11

Artificial Immune Systems in Bioinformatics

Vitoantonio Bevilacqua[1], Filippo Menolascina[12], Roberto T. Alves[3], Stefania Tommasi[2], Giuseppe Mastronardi[1], Myriam Delgado[3], Angelo Paradiso[2], Giuseppe Nicosia[4], and Alex A. Freitas[5]

[1] Polytechnic of Bari, Via E. Orabona 4, 70125 Bari, Italy
bevilacqua@poliba.it
[2] National Cancer Institute 'Giovanni Paolo II', Via F. Hahnemann 10, 70126 Bari, Italy
f.menolascina@ieee.org
[3] Federal Technological University of Paraná, Av. 7 de setembro, 3165 Curitba, Brazil
r.t.alves@gmail.com
[4] University of Catania, Viale A. Doria 6, 95125 Catania, Italy
g.nicosia@dmi.unict.it
[5] Computing Laboratory University of Kent, Canterbury, CT2 7NF, UK
a.a.freitas@kent.ac.uk

Summary. 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 ones coming from computational biology and biochemistry. The chapter will show how AIS have been successfully used in computational biology problems and will give readers further hints about possible implementations in unexplored fields. The main goal of the contribution lays in providing both theoretical foundations and hands-on experience that allow researchers to figure out novel applications of AIS in bioinformatics and, at the same time, providing researchers with necessary insights for implementation in daily research. The contribution will be organised in 5 sections.

11.1 Introduction

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 ones 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 coexistence and integration that reflects a high degree of interaction among several computational intelligence approaches like artificial neural network, evolutionary

T.G. Smolinski et al. (Eds.): Comp. Intel. in Biomed. & Bioinform., SCI 151, pp. 271–295, 2008.
springerlink.com

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 a have a great potential to interact the new born technique with the other previously existing. Strictly speaking evolution and immune system are biologically very correlated 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 stimulate 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 like 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 11.3.1 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/optimisation. 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 11.1 depicts the clonal selection principle.

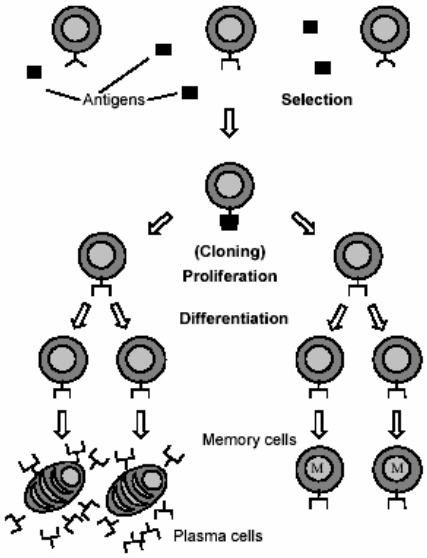

Fig. 11.1. Clonal selection principle in natural immune systems

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 [1]. 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 recognises 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 [1]]:

- 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. 11.2. 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 [1]. 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 [1].

Receptor editing offers the ability to escape from local optima on an affinity landscape. Fig 11.3 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 it is taken 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 optima (*Ab1 **). Because mutations with lower affinity are lost, the antibodies can not go down

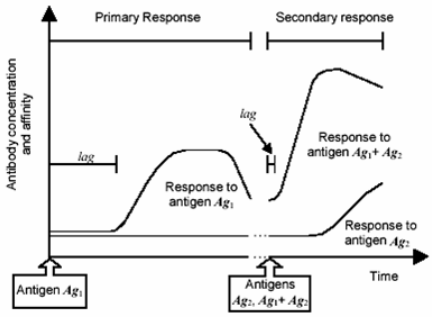

Fig. 11.2. Immune response plotted as antibody concentration over time

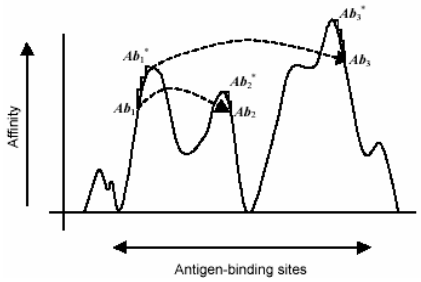

Fig. 11.3. Antibody affinity as function of the specific antigen binding site

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, and concepts of modern immunology to design immunity-based system applications in science and engineering [1]. AIS are adaptive systems in which learning takes place by evolutionary mechanisms similar to biological evolution. These different research areas are tied together: the more we learn from in silico modelling 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 specialised 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. The chapter will show how AIS have been successfully used in computational biology problems and will give readers further hints about possible implementations in unexplored fields. The main goal of the contribution lays in providing both theoretical foundations and hands-on experience that allow researchers to figure out novel applications of AIS in bioinformatics and, at the same time, providing researchers with necessary insights for implementation in daily research.

11.2 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 higher tasks like experimental design and results interpretation at the same time avoiding him minding of 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.

11.2.1 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" [2] (that just like the original Moore's law in electronics, models future progress in biotechnology [3]). 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 didn 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 [4]. 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 [5] 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 grew significant experiences like the GEO (Gene Expression for Omnibus, [7]), SMD (Stanford Microarray Database, [8]) and ArrayExpress [9]. This is the evident that data warehouse can greatly help researchers in reducing the gap providing a valuable aid in filling the last real hole in experimental processes automation: results interpretation.

11.2.2 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"* [6] (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 spacing from Customer Relationship Management (CRM, [10]) to Decision Making Support in medicine [15]. Data mining in bioinformatics, then, can be considered as a useful tool for modelling 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 [56, 57] and cancer evolution modelling [58]. 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* [16]. 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 [40, 41, 45]. Comprehensive software suites 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 suites are well represented by:

- Weka [18]
- Rapid Miner (formerly YALE) [19]
- Orange [20]

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 flavours (BioWeka [21] devoted to biological sequences mining and Weka4WS [22], 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 bioinspired [11, 12, 13] data mining algorithm in a customized version of Weka [14]. 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* flavours 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 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 [55], and renamed to CLONALG in [61] 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 maximise classification accuracy and minimise misclassification accuracy still using clonal selection paradigms.

Immunos. The Immunos [54] algorithm has been mentioned a number of times in AIS literature [37, 38, 39]. 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 [42] algorithm was one of the first AIS technique designed specifically and applied to classification problems. After an initialisation 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 deeply discussed in the next section.

11.2.3 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 [17, 18]. 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 [24] and neural networks [25]. 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. By 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 [23]. 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 [27], and named IFRAIS (Induction of Fuzzy Rules with an Artificial Immune System). IFRAIS 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 [28]. Other algorithms based on AIS for rule induction are discussed in detail in [66, 67].

Current Models

Artificial Immune Systems in Bio-medical Data Mining: IFRAIS Study Case As mentioned earlier, IFRAIS is an AIS that discovers fuzzy classification rules from data. 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 conti-nuous 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 [18]. 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 [34] [35] IFRAIS was successfully employed to discover fuzzy classification rules for female breast cancer familiarity profiling. IFRAIS' results were validated using statistical driven approaches using Gene Ontology through GO Miner [40]. Competitive results obtained by IFRAIS seem to encourage new efforts in this field. A biological interpretation of the results carried out using Gene Ontology is currently under investigation.

11.2.4 Application of AIS based Data Mining in Bioinformatics

As we previously stated several examples of application of AIS based data mining systems in bioinformatics can be retrieved in literature. Artificial Immune Systems-derived algorithms have been employed in familiarity profiling [34], prognosis prediction [58] and estrogen receptor modelling [59] 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 [60]. Previously de Castro and colleagues focused on the use of Hierarchical Artificial Immune Network paradigm for the problem of gene expression clustering [63, 64] and for rearrangement study of gene expression [62]. AIS/K-NNK-NN hybrid data mining algorithm have been tested for cancer classification in [43]. Tsanakova and colleagues, instead, focused on the problem of gene signature finding in the context of diffuse large B-Cell lymphoma [44]. A similar perspective has been reported by Ando and colleagues in [65] for the problem of acute leukemia classification. PCA-AIRS hybrid systems have been employed in the diagnosis of lung cancer [46] and [47]. A hybrid system based on fuzzy weighting pre-processing and AIRS has been described and employed in the diagnosis of heart, hepatitis and thyroid diseases in [48, 49, 50] respectively. Research currently being carried out by Alves and colleagues is mainly focused on the application of a multi-label AIS based data mining system to the problem of protein function prediction [36].

11.3 Immune Algorithms in Structural Bioinformatics and Proteomics

11.3.1 The Multi-objective Immunological Algorithm

Central to the field of protein structural biology is a set of observations, hypothesis and so-called paradoxes. The *Thermodynamic hypothesis* postulates that the

native state of a protein is the state of lowest free energy of the protein system under physiological conditions.

The free energy of a protein can be modelled as function of the different interactions within the protein. These interactions (local, non-local, hydrophobic, entropic effects, hydrogen bonding) depend on the positions of the atoms of the protein. The set of atomic coordinates providing the minimum possible value of the free energy corresponds to the native conformation of the protein. Since the interactions comprising the energy function are highly non-convex, the protein structure prediction (PSP) problem must be tackled as a global optimization problem.

For the past fifty years, the PSP problem has been defined as a *large single-objective optimization problem*, with researchers employing Molecular Dynamics, Monte Carlo methods and Evolutionary Algorithms [71, 69, 72, 73, 70]. In this section, we reason by computational experiments that it would be more suitable to model the PSP problem as a *multi-objective optimization problem*. The goal of the research is to find a set of *equivalent* three-dimensional folded conformations, relying on the observation that the folded state is one of only a small *ensemble* of all possible conformations [74]. We adopt a multi-objective approach in order to obtain "good" non-dominated compact solutions near or inside the folded state.

PAES is a multi-objective optimizer which uses a simple (1+1) local search evolution strategy. Nonetheless, it is capable of finding diverse solutions in the Pareto optimal set because it maintains an archive of non-dominated solutions which it exploits to accurately estimate the quality of new candidate solutions. At each iteration t, a candidate solution c_t and a mutated solution m_t must be compared for dominance. Acceptance is simple if one solution dominates the other. If neither solution dominates the other, the new candidate solution is compared with the reference population of previously archived non-dominated solutions. If the comparison fails to favor one solution over the other, the chosen solution is the one which resides in the least crowded region of the space. A maximum size of the archive is always maintained. The crowding procedure is based on recursively dividing up the M-dimensional objective space in 2^d equal-sized hypercubes, where d is a user defined depth parameter. The algorithm continues until a given, fixed number of *iterations* is reached.

PAES by itself has proved to be a very useful MOEA with successful application in many different fields. However, when applied to the PSP problem, we have observed poor performance both in terms of energy function and final structure obtained. The complexity of the funnel landscape of the PSP problem, which is characterized by a huge number of local minima, coupled with the goal of producing a "good" conformation from a structural point of view ($RMSD$ and DME), clearly poses many problems (e.g., premature convergence, trapping in local minima, etc).

I-PAES [76] is a modified version of PAES with a different solution representation (polypeptide chain) and immune inspired (*cloning* and *hypermutation*) operators. The algorithm starts by initializing a random conformation. The torsion angles (ϕ, ψ, χ_i) are generated randomly from the constraint regions. Next,

```
I-PAES(depth, archive_size, objectives)
1. t := 0;
2. Initialize(c); /*Generate initial random solution*/
3. Evaluate(c); /*Evaluation of initial solution*/
4. AddToArchive(c); /*Add c to archive*/
5. while(not(Termination()))
        /*Start Immune phase*/
6.      (c1^clo, c2^clo) := Cloning(c); /*Clonal expansion phase*/
7.      (c1^hyp, c2^hyp) := Hypermutation(c1^clo, c2^clo); /*Affinity maturation phase*/
8.      Evaluate(c1^hyp, c2^hyp); /*Evaluation phase*/
10.     if(c1^hyp dominates c2^hyp) m := c1^hyp;
10.     else if(c2^hyp dominates c1^hyp) m := c2^hyp;
10.     else m := Best(c1^hyp, c2^hyp); /*min E_charmm selection*/
12.          AddToArchive(Worst(c1^hyp, c2^hyp)); /*max E_charmm selection*/
        /*End Immune phase*/
        /*Start (1+1)-PAES*/
10.     if(c dominates m) discard m;
11.     else if(m dominates c)
12.          AddToArchive(m);
13.          c := m;
14.     else if(m is dominated by any member of the archive) discard m;
15.     else test(c, m, archive_size, depth);
16.     t := t + 1;
17. endwhile
```

Fig. 11.4. Pseudo-code of I-PAES

the energy of the conformation (a point in the landscape) is evaluated. The protein structure in internal coordinates (torsion angles) is transformed in Cartesian coordinates. The CHARMM energy potential of the structure is then computed using routines from TINKER Molecular Modeling Package[1].

Figure 11.4 shows the pseudo-code of the algorithm.

11.3.2 Open Questions in Proteomics

Given a protein with unknown biological function, its function(s) can be determined in a biological laboratory or via theoretical/computational methods. In a biological laboratory, the determination of protein functions is usually performed by experimental methods such as X-ray, crystallography or nuclear magnetic resonance. Theoretical/computational methods include homology modelling (based on previous knowledge) or ab-initio methods [29]. The problem of protein function prediction can be naturally cast as a classification problem. In this context, a protein is considered as an example (record) to be classified, and a list of predefined protein functions that can be assigned to each protein are the classes

[1] http://dasher.wustl.edu/tinker/

to be predicted by the classification algorithm. The ultimate goal is to predict the functions of proteins whose function is not yet known, based on attributes describing characteristics of the proteins. Protein function prediction is a very active research area for several reasons, such as the urgent and crucial need for a better understanding of proteins related to diseases, developing of more effective medical drugs, preventive medicine, etc. In any case, the very large volume of data stored in biological databases makes it infeasible to manually determine the function of each protein in those databases. Hence, several bioinformatics studies have been performed with the aim of developing computational methods for predicting protein function [26]. At present, the biological functions that can be performed by proteins are defined in a structured, standardized dictionary of terms called the Gene Ontology [30]. The GO consists of a dictionary that defines gene products independent from species. GO actually consists of 3 separate "domains" (very different types of GO terms): molecular function, biological process and cellular component. The GO is structurally organized in the form of a direct acyclic graph (DAG), where each GO term represents a node of the hierarchical structure. The inter-node relationships are of the type "is a" or "part of". A "child" node can have one or more parent nodes in the DAG. Several other works have been proposed for predicting the biological functions of proteins according to the GO [31, 32, 33].

Current Models

Towards Protein Function Prediction with AIS for Hierarchical Classification
The vast majority of classification algorithms assign just one class to an example (a protein, in the case of protein function prediction). Such classification algorithms solve a so-called single-label classification problem. However, in the context of protein function prediction, it is often necessary that the algorithm be flexible enough to be able to assign multiple classes (functions) to a protein, characterizing a multi-label classification problem [51]. In addition, protein functions are often defined in a hierarchical fashion, such as the functions included in the Gene Ontology (GO) - briefly discussed earlier. IFRAIS is a single-label, "flat" (non-hierarchical) classification algorithm. Work is ongoing in modifying IFRAIS to be a multi-label hierarchical classification algorithm [36]. One of the extensions being incorporated in the algorithm is to make it consistent with the semantics of the protein function hierarchy in GO. More precisely, when a protein is annotated with a GO term, this means that it contains not only the function specified by that term, but also the functions specified by all other terms which are ancestors of the former term in the GO's function hierarchy. IFRAIS [27] is being modified to guarantee that such hierarchical semantics is preserved in the candidate classification rules throughout the training of the algorithm. Another modification being implemented is to allow the algorithm to solve a multi-label classification problem, so that a single classification rule can predict one or more classes at once. Another research direction being pursued is the development of an AIS for the hierarchical prediction of GPCR (G protein coupled receptors) functions [52]. The AIS being developed in this project is a hierarchical

classification system, but not a multi-label one, since the GPCR classes being predicted are mutually exclusive at each level of the class hierarchy. A distinctive characteristic of this project is that it uses a novel methodology for designing an AIS where, instead of just using the natural immune system as a source of inspiration at a high level of abstraction (as usual in the field of AIS), the design of the AIS is influenced by the computational modelling of some aspect of the natural immune system. Hence, this project tries to achieve a much closer integration between computer science and biology than in previous AIS projects. More precisely, the key aspect of the natural immune system being modelled in the above project is the concept of antigen receptor degeneracy, which, according to [53], is essentially the capacity of a single antigen receptor to bind and recognize many different ligands. Cohen's theory is based on the idea that the degeneracy of different receptors is combined in order to achieve immune specificity. Mendao et al in [52] developed an agent-based computational model of immune degeneracy, and derived from it a high-level degeneracy-based clonal selection algorithm. This algorithm is currently being refined and extended in order to produce a degeneracy-based AIS for hierarchical classification [78].

11.3.3 Results

In the first set of experiments, we apply the approach to six proteins sequences, five extracted from reference [73] and one from [77]: 1ZDD, 1ROP, 1CRN, 1UTG, 1R69 and 1CTF.

Discussion is as follows. First we compare the performance of different versions of the PAES and I-PAES algorithms on the first protein set. Then we study the stability of the approach with respect to the native and predicted secondary structure constraints. Finally, we show specific results for each protein in terms of the obtained observed Pareto optimal sets at different time steps, $\mathcal{P}^{*,t}_{obs}$, and various dynamics of the algorithm during the evolution.

Four different versions of the PAES algorithm have been used [76] featuring dynamic (exponential decay)

The best conformation obtained with I-PAES has $DME = 0.77$Å and $RMSD = 1.92$Å (see figure 11.5).

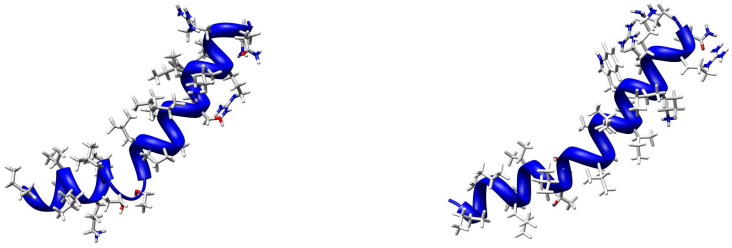

Fig. 11.5. Native (left plot) and predicted (right plot) for 2MLT protein ($DME = 0.77$Å, $RMSD = 1.92$Å)

Table 11.1. Comparative results between I-PAES$_s$, I-PAES$_m$, (1+1)-PAES$_1$ and (1+1)-PAES$_2$. For each protein we report the Protein Data Bank (PDB) identifier, the length (number of residues), the approximate class (α-helix, β-sheet), and the energy values of the native structures. The last three columns show the best results obtained for each protein on 10 independent runs. The DME and $RMSD$ values are measured on C_α atoms from the native structure. Energy values are calculated using the ANALYZE routine from TINKER.

Protein	Algorithm	DME_{min} (Å)	$RMSD_{min}$ (Å)	Min energy ($kcal/mol$)
1ROP(56 aa)	I-PAES$_s$	2.01	4.11	−661.48
class: α	I-PAES$_m$	**1.684**	**3.70**	**−902.36**
energy: *-667.05 kcal/mol*	(1+1)-PAES$_1$	4.91	6.31	2640.77
	(1+1)-PAES$_2$	5.99	8.665	−409.95
1UTG(70 aa)	I-PAES$_s$	4.49	5.11	**282.24**
class: α	I-PAES$_m$	**3.79**	**4.60**	573.89
energy: *-142.46 kcal/mol*	(1+1)-PAES$_1$	4.71	6.04	7563.07
	(1+1)-PAES$_2$	4.82	5.56	397.12
1CRN(46 aa)	I-PAES$_s$	4.13	4.73	**232.29**
class: $\alpha + \beta$	I-PAES$_m$	**3.72**	**4.31**	509.09
energy: *202.73 kcal/mol*	(1+1)-PAES$_1$	4.67	6.18	1653.93
	(1+1)-PAES$_2$	6.05	7.89	509.52
1R69(63 aa)	I-PAES$_s$	5.93	8.42	**211.26**
class: α	I-PAES$_m$	**4.91**	**5.05**	264.56
energy: *-676.53 kcal/mol*	(1+1)-PAES$_1$	5.16	7.59	9037.89
	(1+1)-PAES$_2$	6.88	8.52	659.49
1CTF(68 aa)	I-PAES$_s$	8.08	10.69	**71.55**
class: $\alpha + \beta$	I-PAES$_m$	**6.82**	**10.12**	218.99
energy: *230.08 kcal/mol*	(1+1)-PAES$_1$	9.61	12.09	1424.33
	(1+1)-PAES$_2$	8.84	10.21	617.69

11.4 Proteomic Multiple Sequence Alignments: Refinement Using an Immunological Local Search

11.4.1 Proteomics Multiple Sequence Alignments

The Multiple Sequence Alignment (MSA) of proteins plays a central role in molecular biology, as it can reveal the constraints imposed by structure and function on the evolution of whole protein families [78]. MSA has been used for building phylogenetic trees, for the identification of conserved motifs, to find diagnostic patterns families, and for predicting secondary and tertiary structures of RNA and protein sequences. In order to be able to align a set of biosequences, a reliable objective function for the measurement of an alignment in terms of its biological plausibility through an analytical or computational function is needed.

One of the most important and popular computational sequence analysis problem is to determine if two, or more, biological sequences have common subsequences. However, to check the similarities between two or more sequences, there are two primary issues that need to be faced: the choice of an objective function that assesses the biological alignment quality and the design of an

effective algorithm to optimize the given objective function. The alignment quality is often the limiting factor in biological analyses of amino-acid sequences; defining a proper objective function is a crucial task.

The classical objective function used to measure the biological alignment quality is the *weighted sums-of-pairs* with affine gap penalties [79]: each sequence receives a weight proportional to the amount of independent information that it contains [80] and the cost of the multiple alignment is equal to the sum of the costs of all the weighted pairwise substitutions:

$$\max_{\hat{S}} \left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} WSS(\hat{S}_i, \hat{S}_j) + \sum_{i=1}^{n} AGPS(\hat{S}_i) \right). \tag{11.1}$$

Sequence weights are determined by constructing a guide tree from known sequences.

11.4.2 *IMSA*, an Immunological Algorithm

In this chapter we present an immunological algorithm, IMSA, to tackle the multiple sequence alignment problem. It incorporates two different strategies to create the initial population, as well as new hypermutation operators, specific operators for solving MSA, which insert or remove gaps in the sequences. Gap columns which have been matched are moved to the end of the sequence. The remaining elements (amino acids in this work) and existing gaps are shifted into the freed space.

IMSA considers antigens (Ags) and B cells. The Ag is a given MSA instance, and B cells a set of alignments, that have solved (or approximated) the initial problem. In tackling the MSA Ags and B cells are represented by a sequence matrix. In particular, let

$$\Sigma = \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\} \tag{11.2}$$

be the twenty amino acid alphabet, and let $S = \{S_1, S_2, \ldots, S_n\}$ be the set of $n \geq 2$ sequences with length $\{\ell_1, \ell_2, \ldots, \ell_n\}$, such that $S_i \in \Sigma^*$,. Then an Ag is represented by a matrix of n rows and $max\{\ell_1, \ldots, \ell_n\}$ columns, whereas each B cell is represented by an $(n \times \ell)$ matrix, with $\ell = (\frac{3}{2} \cdot max\{\ell_1, \ldots, \ell_n\})$. By using such a representation *IMSA* was able to develop more *compact alignments*.

11.4.3 Results and Conclusions

To evaluate the biological alignment quality produced by *IMSA*, we tested it using the classical benchmark BAliBASE.

The obtained results showed in the next tables were obtained using a robust experimental protocol : $d = 10, dup = 1, \tau_B = 33, T_{max} = 2 \times 10^5$ and 50 independent runs. Moreover, we used the following substitution matrices:

- BLOSUM45 for Ref1v1 and Ref 3, with $GOP = 14, GEP = 2$;

Table 11.2. Pseudo-code of the proposed hybrid immune algorithm for the MSA

```
IMSA (d, dup, τ_B, T_max)
t ← 0;
FFE ← 0;
N_c ← d × dup;
P^(t) ← Initialize_Population(d);
Strip_Gaps(P^(t));
Evaluate(P^(t));
FFE ← FFE + d;
while (FFE < T_max)do
    P^(clo) ← Cloning (P^(t), dup);
    P^(gap) ← Gap_operators (P^(clo));
    Strip_Gaps(P^(gap));
    Evaluate(P^(gap));
    FFE ← FFE + N_c;
    P^(block) ← BlockShuffling_operators (P^(clo));
    Compute_Weights();
    Normalize_Weights();
    Strip_Gaps(P^(block));
    Evaluate(P^(block));
    FFE ← FFE + N_c;
    (P_a^(t), P_a^(gap), P_a^(block)) = Elitist-Aging(P^(t), P^(gap), P^(block), τ_B);
    P^(t+1) ← (μ + λ)-Selection(P_a^(t), P_a^(gap), P_a^(block));
    t ← t + 1;
end_while
```

Table 11.3. SP values given by several methods on the BAliBase v.1.0 benchmark

Aligner	Ref 1 (82)	Ref 2 (23)	Ref 3 (12)	Ref 4 (12)	Ref 5 (12)	Overall (141)
IMSA	80.7	**88.6**	**77.4**	70.2	82.0	**79.7**
DIALIGN [89]	77.7	38.4	28.8	**85.2**	**83.6**	62.7
CLUSTALX [83]	85.3	58.3	40.8	36.0	70.6	58.2
PILEUP8 [82]	82.2	42.8	33.3	59.1	63.8	56.2
ML_PIMA [86]	80.1	37.1	34.0	70.4	57.2	55.7
PRRP [91]	**86.6**	54.0	48.7	13.4	70.0	54.5
SAGA [94]	70.3	58.6	46.2	28.8	64.1	53.6
SB_PIMA [86]	81.1	37.9	24.4	72.6	50.7	53.3
MULTALIGN [81]	82.3	51.6	27.6	29.2	62.7	50.6

- BLOSUM62 for Ref1v2, Ref 2, Ref 4 and Ref 5, with $GOP = 11, GEP = 1$;
- BLOSUM80 for Ref1v3, with $GOP = 10, GEP = 1$.

Table 11.3 shows the average SP score obtained by the described alignment tools on every instance set of BAliBASE v.1.0. As it can be seen in the table, *IMSA* performs well on the Reference 2 and Reference 3 sets. The values obtained aid to raise the overall score, which is higher compared to the results published by the Bioinformatic platform of Strasbourg.

11.5 Conclusions and Open Questions

In this chapter we have analysed some applications of Artificial Immune System based algorithms in bioinformatics. Of course this is only a partial outlook on the world of 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, AIS are currently undergoing very fast changes resulting in a very dynamical field of reasearch 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 bioinfomatics. This led to exclude interesting projects like BIAS-PROFS coordinated by Freitas and colleagues; even in this case interested readers can find useful information in the references. 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 AIS are considered a novel paradigm but they have been already able to reach significant results in highly complex contexst like Knowledge Discovery in Data bases (section 11.2) and Protein Structure Prediction (section 11.1). Even if 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:

- Large molecules folding prediction;
- Gene networks inference;
- Disease profiling and evolution modelling.

These are only some of the most active areas of AIS based research in bioinformatics. From a theoretical point of view it should be noted that some areas like *danger theory* and *hybrid systems* 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 AIS research in bioinformatics can find useful information in [43, 49].

References

1. de Castro, L.N., Timmis, J.: Artificial Immune Systems: A New Computational Approach. Springer, Heidelberg (2002)
2. Scott, R.: Keynote Speach. TNTYN, San Francisco (2000)
3. Economist, Life 2.0. The new science of synthetic biology is poised between hype and hope. But its time will soon come. August 31, 2006 (2006)
4. Grossman, R., Kamath, C., Kumar, V.: Data Mining for Scientific and Engineering Applications. Springer, Heidelberg (2001)
5. Immon, W.H.: Building the Data Warehouse. John Wiley and Sons, New York (1996)
6. Frawley, W., Piatetsky-Shapiro, G., Matheus, C.: Knowledge Discovery in Databases: An Overview. AI Magazine, 213–228 (Fall 1992)

7. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R.: NCBI GEO: Mining tens of millions of expression profiles–database and tools update. Nucleic Acids Res. (November 11, 2006)
8. Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese, J.C., Nitzberg, M., Wymore, F., Zachariah, Z.K., Brown, P.O., Sherlock, G., Ball, C.A.: The Stanford Microarray Database: Implementation of new analysis tools and open source release of software. Nucleic Acids Res. 35(Database Issue), D766–D770 (2007)
9. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P., Sansone, S.A.: ArrayExpress–a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31(1), 68–71 (2003)
10. Rigby, D.K., Ledingham, D.: CRM Done Right. Harvard Business Review (November 1, 2004)
11. Brownlee, J.: Artificial Immune Recognition System (AIRS) - A Review and Analysis [Technical Report]. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology, Victoria, Australia, Technical Report ID: 1-01 (2005)
12. Brownlee, J.: Clonal Selection Theory and CLONALG - The Clonal Selection Classification Algorithm (CSCA) [Technical Report]. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology, Victoria, Australia, Technical Report ID: 2-01 (2005)
13. Brownlee, J.: Immunos-81 – The Misunderstood Artificial Immune System [Technical Report]. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology, Victoria, Australia, Technical Report ID: 3-01 (2005)
14. Brownlee, J.: Weka Classification Algorithms, `http://sourceforge.net/projects/wekaclassalgos`
15. Siadaty, M.S., Knaus, W.A.: Locating previously unknown patterns in data-mining results: a dual data- and knowledge-mining method. BMC Medical Informatics and Decision Making 6(13) (2006) doi:10.1186/1472-6947-6-13
16. Pool, R., Esnayra, J.: Bioinformatics: Converting Data to Knowledge. Natl. Acad. Press, Washington (2003)
17. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT, Cambridge (1996)
18. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Mateo (2005)
19. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006) (2006)
20. Michalski, R.S., Bratko, I., Kubat, M.: Machine Learning and Data Mining: Methods and Applications. Wiley, Chichester (1998)
21. Gewehr, J.E., Szugat, M., Zimmer, R.: BioWeka-extending the Weka framework for bioinformatics. Bioinformatics 23(5) (March 2007) ISSN:1367-4803
22. Talia, D., Trunfio, P., Verta, O.: Weka4WS: a WSRF-enabled Weka Toolkit for Distributed Data Mining on Grids. In: Jorge, A.M., Torgo, L., Brazdil, P.B.,

Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 309–320. Springer, Heidelberg (2005)
23. Freitas, A.A.: Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer, Berlin (2002)
24. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)
25. Haykin, S.: Neural Networks – A Comprehensive Foundation, 2nd edn. Prentice Hall, Upper Saddle River (1999)
26. Fogel, G.B., Corne, D.W.: Evolutionary Computation in Bioinformatics. Morgan Kaufmann Publishers, San Franciso (2003)
27. Alves, R.T., Delgado, M.R., Lopes, H.S., Freitas, A.A.: An artificial immune system for fuzzy-rule induction in data mining. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiňo, P., Kabán, A., Schwefel, H.-P. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 1011–1020. Springer, Heidelberg (2004)
28. Pedrycz, W., Gomide, F.: An Introduction to Fuzzy Sets: Analysis and Design. MIT Press, Cambridge (1998)
29. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Water, P.: Molecular Biology of the Cell, 4th edn. Garland Science, New York (2002)
30. The Gene Ontology Consortium, The Gene Ontology (GO) Database and Informatics Resource. Nucleic Acids Research 32(1), 258–261 (2004)
31. Vinayagam, A., Konig, R., Moormann, J., Schubert, F., Eils, R., Suhai, S.: Applying Support Vector Machines for Gene Ontology based gene function prediction. BMC Bioinformatics 5, 116–129 (2004)
32. Eisner, R., Poulin, B., Szafron, D., Lu, P., Greiner, R.: Improving Protein Function Prediction using the Hierarchical Structure of the Gene Ontology. In: Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2005)
33. Tu, K., Yu, H., Guo, Z., Li, X.: Learnability-Based Further Prediction of Gene Func-tions in Gene Ontology. Genomics 86, 922–928 (2004)
34. Menolascina, F., Alves, R.T., Tommasi, S., Chiarappa, P., Delgado, M., Bevilacqua, V., Mastronardi, G., Freitas, A.A., Paradiso, A.: Fuzzy Rule Induction and Artificial Immune Systems in Female Breast Cancer Familiarity Profiling. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS (LNAI), vol. 4694, pp. 830–837. Springer, Heidelberg (2007)
35. Menolascina, F., Tommasi, S., Paradiso, A., Cortellino, M., Bevilacqua, V., Mastronardi, G.: Novel Data Mining Techniques in aCGH based Breast Cancer Subtypes Proling: the biological perspective. In: Proc. 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Honolulu, HI, USA, April 1-5, pp. 9–16 (2007)
36. Alves, R.T.: An Artificial Immune System to Hierarchical Multi-label Classification for Predicting Protein Function. Ph.D. Qualifying Exam 42, Federal University of Technology of Paraná -UTFPR, Curitiba, Brazil (2007)
37. Timmis, J., Knight, T., de Castro, L.N., Hart, E.: An Overview of Artificial Computation in Cells and Tissues: Perspectives and Immune Systems. In: Tools for Thought, Anonymous, pp. 51–86. Springer, Heidelberg (2004)
38. Hart, E.: Immunology as a Metaphor for Computational Information Processing: Fact of Fiction. University of Edinburgh (2002)
39. Twycross, J.: An Immune System Approach to Document Classification. University of Sussex (2002)
40. Zeeberg, B.R., et al.: GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data. Genome Biology 4(4), R28 (2003)

41. Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., Mesirov, J.P.: GenePattern 2.0. Nature Genetics 38(5), 500–501 (2006)
42. Watkins, A.B.: A resource limited artificial immune classifier. Mississippi State University (2001)
43. Sahan, S., Polat, K., Kodaz, H., Gunes, S.: A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine 37(3), 415–423 (2007)
44. Tsankova, D., Rangelova, V.: Cancer Outcome Prediction by Cluster-based Artificial Immune Networks. In: Proc. Biomedical Engineering (2007)
45. de la Nava, J.G., Santaella, D.F., Alba, J.C., Carazo, J.M., Trelles, O., Pascual-Montano, A.: Engene: The processing and exploratory analysis of gene expression data. Bioinformatics 19(5), 657–658 (2002)
46. Polat, K., Gunes, S.: Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer. Expert Systems with Applications: An International Journal 34(1) (2008)
47. Polat, K., Gunes, S.: Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm. Expert Systems with Applications: An International Journal 34(1) (2008)
48. Polat, K., Gunes, S., Tosun, S.: Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre-processing. Pattern Recognition 39(11) (2006)
49. Polat, K., Sahan, S., Gunes, S.: 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(4) (2007)
50. Polat, K., Gunes, S.: Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection. Expert Systems with Applications: An International Journal 33(2) (2007)
51. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3(3), 1–13 (2007)
52. Mendao, M., Timmis, J., Andrews, P.S., Davies, M.: The Immune System in Pieces: Computational Lessons from Degeneracy in the Immune System. In: Proc. 2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007), Honolulu, HI, USA (2007)
53. Cohen, I.R., Hershberg, U., Solomon, S.: Antigen-receptor degeneracy and immunological paradigms. Molecular Immunology 40, 993–996 (2004)
54. Carter, J.H.: The immune system as a model for classification and pattern recognition. Journal of the American Informatics Association 7 (2000)
55. de Castro, L.N., von Zuben, F.J.: The Clonal Selection Algorithm with Engineering Applications. In: GECCO 2000, Workshop on Artificial Immune Systems and Their Applications, Las Vegas, USA, pp. 36–37 (2000)
56. Larranaga, P., Gallego, M.J., Sierra, B., Urkola, L., Michelena, M.J.: Bayesian Networks, Rule Induction and Logistic Regression in the prediction of the survival of women suffering from breast cancer. In: Costa, E. (ed.) EPIA 1997. LNCS, vol. 1323, pp. 303–308. Springer, Heidelberg (1997)
57. Bevilacqua, V., Chiarappa, P., Mastronardi, G., Menolascina, F., Paradiso, A., Tommasi, S.: Identification of Tumour Evolution Patterns by Means of Inductive Logic Programming. Journal - Genomics Proteomics and Bioinformatics (in press, 2007)

58. Menolascina, F., Alves, R.T., Tommasi, S., Chiarappa, P., Delgado, M., Bevilacqua, V., Mastronardi, G., Freitas, A.A., Paradiso, A.: Improving Female Breast Cancer Prognosis by means of Fuzzy Rule Induction with Artificial Immune Systems. Journal of Dynamics of Discrete Continuous and Impulsive Systems (to appear, 2007) ISSN:1492-8760
59. Menolascina, F., Alves, R.T., et al.: Induction of Fuzzy Rules with Artificial Immune Systems in aCGH based ER Status Breast Cancer Characterization. In: Proc. GECCO 2007, ACM 978-1-59593-697-4/07/0007 (2007)
60. Menolascina, F., Tommasi, S., Chiarappa, P., Bevilacqua, V., Mastronardi, G., Paradiso, A.: Data mining techniques in aCGH-based breast cancer subtype profiling: an immune perspective with comparative study. BMC Systems Biology 1(suppl. 1), P70 (2007)
61. de Castro, L.N., von Zuben, F.J.: Learning and Optimization IEEE Transactions on Evolutionary Using the Clonal Selection Principle Computation. Special Issue on Artificial Immune Systems 6, 239–251 (2002)
62. de Sousa, J.S., de Gomes, C.T., Bezerra, G.B., de Castro, L.N., von Zuben, F.J.: An Immune-Evolutionary Algorithm for Multiple Rearrangements of Gene Expression Data. Genetic Programming and Evolvable Machines 5(2), 157–179 (2004)
63. Bezerra, G.B., Cançado, G.M.A., Menossi, M., de Castro, L.N., von Zuben, F.J.: Recent advances in gene expression data clustering: a case study with comparative results. Genet. Mol. Res. 4(3), 514–524 (2005)
64. Hruschka, E.R., Campello, R.J.G.B., de Castro, L.N.: Evolving clusters in gene-expression data. Inf. Sci. 176(13), 1898–1927 (2006)
65. Ando, S., Iba, H.: Artificial Immune System for Classification of Cancer. In: Raidl, G.R., Cagnoni, S., Cardalda, J.J.R., Corne, D.W., Gottlieb, J., Guillot, A., Hart, E., Johnson, C.G., Marchiori, E., Meyer, J.-A., Middendorf, M. (eds.) EvoWorkshops 2003. LNCS, vol. 2611, p. 219. Springer, Heidelberg (2003)
66. Castro, P.A.D., Coelho, G.P., Caetano, M.F., von Zuben, F.J.: Designing ensembles of fuzzy classification systems: An immune-inspired approach. In: Jacob, C., Pilat, M.L., Bentley, P.J., Timmis, J.I. (eds.) ICARIS 2005. LNCS, vol. 3627, pp. 469–482. Springer, Heidelberg (2005)
67. Alatas, B., Akin, E.: Mining fuzzy classification rules using an artificial immune system with boosting. In: Eder, J., Haav, H.-M., Kalja, A., Penjam, J. (eds.) ADBIS 2005. LNCS, vol. 3631, pp. 283–293. Springer, Heidelberg (2005)
68. Anfinsen, C.: Principles that govern the folding of protein chains. Science 181, 223–230 (1973)
69. Simons, K.T., Kooperberg, C., Huang, E., Baker, D.: Assembly of of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring function. J. Mol. Biol. 306, 1191–1199 (1997)
70. Hansmann, U.H., Okamoto, Y.: Numerical comparisons of three recently proposed algorithms in the protein folding problem. J. Comput. Chem. 18, 920–933 (1998)
71. Bowie, J.U., Eisemberg, D.: An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. Proc. Natl. Acad. Sci. USA 91, 4436–4440 (1994)
72. Pendersen, J.T., Moult, J.: Protein folding simulations with genetic algorithms and a detailed molecular description. J. Mol. Biol. 169, 240–259 (1997)
73. Cui, Y., Chen, R.S., Wong, W.H.: Protein Folding Simulation using Genetic Algorithm and Supersecondary Structure Constraints. Proteins: Structure, Function and Genetics 31(3), 247–257 (1998)

74. Plotkin, S.S., Onuchic, J.N.: Understanding protein folding with energy landscape theory. Quarterly Reviews of Biophysics 35(2), 111–167 (2002)
75. Foloppe, N., MacKerell Jr., A.D.: All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. J. Comput. Chem. 21, 86–104 (2000)
76. Cutello, V., Narzisi, G., Nicosia, G.: A Class of Pareto Archived Evolution Strategy Algorithms Using Immune Inspired Operators for Ab-Initio Protein Structure Prediction. In: Rothlauf, F., Branke, J., Cagnoni, S., Corne, D.W., Drechsler, R., Jin, Y., Machado, P., Marchiori, E., Romero, J., Smith, G.D., Squillero, G. (eds.) EvoWorkshops 2005. LNCS, vol. 3449, pp. 54–63. Springer, Heidelberg (2005)
77. Dal Palu, A., Dovier, A., Fogolari, F.: Constraint Logic Programming approach to protein structure prediction. BMC Bioinformatics 5(11), 186 (2004)
78. Eidhammer, I., Jonassen, I., Taylor, W.R.: Protein Bioinformatics. Wiley, Chichester (2004)
79. Altschul, S.F., Lipman, D.J.: Trees stars and multiple biological sequence alignment. SIAM J. on App. Maths. 49, 197–209 (1989)
80. Altschul, S.F., Carroll, R.J., Lipman, D.J.: Weights for data related by a tree. J. on Mol. Biol. 207, 647–653 (1989)
81. Corpet, F.: Multiple sequence alignment with hierarchical clustering. Nuc. Acids Research 16, 10881–10890 (1998)
82. Genetics Computer Group, Wisconsin Package v.8 (1993), `http://www.gcg.com`
83. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G.: The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nuc. Acids Research 24, 4876–4882 (1997)
84. Zhou, H., Zhou, Y.: SPEM: Improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 21, 3615–3621 (2005)
85. Do, C.B., Mahabhashyam, M.S.P., Brudno, M., Batzoglou, S.: ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Research 15, 330–340 (2005)
86. Smith, R.F., Smith, T.F.: Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. Prot. Engineering 5, 35–41 (1992)
87. Carrillo, H., Lipman, D.J.: The Multiple Sequence Alignment Problem in Biology. SIAM J. on App. Maths. 48, 1073–1082 (1988)
88. Stoye, J., Moulton, V., Dress, A.W.: DCA: An efficient implementation of the divide-and conquer approach to simultaneous multiple sequence alignment. Bioinformatics 13(6), 625–626 (1997)
89. Morgenstern, B., Frech, K., Dress, A., Werner, T.: DIALIGN: Finding local similarities by multiple sequence alignment. Bioinformatics 14, 290–294 (1998)
90. Morgenstern, B., Frech, K., Dress, A., Werner, T.: DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15, 211–218 (1999)
91. Gotoh, O.: Further improvement in methods of group-to-group sequence alignment with generalized profile operations. Bioinformatics 10(4), 379–387 (1994)
92. Eddy, S.R.: Multiple alignment using hidden Markov models. In: Proc. 3rd Int. Conference on Intelligent Systems for Molecular Biology (ISMB 1995), Cambridge, UK, pp. 114–120 (1995)
93. Edgar, R.C.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nuc. Acids Research 32, 1792–1797 (2004)
94. Notredame, C., Higgins, D.G.: SAGA: Sequence alignment by genetic algorithm. Nuc. Acids Research 24, 1515–1539 (1996)

95. Notredame, C.: COFFEE: An objective function for multiple sequence alignments. Bioinformatics 14, 407–422 (1998)
96. Simossis, V.A., Heringa, J.: PRALINE: A multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nuc. Acids Research 33, 289–294 (2005)
97. Shyu, C., Sheneman, L., Foster, J.A.: Multiple Sequence Alignment with Evolutionary Computation. Gen. Prog. and Evol. Machs. 5, 121–144 (2004)
98. Nguyen, H.D., Yoshihara, I., Yamamori, K., Yasunaga, M.: Aligning Multiple Protein Sequences by Parallel Hybrid Genetic Algorithm. Genome Inf. 13, 123–132 (2002)
99. Cutello, V., Narzisi, G., Nicosia, G., Pavone, M.: Clonal selection algorithms: A comparative case study using effective mutation potentials. In: Jacob, C., Pilat, M.L., Bentley, P.J., Timmis, J.I. (eds.) ICARIS 2005. LNCS, vol. 3627, pp. 13–28. Springer, Heidelberg (2005)
100. Cutello, V., Nicosia, G., Pavone, M.: Exploring the capability of immune algorithms: A characterization of hypermutation operators. In: Nicosia, G., Cutello, V., Bentley, P.J., Timmis, J. (eds.) ICARIS 2004. LNCS, vol. 3239, pp. 263–276. Springer, Heidelberg (2004)