New Microbiologica, **30**, 35-44, 2007

Records matching model for data survey on applied and experimental microbiology

Salvatore A. Reina[1], Vittorio M. Reina[2], Eugenio A. Debbia[1]

[1]Laboratory of Experimental Microbiology and Epidemiology, DISCAT, School of Medicine, University of Genoa, Italy; [2]Freelance, ICT professional, Rome, Italy

SUMMARY

Experimental microbiology yields a huge quantity of raw data which needs to be evaluated and classified in a wide variety of situation from marine research, environmental pollution and pharmacokinetics of antimicrobial agents to epidemiological clinical trials on infectious diseases.

T is indispensable in all kinds of disciplines to validate, transform and correlate data clusters to demonstrate the statistical significance of results. Whether academic or biotechnological, the scientific credibility of a work is strongly affected by the statistical methods and their adequacy.

For a simple univariate analysis, many commercial or open source software products are available to perform sophisticated statistics for discriminant and multi-factorial analysis, but the majority of scientists use statistics partially. This is due to the high competence level required by a multivariate approach. It is known that the choice of a test, correct distribution assumption, valid experimental design and preliminary raw data validation are prejudicial to good science. All kinds of experimentation need analytical interpretation of descriptive evidence so that a classical numerical approach is not enough when raw data are not validated or incomplete.

Microbiologists always wish to quickly discriminate, or correlate, group and data clusters concerning clinical patient profiles, auditing multi-sensor derived numbers, monitoring biosphere indicators on either chemical and physical parameters or dynamics of microbe populations. Mathematical and statistical analysis is essential to distinguish phenotypes or constraints.

Data are in general stored in spreadsheet and database files which change continuously depending on the data collection and scope. We here propose a Records Matching Method (RMM) suitable for any kind of cluster analysis and pattern identification which can be used for either parametric or non parametric analysis without necessarily stating the pre-process statistical assumption on variable distribution.

The RMM is an application of a theoretical approach based on the Unique Factorisation Domain and is explained with an ideal generalisation model and then applied to a real-world microbiological study.

We used an easy mathematical formalism and discuss the possible application of the method as widely applicable to a plethora of taxonomic and phenetic investigations as well as for clinical trials and epidemiology.

Prototyping of the model for a computational automated process are also described in order to devise simple software which can infer on data using a heuristic rules file.

KEY WORDS: Records matching, Unique factorisation domain, Bioinformatics, Experimental microbiology, Statistical process control, Quality assurance, System audit

Received November 2, 2006 *Accepted* December 5, 2006

Corresponding author
Prof. Eugenio A. Debbia
Laboratory of experimental Microbiology and Epidemiology
DISCAT, School of Medicine
University of Genoa, Italy
E-mail: eugenio.debbia@unige.it

INTRODUCTION

The method here described, and its software functional specification, were thought to provide a simple tool for data calculation and experimental analysis in applied and experimental microbiology.

Data analysis achieved with the method is finalised to infer, group, filter or cluster data regardless of the statistical assumption so that it could be applied on either diagnostics, clinics or observational measurements.
Generally speaking data are a collection of records and groups of records are considered datasets. Such a scheme can be generalised to any record profile, thus a dataset is a table where rows are the samples under investigation and each column is a characteristic of the sample. The analogy of this scheme is typically a table of rows by columns and each column is called a "field" of the record (single row)
Besides the discipline and the specific domain, the dataset treated on microbiology and biotechnology needs to be analysed according to several pre-process tasks which allow scientists to sort, classify and categorise groups of records according to descriptive criteria; afterwards, it will be possible to evaluate statistics.
Very often one or more datasets indicate a set of records which share a common meaning and values (descriptive variable and scalar parameter respectively) and for basic science it is essential to discriminate or associate samples according to empirical criteria. A pre-process phase is indispensable especially when considering large datasets in that a validation of record integrity and coherent not null information of each field will impact on the credibility of results.
Usually, a dataset can be grouped and/or filtered starting from a database by utilising the Standard Query Language (SQL) (Reina *et al.*, 1991); this requires good statistics and a high degree of computer competence. Moreover the SQL is only effective in evaluating "identity" criteria rather than matching groups of records according to criteria such as "tolerance" and "proximity".
Applied and experimental microbiology imply several doubtfulness and *"fuzzy"* pieces of evidence [Hanai *et al.*, 2004; Reina *et al.*, 1994]. Sometimes the ability to approximate variables range can leverage the probability of a system adaptability (e.g. environmental sensor automation). In addition, it is restrictive to use a predetermined range of significance for variables and indicators because it would be preferable to dynamically calibrate a variable or a parameter with a "weight factor" which modulates the influence and consideration of that variable or parameter by virtue of the context.
We have already tested logic and mathematical models on several microbiological experiments concerning microorganism growth and taxonomy, Post Antibiotic Effect, MIC and genetics of quinolones (Cavallero *et al.*, 1987; Reina *et al.*, 1991; Reina *et al.*, 1993, Reina *et al.*, 1995) and these experiences led to a unified record matching model. In the specific cases of marine microorganisms identification in environmental polluted mud and HIV-eukaryotic cell interaction model (Reina *et al.*, 1987, Reina *et al.*, 1994) unsupervised Kohonen algorithms were also used (Ruggiero *et al.*, 1993). Almost every experiment design used software computation.
Many advanced software programmes are available to study microbiology with multifactorial and multivariate techniques for pattern matching such as Neural Networks, Bayesian nets and fuzzy logics. As already pointed out for high-performance statistical tools, artificial intelligence and reasoning software are complicated and cumbersome. but it would be desirable to be able to study similarities, proximity, phenotype varieties and cluster analysis in the everyday laboratory setting. We addressed this issue by creating a simple method based on a mathematical model for *cluster analysis* and *pattern matching*. The method is implemented in practice as a set of *software frameworks* which can be easily implemented by anyone regardless of the programming language and the dataset file format.
The method is called Records Matching Method (RMM) because it is formalised with a record profile metaphor and its typical application is based on the recursive comparison of records which can be clustered by means of an algorithm which uses a simple template file containing heuristic rules for each variable and parameter.
In order to let everyone create their own software, functional specifications are provided together with software documents and guideline web coordinates. Several real world applications were used by physicians for clinical trials and epidemiology applied on Assisted Reproductive Medicine ART, andrology and endometriosis surveys. In those experiments the theoretical model previously described (Reina *et al.*, 2006) was verified and tested for its simplicity and suitability so that it is now possible to provide a software front-end

specification for an intuitive easy to use tool with powerful cluster analysis capability.

MODEL AND METHODS

Both experimental and applied microbiology imply an articulated panel of factors analysis of scalar and descriptive information. Variables and parameters are generally referred to diverse typologies of scales and distribution, thus record-to-record comparison as well as dataset correlation equally need parametric and non parametric statistics.

We refer to a record as a pertinent set of information concerning a generic sample which is the object under investigation. Notoriously a record has a typical fields profile which is globally considered in our model as a unique factorisation index (FI). This index has the peculiarity of being at the same time a quantitative and a qualitative expression of that specific record which can be thought of as a fingerprint equivalent of the record as a whole.

If many records, hence a dataset, are serially calculated as an array of unique FI it will be possible to apply univariate analysis to a vector of values. This simplification transforms the study of complex rows by columns dataset to a series of indexes which can then be evaluated according to a heuristics previously defined by the scientist's empirical experience. The interaction between factorised dataset and "weighted" logics inside a heuristic file, will be the mean for which the theoretical model will allow a recursive correlation of FI values according a grouping criteria with dynamic and programmable range criteria. After all, the method will be represented by the ability of associating (or discriminating) samples by reason of their affinity and similarity simply because it is able to determine how much the records are diverse. We shall see that diverse could be analogously considered with the concepts of "*weighted distance*" of two overlapped records *fingerprints* (mathematical abstraction of a pattern).

Because it can compare records contiguity or closeness, the model finds *which,* and estimates *how much*, a subset of records in a wider dataset table, is phenetically similar to a given record called Master Profile (MP). Generally, an MP can be a reference record which is either newly inserted in the database or is one already registered record which is assumed to be a significant paradigm.

An essential step in promoting the model to a method is the definition of a heuristics logic which describes a priori the relevance of each field in the record profile. Before concepts such as correlation, association and dependence can be applied to datasets it is necessary to determine the sense and concurrent relationships between variables relevance. In order to generalise the use of the method as much as possible we shall refer to variables and parameters with homologous fields of a record profile in that their contribution coincides with the descriptive element of our sample. On a practical basis, the fields are the columns of a data table or a spread-sheet and this work will use this scheme to better explain both mathematical model and easily applicable method.

Any experimental discipline uses a variety of analyses on descriptive science based on categorical information formalised in a table where rows represent records (set of studied samples) and columns represent the characters of a sample (informative units, IU). Mathematically, a data table can be formalised ad a matrix of **r** by **c** (***r*** x ***c,*** rows by column) and our aim is to substitute the matrix with a vector containing a series of values equivalent to each row or record.

The transformation cited above is possible with the Unique Factorisation Domain theorem (Artin *et al*., 1991, Dummit *et al*., 1999) which profits from a set of trained matrices containing the relative weights of the fields of a record so that all the range of all the possible values assumed by a field have to be classified. In fact, the matrices will be used to determine the relative distance (*weighted distance*) between the homologous field of two records when compared and computed for their record FI.

An especially useful feature of the factorisation technique is to "summarise" and "persist" a quantitative and qualitative expression of similarity in a two-records comparison by means of a *delta* value which sums the contribution of each single field comparison with its corresponding one on the opposite record.

We now introduce the definition of Matching Level or ML as the value achieved each time a

record-to-record comparison is complete; when operated recursively, this process originates a vector (one-dimension matrix) with all the ML values derived from the difference of two FI values. Such array of ML value will be easy to aliquot, rank and cluster according to cut-off values and/or an arbitrary range of tolerances so that discrete bands of records can be distinguished to confine coherent groups of records on the basis of phenetic closeness and relevance similarity. In the simplest case we can divide two subsets with a cut-off in the middle to separate concordant and non concordant records. This process can be repeated with arbitary cut-offs to trace which samples fall within an acceptable level of similarity.

The use of factorisation gives the RMM a simple way of treating experimental data because the heuristic knowledge is empirically dynamically modified by the expert (heuristic rules file, HRF) so that it can be moulded and adapted to any experiment. Moreover, the calculation algorithm can be reiterated by systematically changing the HRF at every run and saving the corresponding results ML of FI delta's vectors. A *supervised* analysis on a well characterised sample control will be possible. This variant of the method is strictly related to the mathematical model demonstration (Reina *et al.*, 2006) which justifies an ultimate RMM usage to create calibration templates of HRF. Frequently, distinct groups of scientists share observational data typology collected with different survey, yet they wish to compare and evaluate data under a common impartial standard.

Because implementation of the method is easily translatable with a software acknowledged template of HRF, it can be utilised for large multicentre audits of *consensus* trials, still every group could save its own ability to filter, cluster and monitor its data according to specific experimental schemes.

Despite its theoretical simplicity the model of RMM can lead to sophisticated reasoning software applications. In short, the algorithm could indeed be run as a self-evaluation learning system; in such a case the process would be started without pre-defined HRF knowledge and historical repertoires could be scanned to derive a set of rules automatically by inferring on raw-data regardless of the stochastic and homoscedasticity assumption required by pre-process statistics.

A self-referential RMM system would lead to an ideal knowledge scanning system oriented meta and cluster analysis for epidemiology. At the present time these tasks can be accomplished with PCO, hierarchical and cladistical PCA generally available only in high-level statistical software packages.

DISCUSSION

Model theory and applied method

The mathematical treatise of the model and formal definition are discussed elsewhere (Reina *et al.*, 2006), while this work gives an explanatory exposure of the theory with a minimal use of mathematical formalisms in that practical examples will be oriented to experimental applied microbiology.

In order to be correctly applied, the method introduced so far has to be formally defined and modelled. Before a practical approach, we shall describe a totally theoretical example concerning the RMM. A hypothetical example is preliminary because will simplify comprehension and concepts realistic applicability.

We premise that the field of a record must contain non consecutive values with clearly non contiguous rank and meaning. It is also necessary to extend a definition of *weighted-distance* for each field which will give a direct measure of proximity or distance for a comparison of inter-fields as well inter-records entity.

Let us suppose that with the notation:

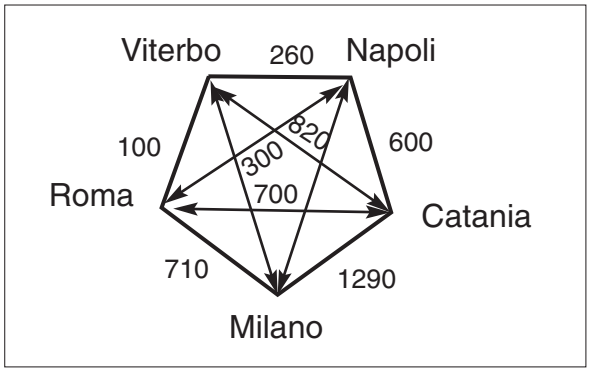
1) C_4 = CITY = {*Rome, Viterbo, Naples, Catania, Milan*} *(m=5)*

we shall refer to the fourth field of a generic record **R**. The field considered is the descriptive value of the name of an Italian city [CITY] while **m** is the number of possible values of the field so that it will be generically noted as ${}_{i}M$ or, to be adherent to the 1) example we shall have:

1a) ${}_{4}M = 5$, meaning with this the set of 5 possible values for the field CITY*;*

Weighted distance

It is possible to intuitively express a "w*eighted-distance*" between two values of the field [CITY] in terms of geographical distance expressed in kilo-



metres. Clearly, this *proximity* measure is reminiscent of the physical distance between cities. We now use a graph to visually represent the possible field's reciprocal relationships. Each arc of the graph subtends a value for every couple of cities. For readability, the graph and its arcs are not proportional, geographical distances are intentionally approximate and original Italian names of the cities are reported. The graph appears as a clear representation of the symbolic relationship of the cities with each other among those considered as possible values of the field in 1) formula.

Let use the notation G_i to globally indicate the graph of all the *"weighted-distances"* for each couple of values of a i*th* field C_i in a generic record R hence we define as:

2) d_i (j,k), j, for k=1.., $_iM$,

the *"weighted-distance"* between the j*th* and the k*th* value of the field C_i in R.

The graph G_i can be represented with its associate matrix defined as M_i containing the d_i (j,k) values for the field C_i in R:

T1)

	Rome	Viterbo	Naples	Catania	Milan
Rome	0	100	300	700	710
Viterbo	100	0	260		
Naples	300	260	0	600	
Catania	700		600	0	1290
Milan	710			1290	0

The matrix in T1) is symmetric and indeed d_i (j, k) = d_i (k, j), albeit a field typology is possible for which the possible values d(j, k) do not necessarily have a linear correlation: thus M_i would not be mirrored in its diagonal line. Each cell contains at the intersection of two possible values of the field the phenetic distance which can be interpreted as an index of affinity and similarity of two values among those possibly assumable by the field.

The example described intentionally uses the geographical distance to emphasise the concept prior to applying the scheme for more general kinds of information.

The Records Matching Method (*RMM*)

To consolidate the theoretical approach we now address the model to a more specific real case. As already stressed, the practical use of the RMM is highly flexible because it can be generalised to any kind of descriptive evidence as long the scientist defines an *a priori* knowledgebase which "informs" the algorithm on the relative significance of the information evaluated and classifies all its assumable values according to an indexed relevance.

We now describe a case of an agent-resistance experiment, but a microbiologist will immediately recognise a much large spectrum of investigation to which the RMM could be applicable with success. This case, taken as a paradigm, is simple but complete since all the possible types of experimental variables and parameters, including casting variants, are treated in detail.

Let us recall the formalism in 1) and consider a *set* of fields which taken together represent a record profile. The record and its composition of field values is obviously our sample. The RMM's finality is to compare two records by determining their affinity and measuring it with a matching level.

Before a record-to-record level of matching we shall explain a field-by-field matching level which is a propedeutic step; a sample is globally evaluated as a result of the single contribution of each of its characters whether it is a variable or a parameter (field).

Consider a record R (C_i) for i=0,..7 which is a sample of an experiment concerning the estimation of the Post Antibiotic Effect (PAE) under several cultural conditions. Briefly, in vitro bacterial growth can show variable fresh outbreaks after antibiotic exposure depending on cultural media and incubation time.

The information characteristic collected, our fields, were recorded to investigate parametric, non parametric experimental outcomes in rela-

tion to phenotype and genotypes. All indicators can also be associated with a descriptive field which records the growth.
Schematically the record's profile can be formalised as follows

S1)

C_0= PAE = {*0|0.10|0.11| 0.12| 0.13| 0.14|* ……|*1.0*}
C_1= PAERange = {*0-0.30 | 0.31- 0.50 | 0.51-0.60 | 0.61-0.90 | 0.91-1.2* }
C_2= Incubation = {*60 min | 120 min | 360 min | 480 min*}
C_3= Resistance = {*R | I | S*}
C_4= Antibiotic = {*Amoxicillin | Meropenem | Ciprofloxacin | Gentamycin | Cefotaxime*}
C_5= Phenotype = {### | *PenS | PenI | PenR | EryS | EryR* M | *ESBL*}
C_6= Genotype = {### | *Pbp | ermB | mefA | ermTR | TEM4*}
C_7= Growth = {### | true | false} or {### | - | +}

Each field gives the opportunity to explain all the cases for which the combination of values can be translated by the model in a unique *"image"* which is the consequence of the *weighted-distance* of each piece of information therefore we describe this pharmacoresistance experiment keeping in mind that any other kind of characters can be applied as well.
The first field *C0(*PAE*),* simply contains continuous values in a range of linear variability and the *weighted-distance* could be calculated very much as for the example previously shown in 2), thus considering a simple absolute *delta* between two values.
In practice the distance of the two records R1 and R2 for the field *C0*(R1) versus *C0*(R2) is the algebraic difference of the values assumed by the two fields, thus if:

3) *C0*(R1) =0.45 vs *C0*(R1) =0.27, then according to formalism in 2) d**(**0.45 **|** 0.27**) =**0.18

This first example concerns linear and continuous measures and as an obvious parametric variable the value itself can be appreciated as a direct measurement of geometric eucledian position. We shall soon see how the model will translate even attribute, binaries and categorical descriptive fields.

The second field *C1(*PAERange*),* again, belongs to the PAE but is expressed as discrete ranks of values rather then a variable single value. For microbiologists this attitude is reminiscent of the MIC in antimicrobial susceptibility experiments, which indeed could be treated in the same way. The field is clearly classified according to 5 ranks (restricted groups of values), so recalling the T1 matrix we can reproduce a second matrix T2 which symbolises the theoretical graph G2 (not reported).

T2)

	0-0.30	*0.31-0.50*	0.50-0.60	0.61-0.90	0.61-1.2
0-0.30	0	1	2	3	4
0.31- 0.50	1	0	1	2	3
0.51-0.60	2	1	0	2	3
0.61-0.90	3	*2*	1	0	1
0.91-1.2	4	3	2	1	0

The matrix shows the relations of the mutual combination of *weighted-distance* between two rank indexes. Recalling 2) we can adapt as follows:

3) C1(R1)= [0.61-0.90] vs C1(R2)= [0.31- 0.50]; thus d(3 | 1) = 2

In this case the delta value is calculated using the ordinal index of the position of the rank. This is reasonable also because the ranges of the ranks arbitrarily decided in their limits are nevertheless sorted in an ascendent way.
It will appear intuitive to microbiologists how the limits of each rank can be arbitrarily decided depending on the experimental needs. There are no prejudices on the way the scales can be split and no forced schemes for regular length. On the contrary, diverse grouping can be decided to intentionally emphasize specific ranges.
Therefore, the phenetic distance can assume all values between 0 and 4. It can be noted how the 0 value means that two records R1 and R2 are identical for the field *C1*; moreover, this latter implication shows a first important corollary of the model which demonstrates its coherence on the contour.
The field *C2* allows us to consider the case of sorted and discrete variables which however do not follow a linear function. For the field *C2 (*Incubation*)* expressed in minutes, the simple difference between values can be calculate in the way that the example 4) shows, but two impor-

tant aspects arise. Firstly the sign of the delta value can be taken into consideration with its negative value, and in a second instance not necessarily the different relative distances from one the indexed field position could reflect the meaningful desired by the investigator on reality.
Let us consider these two situations starting with the ordinary notation:

4) *C2*(R1)= [120] vs *C2*(R2)= [480] cioè d(120 | 480) = -360

Hence we have two possibile choices which can be adopted depending on the a priori empirical judgment of the scientist:

a) to use the delta value the way it is, meaning by this that the difference will be taken on absolute value

5) d(120 | 480) = -360 become |d(120 | 480) | = 360

b) to use a matrix of heuristic indexes to calculate a phenetic distance in a uniformed and predetermined way

As emphasized, the intervals taken as absolute values between consecutive incubation time have no geometrical regularity and there are no regular proportions in the values succession: the three delta values 60|240|120, obtained for the position 60-120-360-480, are non sorted (only crescent or descendent) and scraps are not linear.
To understand how the b) situation can be favourable, we take advantage of the matrices T3a e T3b proposed as follows:

T3a)

	60 min	120 min	360 min	480 min
60 min	0	1	2	3
120 min		0	1	2
360 min			0	2
480 min				0

T3b)

	60 min	120 min	360 min	480 min
60 min	0	4	8	15
120 min		0	4	8
360 min			0	4
480 min				0

The two possibilities will be treated to give different meaning relevance to the diverse experimental situations, but it will be processed by the RMM exactly in same way and the human role will be discriminative.

If we intend to get a linearity between incubation intervals, namely, they will all be considered at the same level and we shall want only cluster and qualitative distinction among various experiments we could use a heuristic table T3a and an example would be:

6) *C2*(R1)= [120] vs *C2*(R2)= [480]; thus **d(**3 | 1**) = 2**

This example clearly implies proportional increment deltas and the maximal separating factor would be

7) [max **d(**0|3**)**] = 3 corresponding to the extremes 60 and 480 minutes.

If the microbiologist prefers a more evident discrimination among incubation times, and even more, he wants to specifically decide which intervals are most relevant to the experiment's duration then an hypothetical heuristic matrix would be the T3b. The inter-distances scheme is identical to that in T3a albeit the weighted indexes were clearly chosen according to an exponential progression.
If we repeat step 6) by applying the T3b heuristics and maintaining the same field values we shall obtain:

7) *C2*(R1)= [120] vs *C2*(R2)= [480]; thus **d(**15 | 4**) = 11**

It appears evident how heuristic matrices can be arbitrarily rendered to fine tune the microbiologist's decisions which are based on the logics on his empirical experience.
Detailed field analysis of the model described so far on the first 3 fields is essentially the same for all the others hence we shall omit the formalism of the heuristic calculation to preferably exhaust all other types of information in the record structured profile designated in S1).
We more briefly complete the plethora of possible fields typology and their *weighted-distance* casting.
The field *C3* (Resistance) is a useful example of how attribute variables can be used as qualitative discrete inter-values.
In such a case it is not that relevant to conserve an ideal sorting along the three symbolic values (Resistant, Intermediate and Susceptible) therefore a simple linear heuristic matrix will adequately fit most cases in that there is no a priori preferable values direction.
For the field *C4* (Antibiotic) all the considerations already assured for the field *C3* are legitimate

since values are not scalar or oriented, but it is plausible to establish a special relevance to privilege one type of antimicrobial agent towards another. For instance, quinolone and cephalosporin could be considered very similar and therefore very close, when compared with ampicillin. This scheme could lead the RMM resolution to a better stratification for clustering purposes because records with ampicillin will tend to segregate more centrifugally in their phenetic score.

Fields C5 and C6 (Genotype and Phenotype) share all previous considerations for the descriptive variables except for the peculiar value Null or symbolically [###]. It is indeed possible that either genotype or phenotype would be unknown (or not definable). This important case, again, could be a subtle clue which needs to be brought to the foreground to perspicuously separate samples.

The C5 and C6 fields are also vital to understand a further concept of the RMM named Extended Matching Score or ExMS which makes it possible to extend the use of indexed w*eighted-distance* by combining two concurrent fields considered to be related in some way. This is exactly the case of the genotype and the phenotype fields in that the expectation of having a specific genotype associated with a phenotype is quite probable. Failing this evidence should raise doubts and it would be optimal to use RMM with an appropriate logics.

The ExMS is helpful in this case and is simple to apply because *delta* indexed values of two variables can simply be multiplied by a factor called "*enhancer*" when a predetermined combination of values belonging to associated fields will occur. We have given an exhaustive treatise on the issue to include with the model the concept of fields "*neighbourhood concurrency*" (Reina *et al.*, 2006).

Lastly, consider field C7 which specifically deals with the case of binary variables (TRUE/FALSE, YES/NO and symbolically +/-). Despite two possible values the RMM heuristic matrix would help in discriminating a third level of information because the Null possibility could be indexed and samples could be diversely interpreted. Possible Null values of a field could have several meanings such as n*ot measurable*, *unknown value* or *not trustworthy* data.

Factorial record index

After a basic level of abstraction which explained the intra-record (inter-field) the model can now be scaled up to an inter-record level.

The concept of *weighted-distance* applied to the fields relationship can be transferred to classify the entire record to bring forward the RMM properly defined which will cumulate all the variable's weights of each field of the record.

This mechanism aims to substitute a sample/record with a unique number which is together a quantitative and qualitative expression of that record.

By recalling the structure in **S1**) we obtain a set of fields representing a record **R** formalised as follows:

8) **R** = $\{C_0 \mid C_1 \mid C_3 \mid C_3 \mid C_4 \mid C_5 \mid C_6 \mid C_7\}$

eventually substituted with nominal definition

9) **R** = {*PAE* | *PAERange* | *Incubation* | *Resistance* | *Antibiotic* | *Phenotype* | *Genotype* | *Growth*}

We can express the content of the record **R** as an equivalent number called Factorial Record Index or FRI.

This number has a series of features that will be useful to give a qualitative and quantitative representation of the record.

By utilising the Unique Factorisation Domain approach (Artin *et al.*, 1991; Dummit *et al.*, 1999) it is possible to achieve a unique number and by reversing the algorithm to go back to all the values of the field of the original record (Reina *et al.*, 2006).

In this paper the FRI will be described with respect to only the practical suitability with the RMM. We recall that the sum of all the weighted-indexes derived from the matrical calculation of each field of a record (e.g. 8 and 9 formulas) is finalised to the comparison between two records.

Each field inside the record profile will have a "*weight*", all fields taken together, will result in an FRI.

We first define a table called Field Weights Table or FWT which is comprehensive of 3-dimensional arrays: the ordinal value of the field in **R** (its relative position in the record profile), its index value and its contained descriptive value.

With reference to what was defined in S1) and supposing all descriptive fields as already classified in a heuristic matrix like T2), we then have a table as follows:

T4)

Field Ordinal	Weight Index	Contained value	FWF
1	1	0.00 - 0.30	1
1	2	0.31 - 0.50	1
1	3	0.51 - 0.60	1,5
1	4	0.61 - 0.90	1
1	5	0.91 - 1.20	1
2	1	60 min	1
2	2	120 min	2
2	3	360 min	1
2	4	480 min	1
3	1	R	1
3	2	I	2,5
3	3	S	1
….	….	….	——
….	….	….	——
7	1	###	0
7	2	True	1
7	3	False	1

Dotted lines signify tacitly omitted fields between *C4* and *C6;* the scheme's meaning remains unaltered. The fourth column is a Field's Weight Factor or FWF and will be essential to manipulate a meticulous logic which differs the importance of one field towards others.

Every row of the table has a *weight* which act as a multiplicative factor so that the expression in 6) can be applied as the difference of two records **R1** and **R2** for that field; hence, that expression was **d(**3 | 1**) = 2** for the Incubation field and now would be revised according to table T4 as follows:

10) **d(**3 | 1**) * FWF (**2|2**) = 2 * 2 = 4**

Basically, when a record is a case of a 2 hours incubation, its relevance during RMM is double in terms of *weighted-distance* with other kinds of duration. This feature of the FWF is extremely important to understand how a scientist can freely design a heuristics made with detailed rules and set up a reasoning *template* for the algorithmic engine of the RMM. A weighted logics, adequately prepared for a specific set of information, is a sort of optical filter which will deflect experimental dataset and re-project it on a screen as a clustered map; in a way a metaphor of the trapezoid that filters a coherent light-wave and separates in wider coloured bands.

The case **FWF** (7|1) is zero, meaning by this that the Factorial Weight Index is also an effective mechanism to selectively exclude a field. This feature is useful when the investigator wants to run a RMM on a dataset considering only part of the record information; he will simply prevent the model from calculating.

The seventh field *C7* (**Growth**) in T4 is a special example because can show how coherent the model would be considering other borderline experimental situations. For instance when the detection of a value was not possible or is not available, this does not mean that there is no evidence of growth, simply the information is not available (e.g. automation and technical accidents). It is obvious that setting the symbolic value of [###]to zero will prevent the sample from being accidentally considered as [*false*], which instead means no growth.

As a last implication, the RMM ignores, namely will not compare, those records which have even only one Null FWF; only [True/False or +/-] are meaningful values.

CONCLUSION

The proposed model of the RMM is suitable to analyse experimental datasets in the daily microbiological routine. The method is finalised to cluster analysis and it represent a simple and customisable alternative to complex modelling software and sophisticated statistics.

Its use and effectiveness are linked to the investigator who decides an a priori set of rules to determine the association level of the experimental measures studied. The rules are represented by simple and intuitive knowledge tables for each variable or parameter of a record. The heuristics can be arbitrarily calibrated and adjusted so that the dataset can be scanned by the RMM algorithm which will recursively process matching records on sample tables.

Mathematical formalism of the model and its basic calculation algorithm are provided in the literature (Reina *et al.*, 2006). Thus the scientist who has programmatic skill can develop his own software program using any programming language.

Virtually any type of dataset and experiments can be processed, but for practical software implementation, the example of source code concerning the modelling discussed in this work is freely distributed by the authors to anyone who

wish to devise the software toolkit. The hope is that several other groups involved in different microbiological fields will adopt the RMM and test its efficacy.

ACKNOWLEDGEMENTS

Authors are grateful to the programmer Carlo Bergamini (Genoa) for the Delphi and MS-VB6 source code engineering and Franco Ameglio (Rome) for his revision on the microbiological aspects related to clinics.

REFERENCES

ARTIN FGM. Algebra, Prentice Hall (1991).

CAVALLERO A., REINA S., SCHITO G.C. (1987). Post Antibiotic Effect induced by Ofloxacin in both gram-positivi and gram-negative bacteria. *Chemoterapia*.

DUMMIT, D.S. FOOTE R.M. (1999). Abstract Algebra, Wiley.

HANAI T., HONDA H. (2004). Application of knowledge information processing methods to biochemical engineering, biomedical and bioinformatics fields. *Adv Biochem Eng Biotechnol*. **91**: 51-73. Review. PMID: 15453192 [PubMed - indexed for MEDLINE].

POLLERA C.F., AMEGLIO F., REINA S. (1987). Changes in serum iron levels following very high dose of cisplatin. *Cancer Chemotherapy and Pharmacology*.

REINA S., DEBBIA E.A., SCHITO G.C. (1991. Ciprofloxacin Induced Modulation of cellular grothw in activated, normal and lymphoid established Cell Lines. The antimicrobial agent resistances: orin treatment and control. Abs. Principato di Monaco. **70**; 25 5.

REINA S., DEBBIA E., SCHITO G.C. (1993). Evaluation of the post antibiotic effect induced by various antibiotics against Staphylocossi and Enterococci. *A.A.M.J.*

REINA S., DEBBIA E. (1993). Genetic recombination by spheroplast fusion in Escherichia coli K12. Cytobios by The Faculty Press. 76 91-95.

REINA S., DEBBIA E.A., SCHITO G.C. (1995). In Vitro Cellular Growth Modulation by quinolone conditioned medium. 93rd General Meeting, Atlanta, Georgia, USA. Session 120. Paper nu. I28.

REINA S., BOERI E., LILLO F., CAO Y., VARNIER E.O. (1991). Automation in AIDS research and diagnostic activity: a Local Area Network with Standard Query Language. 7th European Edition of Conference on Advanced Technology for Clinical Laboratory and Biotechnology. - ATB '91 26-11, B11.

REINA S., MIOZZA F. (1994). Knowledge Data Base System for Twins study. ACTA GENET MED ET GEMELLOL. Ed. Mendel Institute, Rome. **43**: 83-88.

REINA S., REINA V., GIACOMINI M., DEBBIA E. Bio-fouling and micro-organisms identification on polluted materials: a novel Knowledge Data Base System architecture for a heuristic expert system engine. Atti congresso Internazionale dei Biologi, 22-25 settembre 1994. Vieste.

REINA S. Il percent growth rate average (PGRA) migliora l'interpretazione dell'effetto post-antibiotico. 16mo Congresso AMCLI Nov 12-15, 1987.

REINA S., REINA V.M. DEBBIA E.A. Simple method for Records Matching for experimental and diagnostic datasets of patient's records. Berkely-press (April 2006). COBRA Preprint Series. Article 3 - http://www.biostatsresearch.com/cobra/ps/art3.

RUGGIERO C., GIACOMINI M., REINA. S., GAGLIO S. A qualitative process theory based model of the HIV-1 Virus-Cell interaction. Procedings of Medical Informatics Europe 93, Israel. ISBN 965-294-091-7, 147-150.