

An Association Rule-based CLIPS Program for Interactive Prediction of MSC Differentiation *in vitro*

Weiqi Wang,
René Bañares-Alcántara, Zhanfeng Cui
Department of Engineering Science,
University of Oxford, Oxford, OX1 3PJ. UK.
weiqi.wang1983@googlemail.com
{rene.banares, zhanfeng.cui}@eng.ox.ac.uk

Yanbo J. Wang
Information Management Center,
China Minsheng Banking Corp., Ltd.,
87606, Building No. 8, 1 Zhongguancun Nandajie,
Beijing, 100873. China.
wangyanbo@cmbc.com.cn

Frans Coenen
Department of Computer Science,
University of Liverpool,
Ashton Building, Ashton Street, Liverpool, L69 3BX. UK.
coenen@liverpool.ac.uk

Abstract— In this paper, a software toolkit has been developed for *in silico* prediction of the differentiation destiny of Mesenchymal Stem Cells (MSCs) *in vitro*. The software toolkit was developed in CLIPS (C Language Integrated Production System) as an expert system, with a java-based GUI. This toolkit utilizes the rules obtained from previous experimental data via data mining techniques, based on which the prediction is to be made. Thus, the prediction accuracy can be affected by the amount and quality of the rules, which can be improved by both manual adjustment and expansion of the MSC differentiation database in future.

Keywords- mesenchymal stem cell, cell differentiation, CLIPS, data mining, rule based.

I. INTRODUCTION

Mesenchymal Stem Cells (MSCs) are important to tissue engineering and stem cell therapy due to their pluripotent differentiation potentials both *in vivo* and *in vitro* [1], and have become one of the most studied stem cells in the last century. The pluripotency of MSCs includes osteogenesis, chondrogenesis, adipogenesis, myogenesis, tendogenesis, and neurogenesis, besides trans-differentiation [2]. Other discoveries regarding plasticity and immunologic properties of MSC have further increased the interest in their clinical applications [3]. The significance of the application of MSCs in clinical therapy has triggered an urgent need for the prediction of MSC differentiation.

A large number of studies have been carried out with the aim of understanding and predicting MSC differentiation [4]. However, different experiments have adopted different culture conditions under which MSCs grow and differentiate. Those culture conditions include donor species, culture medium, supplement and growth factor, culture dimension (monolayer vs. 3D culture), substrate (for monolayer culture) vs. scaffold (for 3D culture), etc [5]. They resulted in a large yet scattered spectrum of MSC differentiation scenarios and the discrete nature of available data structure for MSCs, which motivates our previous research [6] where those MSC data were analyzed by the Classification Association Rule Mining (CARM) approach [7], with the help of an online

database containing essential experimental data regarding MSC differentiation scenarios [8]. In our latest study [9], a CARM algorithm called CMAR (Classification based on Multiple Association Rules) [10] has been successfully used to obtain rules with useful information on MSC differentiation.

In this study, we aim at utilizing the rules obtained from CMAR (referred as “CMAR rules” below) for prediction on MSC differentiation *in silico*. For this purpose, a rule-based and object-oriented programming tool called CLIPS (C Language Integrated Production System) [11] has been selected to be the coding environment. After the integration of previously obtained CARM rules into the CLIPS routine, together with a java-based GUI, a software toolkit for computational prediction on MSC differentiation has been produced as its first version.

II. PROGRAMMING BACKGROUND: CLIPS, EXPERT SYSTEMS AND RULE-BASED PROGRAMMING

CLIPS is a programming tool which provides a complete environment for the construction of rule and/or object based expert systems [12]. CLIPS was created in 1985, and has now been widely used throughout the government, industry, and academia for its advantages in knowledge representation, portability, extensibility, verification/validation, etc [11].

Unlike traditional programming languages, such as FORTRAN and C that are designed and optimized for the procedural operation of data or digits, CLIPS allows the modeling of information at higher levels of abstraction, which simulates the way by which humans solve complex problems. As a product of the development of artificial intelligence, CLIPS allows programs to be built in a way that they closely resemble human logic in their implementation and are therefore easier to be developed and maintained [12]. “These programs, which emulate human expertise in well defined problem domains, are called expert systems” [13].

Rule-based programming is one of the most commonly used techniques for developing expert systems. In this programming paradigm, rules are used to represent heuristics, which specify actions to be executed for a set of given conditions [13]. A rule consists of an “IF portion” which is a

series of patterns specifying the conditions (often referred as “facts”) which make the rule applicable, and a “THEN portion” specifying actions to be performed once the rule becomes applicable. The process of matching “facts” to patterns is called pattern matching. The expert system tool provides a mechanism which automatically matches “facts” against patterns and determines which rules are applicable. Pattern matching always occurs whenever changes are made to “facts”, thus the actions of applicable rules can be dynamically updated on the execution list, and executed when being instructed to.

In this study, the heuristics that the rules in CLIPS represent is the CMAR rules derived from data mining techniques in our previous study [9]. As CMAR was applied to the data in the online database, hundreds of CMAR rules were obtained and analyzed. These rules were then filtered according to our pre-knowledge in lab, with the portion that we found not to have scientific sense abandoned. The remaining rules were implemented into CLIPS routine after being transformed according to the CLIPS syntax, as elucidated below.

III. PRODUCTION OF THE MSC DIFFERENTIATION PREDICTION TOOLKIT.

A. Outline of the Toolkit

The MSC differentiation prediction toolkit was built on the modification of a CLIPS sample program “WineDemo”, as a demo in the CLIPSJNI (CLIPS with Java Native Interface) package v0.2¹ for CLIPS v6.3. The toolkit consists of two pieces of sub-routine: 1) a CLIPS routine containing the integrated CMAR rules based on which predictions can be made and 2) a java routine constructing the GUI for the toolkit, a snapshot of which is shown in Fig. 1. The current version of the toolkit is v1.0, which can be freely downloaded².

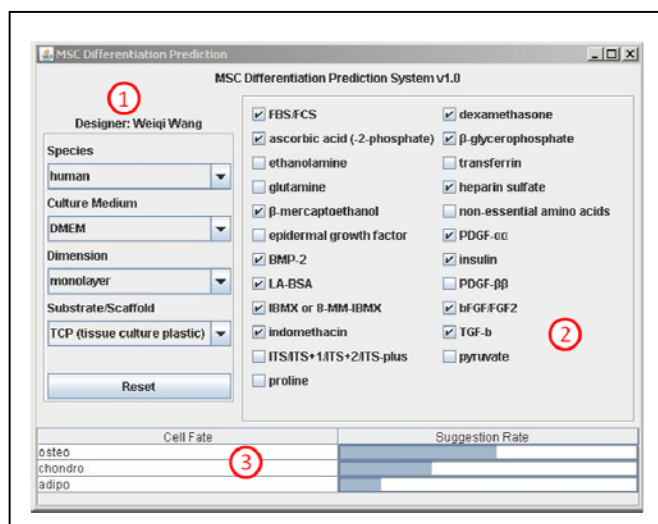


Figure 1. A snapshot of the GUI for the MSC differentiation toolkit. (①: single choice panel, ②: multi choice panel, ③: result panel)

B. GUI of the Toolkit

As shown in Fig. 1, the GUI consists of three major components: single choice panel, multi choice panel and result panel. The single choice panel contains four parameters which have been claimed to be essential information on MSC: donor species of MSC, culture medium to MSC, culture dimension and substrate/scaffold on which MSCs grow. In the current version (v1.0), options for the values of these parameters in the single panel were listed in Table 1. The multi choice panel contains 23 types of chemical reagents as supplements to culture medium, including FBS (Fetal Bovine Serum)/FCS (Fetal Calf Serum), dexamethasone, insulin, ascorbic acid, etc. The result panel is the place where predictions are shown, together with their respect Suggestion Rate (SR). Users should be reminded that predictions given by this toolkit can be more than one, each having a suggestion rate as its complement. The predictions shown in the result panel of the GUI were referred as “DPs” (Deliberate Predictions) below, with an intention to differ from “CPs” (Candidate Predictions), as elucidated below.

C. Integration of CMAR rules into the CLIPS routine

As indicated above, the toolkit was developed as an expert system for the human expertise embedded in the CLIPS routine, which is a selected portion of the 295 CMAR rules obtained from our previous study [9]. The CMAR rules were derived from the online MSC database³, as partially listed in Table 2.

According to our pre-knowledge in lab, most of the CMAR rules contain useful information on MSC differentiation, and were regarded as validated rules in this study. However, some rules does not have scientific sense, thus should be abandoned before integrated into the CLIPS routine. For example, Rule #294: {human} -> {proliferation} [58.04%] says that human MSCs do not differentiate, which is not scientifically sensible according to our pre-knowledge because the lineage to which MSCs committed should be predominantly directed by culture medium and supplements,

TABLE I. OPTIONS FOR THE VALUES OF THE PARAMETERS IN THE SINGLE PANEL

| Parameters in the single panel | Options |
|--------------------------------|--|
| donor species | human, rat, mouse, rabbit |
| culture medium | DMEM (Dulbecco’s Modified Essential Medium), DMEM-LG (DMEM with Low Glucose), DMEM-HG (DMEM with High Glucose), α -MEM (α -Minimum Essential Medium), RPMI 1640 (Roswell Park Memorial Institute Medium 1640), IMDM (Iscove’s modified Dulbecco’s medium) |
| culture dimension | monolayer or 3D |
| substrate/scaffold | Substrate: TCP (tissue culture plastic), gelatine-coated plastic and ornithine-fibronectin coated plastic Scaffold: none |

¹ Available at <http://clipsrules.sourceforge.net/CLIPSJNIBeta.html>.

² Available at http://www.oxford-tissue-engineering.org/MSCPrediction/MSCDiff_CLIPSJNI.rar

³ <http://www.oxford-tissue-engineering.org/forum>

TABLE II. A PARTIAL VIEW OF THE CMAR RULES DERIVED FROM [9]

| CMAR Rules | |
|---|--|
| (1) {DMEM + dexamethasone + β -glycerophosphate} -> {osteo} [100.0%] | |
| (2) {DMEM + "ascorbic acid (-2-phosphate)" + dexamethasone + β -glycerophosphate} -> {osteo} [100.0%] | |
| (3) {human + DMEM + "ascorbic acid (-2-phosphate)" + dexamethasone} -> {osteo} [100.0%] | |
| (4) {human + DMEM + dexamethasone + β -glycerophosphate} -> {osteo} [100.0%] | |
| (5) {DMEM + FBS + β -glycerophosphate} -> {osteo} [100.0%] | |
| | |
| (33) {human + "ascorbic acid (-2-phosphate)" + insulin + 2D + TCP} -> {chondro} [100.0%] | |
| (34) {human + "ascorbic acid (-2-phosphate)" + insulin + TGF- β + 2D + TCP} -> {chondro} [100.0%] | |
| (35) {indomethacin} -> {adipo} [100.0%] | |
| (36) {indomethacin + 2D} -> {adipo} [100.0%] | |
| (37) {"IBMX or 8-MM-IBMX"} -> {adipo} [100.0%] | |
| | |
| (291) {BMP-2 + TCP} -> {osteo} [58.33%] | |
| (292) {TCP} -> {proliferation} [58.12%] | |
| (293) {2D + TCP} -> {proliferation} [58.12%] | |
| (294) {human} -> {proliferation} [58.04%] | |
| (295) {2D} -> {proliferation} [56.65%] | |

rather than animal species. Thus, 13 CMAR rules which do not contain information on culture medium or supplement, such as Rule #294, were filtered out and abandoned. 283 rules remained after the filtration and were integrated into the CLIPS routine according to the CLIPS syntax to act as the heuristics in expert system. After the integration, the antecedent of a CMAR rule which represents culture conditions of MSCs becomes the "IF portion" of the corresponding CLIPS rule, and the class of the CMAR rule, i.e., the prediction to the MSC differentiation fate based on this specific CMAR rule, becomes the "THEN portion".

In this study, the CLIPS routine contains 283 CLIPS rules, which are the respect 283 CMAR rules. These CMAR rules represent a sub-portion of the MSC differentiation pattern in reality, as they were derived from the current MSC database which contains the experimental data. On the other hand, user-manipulated rules can always be added into the CLIPS routine to interfere the prediction, whereas in the current stage we only use these CMAR rules.

D. Working Mechanism of the Toolkit

The working mechanism of the toolkit was shown in Fig.2. The Toolkit needs the support of CLIPSJNI package to get started. Once started, the GUI of the toolkit will appear and simultaneously start the CLIPS routine in the background. The GUI is implemented with a real-time listener so that any update of the user input (e.g., selection/un-selection of a checkbox) will be immediately noticed and transferred to the CLIPS routine. The CLIPS routine will not

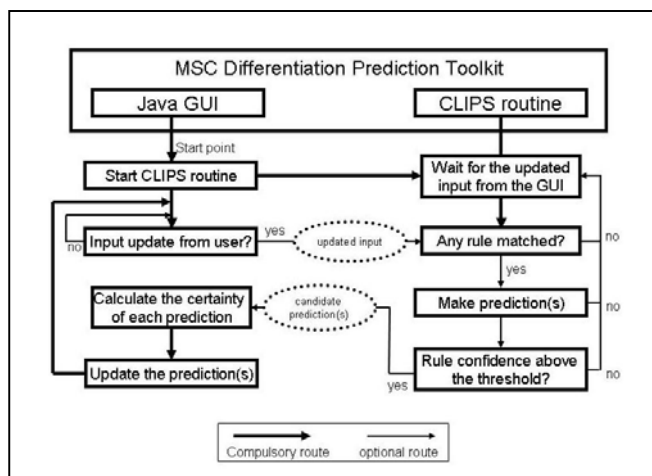


Figure 2. The working mechanism of the toolkit.

take action until receiving the updated input from the GUI. Once an input update is received, the CLIPS routine initiates the pattern matching mechanism of CLIPS, during which the received input from the GUI is treated as a "fact". If no rule is matched after the pattern matching, the CLIPS routine remains idle; otherwise all the matched rules become applicable, followed by their actions executed, which is to make their respect predictions. These predictions are then checked one by one for their confidence values and anyone with a confidence value lower than a given threshold is filtered out. The remaining predictions are transferred to the GUI as candidate predictions (CPs). Hence, the threshold is named "CP threshold". In the current version, the CP threshold was set to be 25% as a sample value, which can be changed according to the necessity in future. DPs, which are to be shown on the GUI result panel, are then generated out of all the CPs, with their respect suggestion rates calculated by the underlying java routine of the GUI (details for the calculation were elucidated below). In the last step, the DPs with their respect suggestion rates are updated to the GUI as the prediction results of the toolkit.

E. Calculation of the Suggestion Rate

The calculation of suggestion rates of DPs is based on confidence values of CPs. Readers are reminded that DPs made by the toolkit can be more than one, each with its corresponding suggestion rate. The reason for the generation of multi-DPs as the prediction results is that, in reality, chemical agents may induce MSCs to differentiate along one lineage while inhibiting them against others, the interplay of which makes it difficult to guarantee the differentiation fate of MSCs in the end. Thus, it is reluctant to stick on only one DP without taking account of other possibilities. As a consequence, an alternative way is adopted in this study, which is to include all the possible differentiation fates of MSCs as multi-DPs.

The generation of DPs is dependent on the composition of CPs. In this study, CPs derived from applicable rules in the CLIPS routine can cover up to four categories due to the four classes covered by the 283 CMAR rules, which are

“osteogenesis”, “chondrogenesis”, “adipogenesis” and “proliferation without differentiation”, respectively. Thus, CPs from different rules may differ from each other, with some of them overlapped meanwhile. As a consequence, two possible scenarios for the composition of CPs can be summarized as follows:

- Scenario 1 – homogeneous CPs

In this scenario, CPs consist of one same category (e.g., “osteogenesis”), yet their confidence values can be different. As a result, only one DP will be generated, which is same to the CPs. For example, if all CPs suggest “osteogenesis”, the only DP is also “osteogenesis”, with its suggestion rate (SR_O) calculated as follows:

$$SR_O = MAX(Co_i) \quad (1)$$

where Co_i is the confidence value of each CP suggesting “osteogenesis”.

The Eq. (1) is applicable to all DPs (SR_O for “osteogenesis”, SR_C for “chondrogenesis”, SR_A for “adipogenesis” and SR_P for “proliferation without differentiation”).

- Scenario 2 – heterogeneous CPs

In this scenario, CPs consist of more than one category. In this scenario, all the CPs suggesting “proliferation without differentiation” will be ignored when generating DPs. The reason is that if MSCs can be induced into any types of differentiation, the “proliferation without differentiation” must be a false prediction. Consequently, the calculation for suggestion rates in this scenario only applies to the other three categories of predictions (i.e., “osteogenesis”, “chondrogenesis” and “adipogenesis”). As an example, suppose the CPs consist of all the three categories “osteogenesis”, “chondrogenesis” and “adipogenesis”, in the current version (v1.0) of the toolkit, the formula for calculation of suggestion rate for “osteogenesis” (SR_O) is set to be:

$$SR_O = MAX(Co_i) \cdot \frac{\sum Co_i}{\sum Co_i + \sum Cc_j + \sum Ca_k} \quad (2)$$

where Co_i , Cc_j , Ca_k is the confidence value of each individual CP suggesting “osteogenesis”, “chondrogenesis” and “adipogenesis”, respectively.

In Eq. (2), if either $\sum Cc_j = 0$ or $\sum Ca_k = 0$, then the equation implies the situation where only two categories are covered by all CPs. Similarly, if $\sum Cc_j = 0$ and $\sum Ca_k = 0$, then Eq. (2) becomes Eq. (1), which implies the scenario 1.

Eq. (2) is applicable to the DPs of “chondrogenesis” and “adipogenesis”, but not “proliferation without differentiation”.

F. Summarization

In this section, the production of the MSC differentiation prediction toolkit has been described. A CLIPS sample program “WineDemo” was chosen to be the antetype for the toolkit, with its GUI reconstructed. 283 CMAR rules were integrated into the CLIPS routine after being filtered according to our pre-knowledge on MSC differentiation. Users are reminded that all the CMAR rules utilized in the current version of toolkit were derived from

experimental data *in vitro*, thus the DPs made by the toolkits are for MSC differentiation *in vitro* only. The calculation of suggestion rate to DPs was designed and implemented into the java routine, which dominates the working mechanism of the toolkit. After the toolkit was built, tests have been made to examine its performance, the results of which were elucidated in the next section.

IV. TESTS FOR THE TOOLKIT AND RESULTS

The MSC differentiation prediction toolkit can be run on most operation systems with the support from the CLIPSJNI package (included with the toolkit), and users have to install java on their computers before running the program. To run the programme, go to the directory of “Toolkits\MSCDiff Toolkit” in the downloaded package. Windows XP users can simply click “run.bat” file in this directory, or type the following command in the “Command Prompt” application (select Start > All Programs > Accessories > Command Prompt):

```
java -cp .;../CLIPSJNI.jar -Djava.library.path=../ MScDiff
```

For Mac OS X users, type the following command in the “Terminal” application (located in the Applications/Utilities directory)

```
java -cp ../CLIPSJNI.jar -Djava.library.path=../ MScDiff
```

For users of other operation systems, please refer to “instructions.pdf” for instructions.

After the toolkit has been started, tests were executed to examine its performance. All tests were performed on Microsoft Windows XP Professional (Service Pack 2), with CLIPS (Quicksilver Beta) and java 1.6.0 installed. One test was chosen to be an example, the result of which was shown in Fig. 1. In this test, several culture conditions were selected for the prediction to human MSC differentiation in monolayer culture. These culture conditions included culture medium of DMEM, substrate of TCP and several randomly chosen growth factors as supplements to the culture medium. The DPs given by the toolkit included chondrogenesis, osteogenesis and adipogenesis, with a descending order in terms of their suggestion rates. To check the causes to the DPs (i.e., the applicable rules and the consequent CPs), we embedded a debugging switch in the java routine. After the switch was turned on, it was found that 35 rules suggesting “osteogenesis”, 48 rules suggesting “chondrogenesis”, 6 rules suggesting “adipogenesis” and 2 rules suggesting “proliferation without differentiation” became applicable, out of the 283 rules in total. However, as elucidated in the working mechanism of the toolkit, the DPs in this test did not include “proliferation without differentiation”. The respect suggestion rates for the three DPs were calculated according to Eq. (2).

V. DISCUSSIONS

A. Design of Eq. (2)

The calculation of suggestion rates is done using Eq. (2). Although we are aware that Eq. (2) has limitations, it is extremely difficult or even impossible to design a perfect equation. It was decided that a “weight factor” for each DP

ought to be calculated in terms of: 1) the number of its supporting rules; and 2) the confidence value of each supporting rule. In the current version of the toolkit the “weight factor” is affected by the sum of confidence values of the rules supporting the corresponding DP and the sum of confidence values of all applicable rules.

B. The CP Threshold

As mentioned in Section III D, a CP threshold was set in the CLIPS routine as a part of its working mechanism. In this study, the threshold value is set to be 25%. The reason is that there only 4 differentiation destinies were taken into account, which are osteogenesis, chondrogenesis, adipogenesis and proliferation without differentiation; thus, any rule with a confidence value lower than 25% is regarded to be incompetent in a sense that it performs no better than a random guess, hence it is eligible to contribute to CPs. However, because each of the 283 CMAR rules has a confidence value higher than 50%, the threshold value of 25% in this study is just to show an example, which can be changed in future according to the need.

C. Evaluation of the Toolkit

As the DPs with their suggestion rates are generated according to the 283 CMAR rules, the accuracy of each DP is affected by the accuracy of each CMAR rule that supports it. Unfortunately, due to the mechanism of Weighted Chi Squared (WCS) [14] adopted by the CMAR algorithm, the accuracy of an individual CMAR rule is not retrievable. However, as reported in [9], the average accuracy of all the 295 CMAR rules (including the abandoned 13 rules) is 90.4%, which is satisfactory. Another factor affecting the DP accuracy is the size of the online database from which the CMAR rules were abstracted. The current database has 501 records and will be constantly expanded. With the expansion of the database, more CMAR rules are expected to be explored, with an intension to improve the accuracy of DPs. Additionally, DPs made by the current version of the toolkit do not involve consideration of the doze of the chemical reagents that are supplemented to the culture medium, because the CMAR rules do not have the regarding information. This is due to the fact that the current available data on MSCs are extremely limited. Neither the involvement of different chemicals in the intracellular pathways related to MSC differentiation nor the sensibility of the doze of chemicals to those pathways is clearly known in the current stage. Thus, DPs given by the current version of the toolkit should be regarded as an intuitive suggestion rather than a rigorous diagnose. With the development of human’s knowledge on MSC differentiation and the consequent expansion of the database, information on the doze of chemicals is expected to be taken into account in future versions of the toolkit.

VI. CONCLUSIONS AND FUTURE WORK

In this study, a software toolkit for prediction of MSC differentiation was developed on the antetype of a CLIPSJNI sample program. This toolkit utilizes 283 integrated CMAR rules obtained from our previous study to make predictions

on user-given conditions. The performance of the toolkit, as its first version, was tested and showed a satisfactory function in terms of guidance to MSC culture. With the framework of the toolkit settled, its power depends on the amount and quality of the integrated rules, which can be improved by both manual adjustment and expansion of the MSC differentiation database in future.

ACKNOWLEDGMENT

The authors would like to thank Prof. Jian Lu from the School of Physics & Astronomy at the University of Manchester, and the following colleagues from the Department of Engineering Science at the University of Oxford for their valuable suggestions to this study: Dr. Cathy Ye, Paul Raju, Dr. Shengda Zhou, Dr. Renchen Liu, Nuala Trainor, Zhen Lu, and Jinnan Zhang.

REFERENCES

- [1] A. Alhadlaq and J.J. Mao, “Mesenchymal stem cells: isolation and therapeutics”. *Stem Cells Dev*, vol. 13, 2004, pp. 436-48.
- [2] A.I. Caplan, “Mesenchymal Stem-Cells”. *Journal of Orthopaedic Research*, vol. 9, 1991, pp. 641-650.
- [3] R.J. Deans and A.B. Moseley, “Mesenchymal stem cells: biology and potential clinical uses”. *Exp Hematol*, vol. 28, 2000, pp. 875-84.
- [4] A.R. Derubeis and R. Cancedda, “Bone marrow stromal cells (BMSCs) in bone engineering: limitations and recent advances”. *Ann Biomed Eng*, vol. 32, 2004, pp. 160-5.
- [5] J.J. Minguell, A. Erices, and P. Conget, “Mesenchymal stem cells”. *Exp Biol Med (Maywood)*, vol. 226, 2001, pp. 507-20.
- [6] W. Wang, Y.J. Wang, Q. Xin, R. Bañares-Alcántara, Z. Cui, and F. Coenen, “A Comparative Study of Using CARM Approaches in Mesenchymal Stem Cell Differentiation Analysis”, in *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains*, A.V.S. Kumar, Editor. in press, IGI Global Publishing: USA.
- [7] F. Coenen, P. Leng, and L. Zhang, “Threshold Tuning for improved Classification Association Rule Mining”. *Proc. 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2005. Hanoi, Vietnam: Springer-Verlag.
- [8] W. Wang, Y.J. Wang, R. Bañares-Alcántara, Z. Cui, and F. Coenen. “Construction and Application of a Public-Domain Mesenchymal Stem Cell Database”. *Proc. IEEE 4th International Symposium on Biomedical Engineering (ISBME)*, 2009. Bangkok, Thailand.
- [9] W. Wang, Y.J. Wang, R. Bañares-Alcántara, F. Coenen, and Z. Cui, “Analysis of mesenchymal stem cell differentiation in vitro using classification association rule mining”. *Journal of Bioinformatics and Computational Biology*, vol. 7, 2009, pp. 905 - 930.
- [10] W. Li, J. Han, and J. Pei. “CMAR: Accurate and Efficient Classification based on Multiple Class-association Rules”. *Proc. 2001 IEEE International Conference on Data Mining*. 2001. San Jose, CA, USA: IEEE Computer Society.
- [11] J.A. Cannataci, “Liability for medical expert systems: an introduction to the legal implications”. *Med Inform (Lond)*, vol. 14, 1989, pp. 229-41.
- [12] P.H. Bartels and J.E. Weber, “Expert systems in histopathology: Introduction and overview”. *Anal Quant Cytol Histol*, vol. 11, 1989, pp. 1-7.
- [13] R.D. Semmel, “Expert systems: a classical introduction”. *J Clin Eng*, vol. 13, 1988, pp. 185-94.
- [14] F. Coenen, “LUCS KDD implementation of CMAR (Classification based on Multiple Association Rules)”. <http://www.csc.liv.ac.uk/~frans/KDD/Software/CMAR/cmar.html>. 2004. Department of Computer Science, The University of Liverpool, UK.