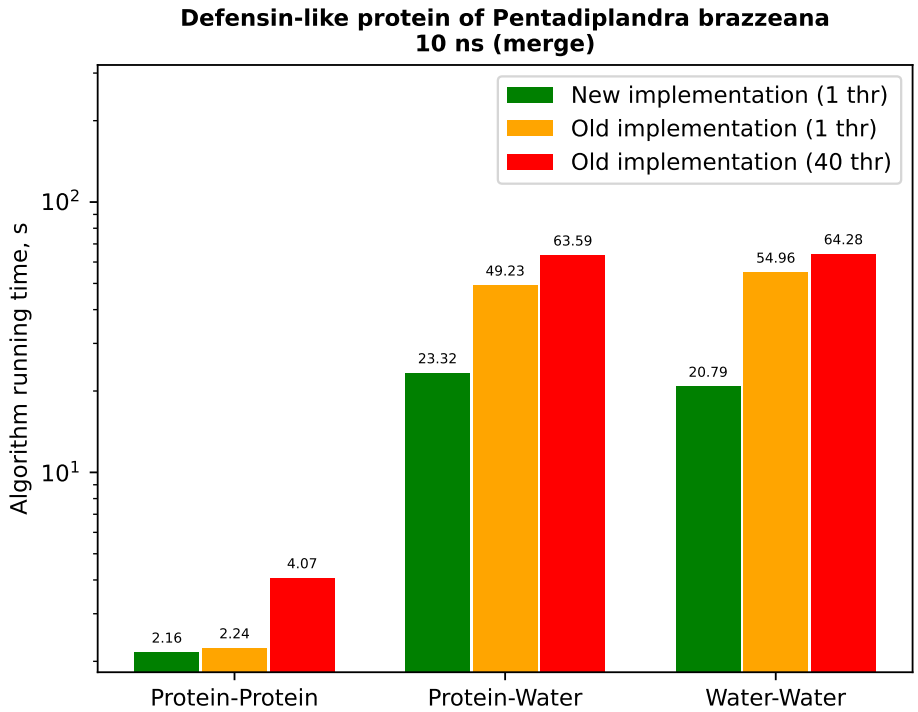
Our standalone implementation of the algorithm for determining hydrogen bond networks in Gromacs is available at the GitLab repository: `https://gitlab.com/bio-pnpi/gmx-hbond`. You can also look at the measured numerical results of speed comparisons (in the form of graphs and tables) at `https://gitlab.com/bio-pnpi/gmx-hbond-refdata`.

Acknowledgments

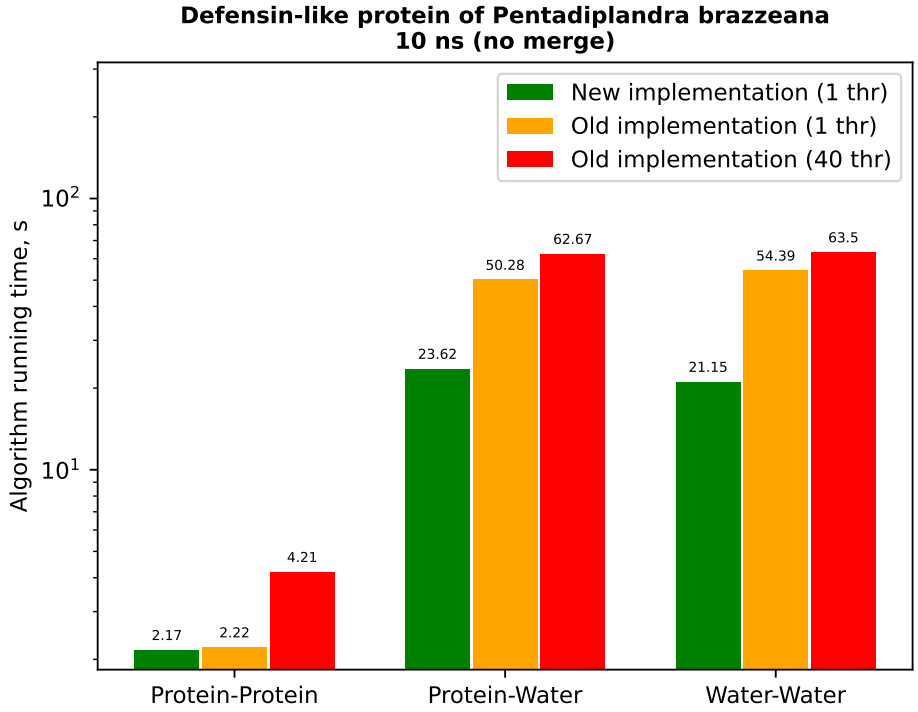
This research is supported by the Ministry of Education and Science of the Russian Federation in the framework of the basic part of the state order (projects № 121060200127-6 and No 2.8629.2017).

Supporting Information Available

The speeds of the old and new implementations of the algorithm for determining hydrogen bonds were tested on arbitrary trajectories of different sizes for the defensin-like protein of Pentadiplandra brazzeana, Sars-Cov-2 Non-Structural Protein 12 and T7 RNA polymerase. In all cases, the old implementation ran in one and 40 threads, while the new implementation ran in one thread. When testing the running time, each algorithm was given a trajectory with instructions to determine hydrogen bonds inside protein molecules, inside water molecules, and between protein and water molecules. Please note that in order to reduce the usage of space, not all obtained graphs for comparing the running time of the algorithms are given below. You can view the resulting graphs for all trajectories in the GitLab repository by following the link provided in the "Data Availability" section.

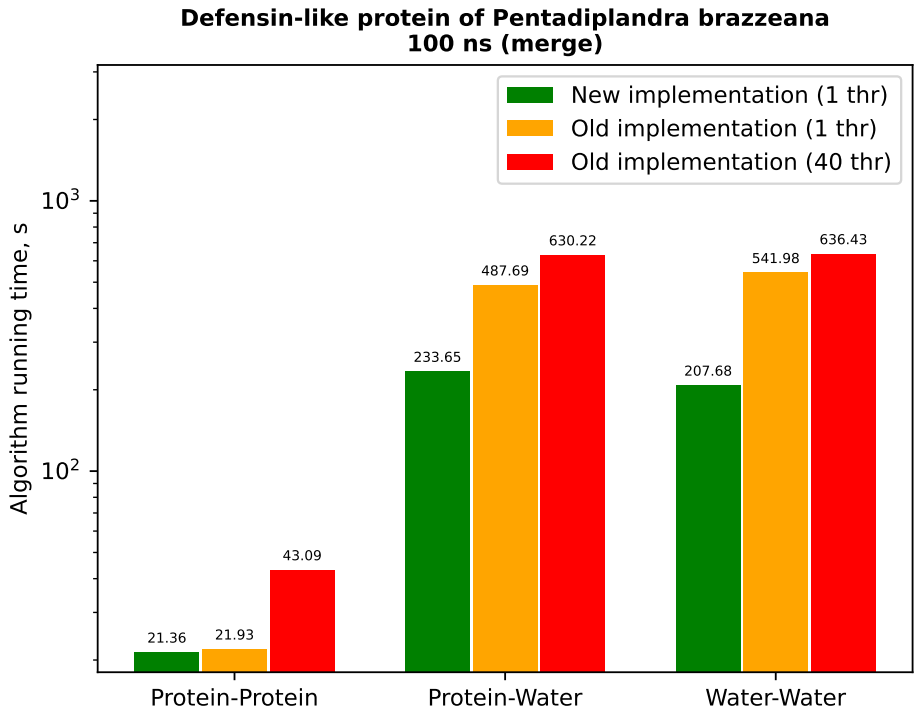


(a)

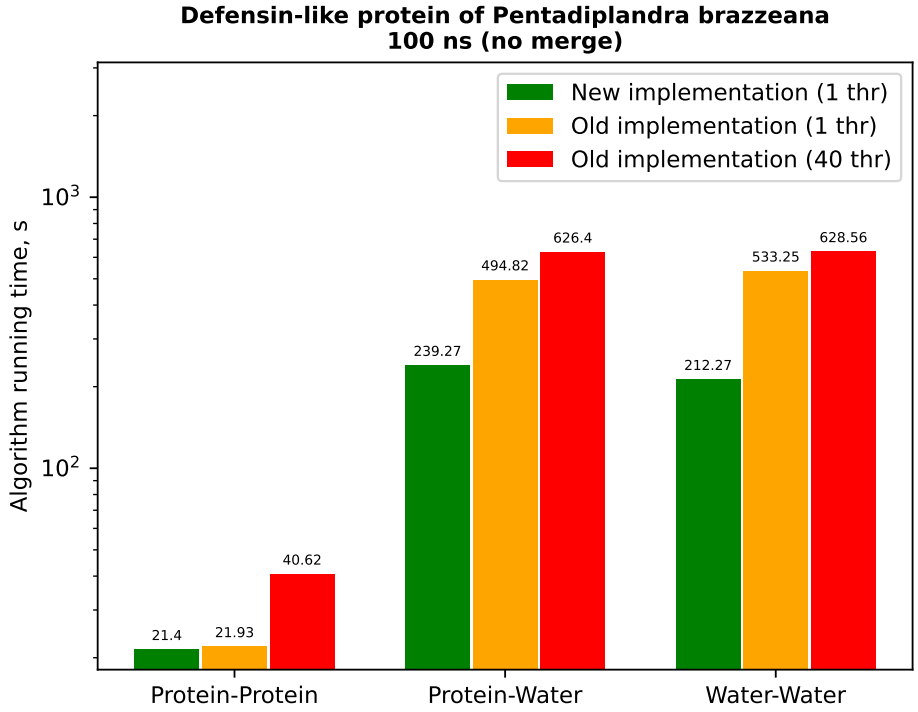


(b)

Figure 7: Comparison of the running time of the algorithms on the trajectory of the defensin-like protein Pentadiplandra brazzeana with a duration of 10 ns with (7a) and without (7b) the use of the hydrogen merge option.

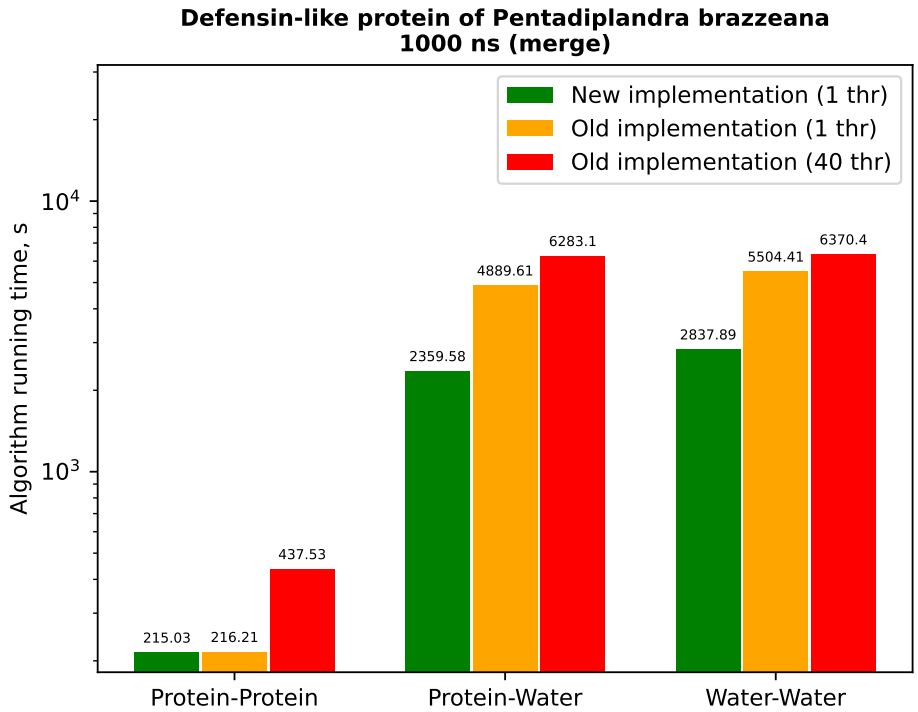


(a)

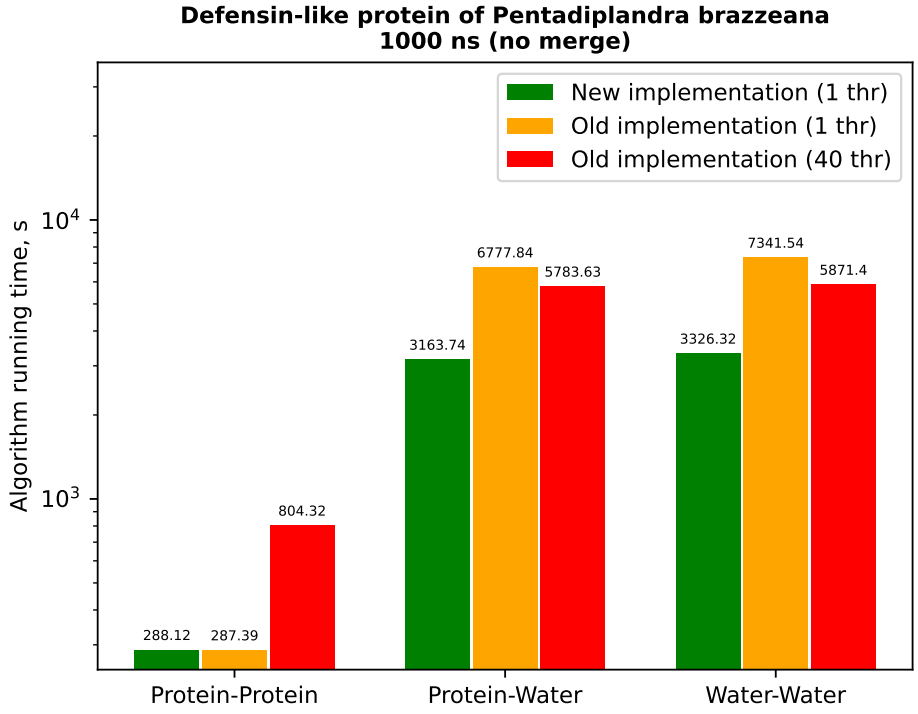


(b)

Figure 8: Comparison of the running time of the algorithms on the trajectory of the defensin-like protein Pentadiplandra brazzeana with a duration of 100 ns with (8a) and without (8b) the use of the hydrogen merge option.



(a)



(b)

Figure 9: Comparison of the running time of the algorithms on the trajectory of the defensin-like protein Pentadiplandra brazzeana with a duration of 1000 ns with (9a) and without (9b) the use of the hydrogen merge option.

Thus, according to the obtained data, we can conclude that the re-implementation of the algorithm for determining hydrogen bond networks works, in general, faster than the previous implementation. It is important to note that re-implementation is sometimes somewhat slower on small trajectories. This happens due to the peculiarities of the organization of the structures of the collections of atoms inside the trajectory analysis module. But, despite the slight delay on small trajectories, with an increase in the size of the trajectories arriving at the input, a significant difference in the speed of algorithm execution becomes noticeable in favor of the new implementation.

A formula for calculating the memory occupied for storing data on hydrogen bonds in the old and new implementations was derived. According to the calculations, the worst forecast for the amount of memory V (in bytes) occupied by the data store in the old implementation is:

$$V = 72m + 60n^2,$$

while in the new implementation it is:

$$V = m(64n + 32),$$

where m is the number of frames and n is the number of atoms in the investigated trajectory.

From this we can conclude that the amount of occupied memory in the old implementation grows quadratically from the number of atoms, while in the new one it grows linearly. Therefore, in the old implementation the limiting factor is the total size of the system, while in the new it is the length of the trajectory. However, if in the old implementation this limitation cannot be bypassed, thereby becoming an insurmountable boundary for analysis, in the new implementation this limitation can be circumvented by splitting the trajectory under study into successive fragments. Additionally, the new implementation allows storing, in addition to the list of hydrogen bonds, the numbers of frames in which they were found, while in the old implementation this possibility was not implemented.

It's also important to note that the old implementation supports multithreading, while the new implementation doesn't (at the moment). However, it was noticed that when even 40 parallel threads were connected to perform calculations in large systems (such as the molecular dynamics trajectory for 10000 and 100000 frames), the running time of the old implementation was still significantly longer than the running time of the new implementation. It was noticed that with an increase in the number of threads in the old implementation, sometimes the calculation time increased when evaluating hydrogen bond networks within a protein and between a protein and water, although it was expected that an increase in the number of threads should, on the contrary, dramatically increase the calculation speed. This indicates the imperfection of the implementation of the algorithm in the old version. Despite all of the above, the creation of a multi-threaded mode for the new implementation is in the plans for the further development of scientific work in order to further reduce the calculation time of the developed software algorithm.

References

(1) Helmenstine, A. M. Hydrogen Bond Definition and Examples. *ThoughtCo* **2023**,

(2) Pace, C. et al. Contribution of hydrogen bonds to protein stability. *Protein science : a publication of the Protein Society* **2014**, *23*.

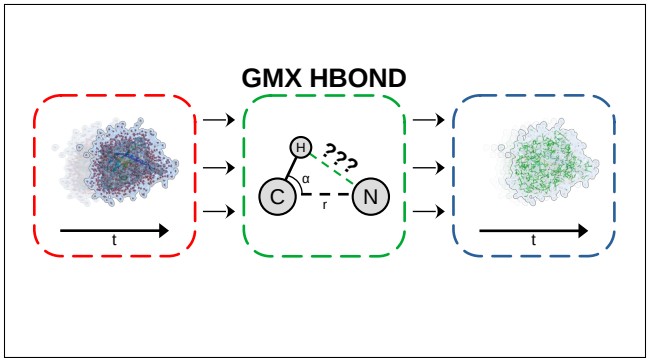
(3) Hubbard, R. E.; Haider, M. K. Hydrogen Bonds in Proteins: Role and Strength. **2010**,

(4) Abraham, M. et al. GROMACS 2023.1 Manual (2023.1).

(5) Dannenberg, J. J. An Introduction to Hydrogen Bonding By George A. Jeffrey (University of Pittsburgh). *Journal of the American Chemical Society* **1998**, *120*, 5604–5604.

(6) Steiner, T. The Hydrogen Bond in the Solid State. *Angewandte Chemie International Edition* **2002**, *41*, 48–76.

(7) Yao, L.; Vögeli, B.; Ying, J.; Bax, A. NMR Determination of Amide N-H Equilibrium Bond Length from Concerted Dipolar Coupling Measurements. *Journal of the American Chemical Society* **2008**, *130*, 16518–16520.

(8) Demaison, J.; Herman, M.; Lievin, J. The equilibrium OH bond length. *International Reviews in Physical Chemistry* **2007**, *26*, 391–420.

TOC Graphic