Galaxy workflow for dnpatterntools

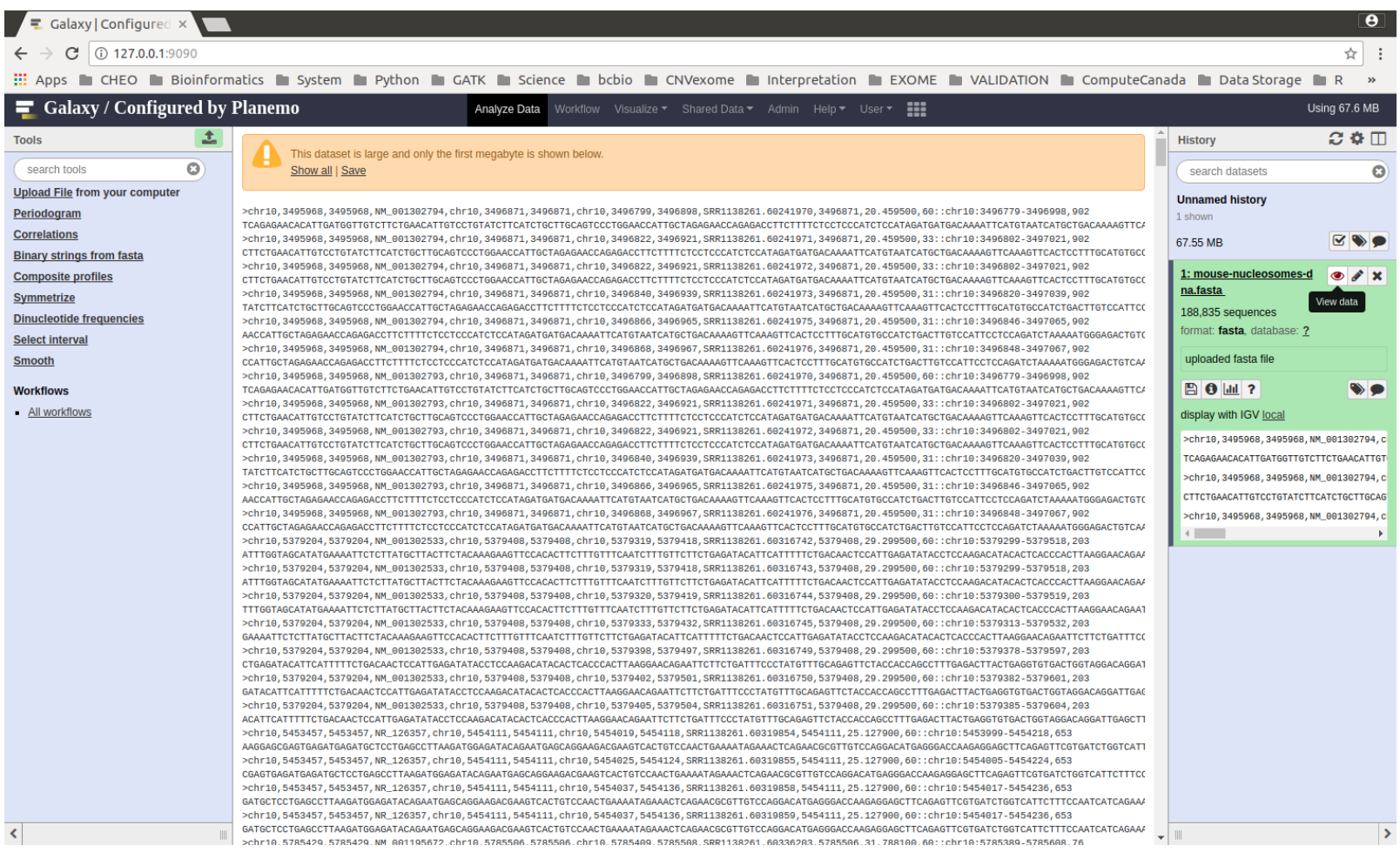
Using dnpatterntools in Galaxy.

Demo configured and served by Planemo.

Step 1. Upload fasta file

Using **Get data** upload fasta file . The mouse-nucleosomes-dna.fasta from the tools test-data directory contains 188,835 dna sequences of +1 nucleosomes obtained from MNase-Seq H3 data of control mouse brain (GEO GSE54263) aligned to *mm9* .

Figure 1. Upload fasta.



Step 2. Compute dinucleotide frequency

Using **Dinucleotide frequencies** tool compute positional frequency of dinucleotide occurrences on forward and complementary sequences for all 16 dinucleotides. Figure 2 shows tool's interface, Figure 3 shows plots of resulting frequency profiles along all fasta sequence positions. Column names identify dinucleotide. Extension (f) stands for forward and (r) for complementary sequence. A peak identifies a cleavage site.

Figure 2. Dinucleotide frequencies tool interface

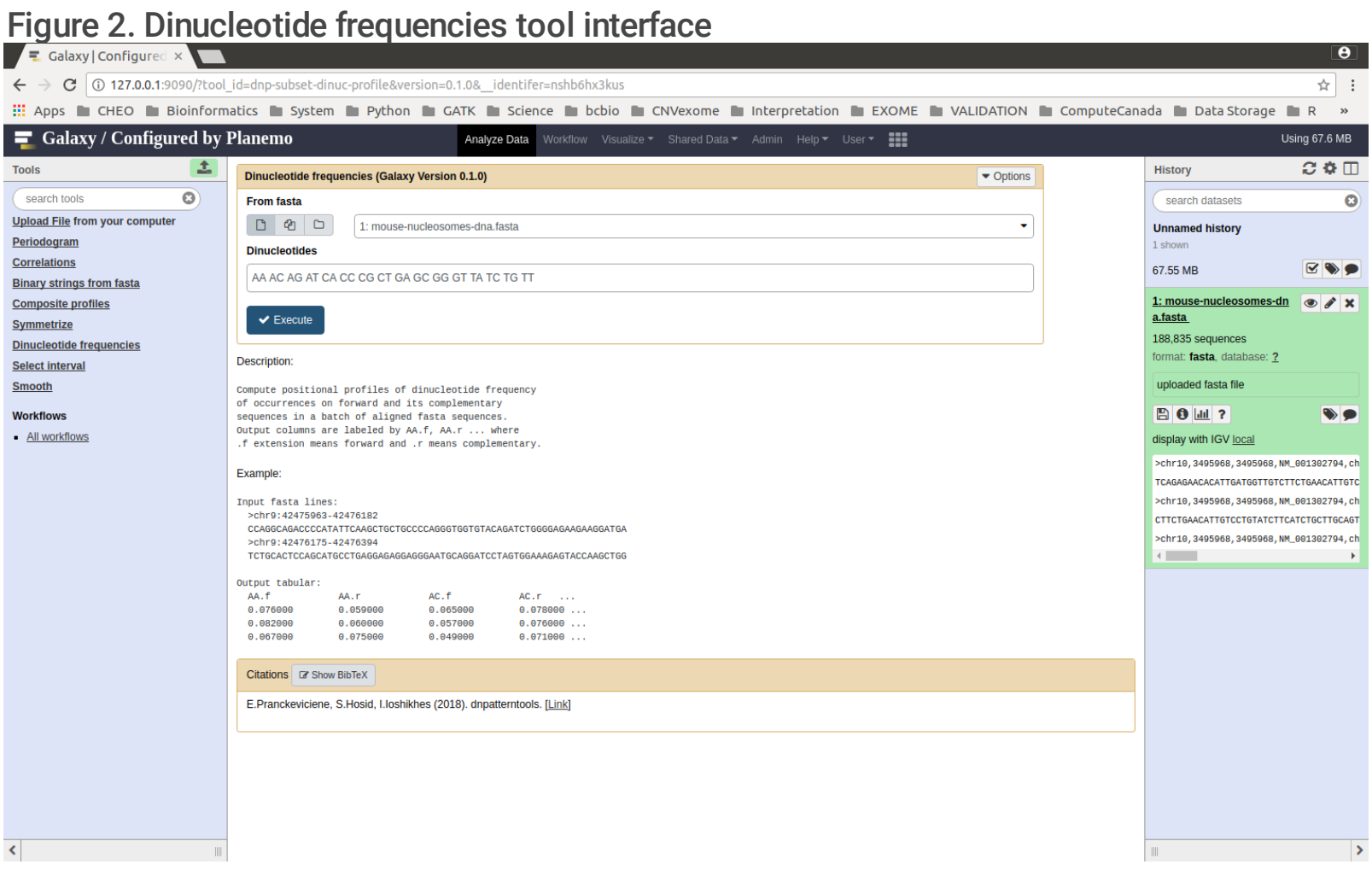
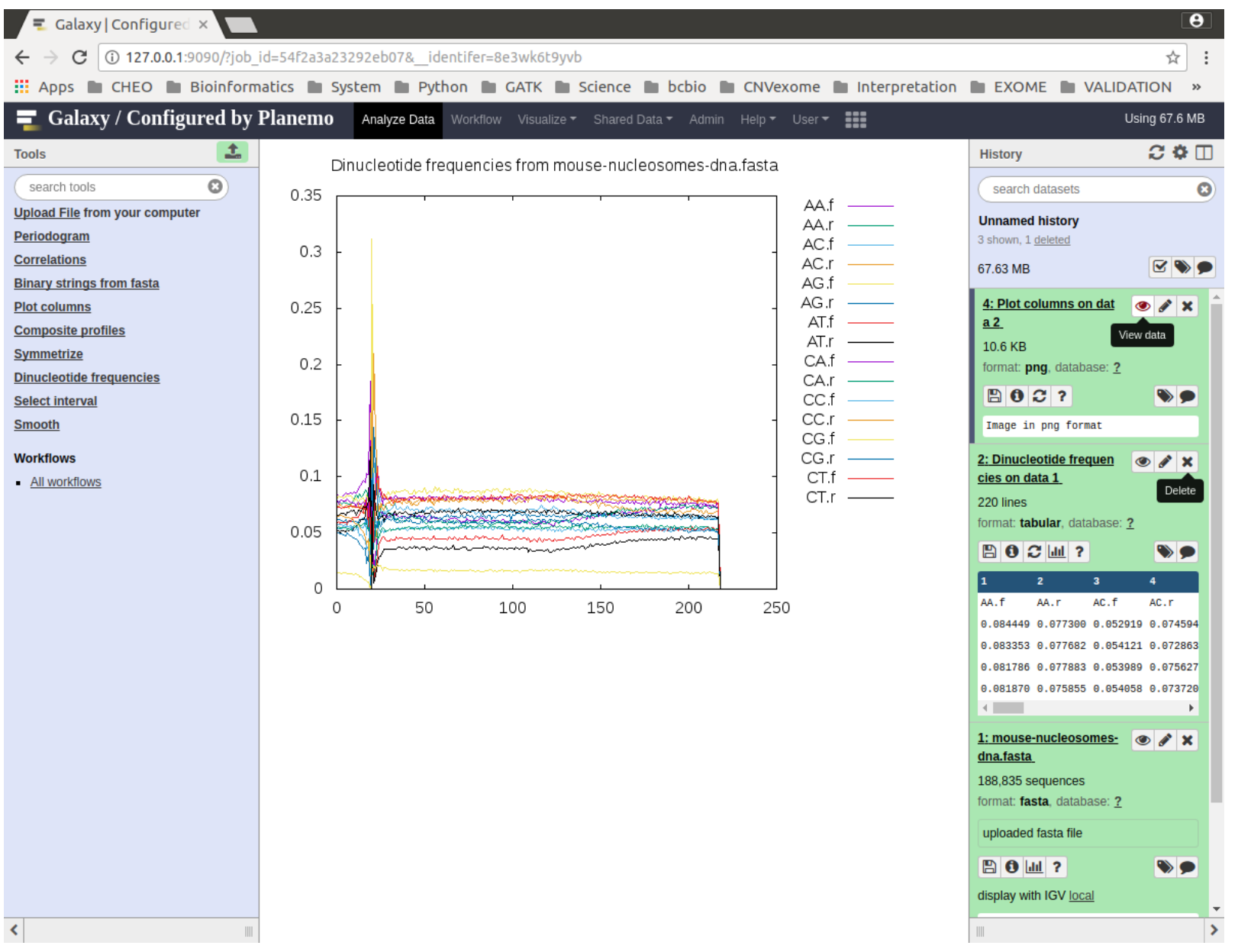


Figure 3. Plot of dinucleotide frequency profiles- output of Dinucleotide frequencies tool



Step 3. Compute correlations between the dinucleotide frequency profiles from forward and complementary sequences

Using **Correlations** tool compute correlations between dinucleotide frequency profiles on forward and complementary sequences. The following figures 4 and 5 show interface and output of the Correlations tool. Input for the Correlations tool is frequency profiles computed in Step 2. Correlations are computed for all dinucleotides. In the output a first column named "0" contains average correlation of all dinucleotides. The **Correlations** tool on data 2 also prints dinucleotide, its position of maximum positive correlation and a value of maximum positive correlation. The position of the maximum positive correlation most proximal to the cleavage site identifies a start position of a nucleosome in a batch of sequences aligned by an experimental end. Determination of the start position of a nucleosome step can't be fully automated. In this example a start position of a nucleosome is 25 based on a very prominent peak of maximum correlation in AA and TT as shown in a plot of correlation results.

Figure 4. Correlations tool interface

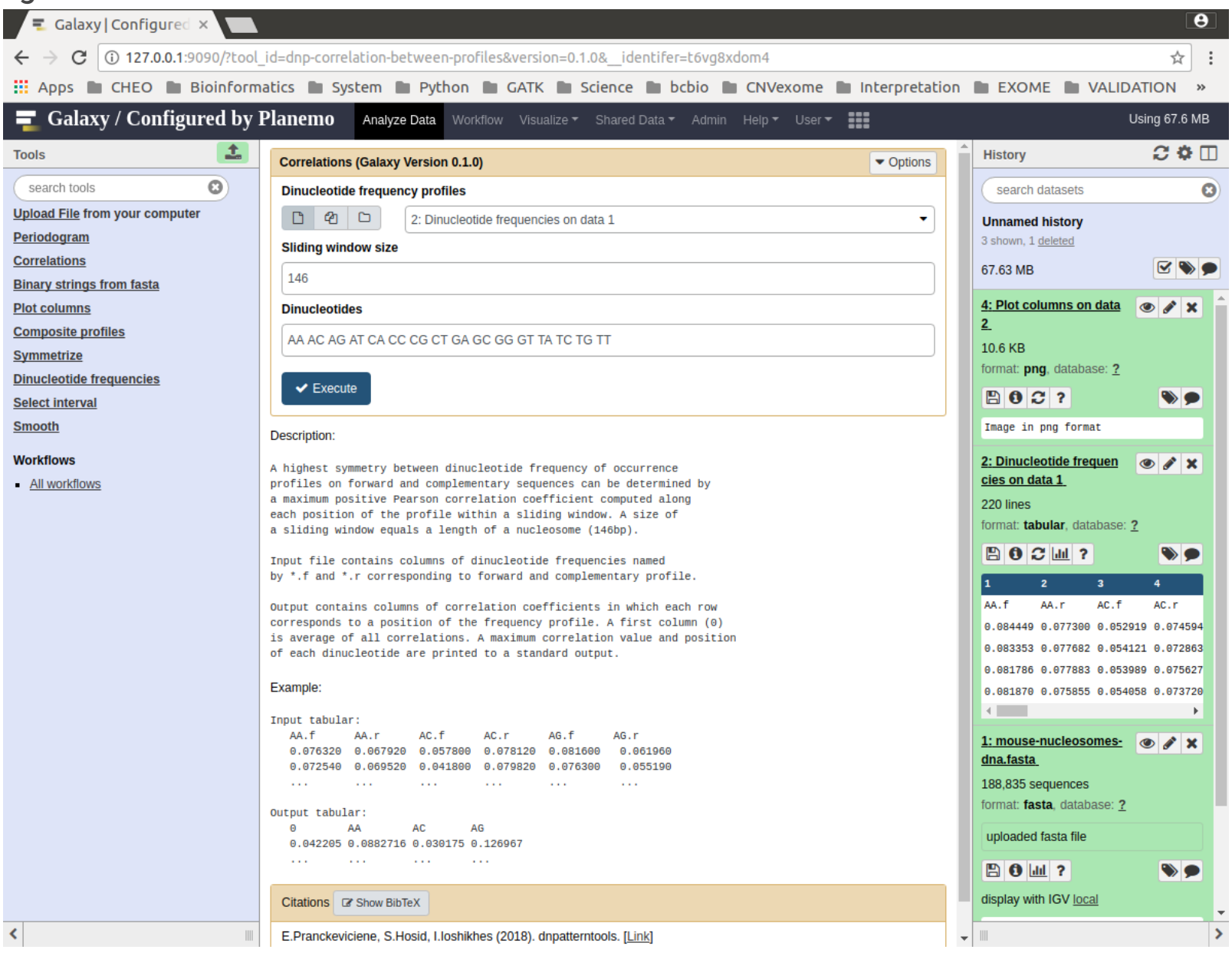


Figure 5. Correlation tool output

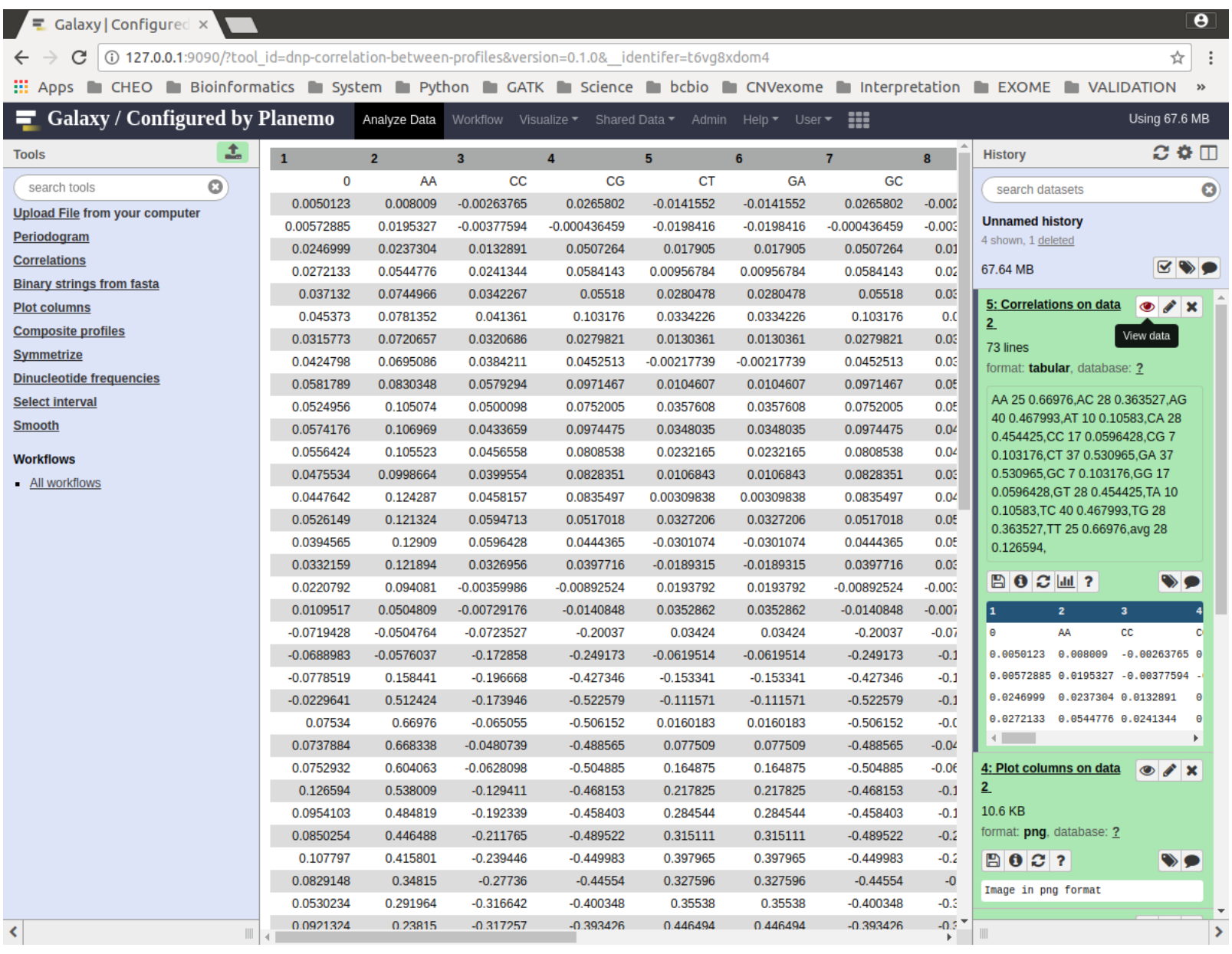
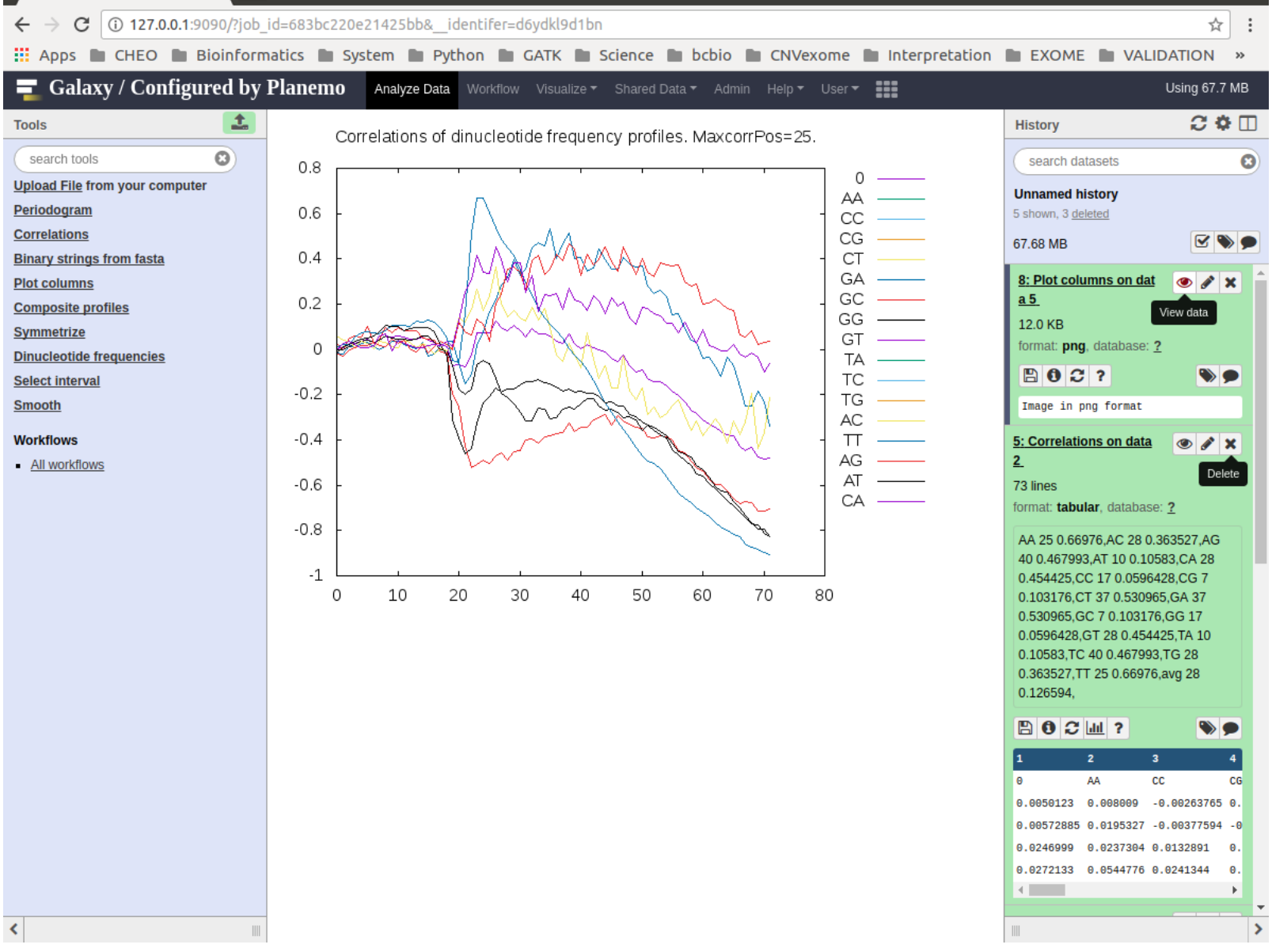


Figure 6. Plot of the Correlations tool output



Step 4. Select interval

Using **Select interval** tool from the dinucleotide frequencies computed in Step 2 select interval from the profile which belongs to a nucleosome. Start position is a position identified through maximum positive correlation in Step 3. In this example It is 25. The size of selection is a length of a nucleosome - 146 base pairs. The interval is selected for all dinucleotides. A column with a position information is added to the selection, shown in Figure 8.

Figure 7. Select interval tool interface

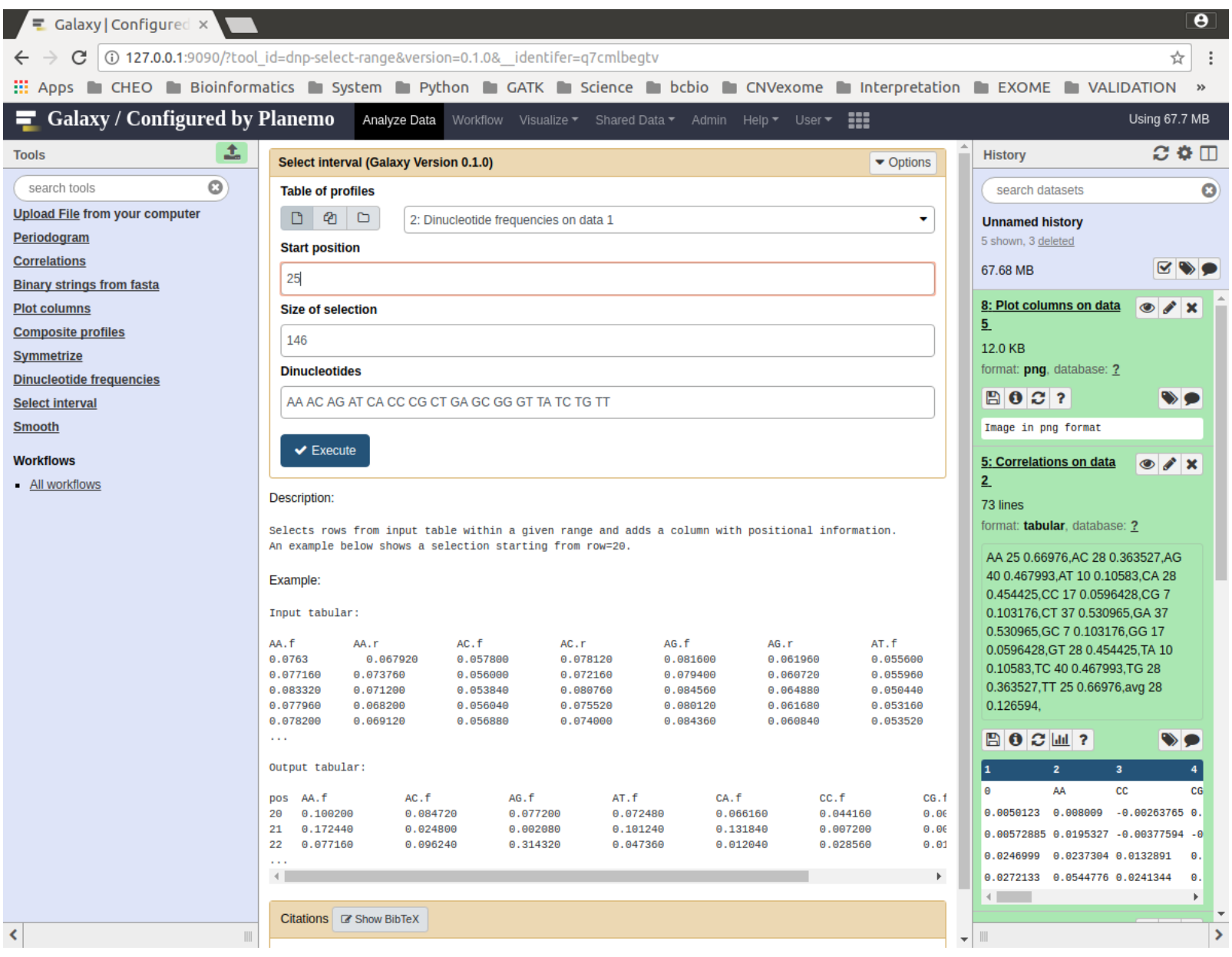
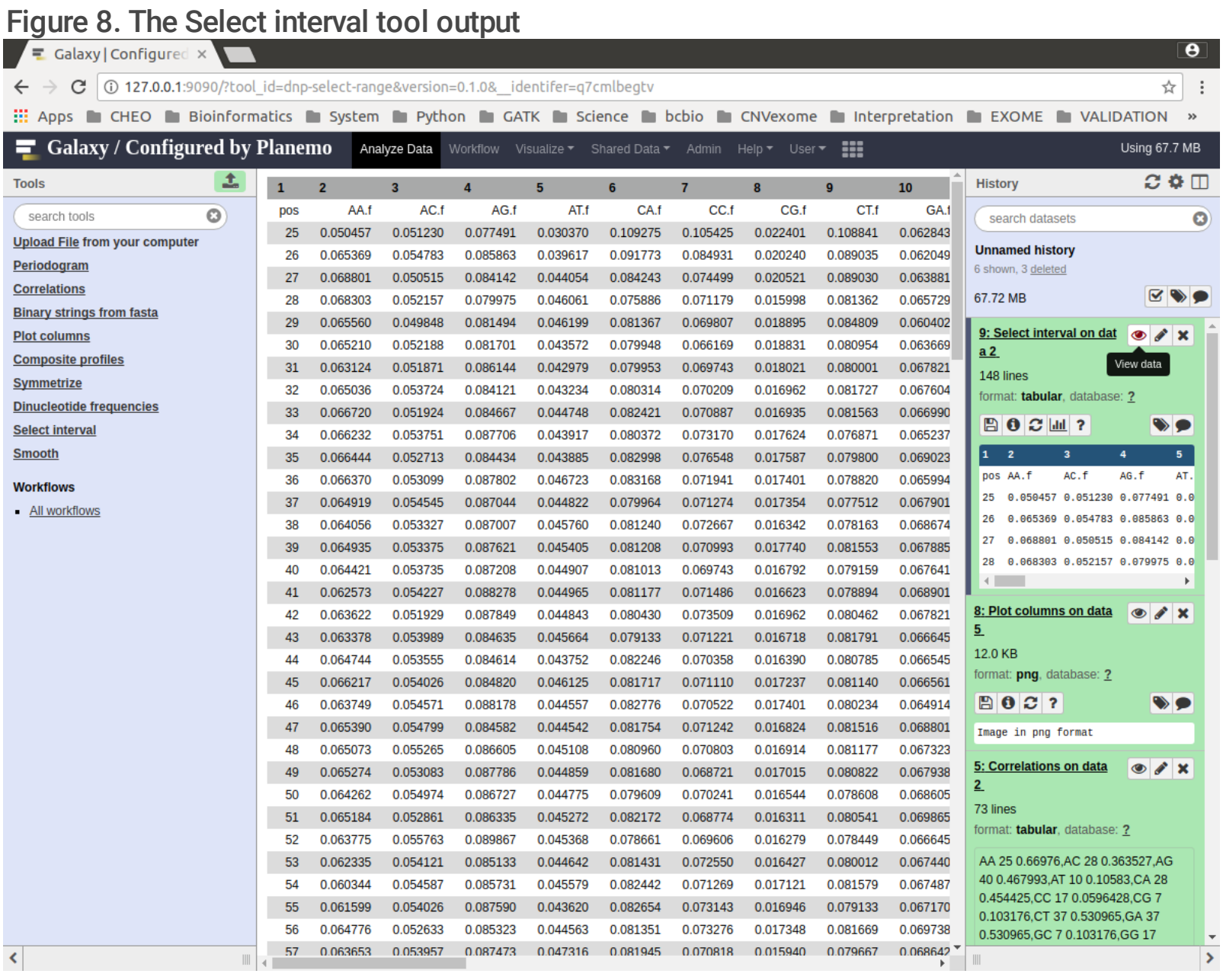


Figure 8. The Select interval tool output



1	2	3	4	5	6	7	8	9	10
pos	AA.f	AC.f	AG.f	AT.f	CA.f	CC.f	CG.f	CT.f	GA.f
25	0.050457	0.051230	0.077491	0.030370	0.109275	0.105425	0.022401	0.108841	0.062843
26	0.065369	0.054783	0.085863	0.039617	0.091773	0.084931	0.020240	0.089035	0.062049
27	0.068801	0.050515	0.084142	0.044054	0.084243	0.074499	0.020521	0.089030	0.063881
28	0.068303	0.052157	0.079975	0.046061	0.075886	0.071179	0.015998	0.081362	0.065729
29	0.065560	0.049848	0.081494	0.046199	0.081367	0.069807	0.018895	0.084809	0.060402
30	0.065210	0.052188	0.081701	0.043572	0.079948	0.066169	0.018831	0.080954	0.063669
31	0.063124	0.051871	0.086144	0.042979	0.079953	0.069743	0.018021	0.080001	0.067821
32	0.065036	0.053724	0.084121	0.043234	0.080314	0.070209	0.016962	0.081727	0.067604
33	0.066720	0.051924	0.084667	0.044748	0.082421	0.070887	0.016935	0.081563	0.066990
34	0.066232	0.053751	0.087706	0.043917	0.080372	0.073170	0.017624	0.076871	0.065237
35	0.066444	0.052713	0.084434	0.043885	0.082998	0.076548	0.017587	0.079800	0.069023
36	0.066370	0.053099	0.087802	0.046723	0.083168	0.071941	0.017401	0.078820	0.065994
37	0.064919	0.054545	0.087044	0.044822	0.079964	0.071274	0.017354	0.077512	0.067901
38	0.064056	0.053327	0.087007	0.045760	0.081240	0.072667	0.016342	0.078163	0.068674
39	0.064935	0.053375	0.087621	0.045405	0.081208	0.070993	0.017740	0.081553	0.067885
40	0.064421	0.053735	0.087208	0.044907	0.081013	0.069743	0.016792	0.079159	0.067641
41	0.062573	0.054227	0.088278	0.044965	0.081177	0.071486	0.016623	0.078894	0.068901
42	0.063622	0.051929	0.087849	0.044843	0.080430	0.073509	0.016962	0.080462	0.067821
43	0.063378	0.053989	0.084635	0.045664	0.079133	0.071221	0.016718	0.081791	0.066645
44	0.064744	0.053555	0.084614	0.043752	0.082246	0.070358	0.016390	0.080785	0.066545
45	0.066217	0.054026	0.084820	0.046125	0.081717	0.071110	0.017237	0.081140	0.066561
46	0.063749	0.054571	0.088178	0.044557	0.082776	0.070522	0.017401	0.080234	0.064914
47	0.065390	0.054799	0.084582	0.044542	0.081754	0.071242	0.016824	0.081516	0.068801
48	0.065073	0.055265	0.086605	0.045108	0.080960	0.070803	0.016914	0.081177	0.067323
49	0.065274	0.053083	0.087786	0.044859	0.081680	0.068721	0.017015	0.080822	0.067938
50	0.064262	0.054974	0.086727	0.044775	0.079609	0.070241	0.016544	0.078608	0.068605
51	0.065184	0.052861	0.086335	0.045272	0.082172	0.068774	0.016311	0.080541	0.069865
52	0.063775	0.055763	0.089867	0.045368	0.078661	0.069606	0.016279	0.078449	0.066645
53	0.062335	0.054121	0.085133	0.044642	0.081431	0.072550	0.016427	0.080012	0.067440
54	0.060344	0.054587	0.085731	0.045579	0.082442	0.071269	0.017121	0.081579	0.067487
55	0.061599	0.054026	0.087590	0.043620	0.082654	0.073143	0.016946	0.079133	0.067170
56	0.064776	0.052633	0.085323	0.044563	0.081351	0.073276	0.017348	0.081669	0.069738
57	0.063653	0.053957	0.087473	0.047316	0.081945	0.070818	0.015940	0.079667	0.068642

Step 5. Symmetrize

Using **Symmetrize** tool superimpose dinucleotide frequency profiles from forward and complement sequences with respect to a dyad position that is a central position of the frequency profile. Figure 9 and 10 shows the tool interface and an output. The first column of the output contains a position upstream (-) and downstream (+) relative to a dyad position.

Figure 9. Symmetrize tool interface

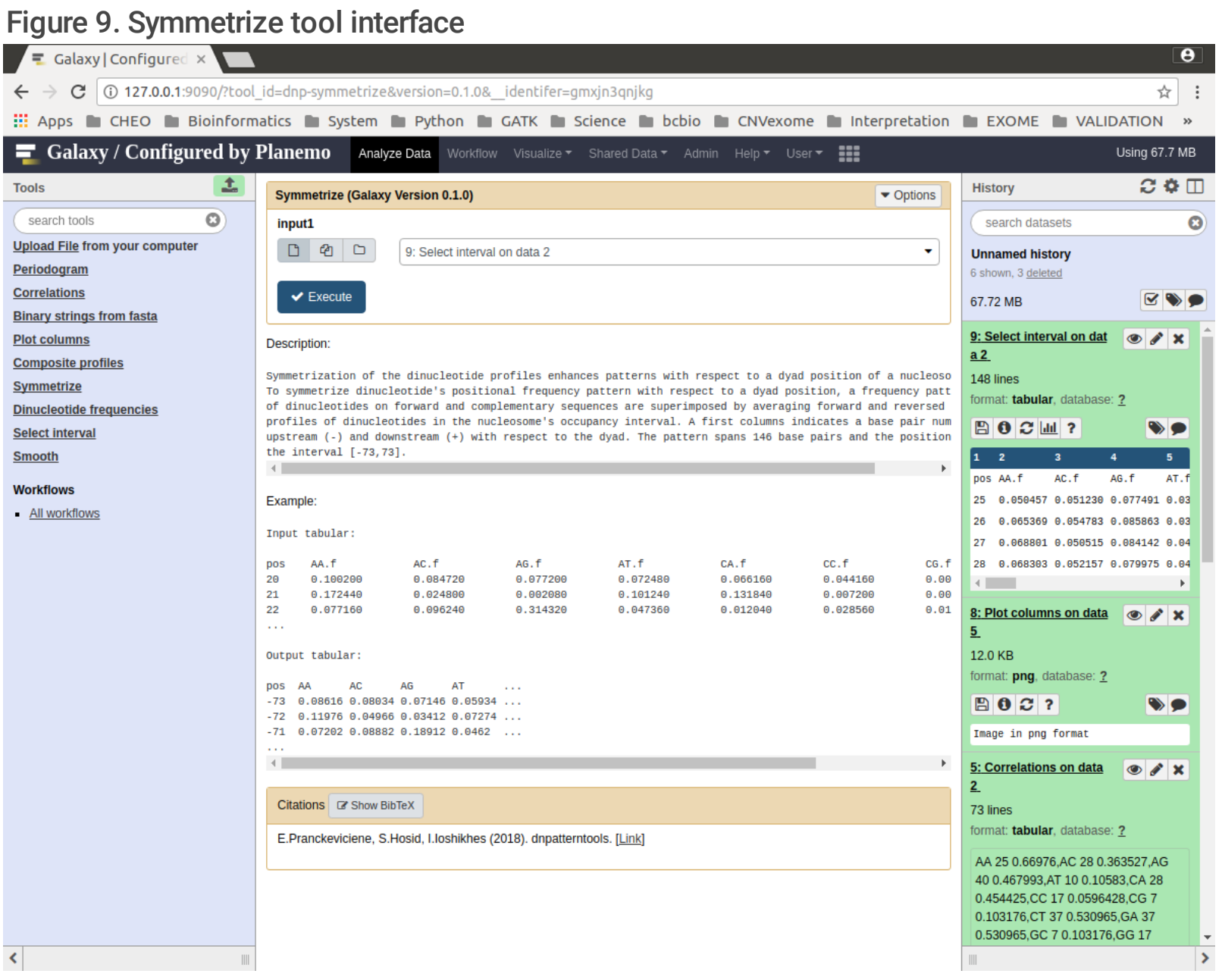
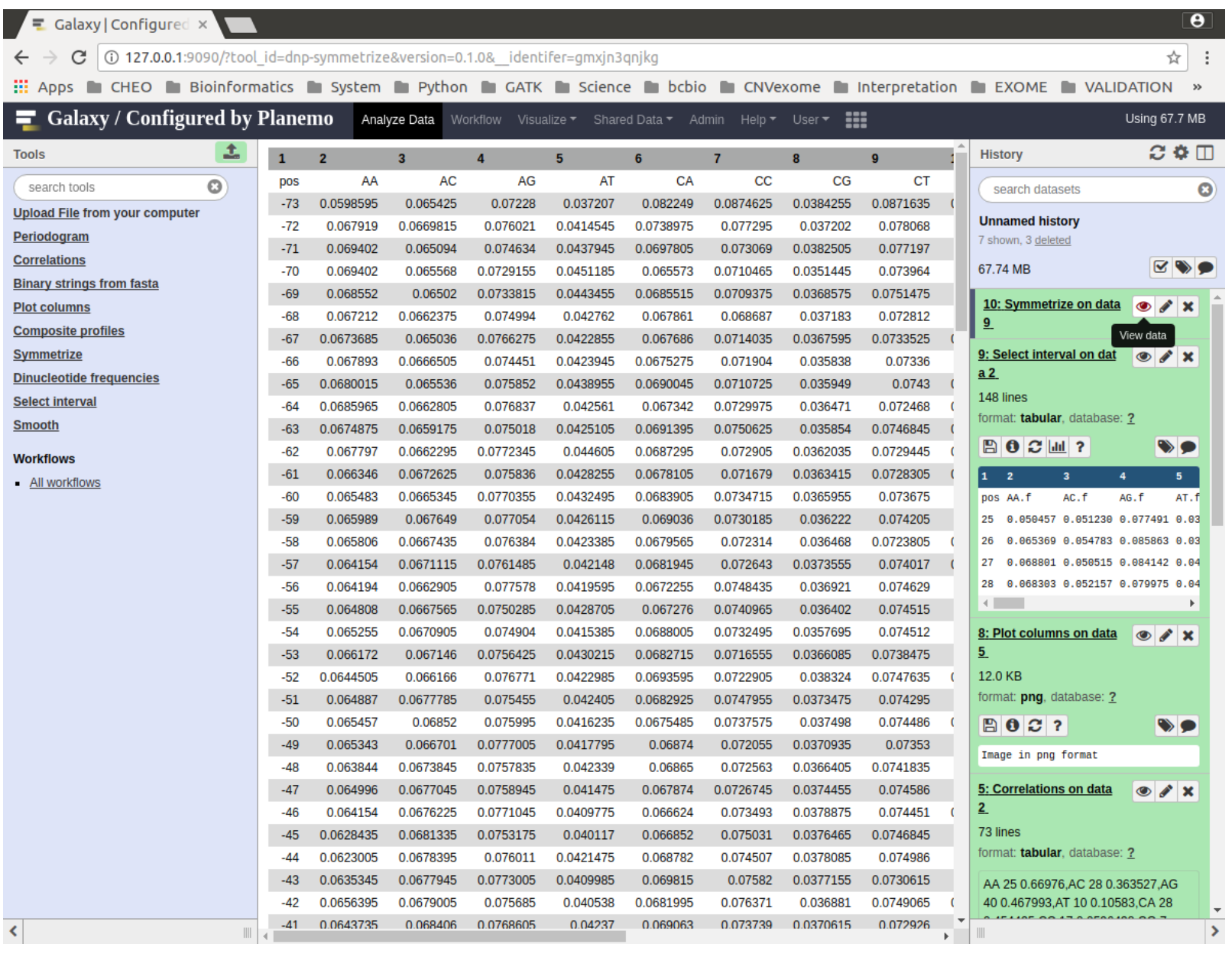


Figure 10. Output of Symmetrize tool

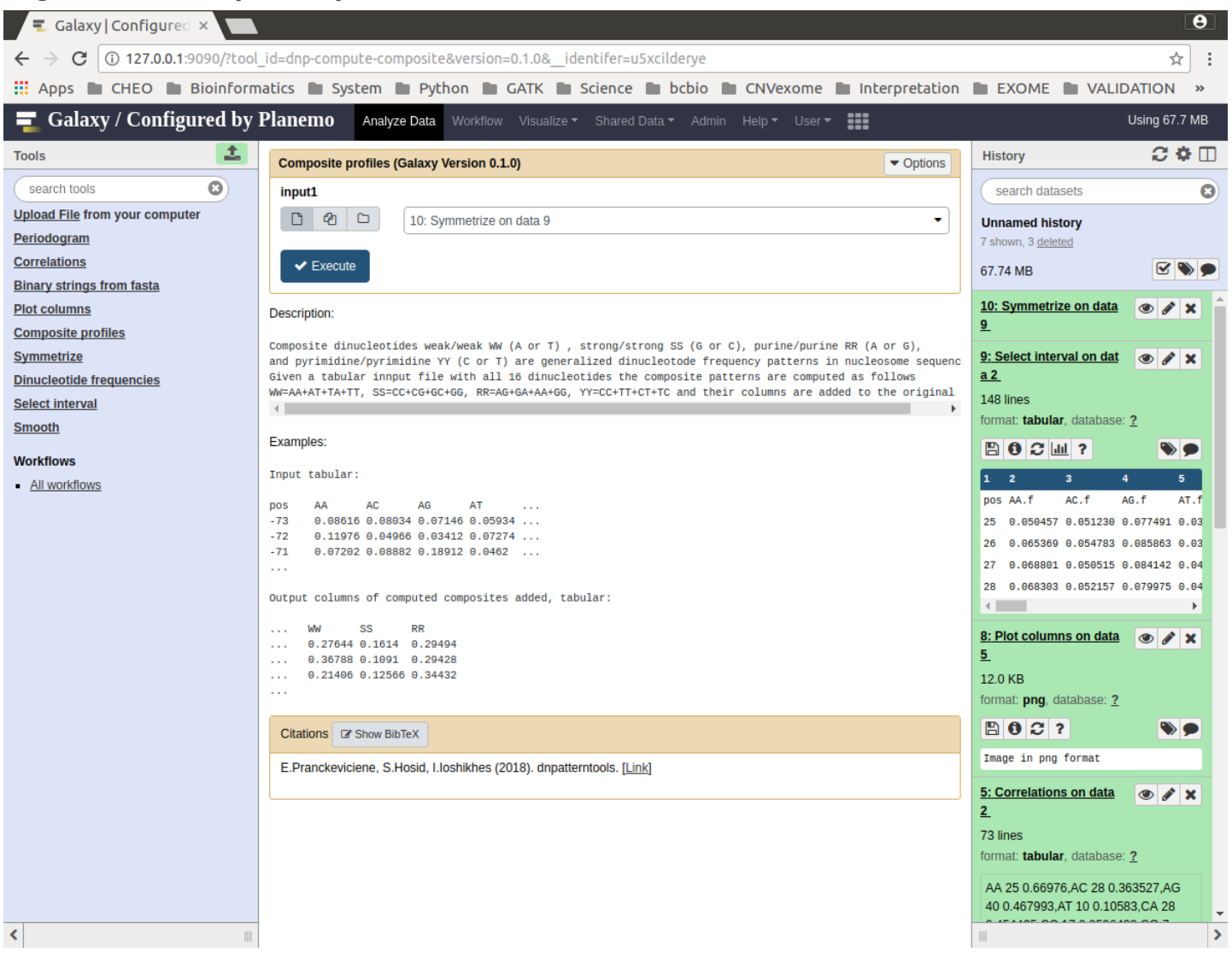


1	2	3	4	5	6	7	8	9
pos	AA	AC	AG	AT	CA	CC	CG	CT
-73	0.0598595	0.065425	0.07228	0.037207	0.082249	0.0874625	0.0384255	0.0871635
-72	0.067919	0.0669815	0.076021	0.0414545	0.0738975	0.077295	0.037202	0.078068
-71	0.069402	0.065094	0.074634	0.0437945	0.0697805	0.073069	0.0382505	0.077197
-70	0.069402	0.065568	0.0729155	0.0451185	0.065573	0.0710465	0.0351445	0.073964
-69	0.068552	0.06502	0.0733815	0.0443455	0.0685515	0.0709375	0.0368575	0.0751475
-68	0.067212	0.0662375	0.074994	0.042762	0.067861	0.068687	0.037183	0.072812
-67	0.0673685	0.065036	0.0766275	0.0422855	0.067686	0.0714035	0.0367595	0.0733525
-66	0.067893	0.0666505	0.074451	0.0423945	0.0675275	0.071904	0.035838	0.07336
-65	0.0680015	0.065536	0.075852	0.0438955	0.0690045	0.0710725	0.035949	0.0743
-64	0.0685965	0.0662805	0.076837	0.042561	0.067342	0.0729975	0.036471	0.072468
-63	0.0674875	0.0659175	0.075018	0.0425105	0.0691395	0.0750625	0.035854	0.0746845
-62	0.067797	0.0662295	0.0772345	0.044605	0.0687295	0.072905	0.0362035	0.0729445
-61	0.066346	0.0672625	0.075836	0.0428255	0.0678105	0.071679	0.0363415	0.0728305
-60	0.065483	0.0665345	0.0770355	0.0432495	0.0683905	0.0734715	0.0365955	0.073675
-59	0.065989	0.067649	0.077054	0.0426115	0.069036	0.0730185	0.036222	0.074205
-58	0.065806	0.0667435	0.076384	0.0423385	0.0679565	0.072314	0.036468	0.0723805
-57	0.064154	0.0671115	0.0761485	0.042148	0.0681945	0.072643	0.0373555	0.074017
-56	0.064194	0.0662905	0.077578	0.0419595	0.0672255	0.0748435	0.036921	0.074629
-55	0.064808	0.0667565	0.0750285	0.0428705	0.067276	0.0740965	0.036402	0.074515
-54	0.065255	0.0670905	0.074904	0.0415385	0.0688005	0.0732495	0.0357695	0.074512
-53	0.066172	0.067146	0.0756425	0.0430215	0.0682715	0.0716555	0.0366085	0.0738475
-52	0.0644505	0.066166	0.076771	0.0422985	0.0693595	0.0722905	0.038324	0.0747635
-51	0.064887	0.0677785	0.075455	0.042405	0.0682925	0.0747955	0.0373475	0.074295
-50	0.065457	0.06852	0.075995	0.0416235	0.0675485	0.0737575	0.037498	0.074486
-49	0.065343	0.066701	0.0777005	0.0417795	0.06874	0.072055	0.0370935	0.07353
-48	0.063844	0.0673845	0.0757835	0.042339	0.06865	0.072563	0.0366405	0.0741835
-47	0.064996	0.0677045	0.0758945	0.041475	0.067874	0.0726745	0.0374455	0.074586
-46	0.064154	0.0676225	0.0771045	0.0409775	0.066624	0.073493	0.0378875	0.074451
-45	0.0628435	0.0681335	0.0753175	0.040117	0.066852	0.075031	0.0376465	0.0746845
-44	0.0623005	0.0678395	0.076011	0.0421475	0.068782	0.074507	0.0378085	0.074986
-43	0.0635345	0.0677945	0.0773005	0.0409985	0.069815	0.07582	0.0377155	0.0730615
-42	0.0656395	0.0679005	0.075685	0.040538	0.0681995	0.076371	0.036881	0.0749065
-41	0.0643735	0.068406	0.0768605	0.04237	0.069063	0.073739	0.0370615	0.072926

Step 6. Composite profiles

Using **Composite profiles** tool compute composite profiles of Weak/Weak WW (W = A or T) Strong/Strong SS (S = C or G) Purine/Purine RR (R = A or G) and Pyrimidine/Pyrimidine YY (Y=C or T) dinucleotides. The composite profiles represent most prominent general periodic features of nucleosome's DNA patterns. Figure 11 shows interface of **Composite profiles** tool.

Figure 11. Composite profiles tools interface



Step 7. Smooth

The **Smooth** tool is used to enhance dinucleotide frequency patterns by applying moving average smoothing, The smoothing removes noisy peaks. The default smoothing window size is 3 positions. To remove a noisy signal from the ends of the dinucleotide frequency patterns of nucleosomes the profiles are trimmed by a given number of positions (default 4). Figures 12 an 13 show **Smooth** tool interface and output.

Figure 12. Smooth tool interface

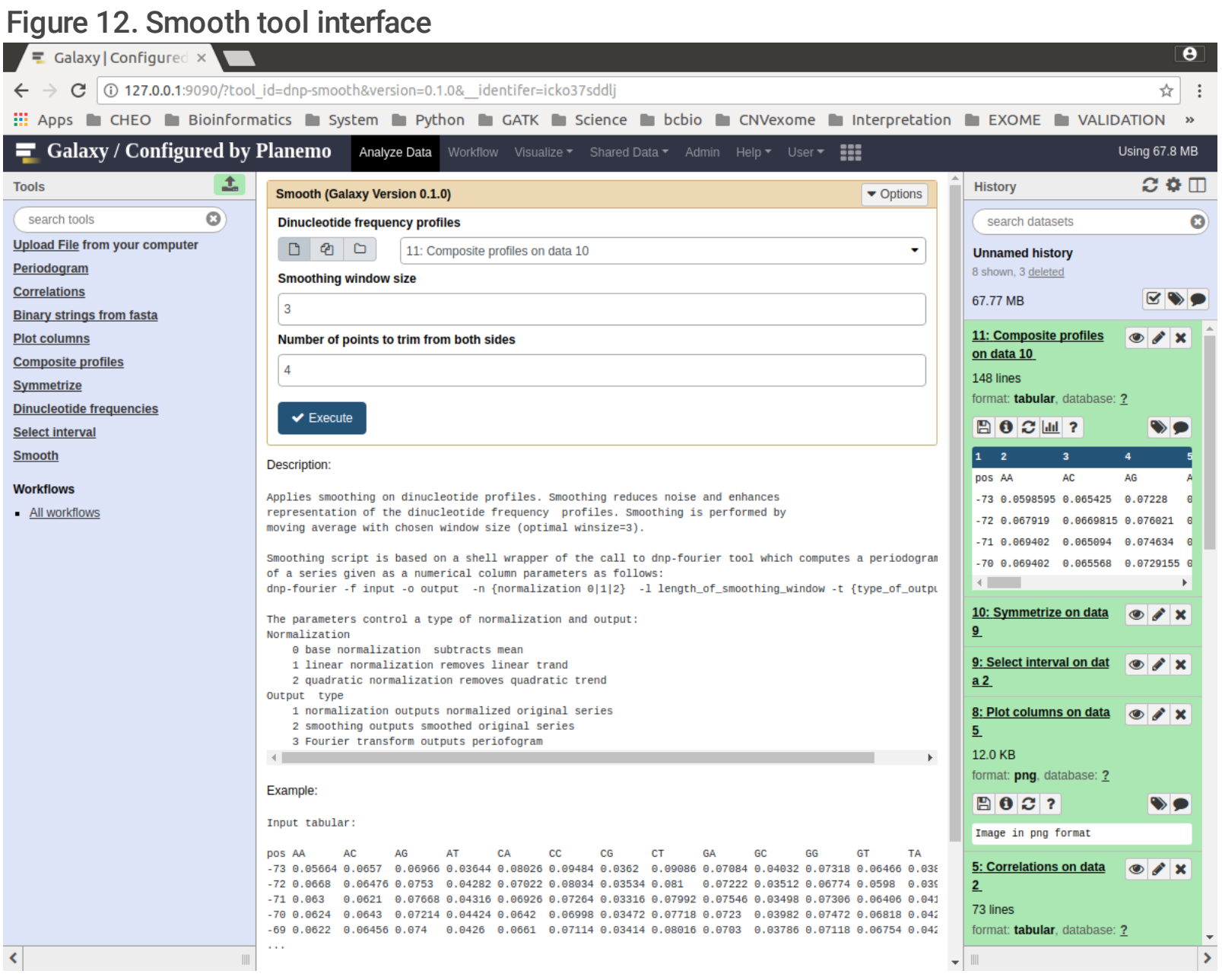
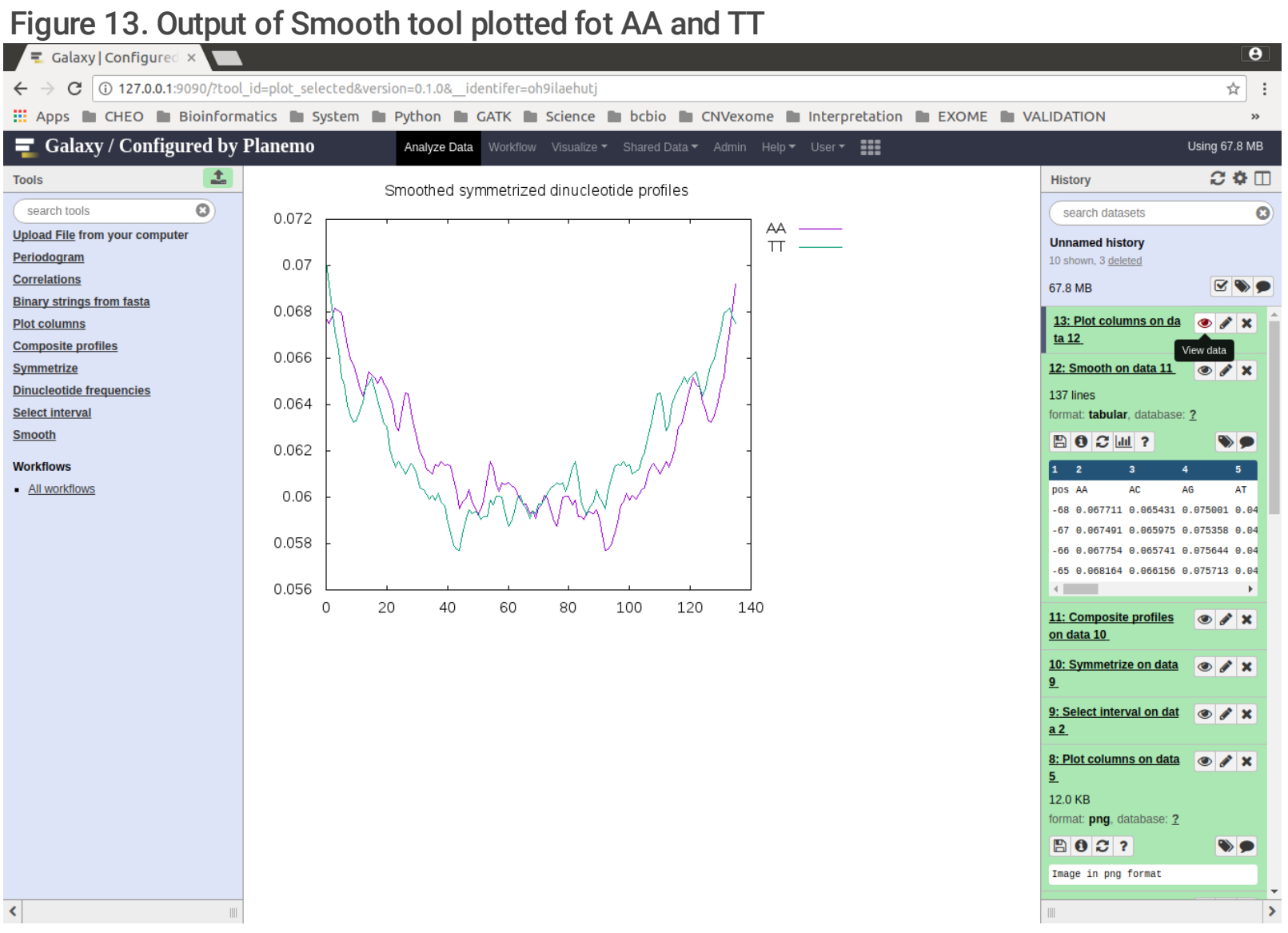


Figure 13. Output of Smooth tool plotted fot AA and TT



Step 8. Compute periodogram

The **Periodogram** tool computes Fourier Transform on dinucleotide frequency patterns. The tool performs normalization - trend removal- (quadratic by default) and smoothing. To remove low frequency component caused by a noisy signal at the ends of the dinucleotide frequency profile it is trimmed (default value is 4 positions from both ends). The operation is applied to all columns of the input data. In the output table of the **Periodogram** tool a first column indicates period in base pairs. Figures 14 and 15 show interface and output of the tool.

Figure 14. Periodogram tool interface

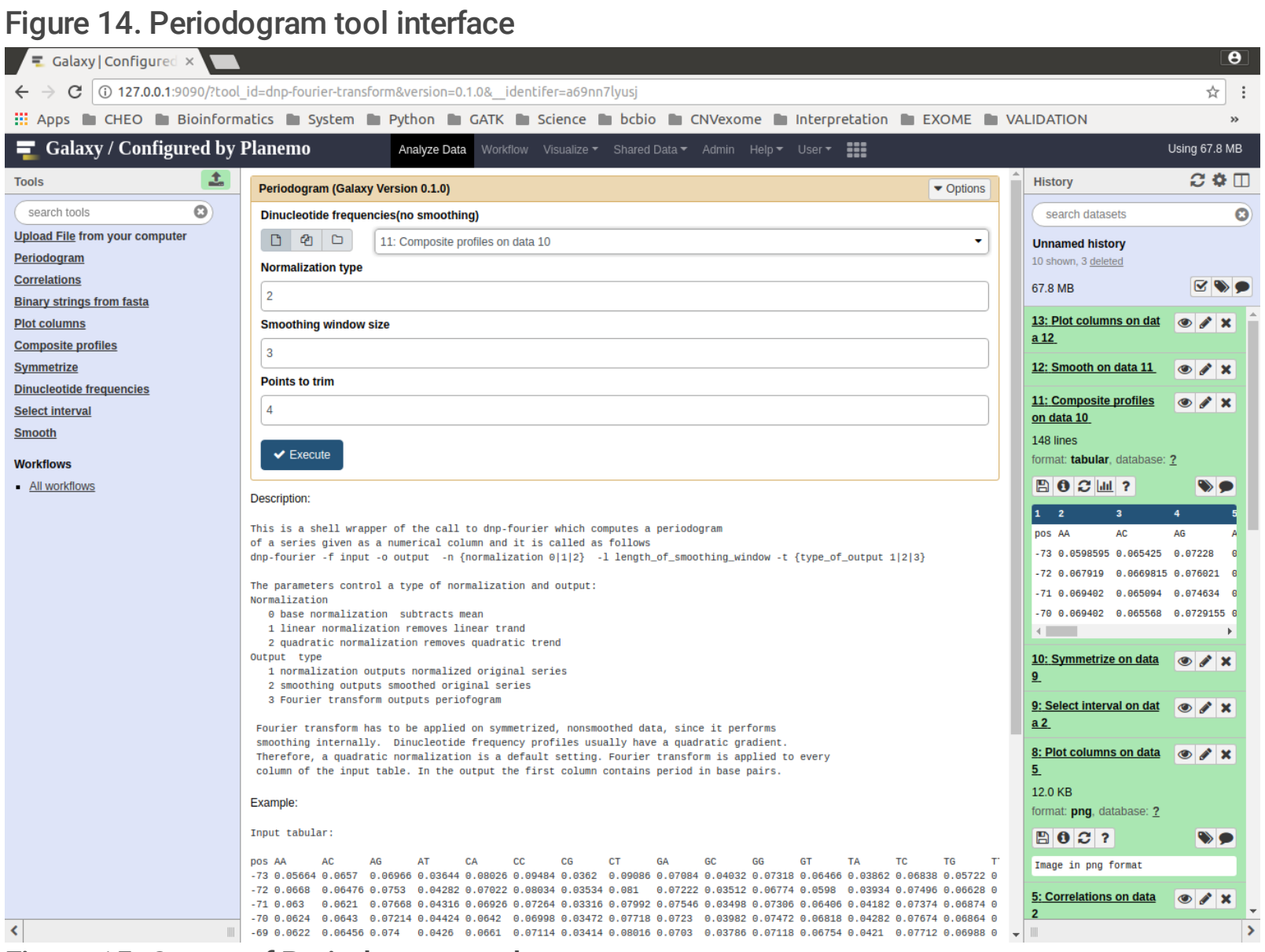


Figure 15. Output of Periodogram tool