Processing of ICR transients with non-FT Genetic Algorithm

Marc HAEGELIN

Miniaturization for Synthesis, Analysis & Proteomics UAR 3290

EU_FTICR EUS December 12-16 2022, Lille

University of Lille

Presentation plan

FT-ICR mass spectrometry

- FT-ICR is a high-resolution mass spectrometer
- Signals are exponentially decaying sine waves and white noise

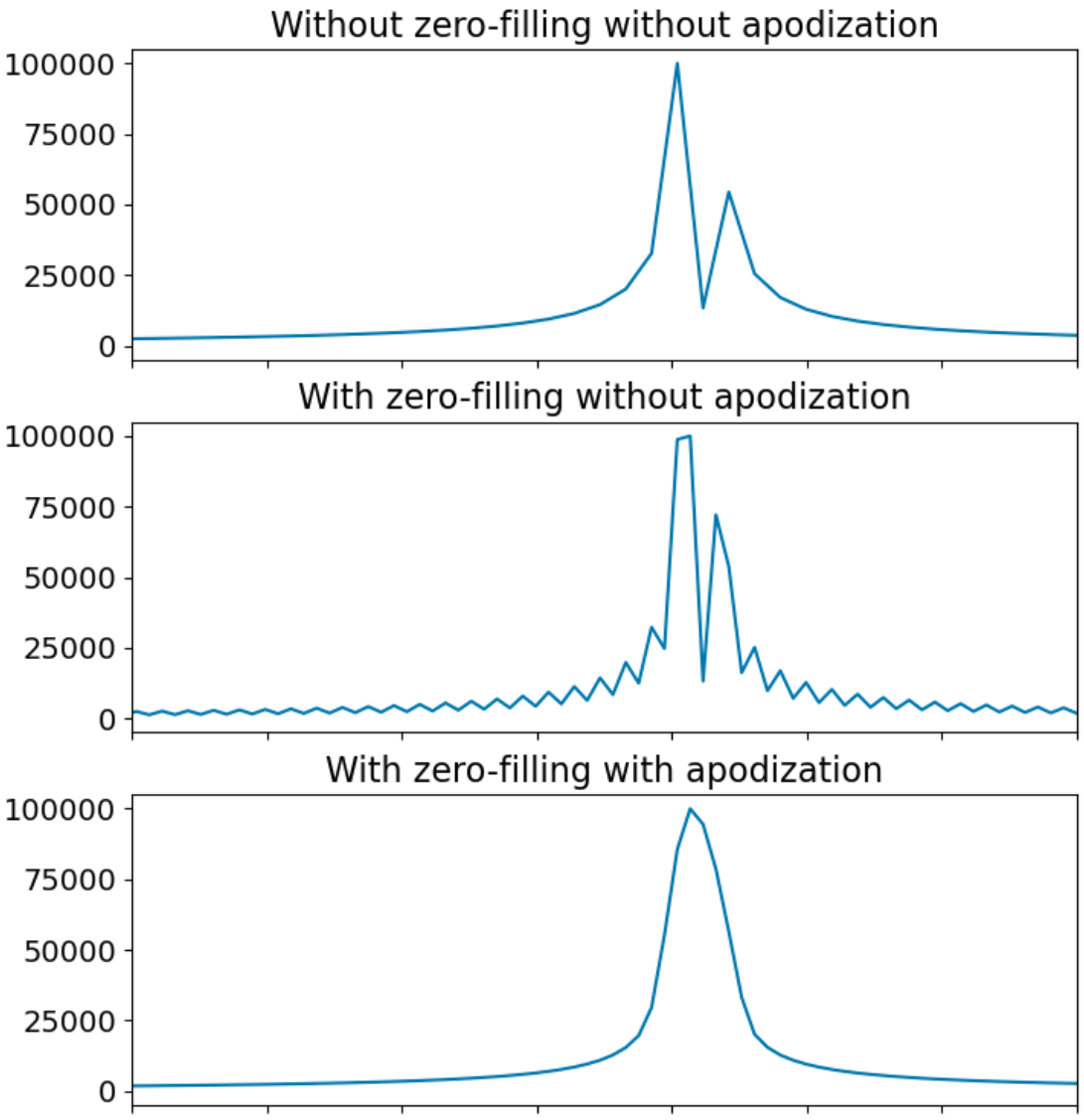
$$s(t) = \sum_{n=0}^{K} [\, e^{-dec_n t} amp_n . sin(\, 2.\pi.freq_n . t + ph_n) \,] + \xi(t)$$

- The signal is zero-filled, apodized, then Fast Fourier Transformed (FFT)

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-2i\pi n \frac{k}{N}}$$

FT-ICR MS SolariX XR 9.4 Tesla (Bruker Daltonics)

Off-grid and on-grid frequencies

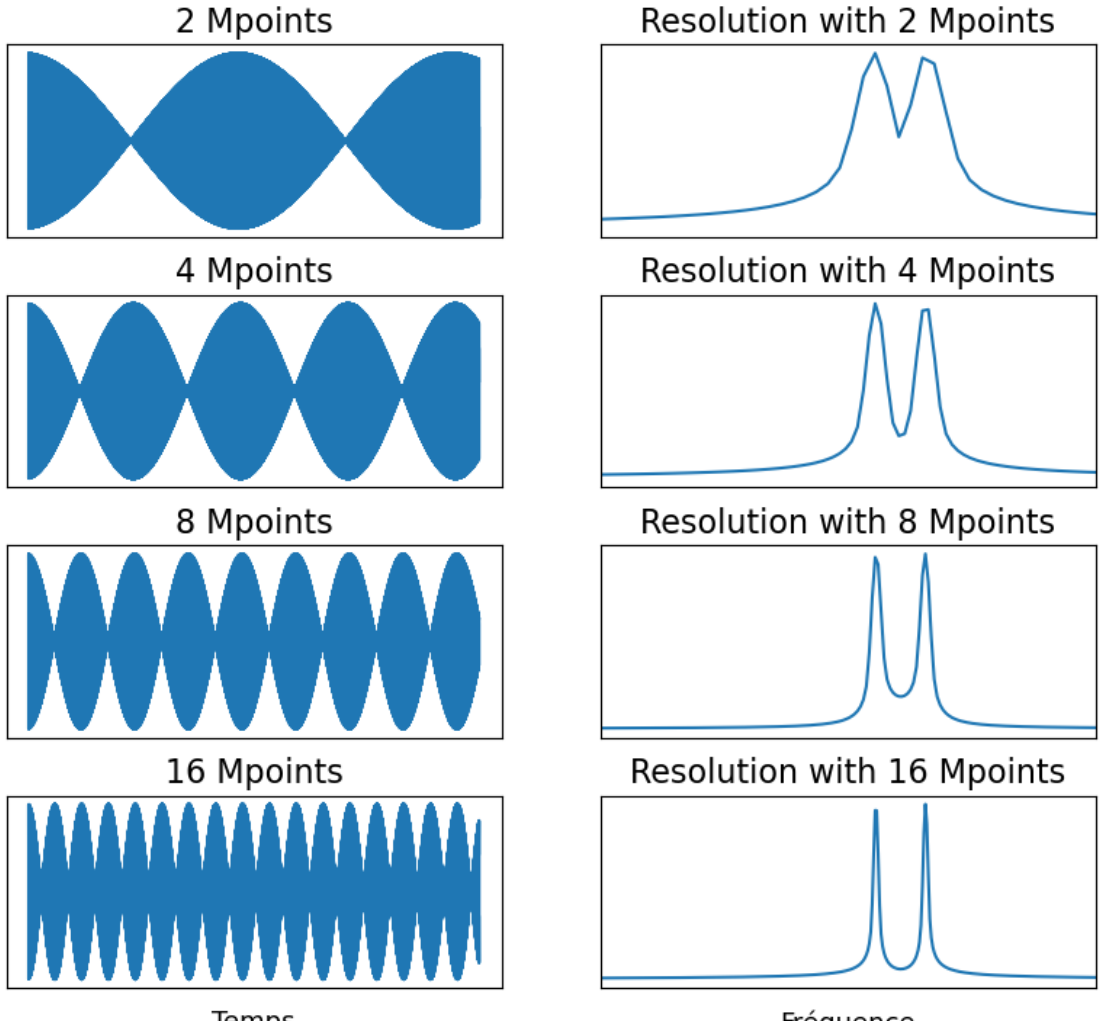


- FFT has a grid of width 1/N with N points
- Frequencies can be either on-grid or off-grid, which makes the frequencies harder to be extracted
- The spectrum can be smoothed thanks to zero-filling (which creates side-lobes that can hide small peaks of the distribution)
- The apodization removes the side-lobes (but modifies the mass isotopic ratios and widens the peaks)

EU FT-ICR MS MSAP cnrs Université de Lille

Resolution limit

Low resolving power



High resolving power

- Resolution is defined as the full width at half maximum (FWHM) of the peak
- Resolution is proportional to the number of points
- Higher number of points means longer acquisition times
- Zero-filling smooths the transient but does not improve resolution

EU FT-ICR MS

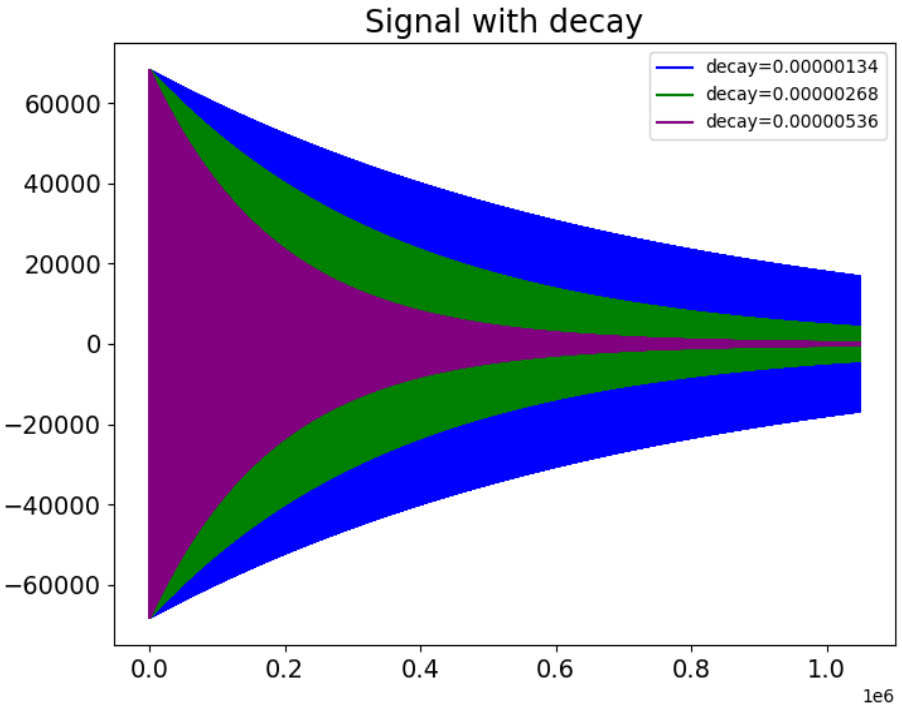
MSAP

cnrs

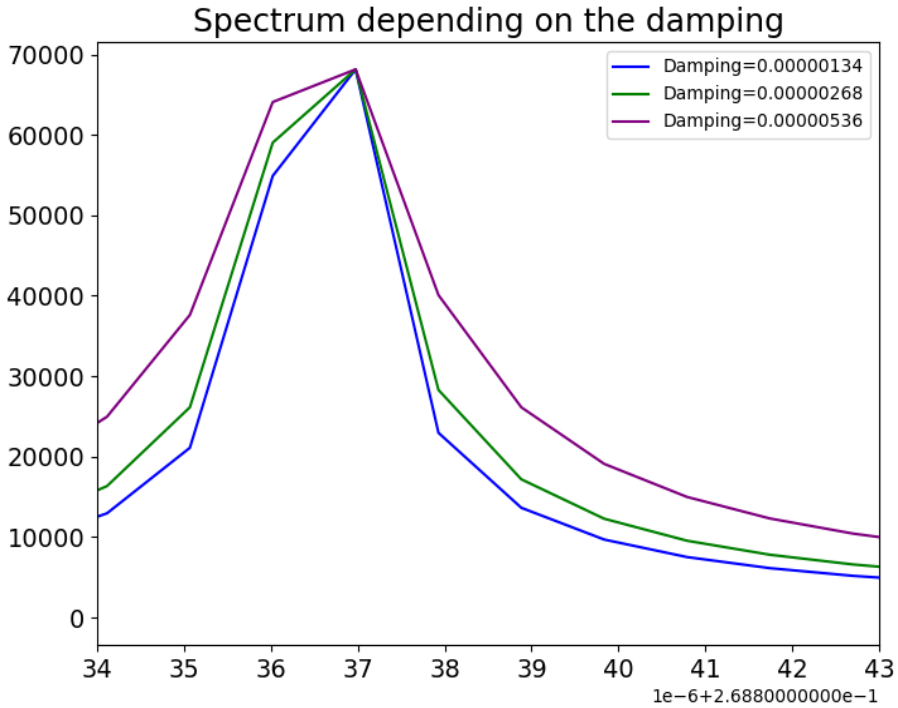
Université de Lille

Decay influence on peak width

- FFT of exponentially decaying sines results in lorentzian peaks shape whose width are wider with higher decays and whose phases are modified



FFT



Time domain

Frequency domain

EU FT-ICR MS

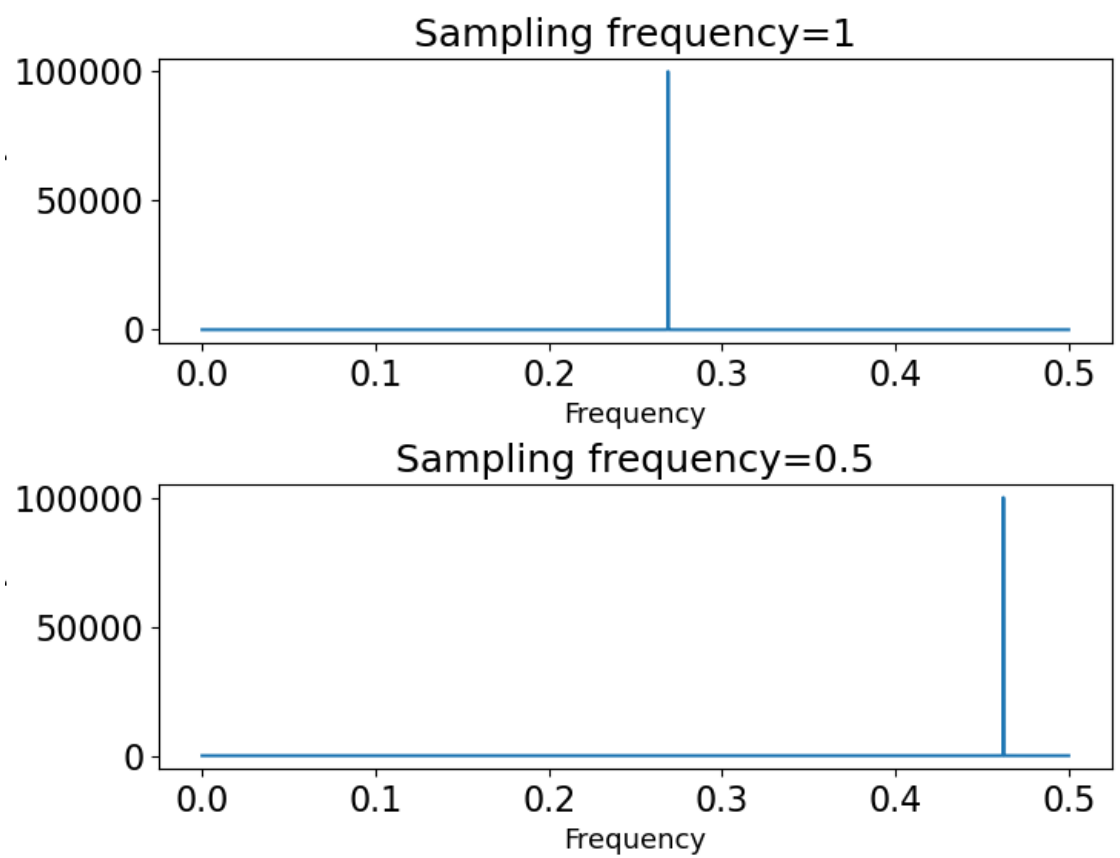
MSAP

cnrs

Université de Lille

Aliasing

- Shannon-Nyquist sampling theorem states that we need to sample at least at twice the highest frequency in the spectrum (Nyquist frequency) otherwise there is a peak shift called aliasing



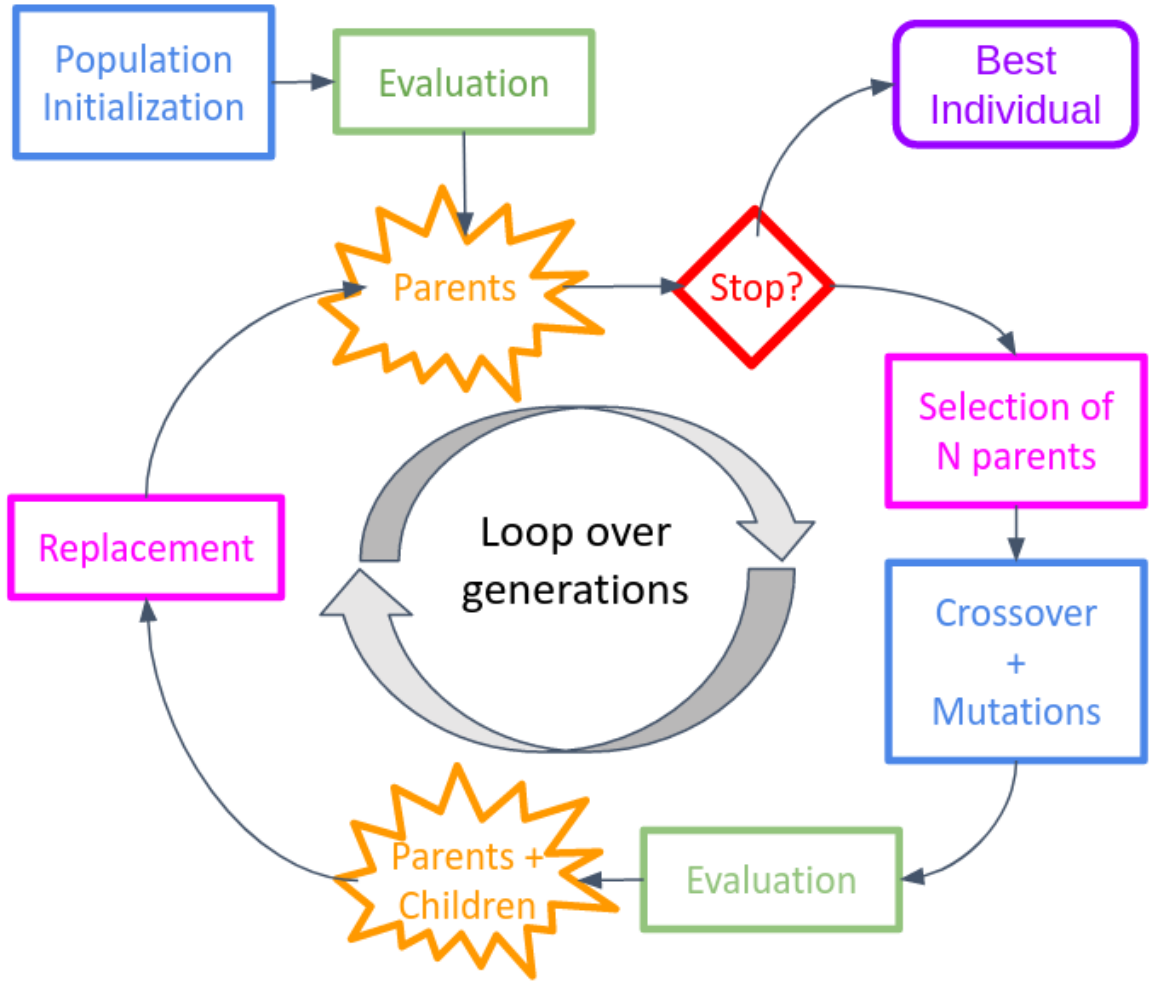
One signal peak which represents
100000.0 sin(2π . **0.26883656** . t + 1.4112345)
depending on the sampling frequency

NB : 0.25 < 0.26883656 < 0.5

Presentation plan

Evolutionary algorithm (EA)

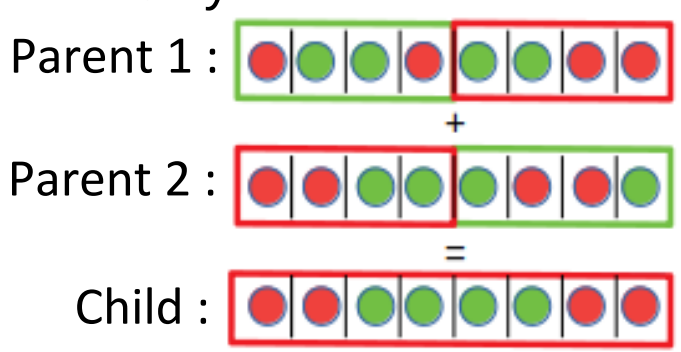
- Random initialization of the parents
- Evaluation of the parents
- For each generation i
  - Select best individuals from i-1th generation
  - Crossover with probability p_c
  - Mutation with probability p_m
  - Evaluation of the children
  - Reduction of the population to original size
- Stopping criterion
  - Time
  - Nb. of generations
  - Convergence



Example of an EA

- An individual is composed of red and green balls

- We want the individual with only green balls so we select the best individuals and recombine



Each parent is cut into two and exchange its second part with the other to produce children

- Then we mutate the children (red balls can become green or green balls become red)

- Finally we reduce the population until we get the individual with only green balls and stop

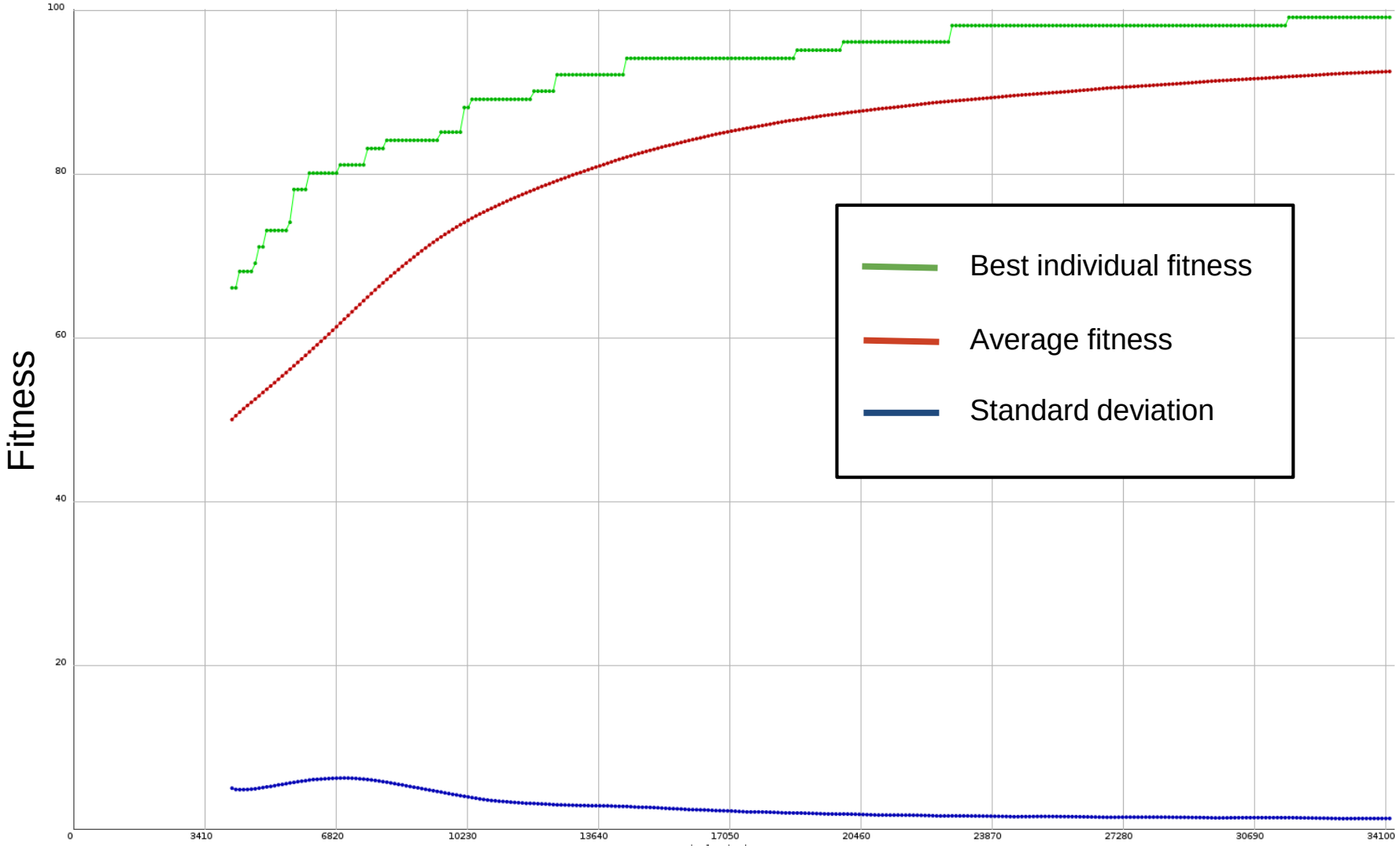
EU FT-ICR MS

MSAP

cnrs

Université de Lille

Execution of this simple algorithm (100 balls)



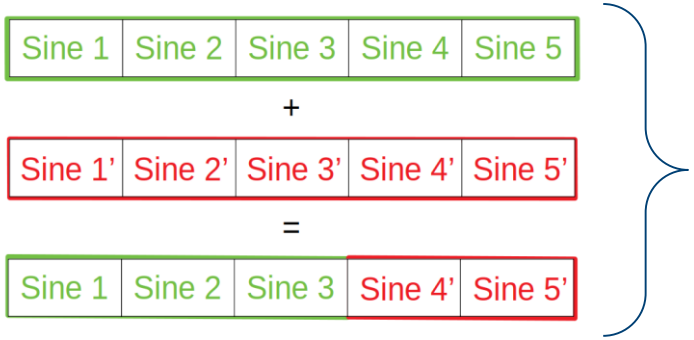
	Number of trials to reach optimum
Random search	≃1.2E30
EA	≃35.000

EAs are **stochastic** not random, like evolution by natural selection

Sinus_it algorithm

- A population is made of individuals composed of N sines (amplitude, frequency, phase, decay)

- We select the best individuals and recombine them with probability p_c



Each parent is cut into two and exchange its second part with the other to produce children

- Then we mutate the children created with probability p_m

- We reduce the population and loop until we reach the stopping criterion

EU FT-ICR MS

MSAP

cnrs

Université de Lille

Materials

- Sinus_it is written in C++/CUDA from Nvidia and the EASENA platform built by CSTB (ICube)
- Sinus_it was tested on the following hardware architecture

	CPU	GPU
Model	Intel Xeon Core i7	1× RTX 2080 Ti
Nb. of cores	48	4352
Frequency	2.1 Ghz	1.8 Ghz
RAM	128 Go	11 Go

Nvidia RTX 2080 Ti Founders Edition

EU FT-ICR MS

MSAP

cnrs

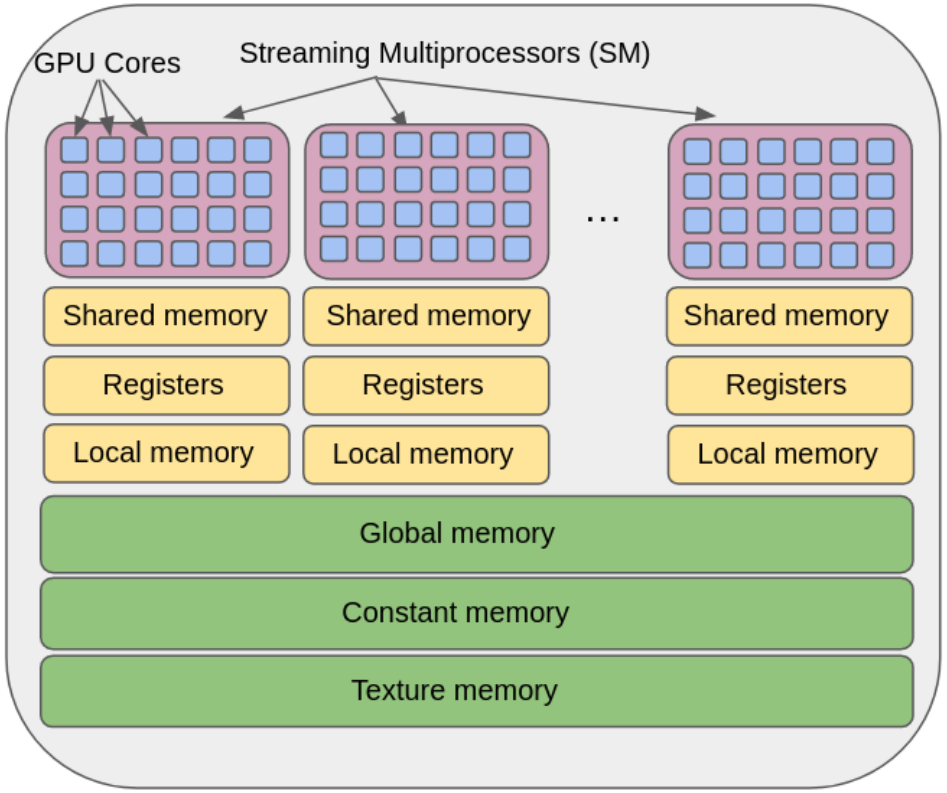
Université de Lille

Parallel programming on GPUs

- GPUs programming is a mix between
  - Single Instruction Multiple Data (SIMD) the same instruction being executed in parallel on multiple data
  - Single Program Multiple Data (SPMD) the same program being executed by the distributed streaming multiprocessors (SMs) independently

- Each SM has dedicated registers, shared memory and local memory, and all of them have access to global memory, constant memory or texture memory

- Each GPU Core has SFUs, Single Precision floating-point units and Double Precision floating-point units



GPU memory hierarchy architecture

EU FT-ICR MS

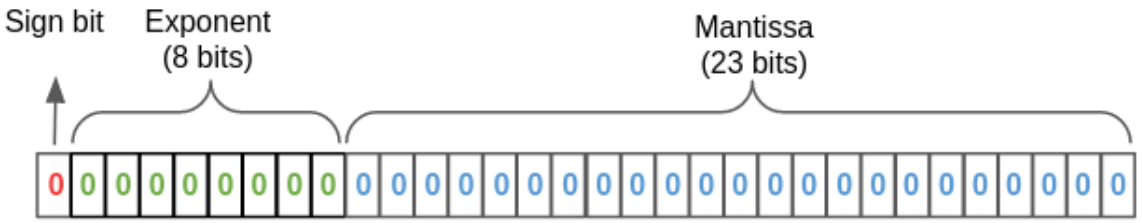
MSAP

cnrs

Université de Lille

Single and double precision

- Floating-point numbers can be represented either in single precision (32 bits) or in double precision (64 bits)



Single precision floating-point representation

- GPU hardware are quicker in single precision (in particular gaming cards)

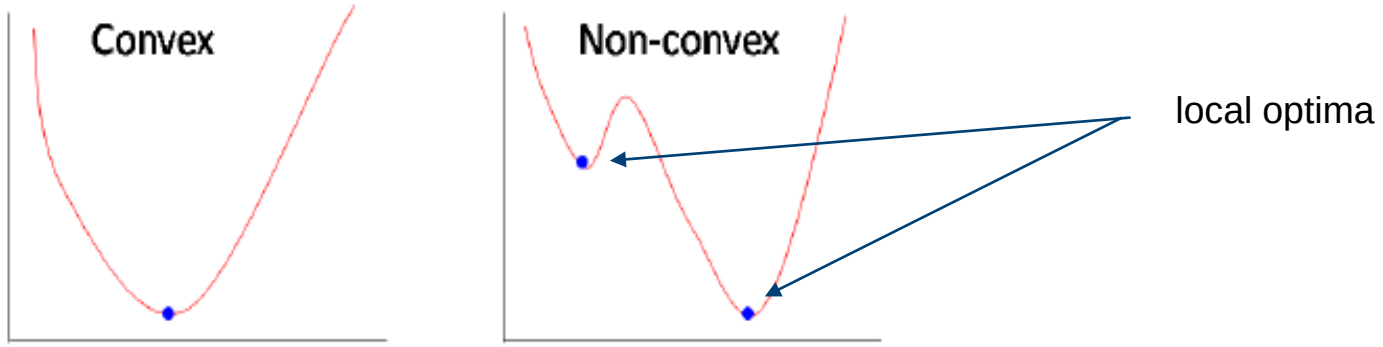
	Nvidia RTX 2080 Ti	**Nvidia P100**
Single Precision	13.5 TFLOPS	9.3 TFLOPS
Double Precision	420 GFLOPS	4.7 TFLOPS

- We use double precision (15.95 significative digits) instead of single precision (7.22 significative digits), there is a tradeoff between accuracy and speed

EU FT-ICR MS MSAP cnrs Université de Lille

Exploration / Exploitation dilemma

- The search space may or may not be convex



- Sinus_it search space is not convex and we want to find the global optimum.
- The evaluation function is the L^2 norm between the approximation of the individual and the signal
- We need to balance between
  - Exploration : Finding new valleys in the search space (initialization, mutation)
  - Exploitation : Finding the local minimum of the available valleys (crossover, selection/reduction)

Changing parameters (1)

- Number of sines
  - Fixed number of sines : the number of peaks is known and fixed a priori which can be used for quantitation, possibly with fixed frequencies.
  - Dynamic number of sines : the number of peaks has an upper limit, starts with 1 sine, seeks to converge, then iteratively increments until there no progress anymore.

- Isotopic structure
  - Coarse mode : Extraction of the M,M+1,M+2 and s.o. isotopic distributions, the amplitudes being the sum of the small peaks within and the other parameters being the weighted average. Retrieval of the phase/frequeny correlation is performed.
  - Fine mode : Extraction of the fine isotopic structure belonging to the peaks of the coarse mode, by using the phase/frequency correlation previously found.

Changing parameters (2)

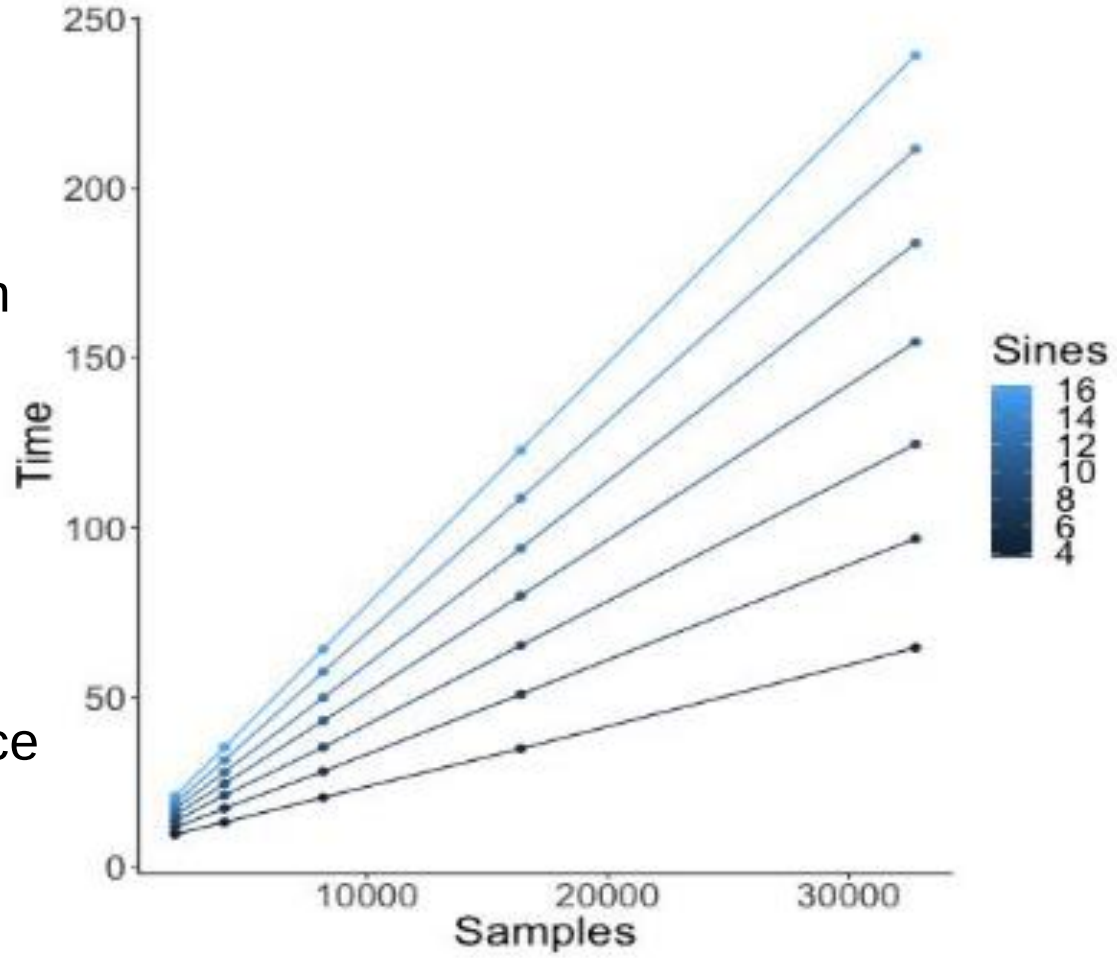
- Boundaries : the amplitudes, frequencies, phases and decay parameters can change within given limits (minimum and maximum values) previously set. It can influence convergence speed of Sinus_it.
- Part of the transient : It is possible to work anywhere in the transient on a small portion of it (useful when the transient has some more imperfections in certain parts than in others).
- Maximal resolution : Estimation of the maximal closeness between peaks (merging sines otherwise)
- Sampling method
  - FULL : All points in the selected part of the transient are considered
  - NUS : Points are randomly sampled at an average rate (for instance 1 point out of 32 is NUS 32)
  - GRS : Begin with a small number of points N and adds N more points every G generations

Stable parameters

- Number of parents and offsprings : It was fixed at 134912 because it is a multiple of the number of threads on the GPU card (64 cores distributed among 68 streaming multiprocessors), hence results in optimal parallelization.
- Selection and reduction : We choose the best individual between 20 randomly chosen individuals in a random draw with replacement (Tournament 20) within parents and offsprings together. There is a balance between exploration and exploitation with the number of individuals.
- Crossover probability : It is chosen to be 1 in order to have the greatest possible exploitation of the solutions.
- Mutation probability : It was set to be 1/4N by parameter for every individual, such that on average only one parameters changes, which avoids destructive mutations and allows not to get stuck in a local minimum.
- Elitism : We chose to always keep the best individual to the next generation within parents and children as a whole, so as to avoid losing good individuals (or even the best individual) when the algorithm reaches it.

Execution time

- Sinus_it execution time is linear with the number of sines and the number of samples
- The number of points can become very huge (it can reach 16 Mpoints = 16×1024×1024 = **16777216** points)
- The number of sines can also increase with the experiment
- The execution time can be huge before convergence on long transients (more than 3 days)



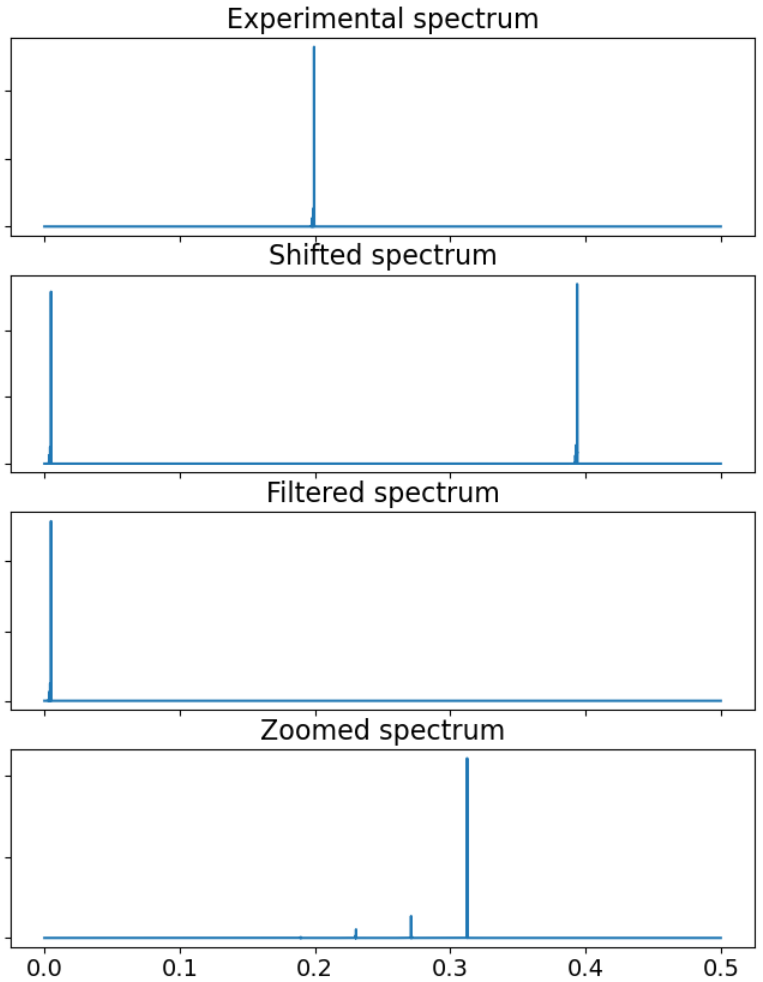
EU FT-ICR MS

MSAP

cnrs

Université de Lille

Narrow-band spectra (zoomFFT)



Experimental glutathione (4 Mpoints = 4×1024×1024 points)
All the peaks are in the area 0.19725-0.19950

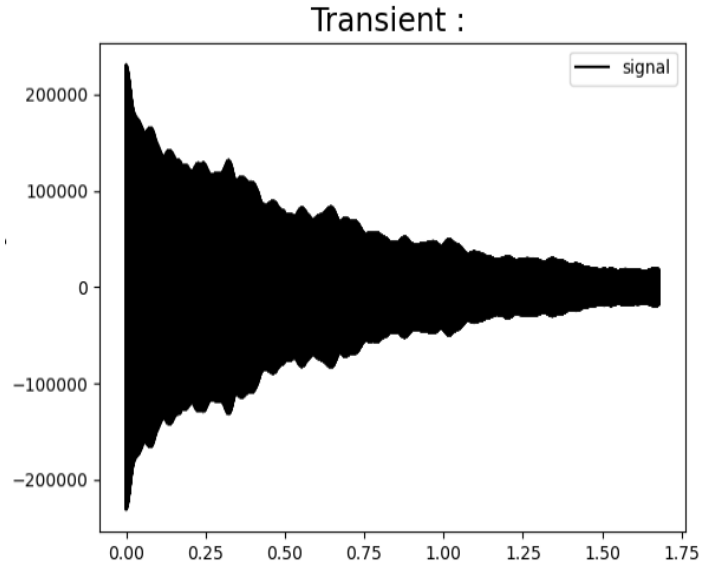
Multiplying by constant sine sine(2×π×**0.19446875**×t) creates the positive and negative components

Low-pass filtering (Butterworth filter with 0 phase)

Decimation by downsampling ratio 64 that creates a 64 kpoints signal
Frequencies are multiplied by 64.
Sinus_it speedup = 64

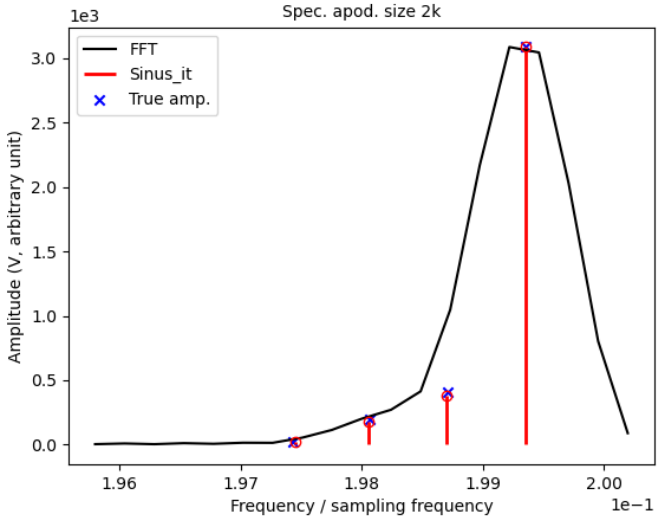
Original frequency = zoomed frequency / 64 + **0.19446875**

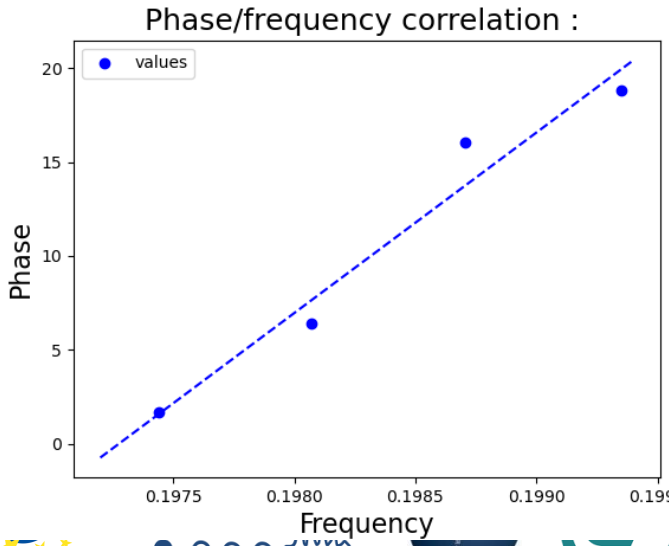
Sinus_it workflow



Apply zoomFFT (butterworth filter) which only keeps sines waves **within a given frequency range** and resolve the peaks with Sinus-it

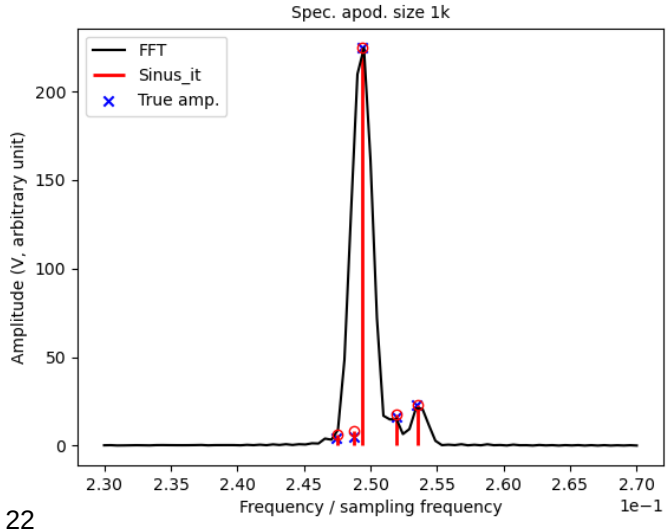
Coarse analysis





Using **phase/frequency correlation** to resolve the fine structure

zoomFFT



Fine analysis

Presentation plan

State of the art (FDM)

- Filter Diagonalization Method : Diagonalization of an Hamiltonian evolution operator

$$C(t) = \sum_k d_k e^{-it\omega_k}$$

$$C(t_n) = (\phi_0, \hat{U}^n \phi_0) = (\phi_0, e^{-int\hat{H}} \phi_0)$$

$$U_{jj'}^{(p)} = (\Psi_j, \hat{U}^p \Psi_{j'}), \; j = 1, 2, \ldots, N_{win}$$

$$U^{(p)} B_k = u_k^p U^{(0)} B_k$$

$$u_k = e^{-it\omega_k}$$

$$d_k = \left[\sum_j \sum_{n=0}^{M-1} c_n z_j^{-n} [B_k]_j \right]^2$$

- ω_k : complex frequencies / d_k : complex amplitudes
- $C(t)$ signal / $\hat{U}$: evolution operator / $\hat{H}$: Hamiltonian
- $\Psi_j, j = 1, 2, \ldots, N_{win}$: basis diagonalizing $\hat{U}^p$ / u_k eigenvalues
- B_k : eigenvectors / z_j complex value on the unit circle

State of the art (compressed sensing)

- Finds the equation of a randomly sampled sparse signal
- Using the L^1 and L^2 norm with a Fourier basis to minimize the expression

$$||s(t) - \lambda_i g_{ij}||_2 + \alpha ||\lambda_i||_1$$

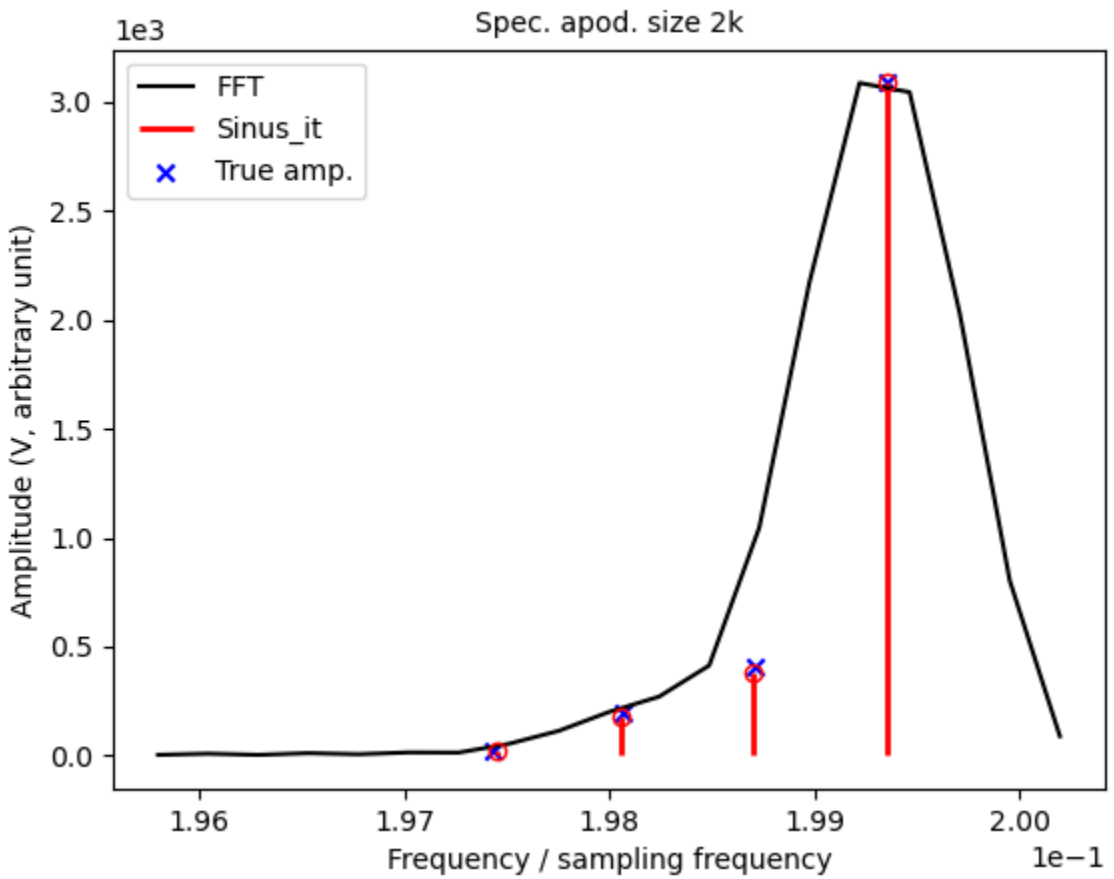
- α the penalty coefficient
- g_{ij} a basis (for instance Fourier basis)
- λ_i the coefficients in that basis
- $s(t)$ the signal

- Bypasses the Shannon-Nyquist limit (can see frequencies at 1/4th of the Nyquist frequency)

Presentation plan

All Sinus_it results shown in the following slides were obtained by processing a 4M points ***glutathione*** *transient recorded at the* ***University of Rouen****, France on a* ***12 Tesla Bruker SolariX*** *FTICR fitted with an harmonized cell by the Pr. Carlos AFONSO team in the frame of the second* ***EU_FT-ICR_MS round robin test****.*

Sinus_it: Glutathione, 2k from full 4 Mpoints transient

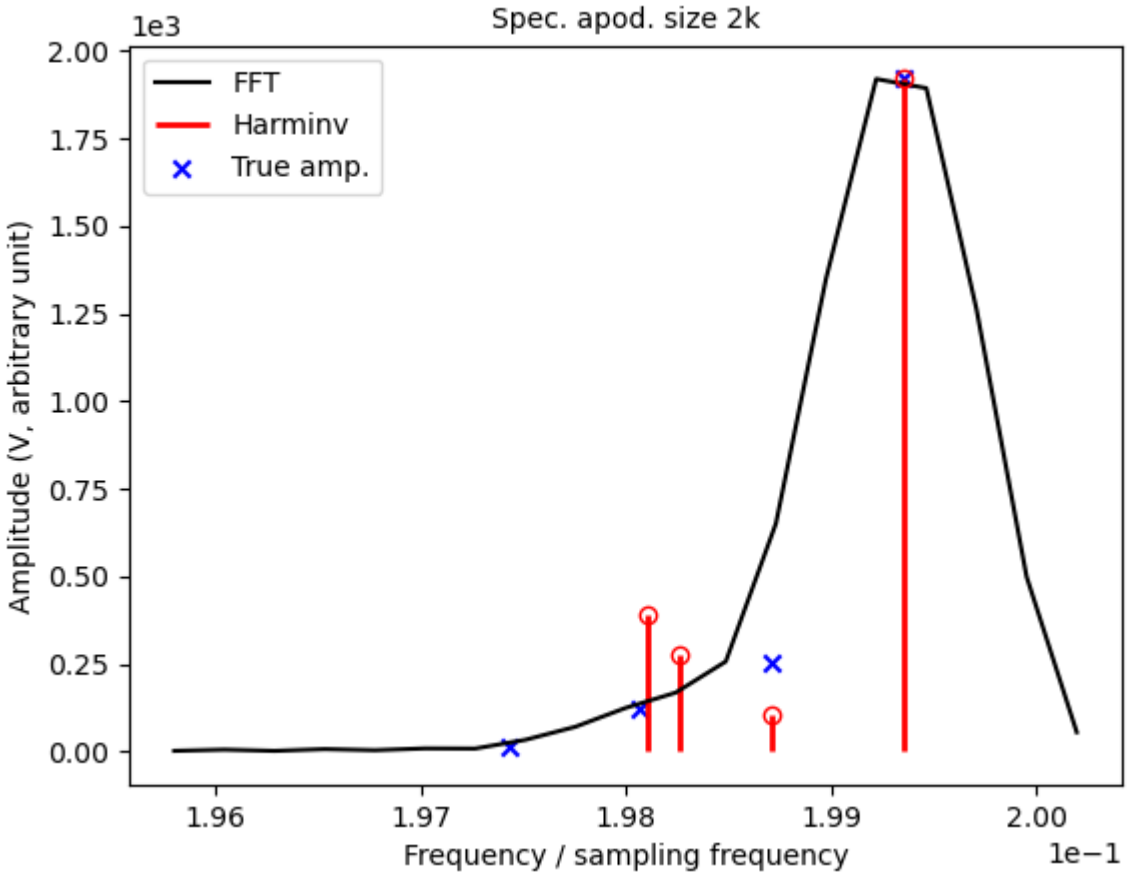


	True	Sinus_it	Error
Amplitude 1	1924.776489	1924.776489	0
Frequency 1	0.199351504	0.199351757	1.2E-6
Amplitude 2	252.859247	235.580429	6.8E-2
Frequency 2	0.19870484	0.19870433	2.5E-6
Amplitude 3	125.228977	110.411521	1.1E-1
Frequency 3	0.198066665	0.19805787	4.4E-5
Amplitude 4	14.320436	14.426076	7.37E-3
Frequency 4	0.197428549	0.197448968	1.0E-4

Mise en forme à reproduire sur toute les tables

Execution time : **2 hours 5 minutes** on an Intel Xeon 2.1 Ghz server with 128 Go of memory RAM, embedded with an Nvidia RTX 2080 Ti graphics card

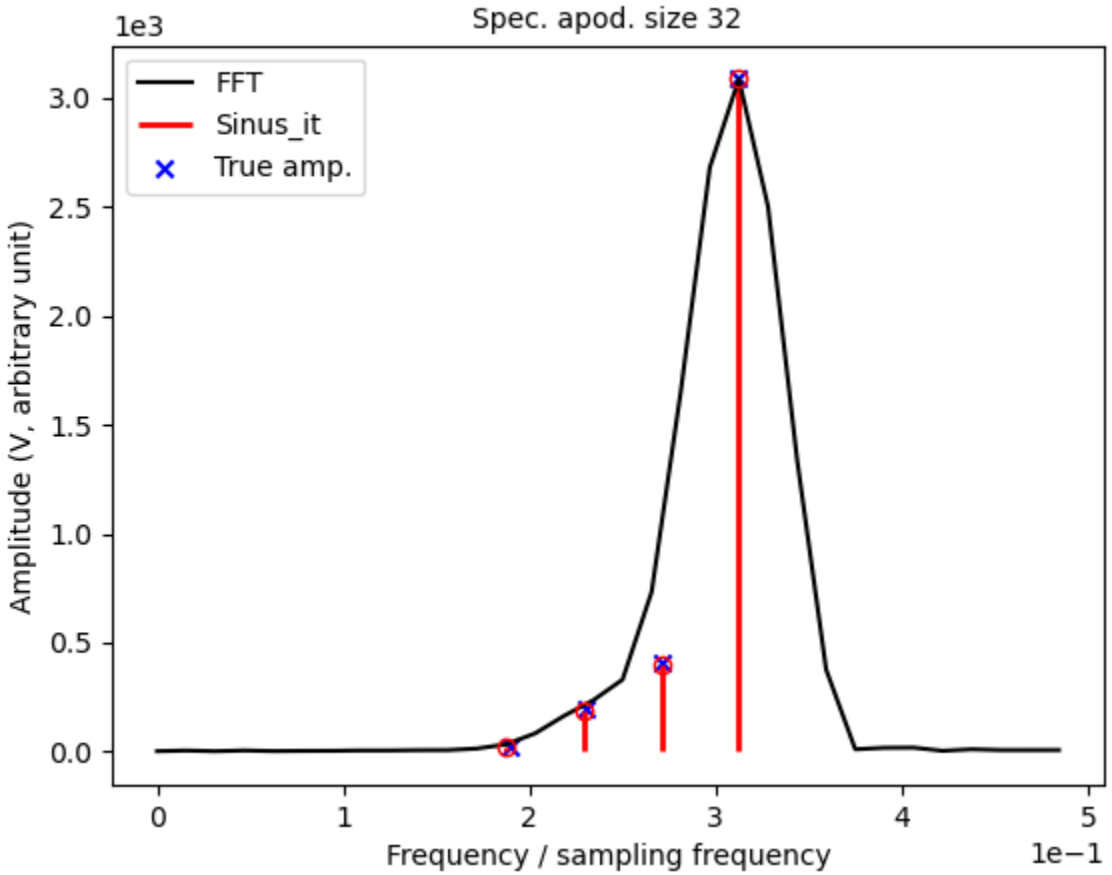
FDM: Glutathione, 2k from full 4 Mpoints transient



	True	FDM	Error
Amplitude 1	1919.474446	1919.474446	0
Frequency 1	0.199351504	0.199351056	2.2E-6
Amplitude 2	252.162713	105.6710597	5.8E-1
Frequency 2	0.19870484	0.198710769	2.9E-5
Amplitude 3	124.884017	279.1216812	1.23
Frequency 3	0.198066665	0.198256436	9.5E-4
Amplitude 4	14.280989	388.4244202	26.19
Frequency 4	0.197428549	0.198107328	3.4E-3

Execution time : **1 minute** on an Intel Xeon 2.1 Ghz server with 128 Go of memory RAM, embedded with an Nvidia RTX 2080 Ti graphics card

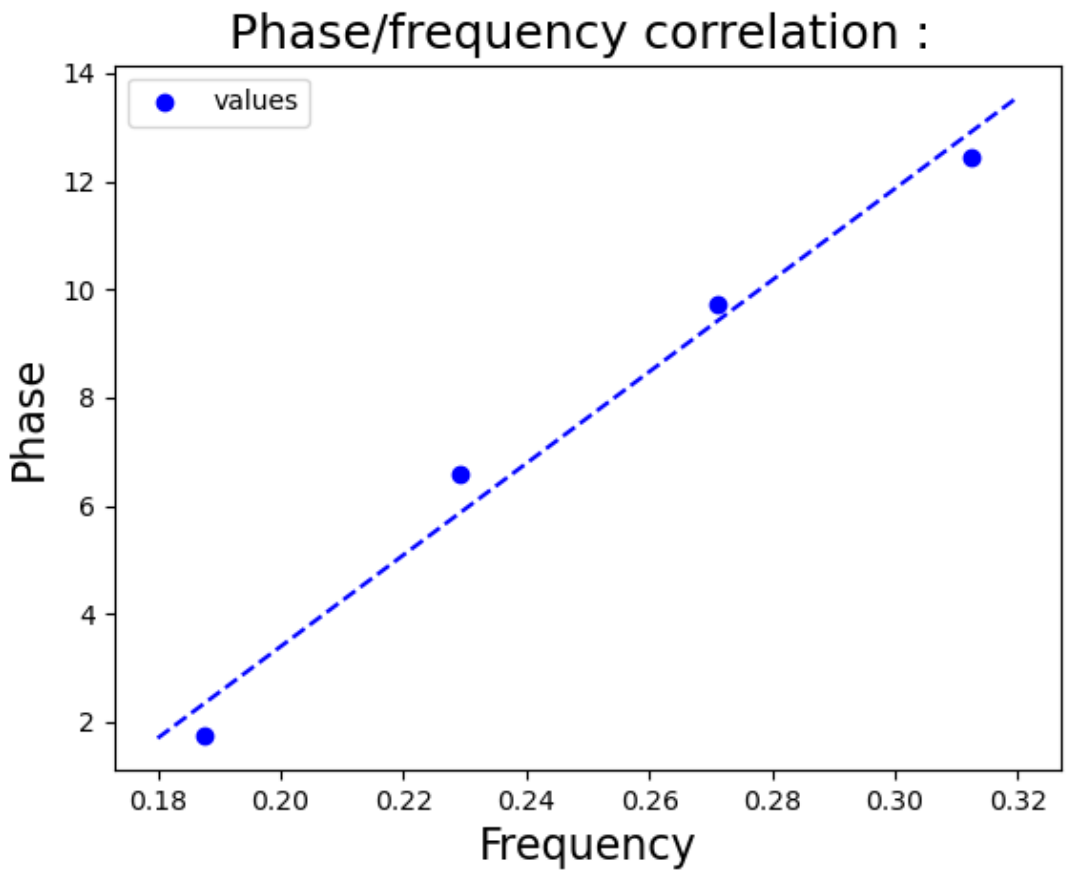
Sinus_it: Glutathione 32 points from 4 Mpoints after ZoomFFT 64 (eq. 2k full)



	True	Sinus_it	Error
Amplitude 1	1943.875122	1943.875122	0
Frequency 1	0.312496284	0.312419056	2.4E-4
Amplitude 2	255.3668747	253.5423583	7.1E-3
Frequency 2	0.271109789	0.271021336	3.2E-4
Amplitude 3	126.685153	117.5560607	7.2E-2
Frequency 3	0.230266530	0.229053959	5.2E-3
Amplitude 4	14.462430	13.31410408	7.97E-2
Frequency 4	0.189427107	0.187607944	9.6E-3

Execution time : **20 minutes** on an Intel Xeon 2.1 Ghz server with 128 Go of memory RAM, embedded with an Nvidia RTX 2080 Ti graphics card

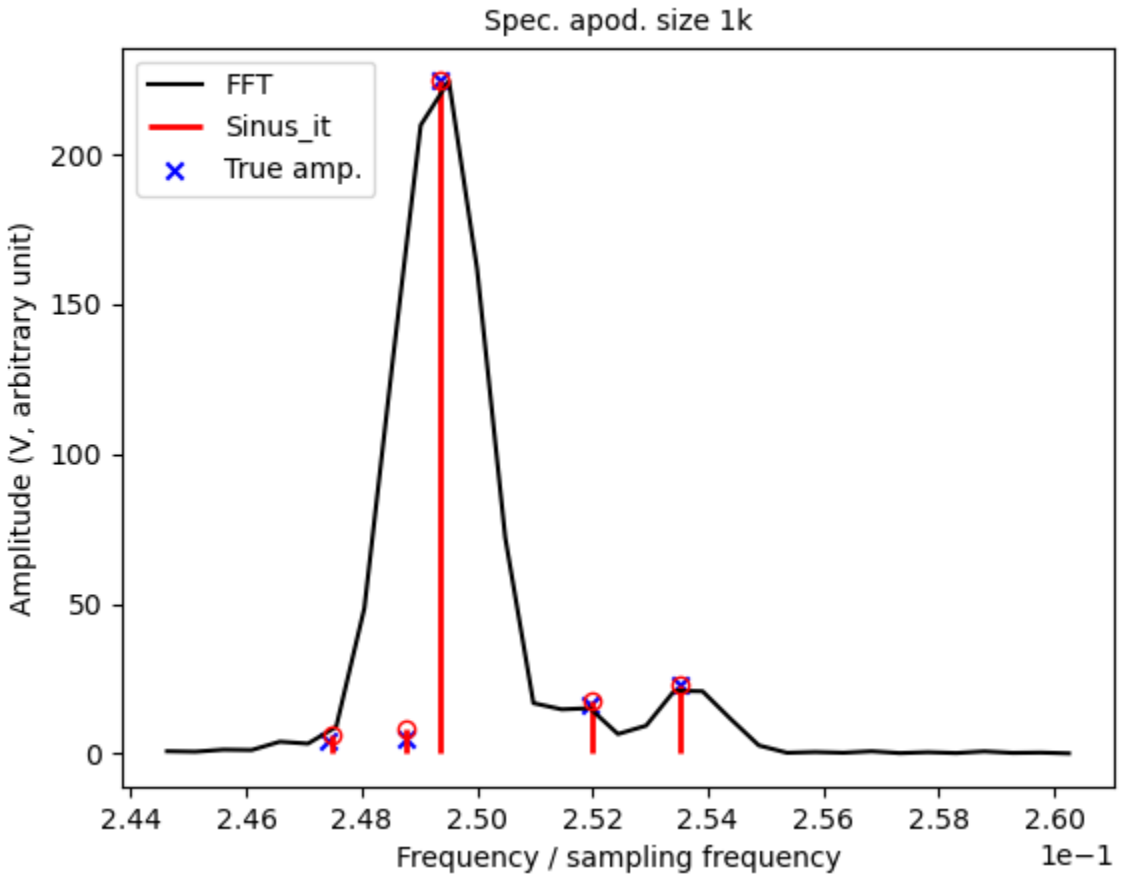
Phase vs frequency correlation



- Phase/Frequency correlation extraction with **32 points** in coarse analysis of experimental glutathione
- Sinus_it finds the **real value** of the phases
- Correlation coefficient = **0.99**
- Need to add **2π×t** to make correlation appear because of the modulos
- In FT-ICR MS, the phase frequency relationship is quadratic on the whole frequency range, but quasi-linear on a small range

Sinus_it: Glutathione 1st isotope 1k from 4M transient, ZoomFFT 1024 (eq. 1M full)

Fixed number of sines & phase/frequency correlation



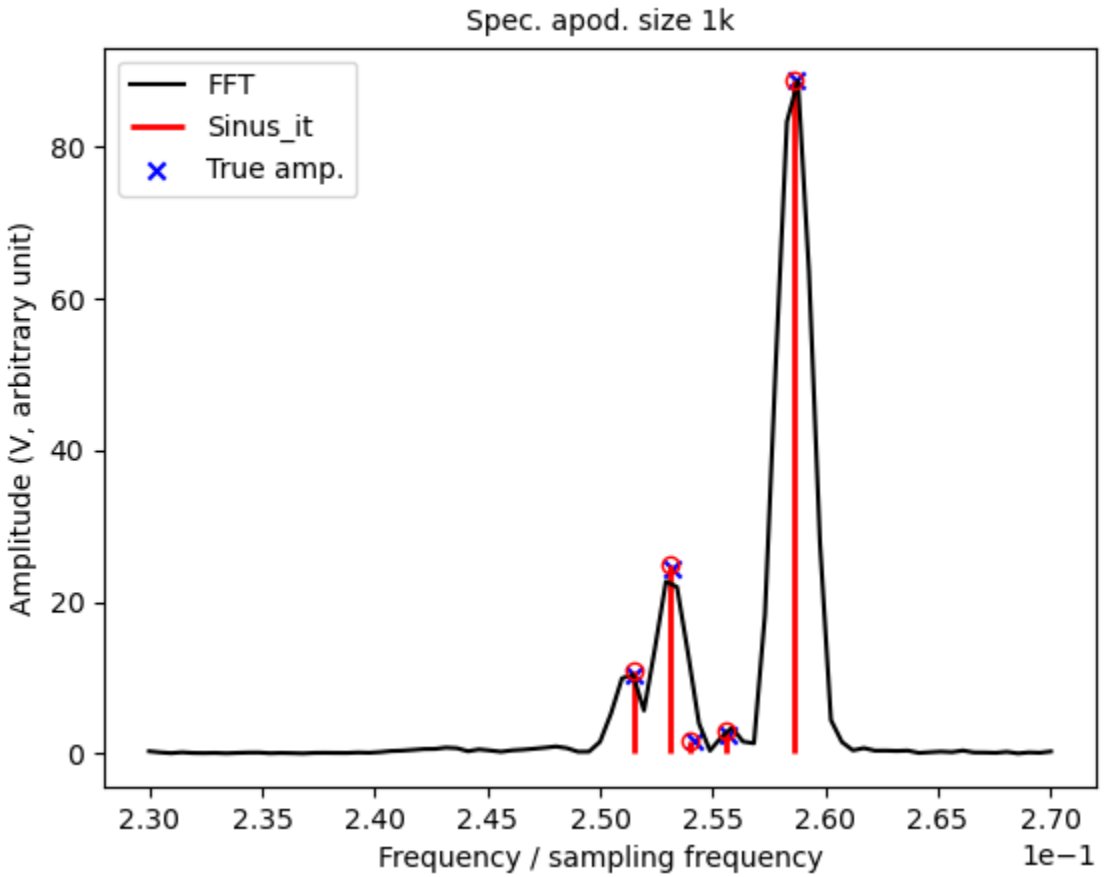
	True	Sinus_it	Error
Amplitude 1	223.271530	223.271530	0
Frequency 1	0.249345020	0.249374181	1E-4
Amplitude 2	22.62654664	22.73986409	5E-3
Frequency 2	0.253505482	0.253532171	1E-4
		17.36123890	6E-2
Frequency 3	0.249345020	0.249374181	3E-4
Amplitude 4	4.72880916	8.217349565	7E-1
Frequency 4	0.248777578	0.248682469	3E-4
Amplitude 5	4.27496060	5.936296485	3E-1
Frequency 5	0.247421523	0.247358560	2E-4

Mettre l'erreur après de-zooming

Execution time : **12 minutes** on an Intel Xeon 2.1 Ghz server with 128 Go of memory RAM, embedded with an Nvidia RTX 2080 Ti graphics card

Sinus_it: Glutathione 2nd isotope 1k from 4M transient, ZoomFFT 1024 (eq. 1M full)

Fixed number of sines & phase/frequency correlation



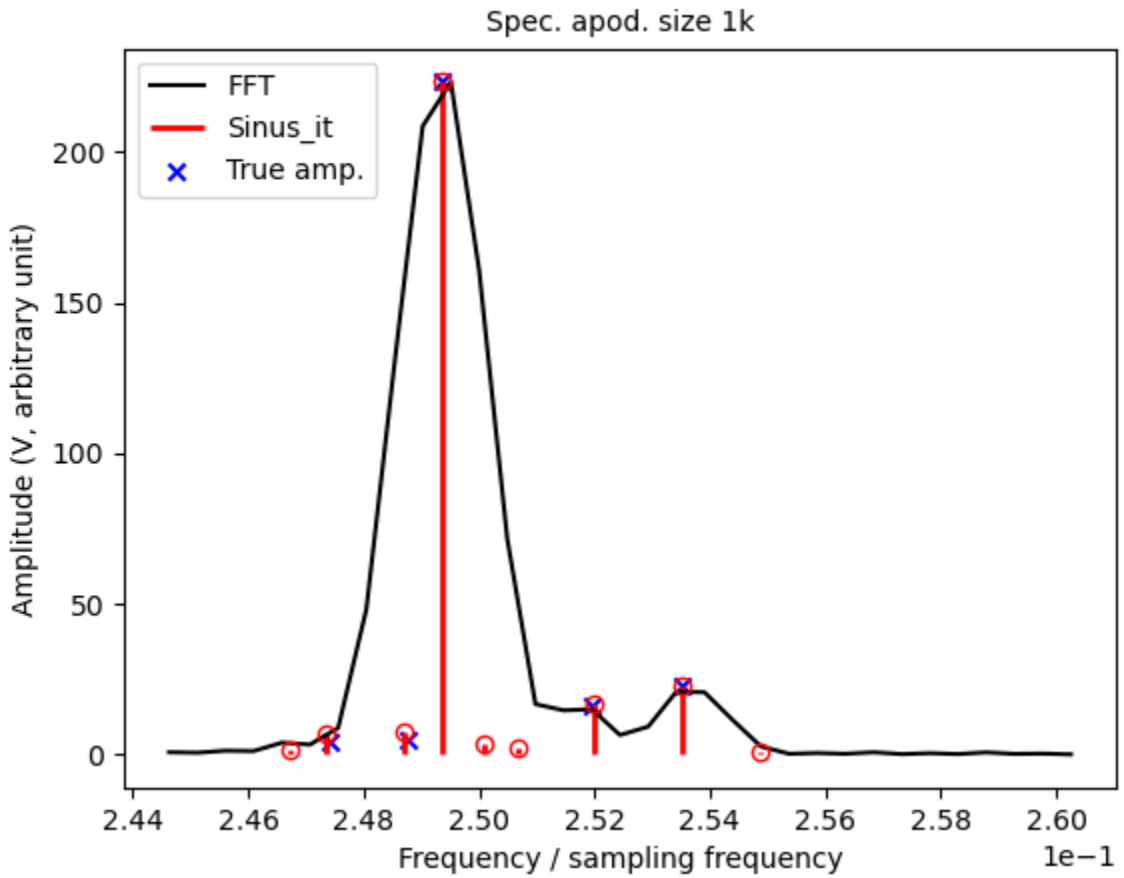
	True	Sinus_it	Error
Amplitude 1	88.65662384	88.65662384	0
Frequency 1	0.258683520	0.258649869	1.3E-4
Amplitude 2	24.43204256	24.94586181	2.1E-2
Frequency 2	0.253158016	0.253079324	3.1E-4
		10.91672229	4.7E-2
Frequency 3	0.251545216	0.251528084	6.8E-5
Amplitude 4	2.360991171	2.975147008	2.6E-1
Frequency 4	0.255679104	0.255635380	1.7E-4
Amplitude 5	1.686422265	1.722096204	2.1E-2
Frequency 5	0.254141056	0.254017502	4.8E-4

Mettre l'erreur après de-zooming

Execution time : 12 minutes on an Intel Xeon 2.1 Ghz server with 128 Go of memory RAM, embedded with an Nvidia RTX 2080 Ti graphics card

Results (Experimental glutathione 1st isotope from 4k zoom of a 4M transient)

Dynamic number of sines & phase/frequency correlation



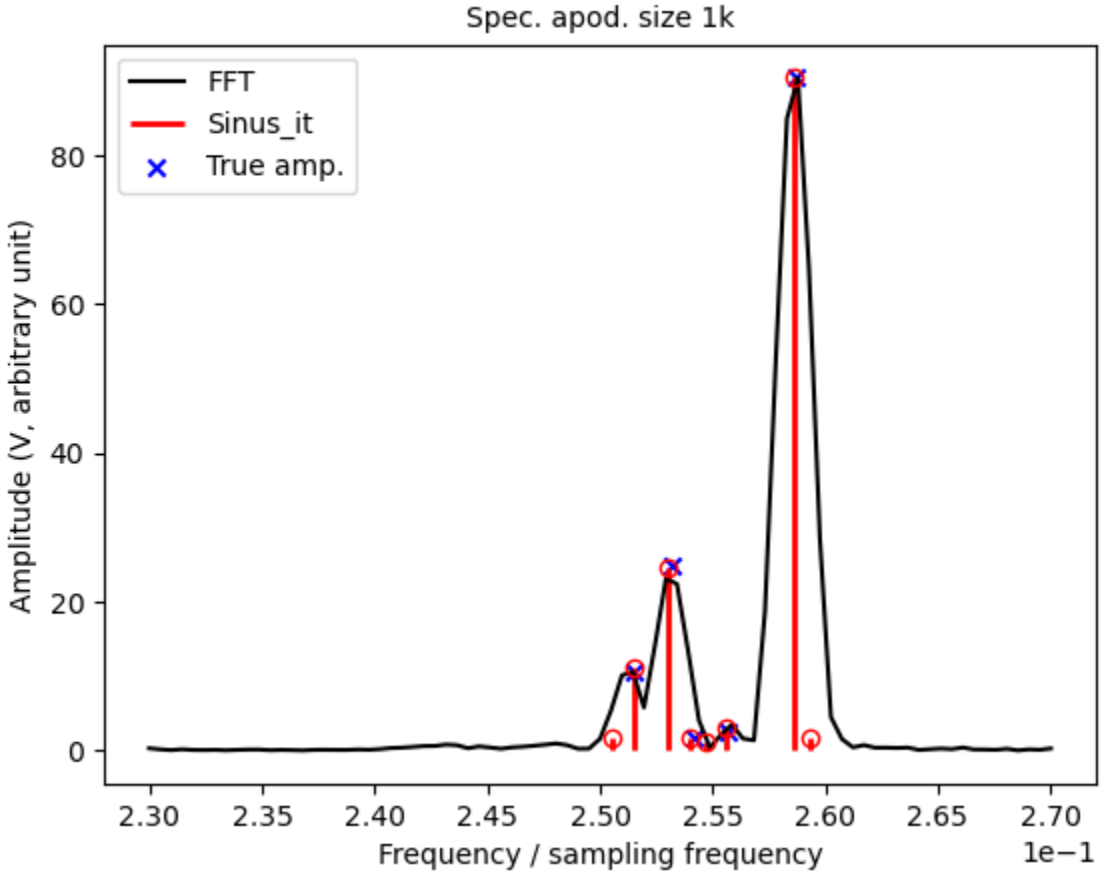
	True	Sinus_it	Error
Amplitude 1	223.2715301	223.2715301	0
Frequency 1	0.249345020	0.249374181	1.1E-4
Amplitude 2	22.62654665	23.07601356	1.9E-2
Frequency 2	0.253505482	0.253532171	1E-4
		16.56901359	1.5E-2
Frequency 3	0.251956480	0.251982152	1E-4
Amplitude 4	4.728809168	7.670573234	6.2E-1
Frequency 4	0.248777578	0.248682469	3.8E-4
Amplitude 5	4.27496061	7.004225254	6.3E-1
Frequency 5	0.247421523	0.247358560	2.5E-4

Mettre l'erreur après de-zooming

Execution time : **12 minutes** on an Intel Xeon 2.1 Ghz server with 128 Go of memory RAM, embedded with an Nvidia RTX 2080 Ti graphics card

Sinus_it: Glutathione 2nd isotope 1k from 4M transient, ZoomFFT 1024 (eq. 1M full)

Dynamic number of sines & phase/frequency correlation



	True	Sinus_it	Error
Amplitude 1	90.42452239	90.42452239	0
Frequency 1	0.258683520	0.258627384	2.1E-4
Amplitude 2	24.91924105	24.61579895	1.2E-2
Frequency 2	0.253158016	0.253058731	3.9E-4
		11.15986537	4.9E-2
Frequency 3	0.251545216	0.251532524	5E-5
Amplitude 4	2.408071612	3.169341087	3.1E-1
Frequency 4	0.255679104	0.255593299	3.3E-4
Amplitude 5	1.720051151	1.658804059	3.5E-2
Frequency 5	0.254141056	0.254038721	4E-4

Mettre l'erreur après de-zooming

Execution time : 20 minutes on an Intel Xeon 2.1 Ghz server with 128 Go of memory RAM, embedded with an Nvidia RTX 2080 Ti graphics card

Conclusion

- Sinus_it achieves a greater resolution than FFT (4× peak sharpening)
- Sinus_it find the real phases and amplitudes, close to the real mass isotopic ratios
- Sinus_it performs better than FDM on small or noisy transients
- The zoomFFT algorithm can reduce the number of points used and so speed-up ZoomFFT
- Further parallelization will permit to apply Sinus_it to a complex spectrum

Thank you for your attention !

MSAP team

Dr Christian ROLANDO

Pr Ahmed MAZZAH

Dr Fabrice BRAY

Stéphanie FLAMANT

Icube team (Strasbourg)

Pr Pierre COLLET

Flous!

French-Azerbaijani University (Baku)

Dr Ulviya ABDULKARIMOVA

EU FT-ICR MS

MSAP

cnrs

Université de Lille