# Knowledge Base for Evidence Based Medicine with Bioinformatics Components

Witold Jacak[1], Karin Pröll[1], and Jerzy Rozenblit[2]

[1] Department of Software Engineering,
Upper Austria University of Applied Sciences,
Hagenberg, Austria
[2] Department of Electrical and Computer Engineering
University of Arizona,
Tucson, USA

**Abstract.** This paper presents an approach for a multilevel knowledge base system for evidence-based medicine. A sequence of events called patient trial is extracted from computer patient records. These events describe one flow of therapy for a concrete disease. Each event is represented by state and time. We introduce a measure between states, which is used to calculate the best alignment between different patient trials. The alignment measure calculates the distance between two sequences of patient states, which represents the similarity of the course of disease. Based on that similarity- value classes are introduced by using specific clustering methods. These classes can be extended by gene expression data on micro-arrays leading to finer clustering containing similar trials - called trial families. For easy checking if a new trial belongs to a family we use profiles of Hidden Markov models to detect potential membership in a family.

## 1 Introduction

Knowledge based system techniques and applications will be one of the key technologies of new medicine approaches. A variety of intelligent techniques have been initiated to perform intelligent activity such as diagnosis. Among them knowledge based techniques are the most important and successful branch. Especially, in new clinical information systems (CIS), called evidence based medical systems (EBM) the construction of an appropriate knowledge base is one of the most important problems. Evidence based medicine/healthcare is looked upon as a new paradigm, replacing the traditional medical paradigm, which is based on authority. It depends on the use of randomised controlled trials, as well as on systematic reviews (of a series of trials) and meta-analysis, although it is not restricted to these. There is also an emphasis on the dissemination of information, as well as its collection, so that evidence can reach clinical practice. It therefore has commonality with the idea of research-based practice [3], [4], [5].

Evidence based medicine is the integration of *best research evidence* with *clinical expertise* and *patient values*. (see Fig.1.)

R. Moreno-Díaz and F. Pichler (Eds.): EUROCAST 2003, LNCS 2809, pp. 163–173, 2003.

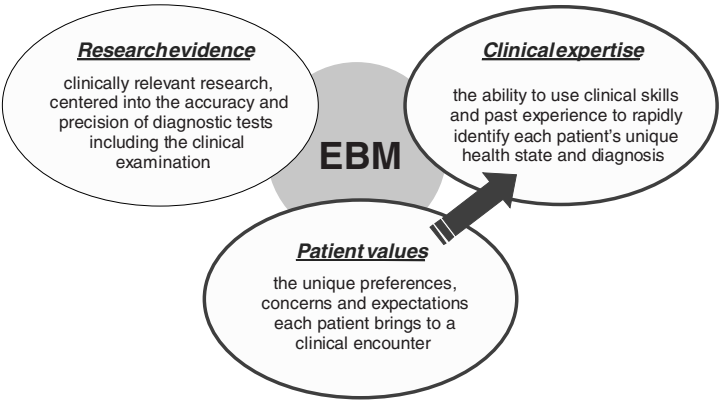


**Fig. 1.** Evidence based medicine (EBM) is the integration of best research evidence with clinical expertise and patient values.

- by *best research evidence* we mean clinically relevant research, often from the basic sciences of medicine, but especially from patient centered clinical research into the accuracy and precision of diagnostic tests (including the clinical examination), the power of prognostic markers, and the efficacy and safety of therapeutic, rehabilitative, and preventive regimens. New evidence from clinical research both invalidates previously accepted diagnostic tests and treatments and replaces them with new ones that are more powerful, more accurate, more efficacious, and safer.
- by *clinical expertise* we mean the ability to use our clinical skills and past experience to rapidly identify each patient's unique health state and diagnosis, their individual risks and benefits of potential interventions, and their personal values and expectations.
- by *patient values* we mean the unique preferences, concerns and expectations each patient brings to a clinical encounter and which must be integrated into clinical decisions if they are to serve the patient.

When these three elements are integrated, clinicians and patients form a diagnostic and therapeutic alliance, which optimises clinical outcomes and quality of life.
This article focuses the consideration on the multilevel approach for constructing an evidence based knowledge system for classification of different therapies and their effectiveness for clinical trials.

## 2 Multilevel EBM Knowledge Base System

In the first phase, a sequence of events called *patient trial* will be extracted from computer patient records (CPR). These events describe only one flow of therapy of a concrete disease. Each event is represented as a pair *(state, time)*. The *state* contains not only standard numeric parameters but also images (MR, RT, or photo) and text based linguistic descriptions. Based on such *state* we introduce the measure between

states of different patient. We assume that each patient's state is represented in the global state space. Based on the measure the system calculates the best alignment between different patient trials. The alignment measure (score) calculates the distance between the two sequences of patient states, which represents the similarity of the flow of the therapy [2].

This procedure is applied to each pair of patient trials stored in the Hospital Information Systems (HIS) concerning similar diseases. Based on the value of similarity a semantic network is constructed and divided into full-connected partitions. Each of these partitions represents the class of the similar therapy and can be used for computer-aided decision-making in evidence based medicine diagnostic. The clustering only performed on the basis of medical data can be extended by the additional clustering process based on gene expression data for these same patients sets. The general structure of the system is sketched in the Fig.2.

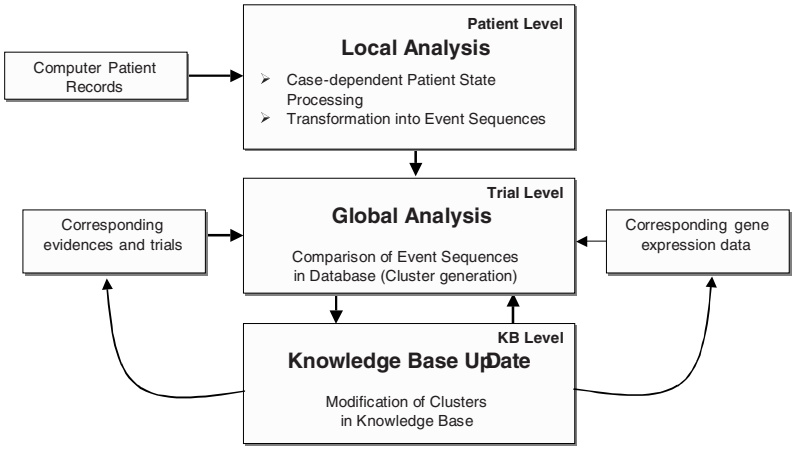


**Fig. 2.** Patient Values Knowledge Base for Evidence Based Medicine

## 3 Patient Record Level

On the patient level of knowledge base the data form patient records should be preprocessed to obtain the compact representation of course of disease. Normally we have different sources of medical information concerning a couple of numerical data representing the labor test results, some RT images, linguistic text describing diagnosis and therapy. The general patient record can described as

$$Patient\ Record = \{(v_j | j=1,..,n),\ \{image_1,\ldots image_k\}, diagnosis, therapy\}$$

In the first phase the patient record should be transformed into sequence of events $e$ representing the time course of a patient healing process called patient trials.

$$Trial=(e_1,\ldots,e_n)=((state_1,\ time_1),..,(\ state_n,\ time_n))$$

In order to compare the different trials it is necessary to introduce a formal description of states to allow the calculation of the distance or similarity between two states. It is

obvious that the vector of numerical data $(v_j \mid j=1,..,n)$ is easy to compare. To make it possible to measure the similarity between images we propose to map images with respective 3D models which can be used to obtain additional numerical data $(w_j \mid j=1,..,k)$ describing images. Such methods are very useful for dermatological diseases.

Based on 3D models we cannot only calculate various geometrical parameters as for example field, contour length or shape but we also can automatically perform a classification of infected anatomical regions and a coding of disease (Fig. 3.). Additionally we assume that diagnosis will be transformed into standard code for example ICD-10.

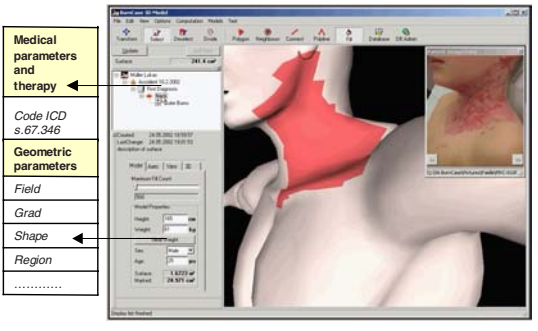


**Fig. 3.** Case-dependent Patient State containing numerical data and 3D-model data

After these preprocessing steps we can represent the state as a vector of numerical parameters and codes.

$$e^i = (state^i, time^i) = (((v^i_j \mid j=1,..,n), (w^i_j \mid j=1,..,k), diag\_code^i), time^i)$$

Based on such *state* we introduce the measure of distance

$$\rho : S \times S \rightarrow R^+$$

between states of different patient (see Fig.4.).

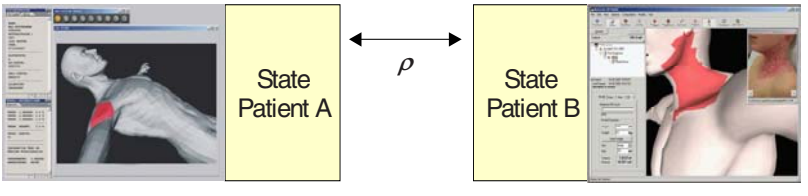


**Fig. 4**. The local measure between states of different patient.

We assume that each patient's state is represented in this global state space *S*. This distance represents the similarity of two states form different trials of the course of disease. An example of a patient trial for a dermatological disease is presented in Fig.5.


**Fig. 5.** Example of a patient trial

The images above are extracted from a concrete patient trial showing the healing process of a dermatological case in the neck region. Each state is based on a 3D model of the affected part of the body and a set of numeric and code parameters representing results of assessments or treatment protocols. The 3D Model and the parameters are used to build the state vector at each point of time, which are used to calculate the measure between states of treatment in different patient trials.

## 4 Trial Level : Alignment of Trials

The most basic trials analysis task is to ask if two trials are related. There are many positions at which the two corresponding patient states (residues) are very similar. However, in the case when the one of the trial has extra residues, then in a couple of places gaps have to be inserted to the second trial to maintain the alignment across such regions. When we compare sequences of the patient states, the basic processes are considered as substitution, which changes residues in the sequence, and insertion or deletion, which add or remove residues. Insertions and deletions are referred to as gaps. An example for state alignment can be seen in Fig.6.

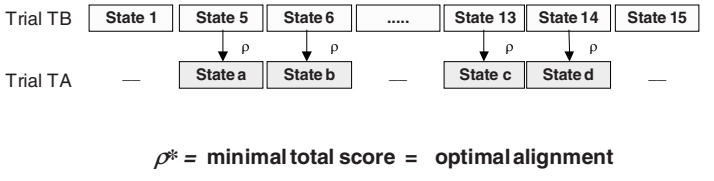


**Fig. 6.** Alignment between two trials

Based on the distance $\rho$ the system calculates the alignment score $\rho^*$ which described the similarity (distance) between different patient trials.

$$\rho^* : S^* \times S^* \rightarrow R^+$$

where $S^*$ is the set of ordered sequences of the states from the state space $S$. Each sequence

$$x = (x_1, \ldots, x_n) \in S^*$$

has the respective events sequence describing the concrete patient trial $(e_1, \ldots e_n)$. The measure $\rho^*(x,y)$ calculates the distance between these two sequences, which describes the similarity of the flow of the therapy. For solving this problem we use a special sequence-matching algorithm.

The global measurement (score) we assign to an alignment will be the sum of terms for each aligned (similar) pair of states (residues), plus terms of each gaps (see Fig.7.). We will consider a pair of trials (patient states sequences) $x$ and $y$ of lengths $n$ and $m$, respectively.

Let $x_i$ be the $i$th state in $x$ and $y_j$ be the $j$th symbol of $y$. Given a pair of trials, we want to assign a score to the alignment that gives a measure of the relative likelihood that the trials are related as opposed to being unrelated. For each two states $x_i$ and $y_j$ we can use $\rho$ as measure of similarity of this residues. Let $n > m$ then we should add the gaps g in the second trial to find the best alignment.

### 4.1 Gap Penalties

We expect to penalize gaps. Each state $x_i$ in the sequence $x$ has additional parameter $\tau_i$ which represent time interval between the state $x_{i-1}$ and $x_i$. The first state $x_1$ has $\tau_1=0$. For finding the penalty value of gap at the $i$th position of the trial $y$ we use the knowledge about time intervals associated with each state. Let the last ungaped substitution with state $y_l$ has place on $(i-k)$th position in the trial $x$.

| | | | | |
|---|---|---|---|---|
| $x_{i-k}$ | $x$ | $x$ | $x_{i-1}$ | $x_i$ |
| $y_l$ | $g$ | $g$ | $g$ | $g$ |

**Fig. 7.** Gaps Insertion

The standard cost associated with a gap is given by

$$\rho(x_i, g) = K \exp(-(|\tau^x_i - \tau^y_l|)) = pen(x_i, y_l)$$

where $\tau^x_i$ is the time interval associated with state $x_i$ and $\tau^y_l$ is the time interval associated with state $y_l$ from trial $y$ which was aligned with the state $x_{i-k}$ from the trial $x$.

The long insertions and deletions with different intervals of time are penalized less as those where the intervals of the time is quite the same.

### 4.2 Alignment Algorithm

The problem we consider is that of obtaining optimal alignment between two patients trials, allowing gaps. The problem can be defined as follows:

*Find the best alignment between sequence x* and y* (x*, y* represent sequences x and y extended of necessary gaps) such that global score*

$$\rho^*(x,y) = \Sigma(\rho(x^*_i, y^*_i)) = min$$

*where* $x^*_i = x_i$ *or gap g and*
*if* $x_i$ *is aligned to* $y_l$ *and* $x_{i+k}$ *is aligned to* $y_u$ *then* $l<u$.

We can use the known dynamic programming algorithm, which has many applications in the biological sequences analysis [1].

The idea is to build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences.

We construct a matrix $F$ indexed by $i$ and $j$, one index for each trial, where value $F(i,j)$ is the score of the best alignment between the initial segment $x_1 \ldots x_i$ of $x$ up to $x_i$ and the initial segment $y_1 \ldots y_j$ of $y$ up to $y_j$. We can build $F(i,j)$ recursively. We begin by initializing $F(0,0) = N$ ($N$ is the large number). We then proceed to fill the matrix from top left to bottom right. If $F(i-1,j-1)$, $F(i-1,j)$ and $F(i,j-1)$ are known it is possible to calculate $F(i,j)$.

Let us assume that we are only interested in matches scoring ρ less than some threshold $T$, it means that similarity between two states of the patients is very high. $F(i,0)$ is the minimal (best) sum of completed match scores to the subsequence $x_1,\ldots x_i$ assuming that $x_i$ is in an unmatched region.

It is clear that we expect that one trial contains the other or that they overlap. It means that we want a match to start on the top or left border of the matrix and finish on the right or bottom border. The initialization equations are that $F(i,0) = 0$ for $i= 1,\ldots, n$ and $F(0,j) = 0$ for $j = 1,..,m$. Now we calculate recursively the matrix value as

$$F(i,0) = \min \begin{cases} F(i-1,0) \\ F(i-1,m)+T \end{cases}$$

$$F(i,j) = \min \begin{cases} F(i-1,j-1) + \rho(x_i, y_j) \\ F(i-1,j) + pen(x_{i-1}, y_j) \\ F(i,j-1) + pen(x_i, y_{j-1}) \end{cases}$$

The calculation steps are presented in Fig.8.

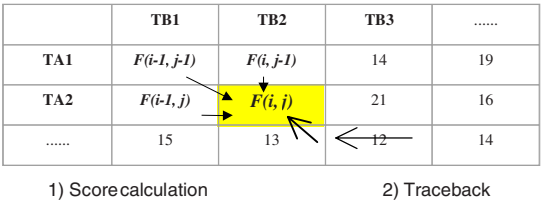


**Fig. 8.** Matrix for alignment calculation

Let $F_{min}$ be the minimal value on the right border $(i,m)$ for $i = 1,..n$, and the bottom border $(n,j)$ $j = 1,..,m$. This minimal score is the measure of the similarity between the complete two trials x and y. i.e.

$$\rho^*(x,y) = F_{min}$$

To find the alignment itself we must find the path of choices that led to the minimal value. The procedure for doing this is known as a traceback. The traceback starts from the minimal point and continues until the top or left edge is reached.

## 5 Clustering of the Trial Space

Based on the alignment score $\rho$ we can define similarity classes on the set of trials. The similarity class $C$ is defined as follows:

- $C \subset$ *Set of Trials*
- $(\forall\ x, y \in C\ )(\ \rho\,(x,y) < \varepsilon\ )$
- $card\ (\ C) \rightarrow max$

where $\varepsilon$ is the threshold value discriminating the similarity of two trials. For construction of similarity classes we build the graph which nodes are trials and its arcs is the relation $\rho$. The threshold value cuts different arcs. The similarity class is the maximal subgraph of the graph for which its nodes are full connected (see Fig.9.). There are many algorithms for constructions similarity classes [7].

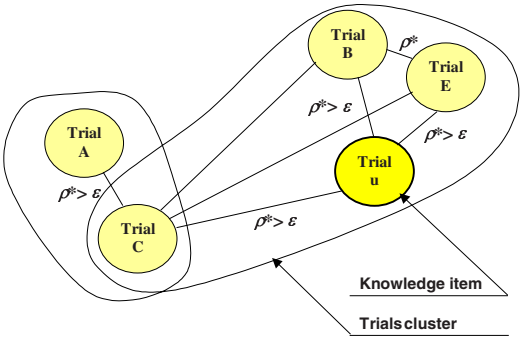


**Fig. 9.** Trials clustering

## 6 Gene Expression Clustering

As an additional support for diagnostics gene expression data can be used. Using micro-arrays, we can measure the expression levels of more genes simultaneously. These expression levels can be determined for samples taken at different time points during a biological process or for samples taken under different conditions. For each gene the arrangement of these measurements into a vector leads to what is generally called an expression profile. The expression vectors can be regarded as data points in high dimensional space. Cluster analysis in a collection of gene expression vectors aims at identifying subgroups (clusters) of such co-expressed genes, which have a higher probability of participating in the same pathway. The clusters can be used to validate or combine the cluster to prior medical knowledge. Many clustering methods are available, which can be divided into two groups: first and second generations algorithms. The first generation algorithms are represented by hierarchical clustering algorithms, K-means clustering algorithms or self-organizing maps. These algorithms are complicate in use and often require the predefinition of more parameters that are hard to estimate in biomedical praxis. Another problem is that first generation

clustering algorithms often force every data to point into a cluster. It can lead to lack of co-expression with other genes. Recently new clustering algorithms have started to tackle some of limitations of earlier methods. To this generation of algorithms belong: Self-organizing tree algorithms, quality based clustering and model-based clustering [2], [6], [8], [9]

Self-organizing tree algorithms combine both: self-organizing maps and hierarchical clustering. The gene expression vectors are sequentially presented to terminal nodes of a dynamic binary tree. The greatest advantage is that the number of clusters does not have to be known in advance. In quality based clustering clusters are produced that have a quality guarantee, which ensures that all members of cluster should be co-expressed with all members of these cluster. The quality guarantee itself is defined as a fixed threshold for a maximal distance between two points between clusters. Based on these methods it is possible to generate the clusters on the gene expression states.

Let $G$ be a gene expression cluster, which contains the data obtained from micro-arrays. Each micro-array is connected with one or more patient trials. The cluster $G$ can be transformed as a cluster $GT$ to the patient trial space. This results in two patterns, which can be used for classifying each trial. On the one side the pattern based on trial alignment - on the other side the pattern based on gene expression can be used. Both patterns can be combined for creation fine classes containing very similar trials with high co-expression of genes [9].

## 7 Profiles of Patient Trials Families

In the previous sections we have already created a set of trials belonging to a particular cluster. Such a cluster is called a trial family. Trials in a family are diverged from each other in the primary sequence of the course of a disease. For easy checking if a new trial belongs to a family we propose to use statistical features of the whole set of trials in the cluster. With such a description it is possible to decide how strong the given trial is related to a cluster. We will develop a particular type of a Markov model, which is well suited for modeling multiples alignment in a group [1].

The profile of the trials family with maximal length $n$ should describe the probability of an observation of a specific state of a patient in $i$-th position of the trial. It means that the profile can be formalized as:

> *The profile P of length n based on states set S is the matrix*
> $P = [\pi_i(s) \mid i = 1,..,n \text{ and } s \in S]$
> *of probabilities.* $\pi_i(s)$ *is the probability that s occurs on position i in the sequence of states.*

The approach is to build a hidden Markov model with a repetitive structure of states but different probabilities in each position. The key idea behind profile HMM is that we can use the same structure, but set the transition and emission probabilities to capture specific information about each position in the whole family of trials. The model represents the consensus sequence for the family, not the sequence of any particular member.

A probabilistic model would be to specify probabilities $\pi_i(s)$ of observing the state $s$ in position $i$ of trial. Such a model is created by Hidden Markov models with insertion and deletion where for each matched state the emission distribution for each patient state is estimated - for insert regions the emission and transition distribution is estimated and for deletion only the transition one. One of the main purposes of developing profiles of Hidden Markov models is to use them to detect potential membership in a family by obtaining significant matches of a given trial to a profile.

To find out if the observation trial belongs to the trials family we look for the most probable path in the Hidden Markov Model in the family. The most probable path in the HMM can be found recursively with standard Viterbi algorithm. Using this algorithm we can calculate the maximal probability a the given trial belongs to the cluster. The most difficult problem faced when using HMM is specifying the model in the first place. There are two parts to do this: the design of the structure, i.e. what states there are and how they are connected and the assignment of the parameter values, the transition and emission probabilities. The estimation of these parameters can be performed by using the Baum-Welch training algorithm. The choice of the length of the model corresponds more precisely to the decision on which multiple alignment columns in trial family to assign to match states, and which to assign insert states. A simple rule working well is that columns that are more than half gaps should be modeled by inserts.

The family profile can be used to predict the course of the disease based only on an initial subsequence of the trails. In this sense the profiles are very important tools in decision support for medical therapy.

## 8 Final Remarks

Evidence-based Systems gain a more important role in decision support for medicall praxis. This paper presents a new approach for a multilevel knowledge base system for evidence-based medicine combined with bioinformatics components. In the first level a sequence of events called patient trial is extracted from computer patient records. These events describe one flow of therapy for a concrete disease. Each event is represented by state and time. We introduce a measure between states, which is used to calculate the best alignment between different patient trials. The alignment measure calculates the distance between two sequences of patient states, which represents the similarity of the course of disease. On the second level based on that similarity-value classes are introduced by using specific clustering methods. These classes can be treated more exactly by combining the information about gene expression data on micro-arrays. This leads to finer clustering containing similar trials - called trial families. For easy checking if a new trial belongs to a family we use profiles of Hidden Markov models to detect potential membership in a family. This knowledge base can be used to support diagnostic in hospital praxis.

## References


1. Durbin R., Eddy S., Krogh A. Mitchison G.: Biological sequence analysis, Cambridge University Press (1998), UK
2. Moreau Y., De Smet F., Thijs G., Marchal K., De Moor B.: Functional Bioinformatics of Microarray Data: From Expression to Regulation, Special Issue on: Bioinformatics, Part1, Proceedings of the IEEE, Vol. 90, Number 11, (2002), 1722–1743
3. Gosling A. S., Westbrook J. I., Coiera E. W.: Variation in the use of online clinical evidence: a qualitative analysis, International Journal of Medical Informatics, 69; 1, (2003),1–16.
4. Warner HR**,** Sorenson DK, Bouhaddou O. : Knowledge engineering in health informatics. Springer-Verlag, New York, NY. (1997)
5. Friedman CP, Wyatt JC.; Evaluation Methods in Medical Informatics. Springer, New York (1997).
6. Kohonen, T. Self-organizing Maps, Springer-Verlag, (1997), Berlin, Germany.
7. Kaufmann L., Rousseeuw P.J. Finding groups in Data: An Introduction to Cluster Analysis, Wiley, (1990), New York, USA
8. Moreau Y., De Smet F., Thijs G., Marchal K., De Moor B. : Adaptive quality-based clustering of gene expression profiles, Bioinformatics, Vol.18, no.5 (2002)
9. Aris V, Recce M.: A method to improve detection of disease using selective expressed genes in microarray data, Methods of Microarray Data Analysis , eds. SM Lin, KF Johnson (Kluwer Academic) (2002), pp. 69–80