An Intelligent Tool to support Biomedical Applications

Teresa M. A. Basile\(^{(1)}\), Floriana Esposito\(^{(1)}\), Laura Caponetti\(^{(1)}\)

\(^{(1)}\) Università degli Studi di Bari - Dipartimento di Informatica- Bari, Italy
{basile, esposito, laura}@di.uniba.it

Keywords: Multi-relational learning, biological image processing.

Abstract. The definition of tools able to extract knowledge from structured biological data in order to support scientists research is increasing as shown by the popularity reached in the field of bioinformatics. Generally the existing tools work at different level according to their data starting point, i.e. images or clinical/molecular information. In both cases, each approach neglects the complementary aspects identified and analyzed by the other approach. On the other hand, some approaches work with both these kinds of information, but they assume that all the information on both features extracted from the images and the clinical/molecular data is already provided. In this paper we would provide an intelligent framework able to deal at the same time with both the aspects of automatically analyze bio-medical images in order to extract the relevant features to the problem at hand and put them together with the set of clinical/molecular data of the biological system in a multi-relational learning framework able to discover hidden relationships between variables involved in a biology system, similarities between the individuals belonging to a same system and peculiarities of the cellular systems to exploit as markers in order to assess/predict their “quality” in specific tasks. In particular, we focus our attention on the domain of assisted reproductive techniques with particular interest on the field of intracytoplasmic sperm injection.

1 Introduction

Machine learning has become a rapidly developing and increasingly aspect of many biomedical applications involving clinical information systems and clinical decision support systems.

In the field of assisted reproductive technologies, ICSI (IntraCytoplasmic Sperm Injection) fertilization is a medically-assisted reproduction technique, enabling infertile couples to achieve successful pregnancy. In this field crucial points are: the analysis of clinical data of the patient, aimed at adopting an appropriate stimulation protocol to obtain an adequate number of oocytes, and the selection of the best oocytes to fertilize. The main goal is the identification of some factors useful to prognostic a pregnancy.

Generally this analysis is manually performed by the clinicians and is based on the subjective experience taking in account only few cases previously examined. Thus a learning system able to exploit past experiences to suggest possible modifications to an ICSI treatment plan could be useful to aid clinicians in making decisions, for example, about the stimulation protocol to be carried out in order to obtain good quality oocytes. Once the system’s knowledge base is populated with a sufficient number of past cases, it can be used to explore and discover interesting relationships among data, thereby achieving a form of knowledge mining.

In this work we present a multi-relational learning approach able to deal with clinical data and relevant features extracted from oocyte images with the aim of discovering new information useful to support the clinicians both in the stimulation protocol definition for new unseen patients and in the selection of oocytes between new unseen oocytes. Due to the presence of strong relationships among different stage of the process, multi-relational learning techniques that are able to take into account the relationships existing
among all the entities involved in the process seem to be the most suitable approaches in this and similar medical application domains. Furthermore, the human-comprehensible results produced by the multi-relational techniques can be used as an advisor to the clinicians during their work in order to help them in determining what knowledge sources are relevant for a treatment plan.

This paper is organized as follows. In Section 2 the ICSI application domain is presented along with its features. Section 3 and Section 4 present the multi-relational learning techniques exploited in such domain and the framework for image analysis and knowledge extraction from data. Finally Section 5 reports the preliminary experimental results on real data.

2 Problem description

Infertility is becoming a frequent problem in the last decades and many assisted reproductive techniques have been designed to overcome it. One of these techniques is the intracytoplasmic sperm injection (ICSI) technique in which a single sperm is directly injected into an oocyte. After the procedure, the oocyte is placed into cell culture and checked on the following day for signs of fertilization. The fertilized oocyte grows in a laboratory for one to five days, then it is placed in the woman’s uterus.

Due to ethical and medical reasons a specified number of embryos have to be selected and hence transferred in woman’s uterus. As a consequence, even the number of oocytes to fertilize could be under such a restriction and clinicians prefer to appropriately select the most promising oocytes among all the oocytes taken from the woman. Fig. 1(left) shows a complete overview of the procedure.

The oocytes selection is manually done by non-invasive examination based on simple methods and observation focused on morphology and dynamics of the oocyte. A set of morphological parameters to be examined are present in medical literature such as oocyte/cytoplasm dimension, perivitelline space and zona pellucida thickness, first polar body conformation, and more subtle abnormalities (dysmorphisms) of cytoplasm such as central granularity, inclusions and vacuolation.

However, these variables are not the unique and independent parameters involved in the process (see fig. 1(right)). Indeed, in general, before the ICSI procedure, an hormone stimulation protocol of the female patient, consisting of a set of pharmacological treatments, is carried out in order to ensure the development of multiple preovulatory follicles to obtain multiple oocytes to aspirate. In this phase, the couples’ health conditions and characteristics have to be taken into account as well.

Some works faced the problem of introducing systems to support clinicians in their work. Some approaches work with low level features extracted from oocyte images to assess their quality, such as [1] that proposes a method to evaluate the oocyte diameter, [2] that presents an approach to evaluate the quantification of oocyte cytoplasm...
morphology or [3] that defines a quantitative evaluation of birefringence properties of the zona pellucida. Other approaches work with higher level characteristics such as the clinical data of the patients with the aim of grasping structural patterns that define the peculiarities of the patient.

Few approaches are presented in literature that work with both these kinds of information, but they consider that all the information on both features extracted from the images and the clinical data are available [4, 5, 6]. Furthermore, they exploit an attribute-value description of the data thus losing the relationships existing between oocytes and patients data. Indeed, an important aspect and commonly neglected by these approaches is that each set of variables, i.e. clinical patient data and image features cannot be considered as a stand-alone set since relationships between such sets of data can occur [7]. For example, clinical data of the patient are related both to the oocyte quality and to the implantation success rate; the oocyte quality, intended as its maturity, plays a fundamental role in the embryo development; finally the correct embryo development is crucial for the successive transfer and implantation success [8, 9]. For these reasons, multi-relational learning techniques able to take into account the relationships existing among all the entities involved in the process seem to be the most suitable approaches in this and similar medical application domains.

3 Multi-relational Learning Approach

The learning approaches we consider to tackle the knowledge extraction task in the ICSI application domain aim at solving the problem of relational clustering as well as the task of relational rule induction. The representation language used in this work for the multi-relational descriptions is Datalog [10], a first-order logic language. The first-order alphabet consists of a set of constants \( C \), a set of variables \( V \), a set of function symbols \( F \), and a non-empty set of predicate symbols \( P \). A multi-relational description is made up of a set of predicate symbols \( p \in P \) applied to \( n \) terms \( t_i \), \( (t_i \in \{C \cup V\}) \): \( p(t_1, \ldots, t_n) \). Multi-relational descriptions are said to be ground whenever they do not contain variables A Datalog description is a multi-relational description in which only variables and constants are used as predicate arguments.

3.1 Multi-Relational Clustering

Clustering is an unsupervised learning technique used to find a partition of a set of objects into clusters so that the objects within each cluster are similar to each other. The similarity between objects can be determined using various distance measures.

Relational clustering works on relational data (i.e., objects with a first-order description as representation language) and uses distance measures that are generally more complex than those used in the case of attribute-value representations. Indeed, the generic Euclidean distance cannot be applied to relational representations of the data as they are not represented by a feature vector of a fixed number of measurements.

Here we use the distance function and the modification of a partitional clustering algorithm, named Approximate Partition Around Medoids (APAM) both introduced in [11], and here briefly reported.

As to the distance function an adaptation of the Tanimoto metric to the case of relational descriptions is exploited. The Tanimoto metric between two objects \( O_1 \) and \( O_2 \) considers the number \( n_{11} \) of elements in \( O_1 \), the number \( n_{2} \) of elements in \( O_2 \) and the number \( n_{12} \) of the elements that are present in both the sets. These values are combined in the following formula: \( d_T(O_1, O_2) = \frac{n_{11} + n_{2} - 2n_{12}}{n_{11} + n_{2} - n_{12}} \). The resulting distance value ranges in \([0, 1]\); a value close to 0 implies similarity and a value close to 1 implies a dissimilarity among the two descriptor sets compared.

The data in a multi-relational description language are represented by a set of literals describing both properties and relationships of the objects involved in the description.
Hence, the distance function has to find the common components (objects properties and relationships) among data.

Before introducing the Tanimoto metric adaptation to the multi-relational case, let us introduce the concept of renaming of a multi-relational description.

Let $D$ be a multi-relational description, represented as a Datalog ground description, and let $\text{consts}(D)$ be the set of the constants (objects) appearing in $D$. One can write a new multi-relational description $D'$ from $D$ by changing one or more constants in $D$, i.e. by renaming. In particular, $D' = D\sigma^{-1}\theta$ may be obtained by applying an antisubstitution $\sigma^{-1}$ (i.e., a mapping from constants onto variables) and a substitution $\theta$ (i.e., a mapping from variables onto constants) under Object Identity\(^1\) to $D$.

Formally, a renaming of a multi-relational description $D$, indicated by $R(D)$, is a ground description obtained by applying a substitution $\theta = \{ V_1/t_1, V_2/t_2, \ldots, V_n/t_n \}$ to $D\sigma^{-1}$, i.e. $R(D) = D\sigma^{-1}\theta$, such that $\sigma^{-1}$ is an antisubstitution, $\{ V_1, V_2, \ldots, V_n \} \subseteq \text{vars}(D\sigma^{-1})$, and $\{ t_1, t_2, \ldots, t_n \}$ are distinct constants of $\text{consts}(D)$, $n = \text{consts}(D)$.

The set of renamings $S = \text{ren}(k, E, C)$ are generated randomly choosing $k$ renamings of the multi-relational description $D$ onto the set of constants $C$:

$$\text{ren}(k, E, C) = S = S_i \cup \{ R(D)_{(C)} \} \quad i = 1, \ldots, k$$

Now we can introduce the Tanimoto metric adaptation. Given two multi-relational descriptions $D_1 = \{ l_{11}, l_{12}, \ldots, l_{1n} \}$ and $D_2 = \{ l_{21}, l_{22}, \ldots, l_{2m} \}$, let $C_1$ and $C_2$ the set of objects involved in the description $D_1$ and $D_2$ respectively, and $C = \text{max}(C_1, C_2)$. Then, the number of literals in common between $D_1$ and $D_2$ is approximated by:

$$s_{\cap}(D_1, D_2, \alpha) = \sum_{i=1}^{n} \frac{|R_{D_1} \cap R_{D_2}|}{|D_1|} \frac{|R_{D_2}|}{|D_2|},$$

where $R_{D_1} = \text{ren}(1, D_1, C)$ is a fixed renaming of the multi-relational description $D_1$, $R_{D_2} \in \text{ren}(\alpha, D_1, C)$ is a renaming of the multi-relational description $D_2$, and $\alpha > 0$ is the parameter governing the approximation. In other words, $s_{\cap}(D_1, D_2, \alpha)$ is the mean of the number of common literals in $D_1$ and $D_2$ for each of the $\alpha$ renamings of $D_2$.

Now we can use the Tanimoto metric to define the distance between two multi-relational descriptions $D_1$ and $D_2$ as follows:

$$d_{\cap}(D_1, D_2, \alpha) = \frac{|D_1| + |D_2| - 2s_{\cap}(D_1, D_2, \alpha)}{|D_1| + |D_2| - s_{\cap}(D_1, D_2, \alpha)},$$

where $|D_i|$ is the number of components in the multi-relational description $D_i$.

As regards the partitional clustering algorithms, they use the following generic schema:

1. randomly choose $k$ representatives for clusters;
2. iteratively improve these initial representatives until the change in the objective function from one iteration to the next drops below a given threshold:
   (a) assign each object to the cluster it “fits best” in the current clustering
   (b) compute new cluster representatives using these new assignments

One of the most well-known and commonly used partitioning method is the $k$-medoids clustering algorithm. Traditional $k$-medoids clustering algorithm seeks to find $k$ medoids among the objects in the data set minimizing, for a given clustering solution $C$, the following objective function:

$$\text{tightness}(C) = \frac{1}{n} \sum_{i=1}^{n} d(\mathbf{x}_i, \mu_i),$$

where $\mu_i$ is the medoid of the cluster the object $\mathbf{x}_i$ belongs to and $d(\cdot, \cdot)$ is the distance.

The $k$-medoids clustering algorithm PAM on which APAM is based starts with a set of clusters containing the medoids of the complete data set, and greedily inserts new objects into this set of clusters while minimizing the above objective function. Then, it tries

\(^{1}\text{In the Object Identity framework, within a description, terms that are denoted with different symbols must be distinct, i.e. they must represent different objects of the domain.}\)
to improve the previously obtained clustering by exploring all possible replacements of medoids by non-medoids picking the replacement that enhances the fitness function. If no such fitness improving replacement can be found, the procedure terminates.

APAM, the approximate relational clustering variant of PAM, uses the following objective function:

$$J_{tightness}(\mathcal{C}, \alpha) = \frac{1}{n} \sum_{i=1, \ldots, n} d_{T_i}(x_i, \mu_i, \alpha).$$

Similarly to PAM, it starts by randomly selecting $k$ medoids and finding the first clustering solution $\mathcal{C}$ by associating each non-medoid instance to the cluster whose medoid is more similar. Then, it iteratively tries to swap a medoid with a non-medoid object, exploring all possible replacements, in order to minimize the value of the objective function $J_{tightness}(\cdot, \cdot)$. It terminates if no replacement can be found that leads to a clustering with a better (lower) objective value with respect to $J_{tightness}(\cdot, \cdot)$.

### 3.2 Multi-Relational Rule Induction

Rule induction is a supervised learning technique concerning the extraction of a set of formal rules from a set of labelled observations. One of the rule induction paradigms able to deal with first-order (multi-relational) representation language is the Inductive Logic Programming (ILP) framework [12]. In this setting, given a background knowledge and a set of labelled (positive and negative) observations represented as a logical database of facts, the aim is to derive a hypothesised logic program (set of rules) which entails all the positive and none of the negative observations. The derivation of the set of rules is performed by exploring the lattice-based concepts by means of some operators such as refinement, least general generalisation and inverse resolution.

In this work we adopted the ILP system INTHELEX [13] that here we briefly describe. INTHELEX is a learning system for the induction of hierarchical theories (set of rules) from positive and negative observations which focuses the search for refinements by exploiting the Object Identity bias on the generalization model. It is fully and inherently incremental: this means that, in addition to the possibility of taking as input a previously generated version of the theory, learning can also start from an empty theory and from the first available observation; moreover, at any moment the theory is guaranteed to be correct with respect to all of the observations encountered thus far.

The system can learn simultaneously various concepts, possibly interrelated, and is based on a closed loop architecture, i.e. correctness is checked on any new example and, in case of failure, a revision process is activated on it, in order to restore the correctness. The system deals with theories expressed as sets of Datalog descriptions. It adopts a full memory storage strategy, i.e. it retains all the observations in order to guarantee correctness of the learned theories on all of them.

The process of theory refinement, as performed by the system, is now briefly summarized. The system exploits a previous theory (if any) and a memory of all the past (positive and negative) observations that led to the current theory. The new observations are exploited incrementally to modify incorrect hypotheses according to a data-driven strategy. In particular, when a positive observation is not covered, a revision of the theory to restore its completeness is performed as follows:

- replacing a rule in the theory with one of its least general generalizations against the problematic observation;
- adding a new rule to the theory, obtained by properly turning constants into variables in the problematic example;
- adding the problematic observation as a positive exception.

When, on the other hand, a negative observation is covered, the system outputs a revised theory that restores consistency by performing one of the following actions:
- adding positive information able to characterize all the past positive observations (and exclude the problematic one) to the rules that concur to the example coverage;
- adding negative information to discriminate the problematic observation from all the past positive ones to rule in the theory that covers the problematic observation;
- adding the problematic observation as a negative exception.

4 The Framework

The general framework we propose, depicted in fig. 2, is made up of a module devoted to the image processing based on morphological operators, for the morpho-structural features extraction task and a module for the knowledge extraction form both clinical and image-based data, that exploits multi-relational learning techniques.

4.1 Knowledge Extraction

The knowledge extraction step involves the definition of the data representation of both clinical and morpho-structural features extracted from the images and the identification of the multi-relational learning algorithms to build a model able to solve the issues concerning the predictivity of the goodness of the oocytes to choose as the most promising for the specific task of fertilization and the identification of similarities among situations such as stimulation protocols under specific patients’ health conditions.

As to the data representation, the general information on a patient and the clinical data about the couple diseases before the ICSI procedure starts is followed by the set of data describing the ovarian stimulation protocol carried out according to the couple’s characteristics and by a set of clinical data of the patients observed after the therapeutic plan has been taken place. Successively, the data about the oocyte aspiration phase is introduced. For each patient a set of “i” (a value varying from one patient to another) oocytes is obtained and each oocyte is described according to the own automatically extracted morpho-structural features. Hence, once the images are elaborated, the information extracted along with the clinical data are automatically represented in a multi-relational description language as reported in Table 1: for each entity involved in the domain, i.e. patient, ovaric stimulation protocol, hormon pharmacological treatment, oocyte aspired and oocyte components, a set of descriptive attributes are reported along with the existing relationships (italic font in the table).

As to the knowledge extraction phase, we propose the exploitation of two submodules devoted to specific tasks to solve. The first one concerns the possibility to apply clustering techniques to identify similarities among situations. Indeed, the aggregation of patients that show a similar behavior could be useful to better understand the conditions under which a pregnancy could be obtained. Once the clustering has been taken place, for each cluster it is possible to induce a set of rules able to identify relationships between stimulation protocol and health conditions or between number and quality of oocytes obtained. Thus the model can support the clinicians in the stimulation protocol definition for new unseen patients that show a similarity degree with a cluster. On the other hand, the induction mechanism would infer a set of rules to automatically classify unseen oocytes for similar patients in order to support oocyte selection.
Due to the complexity of data in our application domain, the multi-relational techniques previously presented were exploited. In particular, we use (APAM) \cite{11} as it is very robust with respect to the existence of outliers (i.e., data points that are very far away from the rest of the data points). This is a fundamental characteristic for our application domain as clinicians can adopt very different stimulation protocols according to their experience and, more importantly, according to the patients’ health problems and characteristics. Furthermore, the APAM algorithm is based on an approximate evaluation of the clustering membership thus allowing to tackle the uncertainty in the data.

As to concern the inference process, the incremental multi-relational inductive logic system \cite{13} was exploited as its incremental capability makes it able to learn a satisfiable model even with few examples and, more importantly, to revise the learned rules as new examples are provided without restart the learning step from scratch.

### 4.2 Image-based features extraction

The features extraction module is oriented to extract some relevant morpho-structural features from oocyte images, such as the measures of oocyte and cytoplasm diameters. This can be addressed as an image segmentation problem. Since we are interested in extracting the shape of the oocyte from the image, we employ a segmentation method better suited for shape analysis, that is mainly based on mathematical morphology \cite{14}.

Basic concept of mathematical morphology is the structuring element on which the operators are based. Given a two-dimensional binary image $X$, that is a subset of the 2-D integer space $Z^2$, a structuring element is a particular set $B \subset Z^2$, that gets translated over the image $X$ and whose relationships with this image are studied at each location. In the following, we denote $B_x$ the translation of $B$ by $x$: $B_x = \{b + x \mid b \in B\}$.

The basic operations of mathematical morphology are dilation and erosion. The dilation of an image $X \subset Z^2$ by a structuring element $B$, denoted by $\oplus$, is the set of points $x \in Z^2$ such that the translation of $B$ by $x$ has a non-empty intersection with set $X$: $X \oplus B = \{x \in Z^2 \mid X \cap B_x \neq \emptyset\}$. The erosion of $X$ by a structuring element $B$, denoted by $\ominus$, is the set of points $x \in Z^2$ such that the translation of $B$ by $x$ is included in $X$: $X \ominus B = \{x \in Z^2 \mid B_x \subseteq X\}$.
From the erosion and dilation operators, two fundamental morphological operations can be derived as follows: the opening of $X$ by $B$, denoted by $\circ$, is the union of all the translations of the structuring element that fit inside the image $X$, i.e. $X \circ B = \bigcup \{B_x | B_x \subset X\} = ((X \ominus B) \oplus B)$. The dual operation of the opening is the closing, denoted by $\bullet$, which is defined as: $X \bullet B = ((X \oplus B) \ominus B)$.

On such operators we designed a procedure able to extract the region containing the oocyte, and its diameter, taking out elements which are not of interest (e.g. the holding/injection pipettes, that are visible in many images), along with a good approximation of the cytoplasm diameter. Specifically, the proposed procedure works as follows.

**Preprocessing.** The preprocessing consists of finding edges in the image, by means of Sobel operator [15] and successively binarize the result (see fig. 3). After the binarization, elements that are not of interest surrounding the image borders, such as the holding pipette and the injection pipette, have to be taken out. This can be done by firstly selecting a point $p$ in the region of the border and, starting form it, by finding the connected components$^2$. This step uses an extraction of connected components algorithm [15] based on dilation and intersection of the set of pixels of the binary image.

**Oocyte region detection.** After preprocessing, the binary image shows segments of high contrast that do not quite delineate the outline of the object of interest (fig. 3d). Indeed gaps in the segments surrounding the object are evident. These gaps will disappear as soon as the image is dilated twice using circular structuring elements.

The dilated image shows the outline of the object quite nicely, but there are still holes in the interior of the object. The filling of these holes is performed by starting from a point $p$ in the region to fill and iteratively dilating it and intersecting the resulting dilation with the complement of the image to fill [15]. Finally, in order to make the segmented object look natural, the corresponding region is smoothed by an opening-closing operation with a circular structuring element.

Now, by subtracting the obtained image from the original one, the region of the oocyte on a black-background is achieved (see fig. 4a). Finally, in order to obtain the smallest rectangle that contains every point of the object, the center of mass of the oocyte region is calculated and, starting from it, the 4-directional Euclidean distances, until a pixel background is encountered, are computed (fig. 4b). The mean of these values represents the diameter of the oocyte region and the minimum and maximum $x$ and $y$ coordinates of the 4-directional Euclidean distances are the starting points from which to extract the bounding rectangular region containing the oocyte (fig. 4c).

**Cytoplasm region detection.** As according to medical literature the cytoplasm dimension is about the 66% of the oocyte dimension [16], this value can be used to approximate the cytoplasm diameter. A more accurate measure of the diameter of the cytoplasm has been obtained by considering that the shape of the cytoplasm can be approximated by a circumference. At this aim the Hough transform is applied to the binary image so as to detect the best circle fitting the shape of the oocyte cytoplasm. This has been done by searching for circles of radius $r$, varying from $(OD/2 - \delta)$ to $(OD/2 + \delta)$, where $\delta$ has been choosen equal the 10% of the oocyte dimension $OD$. The resulting cytoplasm detection is shown in fig. 4d.

---

$^2$Two pixels are connected in $S \subseteq A$ if there exists a path between them made up of pixels belonging to $S$. The set of pixels connected to the pixel $p \in S$ is known as connected component of $S$. 
Figure 4: Oocyte region detection. (a) Detected region (b) Oocyte Diameter (c) Oocyte region extraction (d) Cytoplasm detection

5 Evaluation and Discussion

The overall framework was tested on a set of data provided by the Department of Endocrinology and Molecular and Clinical Oncology of the University Federico II of Naples including clinical data of the patients along with the corresponding light microscope images of the oocytes. The dataset consisted of about 30 patients and 100 oocytes images. The image processing devoted to the extraction of morpho-structural features from the oocyte images was able to correctly extract the measures of the oocyte and cytoplasm diameters in the 90% of the cases with a good approximation with respect to the real diameters manually extracted.

As to concern the experimental outcomes of the applications of the multi-relational techniques, they revealed some interesting features that correlate health conditions to the stimulation protocol, that could confirm in some cases the medical literature, such as the identification of couples with some female infertility factors and severe male infertility factors for which the patients were subjected to a long stimulation protocol with respect to the other patients and obtain the greater number of oocytes with medium oocyte/cytoplasm dimensions (Rule 1 and Rule 3 in fig. 5). Furthermore, in couples characterized by male infertility factors the patients were subjected to a short stimulation protocol (Rule 2 in fig. 5). On the other hand, it was revealed that the number of aspired oocytes is in relations with the oocyte dimension, in the sense that many obtained oocytes (5-7) corresponds to a medium diameter of oocytes (156-164 $\mu$m).

This outcomes on preliminary experiments showed that multi-relational techniques could be able to grasp hidden relationships in data. Better results could be obtained by extending and considering the set of clinical data with more parameters in both the stimulation protocol and in the definition of health conditions, and by extending the image processing module in order to extract more features from the oocyte images and from other images that follow the oocyte development after fertilization, i.e. zygote and embryo images.

6 Conclusion and Future Work

In this paper an intelligent tool to support bio-medical applications is presented. The existing approaches work at different level according to the data starting point, i.e. images or clinical data. On the contrary, the proposed framework involves both data with the aim of put together the automatically extracted morpho-structural data of the image and the clinical data with the aim of further elaboration steps devoted to discovery relationships among data.

Future work will concern the extension of the clinical data to elaborate by considering more parameters in both the stimulation protocol and in the definition of health conditions, and the extension of the image processing module in order to extract more features from the oocyte images and from other images that follow the oocyte development after fertilization, i.e. zygote and embryo images. Finally, an exhaustive experimental phase is planned to be carried out.

Acknowledgment

This work is partially supported by the italian project “MIUR-FAR 08-09: The Molecular Biodiversity LABoratory Initiative (MBlab-DM19410).”
Rule 1
age_young(B),
basal_fsh_low(B),
oocyte_aspired_many(B),
female_infertility(B,D), type_thyroid(D,E),
stimulation_protocol(B,O),
hormon_stimulation_agGRH(O,P), specification_long(P),
hormon_stimulation_antagGRH(O,Q), specification_none(Q),
hormon_stimulation_RFH(O,R), specification_yes(R),
hormon_stimulation_UFSH(O,S), specification_none(S),
hcg_pred(O,T), hcg_dose_medium(T),
aspiration_timing_later(T),
is_oocyte_of(B,A),
is_oocyte_of(B,U),
dimension_ovo_medium(U),
is_oocyte_of(B,Y),
dimension_ovo_medium(Y),
is_oocyte_of(B,C1),
dimension_ovo_medium(C1).

Rule 2
age_hgh(B),
oocyte_aspired_medium(B),
male_infertility(B,K), type_oligo(K,L), specification_normal(L),
stimulation_protocol(B,O),
hormon_stimulation_agGRH(O,P), specification_short(P),
hcg_pred(O,Q), hcg_dose_medium(Q),
aspiration_timing_later(Q).

Figure 5: Sample learned rules

References