Electronic Supplementary Information:

Protein-ligand docking using fitness learning-based artificial bee colony with proximity stimuli

Shota Uehara, Kazuhiro J. Fujimoto*, Shigenori Tanaka**

Department of Computational Science, Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Hyogo 657-8501, Japan

*E-mail: uehara@eniac.scitec.kobe-u.ac.jp
*E-mail: fujimoto@ruby.kobe-u.ac.jp
*E-mail: tanaka2@kobe-u.ac.jp

Tel: +81 078 803 6620, Fax: +81 078 803 6621

Fitness learning-based artificial bee colony with proximity stimuli

Fitness learning-based artificial bee colony with proximity stimuli (F*l*ABCps) proceeds in the same way as the classical ABC through the employed bee phase, the onlooker bee phase and the scout bee phase. The position of a food source represents a solution vector to the optimization problem, and the quality of a food source (nectar amount) corresponds to the fitness value calculated with the scoring function. The number of food sources *SN* is equal to the number of the employed bees or the onlooker bees. The three kinds of bees search for a global optimum point in *D*-dimensional real parameter space, where *D* corresponds to the number of optimization parameters. Each *D*-dimensional solution vector at the *SN* food sources is described as

$$\boldsymbol{\theta}_i^C = \left[\theta_{i,1}^C, \theta_{i,2}^C, \theta_{i,3}^C, \cdots, \theta_{i,D}^C\right], \tag{S1}$$

where i=1,2,…,*SN* is the index of food sources and C=0,1,…,*MCN* (maximum count number) is the current cycle number. In the beginning of optimization (C=0), each parameter of food sources is initialized with uniformly distributed random numbers which are restricted to certain ranges. After the initialization, the following procedures are repeated in each cycle until the termination criteria is satisfied.

First, fitness value of all the food sources is calculated as

$$fitness_i = \begin{cases} 1/(1+f_i) & \text{if } f_i \geq 0 \\ 1+\mathrm{abs}(f_i) & \text{if } f_i < 0\,, \end{cases} \tag{S2}$$

where f_i is an actual value of scoring function to be optimized ($f_i = F(\boldsymbol{\theta}_i^C)$). Since we consider a minimization condition, a food source with the lower score of function has a higher fitness value. Fitness learning mechanism is described with mixing with the elite components

$$\theta_{i,j}^{FC} = \begin{cases} \theta_{r1,j}^C & \text{if } fitness_{r1} \geq fitness_{r2} \\ \theta_{r2,j}^C & \text{if } fitness_{r1} < fitness_{r2}\,, \end{cases} \tag{S3}$$

where j=1,2,…,D is D-dimensional parameter index of a food source. *r1* and *r2* are different indices of the elite food sources randomly selected from the top q% of the population. The value of q is varied from the top 20% (0.2) members initially, to the 10% (0.1) at the end of the cycle. This variation of q that occurs nonlinearly is given by

$$q = 0.2 - 0.1\left(\frac{e^{m \cdot C/MCN} - 1}{e^{m} - 1}\right), \tag{S4}$$

with an uniform random number m, lying in the range [0,1]. In addition, F*l*ABCps uses a selective parameter scheme for multi-dimensional perturbation, based on the Rechenberg's 1/5th mutation rule (see *ref.*24 in the main text). The perturbation parameters for the positional modification are selected by

$$J^* = \{j_1, j_2, j_3, \cdots, j_n\}; \;\; j_v \epsilon \{1,2,\cdots,D\}; \;\; 1 \le n \le \left[\frac{1}{5}D\right]. \tag{S5}$$

Here, n is a random integer corresponding to the number of components of J^*. It is noted that J^* is a subset of D-dimensional parameters which are composed of randomly selected n indices.

In the employed bee phase, the employed bee seeks a new food source $\boldsymbol{v}_i^C$ around the assigned food source $\boldsymbol{\theta}_i^C$ by using the perturbation parameters J^*. The positional modification scheme in F*l*ABCps is performed with a combination of the directive component towards adjacent food sites and the fitness learning mechanism of $\boldsymbol{\theta}_i^{FC}$ (eqn S3)

$$v_{i,J^*}^C = \theta_{i,J^*}^C + \phi_G\left(\theta_{k_i,J^*}^{NC} - \theta_{i,J^*}^C\right) + \phi_C\left(\theta_{i,J^*}^{FC} - \theta_{i,J^*}^C\right), \tag{S6}$$

where $\boldsymbol{\theta}_{k_i}^{NC}$ represents one of the kth-nearest food sources from $\boldsymbol{\theta}_i^C$ according to the Euclidean distance. The parameter k is a random integer, lying in the range $[0, \sqrt{SN/2}]$. Two control parameters, ϕ_G and ϕ_C, are different random numbers, generated as

$$\begin{aligned} \phi_G &= N(\mu, \sigma^2); \qquad \mu = 0, \sigma^2 = 1, \\ \phi_C &= Q(r; x_0, \gamma); \qquad x_0 = 0, \gamma = 0.5, r \in (0,1), \end{aligned} \tag{S7}$$

where $N(\mu, \sigma^2)$ denotes the Gaussian distributed number with mean μ and variance σ^2; $Q(r; x_0, \gamma)$ denotes the quantile function of Cauchy distribution with location x_0, scale γ and restrict range r. The Gaussian distribution has a short tail property, and is suitable for the fine local search. On the other hand, the Cauchy distribution has a far wider tail than the Gaussian distribution, and is useful when the global optimum is far away from the current search point. If a new food source $\boldsymbol{v}_i^C$ has the higher fitness value than the current food source $\boldsymbol{\theta}_i^C$, the employed bee updates $\boldsymbol{\theta}_i^C$ to $\boldsymbol{v}_i^C$.

In the onlooker bee phase, the onlooker bee performs a probabilistic selection of the food source for exploitation. In the classical ABC, a probability of a food source to be selected is calculated with the fitness values, given by

$$p_i = \frac{fitness_i}{\sum_{l=1}^{SN} fitness_l}. \tag{S8}$$

The selection scheme using eqn S8 sometimes causes the overcrowding of the onlooker bees at the best-so-far food source, which results in the premature convergence. To circumvent this problem, F*l*ABCps introduces a weighted probability based on the proximity-based stimuli

$$p_i^w = \frac{1}{2m_i} \sum_{l=1}^{m_i} \left(p\left(N_l^i\right) + p(F_l) \right), \tag{S9}$$

where $p(\cdot)$ represents the probability (eqn S8) of a selected food source taken as an argument $(p(i) = p_i)$. N_l^i is an index representing the *l*th-nearest food sources calculated with the Euclidean distance from the *i*th food source. Similarly, F_l is an index which refers to the *l*th-best food source calculated with the fitness value. The parameter m_i is a random integer, lying in the range $[0, SN/\sqrt{D}]$. If the weighted probability p_i^w is larger than p_i, the *i*th food source is selected by an onlooker bee for exploitation. The onlooker bee searches for a new food source $\boldsymbol{v}_i^C$ around the selected food source $\boldsymbol{\theta}_i^C$ using eqn S6, and updates $\boldsymbol{\theta}_i^C$ to $\boldsymbol{v}_i^C$ with the greedy selection in the same way as the employed bee. This selection is repeated until all the onlooker bees are assigned to any of the food sources.

In the scout bee phase, the food source that cannot be improved anymore is replaced randomly by a scout bee. To find these exhausted food sources, a trial counter t_i is used at each *i*th food source. If an employed or onlooker bee is unable to improve the previous fitness value of the *i*th food source, t_i is increased by unity. The trial counter t_i is reset to zero when the *i*th food source is successfully improved. When t_i reaches the maximum trial number *limit*, the *i*th food source is replaced with random numbers and t_i is reset to zero.

Details of the docking simulations

A complete pseudo-code of F*l*ABCps is described in Table S1. Setting parameters of the five algorithms for docking experiments are summarized in Table S2. Docking results of five algorithms for 85 complexes in Astex diverse set are given in Table S3.

Table S1. Complete pseudo-code of F*l*ABCps

Initialization:

Set the parameters *SN*, *limit* and *MCN*.
Generate initial food sources $\{\boldsymbol{\theta}^0\}$ with random numbers.
Evaluate the fitness value of all food sources according to the eqn S2.
Set the value of trial counter t_i to 0 for all food sources.

While *C*<*MCN*

Increment the current cycle number *C* by 1.
Update the value of *q* by eqn S4.

//Calculate fitness learning sources.
Sort the food sources $\{\boldsymbol{\theta}^{\mathrm{C}}\}$ order in the fitness values in descending.
For *i*=1 to *SN*
Initialize the fitness learning source $\boldsymbol{\theta}_{\mathrm{i}}^{\mathrm{F}}$.
For *j=1 to D*
Chose different indicators *r1* and *r2* randomly in the top range [1,ceil(*q*/100)*SN*].
Set the value of $\theta_{\mathrm{i,j}}^{\mathrm{F}}$ by using eqn S3.
End for
End for

//Employed bee phase
For *i=1* to *SN*
Simulate the perturbation parameters J^* by eqn S5.
Obtain the k-nearest neighbor $\boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{NC}}$ with respect to food source $\boldsymbol{\theta}_{\mathrm{i}}^{\mathrm{C}}$.
Search a new food source $\mathbf{v}_{\mathrm{i}}^{\mathrm{C}}$ by using eqn S6.
Evaluate the fitness value of $\mathbf{v}_{\mathrm{i}}^{\mathrm{C}}$ according to eqn S2.
Perform the greedy selection between $\boldsymbol{\theta}_{\mathrm{i}}^{\mathrm{C}}$ and $\mathbf{v}_{\mathrm{i}}^{\mathrm{C}}$ according to their fitness values.
Update the trial counter t_i.
End for

//Onlooker bee phase
Obtain *p* according to eqn S8.
For *i*=1 to *SN*
Obtain p_i^w by eqn S9 using $\mathbf{N}^{\mathrm{i}}$ *and* **F**.
End for
Set *i=1*and *l=0.* *//Then l corresponds to the index of onlooker bees.*
While *l*<*SN*
If $p_i < p_i^w$
Increment the value of *l* by 1.
Simulate the perturbation parameters J^* by eqn S5.
Obtain the k-nearest neighbor $\boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{NC}}$ with respect to food source $\boldsymbol{\theta}_{\mathrm{i}}^{\mathrm{C}}$.
Search a new food source $\mathbf{v}_{\mathrm{i}}^{\mathrm{C}}$ by using eqn S6.
Evaluate the fitness value of $\mathbf{v}_{\mathrm{i}}^{\mathrm{C}}$ according to eqn S2.
Perform the greedy selection between $\boldsymbol{\theta}_{\mathrm{i}}^{\mathrm{C}}$ and $\mathbf{v}_{\mathrm{i}}^{\mathrm{C}}$ according to their fitness values.
Update the trial counter t_i.
End if
Increment the value of *i* by 1.
If *i*>*SN* *//Food site selection is repeated till all onlooker bees have been allocated.*
i=1.
End if
End while

//Scout bee phase
For *i*=1 to *SN*
If $t_i \geq limit$
Reinitialize the food source $\boldsymbol{\theta}_{\mathrm{i}}^{\mathrm{C}}$ with random number.
Evaluate the fitness value of $\boldsymbol{\theta}_{\mathrm{i}}^{\mathrm{C}}$ according to eqn S2.
Reset the trial counter t_i to 0.
End if
End for

End while

Table S2. Setting parameters of the five algorithms for docking experiments

FlABCps	
Number of food soures, *SN*	500
Maximum trial number, *limit*	200
ABC	
Number of food sources, *SN*	500
Maximum trial number, *limit*	200
SODOCK	
Number of particles, *Np*	500
Number of immediate neighbors, *K*	4
Inertia weight, *w*	0.9~0.4 (liner decreasing)
Cognitive weight, *c1*	2.0
Social weight, *c2*	2.0
Maximal velocity, *Vmax*	2.0 Å (for translation)
	1.0, 180 deg (for orientation)
	50 deg (for conformation)
Maxmal steps of local search	50
PSO	
Number of particles, *Np*	150
Inertia weight, *w*	0.9~0.4 (liner decreasing)
Cognitive weight, *c1*	2.0
Social weight, *c2*	2.0
Maximal velocity, *Vmax*	2.0 Å (for translation)
	1.0, 180 deg (for orientation)
	50 deg (for conformation)
LGA	
ga_pop_size	150
ga_elitism	1
ga_mutation_rate	0.02
ga_crossover_rate	0.8
ga_window_size	10
ga_cauchy_alpha	0.0
ga_cauchy_beta	1.0
sw_max_its	300
sw_max_succ	4
sw_max_fail	4
sw_rho	1.0
sw_lb_rho	0.01
ls_search_freq	0.06

Table S3. Docking results of five algorithms for 85 complexes in Astex diverse set

PDB	N_r [a]	Root mean square deviation [Å] [b]					Value of scoring function [kcal/mol] [c]				
		FlABCps	ABC	SODOCK	PSO	LGA	FlABCps	ABC	SODOCK	PSO	LGA
1g9v	7	3.80	3.62	7.08	6.78	6.61	-10.75	-10.72	-10.50	-10.84	-10.77
1gkc	13	1.07	0.95	2.70	0.76	0.91	-12.70	-12.87	-12.24	-13.01	-11.67
1gm8	6	2.74	2.83	2.83	3.04	7.87	-12.25	-11.86	-12.00	-11.04	-12.41
1gpk	1	2.69	2.65	2.12	3.14	2.06	-9.54	-9.50	-8.13	-7.28	-7.27
1hnn	2	1.09	1.10	1.07	0.90	1.00	-11.98	-11.92	-10.06	-10.02	-9.93
1hp0	6	0.48	1.68	1.68	1.58	8.00	-7.34	-7.34	-6.74	-6.15	-7.08
1hq2	3	0.45	0.42	0.36	0.27	0.37	-10.17	-10.17	-10.03	-10.06	-10.05
1hvy	10	1.00	1.40	1.29	0.95	1.51	-14.58	-14.79	-13.54	-16.78	-15.27
1hwi	11	3.75	0.53	0.60	0.58	0.82	-16.19	-16.00	-15.10	-15.62	-15.39
1hww	3	0.22	0.22	0.22	0.26	0.43	-10.55	-10.55	-10.48	-9.91	-9.89
1ia1	4	0.70	0.61	0.71	0.53	0.50	-8.42	-8.40	-7.97	-7.83	-7.80
1ig3	6	0.83	0.79	0.87	0.80	2.80	-10.51	-10.46	-9.84	-8.67	-9.60
1j3j	4	0.62	0.60	0.58	0.53	0.53	-7.63	-7.62	-7.19	-7.22	-7.22
1jd0	4	4.42	4.47	4.38	4.23	4.60	-8.50	-8.40	-7.36	-7.01	-6.99
1jje	7	1.48	1.20	1.30	2.60	1.85	-28.76	-27.27	-24.93	-24.24	-23.41
1jla	7	0.93	0.80	0.83	0.84	0.80	-15.09	-15.12	-14.91	-14.82	-14.77
1k3u	7	0.56	0.53	0.74	0.56	1.36	-11.64	-11.39	-10.30	-8.20	-7.13
1ke5	5	0.78	0.81	0.82	0.54	0.61	-13.11	-13.09	-12.53	-12.41	-12.23
1kzk	14	0.93	0.88	1.19	0.80	0.82	-21.07	-21.08	-20.06	-20.84	-17.06
1l2s	4	0.93	0.90	0.68	0.92	0.94	-12.41	-12.17	-10.65	-10.49	-10.48
1l7f	12	0.83	0.89	0.74	0.48	2.88	-16.46	-16.56	-14.38	-10.34	-13.59
1lpz	9	0.56	0.28	0.66	0.57	1.00	-15.41	-14.90	-14.91	-14.14	-13.53
1lrh	2	0.65	0.96	0.59	0.79	0.75	-13.00	-12.78	-11.75	-10.83	-10.78
1m2z	7	0.55	0.55	0.47	0.51	0.79	-18.12	-18.11	-16.94	-14.88	-15.04
1meh	8	1.46	1.53	1.08	7.22	5.94	-8.34	-8.14	-8.13	-7.14	-7.23
1mmv	10	0.66	0.77	0.54	0.47	2.81	-14.55	-14.42	-13.64	-13.38	-13.56
1mzc	7	1.02	0.89	0.86	1.07	1.22	-17.39	-15.84	-16.35	-15.93	-15.89
1n1m	4	0.87	0.91	0.98	0.80	0.59	-10.84	-10.70	-9.25	-9.24	-9.20
1n2j	5	7.25	8.70	6.86	8.22	6.38	-12.98	-12.28	-11.25	-8.20	-7.98
1n2v	3	3.29	3.57	3.13	2.01	2.61	-7.87	-7.71	-7.63	-7.39	-7.37
1n46	5	0.82	0.82	0.59	0.64	0.53	-15.89	-15.84	-15.80	-15.50	-15.48
1nav	6	0.92	0.54	0.84	0.42	0.51	-14.22	-14.13	-14.02	-13.21	-13.15
1of1	4	0.44	0.48	0.46	0.45	1.70	-10.53	-10.53	-10.33	-9.07	-10.26
1of6	5	0.76	0.86	0.70	0.64	0.60	-16.16	-15.40	-10.43	-10.06	-10.02
1opk	5	0.61	0.61	0.65	0.86	0.82	-14.84	-14.84	-14.76	-14.46	-14.49
1oq5	5	0.76	0.76	0.64	0.55	3.18	-11.30	-11.22	-10.06	-9.87	-10.06
1owe	6	1.87	1.83	2.01	1.96	2.24	-10.20	-9.68	-9.88	-10.59	-10.21
1oyt	6	0.67	0.65	0.59	0.55	0.50	-9.93	-9.89	-9.08	-8.21	-8.38
1p2y	1	1.84	1.84	1.84	1.83	1.54	-6.73	-6.73	-6.73	-6.73	-6.72
1p62	5	0.62	0.60	0.59	0.51	1.00	-10.49	-10.49	-10.20	-10.36	-10.35
1pmn	7	0.85	0.78	0.78	0.75	4.03	-13.92	-13.88	-13.93	-11.90	-13.60
1q1g	5	1.03	1.01	0.98	0.52	1.10	-13.72	-13.71	-13.56	-13.10	-13.00
1q41	3	0.50	0.48	0.48	0.49	0.68	-10.98	-10.98	-10.97	-9.30	-9.30
1q4g	3	1.05	1.12	0.94	0.98	0.75	-12.59	-12.58	-11.48	-10.88	-10.88
1r1h	12	1.33	0.87	1.21	0.69	8.19	-21.51	-21.94	-21.42	-17.79	-16.63
1r55	12	0.69	0.84	0.96	0.68	0.72	-15.19	-14.54	-13.60	-13.59	-13.00
1r58	12	5.64	5.57	6.99	7.08	7.83	-12.16	-12.03	-11.71	-12.33	-11.22
1r9o	3	3.63	4.39	3.38	6.07	6.12	-16.49	-11.16	-11.86	-11.79	-11.83
1s19	11	0.42	0.73	0.73	0.85	0.54	-18.40	-18.39	-17.69	-17.70	-17.62

1s3v	8	0.54	0.58	0.62	0.60	0.56	-13.77	-13.74	-12.87	-12.76	-12.74
1sg0	6	1.05	0.84	0.84	0.84	1.95	-7.19	-7.18	-7.09	-7.00	-7.09
1sj0	8	0.66	0.73	0.70	0.82	2.60	-19.45	-19.43	-19.21	-17.39	-19.18
1sq5	9	2.27	2.39	2.45	2.60	3.48	-12.18	-12.09	-12.46	-9.92	-9.40
1sqn	1	1.23	1.23	1.23	1.21	1.33	-12.82	-12.82	-12.82	-11.49	-11.49
1t40	7	0.44	0.34	0.44	0.27	6.53	-14.19	-14.17	-14.10	-11.39	-12.41
1t46	8	0.47	0.46	0.46	0.59	0.71	-18.60	-18.60	-18.51	-18.21	-18.12
1t9b	6	3.26	1.28	4.90	3.46	3.60	-6.08	-6.06	-5.94	-5.88	-5.70
1tow	4	1.13	1.14	1.13	1.17	0.89	-11.45	-11.46	-10.81	-9.51	-9.53
1tt1	4	1.00	1.07	1.11	0.41	0.25	-15.75	-15.83	-13.42	-13.09	-13.06
1tz8	7	3.51	3.49	3.43	0.75	0.73	-7.32	-7.34	-6.86	-6.91	-6.87
1u1c	7	0.94	0.87	0.93	0.91	1.21	-7.60	-7.57	-7.52	-7.01	-7.54
1u4d	2	0.83	0.82	0.82	0.76	1.20	-8.68	-8.68	-8.31	-7.49	-7.49
1uml	13	1.19	1.42	0.89	1.52	3.30	-14.82	-15.04	-15.23	-14.15	-13.91
1unl	9	0.82	1.01	0.87	0.77	3.96	-13.02	-12.91	-12.60	-11.81	-11.80
1uou	3	0.76	0.79	0.71	0.75	0.90	-9.87	-9.67	-8.99	-8.85	-8.84
1v0p	9	1.16	0.89	5.37	4.68	1.17	-13.21	-13.27	-13.85	-13.06	-12.94
1v48	10	4.17	0.57	0.53	0.58	3.82	-7.87	-7.78	-7.67	-7.01	-7.49
1v4s	6	1.61	1.38	1.60	0.53	0.48	-6.42	-6.43	-5.96	-6.57	-6.69
1vcj	9	1.10	0.83	0.85	0.57	3.38	-18.37	-17.98	-16.00	-10.99	-14.52
1w1p	0	0.19	0.19	0.19	0.20	0.19	-5.89	-5.89	-5.90	-5.55	-5.55
1w2g	4	0.87	0.97	0.65	1.88	6.78	-6.25	-6.19	-5.89	-6.12	-6.29
1x8x	5	0.89	0.73	0.67	0.98	0.81	-16.23	-15.29	-10.54	-10.79	-10.67
1xm6	5	1.45	1.59	1.52	1.41	2.47	-10.18	-10.17	-10.14	-10.05	-10.02
1xoq	8	3.16	3.21	3.13	3.00	3.01	-11.72	-11.32	-11.55	-11.24	-11.09
1xoz	1	0.43	0.43	0.43	0.42	0.44	-13.13	-13.13	-13.12	-11.99	-11.99
1y6b	9	0.71	0.76	0.69	0.59	0.86	-10.51	-10.48	-10.52	-10.35	-10.16
1ygc	16	0.60	0.68	0.61	0.60	1.34	-17.01	-17.09	-16.83	-16.93	-15.12
1yqy	7	0.84	0.73	0.84	0.83	1.82	-13.58	-13.59	-13.60	-12.14	-12.91
1yv3	2	0.49	0.49	0.50	0.50	0.54	-8.52	-8.52	-8.51	-7.54	-7.54
1yvf	8	0.71	1.02	0.86	0.65	4.55	-14.11	-13.80	-12.95	-13.18	-12.58
1ywr	6	0.77	0.77	0.77	0.86	0.75	-16.02	-15.91	-15.71	-15.51	-15.46
1z95	8	0.61	0.67	0.71	0.67	0.78	-16.87	-16.08	-16.01	-15.67	-15.34
2bm2	8	0.56	0.58	0.50	1.55	5.46	-14.56	-14.57	-14.58	-13.15	-14.11
2br1	8	1.53	1.08	3.06	4.80	3.13	-11.77	-11.47	-11.25	-11.46	-11.38
2bsm	8	0.84	0.71	0.79	0.58	4.20	-13.44	-13.38	-13.08	-11.15	-13.15

[a]N_r represents the number of rotational bonds for ligands. [b]The root mean square deviation (RMSD) between the docking ligand and the crystal ligand. [c]The energy score of AutoDock energy function 4.2 obtained from the docking calculations.