Supplementary Information

MPF-BML: A standalone GUI-based package for maximum entropy model inference

Ahmed A. Quadeer[1], Matthew R. McKay[1,2], John P. Barton[3], Raymond H. Y. Louie[4,5,]*

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, [2]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, [3]Department of Physics and Astronomy, University of California, Riverside, USA, [4]The Kirby Institute, University of New South Wales, Sydney, Australia, [5]School of Medical Sciences, University of New South Wales, Sydney, Australia.

*To whom correspondence should be addressed.
Contact: rlouie@kirby.unsw.edu.au

Table of contents

Supplementary Text

Text S1. Inputs

Besides the mandatory input, optional inputs are: (i) the maximum fraction of gaps above which the variable is removed (default value set to 0.5); (ii) an optional sub-sample number to train MPF (used if MPF is running slow); (iii) an optional input file in csv/xls format comprising of sample weights to reduce sampling bias (each sample is weighted according to a maximum fraction threshold of similar sequences as default); and (iv) an optional file comprising model parameters to initialize the method.

For (iii), reweighting is performed based on the provided threshold $0 < x < 1$ as follows (Morcos 2011): Two sequences with Hamming distance less than xN (where N is the length of protein) are considered to carry almost the same information, and vice versa. Thus, for each sequence m, a weight $w_m = 1/z_m$ is calculated, where z_m are all the sequences in the MSA having a hamming distance less than xN with the sequence m. A threshold value of 0.1 is used as default if the user does not provide any weight file.

For (iv), the user can provide an optional input maximum entropy model to initialize the MPF-BML method. This model may be obtained from some other inference method or from a previous run of the MPF method (the model obtained after running MPF is automatically saved once the MPF method converges) for which the user is interested in running either BML method alone or both MPF (again) and BML methods with a different set of parameters. For the latter, the user needs to check the provided “Run only BML step on user-provided parameters” option in the “Input information” panel.

Text S2. MPF-BML parameters

Here, we provide details of all parameters available in the “Parameters” panel of the MPF-BML GUI.

Step 1: Mutant combining

phi_opt: Mutant combining factor which indicates the ratio of entropy with grouping of states to the entropy without grouping. See (Louie *et al.*, 2018) for details. A value of 0 indicates each site is coarse-grained to only two states, a value of 1 indicates all the states from the original MSA are kept.

Step 2: MPF

L1 reg: L1 regularization parameter. Higher value indicates more couplings will be set to zero.

L2 reg: L2 regularization parameter. Higher value indicates high couplings values will be more substantially reduced to zero.

Grad. tol: If the maximum gradient of the MPF function falls below this tolerance level, the MPF algorithm will terminate. Smaller value will require more time for the algorithm to converge but will converge with more accuracy.

Func. tol: If the absolute difference between the MPF function in the previous and current iteration falls below this tolerance level, the MPF algorithm terminates. Smaller value will require more time for the algorithm to converge but will converge with more accuracy.

Max iter: Maximum number of iterations before the MPF algorithm terminates.

Step 3: BML

Max iter: Maximum number of iterations before the BML algorithm terminates.

Max eps: Maximum epsilon value for the individual frequencies, pairwise frequencies and connected correlations. The epsilon value is chosen to balance between underfitting and overfitting the individual frequencies, pairwise frequencies and connected correlations. Higher value will mean less accuracy but faster convergence. See (Louie *et al.*, 2018) for details.

Param. tol: If the sum absolute difference in parameter values in the previous and current iteration falls below this tolerance level, the BML algorithm terminates. Smaller value will require more time for the algorithm to converge but will converge with more accuracy.

Grad. tol: If the sum absolute difference in the gradient of the parameter values in the previous and current iteration falls below this tolerance level, the BML algorithm terminates. Smaller value will require more time for the algorithm to converge but will converge with more accuracy.

Thinning: Keep every x samples in the MCMC algorithm, where x is the thinning parameter. Higher value will mean less samples, thus less accuracy but faster speed.

Burn-in: Remove the first x samples in the MCMC algorithm, where x is the burn-in parameter. Higher value will mean the MCMC algorithm will produce more accurate output, but slower speed.

Cores: Number of cores to use in the MCMC algorithm. If set to a value greater than or equal to the maximum number of cores (N_c) available in the system, the application resets it to N_c – 1 to avoid using all available resources.

MCMC samples: Number of samples in the MCMC algorithm before thinning and burn-in. Higher value will mean the MCMC algorithm will produce more accurate output, but slower speed.

The following are RPROP specific parameters, See (Riedmiller and Braun, 1993) for more details.

Gamma h: Initial weight update value for field (h) parameters. Set larger for larger step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

Gamma J: Initial weight update value for coupling (J) parameters. Set larger for larger step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

h pos: Weight increase factor for the fields with positive gradient. Set larger for larger step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

h neg: Weight increase factor for the fields with negative gradient. Set larger for larger step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

J pos: Weight increase factor for the couplings with positive gradient. Set larger for larger step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

J neg: Weight increase factor for the couplings with negative gradient. Set larger for larger step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

h delta max: The maximum weight update per iteration for the fields. Set larger for larger possible step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

h delta min: The minimum weight update per iteration for the fields. Set larger for larger possible step sizes. Smaller value will mean initial faster convergence of algorithm, but converged values may be less accurate.

J delta max: The maximum weight update per iteration for the couplings. Set larger for larger possible step sizes. Higher value will mean initial faster convergence of algorithm, but converged values may be less accurate.

J delta min: The minimum weight update per iteration for the couplings. Set larger for larger possible step sizes. Smaller value will mean initial faster convergence of algorithm, but converged values may be less accurate.

Text S3. Contact prediction using the inferred couplings

A tab-delimited text file is also provided as an output with pairs of positions arranged in descending order according to their computed Frobenius norm, a metric representative of the pairs in contact. This metric is calculated using the inferred couplings (Cocco *et al.*, 2018) and the higher the value of this metric, the higher is the chance of this pair to be in contact, and vice versa.

Supplementary Tables

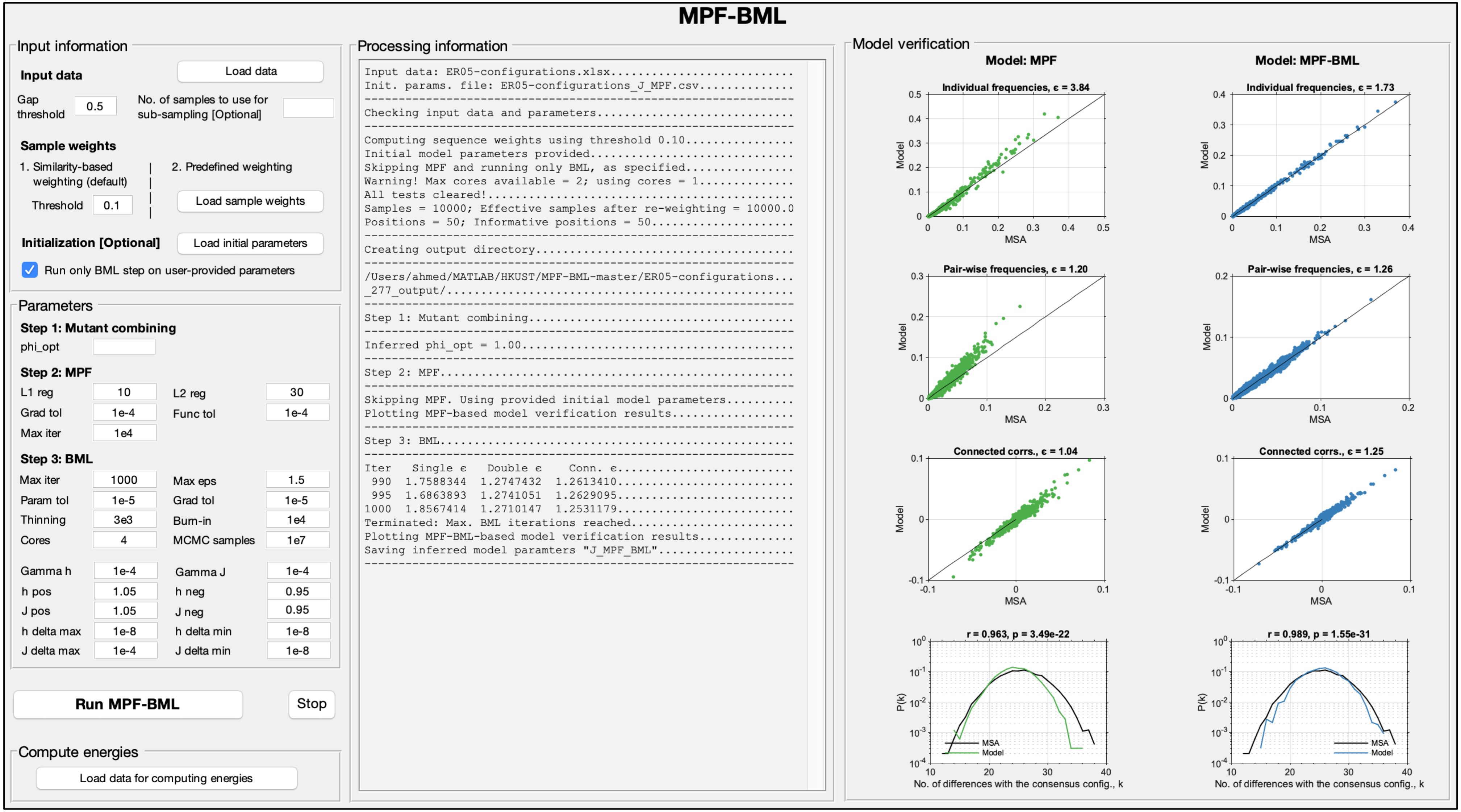
Table S1. Summary of the input datasets analyzed using MPF-BML.

Dataset	Type of data	Format	Weights available	Number of samples	Number of variables	Number of parameters	Associated figure
HCV E2 surface glycoprotein	Amino acid sequence data	FASTA	Yes	3,363	352	~ $2 x 10^6$	Fig. 1
Erdos-Renyi random graphs (ER05)	Synthetic categorical data	Microsoft Excel Open XML	No	10,000	50	~ $5 x 10^4$	Fig. S1
HIV p7 nucleocapsid protein	Amino acid sequence data	FASTA	No	4,131	71	~ $1 x 10^5$	Fig. S2
Trypsin inhibitor protein family (PF00014)	Amino acid sequence data	FASTA	No	4,915	53	~ $5 x 10^4$	Fig. S3
Breast cancer data	Somatic mutations (binary data)	Microsoft Excel Open XML	No	2,327	100	~ $5 x 10^3$	Fig. S4
Wisconsin breast cancer data (UCI repository)	Categorical data	Microsoft Excel Open XML	No	699	9	~ $3 x 10^3$	Fig. S5
Chess data (UCI repository)	Categorical data	Microsoft Excel Open XML	No	3,196	36	~ $5 x 10^2$	Fig. S6

Table S2. Format of the inferred maximum entropy model parameters using MPF-BML. Values are shown for a synthetic test data set. The first row and column show the change of configuration at a position with A/1/B denoting a change from configuration A to configuration B at position 1. The diagonal entries of the matrix represent the inferred fields while the non-diagonal entries represent the inferred couplings. These parameters are saved automatically as a tab delimited file.

	M/1/-	R/2/K	R/2/T	V/3/A	K/4/M	K/4/T	G/5/E	I/6/T	I/6/M	R/7/K	R/7/M	K/8/R	N/9/S
M/1/-	6.43877	-0.31164	0.487343	-0.8045	0.325733	-0.14218	-0.27445	0.150707	0.668197	0.068006	0.28129	0.27983	0.071754
R/2/K	-0.31164	4.690157	0.004776	0.246884	0.266135	-1.31717	-0.72883	0.50299	-0.35844	-0.53223	1.163576	-0.03381	0.365269
R/2/T	0.487343	0.004776	-1.0441	-0.12154	0.953398	0.620634	0.588852	0.322001	-0.88575	0.614748	0.12424	0.870705	0.503503
V/3/A	-0.8045	0.246884	-0.12154	4.172256	0.283196	0.787504	-0.00353	-0.26528	0.033055	-0.44571	-1.2515	-0.23769	1.322479
K/4/M	0.325733	0.266135	0.953398	0.283196	2.792618	-2.1E-05	-0.95053	0.333316	1.192085	0.744931	-0.27076	-0.62612	0.34005
K/4/T	-0.14218	-1.31717	0.620634	0.787504	-2.1E-05	3.632102	-0.1222	0.800195	0.250115	0.937377	-0.27037	-0.09487	1.162473
G/5/E	-0.27445	-0.72883	0.588852	-0.00353	-0.95053	-0.1222	2.732045	-0.03059	-1.84851	0.655304	0.064181	-0.39161	-0.14443
I/6/T	0.150707	0.50299	0.322001	-0.26528	0.333316	0.800195	-0.03059	3.744127	-5.5E-05	0.274624	0.329657	0.040394	-0.69886
I/6/M	0.668197	-0.35844	-0.88575	0.033055	1.192085	0.250115	-1.84851	-5.5E-05	3.82687	1.352347	-0.40026	1.040857	0.853013
R/7/K	0.068006	-0.53223	0.614748	-0.44571	0.744931	0.937377	0.655304	0.274624	1.352347	5.206358	-1E-04	-0.8084	-0.01596
R/7/M	0.28129	1.163576	0.12424	-1.2515	-0.27076	-0.27037	0.064181	0.329657	-0.40026	-1E-04	5.098788	0.930253	-0.61156
K/8/R	0.27983	-0.03381	0.870705	-0.23769	-0.62612	-0.09487	-0.39161	0.040394	1.040857	-0.8084	0.930253	4.569314	-0.99103
N/9/S	0.071754	0.365269	0.503503	1.322479	0.34005	1.162473	-0.14443	-0.69886	0.853013	-0.01596	-0.61156	-0.99103	5.057654

Supplementary Figures



MPF-BML

Input information

Input data

Load data

Gap threshold 0.5

No. of samples to use for sub-sampling [Optional]

Sample weights

1. Similarity-based weighting (default)

Threshold 0.1

2. Predefined weighting

Load sample weights

Initialization [Optional]

Load initial parameters

☑ Run only BML step on user-provided parameters

Parameters

Step 1: Mutant combining

phi_opt

Step 2: MPF

L1 reg	10	L2 reg	30
Grad tol	1e-4	Func tol	1e-4
Max iter	1e4		

Step 3: BML

Max iter	1000	Max eps	1.5
Param tol	1e-5	Grad tol	1e-5
Thinning	3e3	Burn-in	1e4
Cores	4	MCMC samples	1e7
Gamma h	1e-4	Gamma J	1e-4
h pos	1.05	h neg	0.95
J pos	1.05	J neg	0.95
h delta max	1e-8	h delta min	1e-8
J delta max	1e-4	J delta min	1e-8

Run MPF-BML

Stop

Compute energies

Load data for computing energies

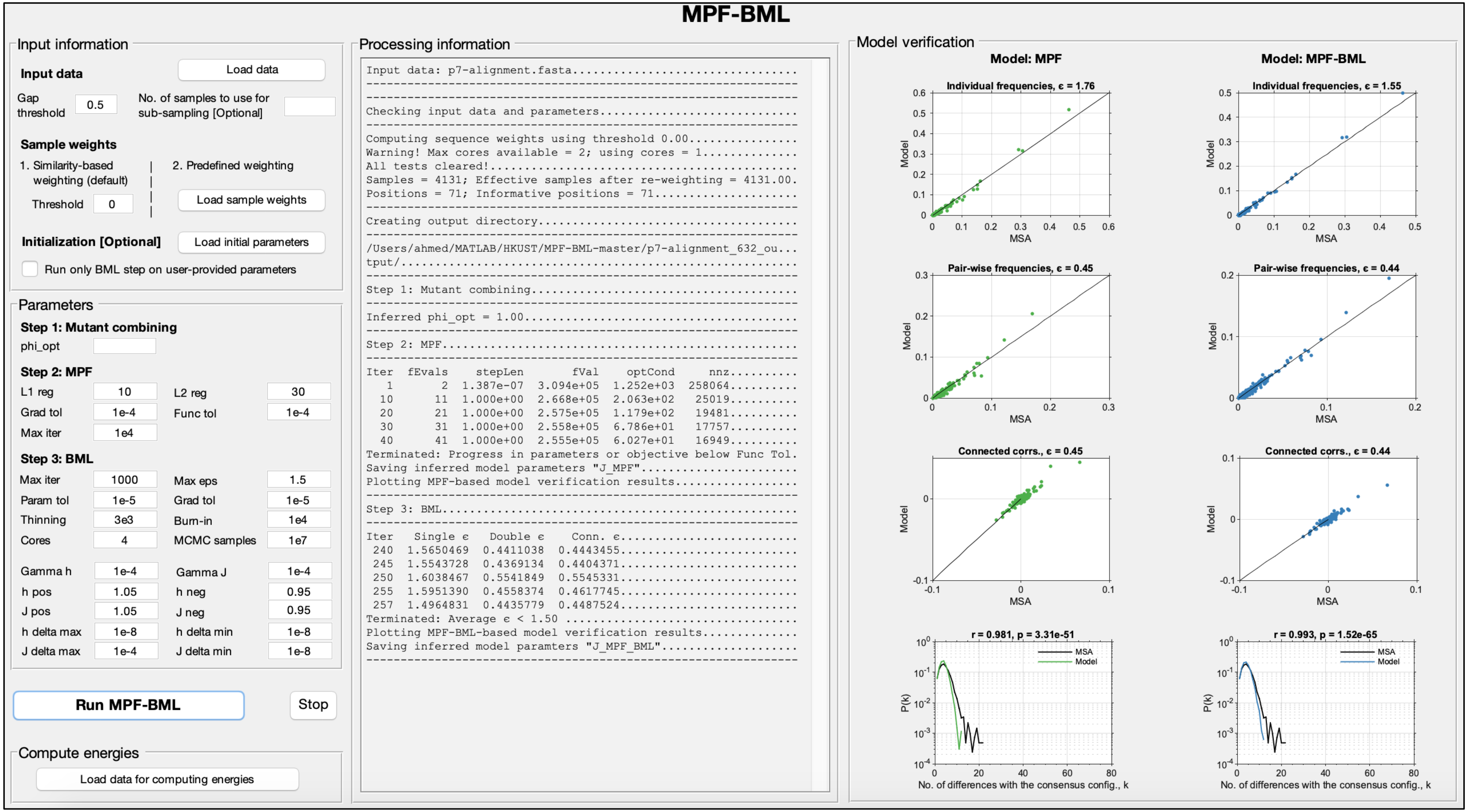
Processing information

```
Input data: ER05-configurations.xlsx...........................
Init. params. file: ER05-configurations_J_MPF.csv.............
---------------------------------------------------------------
Checking input data and parameters............................
---------------------------------------------------------------
Computing sequence weights using threshold 0.10...............
Initial model parameters provided.............................
Skipping MPF and running only BML, as specified...............
Warning! Max cores available = 2; using cores = 1.............
All tests cleared!............................................
Samples = 10000; Effective samples after re-weighting = 10000.0
Positions = 50; Informative positions = 50....................
---------------------------------------------------------------
Creating output directory.....................................
---------------------------------------------------------------
/Users/ahmed/MATLAB/HKUST/MPF-BML-master/ER05-configurations...
_277_output/..................................................
---------------------------------------------------------------
Step 1: Mutant combining......................................
---------------------------------------------------------------
Inferred phi_opt = 1.00.......................................
---------------------------------------------------------------
Step 2: MPF...................................................
---------------------------------------------------------------
Skipping MPF. Using provided initial model parameters.........
Plotting MPF-based model verification results.................
---------------------------------------------------------------
Step 3: BML...................................................
---------------------------------------------------------------
Iter   Single ϵ   Double ϵ    Conn. ϵ.........................
 990  1.7588344  1.2747432  1.2613410.........................
 995  1.6863893  1.2741051  1.2629095.........................
1000  1.8567414  1.2710147  1.2531179.........................
Terminated: Max. BML iterations reached.......................
Plotting MPF-BML-based model verification results.............
Saving inferred model paramters "J_MPF_BML"...................
---------------------------------------------------------------
```

Model verification

Model: MPF
Model: MPF-BML
Individual frequencies, ϵ = 3.84
Individual frequencies, ϵ = 1.73
Pair-wise frequencies, ϵ = 1.20
Pair-wise frequencies, ϵ = 1.26
Connected corrs., ϵ = 1.04
Connected corrs., ϵ = 1.25
r = 0.963, p = 3.49e-22
r = 0.989, p = 1.55e-31
Model
MSA
P(k)
No. of differences with the consensus config., k
MSA
Model
0
0.1
0.2
0.3
0.4
0.5
-0.1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10
20
30
40

Figure S1. Model inferred using the MPF-BML package for Erdos-Renyi random graphs (ER05), analyzed in (Barton *et al.*, 2016). Here, the model is initialized by a set of parameters previously inferred using the MPF method.


MPF-BML
Load data
Load sample weights
Load initial parameters
Run MPF-BML
Stop
Load data for computing energies
Model verification
Model: MPF
Model: MPF-BML
Individual frequencies, ϵ = 1.76
Individual frequencies, ϵ = 1.55
Pair-wise frequencies, ϵ = 0.45
Pair-wise frequencies, ϵ = 0.44
Connected corrs., ϵ = 0.45
Connected corrs., ϵ = 0.44
r = 0.981, p = 3.31e-51
r = 0.993, p = 1.52e-65
Model
MSA
P(k)
No. of differences with the consensus config., k
MSA
Model
0
0.1
0.2
0.3
0.4
0.5
0.6
-0.1
20
40
60
80
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}

Input information

Input data

Gap threshold: 0.5

No. of samples to use for sub-sampling [Optional]:

Sample weights

1. Similarity-based weighting (default) | 2. Predefined weighting

Threshold: 0

Initialization [Optional]

□ Run only BML step on user-provided parameters

Parameters

Step 1: Mutant combining

phi_opt:

Step 2: MPF

L1 reg	10	L2 reg	30
Grad tol	1e-4	Func tol	1e-4
Max iter	1e4		

Step 3: BML

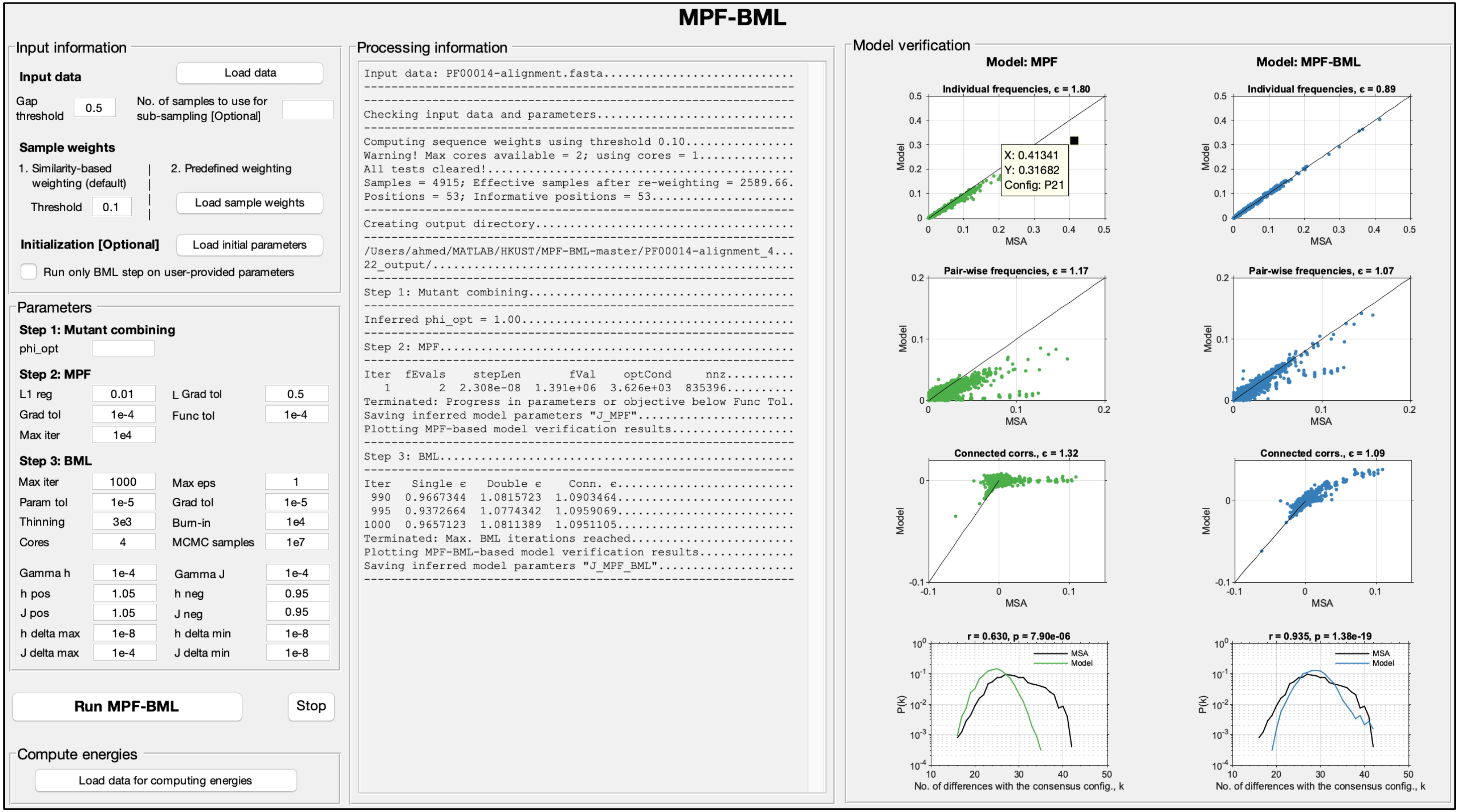
Max iter	1000	Max eps	1.5
Param tol	1e-5	Grad tol	1e-5
Thinning	3e3	Burn-in	1e4
Cores	4	MCMC samples	1e7
Gamma h	1e-4	Gamma J	1e-4
h pos	1.05	h neg	0.95
J pos	1.05	J neg	0.95
h delta max	1e-8	h delta min	1e-8
J delta max	1e-4	J delta min	1e-8

Compute energies

Processing information

```
Input data: p7-alignment.fasta.................................
---------------------------------------------------------------
---------------------------------------------------------------
Checking input data and parameters.............................
---------------------------------------------------------------
Computing sequence weights using threshold 0.00................
Warning! Max cores available = 2; using cores = 1..............
All tests cleared!.............................................
Samples = 4131; Effective samples after re-weighting = 4131.00.
Positions = 71; Informative positions = 71.....................
---------------------------------------------------------------
Creating output directory......................................
---------------------------------------------------------------
/Users/ahmed/MATLAB/HKUST/MPF-BML-master/p7-alignment_632_ou...
tput/..........................................................
---------------------------------------------------------------
Step 1: Mutant combining.......................................
---------------------------------------------------------------
Inferred phi_opt = 1.00........................................
---------------------------------------------------------------
Step 2: MPF....................................................
---------------------------------------------------------------
Iter  fEvals    stepLen        fVal    optCond     nnz.........
   1       2  1.387e-07  3.094e+05  1.252e+03  258064.........
  10      11  1.000e+00  2.668e+05  2.063e+02   25019.........
  20      21  1.000e+00  2.575e+05  1.179e+02   19481.........
  30      31  1.000e+00  2.558e+05  6.786e+01   17757.........
  40      41  1.000e+00  2.555e+05  6.027e+01   16949.........
Terminated: Progress in parameters or objective below Func Tol.
Saving inferred model parameters "J_MPF".......................
Plotting MPF-based model verification results..................
---------------------------------------------------------------
Step 3: BML....................................................
---------------------------------------------------------------
Iter   Single ϵ   Double ϵ    Conn. ϵ.........................
 240  1.5650469  0.4411038  0.4443455.........................
 245  1.5543728  0.4369134  0.4404371.........................
 250  1.6038467  0.5541849  0.5545331.........................
 255  1.5951390  0.4558374  0.4617745.........................
 257  1.4964831  0.4435779  0.4487524.........................
Terminated: Average ϵ < 1.50 ..................................
Plotting MPF-BML-based model verification results..............
Saving inferred model paramters "J_MPF_BML"....................
---------------------------------------------------------------
```

Figure S2. Model inferred using the MPF-BML package for HIV p7 nucleocapsid protein, analyzed in (Barton *et al.*, 2016).



MPF-BML

Input information

Input data

Gap threshold 0.5

No. of samples to use for sub-sampling [Optional]

Sample weights

1. Similarity-based weighting (default)

Threshold 0.1

2. Predefined weighting

Initialization [Optional]

☐ Run only BML step on user-provided parameters

Parameters

Step 1: Mutant combining

phi_opt

Step 2: MPF

L1 reg	0.01	L Grad tol	0.5
Grad tol	1e-4	Func tol	1e-4
Max iter	1e4		

Step 3: BML

Max iter	1000	Max eps	1
Param tol	1e-5	Grad tol	1e-5
Thinning	3e3	Burn-in	1e4
Cores	4	MCMC samples	1e7
Gamma h	1e-4	Gamma J	1e-4
h pos	1.05	h neg	0.95
J pos	1.05	J neg	0.95
h delta max	1e-8	h delta min	1e-8
J delta max	1e-4	J delta min	1e-8

Compute energies

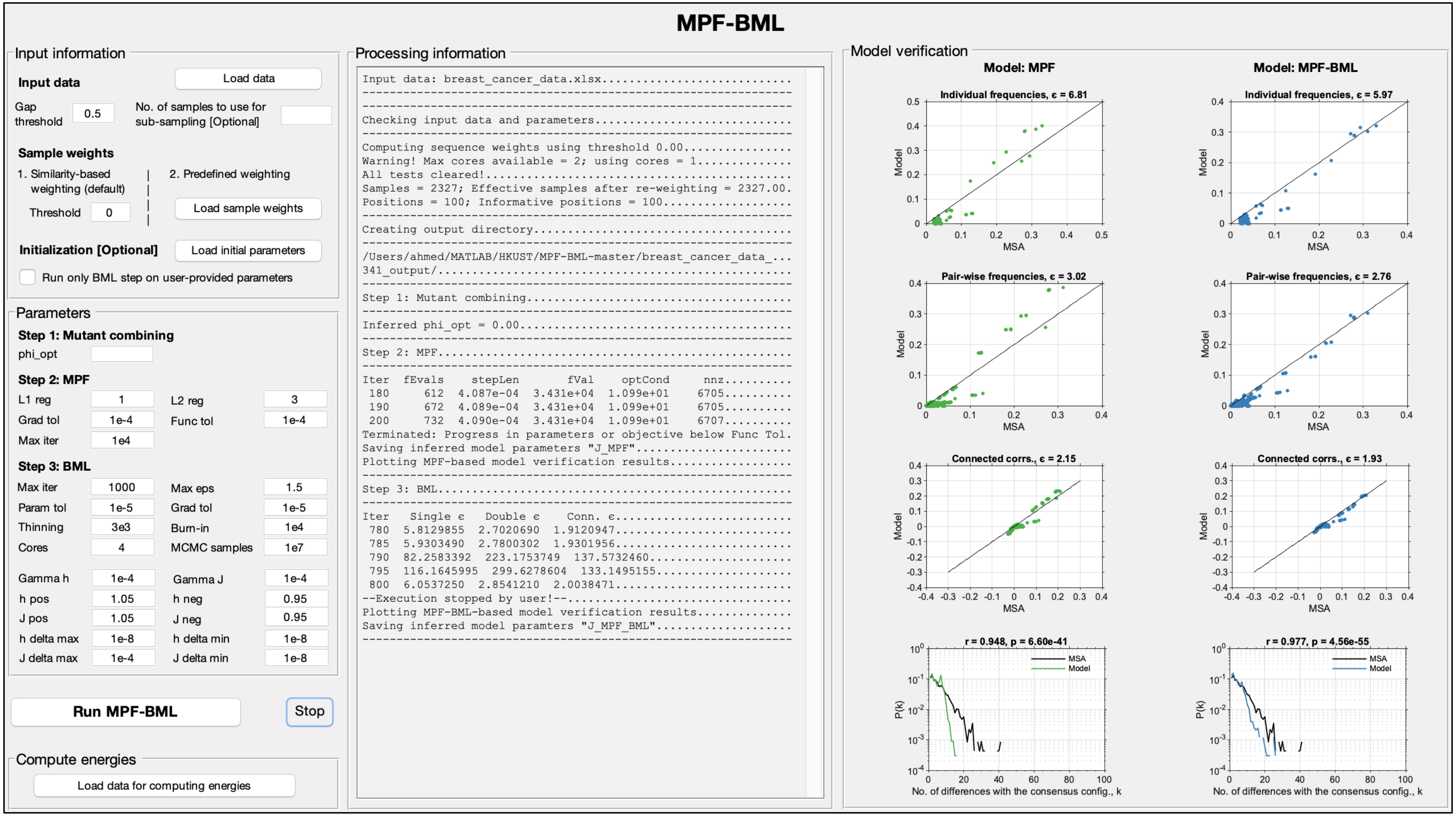
Processing information

```
Input data: PF00014-alignment.fasta............................
---------------------------------------------------------------
---------------------------------------------------------------
Checking input data and parameters.............................
---------------------------------------------------------------
Computing sequence weights using threshold 0.10................
Warning! Max cores available = 2; using cores = 1..............
All tests cleared!.............................................
Samples = 4915; Effective samples after re-weighting = 2589.66.
Positions = 53; Informative positions = 53.....................
---------------------------------------------------------------
Creating output directory......................................
---------------------------------------------------------------
/Users/ahmed/MATLAB/HKUST/MPF-BML-master/PF00014-alignment_4...
22_output/.....................................................
---------------------------------------------------------------
Step 1: Mutant combining.......................................
---------------------------------------------------------------
Inferred phi_opt = 1.00........................................
---------------------------------------------------------------
Step 2: MPF....................................................
---------------------------------------------------------------
Iter  fEvals    stepLen       fVal    optCond      nnz.........
   1       2  2.308e-08  1.391e+06  3.626e+03  835396.........
Terminated: Progress in parameters or objective below Func Tol.
Saving inferred model parameters "J_MPF".......................
Plotting MPF-based model verification results..................
---------------------------------------------------------------
Step 3: BML....................................................
---------------------------------------------------------------
Iter   Single ϵ   Double ϵ    Conn. ϵ..........................
 990  0.9667344  1.0815723  1.0903464..........................
 995  0.9372664  1.0774342  1.0959069..........................
1000  0.9657123  1.0811389  1.0951105..........................
Terminated: Max. BML iterations reached........................
Plotting MPF-BML-based model verification results..............
Saving inferred model paramters "J_MPF_BML"....................
---------------------------------------------------------------
```

Model verification

Model: MPF
Model: MPF-BML
Individual frequencies, ϵ = 1.80
Individual frequencies, ϵ = 0.89
X: 0.41341
Y: 0.31682
Config: P21
Model
MSA
0
0.1
0.2
0.3
0.4
0.5
Pair-wise frequencies, ϵ = 1.17
Pair-wise frequencies, ϵ = 1.07
Connected corrs., ϵ = 1.32
Connected corrs., ϵ = 1.09
-0.1
r = 0.630, p = 7.90e-06
r = 0.935, p = 1.38e-19
MSA
Model
P(k)
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10
20
30
40
50
No. of differences with the consensus config., k

Figure S3. Model inferred using the MPF-BML package for Trypsin inhibitor protein family (PF00014), analyzed in (Barton *et al.*, 2016).



Input data

Gap threshold 0.5

No. of samples to use for sub-sampling [Optional]

Sample weights

1. Similarity-based weighting (default)

Threshold 0

2. Predefined weighting

Initialization [Optional]

☐ Run only BML step on user-provided parameters

Step 1: Mutant combining

phi_opt

Step 2: MPF

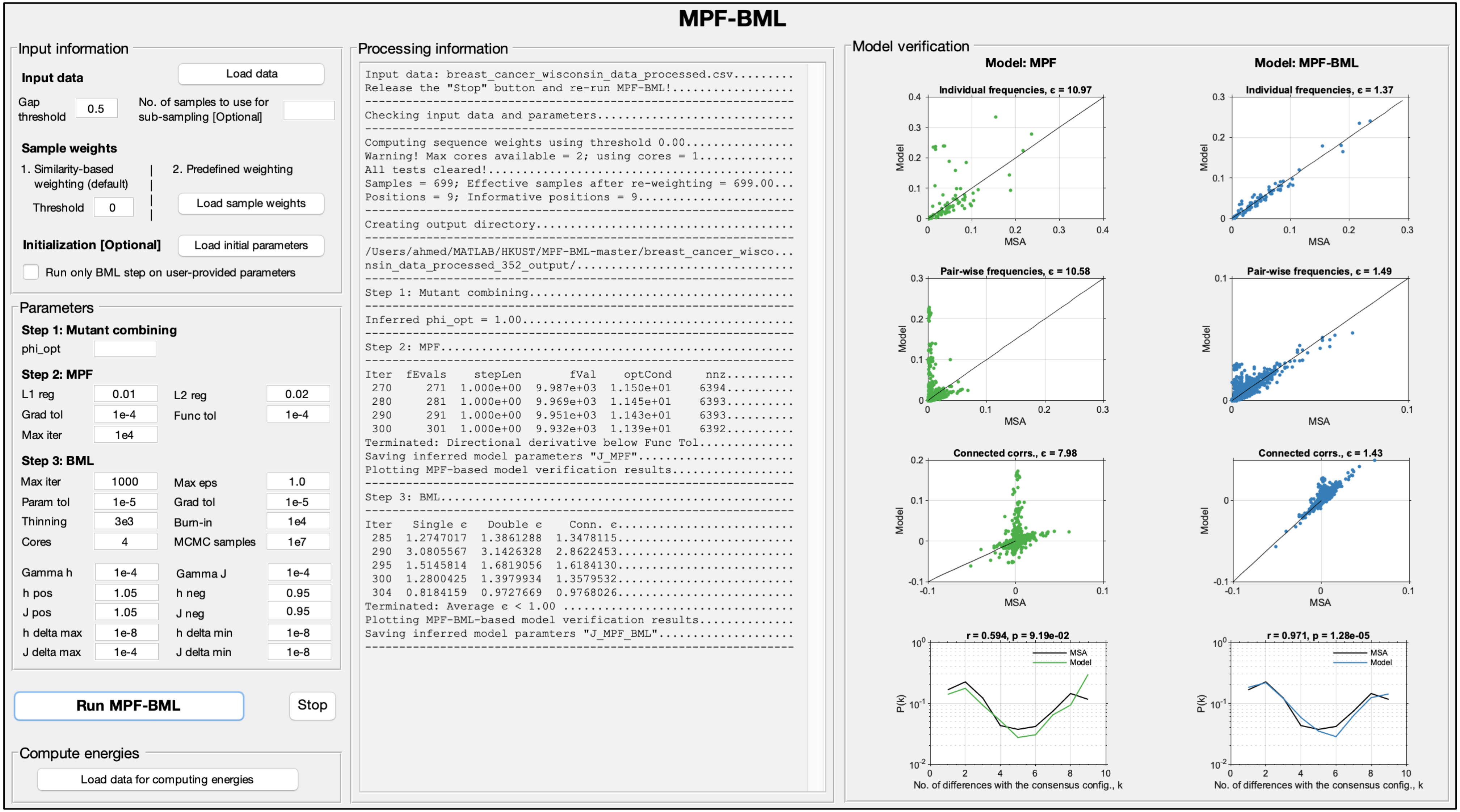
L1 reg	1	L2 reg	3
Grad tol	1e-4	Func tol	1e-4
Max iter	1e4		

Step 3: BML

Max iter	1000	Max eps	1.5
Param tol	1e-5	Grad tol	1e-5
Thinning	3e3	Burn-in	1e4
Cores	4	MCMC samples	1e7
Gamma h	1e-4	Gamma J	1e-4
h pos	1.05	h neg	0.95
J pos	1.05	J neg	0.95
h delta max	1e-8	h delta min	1e-8
J delta max	1e-4	J delta min	1e-8

```
Input data: breast_cancer_data.xlsx...........................
--------------------------------------------------------------
--------------------------------------------------------------
Checking input data and parameters............................
--------------------------------------------------------------
Computing sequence weights using threshold 0.00...............
Warning! Max cores available = 2; using cores = 1.............
All tests cleared!............................................
Samples = 2327; Effective samples after re-weighting = 2327.00.
Positions = 100; Informative positions = 100..................
--------------------------------------------------------------
Creating output directory.....................................
--------------------------------------------------------------
/Users/ahmed/MATLAB/HKUST/MPF-BML-master/breast_cancer_data_...
341_output/...................................................
--------------------------------------------------------------
Step 1: Mutant combining......................................
--------------------------------------------------------------
Inferred phi_opt = 0.00.......................................
--------------------------------------------------------------
Step 2: MPF...................................................
--------------------------------------------------------------
Iter  fEvals    stepLen        fVal     optCond      nnz..........
 180     612  4.087e-04  3.431e+04  1.099e+01    6705..........
 190     672  4.089e-04  3.431e+04  1.099e+01    6705..........
 200     732  4.090e-04  3.431e+04  1.099e+01    6707..........
Terminated: Progress in parameters or objective below Func Tol.
Saving inferred model parameters "J_MPF"......................
Plotting MPF-based model verification results.................
--------------------------------------------------------------
Step 3: BML...................................................
--------------------------------------------------------------
Iter   Single ϵ   Double ϵ    Conn. ϵ.........................
 780  5.8129855  2.7020690  1.9120947.........................
 785  5.9303490  2.7800302  1.9301956.........................
 790  82.2583392  223.1753749  137.5732460....................
 795  116.1645995  299.6278604  133.1495155...................
 800  6.0537250  2.8541210  2.0038471.........................
--Execution stopped by user!--................................
Plotting MPF-BML-based model verification results.............
Saving inferred model paramters "J_MPF_BML"...................
--------------------------------------------------------------
```

Figure S4. Model inferred using the MPF-BML package for whole genome somatic mutation (100 most highly expressed genes) breast cancer data (Bamford *et al.*, 2004) obtained from http://cancer.sanger.ac.uk/.



MPF-BML

Input information

Input data

Gap threshold: 0.5

No. of samples to use for sub-sampling [Optional]:

Sample weights

1. Similarity-based weighting (default)

Threshold: 0

2. Predefined weighting

Initialization [Optional]

□ Run only BML step on user-provided parameters

Parameters

Step 1: Mutant combining

phi_opt:

Step 2: MPF

L1 reg	0.01	L2 reg	0.02
Grad tol	1e-4	Func tol	1e-4
Max iter	1e4		

Step 3: BML

Max iter	1000	Max eps	1.0
Param tol	1e-5	Grad tol	1e-5
Thinning	3e3	Burn-in	1e4
Cores	4	MCMC samples	1e7
Gamma h	1e-4	Gamma J	1e-4
h pos	1.05	h neg	0.95
J pos	1.05	J neg	0.95
h delta max	1e-8	h delta min	1e-8
J delta max	1e-4	J delta min	1e-8

Compute energies

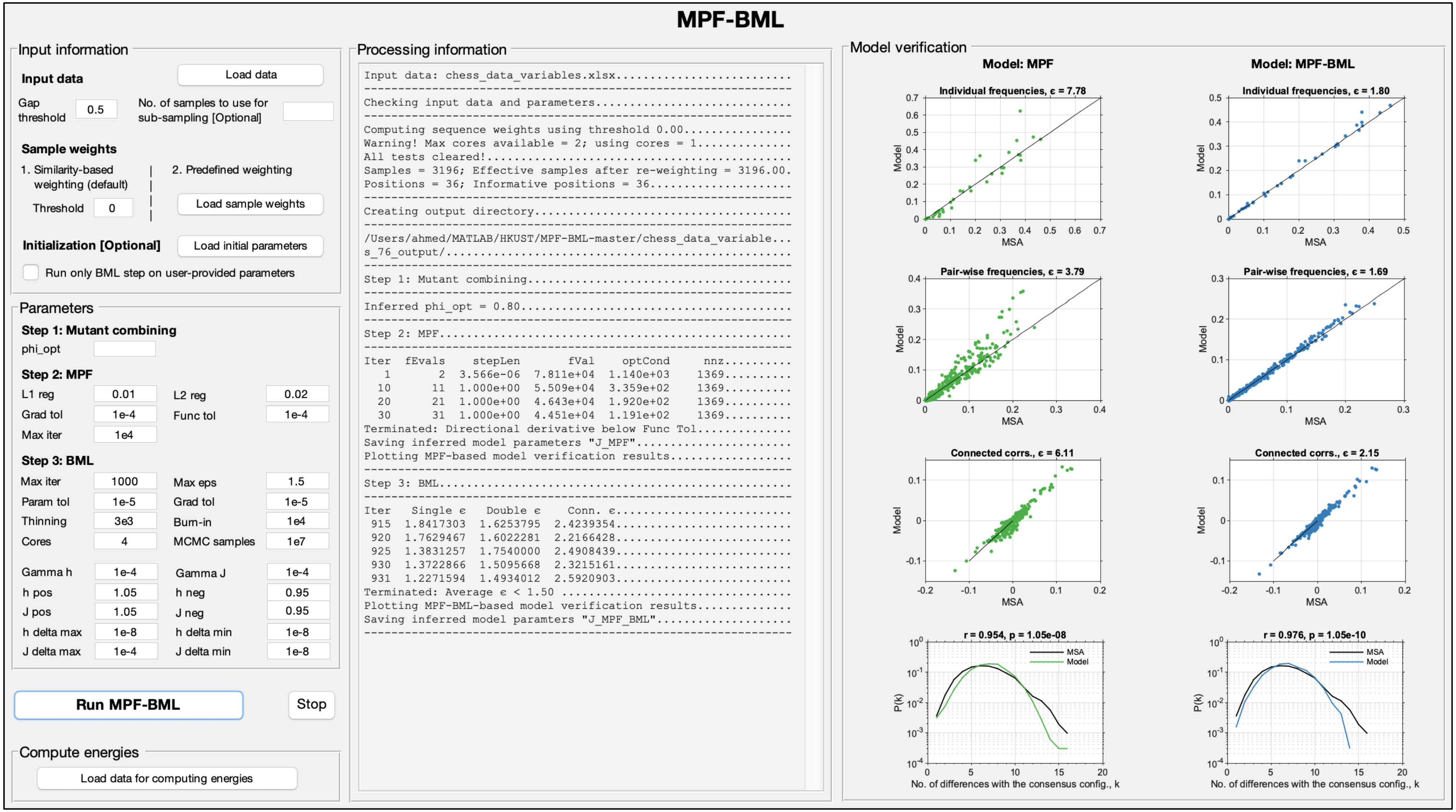
Processing information

```
Input data: breast_cancer_wisconsin_data_processed.csv.........
Release the "Stop" button and re-run MPF-BML!..................
---------------------------------------------------------------
Checking input data and parameters.............................
---------------------------------------------------------------
Computing sequence weights using threshold 0.00................
Warning! Max cores available = 2; using cores = 1..............
All tests cleared!.............................................
Samples = 699; Effective samples after re-weighting = 699.00...
Positions = 9; Informative positions = 9.......................
---------------------------------------------------------------
Creating output directory......................................
---------------------------------------------------------------
/Users/ahmed/MATLAB/HKUST/MPF-BML-master/breast_cancer_wisco...
nsin_data_processed_352_output/................................
---------------------------------------------------------------
Step 1: Mutant combining.......................................
---------------------------------------------------------------
Inferred phi_opt = 1.00........................................
---------------------------------------------------------------
Step 2: MPF....................................................
---------------------------------------------------------------
Iter  fEvals    stepLen       fVal    optCond      nnz..........
 270     271  1.000e+00  9.987e+03  1.150e+01    6394..........
 280     281  1.000e+00  9.969e+03  1.145e+01    6393..........
 290     291  1.000e+00  9.951e+03  1.143e+01    6393..........
 300     301  1.000e+00  9.932e+03  1.139e+01    6392..........
Terminated: Directional derivative below Func Tol..............
Saving inferred model parameters "J_MPF".......................
Plotting MPF-based model verification results..................
---------------------------------------------------------------
Step 3: BML....................................................
---------------------------------------------------------------
Iter   Single ϵ   Double ϵ    Conn. ϵ.........................
 285  1.2747017  1.3861288  1.3478115.........................
 290  3.0805567  3.1426328  2.8622453.........................
 295  1.5145814  1.6819056  1.6184130.........................
 300  1.2800425  1.3979934  1.3579532.........................
 304  0.8184159  0.9727669  0.9768026.........................
Terminated: Average ϵ < 1.00 ..................................
Plotting MPF-BML-based model verification results..............
Saving inferred model paramters "J_MPF_BML"....................
---------------------------------------------------------------
```

Model verification

Load data
Load sample weights
Load initial parameters
Run MPF-BML
Stop
Load data for computing energies
Model: MPF
Model: MPF-BML
Individual frequencies, ϵ = 10.97
Individual frequencies, ϵ = 1.37
Pair-wise frequencies, ϵ = 10.58
Pair-wise frequencies, ϵ = 1.49
Connected corrs., ϵ = 7.98
Connected corrs., ϵ = 1.43
r = 0.594, p = 9.19e-02
r = 0.971, p = 1.28e-05
Model
MSA
0
0.1
0.2
0.3
0.4
-0.1
P(k)
10^0
10^{-1}
10^{-2}
2
4
6
8
10
MSA
Model
No. of differences with the consensus config., k

Figure S5. Model inferred using the MPF-BML package for the classical Wisconsin breast cancer data (Wolberg and Mangasarian, 1990) obtained from the UCI repository (Asuncion and Newman, 2007).



MPF-BML

Input information

Input data

Load data

Gap threshold: 0.5

No. of samples to use for sub-sampling [Optional]:

Sample weights

1. Similarity-based weighting (default)

Threshold: 0

2. Predefined weighting

Load sample weights

Initialization [Optional]

Load initial parameters

□ Run only BML step on user-provided parameters

Parameters

Step 1: Mutant combining

phi_opt:

Step 2: MPF

L1 reg: 0.01

L2 reg: 0.02

Grad tol: 1e-4

Func tol: 1e-4

Max iter: 1e4

Step 3: BML

Max iter: 1000

Max eps: 1.5

Param tol: 1e-5

Grad tol: 1e-5

Thinning: 3e3

Burn-in: 1e4

Cores: 4

MCMC samples: 1e7

Gamma h: 1e-4

Gamma J: 1e-4

h pos: 1.05

h neg: 0.95

J pos: 1.05

J neg: 0.95

h delta max: 1e-8

h delta min: 1e-8

J delta max: 1e-4

J delta min: 1e-8

Run MPF-BML

Stop

Compute energies

Load data for computing energies

Processing information

```
Input data: chess_data_variables.xlsx..........................
---------------------------------------------------------------
Checking input data and parameters.............................
---------------------------------------------------------------
Computing sequence weights using threshold 0.00................
Warning! Max cores available = 2; using cores = 1..............
All tests cleared!.............................................
Samples = 3196; Effective samples after re-weighting = 3196.00.
Positions = 36; Informative positions = 36.....................
---------------------------------------------------------------
Creating output directory......................................
---------------------------------------------------------------
/Users/ahmed/MATLAB/HKUST/MPF-BML-master/chess_data_variable...
s_76_output/...................................................
---------------------------------------------------------------
Step 1: Mutant combining.......................................
---------------------------------------------------------------
Inferred phi_opt = 0.80........................................
---------------------------------------------------------------
Step 2: MPF....................................................
---------------------------------------------------------------
Iter  fEvals    stepLen       fVal    optCond     nnz..........
   1       2  3.566e-06  7.811e+04  1.140e+03    1369..........
  10      11  1.000e+00  5.509e+04  3.359e+02    1369..........
  20      21  1.000e+00  4.643e+04  1.920e+02    1369..........
  30      31  1.000e+00  4.451e+04  1.191e+02    1369..........
Terminated: Directional derivative below Func Tol..............
Saving inferred model parameters "J_MPF".......................
Plotting MPF-based model verification results..................
---------------------------------------------------------------
Step 3: BML....................................................
---------------------------------------------------------------
Iter   Single ϵ   Double ϵ    Conn. ϵ..........................
 915  1.8417303  1.6253795  2.4239354..........................
 920  1.7629467  1.6022281  2.2166428..........................
 925  1.3831257  1.7540000  2.4908439..........................
 930  1.3722866  1.5095668  2.3215161..........................
 931  1.2271594  1.4934012  2.5920903..........................
Terminated: Average ϵ < 1.50 ..................................
Plotting MPF-BML-based model verification results..............
Saving inferred model paramters "J_MPF_BML"....................
---------------------------------------------------------------
```

Model verification

Model: MPF
Model: MPF-BML
Individual frequencies, ϵ = 7.78
Individual frequencies, ϵ = 1.80
Model
MSA
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Pair-wise frequencies, ϵ = 3.79
Pair-wise frequencies, ϵ = 1.69
Connected corrs., ϵ = 6.11
Connected corrs., ϵ = 2.15
-0.2
-0.1
r = 0.954, p = 1.05e-08
r = 0.976, p = 1.05e-10
P(k)
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
MSA
Model
5
10
15
20
No. of differences with the consensus config., k

Figure S6. Model inferred using the MPF-BML package for the classical chess data set obtained from the UCI repository (Asuncion and Newman, 2007).

Supplementary References

Asuncion,A. and Newman,D.J. (2007) UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. *Univ. California, Irvine, Sch. Inf. Comput. Sci.*

Bamford,S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355–358.

Barton,J.P. *et al.* (2016) ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics*, **32**, 3089–3097.

Cocco,S. *et al.* (2018) Inverse statistical physics of protein sequences: a key issues review. *Reports Prog. Phys.*, **81**, 032601.

Louie,R.H.Y. *et al.* (2018) Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci.*, **115**, E564–E573.

Riedmiller,M. and Braun,H. (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *IEEE Int. Conf. Neural Networks*, **1**, 586–591.

Wolberg,W.H. and Mangasarian,O.L. (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Natl. Acad. Sci.*, **87**, 9193–9196.