2B4-03

トピックに依存した文書ランキング：文書引用ネットワーク分析

Topic-dependent document ranking: Bayesian approach to citation-network analysis

岡本洋[*1 *2] 坪下幸寛[*1] 園田隆志[*1]
Hiroshi Okamoto Yukihiro Tsuboshita Takashi Sonoda

[*1] 富士ゼロックス(株)研究技術開発本部
Research & Technology Group, Fuji Xerox Co., Ltd.

[*2] 理化学研究所 脳科学総合研究センター†
RIKEN Brain Science Institute

We propose a method of citation-network analysis for evaluating the topic-dependent importance of individual scientific papers. The method is based on the algorithm replicating the mechanism of memory retrieval in the brain. Information retrieval by this algorithm conforms to a type of Bayesian inference, which ensures its optimality as probabilistic inference.

1. Introduction

Citation of a scientific paper in another scientific paper denotes that research activity described in the latter is under the influence of that in the former (Fig. 1, left). Some papers are cited many times, which means that these papers have broad impact upon subsequent studies. Accordingly, the easiest way of evaluating the importance of a paper in terms of citation is to count how many times it is cited. The importance of a paper defined in this way is exactly proportional to the number of citation. The impact factor, a measure for the influence of a journal, is also calculated with the same idea [1].

Nevertheless, we believe that a citation in a more important paper is more valuable than that in a less important paper. Taking account of the value of each citation will therefore provide a more appropriate definition of the importance of individual papers. The most sophisticated method adopting such an idea is the PageRank algorithm used by the Google search engine [2, 3]. This algorithm assigns higher scores of importance to web pages that are linked from more numerous and more important pages.

The PageRank algorithm defines the importance of individual web pages only from the entire link structure of the World Wide Web. However we often consider the importance of scientific papers as what varies depending on the context or user's interest [2, 4]. Here we propose a novel method of citation analysis to evaluate the importance of individual papers in a topic-dependent manner. This method is based on the algorithm modelling the mechanism of memory retrieval in the brain, which turns out to be equivalent to Bayesian inference.

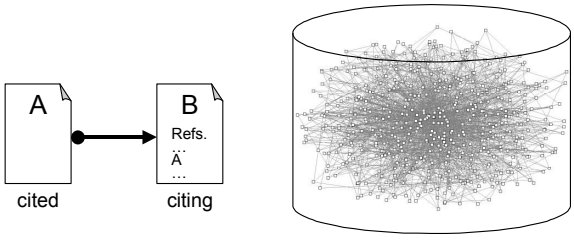

Fig. 1: Citation network

連絡先：岡本洋，富士ゼロックス(株)研究技術開発本部，
〒220-8668 神奈川県横浜市みなとみらい 6 丁目 1 番．
E-mail: hiroshi.okamoto@fujixerox.co.jp
† 客員研究員

2. Methods

2.1 Citation Network and Spreading Activation

Consider a large network consisting of papers as nodes and citation relations between papers as links (Fig. 1, right). Each node has an instantaneous value of 'activity'. Activities spread along links from nodes to nodes, which will be referred to as 'spreading activation' [5]. We define the importance of a paper by the value of activity finally acquired by the paper. It should be noted that the PageRank algorithm is based on a similar idea of spreading activation. (The difference between the PageRank algorithm and ours will be described later.)

2.2 Topic-Dependent Extraction of the Importance

Let $\mathbf{A}=\left(A_{ij}\right)$ be the adjacency matrix of a citation network: If paper j cites paper i, $A_{ij}=1$; otherwise $A_{ij}=0$. Let x_i be the activity (output) of node i corresponding to paper i. The input to node i is given by $I_i=\sum_{j=1}^{N}T_{ij}x_j$, where $T_{ij}\equiv A_{ij}\big/\sum_{k=1}^{N}A_{kj}$ corresponds to the transition matrix in the PageRank algorithm.

We assume a multi-hysteretic input/output (I/O) relationship (Fig. 2, see also Appendix) for each node. Hence the time evolution of spreading activation is defined by the following set of rules:

(I) If $x_i(t)<I_i(t)-\kappa/2$, $x_i(t+1)=I_i(t)-\kappa/2$;

(II) if $I_i(t)-\kappa/2\le x_i(t)\le I_i(t)+\kappa/2$, $x_i(t+1)=x_i(t)$;

(III) if $I_i(t)+\kappa/2<x_i(t)$, $x_i(t+1)=I_i(t)+\kappa/2$.

Here, κ represents the width of hysteresis.

Because of the hysteretic property of the I/O relationship, the iteration of (I)-(III) finally results in a steady state that is continuously dependent on the initial state [6]; that is, the spreading activation yields continuous attractors. This means that, if a given topic is represented by the initial state $\vec{x}(0)$, information specific to this topic can be retrieved as a continuous attractor $\lim_{t\to\infty}\vec{x}(t)$ [7, 8]. Note that, at the limit $\kappa\to 0$, the topic dependence (i.e., continuous dependence of attractors on the initial state) disappears and the iteration of (I)-(III) becomes identical to the PageRank algorithm.

We devised the above algorithm by the analogy of memory retrieval in the brain. Noteworthy is that topic-dependent retrieval of information by this algorithm is equivalent to Bayesian inference. These are elaborated in Appendix.

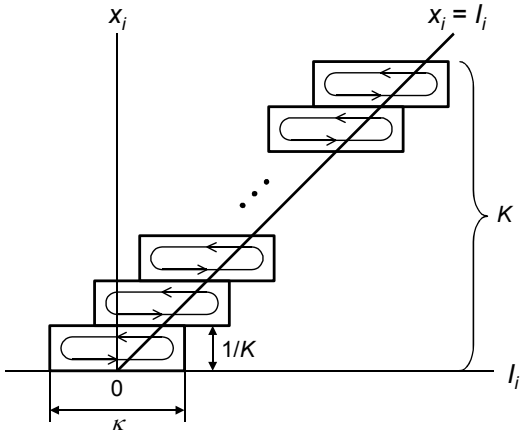

Fig. 2: Multi-hysteretic input/output relationship

2.3 Bibliographic Data

We prepared bibliographic information of papers published in major neuroscience journals. This includes for each paper: Identification data (ID) uniquely assigned to the paper; author(s); title; journal; volume; pages; year; IDs of cited papers; abstract; and so forth. Among them, only IDs of citing and cited papers are necessary to construct a citation network (Fig. 1).

2.4 Seed Documents

A given topic is fed into the algorithm through 'seed documents' prepared by a user. Seed documents are a set of documents forejudged to be relevant to the topic. Since user's knowledge about the topic is incomplete, seed documents might lack some documents relevant to the topic or include irrelevant ones. Nonetheless, the proposed algorithm can restore the documents truly relevant to the topic through Bayesian inference (Appendix).

Seed documents are encoded by the initial state, as follows: If paper i is a seed document, $x_i(0) = \rho > 0$; otherwise $x_i(0) = 0$.

2.5 Topic-Dependent Ranking

Papers that are highly activated in the steady state are regarded as what are truly relevant to the topic. Sorting these papers in descending order of acquired values of activity gives topic-dependent paper ranking.

2.6 Visualization

Papers highly activated in the steady state tend to be mutually connected by citation relations, thus forming a subnetwork of the whole citation network. Visualizing this subnetwork gives an overview of a 'genealogy' in the research field of the topic.

3. Results

Below we empirically demonstrate the use and the benefit of the proposed method of citation-network analysis. Theoretical evaluation of the algorithm is given in Appendix.

A whole citation network (Fig. 1) was constructed from papers published in major neuroscience journals. Then we took for example an emerging topic in neuroscience, expressed by the phrase "graded persistent activity and neural integrator". A set of 10 papers with abstracts showing high scores of word matching to this phrase was chosen as seed documents.

Table 1 shows the top 20 in the ranking obtained by our algorithm. Interviewing neuroscientists engaged in this research field, we confirmed that the obtained ranking was consistent with their expert knowledge. In particular, the paper by Seung et al. (2000) is not highly ranked by word matching and is dropped from the seed documents, but it is ranked first by our method. Indeed, this paper is widely acknowledged as what has marked the beginning of the research field.

Fig. 3 visualizes citation relations among the top 30 in the ranking. Each document icon symbolizes a paper and its size expresses the activity it has acquired, namely, the topic-dependent importance assigned to this paper. Icons are sorted in chronological order from the top to the bottom. Each arrowed line represents the citation relation between two papers. An arrow is directed from a citied to a citing paper, which denotes that the latter is under the influence of the former (Fig. 1, left). When an icon is clicked, bibliographic information of the

Table 1: Top 20 in the topic-dependent ranking

Rank	Title	Authors	Activity	Seed? 1(Y)/0(N)	Journal Vol. pages year
1	STABILITY OF THE MEMORY OF E	SEUNG HS,LEE DD,REIS BY,T	0.073831	0	NEURON 26 259-271 2000
2	IN VIVO INTRACELLULAR RECORD	AKSAY E,GAMKRELIDZE G,SE	0.073822	1	NAT NEUROSCI 4 184-193 2001
3	MODEL FOR A ROBUST NEURAL I	KOULAKOV AA,RAGHAVACH	0.069003	1	NAT NEUROSCI 5 775-782 2002
4	SYNAPTIC MECHANISMS AND NET	COMPTE A,BRUNEL N,GOLD	0.06481	0	CEREB CORTEX 10 910-923 2000
5	SYNAPTIC REVERBERATION UNDE	WANG XJ	0.061916	0	TRENDS NEUROSCI 24 455-463 2001
6	A MODEL OF VISUOSPATIAL WOR	CAMPERI M,WANG XJ	0.058597	0	J COMPUT NEUROSCI 5 383-405 1998
7	ROBUST PERSISTENT NEURAL AC	GOLDMAN MS,LEVINE JH,MA	0.057205	1	CEREB CORTEX 13 1185-1195 2003
8	A RECURRENT NETWORK MODEL	MILLER P,BRODY CD,ROMO	0.056	1	CEREB CORTEX 13 1208-1218 2003
9	BRAIN CALCULUS: NEURAL INTEG	MCCORMICK DA	0.055254	0	NAT NEUROSCI 4 113-114 2001
10	TIMING AND NEURAL ENCODING C	BRODY CD,HERNANDEZ A,ZA	0.055159	0	CEREB CORTEX 13 1196-1207 2003
11	HISTORY DEPENDENCE OF RATE	AKSAY E,MAJOR G,GOLDMA	0.052381	1	CEREB CORTEX 13 1173-1184 2003
12	SYNAPTIC BASIS OF CORTICAL P	WANG XJ	0.050142	0	J NEUROSCI 19 9587-9603 1999
13	MATCHING PATTERNS OF ACTIVI	CHAFEE MV,GOLDMAN-RAK	0.048867	0	J NEUROPHYSIOL 79 2919-2940 1998
14	BASIC MECHANISMS FOR GRADE	BRODY CD,ROMO R,KEPECS	0.047541	0	CURR OPIN NEUROBIOL 13 204-211 2003
15	ROBUST SPATIAL WORKING MEM	RENART A,SONG PC,WANG X	0.042419	0	NEURON 38 473-485 2003
16	NEURAL BASIS OF A PERCEPTUA	SHADLEN MN,NEWSOME WT	0.04152	0	J NEUROPHYSIOL 86 1916-1936 2001
17	CORRELATED DISCHARGE AMON	AKSAY E,BAKER R,SEUNG HS	0.040958	1	J NEUROSCI 23 10852-10858 2003
18	TEMPORAL STRUCTURE IN NEUR	PESARAN B,PEZARIS JS,SAH	0.040258	0	NAT NEUROSCI 5 805-811 2002
19	TURNING ON AND OFF WITH EXCI	GUTKIN BS,LAING CR,COLBY	0.038298	0	J COMPUT NEUROSCI 11 121-134 2001
20	DYNAMICS AND PLASTICITY OF S	BRUNEL N	0.037095	0	CEREB CORTEX 13 1151-1161 2003

corresponding paper is displayed in a pop-up window. The interviewees again acknowledged that the visualized network well represents how this research field has evolved.

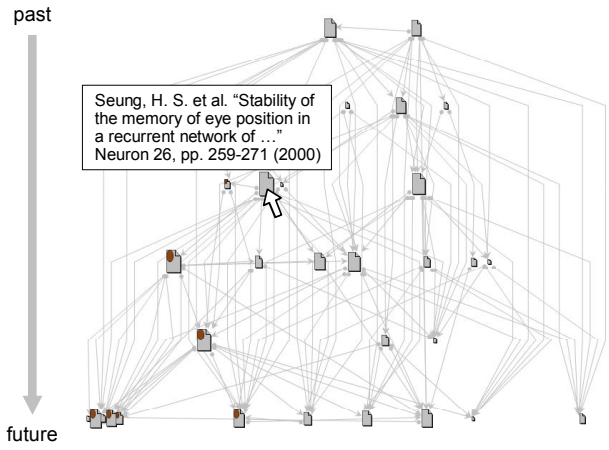

Fig. 3: Visualization of citation relations between extracted documents

4. Discussion

One problem overwhelming modern scientists is a tremendous number of papers being published every year. Even for a narrowed topic, what one has to read often exceeds what one can read. It is therefore critical to efficiently select papers to read from a pile of documents and prioritize them. The topic-dependent ranking of papers demonstrated here will relieve this problem. With this ranking method, one can get a list of papers to read in order of priority (Table 1).

The ranking, however, is just one-dimensional alignment. Papers extracted by the proposed method have a higher dimensional structure as a network. Visualizing this structure (Fig. 3) reveals which papers are central or subsidiary and which relations between papers are mainstream or tributary. Such a visualized network will serve as a chart for exploring the research field.

Acknowledgement

We purchased bibliographic information (Science Citation Index Expanded) from Thomson Scientific (presently Thomson Reuters) and used it under the license agreement. This study was partly supported by JSPS, KAKENHI (20500279).

Appendix

Below we demonstrate the following: The algorithm used for our topic-dependent document ranking replicates the mechanism of memory retrieval in the brain; information retrieval by this algorithm conforms to a type of Bayesian inference.

Memory retrieval in conventional neural-network information processing has been implemented by dynamical systems with discrete attractors [9, 10]. However, recent neurophysiological findings of graded persistent activity [6] suggest that memory retrieval in the real brain is more likely to be described by dynamical systems with attractors that continuously depend on the initial state [11-13]. Theoretical [14-17] as well as experimental [18, 19] studies have demonstrated that neurons with a multi-hysteretic response property can generate robust continuous attractors.

Consider a network of N multi-hysteretic neurons and let T_{ij} be the strength of connection from neuron j to neuron i. A multi-hysteretic response property of a single neuron (say, neuron i) can be modelled by stacked bistable compartments; each compartment is either in the 'on' or 'off' state, as illustrated in Fig. 2. The following one-step processes describe the state transition of the multi-hysteretic neuron:

$$S(0,i,t) \underset{R(1;i,t)}{\overset{G(0;i,t)}{\rightleftarrows}} S(1,i,t) \underset{R(2;i,t)}{\overset{G(1;i,t)}{\rightleftarrows}} \cdots \underset{R(K;i,t)}{\overset{G(K-1;i,t)}{\rightleftarrows}} S(K,i,t)$$

Here, $S(k,i,t)$ symbolizes the state of neuron i where the lowest k compartments are in the 'on' states at time t, and $G(k;i,t)$ and $R(k;i,t)$ are transition rates, given by

$$G(k;i,t) = \left[1 + \tanh \beta \left(I_i(t) - k/K - \kappa/2\right)\right]/2 \quad (1a)$$

$$R(k;i,t) = \left[1 - \tanh \beta \left(I_i(t) - k/K + \kappa/2\right)\right]/2 \quad (1b)$$

with $I_i(t) = \sum_{j=1,\, j\neq i}^{N} T_{ij} x_j(t-1)$ being the input to the neuron. The $\vec{x}_i(t)$ is the output from neuron i, defined as the number of 'on' compartments divided by K, the total number of compartments.

Taking $K \to \infty$ and $\beta \to \infty$, we have the conditional probability of $\vec{x}(t)$ defining interaction between neurons via the connection,

$$P_I\left(\vec{x}(t) \mid \vec{x}(t-1)\right) \sim \prod_{i=1}^{N} p_{I_i(t)}\left(x_i(t)\right) \quad (2a)$$

with

$$p_{I_i(t)}\left(x_i(t)\right) = \begin{cases} 1/\kappa & \text{for} \quad I_i(t) - \kappa/2 \le x_i(t) \le I_i(t) + \kappa/2; \\ 0 & \text{otherwise.} \end{cases} \quad (2b)$$

These expressions indicate that $\vec{x}(t)$ is in the N-dimensional hypercube with its centre at $\vec{I}(t)$ and the side length of κ; that is, the interaction between neurons makes $\vec{x}(t)$ closer to $\vec{I}(t)$, tolerating their difference within the margin of κ.

We assume that individual neurons are subject to independent and identically distributed (i.i.d.) Gaussian noise. Transition from $\vec{x}(t-1)$ to $\vec{x}(t)$ driven by i.i.d. Gaussian noise alone (namely, in the absence of the interaction) is defined by the conditional probability:

$$P_{i.i.d.}\left(\vec{x}(t) \mid \vec{x}(t-1)\right) = \prod_{i=1}^{N} p_G\left(x_i(t) \mid x_i(t-1)\right) \quad (3a)$$

where

$$p_G\left(x_i(t) \mid x_i(t-1)\right) \sim \exp\left[-\beta_G \left(x_i(t) - x_i(t-1)\right)^2 / 2\right]. \quad (3b)$$

Thus, transition from $\vec{x}(t-1)$ to $\vec{x}(t)$ by Gaussian noise in the presence of the interaction between neurons is defined by the combined conditional probability:

$$P\left(\vec{x}(t) \mid \vec{x}(t-1)\right) \sim P_I\left(\vec{x}(t) \mid \vec{x}(t-1)\right) P_{i.i.d.}\left(\vec{x}(t) \mid \vec{x}(t-1)\right). \quad (4)$$

Let $\vec{x}(0)$ represent a cue presentation at $t = 0$ and $\{\vec{x}(1), \vec{x}(2), \cdots\}$ be a temporal sequence generated in response to $\vec{x}(0)$. The conditional probability of $\vec{x}(T)$ for $\vec{x}(0)$ is hence expressed in the path-integral formulation:

$$P(\vec{x}(T)\,|\,\vec{x}(0)) \sim \int \mathbf{D}\vec{x} \prod_{t=1}^{T} P_I(\vec{x}(t)\,|\,\vec{x}(t-1)) P_{i.i.d.}(\vec{x}(t)\,|\,\vec{x}(t-1)), \quad (5a)$$

$$\mathbf{D}\vec{x} = \prod_{t=1}^{T-1} d\vec{x}(t). \quad (5b)$$

This formulation can be evaluated, as follows:

(i) For a given $\vec{x}(t-1)$, define $\vec{x}(t)$ by

$$x_i(t) = \arg\max_x p_G(x\,|\,x_i(t-1)) p_{I_i(t)}(x) \quad (i = 1, \cdots, N). \quad (6)$$

(ii) Repeat (i) for $t = 1, 2, \cdots$.

Note that the obtained sequence $\{\vec{x}(1), \vec{x}(2), \cdots\}$ represents 'the most probable path' in the path-integral formulation (5).

It can be easily checked that the process (i)-(ii) is equivalent to the iteration of (I)-(III) in the main text. Accordingly, the above process gives a continuous attractor. Now we assume that $\vec{x}(t)$ reaches this attractor at $t = T$. The $\vec{x}(T)$ will therefore represent the memory retrieved in response to a cue presentation represented by $\vec{x}(0)$.

Next we show that the path-integral formulation (5) conforms to a Bayesian formula. Since $\vec{x}(T)$ is a fixed point, multiplying the RHS of (5a) by

$$P(\vec{x}(T)) \equiv \prod_{i=1}^{N} p_{I_i(T)}(x_i(T)) \quad (7)$$

places no further probabilistic constraint on $\vec{x}(T)$; hence

$$P(\vec{x}(T)\,|\,\vec{x}(0)) \sim P(\vec{x}(T)) \int \mathbf{D}\vec{x} \prod_{t=1}^{T} P_I(\vec{x}(t)\,|\,\vec{x}(t-1)) P_{i.i.d.}(\vec{x}(t)\,|\,\vec{x}(t-1)). \quad (8)$$

Conversely, (8) defines the condition for that $\vec{x}(T)$ is a fixed point of the process (i)-(ii).

With the equality $P_{i.i.d.}(\vec{x}(t)\,|\,\vec{x}(t-1)) = P_{i.i.d.}(\vec{x}(t-1)\,|\,\vec{x}(t))$, the RHS of (5a) is rewritten as

$$\int \mathbf{D}\vec{x} \prod_{t=1}^{T} P_I(\vec{x}(t)\,|\,\vec{x}(t-1)) P_{i.i.d.}(\vec{x}(t-1)\,|\,\vec{x}(t)). \quad (9)$$

The factor $P_I(\vec{x}(t)\,|\,\vec{x}(t-1)) P_{i.i.d.}(\vec{x}(t-1)\,|\,\vec{x}(t))$ corresponds to the following procedure: Generate $\vec{x}(t-1)$'s from $\vec{x}(t)$ by i.i.d. Gaussian noise according to the probability distribution $P_{i.i.d.}(\vec{x}(t-1)\,|\,\vec{x}(t))$; among them, select by $P_I(\vec{x}(t)\,|\,\vec{x}(t-1))$ those defining the hypercube in which $\vec{x}(t)$ is contained. Thus the expression (9) can be regarded as $P(\vec{x}(0)\,|\,\vec{x}(T))$, the conditional probability of $\vec{x}(0)$ (or in the nomenclature of Bayesian inference, the 'likelihood' function of $\vec{x}(T)$); this describes how $\vec{x}(T)$ is degraded to $\vec{x}(0)$ by Gaussian noise in the presence of the interaction between N elements. Further regarding $P(\vec{x}(T)\,|\,\vec{x}(0))$ as the posterior probability of $\vec{x}(T)$ for a given $\vec{x}(0)$ and $P(\vec{x}(T))$ as the prior probability, we see that (5) conforms to the Bayesian formula:

$$\underbrace{P(\vec{x}(T)\,|\,\vec{x}(0))}_{\text{posterior}} \sim \underbrace{P(\vec{x}(T))}_{\text{prior}} \underbrace{P(\vec{x}(0)\,|\,\vec{x}(T))}_{\text{likelihood}}. \quad (10)$$

The algorithm formulated above enables Bayesian inference in the following situation: A number of elements (e.g. documents) interact with each other (e.g. via citation relations), and $\vec{x}(0)$ is given as observed data (e.g. seed documents signifying a topic), which might be generated as a result of the corruption of the original data $\vec{x}(T)$ (e.g. correct documents relevant to the topic).

Statistical-mechanical formulation of Bayesian inference conventionally assumes that i.i.d. noise alone contributes to the corruption of the original data [20]. This assumption, though it might simplify calculation, is inadequate when interaction between elements significantly affects the corruption. Degradation from the correct documents to seed documents is such a typical example; a pair of documents linked by citation might appear in or disappear from seed documents not independently but in a correlated manner. For appropriate inference of the correct documents from seed documents, therefore, the proposed algorithm is more suitable than the conventional ones.

References

1. Garfield, E. Science 122, 108-111 (1955).
2. Page, L. et al. Stanford Digital Library Technologies Project (1998). http://www-db.stanford.edu/~backrub/pageranksub.ps
3. Maslov, S. & Redner, S. Journal of Neuroscience 28, 11103-11105 (2008).
4. Haveliwala, T. IEEE Transaction on Knowledge and Data Engineering 15, 784-796 (2003).
5. Collins, A. M. & Loftus, E. F. Psychological Review 82, 407-428 (1975).
6. Okamoto, H., Tsuboshita, Y. & Fukai, T. Brain & Neural Networks 12, 235-248 (2005) (in Japanese), and references therein.
7. Tsuboshita, Y. & Okamoto, H. Neural Networks 20, 705-713 (2007).
8. Tsuboshita, Y. & Okamoto, H. Neural Networks 22, 922-930 (2009).
9. Hopfield, J. J. Proceedings of the National Academy of Science USA 79, 2554-2558 (1982).
10. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. Nature Neuroscience 3 (supplement), 1184-1191 (2000).
11. Seung, H. S. et al. Neuron 26, 259-271 (2000).
12. Okamoto, H. et al. Journal of Neurophysiology 97, 3859-3867 (2007).
13. Okamoto, H. & Fukai, T. PLoS Computational Biology 5, e1000404 (2009).
14. Fransen, E. et al. Neuron 49, 735-746 (2006).
15. Goldman, M. S. et al. Cerebral Cortex 13, 1185-1195 (2003).
16. Loewenstein, Y. & Sompolinsky, H. Nature Neuroscience 6, 961-967 (2003).
17. Teramae, J. & Fukai, T. Journal of Computational Neuroscience 18, 105-121 (2005).
18. Egorov, A. V. et al. Nature 420, 173-178 (2002).
19. Winograd, D., Destexhe, A. & Sanchez-Vives, M. V. Proceedings of the National Academy of Science USA 105, 7298-7303 (2008).
20. Nishimori, H. Statistical Physics of Spin Glasses and Information Processing. Oxford University Press (2001).