Parameter Estimation via Analysis of Fuzzy Clusters (PEAF): An Algorithm to Estimate Parameters of Agent-Based Models

Shahab Sheikh-Bahaei[1]

[1]Joint Graduate Group in Bioengineering
University of California
Berkeley and San Francisco, CA 94143-0446.

C. Anthony Hunt[1,2]

[2]Department of Biopharmaceutical Sciences
University of California
San Francisco, CA 94143-0446

Abstract

Biologically focused, agent-based models need many parameters in order to simulate system dynamics. It is often essential to explore the consequences of many parameter vectors before satisfactorily representing phenomena. In this work we propose a simple algorithm based on fuzzy clustering to *estimate* model parameter values for new situations utilizing the characteristics of previously simulated conditions. The estimated parameters can be used to *predict* the behavior of the system in a new situation. Using limited data, we successfully applied the algorithm to estimate parameter values of an agent-based model of hepatocytes (liver cells). Predictions provide acceptable correlations with observed values ($p < 0.05$, $R^2 = 0.65$).

Keywords: parameter estimation, prediction, Agent-Based modeling, fuzzy clustering

1 INTRODUCTION

Agent-based modeling is being used in a variety of fields: examples include social sciences [1-5], supply chain optimization and logistics; modeling of consumer behavior; distributed computing; workforce management; traffic management; portfolio management; complex systems, artificial life, genetic programming and genetic evolution [6–11]; bacterial chemotaxis signaling pathways [11,12]; population ecology [13–15]; social and economic systems [16–18]; and cellular behavior [19-23].

Agent-based models commonly require many parameters. Together, they determine the global dynamics of the system. Small changes made to one parameter can lead to an important change of the dynamics of the entire system. Consequently, identifying informative and plausibly realistic regions of parameter space for exploration can be time-intensive [24]. Several automated techniques have been used, including the Nelder and Mead Simplex Method [25] and Genetic Algorithms [24]. Once parameter vectors have been identified that are suitable for several situations, one can become interested in predicting system behavior for a new situation. In this paper we propose a method to estimate such parameters based on previously seen cases in order to predict system behavior for a new situation. The proposed method uses the Fuzzy-c-Means [26] classification algorithm. As a proof of concept we apply the method to an agent-based model of hepatocytes, and make predictions.

2 METHODS

2.1 Parameters of Agent-Based Models

Parameters in biologically focused, agent-based simulation models can be of different natures. Some map directly to real-world, measurable counterparts and some are simulation-specific with no direct real-life counterpart. Some of the former can be extracted from domain-specific knowledge (either experimental or theoretical). Others are design-specific.

A model's behavior space is expected to overlap somewhat with the behavior space of the referent system. Achieving that requires that model parameters be appropriately tuned (adjusted) to represent desired real-life *situations*. Each real-life *situation* has measurable properties (phenotypic attributes), which define its unique characteristics (phenotype). Each simulated *situation* is similarly characterized by its unique simulation parameters.

In real-life *situations*, a causal relationship exists between generative mechanisms and measured properties. A similar mapping exists for hepatocytes simulations. We follow an axiom that in many cases a mapping exists between the space of selected, measured properties and the space of simulation parameter values. Figure 1 illustrates this axiom: three different real-life *situations* are shown, two of which are closer together in the space of measurable properties. The arrangement of *simulated* phenomena relative to the arrangement of simulation parameters may differ from that of real life, even though their relative distances to each other are more or less similar. Nevertheless, as a first approximation, we assume that the relationship between a new phenomenon and its acceptable simulation parameters can be approximated

from its position relative to acceptable, previously simulated *situations*.

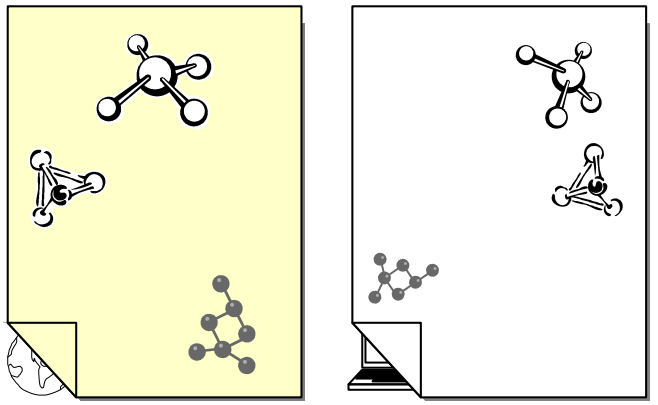


Figure 1. System *situations* with related, measurable properties and generating mechanisms are expected to have similarly related, in silico properties and generating mechanisms.

2.2 The PEAF Algorithm

In this section we present an algorithm that uses fuzzy clustering to estimate simulation parameters for a new *situation* given the tuned parameters of several, previously validated *situations*. A fuzzy classifier provides a measure of the degree to which a pattern fits within a class. There are several techniques for fuzzy pattern recognition. In this work, we use a fuzzy pattern recognition technique introduced in [26]: the Fuzzy c-Means iterative algorithm.

The inputs of the Fuzzy c-Means algorithm are: 1) the set of n data points to be clustered, 2) number of clusters c, and 3) a parameter m known as the fuzzy exponent. As recommended by [27] we always set m = 2. The output of Fuzzy c-Means algorithm, U, is a c-by-n matrix, containing the values of the membership functions of the fuzzy clusters.

In general, for a data set, *S,* containing *n situations* $S = \{c_1, c_2,\ldots, c_n\}$, the following PEAF algorithm is proposed to estimate the simulation parameters of a new *situation*, c_{n+1}:

Step 1. Let $q = n$, and $S_{new} = \{c_1, c_2 \ldots c_n, c_{n+1}\}$.

Step 2. If $q = 1$ go to step 4. Else, classify S_{new} into q clusters using Fuzzy c-Means algorithm.

Step 3. If c_{n+1} is not in the same group with *at least* another member then decrease q to *q-1*. Repeat steps 2 and 3.

Else, let *G-value* be the number of groupmates of c_{n+1}. Go to step 4.

Step 4. Call the q groups $G_1, G_2 \ldots G_q$ where $c_{n+1} \in G_1$. Let μ_k be the membership degree of c_{n+1} to G_k. Estimate the simulation parameters of c_{n+1} as:

$$\hat{\vec{P}}_X = \sum_{k=1}^{q} \mu_k \cdot \vec{P}_{Gk} \quad (2)$$

where P_{Gk} is the weighted average parameter vector of all the members of group k:

$$\vec{P}_{Gk} = \frac{\sum_{j=1}^{m} \mu_j \cdot \vec{P}_j}{\sum_{j=1}^{m} \mu_j} \quad (3)$$

The accuracy or usefulness of the resulting estimates depends, of course, on how many *situations* similar to c_{n+1} exist in the data set, i.e. the higher the *G-value*, the better the accuracy.

2.3 Estimating the Parameters of an Agent-Based Model

In this section we show how the PEAF algorithm can be used to estimate the simulation parameters of an agent-based model in order to make predictions. In this model, a *situation* is characterized by hepatocytes behavior in the presence of a particular compound. In Silico Hepatocyte (ISH) is an agent-based model of hepatocytes [28]. The cells are simulated on a 2D grid; it represents the culture dish. When hepatocytes are exposed to different simulated drug compounds, they metabolize and eliminate them, as in vivo. Consequently, simulation parameter values that are sensitive to physicochemical properties (PCPs) need to be different for each drug. The goal is to estimate the PCP-sensitive parameter values to enable simulating the metabolic and transport properties of a new drug given the parameter values similarly used and validated for several previously studied drugs.

To demonstrate, consider the four compounds shown in Table 1. The following PCPs were considered: molecular weight, logP, hydrogen bond donor count, hydrogen bond acceptor count, rotatable bond count, tautomer count, pKa, TPSA, volume, GPCR ligand, ion channel modulator, kinase inhibitor, and nuclear receptor ligand. The classification results for the PCPs of the four compounds clustered to two and three classes using the Fuzzy c-Means algorithm based on their PCPs are shown in Table 2. The classification results show that when divided into two groups, taurocholate, enkephalin, and methotrexate have more membership in the same group while salicylate belongs primarily to another. However, when divided to three groups, taurocholate and methotrexate have membership in the same group, whereas enkephalin and salicylate belong primarily to different groups.

Table 1. Physicochemical Properties of Salicylate, Taurocholate, Methotrexate and Enkephalin.

Property[1]	Sal.	Taur.	Meth.	Enkeph.
MW[2]	140.1	515.7	454.4	645.8
logP	2.24	0.01	-1.28	2.01
HBD[2] count	2	5	5	7
HBA[2] count	3	7	12	8
RB[2] count	1	7	9	7
Tautomer count	4	2	24	32
pKa	2.97	1.8	4.7	10
TPSA[2]	57.5	144.1	210.6	199.9
Volume	119.1	483.1	387.4	569.7
GPCR[2] ligand	-0.44	-0.26	0.22	-0.19
IC[2] modulator	-0.08	-0.15	0.02	-1.05
Kinase inhibitor	-0.65	-0.47	0.11	-0.84
NR[2] ligand	-0.58	-0.08	-0.36	-0.58

Table 2. Fuzzy Classification Results of Salicylate, Taurocholate, Enkephalin and Methotrexate Based on Their Physicochemical Properties (Table 1). C: number of clusters.

C	Group	Sal.	Taur.	Meth.	Enkeph.
3	1	**0.9981**	0.0997	0.0492	0.0096
	2	0.0011	**0.5492**	**0.8639**	0.0291
	3	0.0007	0.3511	0.0869	**0.9614**
2	1	**0.9862**	0.0736	0.2308	0.1529
	2	0.0138	**0.9264**	**0.7692**	**0.8471**

Consider this task: based on the information in Table 2, we want to estimate the PCP-sensitive parameter values of enkephalin given the corresponding parameters of the other three compounds. There was no point in clustering the four drugs to four clusters, so we started with three. When c = 3, the fuzzy c-means algorithm provides no useful information about similarity of enkephalin to others: no other compound was in the same group with enkephalin (it however tells us about the dissimilarity of enkephalin to others). Thus, we took an additional step and clustered the compounds to two groups. When c = 2, enkephalin has two other groupmates. In that case, the best guess is that the PCP-sensitive parameter values for enkephalin are closer to those of its groupmates, taurocholate and methotrexate, than to salicylate. An intuitive way to estimate a parameter vector value for enkephalin is:

$$\hat{P}_{Enkeph.} = 0.8471\frac{(0.9264P_{Taur.} + 0.7692P_{Meth.})}{(0.9264 + 0.7692)} + 0.1529(0.9862P_{Sal.})$$

where P_x is the simulation parameter vector of compound x. In this example the *G-value* is 2 because two compounds ended up within the cluster with enkephalin.

3 RESULTS

The PEAF algorithm was used to iteratively predict the *clearance*[3] of the 50 drugs (listed in Table 3) in a leave-one-out process. Figure 2 shows that relative prediction error decreases as the *G-value* increases: the predictions are more accurate for compounds with more members in their clusters. Figure 3 shows the distribution of the prediction error. Compounds with *G-value* greater than 1 are located close to zero.

4 CONCLUSION AND DISCUSSION

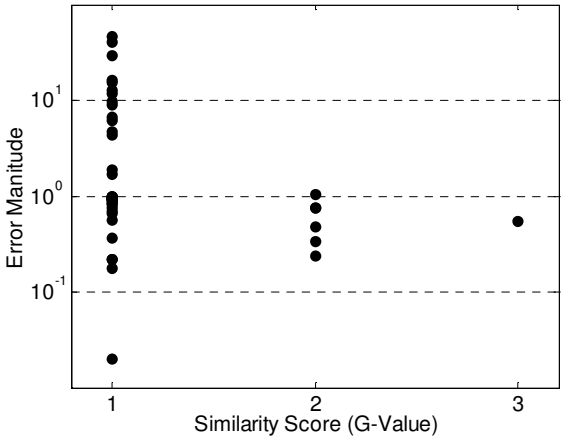
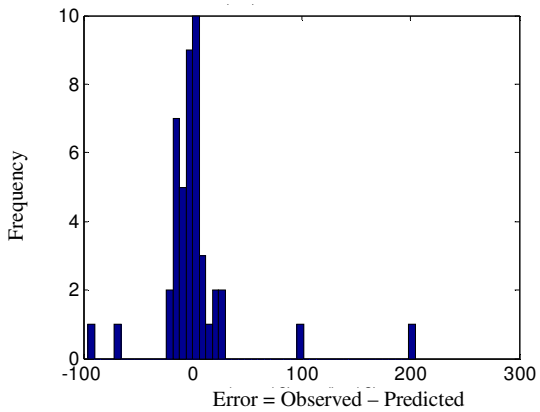


Figure 2. Prediction error decreases as *G-value* increases



We proposed a simple algorithm, called PEAF, that is based on fuzzy clustering to estimate the PCP-sensitive parameter values of agent-based models. When the

[1] Property values were obtained from the following sources. http://www.molinspiration.com/cgi-bin/properties; http://www.syrres.com/esc/est_kowdemo.htm; and http://ibmlc2.chem.uga.edu/sparc/index.cfm.

[2] MW: molecular weight; HBD: hydrogen bond donor; HBA: hydrogen bond acceptor, RB: rotatable bond, TPSA: topological polar surface area, GPCR: G-protein-coupled receptor, NR: nuclear receptor, IC: ion channel.

[3] Clearance is one of the most important pharmacokinetic parameters of a drug. It represents the speed of elimination of the drug from the body.

model's parameter values are properly tuned, it is capable of mimicking its referent. The PEAF algorithm utilizes the Fuzzy c-Means (FCM) algorithm to cluster previously encountered *situations* based on their measurable properties. The algorithm works based on the assumption that similarity in the space of measurable properties maps to the similarity in the parameter space of the simulation model. PEAF offers important advantages: 1) Because Fuzzy c-Means algorithm is unsupervised, the problem of over-fitting to the training data is minimized. That is particularly important in cases with small data sets. 2) The PEAF algorithm has no parameters. 3) It is relatively easy to implement. 4) The algorithm calculates a similarity score (the *G-value*) which correlates with the accuracy of its estimates: the higher the score, the higher the expected precision. As a result, the algorithm can advise in advance on the accuracy of its predictions.

As a proof of concept, we utilized the PEAF algorithm to estimate the PCP-sensitive parameter values of the ISH to *predict* the behavior of the referent system when it is introduced to a new compound not previously encountered. Note that parameter prediction is a direct mapping from the space of PCPs to the ISH parameter space, whereas parameter tuning draws its information from the biological behavior space. The estimated parameter values were fed to the ISH to enable it to make predictions. The predictions were compared to the observed measurements. For the seven of fifty compounds with G-values > 1 the predictions correlate nicely with the observed values ($p < 0.05$, $R^2 = 0.65$). We expect that as the number of drugs in the database increases, the probability that a compound similar to the new one of interest will exist in the database will increase; as a result better predictions can be anticipated.

ACKNOWLEDGMENTS

This research was funded in part by the CDH Research Foundation (of which CAH is a Trustee) and CAH. We thank Nasim Sassan for assisting with evaluation of *in vitro* systems and data. We thank Pearl Johnson, Glen Ropella and members of the BioSystems Group for helpful discussion and commentary.

REFERENCES

[1] Paul E.Johnson, "Simulation Modeling in Political Science," American Behavioral Scientist, vol. 42 (10), pp. 1509-1530, 1999.

[2] Michael W. Macy, "From Factors to Actors: Computational Sociology and Agent-Based Modeling," *Annual Review of Sociology, vol* 28, pp. 143-166, 2002.

[3] Lars-Erik Cederman, "Computational Models of Social Forms: Advancing Process Theory," American Journal of Sociology, vol 110 (4), pp 864-893, 2005.

[4] Thomas B. Pepensky,. From Agents to Outcomes: Simulation in International Relations. European Journal of International Relations, vol 11 (3), pp367-394. 2005

[5] Leigh Tesfatsion, and Kenneith L Judd.. forthcoming. *Handbook of Computational Economics. Vol. 2: Agent-Based Computational Economics.* Elsevier. 2006

[6] C.W. Reynolds, "Flocks, herds, and schools: A distributed behavioral model," Computer Graphics, vol 21 (4), pp 25–34, 1987.

[7] M. Scheutz, P. Schermerhorn, "Many is more but not too many: Dimensions of cooperation of agents with and without predictive capabilities," Proceedings of IEEE/WIC IAT-2003, IEEE Computer Society Press, 2003.

[8] V. Trianni, T. H. Labella, M. Dorigo, "Evolution of direct communication for a swarmbot performing hole avoidance," Proceedings of the 4th Intl. Workshop on Ant Colony Optimization and Swarm Intelligence, pp. 131–142, 2004

[9] P. Schermerhorn, M. Scheutz, "The effect of environmental structure on the utility of communication in hive-based swarms, in: IEEE Swarm Intelligence Symposium 2005, 2005.

[10] P. Schermerhorn, M. Scheutz, "The utility of heterogeneous swarms of simple uavs with limited sensory capacity in detection and tracking tasks, in: IEEE Swarm Intelligence Symposium 2005, 2005.

[11] S. S. Andrews, D. Bray, "Stochastic simulation of chemical reactions with spatial resolution and single molecule detail," Physical Biology, vol 1, pp 137-151.

[12] T. S. Shimizu, "The spatial organisation of cell signalling pathways - a computer-based study," Ph.D. thesis, University of Cambridge, 2002.

[13] V. Grimm, "Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future?" Ecological Modelling vol 115 (2-3), pp 129–148, 1999.

[14] S. F. Railsback, B. C. Harvey, R. H. Lamberson, D. E. Lee, N. J. Claasen, S. Yoshihara, "Population-level analysis and validation of an individual-based cutthroat trout model," Natuaral Resource Modeling, vol 15 (1), pp 83–110, 2002

Table 3. The Clearance values of 50 drugs [29] and their predicted values. Compounds with G-value>1 are shown in Bold.

	Drug Name	Clearance (μL/min/10^6 cells)		G-value
		Observed	Predicted	
1	Bromocriptine	37	7.039	1
2	Caffeine	3.3	103.8	1
3	Carbamazepine	2.0	99.02	1
4	Cimetidine	1.2 ± 0.4	0.124	1
5	Cyclosporin A	3.5 ± 1.5	18.36	1
6	Diazepam	0.3	13.46	1
7	**Ethinylestradiol**	**7 ± 2.0**	**9.814**	**2**
8	Famotidine	< 1	17.38	1
9	Isradipine	18	6.214	1
10	Lorazepam	1.0	17.37	1
11	Midazolam	14 ± 8.0	1.051	1
12	Nifedipine	5.6 ± 1.5	13.52	1
13	Nitrendipine	7.4 ± 3.5	5.226	1
14	Omeprazole	1.7	47.68	1
15	Prazosin	2.3 ± 1.7	15.49	1
16	Propofol	107 ± 26	9.189	1
17	Ritonavir	2.1 ± 3.0	1.890	1
18	Temazepam	2.0	0.043	1
19	Triazolam	1.0	14.18	1
20	Zileuton	2.1 ± 1.8	2.521	1
21	Acebutolol	1.8 ± 1.5	16.13	1
22	Atenolol	< 1	0.815	1
23	Bepridil	2.0	18.97	1
24	Betaxolol	2.5 ± 1.0	17.19	1
25	Bisoprolol	1.6 ± 1.4	14.98	1
26	Carvedilol	35 ± 11	9.964	1
27	**Chlorpheniramine**	**2.8 ± 1.3**	**5.333**	**2**
28	Clozapine	6.0	1.846	1
29	Codeine	23	1.878	1
30	Desipramine	3.0	7.335	1
31	Dextromethorphan	7.6 ± 8.1	7.371	1
32	Diltiazem	9.0 ± 0.5	0.066	1
33	**Diphenhydramine**	**6.0**	**10.25**	**2**
34	**Doxepin**	**13**	**6.090**	**2**
35	Fluoxetine	1.0	13.16	1
36	Granisetron	9.0 ± 8.7	2.146	1
37	**Imipramine**	**8.0 ± 2.5**	**10.33**	**2**
38	Metoprolol	7.0 ± 2.9	12.40	1
39	Morphine	24	0.354	1
40	Nadolol	< 1	21.78	1
41	Naloxone	216	12.35	1
42	Ondansetron	1.4 ± 0.5	1.202	1
43	**Pindolol**	**2.8 ± 1.0**	**0.662**	**2**
44	Pirenzepine	< 1	1.167	1
45	Propranolol	10 ± 0.5	7.073	1
46	Ranitidine	1.0 ± 0.0	0.700	1
47	Scopolamine	7.0	0.083	1
48	Triprolidine	4.3 ± 3.3	6.997	1
49	**Verapamil**	**18 ± 12**	**26.97**	**3**
50	Cetirizine	< 1	28.75	1

[15] M. Scheutz, P. Schermerhorn, "Predicting population dynamics and evolutionary trajectories based on performance evaluations in alife simulations," in: Proceedings of GECCO 2005, 2005.

[16] B. J. L. Berry, L. D. Kiel, E. Elliot, "Adaptive agents, intelligence, and emergent human organization: Capturing complexity through agent-based modeling," Proceedings of the National Academy of Science, vol 99, pp 7187–7188, 2002.

[17] R. Conte, "Agent-based modeling for understanding social intelligence," Proceedings of the National Academy of Science, vol 99, pp 7189–7190, 2002.

[18] P. Schermerhorn, M. Scheutz, "Implicit cooperation in conflict resolution for simple agents," Agent 2003, 2003.

[19] C.A. Hunt, G.E.P. Ropella, L. Yan, D.Y. Hung, and M.S. Roberts. "Physiologically Based Synthetic Models of Hepatic Disposition," J Pharmacokin Pharmacodyn, vol 33(6), pp 737-72, 2006.

[20] M.R. Grant, K.E. Mostov, T.D. Tlsty, and C.A. Hunt. "Simulating Properties of In Vitro Epithelial Cell Morphogenesis," PLoS Computational Biology, vol 2(10): e129, pp 1193-1209, 2006.

[21] Y. Liu and C.A. Hunt., "Mechanistic study of the interplay of intestinal transport and metabolism using the synthetic modeling method," Pharm Res., vol 23(3), pp 493-505, 2006.

[22] Y. Liu and C.A. Hunt. "Studies of Intestinal Drug Transport Using an In Silico Epithelio-Mimetic Device," Biosystems, vol 82(2), pp154-167, 2005.

[23] A. Qutub and C.A. Hunt. "Glucose transport to the brain: A systems model," Brain Research Rev, vol 49(3), pp 595-617, 2005.

[24] Benoit Calvez and Guillaume Hutzler , "Parameter Space Exploration of Agent-Based Models," Springer Berlin / Heidelberg, vol. 3684 , pp 633-639 , 2005.

[25] Nelder, J.A., and R. Mead., "A Simplex Method for Function Minimization," *Computer Journal*, vol 7, pp 308-313. 1965.

[26] Bezdek, J.C., R. Ehrlich and W. Full. "FCM: The fuzzy c-means clustering algorithm," *Computers and Geoscience*, vol 10, pp 191-203, 1984.

[27] Pal NR, JC Bezdek., "On cluster validity for the fuzzy c-means model," IEEE Trans Fuzzy Syst, vol 3(3), pp 370–9. 1995.

[28] S. Sheikh-Bahaei and C. A. Hunt. "Prediction of In Vitro Hepatic Biliary Excretion Using Stochastic Agent-Based Modeling and Fuzzy Clustering," In: L. F. Perrone, et al., eds., Proceedings of the 38th conference on Winter simulation conference, pp 1617 – 1624, 2006.

[29] DF McGinnity, MG Soars, RA Urbanowicz, and RJ Riley, "Evaluation of fresh and cryopreserved hepatocytes as in vitro drug metabolism tools for the prediction of metabolic clearance," Drug Metab Dispos, vol 32, pp 1247–1253, 2004.