1

Induction of Fuzzy Rules by Means of Artificial Immune Systems in Bioinformatics

Filippo Menolascina[1], Vitoantonio Bevilacqua[1], Mariadele Zarrilli[1], and Giuseppe Mastronardi[1]

Polytechnic of Bari, Via E. Orabona 4, 70125 Bari, Italy f.menolascina@ieee.org

Summary. Fuzzy Rule Induction (FRI) is one of the main areas of research in the field of computational intelligence. Recently FRI has been successfully employed in the field of data mining in bioinformatics [34, 38]. Thanks to its flexibility and potentialities FRI allowed researchers to extract rules that can be easily modeled in natural language and submitted to experts in the field that can validate their accuracy or consistency. The process of FRI can result to be highly complex from a computational complexity point of view and, for this reason, several alternative approaches to accomplish this process have been proposed ranging from iterative and simultaneous algorithms [22] to Genetic Algorithms and Ant Colony Optimization based approaches [22]. In this chapter we will focus on a specific application of type-1 (T1) and type-2(T2) fuzzy systems to data mining in bioinformatics in which FRI is carried out using a novel and promising computational paradigm, namely Artificial Immune Systems (AIS). In order to provide the reader with the necessary theoretical background we will go through a brief introduction to the fields of AIS and T2 Fuzzy Systems, then we will set up the scientific context and describe applications of these concepts to real world cases. Conclusions and cues for future work in this fascinating field will be provided in the end.

1.1 Artificial Immune Systems

Artificial Immune Systems (AIS) represent one of the most recent and promising approaches in the branch of bio-inspired techniques. Although this open field of research is still in its infancy, several relevant results have been achieved by using the AIS paradigm in demanding tasks such as the those coming from computational biology and biochemistry. Artificial immune systems (AIS) can be defined as computational systems inspired by theoretical immunology, observed immune functions, principles and mechanisms in order to solve problems. Their development and application domains follow those of soft computing paradigms such as artificial neural networks (ANN), evolutionary algorithms (EA) and fuzzy systems (FS). Soft computing was the term coined to address a new trend of co-existence and integration that reflects a high degree

of interaction among several computational intelligence approaches like artificial neural network, evolutionary algorithms and fuzzy systems. The idea of integrating different computational intelligence paradigms in order to create hybrids combining the strengths of different approaches is not new. Following the previous concepts when in 2002 de Castro and Timmis introduced AIS as a new soft computing paradigm they gave birth to a new challenge to have a great potential to interact the new born technique with others. Strictly speaking evolution and immune system are biologically closely related to each other. In fact the process of natural selection can be seen to act the immune system at two levels. First recall that lymphocytes multiply based on their affinity with a pathogen. The higher affinity lymphocytes are selected to reproduce, a process usually named immune microevolution. The mechanism of immune microevolution is very important. The clonal selection principle presupposes that a very large number of *B-cells* containing antigenic receptors is constantly circulating throughout the organism. The great diversity of this repertoire is a result of the random genetic recombination of gene fragments from different libraries plus the random insertion of gene sequences during cell development. This availability of different solutions guarantees that at least one cell will produce an antibody capable of recognizing, thus binding with, any antigen that invades the organism. The antigen-antibody binding stimulates the production of clones of the selected cells, where successive generations result in exponential growth of the selected antibody type. Some of these antibodies remain in circulation even after the immune response ceases, constituting a sort of immune memory. Other cells differentiate in plasma cells, producing antibodies in high rates. Finally during reproduction, some clones suffer an affinity maturation process, where somatic mutations are inserted with high rates (hypermutation) and, combined with a strong selective mechanism, improve the capability (*Ag-Ab* affinity and clone size) of these antibodies to recognize and respond to the selective antigens. Secondly, there is surely an immune contribution to natural selection, which acts by allowing the multiplication of those people carrying genes that are most able to provide maximal defense against infectious diseases coupled with minimal risk of autoimmune diseases. At this time the majority of the immune algorithms currently developed have an evolutionary type of learning of embodied process and several techniques from one strategy have been used to enhance another. **I-PAES presented and discussed in the Section ?? is an example of hybridization between a particular class of evolutionary algorithms called multi-objective and immune inspired operators namely cloning and hypermutaion.**

The success of the AIS paradigm is based on two key properties of its theoretical foundations: recognition and adaptation/optimization. When an animal is exposed to an antigen, some subpopulation of its bone marrow derived cells (*B lymphocytes*) respond by producing antibodies (*Ab*). Each cell secretes a single type of antibody, which is relatively specific for the antigen. By binding to these antibodies (*cell receptors*), and with a second signal from accessory cells, such as the T-helper cell, the antigen stimulates the *B cell* to

proliferate (divide) and mature into terminal (non-dividing) antibody secreting cells, called plasma cells. The process of cell division (mitosis) generates a clone, i.e., a cell or set of cells that are the progenies of a single cell. While plasma cells are the most active antibody secretors, large B lymphocytes, which divide rapidly, also secrete antibodies, albeit at a lower rate. On the other hand, T cells play a central role in the regulation of the *B cell* response and are preeminent in cell mediated immune responses, but will not be explicitly accounted for the development of our model. Lymphocytes, in addition to proliferating and/or differentiating into plasma cells, can differentiate into long-lived B memory cells. Memory cells circulate through the blood, lymph and tissues, and when exposed to a second antigenic stimulus commence to differentiate into large lymphocytes capable of producing high affinity antibodies, pre-selected for the specific antigen that had stimulated the primary response. Fig 1.1 depicts the clonal selection principle.

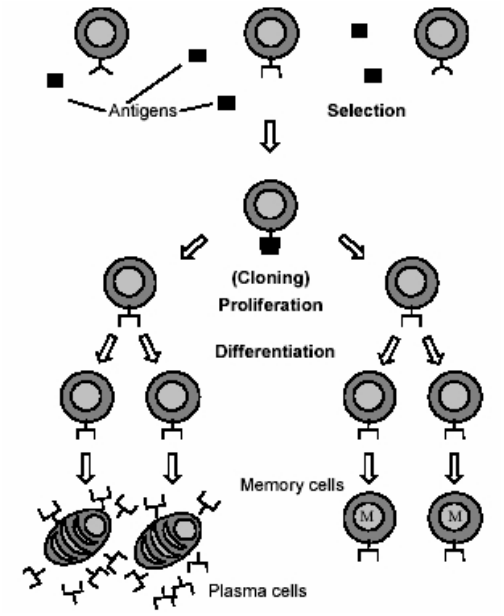

Fig. 1.1. Clonal selection principle in natural immune systems.

The clonal selection and affinity maturation principles are used to explain how the immune system reacts to pathogens and how it improves its capability of recognizing and eliminating pathogens [14]. In a simple form, clonal

selection states that when a pathogen invades the organism, a number of immune cells that recognize these pathogens will proliferate; some of them will become effector cells, while others will be maintained as memory cells. The effector cells secrete antibodies in large numbers, and the memory cells have long life spans so as to act faster and more effectively in future exposures to the same or a similar pathogen. During the cellular reproduction, the cells suffer somatic mutations with high rates and, together with a selective force, the higher affinity cells in relation to the invading pathogen differentiate into memory cells. This whole process of somatic mutation plus selection is known as affinity maturation. To a reader familiar with evolutionary biology, these two processes of clonal selection and affinity maturation are much akin to the (macro-)evolution of species. There are a few basic differences however, between these immune processes and the evolution of species. Within the immune system, somatic cells reproduce in an asexual form (there is no crossover of genetic material during cell mitosis), the mutation suffered by an immune cell is proportional to its affinity with the selective pathogen (the higher the affinity, the smaller the mutation rate), and the number of progenies of each cell is also proportional to its affinity with the selective pathogen (the higher the affinity, the higher the number of progenies). Evolution in the immune system occurs within the organism and, thus it can be viewed as a micro-evolutionary process. As we know, in fact, immunology suggests that the natural Immune System (IS) has to assure recognition of each potentially dangerous molecule or substance, generically called antigen (Ag), by antibodies (Ab). The IS first recognizes an antigen as "dangerous" or external invaders and then adapts (by affinity maturation) its response to eliminate the threat. To detect an antigen, the IS activates a recognition process. In vertebrate organisms, this task is accomplished by the complex machinery made by cellular interactions and molecular productions. The main features of the clonal selection theory that will be explored in this chapter are [14]]:

- Proliferation and differentiation on stimulation of cells with antigens;
- Generation of new random genetic changes, subsequently expressed as diverse antibody patterns, by a form of accelerated somatic mutation (a process called affinity maturation);
- Elimination of newly differentiated lymphocytes carrying low affinity antigenic receptors.

To illustrate the adaptive immune learning mechanism, consider that an antigen *Ag1* is introduced at time zero and it finds a few specific antibodies within the animal (see Fig. 1.2). After a lag phase, the antibody against antigen *Ag1* appears and its concentration rises up to a certain level, and then starts to decline (*primary response*). When another antigen *Ag2* is introduced, no antibody is present, showing the specificity of the antibody response [14]. On the other hand, one important characteristic of the immune memory is that it is associative: *B cells* adapted to a certain type of antigen *Ag1* presents a faster and more efficient secondary response not only to *Ag1*, but also to any

structurally related antigen *Ag1* + *Ag2*. This phenomenon is called immunological cross-reaction, or cross-reactive response. This associative memory is contained in the process of vaccination and is called *generalization capability*, or simply generalization, in other artificial intelligence fields, like neural networks [14].

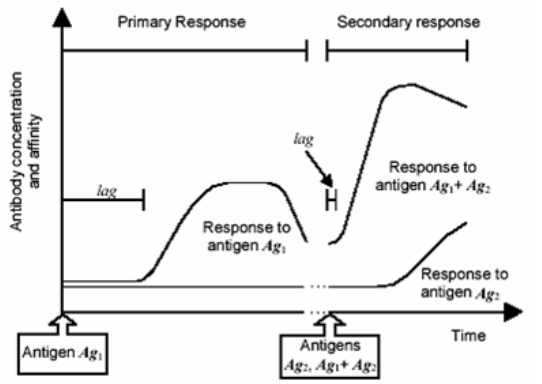

Fig. 1.2. Immune response plotted as antibody concentration over time.

Receptor editing offers the ability to escape from local optima on an affinity landscape. Fig. 1.3 illustrates this idea by considering all possible antigen-binding sites depicted in the x-axis, with the most similar ones adjacent to each other. The Ag-Ab affinity is shown on the y-axis. If we consider a particular antibody (*Ab1*) selected during a primary response, then point mutations allow the immune system to explore local areas around Ab1 by making small steps towards an antibody with higher affinity, leading to a local optimum (*Ab1**). Because mutations with lower affinity are lost, the antibodies can not go down the hill. Receptor editing allows an antibody to take large steps through the landscape, landing in a locale where the affinity might be lower (*Ab2*). However, occasionally the leap will lead to an antibody on the side of a hill where the climbing region is more promising (*Ab3*), reaching the global optimum. From this locale, point mutations can drive the antibody to the top of the hill (*Ab3**). In conclusion, point mutations are good for exploring local regions, while editing may rescue immune responses stuck on unsatisfactory local optima.

Computational immunology is the research field that attempts to reproduce *in silico* the behavior of the natural IS. From this approach, the new field of Artificial Immune Systems (AIS) attempts to use theories, principles,

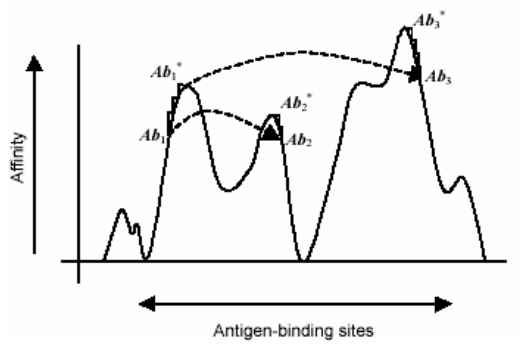

Fig. 1.3. Antibody affinity as function of the specific antigen binding site.

and concepts of modern immunology to design immunity-based system applications in science and engineering [14]. AIS are adaptive systems in which learning takes place using evolutionary mechanisms similar to biological evolution. These different research areas are tied together: the more we learn from *in silico* modeling of natural systems, the better we are able to exploit ideas for computer science and engineering applications.

Thus one wants, first, to understand the dynamics of such complex behavior when they face antigenic attack, and second, one wishes to develop new algorithms that mimic the natural IS under study. Thus the final system may have a good ability to solve computational problems otherwise difficult to be solved by conventional specialized algorithms. The computational and predictive power of AIS offers researchers a promising approach for trying to solve well known and challenging problems like knowledge discovery from huge biological databases (e.g. coming from high throughput platforms) as well as protein folding or function prediction and multiple sequence alignment.

1.2 Type-2 Fuzzy Systems

Type-1 fuzzy sets are characterized by crisp grades of the membership function however, for some reasons, it could be very hard to find the exact membership function for a given fuzzy set and, as a consequence, it is hard to determine an exact membership level for each linguistic variable of the defined universe. It is then necessary to further fuzzify the knowledge base and this is possible only by using fuzzy sets that are fuzzy themselves [33]. Type-2 fuzzy sets are characterized by membership grades that are represented by values in the interval $[0, 1]$. At each value of the primary variable the membership is a func-

tion (and not just a point value), also called secondary membership function, whose domain the primary membership, is in the interval $[0, 1]$ and whose range secondary grades may also be in $[0, 1]$. We can assume, then, that the membership function of a Type-2 Fuzzy Set is three dimensional (see Fig. 1.4). This is a real plus to the theory of Type-1 Fuzzy Sets since it should be evident that such sets are useful in circumstances where uncertainty prevents us from obtaining a sufficiently clear knowledge on the process. As an example

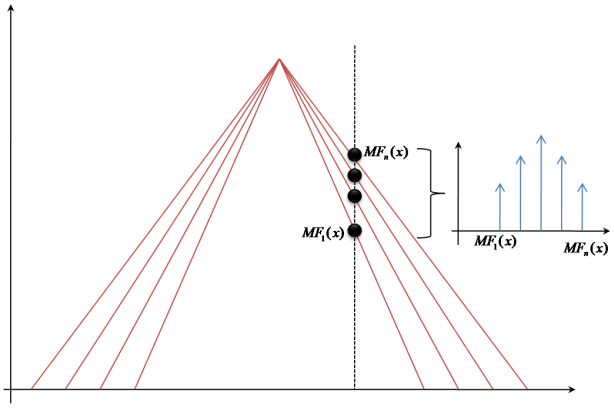

Fig. 1.4. Triangular MFs for a T2FS.

we consider the the well known case of *eye contact* [33]. Let us put the eye contact on a scale of values that goes from 0 to 10. One can say that a term of this universe can be '*some eye contact*'. Suppose we interviewed 100 men and women asking them to set boundaries for this measure on the $0 - -10$ scale. It is unlikely we will get the same results from all of them because words mean different things to different people and this situation is rather frequent even in specialized field like medicine. One approach to using the 100 sets of two end-points is to average the end-point data and to use the average values for the interval associated with some eye contact. We could then construct a triangular (other shapes could be used) membership function (MF), $MF(x)$, whose base end-points (on the x-axis) are at the two average values and whose apex is midway between the two end-points. This type-1 triangular MF can be displayed in two-dimensions. Unfortunately, it has completely ignored the **uncertainties** associated with the two end-points. A second approach is to make use of the average values and the standard deviations for the two end-points. By doing this we are blurring the location of the two endpoints along the x-axis. Now locate triangles so that their base end-points can be anywhere in the intervals along the x-axis associated with the blurred average endpoints. Doing this leads to a continuum of triangular MFs sitting on the x-axis, e.g. picture a whole bunch of triangles all having the same apex point but different

base points, as in Fig. 1.4. For purposes of this discussion, suppose there are exactly 100 (N) such triangles. Then at each value of x, there can be up to N MF values, $MF_1(x), MF_2(x), \cdots, MF_N(x)$. Let us assign a weight to each of the possible MF values, say $w_{x1}, w_{x2}, \cdots, w_{xN}$ (see Fig. 1.4). We can think of these weights as the possibilities associated with each triangle at this value of x. At each x, the MF is itself a function -the secondary MF- ($MF_i(x)$, w_{xi}), where $i = 1, \cdots, N$. Consequently, the resulting type-2 MF is three-dimensional. For more details on T2 Fuzzy Sets the reader is referred to [32] and [28]. From the description we have provided it should be evident that uncertainty handling is a key point of these approaches. Uncertainty plays a major role in bio-medicine and biomedical science since most of the research carried out in this field is experimental and is affected by measurements associated errors. This is why we recently proposed a novel approach to data mining in bioinformatics that tries to face these problems using a coherent algorithmic model. In the next paragraphs we will describe type-1 and type-2 based fuzzy systems for rule inference from bioinformatic databases. We will provide a detailed description of both starting from the type-1.

1.3 Fuzzy-Immunity based Data Mining Systems in Bioinformatics

Recent advances in active fields of research like biotechnology and electronics allowed biomedical research to make a significant step forward in the acquisition of fundamental tools for the elucidation of complex bio-processes like the ones behind cancer or Alzheimer disease. The advent of High-Throughput (HT) platforms has revolutionized the way researchers working in life sciences thought at their role in experiments. HT devices allowed researchers to concentrate on more important tasks like experimental design and results interpretation at the same time allowing him to ignore the hundreds when not thousands of repeats of the same protocols for the different patients or mRNA sequences for instance. Microarrays are, probably, one of the most evident examples of this change of perspectives: gene expression evaluation for a panel of even only a few tens of genes took several days to be completed before their introduction, now we are able to obtain gene expression level for thousands of genes in the time of an overnight hybridization. Together with expression microarrays we can mention copy number monitoring microarrays (commonly referred to as aCGH technique), High-Throughput Sequencers, and Mass Spectrometers. In the next sections we will go through a brief analysis of the main open problems in bioinformatics and will discuss about how they can be addressed using immunity based data mining algorithms. A short introduction on data mining principles and potentialities is given in order to help unexperienced readers understanding concepts behind statements.

1.3.1 Data Bases and Information Retrieval in Biology

Devices coming from the integration of experiences gained in diverse fields like physics, chemistry, biology and engineering, in this way helped researchers in boosting their work and in quickly obtaining results of their experiments. The capabilities of these different kinds of approach pushed the interest for the establishment of data repositories for newly generated results. Data-bases entered the world of biology. Larger and larger amounts of data started to fill public databases (leaving apart literature databases which, of course, need a separated analysis) giving rise to what we can rename "Moore's law in biology" [46] (that just like the original Moore's law in electronics, models future progress in biotechnology [18]). However the main advantages provided by novel devices soon revealed to be their main weak point. The availability of large amount of data as results did not yield of information drawn from these data; this phenomenon characterized both early and more recent years in life sciences research bringing to the so-called "gap". Roughly speaking, researchers indicate, with this term, an estimate of the difference between the amount of available data and the amount of these data that have been sufficiently interpreted [24]. In the recent years we have observed a worrying widening in this gap: this means that we are making quite large investments with a ROI (return on investments) that still keeps low. In order to maximize the information yield of each experiment several alternative solutions have been proposed being probably data warehousing the most successful. Data warehouses are the natural evolution of data bases; described for the first time by William Immon [53]. They are integrated, subject-oriented, time-variant and non-volatile data collection processes implemented with the precise aim to build a unique decision support system. The distinction between data bases and data warehouses is clear: as advanced data bases, data warehouse provide data analysis functionalities that ease the process of knowledge extraction from highly dense data repository. In this context significant experiences like the GEO (Gene Expression for Omnibus [4]), SMD (Stanford Microarray Database [17]) and ArrayExpress [7] have been gained. It is evident that data warehouse can greatly help researchers in reducing the gap by providing a valuable aid in filling the last real hole in experimental processes automation: results interpretation.

1.3.2 Mining the Data: Converting Data to Knowledge

Data mining, also known as Knowledge Discovery in Data-bases (KDD), has been defined as *"The nontrivial extraction of implicit, previously unknown, and potentially useful information from data"* [20] (a more practical definition of data mining will be given in the following section); it uses machine learning, statistical and visualization techniques to discover and present knowledge in a form easily comprehensible to humans. Data mining grew at the border line among statistics, computer science and artificial intelligence and soon

became a golden tool to solve problems ranging from Customer Relationship Management (CRM [31]) to Decision Making Support in medicine [47]. Data mining in bioinformatics, then, can be considered as a useful tool for modeling complex processes allowing researchers speeding the pace towards treatments for diseases like cancer: for instance several works have successfully tried to exploit the potentialities of rule induction systems in breast cancer associated survival [30, 5] and cancer evolution modeling [35]. It can be argued that data mining was born from several diverse disciplines, in the effort of overcoming intrinsic limitations of the single approaches. It is particularly evident if we compare the expressive power of typical statistical inference approaches and propositional or first order logic on the other hand. Huge efforts have been spent, in the recent past, in order to speed up one of the central tasks in current research in bioinformatics, that is, the transformation process that converts *data* in *knowledge* passing through *information* [43]. Data mining software, then, became more and more common: researchers soon realized the valuable aid algorithms could have given to their researchers and the amount of paper describing algorithms for information extraction grew faster and faster [15, 44, 55]. Comprehensive software tools for data mining purposes are currently largely used in bioinformatics and include both open-source and proprietary solutions. Among commercial packages we can list SPSS, SAS, Clementine and E-Miner. Open source tools are well represented by:

- Weka [54]
- Rapid Miner (formerly YALE) [40]
- Orange [39]

In particular Weka has gained a relevant success in the field of data mining due to its flexibility and versatility. Thanks to these characteristics Weka has been customized and redistributed in several different flavors (BioWeka [23] devoted to biological sequences mining and Weka4WS [48], the GRID-enable Weka implementation). Due to a simple but efficient modular organization Weka allowed third-party developers to add functionalities to the core package. It is the case of "Weka Classification Algorithms" project managed by Jason Brownlee who has implemented several bio-inspired [8, 9, 29] data mining algorithm in a customized version of Weka Classification Algorithms [1]. One of the most interesting aspects of this implementation consists in the presence of a wide variety of Artificial Immune System based data mining algorithms. Both the *black* and *white box* flavors are represented in the set of proposed algorithms. The distinction between black and white box algorithms will be described in the following paragraph, however it can be argued that white box approaches provide the user with tools to easily interpret the way it reached a certain results, on the contrary to what happens with black box algorithms (think at how complex is the interpretation of neural network predictions and

[1] http://sourceforge.net/projects/wekaclassalgos

how simple is interpreting rules induced from a dataset). Among black box Immunity based algorithm we can mention:

Clonalg

The Clonal Selection Algorithm, originally called CSA in [12], and renamed to CLONALG in [13] is said to be inspired by the following elements of the clonal selection theory:

- Maintenance of a specific memory set
- Selection and cloning of most stimulated antibodies
- Death of non-stimulated antibodies
- Affinity maturation (mutation)
- Re-selection of clones proportional to affinity with antigen
- Generation and maintenance of diversity

The goal of the algorithm is to develop a memory pool of antibodies that represents a solution to an engineering problem. In this case, an antibody represents an element of a solution or a single solution to the problem, and an antigen represents an element or evaluation of the problem space.

CSCA

The Clonal Selection Classifier Algorithm is an evolution of the concept behind Clonalg since it tries to maximize classification accuracy and minimize misclassification accuracy still using clonal selection paradigms.

Immunos

The Immunos [10] algorithm has been mentioned a number of times in AIS literature [49, 25, 50]. It is claimed as being one of the first immune-inspired classification systems. Immunos tries to mimic in a very precise way the mechanisms underlying immune response to antigen attacks and this has led to a quite complex classification system still under discussion.

AIRS

The Artificial Immune Recognition System [52] algorithm was one of the first AIS technique designed specifically and applied to classification problems. After an initialization phase the algorithm cycles through each antigen (record in the dataset) in order to select best fitting memory cells through a powerful resource competition stage.

On the other hand white box AIS based paradigms can be found in:

- IFRAIS
- AIS based rule induction with boosting

These approaches will be discussed in greater depth in the next section.

1.3.3 Algorithmic Approaches to Data-Mining in Biology

As previously stated data mining is an interdisciplinary research field, involving areas such as machine learning, statistics, databases, expert systems and data visualization, whose main goal is to extract knowledge (or patterns) from real-world data sets [19, 54]. This section focuses on the classification (supervised learning) task of data mining. In essence, the goal of the classification task is to assign each example (data instance or record) to a class, out of a predefined set of classes, based on the values of attributes describing that example. In the context of bioinformatics an example could be, for instance, a protein; the classes could be protein functions; and the attributes describing the protein could be, say, physico-chemical properties of the amino acids composing the protein. It is important that the attributes describing an example are relevant for predicting its class. Hence, it would be a mistake to use a clearly irrelevant attribute, say the name of the patient, as an attribute to predict whether or not a patient will get a certain disease. In bioinformatics, ideally, the classification model should satisfy two requirements. First, it should have a high predictive accuracy, or generalization ability, correctly predicting the class of new examples unseen during the training of the system. Second, it should be comprehensible to users (biologists), so that it can be interpreted in the context of existing biological knowledge and potentially further validated through new biological experiments. Concerning the issue of comprehensibility of the classification model discovered from the data, it should be noted that some classification algorithms are designed to maximize only predictive accuracy, representing the classification model in a way that cannot be understood by the user - therefore ignoring the comprehensibility requirement. Typical examples of algorithms in this category are support vector machines [51] and neural networks [26]. In this case the classification model is a "black box", which does not give the user any insight about the data or explanations about the classification of new examples. In contrast, some classification algorithms use a representation which is comprehensible to the user, therefore returning "knowledge" to the user. In this section we focus on one popular kind of comprehensible representation, namely IF-THEN classification rules, and algorithms that use this kind of representation are called rule induction algorithms [21]. In rule induction algorithms the classification model is represented by a set of classification rules. These rules are of the form: "IF antecedent THEN consequent", where the antecedent represents a conjunction of conditions and the consequent represents the class predicted for all examples (data instances, records) that satisfy the antecedent. Each condition in the antecedent typically specifies a value or a range of values for a given attribute of the data being mined - e.g., "gender = female", "age < 21".

The first AIS for rule induction in the classification task of data mining was proposed in [3], and named IFRAIS (Induction of Fuzzy Rules with an Artificial Immune System). IFRAIS as well as IFRAIS2 will be discussed in

the next section. In this section we just highlight that this system discovers fuzzy classification rules. Fuzzy rules are in general more natural and more comprehensible to human beings than crisp rules, and the fuzzy rule representation also has the ability of coping well with the uncertainties frequently associated with data in biological databases [41]. Other algorithms based on AIS for rule induction are discussed in detail in [1, 11].

Artificial Immune Systems in Bio-medical Data Mining: IFRAIS and IFRAIS2

As mentioned earlier, IFRAIS as well as its Type-2 FS counterpart are AIS that designed to discover fuzzy classification rules from data. From now on we will refer to IFRAIS as the main ideas behind it remained unchanged in IFRAIS 2 unless otherwise stated.

Recall that the rule antecedent is formed by a conjunction of conditions. Each attribute can be either continuous (real-valued, e.g. the molecular weight of a protein) or categorical (nominal, e.g. the name of a species), as usual in data mining. Categorical attributes are inherently crisp, but continuous attributes are fuzzified by using a set of three linguistic terms (low, medium, high). Hence, in the case of continuous attributes, IFRAIS discovers fuzzy rules having conditions such as: "molecular weight is large". IFRAIS discovers fuzzy classification rules by using the sequential covering approach for rule induction algorithms [54]. This is an iterative process which starts with an empty set of rules and the full training set (containing all training examples). At each iteration, IFRAIS is run to discover the best possible classification rule for the current training set, which is then added to the set of discovered rules. Then the examples correctly covered by the discovered rule (i.e. the examples satisfying the antecedent of that rule and having the class predicted by the rule) are removed from the training set, so that a smaller training set is available for the next iteration. This process is repeated until all (or a large part of the) training examples have been covered by the discovered rules. In order to discover classification rules, IFRAIS uses essentially clonal selection and hypermutation procedures. The basic ideas are as follows. Each antibody corresponds to a candidate fuzzy classification rule. During an IFRAIS run, the better the classification accuracy of an antibody, the more likely it is to be selected for cloning. In addition, once an antibody is cloned, the rate of mutation of a clone is inversely proportional to the classification accuracy of the antibody. Hence, the principles of clonal selection and hypermutation drive the evolution of the population of antibody towards better and better classification rules. In IFRAIS2, on the other hand, we are interested in evolving terms with MF that are fuzzy themselves so we handle them using a predefined number of MF for each term and we evolve vectors of these features in place of single attributes (e.g. vectors of mean values or cut-values of MF in place of a single mean or cut-value). In [34, 38] IFRAIS was successfully employed to discover fuzzy classification rules for female breast cancer familiarity

Dataset	IFRAIS2	IFRAIS
CRX	74.82%	69.65%
Monk	91.08%	87.26%
Wine	85.12%	83.26%
Breast Cancer aCGH [38]	87.86%	78.65%
Breast Cancer Gene Expression [34]	87.45%	82.73%

Table 1.1. Results of IFRAIS and IFRAIS 2 on several data sets of varying complexity.

profiling. IFRAIS' results were validated using statistical driven approaches using Gene Ontology through GO Miner [55]. Competitive results obtained by IFRAIS and IFRAIS 2 (Tab. 1.1 show a comparative study of the results of both IFRAIS and IFRAIS 2 on benchmark, as well as, on real world data sets) seem to encourage new efforts in this field. A biological interpretation of the results carried out using Gene Ontology is currently under investigation.

1.3.4 Application of AIS based Data Mining in Bioinformatics

As we previously stated several examples of application of Fuzzy-AIS based data mining systems in bioinformatics can be retrieved in literature. Fuzzy and Artificial Immune Systems-derived algorithms have been employed in familiarity profiling [34], prognosis prediction [35] and estrogen receptor modeling [36] in breast cancer. For a brief comparative overview of the performances of these kinds of systems in the context of aCGH data analysis the reader is referred to [37]. For the AIS counterpart we should note that previously de Castro and colleagues focused on the use of Hierarchical Artificial Immune Network paradigm for the problem of gene expression clustering [6, 27] and for rearrangement study of gene expression [16]. Research currently being carried out by Alves and colleagues is mainly focused on the application of a multi-label Fuzzy-AIS based data mining system to the problem of protein function prediction [2].

1.4 Conclusions and Open Questions

In this chapter we have analyzed some applications of Fuzzy-Artificial Immune System based algorithms in bioinformatics. Of course this is only a partial outlook on the world of Fuzzy-AIS based approaches: interested readers can check references in order to obtain more detailed information about specific aspects of the proposed topics. Furthermore, given their infancy, Fuzzy-AIS are currently undergoing very fast changes resulting in a very dynamical field of research where tens of novel and promising projects are proposed in the time of some months. These aspects forced the authors to select a set of significant experiences to be used as examples of how the algorithms described herein

can be successfully used in the field of bioinformatics. After these necessary statements some conclusions. In this chapter we have learned how novel bio-inspired computational intelligence paradigms can be used in very diverse field of research in bioinformatics. As previously stated Fuzzy-AIS are considered a novel paradigm but they have been already able to reach significant results in highly complex context like Knowledge Discovery in Data bases and Gene signature Prediction. Even if fuzzy-immune-inspired algorithms have been successfully employed in several diverse problems, there are still some strategic fields of research in which solutions seem to be far from being reached, just to name few:

- Gene networks inference;
- Disease profiling and evolution modeling.
- Diagnostic and prognostic disease signature development

These are only some of the most active areas of Fuzzy-AIS based research in bioinformatics. From a theoretical point of view it should be noted that some areas like *hybrid systems* in this field have been exploited with a limited systematic approach in bioinformatics: these areas deserve a comprehensive analytic approach. Readers interested in these promising aspects of the Fuzzy-AIS research in bioinformatics can find useful information in [42, 45].

References

1. B. Alatas and E. Akin, *Mining fuzzy classification rules using an artificial immune system with boosting*, Proceedings of the Conference on Advances in Databases and Information Systems, LNCS 3631, Springer, 2005, pp. 283–293.
2. R. T. Alves, *An artificial immune system to hierarchical multi-label classification for predicting protein function*, Ph.D. Qualifying Exam 42, Federal University of Technology of Paran -UTFPR, Curitiba, Brazil, 2007.
3. R.T. Alves, M.R. Delgado, H.S. Lopes, and A.A. Freitas, *An artificial immune system for fuzzy-rule induction in data mining*, Parallel Problem Solving from Nature, LNCS, vol. 3242, 2004, pp. 1011–1020.
4. T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, *NCBI GEO: mining tens of millions of expression profiles–database and tools update*, Nucleic Acids Res. **35** (2007), D760–D765.
5. V. Bevilacqua, P. Chiarappa, G. Mastronardi, F. Menolascina, A. Paradiso, and S. Tommasi, *Identification of tumour evolution patterns by means of inductive logic programming*, Journal - Genomics Proteomics and Bioinformatics (2007).
6. G.B. Bezerra, G.M.A. Canado, M. Menossi, L.N. de Castro, and F.J. Von Zuben, *Recent advances in gene expression data clustering: a case study with comparative results*, Genet. Mol. Res. **4** (2005), no. 3, 514–524.
7. A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, Lara G.G., A. Oezcimen, P. Rocca-Serra, and S.A. Sansone, *ArrayExpress–a public repository for microarray gene expression data at the EBI*, Nucleic Acids Res. **31** (2003), no. 1, 68–71.

8. J. Brownlee, *Artificial immune recognition system (AIRS) - A review and analysis*, Tech. Report ID: 1-01, Centre for Intelligent Systems and Complex Processes, Faculty of Information and Communication Technologies, Swinburne University of Technology, Victoria, Australia, 2005.
9. ———, *Clonal selection theory and CLONALG - the clonal selection classification algorithm (CSCA)*, Tech. Report 2-01, Centre for Intelligent Systems and Complex Processes, Faculty of Information and Communication Technologies, Swinburne University of Technology, Victoria, Australia, 2005.
10. J.H. Carter, *The immune system as a model for classification and pattern recognition*, Journal of the American Informatics Association **7** (2000), 28–41.
11. P. A. D. Castro, G. P. Coelho, M. F. Caetano, and F. J. Von Zuben, *Designing ensembles of fuzzy classification systems: An immune-inspired approach*, International Conference on Artificial Immune Systems, LNCS 3627, Springer, 2005, pp. 469–482.
12. L. N. de Castro and F. J. Von Zuben, *The clonal selection algorithm with engineering applications*, Genetic and Evolutionary Computation Conference, Workshop on Artificial Immune Systems and Their Applications, 2000, pp. 36–37.
13. ———, *Learning and optimization using the clonal selection principle*, IEEE Transactions on Evolutionary Computation **6** (2002), 239–251.
14. L.N. de Castro and J. Timmis, *Artificial immune systems: A new computational approach*, Springer, 2002.
15. J.G. de la Nava, D.F. Santaella, J.C. Alba, J.M. Carazo, O. Trelles, and A. Pascual-Montano, *Engene: The processing and exploratory analysis of gene expression data*, Bioinformatics (2003), 657–658.
16. J.S. de Sousa, L. de C. T. Gomes, G.B. Bezerra, L.N. de Castro, and F.J. Von Zuben, *An immune-evolutionary algorithm for multiple rearrangements of gene expression data*, Genetic Programming and Evolvable Machines **5** (2004), no. 2, 157–179.
17. J. Demeter, C. Beauheim, J. Gollub, T Hernandez-Boussard, H. Jin, D. Maier, J.C. Matese, M. Nitzberg, F. Wymore, Z.K. Zachariah, P.O. Brown, G. Sherlock, and C.A. Ball, *The stanford microarray database: implementation of new analysis tools and open source release of software*, Nucleic Acids Res. (2007), D766–770.
18. *Life 2.0. the new science of synthetic biology is poised between hype and hope. but its time will soon come*, Economist, September 2006.
19. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*, AAAI/MIT, Cambridge, 1995.
20. W. Frawley, G. Piatetsky-Shapiro, and C. Matheus, *Knowledge discovery in databases: An overview*, AI Magazine **13** (1992), no. 3, 57–70.
21. A. A. Freitas, *Data mining and knowledge discovery with evolutionary algorithms*, Springer-Verlag, Berlin, 2002.
22. M. Galea and Q. Shen, *Iterative vs simultaneous fuzzy rule induction*, IEEE Conference on Fuzzy Systems, 2005, pp. 767–772.
23. J.E. Gewehr, M. Szugat, and R. Zimmer, *BioWeka-extending the Weka framework for bioinformatics*, Bioinformatics **23** (2007), no. 5, 651–653.
24. R. Grossman, C. Kamath, and V. Kumar, *Data mining for scientific and engineering applications*, Springer, 2001.
25. E. Hart, *Immunology as a metaphor for computational information processing: Fact or fiction?*, Ph.D. thesis, University of Edinburgh, 2002.

26. S. Haykin, *Neural networks - a comprehensive foundation*, 2nd edition ed., Prentice Hall, Upper Saddle River, 1999.
27. E.R. Hruschka, R.J.G.B. Campello, and L.N. de Castro, *Evolving clusters in gene-expression data*, Inf. Sci. **176** (2006), no. 13, 1898–1927.
28. http://www.type2fuzzylogic.org/publications/.
29. J.Brownlee, *Immunos-81 - the misunderstood artificial immune system*, Tech. Report 3-01, Centre for Intelligent Systems and Complex Processes, Faculty of Information and Communication Technologies, Swinburne University of Technology, Victoria, Australia, 2005.
30. P. Larranaga, M.J. Gallego, B. Sierra, L. Urkola, and M.J. Michelena, *Bayesian networks, rule induction and logistic regression in the prediction of the survival of women suffering from breast cancer*, Spanish Artificial Intelligence Conference, 1997, pp. 303–308.
31. D. Ledingham and D.K. Rigby, *CRM done right*, Harvard Business Review (2004).
32. J. M. Mendel, *Advances in type-2 fuzzy sets and systems*, Information Sciences **177** (2007), no. 1, 84–110.
33. J.M. Mendel and R.I. John, *Type-2 fuzzy sets made simple*, IEEE Trans. of Fuzzy Systems **10** (2002), no. 2, 117 – 127.
34. F. Menolascina, R. T. Alves, S. Tommasi, P. Chiarappa, M. Delgado, V. Bevilacqua, G. Mastronardi, A. A. Freitas, and A. Paradiso, *Fuzzy rule induction and artificial immune systems in female breast cancer familiarity profiling*, KES, LNAI, vol. 4694, Springer, 2007, pp. 830–837.
35. F. Menolascina, R.T. Alves, S. Tommasi, P. Chiarappa, M. Delgado, V. Bevilacqua, G. Mastronardi, A. A. Freitas, and A. Paradiso, *Improving female breast cancer prognosis by means of fuzzy rule induction with artificial immune systems*, Proceedings of 2007 International Conference on Life System Modeling and Simulation (Shanghai, China), 2007, pp. 1–5.
36. F. Menolascina and et al R. T. Alves, *Induction of fuzzy rules with artificial immune systems in aCGH based ER status breast cancer characterization*, Genetic and Evolutionary Computation Conference, 2007.
37. F. Menolascina, S. Tommasi, P. Chiarappa, V. Bevilacqua, G. Mastronardi, and A. Paradiso, *Data mining techniques in acgh-based breast cancer subtype profiling: an immune perspective with comparative study*, BMC Systems Biology **Suppl. 1** (2007), 70.
38. F. Menolascina, S. Tommasi, A. Paradiso, M. Cortellino, V. Bevilacqua, and G. Mastronardi, *Novel data mining techniques in aCGH based breast cancer subtypes proling: the biological perspective*, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (Honolulu. US), 2007, pp. 9–16.
39. R.S. Michalski, I. Bratko, and M. Kubat, *Machine learning and data mining: Methods and applications*, Wiley, 1998.
40. I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, *YALE: Rapid prototyping for complex data mining tasks*, 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2006, pp. 935–940.
41. W. Pedrycz and F. Gomide, *An introduction to fuzzy sets: Analysis and design*, MIT Press, Cambridge, MA, 1998.

42. K. Polat, S. Sahan, and S. Gunes, *A novel hybrid method based on artificial immune recognition system (airs) with fuzzy weighted pre-processing for thyroid disease diagnosis*, Expert Systems with Applications: An International Journal **32** (2007), no. 4, 1141–1147.
43. R. Pool and J. Esnayra, *Bioinformatics: Converting data to knowledge*, Natl Acad Press, Washington DC, 2003.
44. M. Reich, T. Liefeld, Gould. J., J. Lerner, P. Tamayo, and J.P. Mesirov, *GenePattern 2.0*, Nature Genetics **38** (2006), no. 5, 500–501.
45. S. Sahan, K. Polat, H. Kodaz, and S. Gunes, *A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis*, Computers in Biology and Medicine **37** (2007), no. 3, 415–423.
46. R. Scott.
47. M.S. Siadaty and W.A. Knaus, *Locating previously unknown patterns in data-mining results: a dual data- and knowledge-mining method*, BMC Medical Informatics and Decision Making **6** (2006), 13.
48. D. Talia, P. Trunfio, and O. Verta, *Weka4WS: a WSRF-enabled Weka toolkit for distributed data mining on grids*, European Conference on Principles and Practice of Knowledge Discovery in Databases (Porto, Portugal), LNAI, Springer-Verlag, 2005, pp. 309–320.
49. J. Timmis, T. Knight, L. N. De Castro, and E. Hart, *An overview of artificial immune systems*, Computation in Cells and Tissues: Perspectives and Tools for Thought (R Paton, H Bolouri, M Holcombe, J H Parish, and R Tateson, eds.), Springer, 2004, pp. 51–86.
50. J. Twycross, *An immune system approach to document classification*, Master's thesis, University of Sussex, 2002.
51. V. N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, 1995.
52. A.B. Watkins and L.C. Boggess, *A resource limited artificial immune classifier*, Congress on Evolutionary Computation, vol. 1, 2002, pp. 926 – 931.
53. Immon W.H., *Building the data warehouse*, John Wiley and Sons, New York, 1996.
54. I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and tech-niques*, 2nd edition ed., Morgan Kaufmann, San Mateo, 2005.
55. B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, Bussey K.J., J. Riss, J.C. Barrett, and J.N. Weinstein, *GoMiner: A resource for biological interpretation of genomic and proteomic data*, Genome Biology **4** (2003), no. 4, R28.