Collective Classification of Biomedical Articles using T-Cell Cross-regulation

Alaa Abi-Haidar[1,2] and Luis M. Rocha[1,2]

[1]School of Informatics and Computing, Indiana University, Bloomington IN 47401, USA
[2]FLAD Computational Biology Collaboratorium, Instituto Gulbenkian de Ciência, Oeiras, Portugal
{aabihaid,rocha}@indiana.edu

Abstract

We continue our investigation of a bio-inspired solution for binary classification of textual documents inspired by T-cell cross-regulation in the vertebrate adaptive immune system, which is a complex adaptive system of millions of cells interacting to distinguish between self and nonself substances. In analogy, automatic document classification assumes that the interaction and co-occurrence of thousands of words in text can be used to identify conceptually-related classes of documents—at a minimum, two classes with relevant and irrelevant documents for a given concept (e.g. articles with protein-protein interaction information). Our agent-based method for document classification expands the analytical model of Carneiro et al [5], by allowing us to deal simultaneously with many distinct populations of antigen-specific T-Cells and their collective dynamics. We have previously extended this model to produce a spam-detection system [2; 3]. We have also developed our agent-based model further to apply it to biomedical article classification [4], testing it on a dataset of biomedical articles provided by the BioCreative 2.5 challenge [17]. Here, we study the effect that the sequence of presentation of articles has on classification performance, as well as the robustness of the ensuing T-cell cross-regulation dynamics to initial biases of the proportions of effector and regulatory T-cells. We show that classification is improved when we preserve the original temporal order of biomedical articles, suggesting that our model is capable of tracking the natural conceptual drift of the relevant biomedical literature. We further show that initial biases in the proportions of T-cells are corrected by the dynamics of the model. Our results are useful for biomedical text mining, but they also help us understand T-cell cross-regulation as a potential general principle of classification available to collectives of molecules without a central controller. While there is still much to know about the specifics of T-cell cross-regulation in adaptive immunity, Artificial Life allows us to explore alternative emergent classification principles while producing useful bio-inspired tools.

Introduction

At least since the start of systematic genomic studies, there has been a tremendous growth of scientific publications in the life sciences [13]. Pubmed (`http://pubmed.gov`) now contains a growing collection of more than 19 million biomedical articles. Manually classifying these articles as relevant or irrelevant to a given topic of interest is very time consuming and inefficient for curation of new published articles [14]. Literature (or text) mining offers solutions for automatic biomedical document classification and information extraction from huge collections of text, as well as the linking of numerous biomedical databases and knowledge resources [14; 28]. Because it is very important to validate and assess the quality of proposed solutions, various community-wide competitions and challenges have been organized so that automatic systems can be evaluated against human annotated data sets (e.g. TREC Genomics [10]). One such effort is the BioCreative challenge, which aims to assess biomedical literature mining in real-world scenarios [11; 18; 17]. Machine learning has offered a plethora of solutions to this problem [14; 8], however, even the most sophisticated of solutions often overfit to the training data and do not perform as well on real-world scenarios such as that provided by BioCreative [1; 16]. One of the challenges of biomedical article classification in real-world scenarios is the presence of highly unbalanced classes; typically, there are many more irrelevant than relevant documents, without prior knowledge of class proportions. This was the case of the article classification data set in the Biocreative BC2.5 challenge [17]. While participating teams (including our own team [16]) did not enter bio-inspired solutions, the unbalanced nature of classes and the presence of conceptual drift, which we showed to occur between training to testing data sets [1; 16], may be a good scenario to test classifiers inspired by the vertebrate immune system—which must operate under class-imbalance with permanent drift in the populations of pathogens encountered. Therefore, here we explore the feasibility of using T-Cell cross-regulation dynamics to classify biomedical articles using the real-world scenario provided by the Biocreative 2.5. data set.

The immune system (IS) is a complex biological system made of millions of cells all interacting to distinguish between self and nonself substances, to ultimately attack the latter [12][1]. In analogy, relevant biomedical articles for a

[1]We use the terminology of self/nonself discrimination, though perhaps a more accurate description is classification of harmless

given concept need to be distinguished from irrelevant ones. To perform such a topical classification, we can use the occurrence and co-occurrence of thousands of words in a document. In this sense, words can be seen as interacting in a text in such a way as to allow us to distinguish between relevant and irrelevant documents—in analogy with the interactions among T-cells and antigens that lead to self/nonself discrimination in the immune system, as we describe below.

Our Artificial Life approach is based on the idea that the immune system is a distributed collection of molecular constituents with no central controller [25]. Therefore, its classification ability needs to result from a *collective classification* process, defined as the ability of decentralized systems of many components to classify situations that require global information or coordinated action [20]. Nature is full of examples of collective classification: the dynamics of stomata cells on leaf surfaces are known to be statistically indistinguishable from the dynamics of automata that are capable of performing nontrivial classification [21], biochemical intracellular signal transduction networks are capable of emergent classification [9], quorum sensing in bacteria [33] and social insects [23], etc. We can study collective classification in general models of complex systems such as Cellular Automata, namely by identifying regular patterns in the dynamics that store, transmit and process information [6; 24; 27]. Here, instead of looking at general models of complex systems, we focus on a specific immunological model of T-Cell cross-regulation dynamics [5]. We are are interested in exploring the collective dynamics of this model to: (1) build a novel bio-inspired machine learning solution for document classification, and (2) understand how well collections of T-Cells engaged in cross-regulation perform as a classifier. The first goal entails a bio-inspired approach to computational intelligence, and the second a computational biology experiment, but both are based on artificial life principles. It should be noted that recent work in artificial immune systems (AIS) [30] has lead to a few immune-inspired solutions to document classification in general [32], however, none to our knowledge has been applied to biomedical article classification nor do they employ T-cell cross-regulation dynamics.

We have already proposed an agent-based model of T-cell cross-regulation for spam detection [2; 3]. Our distributed model extendes the original analytical model of T-Cell cross-regulation dynamics [5] to be able to deal with many multiple features simultaneously, and therefore render the model applicable to real-world applications. Our results on spam-detection were comparable to state-of-art text classifiers [2; 3]. However, our initial agent-based implementation of cross-regulation dynamics did not explore important parameter configurations such as the death rate of T-cells or the best training strategies. It also lacked an extensive parameter search for optimized performance. More recently, we started addressing some of these issues on full-text biomedical data from BioCreative, and showed that T-cell death is important to obtain better classification [4]. This is an interesting result, showing that the loss of T-cells rather than hindering, can improve the collective classification of relevant documents. Therefore, the dynamics of T-cell cross-regulation as proposed by Carneiro et al. [5] can lead to the elimination of T-cells that are not useful for classification—even in our extended formulation which contains hundreds of distinct T-cells representing antigens or textual features. We also showed that training exclusively on relevant documents (or self antigens) leads to worse classification performance than training on both relevant and irrelevant documents [4]. This is interesting for tuning the algorithm in text mining settings, but also suggests that T-cell cross-regulation in the vertebrate adaptive immune system can improve from a "training" stage where it is presented with both self and nonself antigens.

Here, we study the importance of the original temporal sequence of bio-medical articles. Text mining classifiers do not typically depend on the sequence of documents they are trained with, but our model of T-cell cross-regulation dynamics does. Therefore, we are interested in ascertaining if the sequence-dependence of ensuing collective dynamics can be used to track the natural change in real-world textual corpora, i.e. concept drift [31]. We also study the effect of biases in the initial T-cell population. This more extensive study allows us to better understand the behavior of T-cell cross-regulation dynamics and establish its capability to classify sequential data. It also leads to a competitive, novel bio-inspired text classification algorithm.

The Immune System as Inspiration

The vertebrate adaptive immune system[2] (IS) is a complex network of cells that distinguishes between self and nonself substances or antigens—usually fragments of proteins that can be recognized by the immune system. When nonself antigens are discovered, an immune response to eliminate them is set in motion. Recognizing self antigens, which obviously should not lead to an (auto)immune response to eliminate them, is resolved by negative selection of T-cells which takes place in the thymus, and removes T-Cells that strongly bind to self antigens—after positive selection of T-Cells that are capable of binding with the major histocompatibility complex (MHC). It is in the thymus that T-cells develop and mature; only T-cells that have failed to bind to self antigens are released (as naive T-cells), while the rest of the T-cells is culled. Mature T-cells are allowed out of the thymus to detect nonself antigens. They do this by binding to

vs. harmful substances, because harmless can also include antigens from bacteria that are necessary for vertebrate bodies, and harmful can also include body's own tumor cells.

[2] A good, though already a bit dated, overview of the vertebrate immune system for the artificial life community is Hofmeyer's [12].

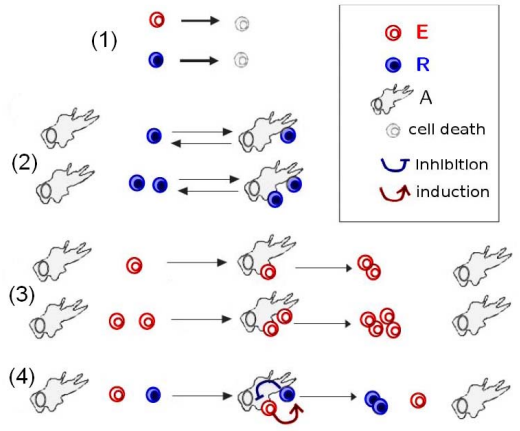

Figure 1: CRM interactions that define the dynamics of APC and E and R T-cells. The model assumes that APC can only form conjugates with a maximum of two T-cells. Adapted from [5].

antigen presenting cells (typically B-cells, macrophages and dendritic cells) that collect and present antigens via MHC after breaking them by lysosome. The specific T-cells that are able to bind to the presented antigens then stimulate B-cells that start a cascade of events leading to antibody production and the destruction of the pathogens or tumors linked to the antigens. However, it is possible that T-cells and B-cells, which are also trained in the thymus and bone marrow, mature before being exposed to all self antigens. Even more problematic is the somatic hypermutation that ensues in lymph nodes after the activation of B-cells. At this stage, it is possible to generate many mutated B-Cell clones that could bind to self antigens. Either situation can cause autoimmunity by generating T-cells capable of attacking self antigens. One way to deal with this problem is by a process called costimulation which involves the co-verification of self antigens by both T-cells and B-cells before the antigen is identified as associated with a nonself pathogen to be attacked. To further insure that the T-cells do not attack self, another type of T-cells known as regulatory T-cells, are formed in the thymus where they mature to avoid recognizing self antigens. These regulatory T-cells have the responsibility of preventing autoimmunity by down-regulating other T-cells that might bind and kill self antigens. Our model is based on this process of T-Cell cross-regulation.

Artificial Immune Systems (AIS) are artificial life tools, inspired by theories and components of the immune system, and applied towards solving computational problems, such as categorization, optimization and decision making [7]. Common AIS techniques are based on specific theoretical models explaining the behavior of the IS such as: Negative Selection, Clonal Selection, Immune Networks and Dendritic Cells [30]. AIS fall in categories: (1) mathematical and computational models to understand IS behavior and (2) engineering of adaptive machine learning algorithms. While our approach fits more immediately in the second category, our goal is also to use our classifier to test the prevailing model of T-cell cross-regulation and therefore also contribute to the first category of the study of AIS.

The Cross-Regulation Model

The *T-cell Cross-Regulation Model* (CRM) [5] is a dynamical system that aims to distinguish between self and nonself protein fragments (antigens) using only four possible interaction rules amongst three cell-types: *Effector T-cells* (E), *Regulatory T-cells* (R) and *Antigen Presenting Cells* (APC). As their name suggests, APC present antigens for the other two cell-types, E and R, to recognize and bind to them. Effector T-cells (E) proliferate upon binding to APC, unless adjacent to regulatory T-cells (R), which regulate E by inhibiting their proliferation. For simplicity, proliferation of cells is limited to duplication in quantity in contrast to having a proliferation rate. T-cells that do not bind to APC die off with a certain death rate. The dynamics of the CRM depend on four interaction rules defined by the following reactions (illustrated in Fig. 1):

$$E \xrightarrow{d_E} \{\} \text{ and } R \xrightarrow{d_R} \{\} \quad (1)$$

$$A + R \rightarrow A + R \quad (2)$$

$$A + E \rightarrow A + 2E \quad (3)$$

$$A + E + R \rightarrow A + E + 2R \quad (4)$$

Reaction (1) defines E and R apoptosis with the corresponding death rates d_E and d_R. The last three proliferation reactions define the maintenance of R (2), the duplication of E (3), and the maintenance of E and duplication of R (4) .

Carneiro et al [5] developed the analytical CRM to study the dynamics of a population of T-cells and APC that present a single antigen associated with a single T-cell population. In [2; 3], we extended the original CRM model to be able to deal with multiple populations of antigens and T-Cells using agent-based modeling. More recently, Sepulveda [26, pp 111-113] extended the original CRM to study analytically multiple populations of T-cells that recognize antigens presented by APC capable of presenting at most two distinct antigens. In our model, explained in detail in the next section, APC are capable of presenting hundreds of antigens to be recognized by T-cells of hundreds of different populations, using the same four interaction rules of the CRM.

The Agent-Based Cross-Regulation Model

In order to adapt CRM to an *Agent-Based Cross-Regulation Model* (ABCRM) for text classification, one has to think of documents as analogous to the organic substances that upon entering the body are broken into constituent pieces. These pieces, known as epitopes, are presented on the surface of Antigen Presenting Cells (APC) as antigens. In the ABCRM, antigens are textual features (e.g. words, bigrams,

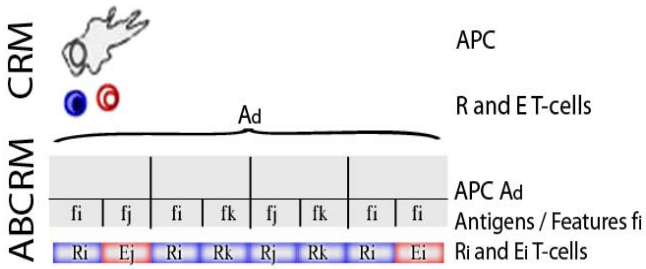

Figure 2: To illustrate the difference between the CRM and the ABCRM, the top part of the figure represents a single APC of the CRM which can bind to a maximum of two T-Cells. The lower part represents the APC for a document d in the ABCRM, which contains many pairs of antigen/feature "slots" where pairs of T-cells can bind. In this example, the first pair of slots of the APC A_d presents the features f_i and f_j; a regulatory T-cell R_i and an effector T-cell E_j bind to these slots, which will therefore interact according to reaction (4)—R_i inhibits E_j and in turn proliferates by doubling. The next pair of slots leads to the interaction of regulatory T-cells R_i,R_k that proliferate via reaction (2), etc.

titles, numbers) extracted from articles and presented by artificial APC such that they can be recognized by a number of artificial Effector T-cells (E) and artificial Regulatory T-cells (R). Individual E and R have receptors for a single, specific (textual) feature: they are *monospecific*. E proliferate upon binding to antigens presented by APC unless suppressed by R; R suppress E when binding in adjacent locations on APC. Individual APC present various document features: they are *polyspecific*. Each APC is produced when documents enter the artificial cellular dynamics, by breaking the former into constituent textual features. Therefore we can say that APC are representative of specific documents whereas E and R are representative of specific features.

A document d contains a set of features F_d; an artificial APC A_d that represents d, presents antigens/features $f_i \in F_d$ to artificial E and R T-cells. E_i and R_i bind to a specific feature f_i on any APC that contains it; if $f_i \in F_d$, then either E_i or R_i may bind to A_d as illustrated in figure 2. In biology, antigen recognition is a more complex process than mere polypeptide sequence matching, but for simplicity we limit our feature recognition to string matching. Once T-cells bind to an APC A_d, every pair of adjacent T-cells on A_d proliferates according to reaction rules (2-4). APC are organized as a sequence of pairs of "slots" of (textual) features, where T-cells, specific for those features, can bind. We use this antigen/feature presentation scheme of pairs of "slots" to simplify our algorithm. In future work we will study alternative feature presentation scenarios. In summary, each T-cell population is specific to and can bind to only one feature presented by any APC. Implementing the algorithm as an Agent-based model (ABM) allows us to deal with the recognition and co-recognition (co-occurrence in the same document/APC) of many features simultaneously, rather than a single one as the original CRM does.

The ABCRM uses incremental learning to first train on N labeled documents (relevant and irrelevant), which are ordered sequentially (typically by time signature) and then test on M unlabeled documents that follow in time order. The sequence in which documents are received affects the artificial cellular dynamics, as incoming APC and T-cells face a T-cell dynamics that depends on the specific documents previously encountered. Therefore, we use publication-time as the default ordering for incoming documents, but we study here if there is an advantage to preserving the original temporal sequence of articles (see below).

Carneiro et al [5] show that both E and R T-cells co-exist in healthy individuals assuming enough APC exist. R T-cells require adequate amounts of E T-cells to proliferate, but not too many that can out-compete R for the specific features presented by APC. "Healthy" T-cell dynamics is identified by observing the co-existence of both E and R features with $R \geq E$. "Unhealthy" T-cell dynamics is identified by observing $E \gg R$, and should result when encountering many irrelevant features in a document—in analogy with encountering many nonself antigens. In other words, features associated with relevant documents should have E and R T-cell representatives in comparable numbers in the artificial cellular dynamics (with slightly more R). In contrast, features associated with irrelevant documents should have many more E than R T-cells. Therefore, when a document d contains features F_d that bind mostly to E rather than R cells, we can classify it as irrelevant—and relevant in the opposite situation.

The ABCRM is controlled by 6 parameters:

- E_0 is the initial number of Effector T-cells generated for all new features
- R_0^- is the initial number of Regulatory T-cells generated for all new features in irrelevant and unlabeled (testing) documents
- R_0^+ is the initial number of Regulatory T-cells generated for all new features in relevant documents
- d_E is the death rate for Effector T-cells that do not bind to APC
- d_R is the death rate for Regulatory T-cells that do not bind to APC
- n_A is the number of slots in which each feature f_i is presented on APC

In the IS, millions of novel T-cells are randomly generated in the thymus every day to attempt to predict future antigens. In our algorithm, in contrast, we generate T-cells only for features (words) occurring in the relevant document corpus. This is reasonable because the space of meaningful words in a language are largely fixed and much smaller than the space of possible polypeptide epitopes in biology. When (textual) features are encountered for the first time, a fixed initial number of E_0 effector T-Cells and R_0 regulatory T-Cells is generated for every new feature f_i. These initial values of T-cells vary for relevant and irrelevant documents

in training and in testing stages. More Regulatory (R_0^+) than Effector T-cells are generated for features that occur for the first time in documents that are labeled relevant in the training stage ($R_0^+ > E_0$), while fewer Regulatory (R_0^-) than Effector T-cells are generated in the case of irrelevant documents ($R_0^- < E_0$). Features appearing in unlabeled documents for the first time during the testing stage are treated as features from irrelevant documents, assuming that new features are irrelevant (nonself) until neutralized by the collective dynamics given their co-occurrence with relevant ones.

Naturally, relevant features will occur in irrelevant documents and vice versa. However, the assumption is that relevant features tend to co-occur more frequently with other relevant features in relevant documents and similarly for irrelevant features. Therefore, the proliferation dynamics defined by the 4 reactions and guided by co-binding to APC slots is expected to correct the erroneous initial bias. But this self-correction has not been proven, and it is one of the issues we test in the present work, as detailed below. The pseudocode for the algorithm is shown below:

ABCRM:
(1) $\forall d$ generate a linear array A_d presenting each $f_i \in F_d$ at n_A arbitrary, randomly distributed slots
(2) Let C_t contain E_k and R_k T-cells for all features f_k in the cellular dynamics at time t.
(3) For an incoming document d, $\forall f_i \in F_d$, if $E_i \notin C_t$ and $R_i \notin C_t$ then,
(4) $E_i = E_0$ (generate E_0 Effector T-cells for f_i)
(5) if d is labeled relevant.
(6) $R_i = R_0^+$ (generate R_0^+ Regulatory T-cells for f_i)
(7) otherwise
(8) $R_i = R_0^-$ (generate R_0^- Regulatory T-cells for f_i)
(9) update C_t with E_i and Ri
(10) Let all E_i, R_i *bind specifically*[3] to matching f_i on A_d:
(11) $\forall$ pairs of adjacent (f_i, f_j) on A_d apply the interaction rules: $(R_i, R_j) \rightarrow R_i + R_j$ $(E_i, E_j) \rightarrow 2.E_i + 2.E_j$ $(E_i, R_j) \rightarrow E_i + 2.R_j$
(12) $\forall R_i$, E_i that bind to A_d, update total number of E_i, R_i
(13) $\forall R_k, E_k \in C_t$ that do not bind to A_d, cull E_k and R_k according to death rates d_E and d_R
(14) If d is unlabeled, Let $R(d) = \sum_{f_i \in F_d}(R_i)$ and $E(d) = \sum_{f_i \in F_d}(E_i)$
(15) Compute the normalized score $S(d) = \frac{R(d)-E(d)}{\sqrt{R^2(d)+E^2(d)}}$
(16) If $S(d) > 0$ then classify d as relevant, else irrelevant.

Data and Feature Selection

The BioCreative (BC) challenge aims to assess the quality of biomedical literature mining algorithms such as article classifiers. The article classification task of Biocreative 2.5 [17] was based on a training data set (T) comprised of 61 full-text articles relevant (P_T) to the topic of *protein-protein interaction* (PPI) and 558 irrelevant ones (N_T). The realistic imbalance between the relevant and irrelevant instances is very challenging for common machine learning techniques, since there are few instances of the topical category of interest to generalize from. Because we cannot predict how imbalanced the validation set will be, we first search for optimal ABCRM parameters on a smaller sample of the training that is balanced in the numbers of relevant and irrelevant documents. For this purpose, we chose the first 60 relevant and sampled 60 irrelevant articles that were published around the same date (uniform distribution between Jan and Dec 2008) as illustrated in figure 3. For final validation we used the entire Biocreative 2.5 testing data set (V) consisting of 63 full-text articles relevant to PPI (P_V) and 532 irrelevant ones (N_V) as also shown in figure 3. Furthermore, we compared our optimized algorithm with a Naive Bayes (NB) [19] and a support vector machine (SVM) classifier [15].

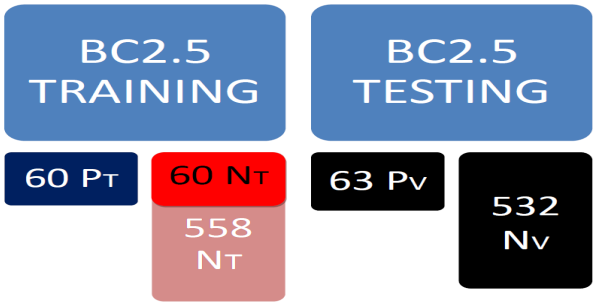

Figure 3: Numbers of relevant (P) and irrelevant (N) documents in the training (T) and testing (V) data sets of the Biocreative 2.5 challenge. In the parameter search stage, we use a balanced set of 60 P_T (blue) and 60 N_T (red) randomly selected articles from the training data set. In the testing stage we use the unbalanced validation set containing 63 P_V (black) and 532 N_V (black) documents. Notice that the validation data was provided to the participants in the classification task of Biocreative 2.5 unlabeled, therefore participants had no prior knowledge of class proportions.

We pre-processed all articles by filtering out common words[4] and porter stemming [22] the remaining words which are all the potential features. We then ranked words/features f extracted from training articles (T)[5] according to two scores: the first one is the average TF.IDF[6] [8], and the second one is the separation score $S(f) = |p_P(f) - p_N(f)|$ where p_P (p_N) is the probability of a feature occurring in a relevant (irrelevant) document of the training set T [1; 16]. The final rank $R(f_i)$ for every feature f_i is given by the product of the ranks obtained from both scores; we used only the 650 top ranked features according

[3]While the features f_i are arbitrarily distributed over the APC A_d, E_i and R_i that are specific to f_i, are sampled from C_t based on the proportions of E_i to R_i

[4]The list of common (stop) words includes 33 of the most common English words from which we manually excluded the word "with", as we know it to be of importance to PPI

[5]For feature extraction we used both the training data of Biocreative 2.5 and Biocreative 2 as described in [16]; all classifiers used the exact same feature set.

[6]TF.IDF is a common text weighting measure to evaluate the importance of a feature/word in a document in a corpus. TF stands for term frequency in a document and IDF for inverse document frequency in the corpus. [8]

Parameter	Range	Step
E_0	[1,7]	1
R_0^-	[3,12]	1
R_0^+	[3,12]	1
d_E	[0.0,0.4]	0.1
d_R	[0.0,0.4]	0.1
n_A	[2,22]	2

Table 1: Parameter ranges used for optimizing the ABCRM

to $R(f_i)$. Features such as "interact", "lysat" and "transfect" were ranked above others for their high ranks according to both scores. See [16] for more details about the feature extraction procedure.

Parameter Search and Robustness

We performed an exhaustive parameter search by training the ABCRM on 60 balanced full-text articles (30 P_T and 30 N_T from BC2.5 training) and testing it on the remaining 60 balanced ones (also 30 P_T and 30 N_T from BC2.5 Training) as illustrated in figure 3[7]. Each run corresponds to a unique configuration of the 6 parameters of the ABCRM. The explored parameter ranges are listed in table 1 which result in a total of 192500 unique parameter configurations for each experiment. Finally, the parameter configurations were sorted with respect to the resulting F-score measure of performance[8], which is a good measure between precision and recall when applied to balanced data [29].

We compiled the performance of the ABCRM on the entire parameter search space for two distinct experiments: (1) effect of **sequence order** of articles, and (2) effect of varying **initial T-cell counts**. In another publication [4] we showed that a positive T-Cell death ratio improves classification, whereas training exclusively on relevant documents lowers the performance. In both experiments, we choose the 50 configurations with highest F-score measure to study the ABCRM performance, because we are interested in identifying the experimental setups that lead to higher **robustness** to parameter changes. We compare experimental outcomes with the paired student t-test; the null hypothesis is that the two samples are drawn from the same distribution. A p-value < 0.01 rejects the null hypothesis, establishing a statistical distinction between the data drawn from two experimental setups—in our case, the data from each experiment are the top 50 F-score values obtained. Finally, we train on both relevant and irrelevant documents as this was shown to be advantageous [4], and search for optimal parameter configurations (including T-Cell death ratios).

The **first** experiment aims to establish how much the sequence order of processing documents impacts performance. In particular, we test if preserving the original temporal order of biomedical documents results in better performance, as this would indicate that the ABCRM can use its sequence-dependent dynamics to track the natural concept or topical drift and thus improve classification. Therefore, we compared the performance of the ABCRM when tested on a sequence of biomedical articles ordered by the original publication, against randomly shuffling the articles. We tested four distinct experimental setups in order to fully explore the influence of document order:

1. Ordered training set ⇒ ordered testing set
2. Ordered training set ⇒ shuffled testing set
3. Shuffled training set ⇒ shuffled testing set
4. Shuffled training set ⇒ ordered testing set

In the case of shuffled sets, we produced 8 runs with distinct random document orderings; in those cases, performance is represented by central tendency and variation.

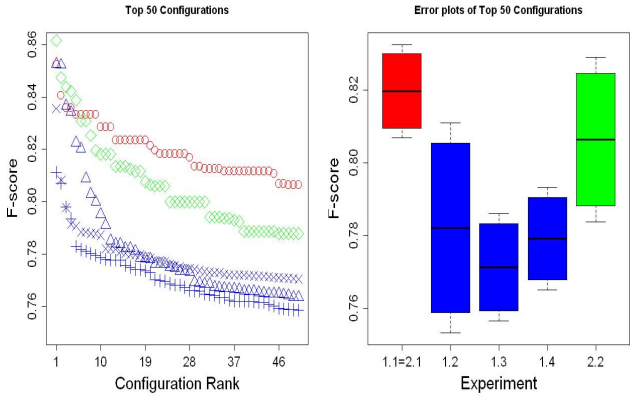

Figure 4: Left: top 50 parameter configurations ranked in terms of F-score for experimental setups 1.1/2.1 (red circles), 1.2 (blue triangles), 1.3 (blue pluses), 1.4 (blue crosses), and 2.2 (green diamonds). Right: mean (line), 95%CI (boxes), and standard deviation (whiskers) of F-scores for top 50 parameter configurations.

The results of this experiment are summarized in figure 4. The robustness of performance of the first experimental setup (preserving temporal order of articles) is significantly above the other setups. Using the paired student t-test as described above, we conclude that the ABCRM is sensitive to article order—i.e. if the articles are shuffled, the performance is worse. While the performance of the best classifier obtained via experimental setup 1.2 is equivalent to the best one obtained for experimental setup 1.2 (F-Score = 0.853, see table 2 and figure 4), that setup is very sensitive to parameter changes and the performance quickly and significantly decreases for subsequent best classifiers (see figure

[7]Notice that this parameter search on the provided labeled training data uses only the information available to the teams participating in Biocreative 2.5 challenge, and none of the testing data whose labels were revealed post-challenge.

[8]F-score = $\frac{2.Precision.Recall}{Precision+Recall}$ where Precision = $\frac{TP}{TP+FP}$ and Recall = $\frac{TP}{TP+FN}$. True Positives (TP) and False Positives (FP) are the classifier's correct and incorrect predictions for relevant documents, while True Negatives (TN) and False Negatives (FN) are the correct and incorrect predictions for irrelevant documents.

Exp.	F-Score	E_0	R_0^+	R_0^-	d_R	d_E	n_A
$1.1 = 2.1$	0.852	2	11	10	0.3	0.2	18
1.2	0.853	2	7	6	0.0	0.0	20
2.2	0.862	3	8	7	0.2	0.1	14

Table 2: Performance and parameters of top classifiers in experiments 1.1, 1,2, 2.1 and 2.2.

4). Indeed, the performance of the top 50 classifiers for experimental setups 1.2, 1.3, and 1.4 is statistically indistinguishable from each other, but is significantly lower than the performance of the top 50 classifiers for experimental setup 1.1. This means that there is indeed a conceptual drift in the Biocreative 2.5 article data stream, and the ABCRM can track it better (and in a more robust manner) when publication date is used as the sequence for processing articles than when the temporal order of articles is shuffled. This also suggests that the process of T-Cell cross-regulation in the IS, as modeled here, can track changing environments.

In the **second** experiment we test the effect of the initial biases introduced when features are first encountered. The initial biases of regulatory T-cells injected in the dynamics for a new feature f_i, depend on whether the first document d where the feature is encountered is labeled irrelevant/unknown (R_0^-) or relevant (R_0^+). Since features will occur in both relevant and irrelevant articles, this initial bias for a feature could be detrimental, as a feature most associated with one class could be first encountered on a document of the opposite class. Therefore, it is important to test if the dynamics of the four reactions and APC feature co-presentation that define the ABCRM can self-correct such erroneous biases. To perform this test, we altered the ABCRM algorithm such that T-cells are incremented appropriately every time a feature occurs in a document, and not just the first time the feature occurs (as the canonical algorithm does). Specifically, every time a feature f_i occurs in a document d, we increment $E_i = E_i + E_0$ and $R_i = R_i + R_0^+$ if d is labeled relevant and $R_i = R_i + R_0^-$ if d is labeled irrelevant or unknown.

The results of this experiment are also summarized in figure 4. The performance of top classifiers obtained for experimental setups 2.1 (same as 1.1) and 2.2 is shown in table 2. While the best overall classifier is obtained with experimental setup 2.2, the performance of both setups is statistically indistinguishable. Indeed, using the paired student t-test as described above, we conclude that this modification does not improve the performance of the ABCRM on the Biocreative data set, thus showing that the initial bias can be corrected by the ABCRM collective dynamics. Because features most associated with a given class tend to co-occur in text with other features most associated with the same class, they will also tend to be co-presented in APC and thus the relevant T-cells will proliferate with similar rates. Therefore, the dynamics of the ABCRM can self-correct initial erroneous biases from the natural textual co-occurrence of features. This shows that T-Cell cross-regulation as modeled here can self-correct initial antigen misclassification by the IS, assuming that antigens from one class (self/nonself) tend to co-occur with antigens from the same class.

Validation and Conclusions

To test the ABCRM on the full, unbalanced testing set of the Biocreative challenge (see figure 3), thus establishing its merit as a bio-inspired biomedical literature mining classifier, we adopted the best parameter configuration from the canonical ABCRM (experimental setup 1.1 and 2.1, see table 2) obtained from the parameter search described above. We compared the ABCRM classifier with the multinomial Naive Bayes (NB) with boolean attributes [19], and the publicly available SVM^{light} implementation of SVM applied to normalized feature counts [15]. All classifiers were tested on the same features obtained from the same data.

	ABCRM	NB	SVM	Mean	StDev.	Median
Precision	0.22	0.14	0.24	0.38		
Recall	0.65	0.71	0.94	0.68		
F-score	0.33	0.24	0.36	0.39	0.14	0.38
Accuracy	0.71	0.52	0.74	0.67	0.30	0.84
AUC	0.34	0.19	0.46	0.43	0.17	0.44
MCC	0.24	0.13	0.31	0.31	0.19	0.33

Table 3: F-Score, Accuracy, AUC and MCC performance of various classifiers when training on the balanced training set of articles and testing on the full unbalanced Biocreative 2.5 testing set. Also shown is the central tendency and variation of all systems submitted to Biocreative 2.5.

Since the F-score and Accuracy are not very reliable for evaluating unbalanced classification [29], we also use the Area Under the interpolated precision and recall Curve (AUC) and Matthew's Correlation Coefficient (MCC). The results are listed in table 3, which also includes the central tendency of the results of all systems submitted by all Biocreative 2.5 participating teams [17; 16]. It should be noted that the ABCRM, NB, and SVM classifiers we tested here, used only single-word features because we wish to establish the feasibility of the method. In contrast, most classifiers submitted to the Biocreative 2.5 challenge (including another method from our group which was one of the top-performing classifiers [16]) used more sophisticated features such as bigrams and problem-specific entities. Therefore, it is not surprising that these methods as tested here performed under the mean of the challenge. Our goal was to establish the ABCRM as a new bio-inspired text classifier to be further improved in the future with more sophisticated features. When we compare its performance to NB and SVM on the exact same single-word features, the results are encouraging. Indeed, based on the given measures, while SVM out-performed the ABCRM, the latter out-performed NB. Therefore, the dynamics T-Cell cross-regulation lead to a competitive collective classification of biomedical articles, which we intend to develop further.

In conclusion, we observed that our algorithm adapts to the initial bias of T-cell populations generated for new fea-

tures, and it performs best when tested on a sequence of articles ordered by publication date—showing that it can track concept drift in the biomedical literature. These properties of our Artificial Life model also show that T-Cell cross regulation is capable of efficient collective classification of nonself antigens and suggest that T-Cell cross-regulation can naturally respond to drift in the pathogen population. Therefore T-Cell cross-regulation defined by the 4 reaction rules and co-presentation of features in APC can be seen as an effective general principle of collective classification available to populations of cells. Clearly, there is still much to do to improve the model. For biomedical literature mining applications, we need to test it with more sophisticated features (as top classifiers in the field do). For our goal of understanding T-Cell cross-regulation in the IS, we need to understand better how memory is sustained in the collective cellular dynamics; for instance, how to sustain regulatory T-Cells, which keep memory of self, in the dynamics even in the presence of very unbalanced scenarios where there are many more nonself instances.

References

[1]Alaa Abi-Haidar, Jasleen Kaur, Ana Maguitman, Predrag Radivojac, Andreas Retchsteiner, Karin Verspoor, Zhiping Wang, and Luis M. Rocha. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. *Genome Biology*, page 9(Suppl 2):S11, 2008.

[2]Alaa Abi-Haidar and Luis M. Rocha. Adaptive spam detection inspired by a cross-regulation model of immune dynamics: A study of concept drift. In *Artificial Immune Systems (Proc. ICARIS)*, volume 5132 of *LNCS*, pages 36–47, 2008.

[3]Alaa Abi-Haidar and Luis M. Rocha. Adaptive spam detection inspired by the immune system. In S. Bullock, J. Noble, R. A. Watson, and M. A. Bedau, editors, *Artificial Life XI: 11th Int. Conf. on the Simulation and Synthesis of Living Systems*. MIT Press, 2008.

[4]Alaa Abi-Haidar and Luis M. Rocha. Biomedical article classification using an agent-based model of t-cell cross-regulation. In *ICARIS 2010: Proc. of the 8th Int. Conf. on Artificial Immune Systems*, LNCS, page In Press., 2010.

[5]J. Carneiro, K. Leon, I. Caramalho, C. van den Dool, R. Gardner, V. Oliveira, M.L. Bergman, N. Sepúlveda, T. Paixão, J. Faro, and J. Demengeot. When three is not a crowd: a crossregulation model of the dynamics and repertoire selection of regulatory cd4 t cells. *Immunological Reviews*, 216(1):48–68, 2007.

[6]James Crutchfield and Melanie Mitchell. The evolution of emergent computation. *PNAS*, 92(23), 1995.

[7]L.N. De Castro and J. Timmis. *Artificial immune systems: a new computational intelligence approach*. Springer Verlag, 2002.

[8]R. Feldman and J. Sanger. *The Text Mining Handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2006.

[9]Tomás Helikar, John Konvalina, Jack Heidel, and Jim A Rogers. Emergent decision-making in biological signal transduction networks. *Proc Natl Acad Sci U S A*, 105(6):1913–1918, Feb 2008.

[10]William Hersh, Ravi Teja Bhupatiraju, and Sarah Corley. Enhancing access to the bibliome: the trec genomics track. *Medinfo*, 11(Pt 2):773–777, 2004.

[11]Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1, 2005.

[12]S.A. Hofmeyr. An Interpretative Introduction to the Immune System. *Design Principles for the Immune System and Other Distributed Autonomous Systems*, 2001.

[13]L. Hunter and K.B. Cohen. Biomedical language processing: What's beyond pubmed? *Molecular Cell*, 21(5):589–594, 2006.

[14]L. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7(2):119–129, Feb 2006.

[15]T. Joachims. *Learning to classify text using support vector machines: methods, theory, and algorithms*. Kluwer Academic Publishers, 2002.

[16]A. Kolchinsky, A. Abi-Haidar, J. Kaur, A. Hamed, and L. M. Rocha. Classification of protein-protein interaction documents using text and citation network features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.*, page In Press, 2010.

[17]M Krallinger. The biocreative ii. 5 challenge overview. In *Proc. the BioCreative II.5 Workshop 2009 on Digital Annotations*, page 19, 2009.

[18]Martin Krallinger and Alfonso Valencia. Evaluating the detection and ranking of protein interaction relevant articles: the biocreative challenge interaction article sub-task (ias). In *Proc. 2nd Biocreative Challenge Evaluation Workshop*, 2007.

[19]V. Metsis, I. Androutsopoulos, and G. Paliouras. Spam Filtering with Naive Bayes–Which Naive Bayes? *Third Conf. on Email and Anti-Spam (CEAS)*, 2006.

[20]Melanie Mitchell. Complex systems: Network thinking. *Artificial Intelligence*, 170(18):1194–1212, 2006.

[21]David Peak, Jevin D. West, Susanna M. Messinger, and Keith A. Mott. Evidence for complex, collective dynamics and distributed emergent computation in plants. *PNAS*, 101(4):918–922, 2004.

[22]MF Porter. An algorithm for suffix stripping. *Program*, 13(3):130–137, 1980.

[23]Stephen C. Pratt. Quorum sensing by encounter rates in the ant temnothorax albipennis. *Behav. Ecol.*, 16(2):488–496, 2005.

[24]L.M. Rocha and W. Hordijk. Material representations: From the genetic code to the evolution of cellular automata. *Artificial Life*, 11(1-2):189–214, 2005.

[25]L.A. Segel and I. Cohen. *Design Principles for the Immune System and Other Distributed Autonomous Systems*. Oxford University Press, 2001.

[26]Nuno H. Sepulveda. *How is the T-cell repertoire shaped*. PhD thesis, Instituto Gulbenkian de Ciencia, 2009.

[27]Cosma Shalizi, Rob Haslinger, Jean-Baptiste Rouquier, Kristina Klinkner, and Cristopher Moore. Automatic filters for the detection of coherent structure in spatiotemporal systems. *Phys.Rev.E*, 73, 2006.

[28]Hagit Shatkay and Ronen Feldman. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–856, 2003.

[29]M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021, 2006.

[30]J. Timmis. Artificial immune systems today and tomorrow. *Natural Computing*, 6(1):1–18, 2007.

[31]Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department Trinity College Dublin*, 4(C):200415, 2004.

[32]J. Twycross and S. Cayzer. An immune system approach to document classification. *Master's thesis, COGS, University of Sussex, UK*, 2002.

[33]Matthew Walters and Vanessa Sperandio. Quorum sensing in escherichia coli and salmonella. *Int. Journal of Medical Microbiology*, 296(2-3):125 – 131, 2006.