INTERNATIONAL JOURNAL OF ENERGY
DOI: 10.46300/91010.2022.16.7 Volume 16, 2022

Multiscale Convergence Optimization in Constrained Molecular Dynamics Simulations

N. M. Nafati, S. Antonczak, J. Topin. J. Golebiowski.
APSM Team. ICN. UMR7272.
Scientific University of Nice. France.

Abstract-The energy analysis is essential for studying chemical or biochemical reactions but also for characterizing interactions between two protagonists. Molecular Dynamics Simulations are well suited to sampling interaction structures but under minimum energy. To sample unstable or high energy structures, it is necessary to apply a bias-constraint in the simulation, in order to maintain the system in a stable energy state. In MD constrained simulations of "Umbrella Sampling" type, the phenomenon of ligand-receptor dissociation is divided into a series of windows (space sampling) in which the simulation time is fixed in advance. A step of de-biasing and statistical processing then allows accessing to the Potential Force Medium (PMF) of the studied process.

In this context, we have developed an algorithm that optimizes the DM computation time regarding each reaction coordinate (distance between the ligand and the receptor); and thus can dynamically adjust the sampling time in each US-Window. The data processing consists in studying the convergence of the distributions of the coordinate constraint and its performance is tested on different ligand-receptor systems. Its originality lies in the used processing technique which combines wavelet thresholding with statistical-tests decision in relation to distribution convergence.

In this paper, we briefly describe a Molecular Dynamic Simulation, then by assumption we consider that distribution data are series of random-variables vectors obeying to a normal probality law. These vectors are first analyzed by a wavelet technique, thresholded and in a second step, their law probability is computed for comparison in terms of convergence.

In this context, we give the result of PMF and computation time according to statistic-tests convergence criteria, such as Kolmogorov Smirnov, Student tTest, and ANOVA Tests. We also compare these results with those obtained after a preprocessing with Gaussian low-pass filtering in order to follow the influence of thresholding. Finally, the results are discussed and analyzed regarding the contribution of the muli-scale processing and the more suited criteria for time optimization.

***Index* Keywords: Molecular Dynamic Simulations. Umbrella Sampling. Potential Force Medium (PMF). Convergence. Ligand – Receptor. Wavelet Thresholding. Statistical-Tests Decision. Normal Probality Law. Kolmogorov Smirnov. Student tTest. ANOVA. Gaussian low-pass filtering.**

I. INTRODUCTION

Molecular Dynamics (MD) is a sampling space based on iterative numerical integration of the equations of Newton motion. Since the time scales accessible to MD simulations are several orders of magnitude less than the time of chemical reactions, we may introduce a bias to increase the likelihood of sampling rare or unlikely events. In fact, when the energy barrier between two states to be sampled is less than KT (K: Boltzmann constant. T: system temperature in °K), the probability can be obtained in simulations at equilibrium. In the case of larger barriers, the state of higher energy will not be reached and a harmonic potential sampled must be inserted in order to obtain the Hamiltonian suitable sampling. The addition of this harmonic potential is called Umbrella-Sampling (U.S.), where the force constant of this potential is another parameter at equilibrium (or reaction coordinate at equilibrium) . This way provides a sampling that does not follow the Boltzmann statistics, but improves sampling in the vicinity of a chosen value of reaction coordinate [9][10][14].

The sampling technique is therefore constrained so as to cut the energy path connecting an initial state to a final state of the reaction in several windows, as shown in the fig1. .

According to Boltzmann statistics [10][13][17][20][23], Potential Strength Medium (PMF) noted F is given by the following formula:

$$\mathbf{F}(\boldsymbol{\xi}) = -\mathbf{K_B T} \ln(\mathbf{P}(\boldsymbol{\xi})) \quad (1)$$

Where P(ξ) is the probability distribution that represents the number of times that the system samples the reaction coordinate. The expression of P(ξ) and hence of the PMF are modified in the case of sampling constraint. The WHAM (Weigthed Average Histogram Analysis) [9][14] software

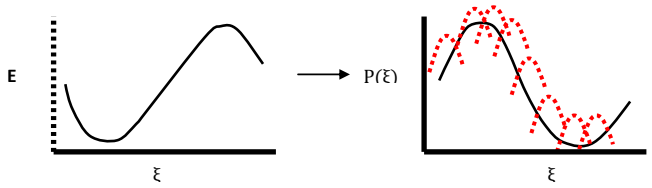

Fig1.: Schematic overall sampling by Umbrella Sampling technique under a harmonic constraint. The left curve shows the energy according to the reaction coordinate. The right one shows the probability as a function of the same parameter. The red curves having the umbrella-shaped sample windows are Umbrella-Sampling.

aims to determine the PMF through the calculation of some constants induced by the harmonic constraint.

As indicated in Fig1., two successive harmonic potential sampling windows, with a recovery given rate, lead to PMF identification up to a constant. The expression of this constant is:

$$\ln\left(\langle e^{\beta V_1(\xi)}\rangle\right) - \ln\left(\langle e^{\beta V_2(\xi)}\rangle\right) \quad (2)$$

The WHAM program corrects the value of PMF in each window by calculating the constants from (2) through the recovery of successive US windows, and thus provides the values of PMF (kcal/mol) depending on the reaction coordinate.

II. Problem Context:

A molecular dynamics simulation [10] consists in several steps of calculation:

-Energy minimization processing.

-Thermalization processing in which the protein complex is heated up to 300 °K. During this phase the thermal noise increa-ses.

-Equilibration phase processing: the energy of the system becomes minimum.

-Production Data phase: data are collected in order to compute the PMF.

The Amber software is used to achieve all these steps, and to reach among other things, different distributions of coordinate reaction (Distance between molecules), expressed in Angstroms. Each distribution is a distance-vector including a deterministic and random component. These distribution-vectors are measured at a given time in a sampling window-time whose size is set in advance. The present random component in each distribution is due to thermal agitation. The random thermal noise explains that noise components due to thermal agitation may have all sorts of values from low frequencies to very high frequencies. It greatly influences the number of iterations of the temporal sampling under a constraint reaction coordinate.

During the production phase, data distributions are collected by moving the US window along the spatial axis sampling. Therefore this sampling is directed or forced because the system is maintained in an energy potential well. Then the distribution distance estimator depends only on the distance, so the directed sampling allows to overcome the other freedom degrees.

To obtain efficient estimators of the distance-distribution, the US windows must be relatively long (approximately about several tens nanoseconds) and this for a large protein complex of approximately 5000 atoms. Consequently, our work has focused on developing algorithms that use short time windows centered around each spatial sampling. Our algorithm combines the wavelet and gaussian filtering processing with statistical convergence criteria. Our simulations have shown that this combination can reduce the cost of computing time.

III. Data formulation:

(2)

In our study, we consider that during a Dynamic Molecular Simulation, each distribution $\boldsymbol{X_i}_{\,0\le i\le N-1}$ is obtained as a stochastic measure which represents a temporal sample at a given distance. In other words, the constrained spatial sampling proceeds as follows: for each distance $\boldsymbol{D_J}_{\,0\le j\le M-1}$ from $\boldsymbol{D_0}$ to $\boldsymbol{D_{M-1}}$, we collect each distribution $\boldsymbol{X_i}_{\,0\le i\le N-1}$ at time $\boldsymbol{t_{i-1}+\Delta t}$ until the convergence of their probability distribution, then we increment the distance and repeat the process until $\boldsymbol{D_{M-1}}$.

The objective of our molecular modeling is to obtain measurements of the vibrational micro-state of the protein complex at a given distance. To do this, we have to accumulate a large number of statistical data which represent the probability distribution of the micro-states system. This statistic depends on the thermodynamics complex.

So, we denote that each distribution is a random measure such that :

$$\boldsymbol{X = X_d + \varepsilon}$$

Where $\boldsymbol{X_d}$ is the determinist component, and $\boldsymbol{\varepsilon}$ is the noise component.

So, at every spatial sampling step, we collected a series of distributions: $\{X_0, X_1, X_2, X_3, X_4, \ldots, X_{N-1}\}$ whose limit mainly depends on the speed of the probability convergence. This series is considered as a set of independent random variables, supposed to follow a normal distribution.

IV. Removing Noise with Wavelet Transformation

A. A brief overview of wavelets

The Continuous Wavelet Transform (CWT) [3][19][23] gives a time-frequency representation of signals. A wavelet

transform is a convolution of a signal X(t) with a set of functions which are generated by translations and expansions of a main function. The main function is known as the mother wavelet and the translated or expanded functions are called wavelets. Mathematically, the CWT is given by:

the translated or expanded functions are called wavelets. Mathematically, the CWT is given by:

$$\boldsymbol{W(a,b) = \frac{1}{\sqrt{a}} \int X(t)\psi(\frac{t-b}{a})\, dt} \quad (3)$$

Here $\boldsymbol{b}$ is the time translation parameter and $\boldsymbol{a}$ is the expansion parameter of the wavelet. Ψ is the mother wavelet which is non-zero only on a certain interval.

The Discrete Wavelet Transform (DWT) is similar to the Fourier transform in that a signal is decomposed in terms of a basis set of functions. In Fourier transforms the basis set involves sines and cosines and the expansion has a single parameter. In wavelet transform the expansion has two parameters and the functions (wavelets) are generated from a single "mother" wavelet [8].

Any signal can be decomposed as:

$$\boldsymbol{X(t) = \sum_a \sum_b c_{ab}\, \psi_{ab}(t)} \quad (4)$$

Where the two-parameter coefficients are given by

$$\boldsymbol{c_{ab}(a,b) = \int X(t)\, \psi_{ab}(t)\, dt} \quad (5)$$

And the wavelets $\boldsymbol{\psi_{ab}}$ obey the condition

$$\boldsymbol{\psi_{ab}(t) = 2^{\frac{a}{2}}\, \psi(2^a t - b)} \quad (6)$$

B. Removing Noise with the DWT

"De-noising" a signal with the DWT involves three steps [1][5][8][15]:

1. Transform the input signal **X** with the DWT (X is a vector whose dimension is equal to N).

2. Force to zero all transform coefficients whose magnitude falls below some percentage of the maximum magnitude. This is a thresholding operation in which the threshold is adaptively computed or not. In our simulation, we used an adaptative soft thresholding with a universal threshold [1][5][15]. given by:

$$\boldsymbol{T = \sigma \sqrt{\frac{2\log(N)}{N}}} \quad (7)$$

Where σ is the standard deviation of X.

3. Inverse DWT.

V. Principle of Statistical Tests

A hypothesis test (or statistical test) is an approach that aims to provide a decision rule which, on the basis of sample results, leads to make a choice between two statistical hypothesis. These two hypothesis are disjoint, in other word mutually exclusive.

Significance level of the test :

There is a risk, agreed in advance, of wrongly rejecting the null hypothesis H_0 when it is true, it is called the significance level of the test and the corresponding probability is noted α:

$$\boldsymbol{\alpha = P\,(\,\text{To Reject } H_0 \mid \text{when } H_0 \text{ is true}\,)} \quad (8)$$

At this level of significance, we affect to the statistic sampling distribution a rejection region of the null hypothesis (also called the critical region or **C**ritical **P**robability ($\mathbf{P_c}$)). The area of this region is the probability α. For example, choosing α = 0.05 means that sampling variable can take in 5% of cases, a value belonging to the rejection area of $\mathbf{H_0}$. The sampling distribution matches to a complementary region, called region of acceptance of $\mathbf{H_0}$ (or region of non-rejection) whose probability is equal to 1 - α.

From this point, we use different stochastic tests as convergence criteria. These criteria are focused on α as a threshold of acceptation or rejection of any convergence hypothesis. We set α error to 0.05 in our simulations.

As we said before, our data are samples (distance distributions) coming from the same population, which follows a probability law supposed to be normal $\mathcal{N}(\mu,\sigma)$. In our simulations we use non-parametric [2][4][6,7][12][16][21] and parametric statistic-tests [11][18][22] in order to find the best statistic-tests convergence criteria.

The statistical tests are:

- The parametric test of student (tTest), to test the sample average convergence, assuming that the variance is known.

- The non-parametric Kolmogorov Smirnov test, another non-parametric alternative to the tTest for independent samples.

- The Anova non-parametric test, to check if the difference between the sample averages can be attributed to random sampling or to the fact that the samples are really significantly different [18].

A. Student Convergence criteria

Convergence of sequences of random variables is an important concept in probability theory and statistics, in particular the study of stochastic processes. For example, several random variables from the same population converge to the same probability.

In this article, we assume that (X_i) is a sequence of real

random variables, and that all these variables are defined on the same probability space.

Let $(\boldsymbol{F_0, F_1, F_2, \ldots\ldots, F_{N-1}})$ be the repartition functions, associated with random variables $(\boldsymbol{X_0,\ X_1,\ X_2\ \ldots\ldots, X_{N-1}})$. **F** is the repartition function of the random variable $\boldsymbol{X}$. In other words, $\boldsymbol{F_i(x)}$ is defined by $\boldsymbol{F_i(x) = P(X_i \leq x)}$ and $\boldsymbol{F}$ by $\boldsymbol{F(x) = P(\ X \leq x)}$. It is said that $\boldsymbol{X_{N-1}}$ converges to $\boldsymbol{X}$ in probability if :

$$\boldsymbol{\lim_{N\to\infty} P(|X_N - X| \geq \epsilon) = 0} \tag{9}$$

The distributions $(\boldsymbol{X_0,\ X_1,\ X_2\ \ldots\ldots, X_{N-1}})$ are considered as a set of independent random variables and are measured at each constraint spatial sampling step. In principle they should converge at a given time. [Our goal is to reduce the iteration numbers and then the cost in time of molecular dynamic simulation.

Student's **tTest** [21] can be used to statistically test the hypothesis of equality of random variables average following a normal distribution and with unknown variance.

So the bilateral symmetrical confidence interval for the mean $\boldsymbol{\mu}$ can be written as:

$$\boldsymbol{IC(n) = \left[\bar{x} - t_{1-\alpha/2}^{n-1}\cdot\sqrt{\frac{S}{n}}\ , \bar{x} + t_{1-\alpha/2}^{n-1}\cdot\sqrt{\frac{S}{n}}\right]} \tag{10}$$

where

tTest is the fractile of order $\boldsymbol{1 - \frac{\alpha}{2}}$ of St(n-1) Student Law. Most **tTest** statistics have the form: $\textbf{tTest} = \frac{\mathbf{Z}}{\mathbf{s}}$, where **Z** and **s** are functions of the data. Typically, $\boldsymbol{Z}$ is designed to be sensitive to the alternative hypothesis (its magnitude tends to be larger when the alternative hypothesis is true), whereas *s* is a scaling parameter that allows the distribution of **tTest** to be determined.

And

$\bar{x} = \frac{1}{n}\ \sum_{i=1}^{n} x_n$ is the average estimator

$\boldsymbol{s = \frac{1}{n-1}\ \sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the unbiased estimator of the variance.

B. Kolmogorov-Smirnov (ksTest) criteria

The **two-sample ksTest** is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions [12][16].

The Kolmogorov–Smirnov test may also be used to test if two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is:

$$\mathbf{D_n = \ sup_x|F_n(x) - F(x)|} \tag{11}$$

where $\mathbf{sup_x}$ is the supremum of the distances set. If the sample comes from a distribution F(x), then $\mathbf{D_n}$ converges to 0 almost surely**.** Kolmogorov strengthened this result, by effectively providing the rate of this convergence.

$$\mathbf{F_n(x) = \frac{1}{n}\sum_{i=1}^{K} I_{X_{i\leq x}}} \tag{12}$$

Where $\mathbf{I_{X_{i\leq x}}}$ is the indicator function, equal to 1 if $\mathbf{X_i \leq x}$ and equal to 0 otherwise.

Under null hypothesis, the sample comes from the distribution **F(x)** and $\sqrt{\mathbf{n}}\mathbf{D_n}$ converges to the Kolmogorov distribution, which does not depend on **F**. This result may also be known as the **Kolmogorov theorem** [12].

C. One-way ANOVA test criteria

In statistics, one-way ANalysis Of VAriance (abbreviated one-way ANOVA) is a technique used to compare mean-values of two or more samples.

ANOVA is a collection of statistical models used to analyze the differences between group mean-values and their associated procedures (such as "variation" among and between groups). In ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the mean-values of several groups are equal, and therefore generalizes tTest to more than two groups.

The normal-model based ANOVA analysis assumes the independence, normality and homogeneity of the variances of the residuals. **Significance testing** ANOVA is a good example for explaining why very many statistical tests represent ratios of explained to unexplained variability. Here, we base this test on a comparison of the variance due to the between-groups variability (called *Mean Square Effect*, or MS_{effect}) with the within-group variability (called *Mean Square Error*, or Ms_{error}). Under the null hypothesis (that there are no mean differences between groups in the population), we would still expect some minor random fluctuation in the mean-values for the two groups when taking small samples. Therefore, under the null hypothesis, the variance estimated based on within-group variability should be about the same as the variance due to between-groups variability. We can compare those two estimates of variance via the *F* test (see also Fisher Distribution), which tests whether the ratio of the two variance estimates is significantly greater than 1. When a statistical test provides a highly significant ratio, we can conclude that the mean-values for the two groups are significantly different from each other.

VI. ALGORITHM ARCHITECTURE:

The Proposed algorithm is :

Step_1: Production by the Program XLeap from Amber of the Coordinate Topology Files of the complex system under study.

Step_2: Initializing input parameters (initial and final distance coordinates of the complex system.

Step_3: Calculation calibration process step.

Step_4: Production step in which molecular dynamic simulation provides data-distribution. Each iteration have to take into account the output file "restart " of the previous step as a coordinate file (Input CRD File).

Step_5: Data outputs Collection, namely the distribution of distance from the file dump.

Step_6: Subband decomposition of collected data and thresholding high frequency components. Or preprocessing by a gaussian filter of the collected data distribution.

Step_7: Data Reconstruction after the suppression of the micro-states noise. Or collecting the smooth data distribution.

Step_8: Evaluation of the convergence statistic-criteria. If the probability distribution doesn't converge then begin the process since step_4. If not the program goes on.

Step_9: Recording of production results and incrementing of the US step.

Step_10: If the final distance is not reached, then loop from step_3, if not, stop the process.

VII. Simulations and Results:

Our simulations have been made on two systems:

-**System_1**: a mixture of one molecule of NaCl (saline) and water in a box of 12 Angstroms long.

-**System_2**: Deca-Alanine, a peptide consisting of ten residues of alanine with an alpha helix form, in vacuum.

Distance distributions are considered by assuming independent random variables following a normal law probability. Their number at a given constrained distance depends strongly on the used convergence criteria.

Distance distributions are considered by assuming independent random variables following a normal law probability. Their number at a given constrained distance depends strongly on the used convergence criteria.

A. Simulation number one:

Below we give the results of the PMF in the case of System_1 and for windows of 16000 steps in time. The spatial sampling is 0.2 Angstrom. The sampling time step is 0.002 ps. The initial distance is equal to 0.5 Angstroms and the final distance is equal to 10 Angstroms. For System_2 we have the same parameters except for the distance varying from 12 to 35 Angstroms.

In this first simulation, we used a simple and effective criteria which is a parametric statistical test of normality, leading to the Confidence Interval of normality with a given confidence level [4,5]. This confidence interval contains 99% of the population when the distribution is following a normal (or Gaussian) law of probability. The character of normality is given by the probability:

$$\mathbf{P}(\mathbf{IC_{Max}} \leq \mathbf{X} \leq \mathbf{IC_{Min}}) \quad (13)$$

$$\textbf{and } \mathbf{IC_{Max}} = \mu + \mathbf{3\sigma} \textbf{ when } \mathbf{IC_{Min}} = \mu - \mathbf{3\sigma}$$

From this, we deduced the following parametric criteria:

$$\mathbf{Criteria\_1} = |\mathbf{IC_{1_{Min}}} - \mathbf{IC_{2_{Min}}}| \quad (14)$$

$$\mathbf{Criteria\_2} = |\mathbf{IC_{1_{Max}}} - \mathbf{IC_{2_{Max}}}| \quad (15)$$

Below we give the PMF simulation in the case of System_1 whose window size is 16000 temporal steps. The convergence normality criteria are: Criteria_1 and Criteria_2. Their values are each fixed to 0.2, 0.1, and 0.05.

Below we give the PMF simulation in the case of System_1 whose window size is 16000 temporal steps. The convergence normality criteria are: Criteria_1 and Criteria_2. Their values are each fixed to 0.2, 0.1, and 0.05.

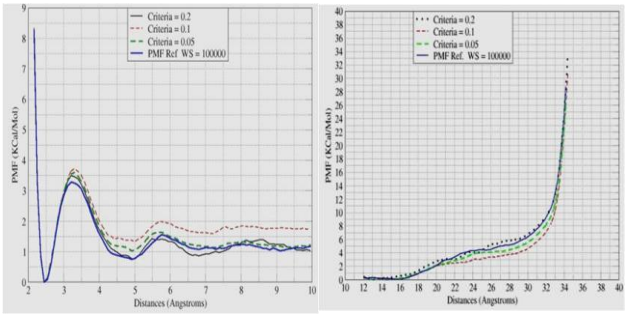

Fig2.: On the left we give the PMFs of system_1, on the right those of system_2; these for different thresholds of the normality criteria. The Umbrella Sampling Window Size (USWS) is fixed to 16000. The blue curve represents the Reference PMF (Ref. PMF) whose USWS is fixed respectively to 40000 for system_1 and 100000 for system_2. All these curves except the Ref. PMF are obtained after the convergence of their probability.

The following figure shows the time cost of these results in function of the threshold criteria values and the Umbrella Sampling Window size.

	System_1		System_2	
Criteria Threshod Values	USWS = 16000	Ref. USWS = 40000	USWS = 16000	Ref. USWS = 100000
0.2	3.232 ns	7.680 ns	40.704 ns	46.400 ns
0.1	4.576 ns		40.736 ns	
0.05	7.296 ns		42.240 ns	

Fig3.: Time costs of simulations for different test-threshold values and sizes of the Umbrella Sampling Window Size.

Observation and Discussion:

we can note that the computation time is low for the criteria threshold value of 0.2 and with PMF close to the PMF reference for both system_1 and sytsem_2 complexes. The used normality convergence criteria indicate that improvement in terms of computation time is possible for adequate test criteria threshold values. Thermal agitation induces duplication of information related to the remaining degrees of freedom. A non-optimal threshold does not filter some parasite micro-states which make information redundant, so the knowing of all the micro-states is not required to measure the PMF. In this context we assume that pre-processing the distributions could improve the convergence speed.

B. Simulation number two:

The purpose of this simulation is to see the impact of the pre-processing by a low-pass Gaussian filter and statistical convergence criteria such as Anova, tTest and skTest. The simulation conditions are identical to that of the previous case.

Note that the pre-processed distributions by a gaussian low pass filter brings a significant gain in terms of time of the probability convergence. The simulation time is reduced by a

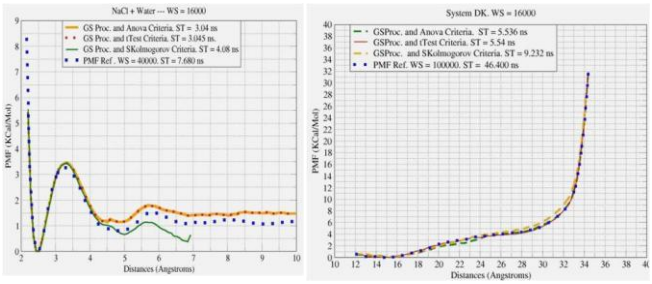

Fig4.: On the Left and right, we give respectively the PMF of system_1 and system_2 for different statistical conver-gence criteria. The distributions of micro-vibration states are pre-processed by Gaussian low-pass filter. One can see the duration of each simulation (Simulation Time=ST) according to statistical criteria. The window size (USWS) is set to 16000 steps for each simulation. Also one gives the reference curve PMF (Ref. PMF) whose size is set to 40000 for system_1, and 100000 for system_2.

factor of two for system_1 and 10 for system_2 when the convergence criteria such as ANOVA and tTest are used. Results of skTest are less satisfactory.

C. Simulation number three:

Here distributions obtained during a sampling time at a given distance are pre-processed by a denoising technique. It involves a wavelet methodology where high frequencies are thresholded (see Fig5.). The used wavelet belongs to the family of bio-orthogonal wavelet (Cubic Spline) and the chosen thresholding method is a universal soft thresholding technique. In order to see the influence of this denoising process, distributions are decomposed and thresholded until level 8. Then the signal is reconstructed and its probability is compared to the previous one by using different convergence criteria such as : skTest, tTest and finally Anova Test.

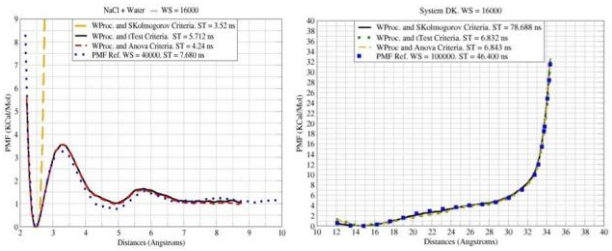

Fig5.: On the left and right we give respectively the PMFs of system_1 and system_2. The PMF curves are represented as a function of distance. The Umbrella Sampling Window Size (USWS) is fixed to 16000. The blue curve (…) represents the Reference PMF (Ref. PMF) whose USWS is fixed to 40000 and 100000 respectively for system_1 and system_2.

Observation and Discussion:

The PMF values in case of multi-scale distributions analysis are shown above (Fig5.). We can see that the cost in computation time is generally less than those of PMFs references. In system_1, a gain of 2ns to 3ns was observed in the case of tTest and ANOVA test criteria. However in the

case of skTest, divergence was observed at the beginning of the simulation.

For the Alanine complex, the computation time of each simulation is measured about 7 times lower than the reference simulation one (46.400 ns). This is in the case of tTest and ANOVA test. Regarding skTest, the computation time increased significantly (about twice the reference time).

It is also noted for the system_2 that convergence is almost perfect with respect to the PMF reference for all statistical tests.

VIII. Conclusion

The performance of the algorithm thus developed, is based on the efficiency of the stochastic convergence criteria, combined with an adaptative pre-processing. This allows to use a succession of short size windows iterated until a convergence. One reminds that a long Umbrella Sampling Window for a big protein complex leads to several days of calculation.

Here distributions of reaction coordinates are pre-processed by a methodology involving a low pass-filter, or a wavelet soft thresholding. It significantly improves and reduces the computation time of the simulations.

The results show that fact of pre-processing all micro-vibrations of high frequencies reduces the computation time. The Umbrella Sampling Method with short windows is used to perform a constrained temporal sampling.

We can assume that the use of narrow bandwidth filters eliminates some useful information (some vibrational frequencies) from the entropy, and then causes the PMF to take inifinite values.

The given algorithm unquestionably allows a reduction of the iteration number and so the time of simulation.

VIIII. REFERENCES

[1] A., Azzalini, M. Farge, and K. Schneider. "Nonlinear wavel et thresholding: a recursive method to determine the optimal denoising threshold," Appl. Comput. Harmon. Anal. 18(2), pp.177–185, Oct. 2004.

[2] V. Bagdonavicius, J. Kruopis, M.S. Nikulin, "Non-parametric tests for complete data," ISTE&WILEY: London&Hoboken, 2011.

[3] H. O. Bartelt, K. H. Brenner, and A.W. Ohman, "The wigner distribution function and its optical production," *Optics Communications*, 32, pp. 32-38,1980.

[4] G. W. Corder & D. I. Foreman, "*Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*," Wiley, 2009.

[5] M. Farge, "Wavelet transforms and their applications to turbulence," Ann. Rev. Fluid Mech. , 24:395–457, 1992.

[6] Gibbons, Jean Dickinson and Chakraborti, Subhabrata "*Nonparametric Statistical Inference*," 4th Ed. CRC, 2003.

[7] T. P. Hettmansperger, J. W. McKean, "*Robust nonparametric statistical methods,"* Kendall's Library of Statistics. 5 (First ed.). London: Edward Arnold.

[8] I. Johnstone and B. Silverman, "Wavelet treshold estimators for data with correlated noise," J. R. Stat. Soc., Ser. B. Stat. Methodol. 59,pp.319–351,1997.

[9] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P.A. Kollman P. A., "Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method," Journal of Computational Chemistry, 16, 11, pp.1339-1350. 1995.

[10] L. Charlier, "Etude des Interactions moléculaires entre molécules odorantes et protéines impliquées dans les premières étapes de la perception olfactive," Thèse de Doctorat. Université de Nice-Sophia Antipolis, Octobre, 2009.

[11] Lehmann E.L., Scheffé H. "Completeness, similar regions, and unbiased estimation," Sankhya: the Indian Journal of Statistics 10 (4), pp. 305–340,1950.

[12] Marsaglia G., Tsang WW., Wang J., "Evaluating Kolmogorov's Distribution," Journal of Statistical Software, 8(18), pp1-4,2003.

[13] Masunov A., and Lazaridis T., "Potentials of mean force between ionisable amino acid side chains in water," J. Am. Chem. Soc. 125, pp1722-1730,2003.

[14] Mills Maria and Andricioaei Loan, "An experimentally guided umbrella sampling protocol for biomolecules," J. Chem. Phys. 129(11): 114101. 2008.

[15] Mostacci E., Truntzer C., Cardot H., Duoroy P., "Méthodes multivariées combinant ondelettes et analyse en composantes principales pour le débruitage de données issues de spectrométrie de masse," *42èmes Journées de Statistique*, Marseille, France. 2010.

[16] Pratt J.W., and Gibbons, J.D., "Concepts of Nonparametric Theory," New York: Springer Verlag, 1981.

[17] Roux B., "The calculation of the potential of mean force using computer simulations," Comp. Phys. Comm., 91, pp.275–282. 1995.

[18] Shorack Galen R., and Wellner Jon A., "*Empirical Processes With Applications to Statistics*," Philadelphie, Society for Industrial & Applied Mathematics, pp.998,2009.

[19] Torrence, C., and Compo G.P., "A Practical Guide to Wavelet Analysis", Bulletin of the American Meteorological Society, 79, pp61-78,1998.

[20] Torrie G. M., and Valleau J. P., "Monte Carlo free energy estimates using non-Boltzmann sampling: application to the sub-critical Lennard-Jones fluid," Chem. Phys. Let., october 1974.

[21] Wasserman, Larry, "*All of nonparametric statistics*," Springer, 2007.

[22] Sawilowsky, Shlomo S., "Fermat, Schubert, Einstein, and Behrens–Fisher: The Probable Difference Between Two Means When $\sigma_1 \neq \sigma_2$,". Journal of Modern Applied Statistical Methods 1 (2), pp.461–472.2002.

[23] Silverman B. W.. "Wavelets in statistics: beyond the standard assumptions, Phil. Trans. R. Soc. Lond. A, 1999.