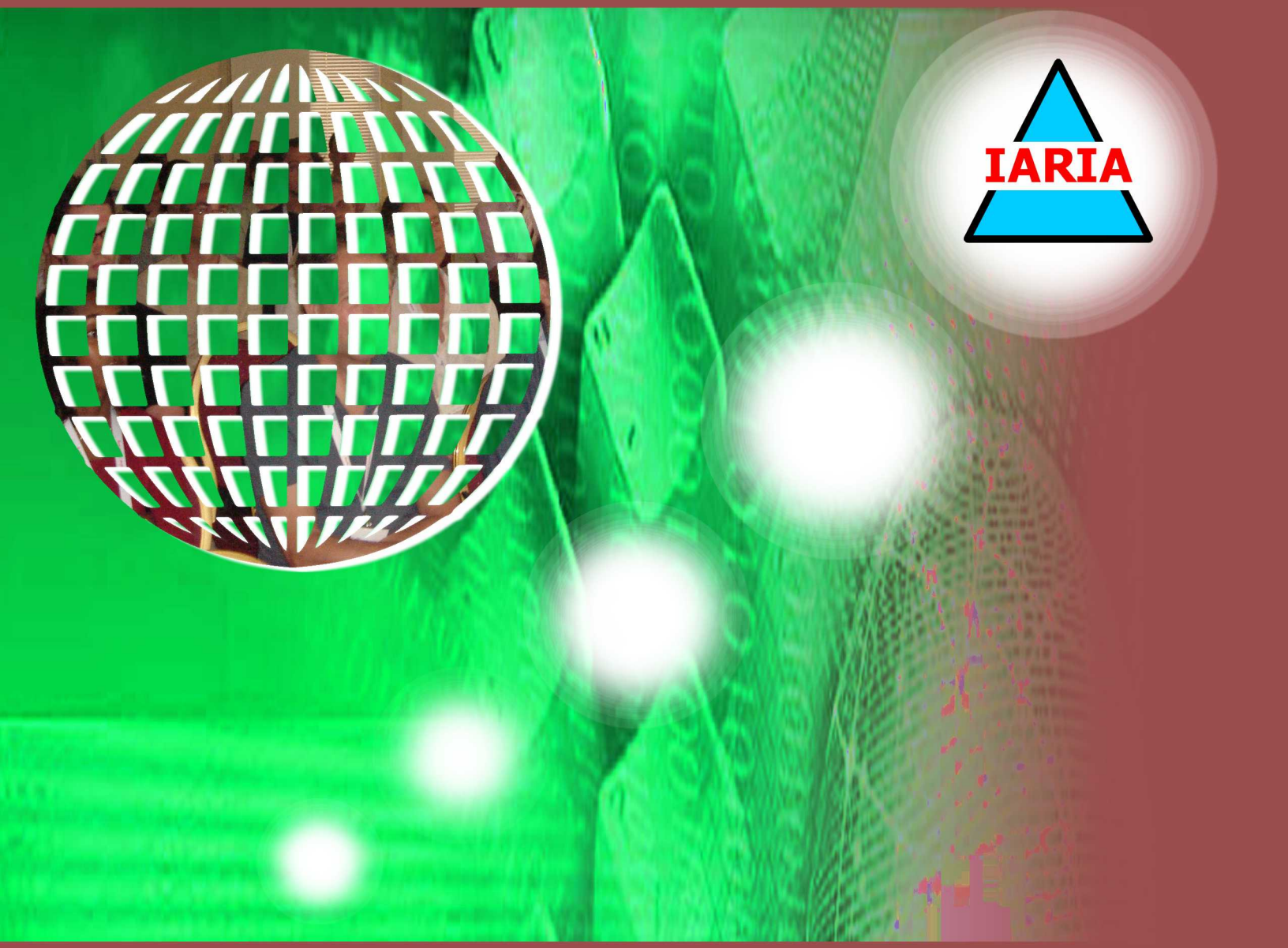
International Journal on Advances in Life Sciences



2009 vol. 1 nr. 1

The *International Journal On Advances in Life Sciences* is Published by IARIA.
ISSN: 1942-2660
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal On Advances in Life Sciences
Volume 1, Number 1, 2009

- Steve Wheeler, University of Plymouth, UK

Advanced Knowledge Representation and Processing

- Freimut Bodendorf, University of Erlangen-Nuernberg Germany
- Borka Jerman-Blazic, Jozef Stefan Institute, Slovenia
- Andrew Kusiak, The University of Iowa, USA
- Selmin Nurcan, University Paris 1 Pantheon Sorbonne, France
- Jeff Riley, Hewlett-Packard Australia, Australia
- Lars Taxen, Linkoping University - Norrkoping, Sweden

International Journal On Advances in Life Sciences
Volume 1, Number 1, 2009

Foreword

The inaugural issue of the International Journal on Advances in Life Sciences compiles a set of extended contributions based on previously awarded work. The enhanced work adds substantial material compared to the initial ideas and provides the reader with additional background, experiments, and a deeper analysis of results.

Of the submitted articles, five were selected to be included in this issue. The subjects presented deal with items ranging from cell level experiments to large scale health network management.

In the first article, Joel Jeffrey presents a new model for a mathematical description of cellular and molecular structure, as well as mechanisms and states. Structural similarity can be mathematically computed and predictions can be made using information stored in repositories.

The second article, Xiaofeng Dai et al. present a stratified mixture model for clustering genes. This proposal allows for the utilization of data coming from multiple data sources, leading to a better clustering accuracy of genes.

Mirela Danubianu et al. present an approach to personalized speech therapy aimed at assisting preschool children with dyslalia. This disorder is one of the more common ones amongst children and it involves the omission, distortion, or substitution of one sound for another. An integrated system assists the decisions of a speech therapist and can benefit from a gradually building knowledge base.

Article four by Rodica-Mariana Ion et al. presents in the fourth article a drug system for the photodynamic treatment of HeLa tumor cells. The combination of TSPP and CisPt was found to have an increased effect of the cells' survival rate caused by cellular lesions.

Finally the fifth article addresses the issue of data migration from a regional network to a national network. In this work, Kari Harno et al. present some of the issues and strategies of such a migration with a focus on the Finnish health care system.

We hope that the contents of this journal will add to your understanding of life sciences, and that you will be inspired to contribute to IARIA's conferences that include topics relevant to this journal.

Elaine Lawrence, Editor-in-Chief
Petre Dini, IARIA Advisory Committees Board Chair

International Journal On Advances in Life Sciences
Volume 1, Number 1, 2009

CONTENTS

Mathematical Description of Biological Structures, Mechanisms, and States

H. Joel Jeffrey
Department of Computer Science,
Northern Illinois University, DeKalb, IL 60115
jeffrey@cs.niu.edu

Abstract—-A new method for mathematically describing cellular and molecular structures, mechanisms, and states is presented. A novel mathematical formulation of structure is developed, and new mathematical formulations of structural complexity and similarity are introduced that take into account differences in composition and structure at all levels of detail and apply equally to structures, mechanisms, and states. A recursive formula for calculation of structural similarity is derived. The methods and mathematical formulations apply equally to cases in which we have complete knowledge and to those in which we have only incomplete or partial information. The formalism and the mathematical similarity definition are the full generalization of sequence and sequence similarity. They enable the creation of repositories of formal multi-level structural descriptions of biological entities and new search capabilities, such as searching for processes or structures similar to a specified one, or with specified structural or compositional deviations.

Keywords- multi-level structure; mathematical models of structure; quantifying structural similarity

1. Introduction

DNA and protein sequence databases are of great value to biologists. They have this value because they are formal. A DNA or protein is specified by a string of characters on a 4- or 20-letter alphabet; a PDB entry is specified by a set of formal atom names and locations. Because specifications are formal, formal measures of sequence similarity are possible, and software such as BLAST is used routinely to find sequences similar to a query sequence.

By contrast, the great majority of descriptions of structure, whether of mechanisms or structures, are not formal. They are in ordinary technical language, in some cases augmented with graphical devices such as interaction diagrams and pictures of key molecular constituents. Because it is represented primarily in natural language, most molecular biological knowledge cannot be handled algorithmically. We can search on amino acid sequence similarity, but we cannot, e.g., query for "all proteins with structure similar to degree *d* to human hemoglobin in the R state," and get back proteins that have subunits similar in number, shape, and inter-relationships to those in R-state hemoglobin.

The situation is even more problematical with respect to biological situations and conditions. The customary concept of state and its formalization represents only a narrow subset of the intuition of biologically important facts or situations, namely those involving only the values of attributes of objects. The more general concept, for which we use the term *state of affairs*, involves processes, objects, and other component states of affairs, related in various ways. When we say, e.g., "the p53 molecule is phosphorylated," or "the DNA is damaged," we are identifying states of affairs.

This paper, an expansion of the work in [1], presents a new formalism for describing biological mechanisms, structures, and states of affairs that is the generalization of the concept of structure to the entire range of processes, structures, and states of affairs encountered in biology. The strings of letters representing DNA and protein sequences are special cases of the formalism. Using the formalism, new mathematical formulations of the concept of shape and structural similarity is developed, one that takes into account differences in composition and structure at all levels of detail and applies equally to structures, mechanisms, and inter-constituent relationships. Additionally, a new mathematical formulation of structural complexity is defined.

One goal of this work is the creation of databases of molecular biological mechanisms, structures, and states analogous to those we now have for genetic and protein sequences, and software systems using the new formulations and other algorithms operating on such multi-level structural knowledge bases.

2. Specification by constituents and relationships

We approach the problem of specifying structure of processes, objects, and states of affairs by considering the thing to be described as comprised of immediate constituents with specific attributes and inter-constituent relationships (temporal, spatial, or some other kind), and identifying the constituents and the relationships with formal names, as in mathematical logic. (Following standard practice in mathematical logic, an attribute is formally a one-place relationship, but here we will, for expository purposes, identify them separately.) We use the abstract term "entity" as a cover term for object, process, or state of affairs. Thus, an entity is specified by specifying its constituents and the relationships between them. The relationships are the formalization of the intuitive idea of structure. Each constituent is itself an entity, and therefore its structure can be elaborated with a second ES, and so on, continuing to any level of description desired. We term the approach *Entity Specification (ES)*.

An entity may be an object (structure), process (mechanism), or states of affairs (the generalization of state). A state of affairs is an entity whose constituents may be any set of objects, processes, and other states of affairs, with the necessary constituent attributes and inter-constituent relationships. States of affairs allow the formalization of complex situations, such as the fact that a p53 molecule is phosphorylated, that the DNA is damaged, the rate or change in rate of a reaction, a concentration of a molecular species, etc. An important case is location: location is an attribute of a process or object, represented formally as with any other attribute. Transport processes thus are processes represented via the same formalism as other processes.

Specifications of the entity (process, object, or state of affairs) are done by identifying the immediate constituents and the relationships and attributes that must be present for the item to conform to the definition, and specifying all constituents, properties, and relationships with formal names and values. For example, the customary high-level description of hemoglobin is that it has two immediate constituents, which are the two αß subunits, and the angle between subunits, which has one value in the R state and another in T. The subunits are further described as having two constituents α_1 and $ß_1$, α_2 and $ß_2$, respectively. Processes have processes and objects as constituents: the steps of the process and the elements involved in it. Constituents of states of affairs may be processes, objects, or other states of affairs. The core of an ES is thus a list of the immediate constituents and a list of the n-place relations among the constituents.

As the hemoglobin example illustrates, a set of Entity Specifications of an entity is formal and multi-level. ESs employ the same logical device used in mathematical logic: the use of formal names that are expanded by use of structured descriptions employing other formal names. Names of entities and relations are formal designators; a formal description of an entity gives further detail, i.e., its constituents and how they are structured.

2.1 Entity Specification

An *Entity Specification* (*ES*) consists of an ordered pair (N, D), where:

- N is the (formal) name of the entity including, optionally, a list of alternate names and/or a numerical ID.
- D is a set of *paradigms*, the major varieties or descriptions of the entity. DNA transcription has two major varieties, eukaroytic and prokayotic. In addition, if is often desirable to specify alternate descriptions due to the state of knowledge of the phenomenon: conjectures, possible alternative mechanisms, etc. The paradigms are the distinct descriptions of the entity.

Each paradigm of D is an ordered triple (C, R, E), where:

- $C = \{(C_i, T_i)\}$, in which C_i are the constituents and T_i is each constituent's classification, an element of the set {P, O, S}, representing "object,", "process," or "state."
- $R = \{r_j\}$ is the set of n-ary *relationships* that must hold between the named constituents. Any relationship may be included, not only those definable in terms of physical locations or quantities. Equations specifying quantitative relationships, including differential equations, are formal relationship names.
- The constituents and their relationships specify the structure of the entity. Additional information specifies particular instances of the entity, by identifying which actual "things" (processes, objects, and states of affairs) fill the roles named by the constituents. This information we term the *eligibilities* for the entity: E is a set of ordered triples (C_j, i, r), in which, for each C_j,
  - C_j is the constituent;
  - i is the name of the actual individual;
  - r is the rule, or condition, under which i takes the role of C_j in the entity N.

2.2 Processes

Processes are multi-step changes in objects and how they are configured, i.e., the relationships between them. In addition, processes may occur in many versions, i.e., combinations of the stages that are all ways of the specific process occurring.

To represent this concept formally, the $\{(C_{i,} T_i)\}$ for a process are:

1. Two constituents, specifying the before-state and after-state.
2. A subset identifying stages, i.e., constituents C_j in which T_j = P. Some stages may be accomplished via two or more alternatives; these alternatives are included in this subset.
3. A subset identifying the objects, i.e., T_j = O.
4. A subset identifying the versions of the process. Each of these version constituents is a state of affairs, i.e., T_k = S, and its constituents are the stages that comprise the version.

The relationships between stages specify those that happen sequentially, in parallel, overlapping, or in any other temporal relationships.

The stages are the steps of the process, and the states are the usual concepts of the before- and after-states of a process.

2.3 Objects

Objects have only object constituents, and in that sense are simpler than entities in general or processes; each constituent of an object is of Type O.

Objects provide clear illustration of multiple paradigms. For example, the large subunit of a ribosome is commonly described as having a roughly spherical main body and three lobes (i.e., with 3 constituents), but it is also described as comprised of two rRNA chains (5s, 23s) and a number of proteins. Fig. 3 shows the ESs of the eukaroytic ribosome and its constituents' sub-constituents.

2.4. States of Affairs

States of affairs are the most general kind of entity, since the constituents may be any object, process, or other state of affairs.

Since there are no restrictions on the constituents of a state of affairs, the general Entity Specification of Sec. 2.1 is the form of a state of affairs. This means that any entity is formally equivalent to a state of affairs.

2.5 Examples

The kinds of entities most directly of interest in biology are mechanisms and structures. We illustrate mechanism ESs with the ES of cell cycle arrest and gene transcription, in which the Constituents are Stages, Versions, and Elements, and the relationships are the constraints on which Stage must complete before initiation of the next one. The eukaryotic ribosome is used to illustrate structure ESs.

2.5.1 Cell cycle arrest

Fig. 1 shows an Entity Specification of the process of damaged DNA stopping the cell cycle. (Formal names similar to ordinary English phrases and sentences are used, with the notational device of square-brackets to indicate use of formal Element names in Stages.)

Stage 4, the general process of gene expression, in this process specifies production of p21, the Individual for Element "a protein" (shown in Fig. 1 in brackets). Thus, what actually occurs is the expression of p21, the phenomenon to be described.

N: [A damaged DNA molecule] stops the cell cycle in [a cell] {P1}
S_{result}: not Occur(S-phase, a cell)
Elements: a damaged DNA molecule: DNA molecule D
a cell: the cell with the damaged DNA
D:
Paradigm 1: eukaroytic cell
Stages:
1. [A damaged DNA molecule] activates [an ATM molecule]
S_{result}: [an activated ATM molecule]
Elements:
a. a damaged DNA molecule: damaged DNA molecule D
b. an ATM molecule: ATM molecule A
c. an activated ATM molecule: activated ATM molecule A
2. [An ATM molecule] phosphorylates [a p53 molecule]
S_{result}: [a phosphorylated p53 molecule]
Elements:
a. a p53 molecule: p53 molecule p53P
b. a phosporylated p53 molecule: phosphorylated p53 molecule p53P
Condition: only after Stage 1
3. [An activated p53 molecule] binds to [DNA] at the p21 coding site
S_{result}: [a phosphorylated p53 molecule] bound to [DNA] at the p21 coding site
Elements:
a. DNA: the DNA of the cell
Condition: only after Stage 2
4. [A cell] produces [molecules of a protein] from [a gene]
S_{result}: [a number molecules] of [a protein]
Elements:
a. a protein: p21
Condition: only after Stage 3
5. [A p21 molecule] inactivates [a cyclin E:cdk2 molecule]
S_{result}: inactivated [cyclin E:cdk2 molecule E2M]
Elements:
a. a cyclin E:cdk2 molecule: cyclin E:cdk2 molecule E2M
b. cyclin E:cdk2 molecule: cyclin E:cdk2 molecule E2M
Condition: only after Stage 4
Versions: 1. 1-2-3-4-5

Figure 1: ES of "Damaged DNA stops the cell cycle"

2.5.2 Gene Transcription

Step 4 in Fig. 1 is gene transcription. Fig, 2 shows, in outline form, the information to be specified in an ES elaboration of it: 7 ESs, at 4 levels of specificity, with a total of 25 processes identified. The top, or overall, level is "a cell produces molecules of a protein from a gene," with three stages, which are its process constituents; Stage 3, with formal name "Ribosomes translate an mRNA transcript of a gene to molecules of the protein," is of course gene transcription, with 4 stages.

Cell cycle arrest and gene transcription illustrate a central feature of ESs and ES methodology, namely the multi-level logical structure of ESs. Any single ES presents a full (formal) specification of the process, object, or state of affairs *at that level of detail*. Further detail is specified by further ESs.

The hierarchical specification technique is the formal analog of the commonly used informal method for describing complex biological processes, namely a hierarchical description elaborating the process structure in finer and finer detail, beginning with a high-level description in terms of a small set of large "steps" and continuing with division of steps into sub-steps, etc.. The outline form of gene expression in Fig. 2 is an example of just such a hierarchical description.

Process: A cell produces molecules of protein from a gene
1. The cell transcribes the gene to an mRNA molecule
2. The mRNA transcript moves to a ribosome in the cytosol
3. Ribosomes translate an mRNA transcript of a gene to molecules of the protein
   3.1. A ribosome initiates translation of the mRNA transcript
      3.1.1 The small ribosomal subunit binds to the mRNA transcript near the start codon
      3.1.2 The small ribosomal subunit moves to the start codon
      3.1.3 A tRNA molecule binds to the start codon on the mRNA transcript
      3.1.4 The large ribosomal subunit arrives at the transcription site.
   3.2. The ribosome adds an amino acid to the peptide chain
      3.2.1 The amino acid on the aminoacyl tRNA molecule binds to the A-site on the large ribosomal subunit
         3.2.1.1 The first nucleotide of the tRNA anticodon forms a hydrogen bond with the first nucleotide of the codon
         3.2.1.2 The second nucleotide of the tRNA anticodon forms a hydrogen bond with the second nucleotide of the codon
         3.2.1.3 The third nucleotide of the tRNA anticodon forms a hydrogen bond with the third nucleotide of the codon
      3.2.2 The large ribosomal subunit joins the amino acid on the P-site with amino acid on the A-site
         3.2.2.1 Peptidyl transferase breaks the bond between the amino acid and the tRNA molecule at the P-site
         3.2.2.2 The P-site amino acid and the A-site amino acid form a peptide bond
         3.2.2.3 The P-site tRNA molecule moves to the E-site on the ribosome
         3.2.2.4 The A-site tRNA molecule moves to the P-site on the ribosome
      3.2.3 The P-site and A-site tRNA molecules move three nucleotides in the 3' direction on the mRNA transcript
      3.2.4 The E-site the releases the tRNA molecule attached to it
   3.3. The ribosome terminates the polypeptide chain.
      3.3.1 The release factor binds to the large ribosomal subunit
      3.3.2. A peptidyl transferase molecule catalyzes the addition of a water molecule to the peptidyl tRNA, breaking the bond holding the polypeptide to the tRNA
   3.4 The large and small subunits of the ribosome dissociate

Figure 2: Process of gene expression (outline form)

Complexity in a process is often due to complex relationships between stages, such as the initiation of one stage only upon completion of another stage, perhaps of an entirely distinct process. Further, often these conditions involve additional factors of several kinds, e.g., a combination of a concentration of a biochemical and the physical location of a ligand. In these cases the formal expressive power of Entity Specification, in particular the formal inclusion of any relationship between constituents, as in mathematical logic, provides the capability of capturing the actual condition.

A different and important source of complexity in biological processes is the common situation in which the "output" of one process is an input to another. Fig. 1 shows an example: Stage 4 specifies the general process of gene expression; the particular instance of this general process is the production of p21, the key molecule in Stage 5. Fig. 1 thus illustrates two additional aspects of Entity Specification as applied to complex processes: 1) specification of a general process instantiated to produce a specific result – in this case, a p21 molecule, and 2) formally specifying the requirement of the presence of an actual object for the process to continue. Thus, Entity Specification represents formally what one can say informally: "the p21 gene is expressed, producing a p21 molecule, which inactivates the cyclin E cdk2 molecule, and so the cell cycle cannot continue."

2.5.3 Ribosome structure

We illustrate object Entity Specifications with an ES of the ribosome, formalizing the customary description into RNA subunits and proteins, and their constituents in turn, as found, for example, in [11].

N: eukaroytic ribosome
Relationships:
a. molecular weight(ribosome.eukaroytic) = 4,200,000

D:
Paradigm: 1
Sub-objects:
1. [SRSU]
2. [LRSU]

Relationships:
a. molecular weight(SRSU) = 1,400,000
b. molecular weight(LRSU) = 2,800,000
c. adjacent(LRSU, SRSU)

N: LRSU
Relationships:
a. molecular weight(LRSU) = 2,800,000

Paradigm: 1
Sub-objects:
1. [5S RNA]
2. [28S RNA]
3. [5.8S RNA]
4. [Protein1]
…
52. [Protein49]

N: 5S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
120. [Nucleotide120]

N: 28S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
4700.[Nucleotide4700]

N: 5.8S RNA
Paradigm: 1

Sub-objects:
1. [Nucleotide1]
…
160. [Nucleotide160]

N: SRSU
Relationships:
a. molecular weight(LRSU) = 1,400,000
Paradigm: 1
Sub-objects:
1. [18S RNA]
2. [Protein1]
…
34. [Protein33]

N: 18S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
900. [Nucleotide1900]

Figure 3: ESs of the eukaroytic ribosome and constituents

This example illustrates two significant aspects of the use of ESs. First, just as with ordinary-English descriptions, completeness is not necessary. ESs may be used to represent as much as is known of the structure of the entity, or as much as is desired for the purpose at hand. The depiction of the ribosomes in [11] includes only the constituents and their weights, and one relationship, namely that the SRSU and LRSU are adjacent; this is the information formalized in Fig. 3. The 5S RNA constituent is described further only by noting that it contains 120 nucleotides, without specifying them, and this formalized by the subobjects Nucleotide1…Nucleotide120 above. Other descriptions of the 5S rRNA constituent of the LRSU of the eukaroytic ribosome specify the particular nucleotides; these are formalized by specifying the nucleotides (A, C, G, U) and their structure with ESs of the constituents of each: the nitrogenous base, the 5-carbon sugar, the phosphate groups, and the positional relationships between them.

Second, it illustrates that there is no single "correct" Specification of an Entity, represented by having multiple Paradigms. This is not a deficit of the ES approach, but is rather a formal representation of the fact that there are often multiple descriptions of the same thing. Thus, we find the LRSU described in terms of the 5S, 28S, and 5.8S rRNA constituents, and we also find it described in terms of constituents of the main body, central protuberance, ridge, stalk, and valley.

Paradigms are the means of formalizing multiple descriptions of the same thing. This is not simply a semantic technicality. In the next section, we will show how to use ESs to mathematically quantify the concepts of complexity and structural differences, and the definitions and algorithms are based on the constituents and relationships *in a description*, i.e., a paradigm. To put it differently, there is no such thing as the "real structure" of an entity – structure, mechanism, or state of affairs. Rather, there are multiple descriptions of the entity, in ordinary English or formal ESs, and it is only meaningful to compare descriptions of entities.

This however does not preclude the discovery of a canonical form of ES, or adoption of standards or conventions for creation of ESs. It may, for example, be desirable to adopt conventions for automatically and uniformly converting protein databases to multi-level ESs representing their secondary, tertiary, and quaternary structure.

2.5.4 DNA and protein sequence databases

In Section 1 we noted that the string representations of DNA and protein sequences are special cases of ESs. In DNA or RNA sequences, the letters "A," "C," "G," and "T" (or "U") are the constituents, and the single relationship is "adjacent." In a protein sequence, the constituents are the letters denoting the amino acids.

In actual sequences, the nucleotides (or amino acids) have relative positions specified by two angles and a distance, and in certain cases the more complete symbolic representation of the sequence including is useful: N_1 (φ_1,θ_1,d_1) N_2 (φ_2,θ_2,d_2) N_3 …. In ES representation of this kind of sequence, the letters are the constituents and the relationships are the three $\varphi(x,y)$, $\theta(x,y)$, and $d(x,y)$.

2.6 Algorithms

Any set of complete descriptions of processes and objects is suitable as the basis of software to analyze and retrieve information about them. When we have complete descriptions, it is relatively straightforward to construct algorithms that answer questions such as:

- How does process P take place, in these conditions?
  - Identify the version that satisfies all the necessary relationships r_i that must be satisfied for the constituent stages to take place
  - Identify the specific individuals that serve as each object.
- What happens if process P does not take place?
  - Find all processes Q in which there is a relationship r_i stating that stage Z of Q can occur only if P has occurred.
- What happens if there are none of object O (such as with knockout experiments)?

- Find all processes P in which O is an individual for element E in stage Z. Since no O is available, Z cannot occur, so all versions of P including Z cannot occur, and if there is no version of P without Z, P itself cannot occur.

These algorithms were successfully implemented and tested in [3].

Things are much more difficult when there is partial information at multiple levels. Many, perhaps most, molecular and cellular processes and structures are not fully understood down to the individual molecule level. This requires formal specifications integrating descriptions (knowledge) at multiple levels, and algorithms designed to operate on incomplete specifications at multiple levels. ESs appear to be the first formalization designed for this multiple-level representation task. Several software systems implementing algorithms for the above queries, and others, have been built based ES knowledge bases [2, 3].

3. Measuring Similarity and Complexity

Biologists routinely use concepts of complexity and similarity of structures and processes in analyzing situations, looking for related structures and processes, and formulating research questions. However, these concepts have, until now, only been articulated in an intuitive, rather than a formal, way and as a result researchers have not been able to use them directly. For example, it would seem obviously valuable to be able to query a database for enzymes similar to a given one, similar at all levels of structure. Retrieval by similarity – the ability to do BLAST searches – is the heart of the value of DNA and protein databases, but such searches are limited to primary sequence similarity.

In this section we use the ES formulation to mathematically define the concepts of complexity and similarity of any two entities. This makes possible the quantification of similarity between structures, processes, or states of affairs that reflects structural differences at every level, not only primary sequence similarity.

We first define the *structural complexity* of an Entity A, with N constituents A_1, ...,A_N and K relationships, recursively as:

$$SC(A) = \sqrt{N^2 + K^2 + \varepsilon \cdot \sum_{i=1}^{N} SC(A_i)^2} \qquad (1)$$

ε is an experimentally-determined multiplier modulating the impact of complexity of constituents, sub-constituents, etc.

In formally defining a similarity measure on pairs of arbitrary entities, such as biological structures or constituents of them, we want to take into account the following intuitions:

- The measure should be responsive to differences in attributes of the entities themselves.
- The measure should be responsive to similarity of structure. Structure differences in an entity are represented by having different relationships among constituents, or in having relationships to a different degree.
- When structure of the constituents of the entities is known, the similarity between A and B should reflect the similarity of the their respective constituents.

Accordingly, we define the structural distance between two entities in terms of the difference of (1) the properties of the constituents, and (2) how much the relationships between the constituents differ, as follows:

Assume we have two entities A and B whose structural similarity is to be calculated. Denote the constituents of A and B by A_1, ..., A_{NA} and B_1, …, B_{NB}, respectively. Let the properties of A and B of interest be p_1,…,p_M. Denote the relationships between A-constituents by r_1, ..., $r_{K,}$ and those between B-constituents by r_{K+1}, ..., r_{K+L}.

First, re-order the constituents of A in order of decreasing complexity, as measured by Formula (1), and similarly with the constituents of B.

We represent the properties of A- and B-constituents in a Property Matrix P, and the relationships between constituents with a Relationship Matrix R. P is defined as follows:

- P has M columns (one for each property of interest).
- Let the top NA rows of P represent the constituents of A, in order, and the next NB rows represent the constituents of B, in order.
- The matrix entries are the values of each constituent on each property p_i.
- If a constituent does not have property p_i, that matrix entry is blank.

	p_1	...	p_M
A_1			
...			
A_{NA}			
B_1			
...			
B_{NB}			

Figure 4: The Property Matrix P

P now represents the properties of interest of the A- and B-constituents. In order to meaningfully compare numerical values representing disparate properties, the value of P must be normalized. Accordingly,

- If any column has a value < 0, re-scale the values of the column by adding the absolute value of the minimum value of the column to each value in it. This makes the minimum value of each column 0.
- Normalize the values of P to the range 1 to 10, by setting
  $p_i(A_j) = 10 * (p_i(A_j) + 1) / (pmax_i+1)$,
  where $pmax_i$ is the maximum value of column i. (The value of 10 is an empirically-determined, selected to emphasize the relative importance of property differences compared to the number of constituents.)
- Set each empty entry of P to 0.

The values of the property matrix P are now between 0 and 10, 0 indicating the component does not have the property of that column, and 1 being the minimum actual property value.

We can now define the p*roperty distance* between any A- and B-constituents, A_i and B_j, by using the Euclidean distance between the corresponding A- and B- rows of P:

$$PD(A, B) = \sqrt{\sum_{i=1}^{M} (p_k(A_i) - p_k(B_j))^2} \qquad (2)$$

Since the rows of P representing properties of A-constituents are sorted in order of most-complex-first, as are the rows of P representing properties of B-constituents, we have a consistent procedure for deciding which A-constituent and B-constituent to compare. For example, if we are calculating the structural similarity of the ribosomes of two species, the calculated value would differ significantly depending on whether the two large subunits and two small subunits are compared, rather than the large subunits being compared to the small, and the ordering ensures that the large are compared to the large, etc.

We now use a similar matrix technique to calculate similarity based on *structure*, rather than properties. Structure is specified by relationships between A- or B-constituents, each relationship r_j being represented by an ordered n-tuples. Each relationship has a specific value. For example, in R-state hemoglobin, the angle between the $\alpha_1\beta_1$ and $\alpha_2\beta_2$ dimers is 15°. Thus, the relationship has the formal name "angle," and angle($\alpha_1\beta_1$, $\alpha_2\beta_2$) = 15.

Denoting the number of A-tuples by NAT, and the number of B-tuples by NBT, we define R as follows:

- R has K+L columns, one for each relationship.
- Each row of R represents one tuple of A- or B-constituents, so there are NAT+NBT rows.
- The matrix entries are the values of the relationships have on the tuples. For example, the entry for the matrix at the row ($\alpha_1\beta_1$, $\alpha_2\beta_2$), column "angle," is 15.
- If a tuple does not have relationship r_k, the corresponding entry of the matrix is blank.

	r_1	...	r_K	r_{K+1}	...	r_{K+L}
A-tuple$_1$						
...						
A-tuple$_{NAT}$						
B-tuple$_1$						
...						
B-tuple$_{NBT}$						

Figure 5: The Relationship Matrix R

The values of R must be normalized in order to be able to make meaningful calculations with the values, as were the values of P:

- If any column has a value < 0, re-scale the values of the column by adding the absolute value of the minimum value of the column to each value in it.
- Normalize the values of R to the range 1 to 10, by setting
  $r_i(A_j) = 10 * (r_i(A_j) + 1) / (rmax_i+1_)$,
  where $rmax_i$ is the maximum value of column i. (As with P, 10 is an empirically-determined value chosen to emphasize the relative importance of relationship differences compared to number of constituents.)
- Set each empty entry of R to 0.

The A-constituent and B-constituent rows of P are ordered, to ensure a consistent calculation procedure. It is necessary to have a consistent scheme

for calculating the Euclidean distance between rows of R as well, for much the same reason. Therefore, for any A-tuple ta_j, let $tb_{\kappa(j)}$ denotes the B-tuple closest to ta_j, using Euclidean distance, i.e., the B-tuple most similar to ta_j.

We can now define the total distance between two Entities A and B in terms of the property distance and the structural distance:

$$TD(A, B) = \sqrt{PD(A, B)^2 + SD(A, B)^2} \qquad (3)$$

The *structural distance* SD(A, B) is defined recursively, using the matrix R, as follows:

Let MC = max(NA, NB) and MT = max(NAT, NBT). Then if both A and B have Descriptions, i.e., specified constituents and relationships, we define the structural distance SD as

$$SD(A, B) = \sqrt{(NA-NB)^2 + \sum_{i=1}^{MC} PD(A_i, B_i)^2 + \sum_{j=1}^{MT} \sum_{i=1}^{K+L} (r_i(ta_j) - r_i(tb_{\kappa(j)}))^2 + \delta \cdot \sum_{i=1}^{MC} SD(A_i, B_i)^2} \qquad (4)$$

If NA > NB, $PD(A_i, B_i) = PD(A_i, 0)$ for i > NA, and similarly if NB > MC.

If NAT > NBT, $r_i(tb_j) = 0$ for NBT < j <= NAT, and similarly if NBT > NAT.

If NA > NB, there is no B-constituent to for the A-constituent, so $SD(A_i, B_i) = SC(A_i,)$, for NB < i <= NA, and similarly if NB > NA.

If either A or B have no Description, SD(A, B) = 0.

δ is an experimentally-determined discount factor reflecting the relative importance of the distance between constituents of A and B. (As with ε, preliminary work indicates a value of approximately 0.7 for δ.)

Intuitively,

- $PD(A_i, B_i)$ measures similarity of properties of each pair of constituents.
- $\sum_{i=1}^{K+L} (r_i(ta_j) - r_i(tb_j))^2$ measures how much the constituents of A and B differ on relationship r_i; and the sum $\sum_{j=1}^{MT} \sum_{i=1}^{K+L} (r_i(ta_j) - r_i(tb_{\kappa(j)}))^2$ measures the total difference in structures A and B, as articulated by the relationships r_i between A- and B-constituents.

If A and B are the same except for differing only in names of constituents and relationships (mathematically, are isomorphic), TD(A, B) = 0.

As the properties of A and B, the number of their constituents, the properties of the constituents, the structure of A and B, and the substructures of A and B diverge, TD(A, B) increases.

3.1 Examples

We illustrate the calculation of SD with two examples: the simple structures H_20 and NH_3, and the more complex case of eukaryotic and prokaryotic ribosomes, which illustrates the recursive calculation and the application of the measure in the presence of incomplete information.

3.1.1. Structural Similarity of H_20 and NH_3

For the purposes of this example, we ignore $PD(H_20, NH_3)$, so $TD(H_20, NH_3) = SD(H_20, NH_3)$, i.e., we calculate similarity due solely to structural differences between H_20 and NH_3. We assume that the properties of interest are atomic mass and electronegativity, and the relationships of interest are distance D and bond angle α between the central atom and non-central ones. (This example illustrates the fact that TD may be considered a class of measures rather than a single one, for the particular similarity values will depend on the properties and relationships included in the calculation. Choice of properties and relationships depends on the particular application.)

We suppose that the member attributes of interest in this case are atomic mass and electronegativity of the constituents, which give the P and R matrices shown in Tables 1 and 2:

	Atomic mass	Electronegativity
O	16	3.44
Hw	1	2.2
Hw	1	2.2
N	14	2.04
Ha	1	2.2
Ha	1	2.2
Ha	1	2.2

Table 1: P matrix for H_20 and NH_3

	D	α
(O, H)	95.84	104.5
(O, H)	95.84	104.5
(N, H)	101.7	107.8
(N, H)	101.7	107.8
(N, H)	101.7	107.8

Table 2: R matrix for H_20 and NH_3

Normalizing P and R and re-ordering rows so that the pairs of most similar rows are adjacent results in Tables 3 and 4:

	Normalized D	Normalized α
(O, H)	9.4	9.7
(N, H)	10.0	10.0
(O, H)	9.4	9.7
(N, H)	10.0	10.0
(N, H)	10.0	10.0

Table 3: Normalized P for H_20 and NH_3

Normalized atomic mass	Normalized electro-negativity
10.0	10.0
8.8	5.9
0.6	6.4
0.6	6.4
0.6	6.4
0.6	6.4
0.6	6.4

Table 4: Normalized R for H_20 and NH_3

From Formula (3) above, $TD(H_20, NH_3) = \sqrt{1 + 18.25 + 201.8} = 14.87$.

Similar calculations with CO_2 yield Table 5:

	H_20	NH_3	CO_2
H_20	0	14.87	7.23
NH_3		0	19.91
CO_2			0

Table 5: TD of H_20, NH_3, and CO_2

3.1.2 Structural similarity of eukaryotic and prokayotic ribosomes

Section 2.5.3 shows an ES of the eukaryotic ribosome. The analogous ES of the prokaryotic ribosome, from [11], is:

N: prokayotic ribosome
Relationships:
a. molecular weight(ribosome.prokayotic) = 2.500,000

D:
Paradigm: 1
Sub-objects:
1. [SRSU]
2. [LRSU]

Relationships:
a. molecular weight(SRSU) = 900,000
b. molecular weight(LRSU) = 1,600,000
c. adjacent(LRSU, SRSU)

N: large ribosomal subunit
Relationships:
a. molecular weight(LRSU) = 1,600,000

Paradigm: 1
Sub-objects:
1. [5S RNA]
2. [23S RNA]
3. [Protein1]

…
36. [Protein34]

N: 5S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
120. [Nucleotide120]

N: 23S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
2900. [Nucleotide2900]

N: SRSU
Paradigm: 1
Relationships:
a. molecular weight(LRSU) = 900,000
Paradigm: 1
Sub-objects:
1. [16S RNA]
2. [Protein1]
…
22. [Protein21]

N: 18S RNA
Paradigm: 1
Sub-objects:
1. [Nucleotide1]
…
1540. [Nucleotide1540]

Denoting the eu- and prokaryotic ribosomes Rib-eu and Rib-pro, from (3) we have

$$TD(Rib_{eu}, Rib_{pro}) = \sqrt{PD(\text{Rib-eu, Rib-pro})^2 + SD(\text{Rib-eu, Rib-pro})^2} =$$

$$\sqrt{4.05^2 + SD(\text{Rib-eu, Rib-pro})^2} =$$

$$\sqrt{16.38 + SD(\text{Rib-eu, Rib-pro})^2}$$

Calculating SD(Rib-eu, Rib-pro) from (4), we have MC = 2 and $\sum \sum (r_i(ta_j) - r_i(tb_{\kappa(j)}))^2 = 0$ because the only constituent relationship is adjacency, which is true of both eukaryotic and prokaryotic ribosomes. Setting $\delta = 0.7$, we have SD(Rib-eu, Rib-pro) =

$$\sqrt{(2-2)^2 + \sum_{i=1}^{2} PD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2 + 0 + 0.7 * \sum_{i=1}^{2} SD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2}$$

From the normalized P matrix, the term

$$\sum_{i=1}^{2} PD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2$$

$$= 4.3^2 + 3.6^2 = 31.1.$$

The term $\sum_{i=1}^{2} SD(\text{Rib-eu}_i, \text{Rib-pro}_i)^2$

$$= SD((\text{LRSU-eu, LRSU-pro})^2 + SD(\text{SRSU-eu, SRSU-pro})^2$$

Again from (4), SD(LRSU-eu, LRSU-pro) =

$$\sqrt{(52-36)^2 + \sum_{i=1}^{52} PD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2 + 0 + 0.7 * \sum_{i=1}^{52} SD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2}$$

The term $\sum PD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2 = 0$, because the ESs here (which are a formalization of the description in [11]) do not include properties of the constituents of the LRSU.

The term $\sum_{i=1}^{52} SD(\text{LRSU-eu}_i, \text{LRSU-pro}_i)^2$

becomes $SD(28S, 23S)^2 + SD(5.8S, 5S)^2 + SD(5S, 0)^2 + 34*0 + 15*1$, because the proteins of the prokaryotic LRSU correspond to 34 of the 49 of the proteins in the eukaryotic LRSU and there is no further specification of those proteins. (Were there such specifications, as there would be with a specification of the ribosomes' structure down to the amino acid or atom level, these terms would not be 0.) The only specification of structure of the rRNA constituents of the LRSU are the numbers of nucleotides in them, so $SD(28S, 23S)^2 + SD(5.8S, 5S)^2 + SD(5S, 0)^2 = (4700-2900)^2 + (160-120)^2 + 120^2 = 3{,}256{,}000$, and SD(LRSU-eu, LRSU-pro) =1509.8.

Similarly, SD(SRSU-eu, SRSU-pro) =

$$\sqrt{(34-22^2 + \sum_{i=1}^{34} PD(\text{SRSU-eu}_i, \text{SRSU-pro}_i)^2 + 0 + 0.7 * \sum_{i=1}^{34} SD(\text{SRSU-eu}_i, \text{SRSU-pro}_i)^2}$$

$$= \sqrt{64 + 0.7 * SD(18S, 16S)^2} = 360.1.$$

Thus, SD(Rib-eu, Rib-pro) =

$$\sqrt{0 + 31.1 + 0.7*(1509.8 + 360.1)} = 114.5, \text{ and}$$

$$TD(Rib_{eu}, Rib_{pro}) = \sqrt{16.38 + SD(\text{Rib-eu, Rib-pro})^2}$$

$$= \sqrt{16.38 + 114.5}$$

$$= 11.4$$

3.2 Discussion

It was noted above (Sec. 2.5.3) that there is no single correct description of an entity. Correspondingly, there is no single "correct" value of TD or SD. Rather, as we have seen in the ribosome example, the calculation depends on the particular properties and relationships chosen as the basis of the calculation, and on the information represented in the particular ESs used, which may reflect information either omitted or unknown. Thus, in use, a researcher first specifies the properties and relationships of interest in the particular investigation, and uses structural similarity search to find structures, mechanisms, etc. similar *in those terms*. For example, it may be of value to find structures with a similar number of constituents in similar positional relationships, without regard to net charge on the overall structure, or structures very similar in shape (as measured by similarity of angle and position relationships) but ignoring the properties of a particular constituent. Or, as in ribosome example, the question of interest may be, "How similar are enzymes A and B in high-level structure, ignoring the fine structure of the proteins in each?"

The work of building large knowledge bases of comprised of Entity Specification of biological knowledge is in its initial stages, and work on building tools to support the creation of ESs is also in its initial stages. Because there are many possible descriptions of an entity, creating ESs that are accurate formalizations of existing, informal, descriptions requires some expertise in biology. This means the work must be done by people with some training in biology, or in collaboration with them. Experience to date, however, indicates that producing good ESs does not require professional-level expertise.

4. Relationship to other work

Entity specifications are based on the "representation formats" of P. G. Ossorio, the Object Unit, Process Unit, and State of Affairs Unit [2]. The representations formats, especially the Process Unit, were the basis of several successful computer systems implementing the algorithms enumerated at the beginning of Sec. 2.6. These included a number of query systems [3] and LDS/UCC, a large system to actually carry out the processes specified [4]. The LDS/UCC system shows the applicability of ESs to simulation, especially when knowledge of the structures and processes is incomplete and at multiple levels of detail, as is commonly the case in biology.

Representation formats have a clear similarity to frames, but are a substantial refinement of the concept of frame. The most important distinction is that the constituents of ES are those that must be present by definition of the entity, whereas a frame is defined simply as "things commonly found together" [5], or as in Protégé [7], a related concept. (Interestingly, while clearly a refinement of frames, Ossorio's work predates the introduction of frames by several years [6].) Entity Specifications may be viewed as a rigorous version of frames, combined with the mathematical logic approach to inclusion of relationships.

Class hierarchies, i.e., ontologies, are the most common representation of biological knowledge. As we have seen, a set of ESs specifying an entity at multiple levels of detail is a hierarchy, and thus a set of ESs has a superficial resemblance to an ontology. However, the resemblance is only superficial, specifically in that both ontologies and ESs are hierarchically structured. Ontologies are designed to represent class membership and inheritance information; ESs are designed to represent structure. In an ontology, a child node represents a particular kind of the parent node, and members of the sub-class inherit attributes defined on the super-class; in ESs, a child node represents a constituent of the larger entity. Properties of entities and relationships between them are not inherited by their constituents. Thus, both the information represented and the fundamental concept denoted by the parent-child relationship in the node hierarchy are entirely different. While relational or attribute knowledge are often included in ontologies, the relations and attributes are any of interest, not those that define the items and its structure.

Frame-based systems such as Protege can be used to define an ontology, but a slot in a Protege frame is not the same as a constituent of an entity. While using frames, descriptions using Protege are nonetheless ontologies, and thus represent class hierarchies, not constituents and inter-constituent relationships.

Some ontologies, such as Gene Ontology [8], include a specific relationship, *part_of*, which specifies that one item is part of another. This is the

same concept as that of entity being a constituent of another. The difference between GO and ES is that while both provide a mechanism for specifying that one item is part of another, only ES provides mechanisms for specifying the other facts about the parts: how the parts are related (the set of n-ary relationships $\{r_j\}$) and eligibilities $\{(C_j, i, r)\}$ that specify rules for which actual thing *i* may serve in the role of each constituent C_j. For example, in both GO and ES we can specify formally that $\alpha_1\beta_1$ and $\alpha_2\beta_2$ are parts (constituents) of hemoglobin, but in ES we can also specify formally that angle between the $\alpha_1\beta_1$ and $\alpha_2\beta_2$ is 15°. Since structure is defined by the relationships $\{r_j\}$, this means ES provides the formal mechanism for specifying all aspects of structures and mechanisms, rather than the bare fact that one thing is a part of another.

The GO relationship *is_a* allows definition of class hierarchies; and *regulates, positively_regulates and negatively_regulates*, identify inter-process relationships. These are the only relationships in GO. Entity Specification incorporates the formal specification of any relationship, as in mathematical logic. GO relationships are therefore special cases or instances of ES relationships.

Peleg *et al* [9] integrates hierarchical process descriptions and participant-role logic, using organization workflow models combined with the Tambis [10] ontology to model biological processes. Tambis has the difficulties of any ontology: the only relationships that can be represented in it are those derivable from the pre-defined base relationships combined by subset/superset and "is part of."

Certain of the concepts in this paper were also presented in [12], in the context of applications in the social sciences.

References

[1] International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies, 2008 (BIOTECHNO '08) Publication Date: June 29 2008-July 5 2008, Bucharest, Romania, pp. 100-108. ISBN: 978-0-7695-3191-5, INSPEC Accession Number: 10091005, Digital Object Identifier: 10.1109/BIOTECHNO.2008.13. Current Version Published: 2008-07-15

[2] P. G. Ossorio, *"What Actually Happens" – The Representation of Real World Phenomena*, *Descriptive Psychology Press*. Ann Arbor, MI: Descriptive Psychology Press, 2005. Originally published by: University of South Carolina Press, Columbia, SC, 1978.

[3] H. J. Jeffrey and A. O. Putman, "MENTOR: replicating the functioning of an organization," in *Advances in Descriptive Psychology*, vol. III, K. E. Davis, Ed. Greenwich, CT: JAI Press, 1983.

[4] H. J. Jeffrey, T. Schmid, H. P. Zeiger, and A. O. Putman, "LDS/UCC: Intelligent Control of the Loan Documentation Process," *Proceedings of the Second International Conference on Industrial & Engineering Applications of Artificial Intelligence and Expert Systems*, University of Tennessee Space Institute, Tullahoma, Tennessee, June 1989. ACM Press, 1989.

[5] M. Minsky, "A Framework for Representing Knowledge," *The Psychology of Computer Vision*, P. Winston, Ed. New York: McGraw-Hill, 1975.

[6] P. G. Ossorio, *State of Affairs Systems: Theory and Technique for Automatic Fact Analysis*. Rome, NY: Air Force Rome Air Development Center, RADC-TR-71-1021971, 1971.

[7] Online: http://www.protege.stanford.edu. Last accessed May 6, 2009.

[8] Online: http://www.geneontology.org. Last accessed May 6, 2009.

[9] M. Peleg, I. Yeh, and R. Altman, "Modeling biological processes using workflow and Petri Net models," *Bioinformatics* 8, No. 6, 2002: 825-837.

[10] P. G. Baker, C. A.Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass, "An ontology for bioinformatics applications," *Bioinformatics* 15, 1999: 510-520.

[11] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., *Molecular Biology of the Cell*, 4th ed., New York: Garland Science (Taylor & Francis), 2002, pp. 343-347.

[12] H. J. Jeffrey, "High-Fidelity Mathematical Models of Social Systems," AGENT 2007 Conference on Complex Interaction and Social Emergence, November 15-17, 2007. (Sponsored by Northwestern University and Argonne National Laboratory, in association with the North American Association for Computational Social and Organizational Science.) http://agent2007.anl.gov/2007pdf/Agent%202007%20Proceedings.pdf. Last accessed May 1, 2009.

A Stratified Beta-Gaussian Finite Mixture Model for Clustering Genes With Multiple Data Sources

Xiaofeng Dai
Department of Signal Processing,
Tampere University of Techonology
Tampere, Finland
Email: xiaofeng.dai@tut.fi

Harri Lähdesmäki
Department of Signal Processing,
Tampere University of Technology
Tampere, Finland
Helsinki University of Technology
Department of Information and Computer Science
Helsinki, Finland
Email: harri.lahdesmaki@tut.fi

Olli Yli-Harja
Department of Signal Processing,
Tampere University of Techonology
Tampere, Finland
Email: yliharja@cs.tut.fi

***Abstract*—This paper presents a stratified mixture model based clustering framework, sBGMM. It is an extension of one of our previously developed models, BGMM (beta-Gaussian mixture model), which can not only cluster genes based on beta and Gaussian distributed data but also convert information from a third data source to the priors based on which genes are pre-partitioned into several groups. By assigning genes in the same pre-group the same prior probabilities of belonging to a certain cluster, sBGMM transfers information from a third data source into the results and allows a high level of flexibility in the choice of the third data source. Different from data sources that are modeled as the component of the joint model, information used for prior construction can be from any sources and of any level of sparsity. Besides the extremely flexible choice of prior, sBGMM can also be extended to other parametric distributed data, which adds even more flexibility to this model-based clustering framework. We developed an expectation maximization algorithm for jointly estimating the parameters of sBGMM, and propose to tackle model selection problem by approximation based model selection criteria, where four well-known penalized methods, Akaike information criterion, a modified Akaike information criterion, the Bayesian information criterion, and the integrated classification likelihood-Bayesian information criterion, are tested and compared. Both simulation and real case study indicate that information from different data sources can reinforce each other and utilizing information from one data source to stratify the model can improve the clustering accuracy especially when the noise is comparatively high in both beta and Gaussian distributed data. Applications with full set of real mouse gene expression data (modeled as Gaussian distribution) and protein-DNA binding probabilities (modeled as beta distribution) not only yield more biologically reasonable results compared to its non-stratified version, but also discovered the relationship between two set of genes and eight TFs, which are all likely to be involved in Myd88-dependent Toll-like receptor 3/4 (TLR-3/4) signaling cascades.**

***Keywords*—stratified finite mixture model; gene clustering; multiple data fusion; prior**

I. Introduction

Gene clustering has become one of the most explosively expanding tools for genome-level data analysis, such as inferring gene functions [34] and identifying genes involved in a particular molecular pathway [28]. Numerous computational methods have been developed for it, among which the most prevalent ones include hierarchical clustering [9], K-means [15], and Self-Organizing Maps [32]. These approaches are generally applied to gene expression data [16], which although have demonstrated their usefulness in applications [29], are over dependent on the similarity among gene expression patterns, rendering the results less accurate due to the varied transcriptional coherence in response to diverse environmental stresses and vulnerable to system or experimental error because of using single data source alone and no reinforcement from other data sources.

Multiple data fusion has been widely applied to many problems in the field of system biology, assuming that information from different data sources reinforce each other and can offer us a general view of the system from different perspectives. Nowadays, as more and more different biological data sources, such as protein-DNA binding probabilities, protein-protein interactions, evolutionary conservations histone modifications and methylation information, et cetera, are becoming available since new experimental techniques keep emerging, it is possible to cluster genes based on multiple data sources and promising to group genes based on multiple criteria. Therefore, how to efficiently utilize heterogeneous data sources has become one of the most challenging problems in this field.

Gene clustering method can be roughly classified into three categories, which are heuristic, iterative relocation and model-based methods [11]. Common restrictions of using methods that belong to the first two categories are the determination of the number of clusters and handling with the outliers, which however can be easily solved by model-based methods. Due to the clear definition of what a cluster is, a subpopulation with a certain distribution, model-based methods handle with outliers by recasting it as the model selection problem and adding one or more components, respectively [11], [17], [24]. Also, model-based method beats the first two approaches in its statistical nature [11].

In the realm of standard model based gene clustering, besides the most commonly used statistically method, Gaussian mixture model (GMM) [3], [10], [12], [18], [20], [24], [31], [37], mixture models of some other distributions have also

been developed to solve various problems, such as using a two-component beta mixture model (BMM) to cluster correlation coefficients [17] and applying multinomial models to high dimensional text clustering [22], [36]. While most works on model based methods are devoted to exploring novel applications or improving the computational complexity of the algorithm regardless of the information source, [25] proposed a GMM that can incorporate priors beyond expression data by allowing genes that share the same biological function to have an equal prior probability while differ from the other genes in gene clustering.

Inspired by the promising results brought out by stratifying the priors in GMM [25] and the obvious superiority of data fusion over using single data sources alone, we developed a stratified joint mixture model, sBGMM, to cluster genes based on beta and Gaussian distributed data and stratify the prior according to a third data source. This algorithm differs from our previously developed joint mixture model, BGMM [7], in its utilization of three data sources by converting information from a third one to the prior of the joint mixture model. Also, it exceeds the work of [25] by integrating multiple data sources. Moreover, besides the flexible framework inherited from BGMM, sBGMM assigns more freedom to the choice of prior, which is not restricted to any distribution or limited to the completeness of the data.

We have previously developed an approximated (optimize the complete log-likelihood instead of its expectation) expectation maximization (EM) algorithm for BMM, and a hybrid EM algorithm, where EM for beta and Gaussian distributions are approximated and standard version, respectively, for sBGMM. Encouraged by the flexibility provided by sBGMM and its excellent simulation performance shown in [1], we further extend the hybrid EM for sBGMM to the standard EM in this paper, and test it under more simulation scenarios and with real data.

Many statistical methods can be applied to solve the model selection problem, where four well known penalized likelihood criteria (which belong to approximation-based model selection criteria [30]), Akaike information criterion (AIC) [2], [4], modified AIC (AIC3) [4], [5], Bayesian information criterion (BIC) [25], [27], and integrated classification likelihood-BIC (ICL-BIC) [17] are compared in sBGMM in this study. Based on the simulation results, where sBGMM was compared with BGMM under different scenarios, ICL, other than AIC3 which is proposed for being used in the approximated version of sBGMM [1], performs best in sBGMM.

The following sections are organized as 'Methods', 'Results', and 'Conclusions', where mixture model based clustering and EM algorithm are heavily discussed in 'Methods', results of performance test with simulations and real data, as well as a real case application are shown in different subsections of 'Results', and in 'Conclusions' we first summarized this work, and then discussed its limitation and possible extensions.

II. METHODS

This section introduces the proposed algorithm, including the clustering framework, EM algorithm, prior construction, model selection, and its initialization and convergence.

A. Stratified beta-Gaussian mixture model clustering framework

In model-based clustering methods, each observation $\mathbf{x}_j$, where $j = 1, \ldots, n$ and n is the number of genes, is drawn from a finite mixture distribution with the prior probability π_i, component-specific distribution $f_i^{(g)}$ and its parameters θ_i. The formula is given as [21]

$$f(\mathbf{x}_j|\theta) = \sum_{i=1}^{g} \pi_i f_i^{(g)}(\mathbf{x}_j|\theta_i), \tag{1}$$

where $\theta = \{(\pi_i, \theta_i) : i = 1, \ldots, g\}$ is used to denote all the unknown parameters, with the restriction that $0 < \pi_i \leq 1$ for any i and $\sum_{i=1}^{g} \pi_i = 1$. Note that g is the number of components in this model. In the following texts, we ignore the superscript (g) from $f_i^{(g)}$ for simplicity.

In order to integrate as many information sources as possible, we propose in this paper an sBGMM

$$f_{(k)}(\mathbf{x}_j|\theta_{(k)}) = \sum_{i=1}^{g} \pi_{(k),i} f_i^{(g)}(\mathbf{x}_j; \theta_i), \tag{2}$$

where $1 \leq k \leq K$. It means that the genes can be partitioned into several groups, say $G_1, \ldots, G_K$, based on additional prior before EM is run, and the K stratified models share the same set of component distributions while differ in their usage of stratum-specific prior probabilities.

Define $\theta = [\pi, \theta_1, \theta_2]^T$, $\pi = \left[\pi_{(1)}, \ldots, \pi_{(K)}\right]^T$, $\theta_1 = [\alpha_{11}, \ldots, \alpha_{gp_1}, \beta_{11}, \ldots, \beta_{gp_1}]^T$, and $\theta_2 = \left[\mu_{11}, \ldots, \mu_{gp_2}, \sigma_1^2, \ldots, \sigma_{p_2}^2\right]^T$, where p_1 and p_2 each represents the dimension of the observations in BMM and GMM, respectively, and $\pi_{(k)} = [\pi_{(k),1}, \ldots, \pi_{(k),g}]$ where $k = [1, \ldots, K]$ for K stratified models. We also denote Y and Z as the observations of beta distributed and Gaussian distributed data, respectively, function f of $\mathbf{y}$ and f of $\mathbf{z}$ as the density function of beta and Gaussian distribution, respectively, and $\mathbf{x} = [\mathbf{y}^T, \mathbf{z}^T]^T$.

Apart from adding the prior, sBGMM is built from BMM and GMM with the assumption that, for each component i, the beta distributed and Gaussian distributed data are independent. In the BMM part, each component is assumed to be the product of p_1 independent beta distributions, whose probability density function is defined as

$$f_i(\mathbf{y}|\theta_{1i}) = \prod_{u=1}^{p_1} \frac{y_u^{\alpha_{iu}-1}(1-y_u)^{\beta_{iu}-1}}{B(\alpha_{iu}, \beta_{iu})}, \tag{3}$$

where $\theta_{1i} = [\alpha_{i1}, \ldots, \alpha_{ip_1}, \beta_{i1}, \ldots, \beta_{ip_1}]$ and $\mathbf{y} = [y_1, \ldots, y_{p_1}]^T$. Likewise, each component is assumed to follow a Gaussian distribution in the GMM part, whose probability density function of each component for each gene is defined

as

$$f_i(\mathbf{z}|\theta_{2i}) = \frac{1}{(2\pi)^{\frac{p_2}{2}}|V|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(\mathbf{z}-\mu_i)^T V^{-1}(\mathbf{z}-\mu_i)\big), \quad (4)$$

where $\theta_{2i} = [\mu_{i1},\ldots,\mu_{ip_2},\sigma_1^2,\ldots,\sigma_{p_2}^2]$, $\mu_i = [\mu_{i1},\ldots,\mu_{ip_2}]^T$, $V = \mathrm{diag}(\sigma_1^2,\sigma_2^2,\ldots,\sigma_{p_2}^2)$ and $|V| = \prod_{v=1}^{p_2}\sigma_v^2$. Notice that diagonal covariance matrix is assumed in the Gaussian part, which is especially useful for high-dimensional data since it can significantly reduce the number of parameters that are needed to be estimated from data.

B. EM algorithm

The standard EM algorithm is applied to estimate the parameters θ in sBGMM iteratively, whose derivation is similar with one of our previously developed model, BGMM, as proposed in [6].

The data log-likelihood (natural logarithm is referred to throughout this paper) can be written as

$$\log L(\theta) = \sum_{j=1}^{n}\log\Big(\Big[\sum_{i=1}^{g}\pi_{(k),i}f_i(\mathbf{x}_j|\theta_i)\Big]\Big), \quad (5)$$

given $X = \{\mathbf{x}_j : j = 1,...,n\}$, whose direct maximization, however, is difficult.

In order to make the maximization of Equation 5 tractable, the problem is casted in the framework of incomplete data. Since we assume that the beta and Gaussian distributed data are independent, the complete data likelihood, L_c, can be factored as

$$L_c(\theta) = f(Y|\mathbf{c},\theta)f(Z|\mathbf{c},\theta)f(\mathbf{c}|\theta). \quad (6)$$

If we define $c_j \in \{1,\ldots,g\}$ as the clustering membership of $\mathbf{x}_j$, then the complete data log-likelihood can be written as

$$\log L_c(\theta) = \sum_{j=1}^{n}\sum_{i=1}^{g}\chi(c_j = i)\log\left(\pi_{(k),i}f_i(\mathbf{x}_j|\theta_i)\right), \quad (7)$$

where $\chi(c_j = i)$ is the indicator function of whether $\mathbf{x}_j$ is from the i^{th} component or not.

In the EM algorithm, E step computes the expectation of the complete data log-likelihood

$$\begin{aligned} Q(\theta|\theta^{(m)}) &= E_{\mathbf{c}|X,\theta^{(m)}}(\log L_c) \\ &= \sum_{j=1}^{n} E_{c_j|\mathbf{y}_j,\mathbf{z}_j,\theta^{(m)}}\left[\log\left(f(\mathbf{y}_j|c_j,\theta_1)\right)\right] \\ &\quad + \sum_{j=1}^{n} E_{c_j|\mathbf{y}_j,\mathbf{z}_j,\theta^{(m)}}\left[\log\left(f(\mathbf{z}_j|c_j,\theta_2)\right)\right] \\ &\quad + \sum_{k=1}^{K}\sum_{j\in G_k} E_{c_j|\mathbf{y}_j,\mathbf{z}_j,\theta^{(m)}}\left[\log\big(f(c_j|\pi_{(k)})\big)\right], \end{aligned} \quad (8)$$

where $\theta^{(m)}$ represents the parameters estimated in the m^{th} iteration, and details of the derivation of Q can be found in [21]. By computing the expectation, Equation 8 becomes

$$Q(\theta|\theta^{(m)}) = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log(\pi_{(k),i}f_i(\mathbf{y}_j|\theta_{1i})f_i(\mathbf{z}_j|\theta_{2i})), \quad (9)$$

where

$$\begin{aligned} \tau_{ji}^{(m)} &= p(c_j = i|\mathbf{x}_j,\theta^{(m)}) \\ &= \frac{\pi_{(k),i}^{(m)}f_i(\mathbf{y}_j|\theta_{1i}^{(m)})f_i(\mathbf{z}_j|\theta_{2i}^{(m)})}{\sum_{i'=1}^{g}\pi_{(k),i'}^{(m)}f_{i'}(\mathbf{y}_j|\theta_{1i'}^{(m)})f_{i'}(\mathbf{z}_j|\theta_{2i'}^{(m)})}, \end{aligned} \quad (10)$$

is the estimated posterior probability of $\mathbf{x}_j$, which belongs to the k^{th} layer according to the prior, coming from component i at iteration m according to Bayes' rule. Note that we can assign each $\mathbf{x}_j$ to the component i_0 that maximizes its estimated posterior probability, i.e., $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$. Also, the assumption that the beta distributed and Gaussian distributed data are independent is carried over to the expected log-likelihood as shown by Equations 8 and 9.

To derive the closed form or numerical optimization formula for updating parameters in sBGMM, we used Lagrange multipliers to solve this constrained optimization problem, with the Lagrangian function shown in Equation 11.

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log\left(f_i(\mathbf{y}_j|\theta_{1i})\right) \\ &\quad + \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log\left(f_i(\mathbf{z}_j|\theta_{2i})\right) \\ &\quad + \sum_{k=1}^{K}\sum_{j\in G_k}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log(\pi_{(k),i}) \\ &\quad + \sum_{k=1}^{K}\lambda_k\left(1 - \sum_{i'=1}^{g}\pi_{(k),i'}\right) \end{aligned} \quad (11)$$

Parameters of BMM part, $\theta_{1i} = [\alpha_{i1},\ldots,\alpha_{ip_1},\beta_{i1},\ldots,\beta_{ip_1}]$ $1 \le i \le g$, are optimized by Newton-Raphson method and updated by

$$\theta_{1i}^{(m+1)} = \theta_{1i}^{(m)} - H^{-1}(\theta_{1i}^{(m)})\nabla_{\theta_{1i}}\mathcal{L}(\theta_{1i}^{(m)}), \quad \theta_{1i} \ge \mathbf{1}, \quad (12)$$

where $H^{-1}(\theta_{1i}^{(m)})$ is the inverse of the Hessian matrix evaluated at $\theta_{1i}^{(m)}$, and $\mathcal{L}(\theta_{1i}^{(m)})$ is the Lagrangian function of $Q(\theta_{1i}^{(m)})$.

Parameters of the GMM part, $\theta_{2i} = [\mu_{i1},\ldots,\mu_{ip_2},\sigma_1^2,\ldots,\sigma_{p_2}^2]$ $1 \le i \le g$, in sBGMM can be estimated by the standard EM algorithm of GMM with diagonal covariance matrix as shown in the following closed form formula

$$\hat{\mu}_{iv}^{(m+1)} = \sum_{j=1}^{n}\tau_{ji}^{(m)}z_{jv}/\sum_{j=1}^{n}\tau_{ji}^{(m)}, \quad (13)$$

$$\hat{\sigma}_v^{2,(m+1)} = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ji}^{(m)}(z_{jv} - \mu_{iv}^{(m)})^2/n, \quad (14)$$

which can be obtained by plugging Equation 4 in Equation 11

and taking the derivatives of Equation 11 with respect to μ_{iv} and σ_v^2, respectively.

Optimization of the prior probability of each gene's clustering membership, π, can be derived by taking the derivative of Equation 11 with respect to $\pi_{(k),i}$, i.e.,

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \pi_{(k),i}} &= \sum_{j \in G_k} \tau_{ji}^{(m)} \frac{1}{\pi_{(k),i}} - \lambda_k, \\ \hat{\pi}_{(k),i}^{(m+1)} &= \sum_{j \in G_k} \tau_{ji}^{(m)} / \lambda_k.\end{aligned}$$

Moreover, since

$$\begin{aligned}1 &= \sum_{i=1}^{g} \pi_{(k),i} \\ &= \sum_{i=1}^{g} \frac{1}{\lambda_k} \sum_{j \in G_k} \tau_{ji} \\ &= \frac{1}{\lambda_k} \sum_{j \in G_k} \sum_{i=1}^{g} \tau_{ji} \\ &= \frac{1}{\lambda_k} \sum_{j \in G_k} 1,\end{aligned}$$

thus, $\lambda_k = n_k$. Consequently, the updates of π is given by

$$\hat{\pi}_{(k),i}^{(m+1)} = \sum_{j \in G_k} \tau_{ji}^{(m)} / n_k, \tag{15}$$

where G_k is the k^{th} group with n_k genes, according to the prior.

From the above equations, it is easy to see that the EM of sBGMM will reduce to the EM of BGMM if $K = 1$, and will further reduce to BMM or GMM, respectively, as p_2 or p_1 equals to zero.

1) Prior construction: Priors of Equation 2 can be determined from any possible data sources. It can be either another complete data source different from what have been used in the component models (BMM and GMM), e.g., the pre-cluster results obtained from PPI data [35], or some incomplete information relevant to our problem, e.g., information retrieved from database. In the following study, we employ a complete PPI data set for simulation test, and obtain a set of incomplete information from a database for real case study. Conversion of PPI data into prior is described below.

PPI data, which is typically a binary square matrix, is first converted into contact matrix (denoted as A) and then transformed into pathlength matrix (denoted as P). Contact matrix is in the form of

$$A = \begin{cases} 1 & \text{if } i \Leftrightarrow j \\ 0 & \text{if } i \not\Leftrightarrow j, \end{cases} \tag{16}$$

where $i \Leftrightarrow j$ means the existence of a connection between node i and j while $i \not\Leftrightarrow j$ denotes the other way around. In the pathlength matrix, the pathlength between nodes i and j is denoted as P_{ij} and characterized as the smallest integer $k \geq 1$ such that $(A^k)_{ij} \neq 0$. P contains all the path lengths for all pairs of nodes which are calculated by the 'pathlength' function of the 'CONTEST' toolbox in matlab [33]. We use the pathlength matrix to pre-cluster the genes (corresponding to the proteins they encode) using a simple hierarchical clustering algorithm which employs Euclidean distance as the distance matrix and nearest neighbor algorithm as the linkage construction method, and matlab function 'clusterdata' is used here for this purpose. Then we assume that genes from the same pre-cluster share the same prior probability $\pi_{(k),i}$ of coming from the same cluster i, and allow them coming from different clusters.

C. Model Selection

Four well-known approximation-based model selection criteria, BIC [25], [27], ICL [17], AIC [2], [4], and AIC3 [4], [5] are compared in sBGMM, according to which the best-performing criterion within the tested scope is chosen. Calculations for the above criteria are defined as

$$AIC = -2 \log L(\hat{\theta}) + 2d, \tag{17}$$

$$AIC3 = -2 \log L(\hat{\theta}) + 3d, \tag{18}$$

$$BIC = -2 \log L(\hat{\theta}) + d \log(nM), \tag{19}$$

$$ICL = -2 \log L(\hat{\theta}) + d \log(nM) - 2 \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ji} \log(\tau_{ji}), \tag{20}$$

where d is the number of free parameters, and M (in equations 19 and 20) is the total amount of the data ($M = \sum_{w=1}^{W} M_w$, M_w is the size of data set w and W is the number of input data sets). Note that $-2 \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ji} \log(\tau_{ji})$ is the estimated entropy of the fuzzy classification matrix $C_{ji} = (\tau_{ji})$ [17].

sBGMM has $K - 1$ more free π_i's than BGMM because of the K stratified layers. In BGMM, the number of free parameters d is the summation of those in BMM and GMM minus one set of redundant free π_i's, which is $d_{BG} = 2gp_1 + p_2 + p_2 g + (g - 1)$. Therefore, the number of free parameters in sBGMM is $d_{sBG} = 2gp_1 + p_2 + p_2 g + K(g-1)$.

D. Initialization and convergence

In this study, parameters α_{iu}'s and β_{iu}'s for each dimension of beta distribution u ($u \in \{1, \ldots, p_1\}$) are initialized by method-of-moments so that their means are randomly distributed within the range of $y_{1u}, \ldots, y_{nu}$ and variances are equal for all clusters (g), μ_{iv}'s and σ_v^2's are obtained from the randomly initialized fuzzy c-means clustering results, and π_i's are initialized with the same random value within each group G_k, and the sum of the probabilities of g components is one.

In order to avoid the possible local maxima, we run the algorithm multiple (100) times with different initial values. The convergence threshold (where Q is used to monitor the convergence) and maximum number of iterations were set to 0.0001 and 100, respectively, for all the tested models, and all the simulations have reached their convergence according to the statistics stored during the simulations.

III. RESULTS

We first tested the performance of sBGMM with artificial and real data, respectively, and then applied it to a real biological case, which are discussed separately below.

A. Performance test with artificial data

According to work done in [19] only part of protein-DNA binding data and gene expression data agree with each other (consisting of the same number of clusters), and data can fall into three regions as illustrated in Figure 1. Beta and Gaussian distributed data may share the same number of underlying clusters (denoted as 'Region 1') or may not, and we denote the scenario that beta distributed data has more underlying clusters as 'Region2', and 'Region3' for the scenario of the other way around. To match the three scenarios, we designed data sets 1 to 3 for data of both beta and Gaussian distributions, respectively, whose parameters are listed in Table I. Each artificial data set is designed to fall into five categories: 'good Beta' (gB), 'bad Beta' (bB), 'good Gaussian' (gG), 'bad Gaussian due to close means' (bG_m), and 'bad Gaussian due to large variances' (bG_v), where 'good' stands for low noise level, and 'bad' means the opposite. The dimensions are designed to be $n = 100$ and $p = 4$ for both data sets. We also designed three PPI data sets to test the influence of different priors on the clustering results. Prior 1 and 2 are constructed based on the same underlying ground truth (the same number of underlying clusters and the same clustering membership of each gene) but differ in their noise levels, while prior 3 contains some mis-clustering (clustering membership of some genes are not consistent with the designed Gaussian and beta distributed data) information and shares the same noise level with Prior 1. All sparsity patterns are shown in Figure 2, where the three priors are denoted as 'T9', 'T2', and 'F9', respectively, with the capital letter representing 'true' or 'false' (meaning that there is or there is not mis-clustering information, respectively), and the following number standing for the noise level (the higher the number the lower the noise), e.g., 9 means the intensity of signal over background is 9. All the simulations are repeated 10 times with randomly generated data sets (including the data used for prior construction).

We used the same scoring system as developed in [7] for performance evaluation, which is denoted as 'E score'

$$\begin{aligned} e_j(r) &= \begin{cases} 1 & if \;\; \hat{z}_{ji} = 1 \;\; and \;\; r_i = T_j \\ 0 & otherwise \end{cases} \\ E &= \max_{r \in R} \sum_{j=1}^{n} e_j(r)/n \qquad (21) \\ R &= \big\{ r = (r_1, \ldots, r_{\hat{g}}) : \forall i \neq j \; r_i \neq r_j; \\ &\quad r_i \in \{1, \ldots, \max\{\hat{g}, g\}\} \big\}. \end{aligned}$$

Notations of in this scoring system are defined as follows. T_j denotes the ground truth clustering membership of data j. R stands for all possible associating ways between the estimated and the true clusters, where r_i is the label of data belonging to component i predicted by the clustering algorithm, and r is chosen from labels $1, 2, \ldots, \max\{\hat{g}, g\}$ ($\hat{g}$ and g are the largest labels in the estimated and ground truth clustering respectively). Denote also e as the individual score of each gene, E as the average score of all the genes for each repetition, 'E score' of each repetition as the one corresponding to the optimal Q, and the final 'E score' of each data set as the median of the 20 'E score's. This scoring system evaluates the overall performance of the model since it not only records the accuracy of the results but also reflects the influence of the criterion for model selection.

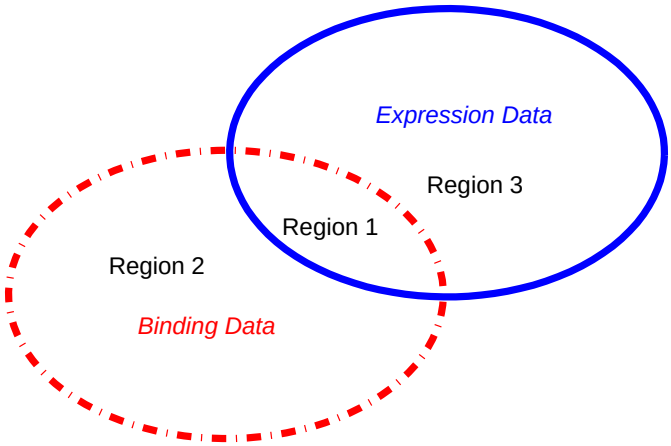


Figure 1. Region divisions of input data. In Region 1 gene expression and protein-DNA binding data have the same number of underlying components; in Region 2 binding data has more components; in Region 3 expression data has more components.

We compared the performance of sBGMM and its non-stratified version, BGMM, with data set 1 to data set 3, each coupled with prior 'T9', 'T2' and 'F9'. Before performance test, we first compared each model selection criterion in handling different scenarios in each model, whose results are shown in Table II. According to the average E scores shown in Table II, there is no universal optimal criterion for sBGMM or BGMM, but ICL is much safer to choose for sBGMM since it selects most of the correct models.

Performance comparison results of sBGMM with its non-stratified version under different scenarios with different priors are shown in Figure 3, where the E scores are calculated with the assumption that the real number of underlying clusters is three (therefore the prior is designed to contain three underlying clusters) and the model is chosen by the criterion that generates the highest average E score under each scenario. It is seen from Figure 3 that sBGMM and BGMM perform equally well when at least one type of data (excluding the prior) contains less noise and has the correct number of underlying clusters for data within 'Region 1', anything combined with 'gB' for data within 'Region 2', and anything combined with 'gG' for data within 'Region 3'. This indicates that our joint models (both sBGMM and BGMM) have the ability to offset the noisy or incorrect information within one type of data by utilizing information from the other one. However, when the noise level, including noise and

Figure 2. Sparsity patterns of the contact matrix of the artificial PPI data sets. (a) 'T9': true prior with noise level equals 9. (b) 'T2': true prior with noise level equals 2. (c) 'F9': false prior with noise level equals 9.

incorrect number of underlying clusters, is too high for both types of data, using additional information becomes important as shown by 'yellow' and 'carmine' in 'Region 1', 'blue', 'yellow' and 'carmine' in 'Region 2', and 'cyan', 'yellow', 'red' and 'carmine' in 'Region 3'. It is also clear that there is no significant difference for using different priors ('T9', 'T2', and 'F9') in sBGMM if the prior does not contain too much mis-clustering information, which means that sBGMM is not sensitive to the noise and is tolerant of small amount of inconsistent information in the prior. All together, these results indicate that adding additional prior can utilize information from more data sources, rendering sBGMM more robust in handling with various scenarios than BGMM.

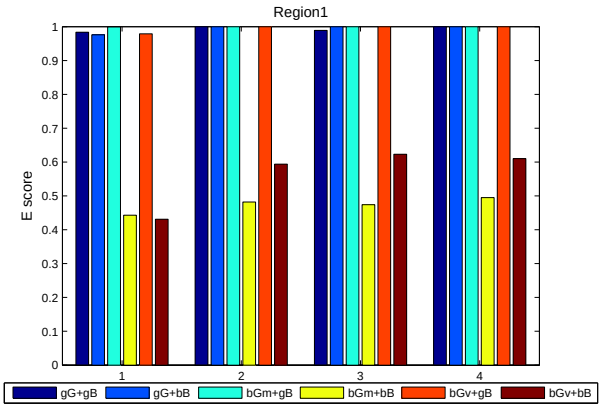
M	P	R	AIC	AIC3	BIC	ICL
BGMM		R1	0.8270	0.8325	0.8459	**0.8464**
		R2	**0.7855**	0.7796	0.7663	0.7723
		R3	0.7763	0.7803	**0.7910**	0.7849
sBGMM	T9	R1	0.8545	0.8680	0.8806	**0.8845**
	T9	R2	0.7434	0.7714	0.8376	**0.8430**
	T9	R3	0.7820	0.7995	0.8188	**0.8270**
sBGMM	T2	R1	0.8459	0.8636	**0.8844**	0.8820
	T2	R2	0.7653	0.8001	0.8305	**0.8391**
	T2	R3	0.7699	0.7941	0.8323	**0.8343**
sBGMM	F9	R1	0.8444	0.8600	**0.8881**	**0.8881**
	F9	R2	0.7430	0.7674	0.8406	**0.8430**
	F9	R3	0.7575	0.7900	**0.8295**	**0.8295**

Note: Values shown here are the averages of E scores over all the tested cases ('gG+gB', 'gG+bB', 'bG$_m$+gB', 'bG$_m$+bB', 'bG$_v$+gB', 'bG$_v$+bB') selected by each criterion in each model. 'P' column shows the priors. 'M' column lists the name of the tested models. 'R' column shows the region that beta and Gaussian distributed data belong to. 'ICL' is short for 'ICL-BIC'. E scores shown in bold face are the selected best criterion with respect to highest average E scores and used in drawing Figure 3. All values are rounded to four decimal points.

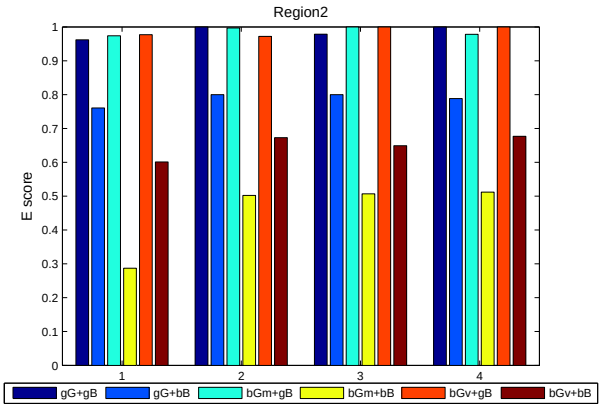
Table II
COMPARISON OF DIFFERENT MODEL SELECTION CRITERIA IN SBGMM AND BGMM.

B. Performance test with real data

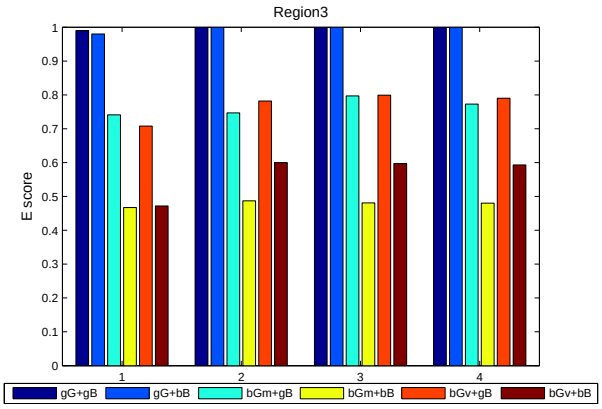
We applied our methods to mouse protein-DNA binding probabilities (modeled as beta distribution) and gene expression data (modeled as Gaussian distribution). The protein-DNA binding data contains the probabilities of 266 TFs binding to 20397 genes, which were calculated with mouse-specific position weight matrices from the TRANSFAC database (the web server and data are available at http://xerad.systemsbiology.net/ProbTF/ [14]). The gene expression data is composed of 1960 genes measured from 95 conditions [26], where six Toll-like receptor (TLR) agonists (C_pG, Pam$_2$CSK$_4$, Pam$_3$CSK$_4$, LPS, poly I:C and R848) were used as the treatments, and four gene knock-out mutants and different time points were included to increase the diversity of the TLR-stimulated gene expression data set and the number of measurements. There are 1766 genes measured in both datasets. We removed the genes whose gene expression profiles have low absolute values (less than 10th percentile) with matlab function 'genelowvalfilter', and then chose genes whose annotations are available through the functional classification tool of DAVID database (whose web server is available at http://david.abcc.ncifcrf.gov/home.jsp [13]). In the end, 673 genes are chosen for the following studies. The chosen protein-DNA binding data (beta distributed data that are used in this study) is composed of the binding probabilities of the 673



(a)



(b)



(c)

Figure 3. Simulation results. Performance comparison of sBGMM with BGMM for (a) region 1 data, (b) region 2 data, and (c) region 3 data. In x-axis of each region, '1' to '4' represent different models, each corresponds to BGMM, sBGMM ('T9'), sBGMM ('T2'), sBGMM ('F9'), accordingly. The y-axis represent the E scores.

genes for six TFs ('Junb', 'Jund1', 'Jun', 'Fos', 'Fosb', and 'Cebpb') which are involved in AP1 gene regulatory network in mouse according to TRED (Transcriptional Regulatory Element Database, which is available at http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home). The Gaussian distributed data that are fitted in the model are the gene expression data of these 673 genes at 23 conditions, which are the midpoint of each time series (selected time point for each treatment are shown in Table III).

Point	Treatment	Time
1	Atf_3^-	0
2	C_pG+Atf_3^-	120
3	LPS+Atf_3^-	240
4	Pam_2CSK_4+Atf_3^-	120
5	poly I:C+Atf_3^-	120
6	$Crem^-$	0
7	LPS+$Crem^-$	240
8	poly I:C+$Crem^-$	360
9	$Myd88^-$	0
10	LPS+$Myd88^-$	60
11	Pam_3CSK_4+$Myd88^-$	60
12	poly I:C+$Myd88^-$	60
13	$TicamI^-$	0
14	LPS+$TicamI^-$	120
15	LPS+Pam_2CSK_4+$TicamI^-$	120
16	no	0
17	C_pG	60
18	LPS	360
19	Pam_2CSK_4	80
20	Pam_3CSK_4	240
21	Pam_3CSK_4+poly I:C	60
22	poly I:C	120
23	R848	120

Note: 'Point' refers to the labels of the x axis; '-' means the mutant strain that does not have the particular gene; time point chosen for the treatment is shown in the column 'Time', and time unit is 'min'. Treatments after point 16 were all applied to the wild type.

Table III
TREATMENTS OF THE GENE EXPRESSION DATA

We constructed two types of priors for real case performance test. Type 1 prior contains two priors (denoted as '$P1_a$', '$P1_b$') which both utilize the transcriptional regulation information stored in TRED. Curation of transcriptional regulation in TRED are done with both experimental evidence and promoter finding tools, and currently involves genes within 36 cancer-related TF families. '$P1_a$' is built from the clustering information of AP1 network, where 10 genes are assigned to two clusters, and the memberships of the rest genes are left unspecified. '$P1_b$' keeps all the clustering membership in '$P1_a$' intact, and specifies the rest memberships from all the other networks in TRED by removing genes that are involved in several networks or form singleton clusters (resulting in 16 more memberships specified). Type 2 priors are obtained from an online classification tool DAVID (the web server is available at http://david.abcc.ncifcrf.gov/gene2gene.jsp [13]), a

database for annotation, visualization and integrated discovery. Different thresholds ('Highest', 'Medium', 'Lowest') were set to obtain different functional classification results, resulting in three different type 2 priors, which are denoted as '$P2_a$', '$P2_b$', '$P2_c$', respectively.

We tested the performance of sBGMM by comparing its performance on clustering the 673 genes (with both types 1 and type 2 priors) with its non-stratified form (BGMM), and both of its component models (BMM, GMM). We employed Gene Ontology (GO) in this study to validate the clustering results. In order to find the most significant annotated terms by looking at the probabilities that the terms are counted by chance, we used the hypergeometric probability distribution to calculate the p-values of gene enrichment score (called 'p-values' for simplicity) for each cluster by each model with each model selection criterion (Bioinformatics Toolbox 3.1 in Matlab). We compared the means and medians of each clustering results by each model with respect to different aspects (molecular function, cellular component, biological process, and all aspects, which are denoted as 'F', 'C', 'P' and 'All'), whose results are shown in Table IV. To see how stable the algorithms work with our test data set, or in other words, whether 100 iterations are enough for convergence, we repeated each set of iterations (100) three times for different models, with one repetition for each model shown in Table IV.

There are at least four pieces of information unveiled by Table IV. First, BGMM works better than BMM and GMM with respect to smaller means and medians of the group p-values. Second, sBGMM can significantly improve the clustering performance compared with BGMM and its component models when the prior is properly chosen. As shown in this table, results generated with type 1 priors are better than those with type 2 priors, whose means and medians are significantly smaller than those of BGMM, BMM and GMM; moreover, sBGMM with type 1 priors generate more stable results than the other models, i.e., two out of three repetitions of sBG_{P1_a} and all three repetitions of sBG_{P1_b} converge to the same clustering, respectively. This is because information delivered by type 1 priors are consistent with that used for choosing TFs of protein-DNA binding probabilities, while type 2 priors, which are the classification results from an online functional classification tool, might group the same gene into another cluster based on its own criteria (DAVID groups genes by measuring the functional relationship of gene pairs based on the similarity of their global annotation profiles [13]). Third, $P1_b$ is denser than $P1_a$, and generates more stable results (three vs. two repetitions converge to the same result), which indicates that the more consistent (consistent with data) information carried out by the prior the more accurate the results will be. However, both type 1 priors used here are quite sparse, therefore, we expect to get even higher accuracy if denser and consistent (consistent with data) prior is available. Fourth, sBGMM with type 1 prior works better than DAVID functional classification tool. As shown in Table IV, all the evaluated quantities of the results obtained from DAVID ($P2_a$, $P2_b$, $P2_c$) are worse than those of sBGMM (coupled with type 1 priors), BGMM, and even some of the results of GMM. Although the improved accuracy can not show the superiority of sBGMM over DAVID since totally different types of data sources are used, the results demonstrate the power of employing gene expression and protein-DNA binding data in gene clustering over relying on global annotation profiles.

C. Biological application with sBGMM

After performance test, we further analyzed all the 1766 genes in our data set. We compared the 1766 genes with the genes involved in all the 36 cancer related gene networks stored in TRED, and decided to extract the information from the network that has the largest overlap with our gene set (NFKB network) for further analysis. There are seven TFs involved in this network, out of which five (which are 'Rel', 'Nfkb1', 'Msx1', 'Rela', 'Myb', and named TF_{normal} for convenience) are available in our data set. Protein-DNA binding probabilities of those five TFs to all the 1766 genes and gene expression data of the 23 midpoint conditions (midpoint of each time series) were chosen as the beta distributed and Gaussian distributed data set, respectively. Genes involved in NFKB network are grouped into six clusters by TRED, among which 42 are present in our data set. We constructed a type '$P1_b$' prior for the whole gene set since it tends to have a more stable behavior compared with '$P1_a$' according to the real case performance test (see the previous subsection).

There are 34 genes that encode TFs (named TF genes) among the whole data set. We first clustered the 34 TF genes with BGMM, for three times, and chose the TF gene cluster which has the smallest enrichment p-values for further analysis. There are 11 genes in the selected TF gene group, out of which eight are repeated clustered together among three repetitions and, for convenience, we call them the 'core TF genes' and denoted as TF_{core} in the following text.

To find the influence of the choice of protein-DNA binding probabilities on the clustering accuracy of sBGMM and find a set of protein-DNA binding probabilities as suitable as possible for further analysis, we first clustered the 1766 genes by sBGMM with binding data corresponding to TF_{normal}, and then re-clustered them with those selected by TF_{core}, each with three repetitions. For comparison purpose, we also clustered the 1766 genes by BGMM with binding data of the core TF genes, and the result of one repetition from each clustering were compared and shown in Table V. Note that the expression data and the prior are the same in the models where they were used.

It is interesting to see from Table V that the group p-values are significantly dropped after using the core TF genes for gene clustering, and the group p-values obtained with sBGMM are overwhelmingly lower than those obtained by BGMM. This means that the core TF genes are more responsible to TLR-stimulated macrophage activation than the TFs chosen based on the prior information obtained from NFKB network, and again demonstrates the superiority of sBGMM over its non-stratified version.

We further analyzed the causal relationship between the set of core TFs and the whole set of genes. We notice that among the eight core TF genes, 'E2F6', 'E2f7', 'Foxm1' and 'Nfatc1' are clustered together with 363 other genes, and 'Rest', 'Rfx5', 'Mxd1' and 'Stat1' fall into the same group with 305 other genes. Moreover, by examining the expression profiles of the two sets of genes under different treatment (shown in Figure 4), it is clear that there is a plateau existed in all profiles from point 26 and 48 where either mutant *Myd88*$^-$ or *TicamI*$^-$ is used, or no treatment is applied or C_pG is added. This indicates that genes *Myd88* and *Ticam1* are crucial for the system (which involves the genes that belong to the four clusters) to response to the external stimuli, and agonist C_pG does not have so much influence on it. Moreover, whenever LPS or poly I:C is added to the wild type (regions between points 5 and 10, 14 and 16, 18 and 22, 23 and 26, 48 and 59, 79 and 87), there is a sharp drop in the red profile while there is a peak in the green curve. This feature indicates that the two set of genes (including the core TF genes) are sensitive to LPS and poly I:C, and behave in an opposite way after being stimulated. Genes that are clustered with 'E2F6', 'E2f7', 'Foxm1' and 'Nfatc1' are activated by them while repressed by TFs 'Rest', 'Rfx5', 'Mxd1' and 'Stat1'; while operation goes the other way around for the other set of genes. Moreover, since poly I:C, LPS and C_pG are TLR-3, TLR-4 and TLR-9 agonists, respectively, and Myd88 and Ticam1 are adaptors involved in TLR-3/4 signaling according to [23], we can deduce that most of the two set of genes (including TF genes) are involved in Myd88-dependent TLR-3/4 signaling cascades.

IV. Conclusion and Future Work

This paper presents a novel method based on stratified beta-Gaussian mixture model, sBGMM, for gene clustering from multiple data sources. In addition to integrating beta distributed and Gaussian distributed data, sBGMM can also facilitate clustering by employing priors which come from a third data source. A stratified version of EM algorithm is developed for jointly estimating parameters from beta and Gaussian distributions, and is used as the core of sBGMM. sBGMM differs from its non-stratified version (BGMM) by setting the same prior probabilities of coming from each cluster to genes that belong to the same layer which are stratified according to the additional prior. In principle, any relevant information can be used as priors, whereas in this study, we built the prior from PPI data in simulations, and retrieved it from database TRED in the real case study. Simulation results show that sBGMM works better than its non-stratified version especially when both beta and Gaussian distributed data contain too much noise, and certain mis-clustering information in the prior is tolerable. In real case study we not only demonstrated the superiority of sBGMM compared with BGMM and a gene annotation based classification method (DAVID functional classification tool) by analyzing 673 genes, but also revealed the relationship of two sets of genes and eight TFs in TLR-stimulated macrophage signaling through analyzing the full data set (1766 genes).

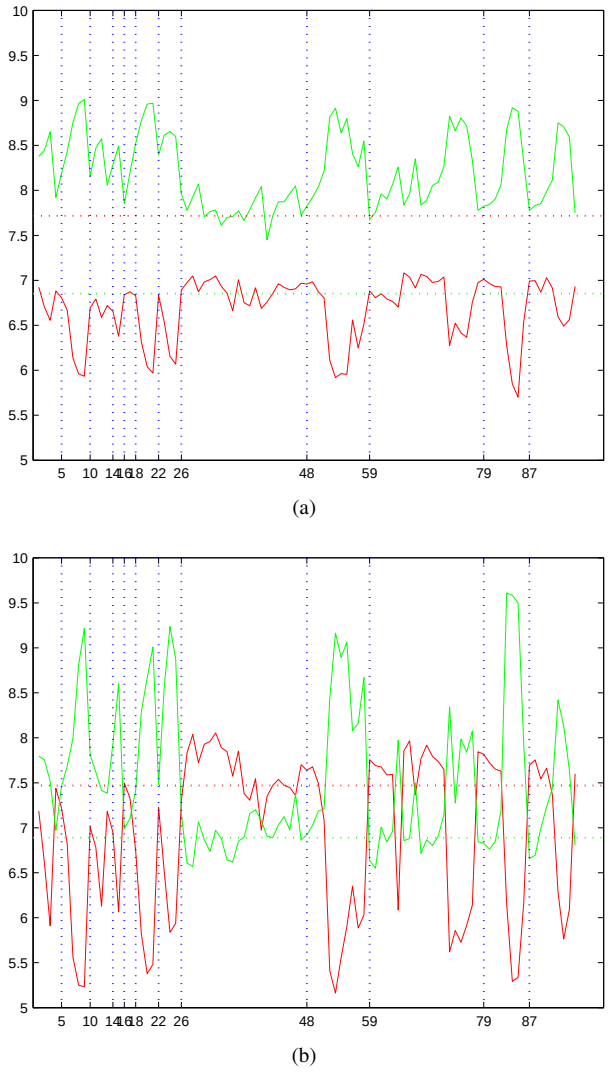


Figure 4. Median gene expression profiles of the (a) interested genes and (b) TFs. Solid curves represent the median expression profile of the genes or TFs. Horizontal dot lines stand for the expression level of wild type without treatment. Vertical blue lines divides the whole plane into different regions, where in each region different treatment is applied.

This work demonstrates one approach of utilizing multiple data sources in gene clustering, and data of other distributions can also be incorporated into this framework by joining EM algorithm of that particular distribution in a similar way. So in a sense, the framework proposed in this paper is applicable to many problems and not limited to the particular problem considered here [8].

Although, sBGMM is tolerant to some mis-clustering information in the prior, its performance might not be improved or even dragged down if the prior is built under different criterion and totally irrelevant to the focused problem. Moreover, although sBGMM is extremely useful when only sparse prior is available, it might not be able to efficiently utilize the third data source whose information is complete (such as PPI data). Moreover, since PPI data is one direct

measure of the regulatory network and is commonly used in gene clustering for many applications, such as inferring gene functions [34] and discovering genes involved in a particular molecular pathway [28], it is important to develop a model that can make as efficient use of PPI data as possible. In the future, instead of utilizing PPI data as prior, we could model it as Bernoulli distribution and treat it as one component of the joint model-based clustering framework.

ACKNOWLEDGMENT

This work was supported by the Academy of Finland (application number 129657, Finnish Programme for Center of Excellence in Research 2006-2011). We would also like to thank the Tampere Graduate School in Information Science and Engineering (TISE) for its financial support in this project.

REFERENCES

[1] X. F. Dai, H. Lähdesmaki, and O. Yli-Harja, *sBGMM: a stratified Beta-Gaussian mixture model for clustering genes with multiple data sources*. International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies (BIOTECHNO 2008), Bucharest, Romania, 29 June - 5 July 2008, pp. 94-99.

[2] H. Akaike, *A new look at the statistical identification model*. IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, 1974.

[3] J. D. Banfield and A. E. Raftery, *Model-based Gaussian and non-Gaussian clustering*. Biometrics, vol. 49, no. 3, pp. 803-821, 1993.

[4] C. Biernacki and G. Govaert, *Choosing models in model-based clustering and discriminant analysis*. J. Statis. Comput. Simul., vol. 64, pp. 49-71, 1999.

[5] H. Bozdogan, *Model Selection and Akaike Information Criterion (AIC): The General Theory and its Analytic Extensions*. Psychometrika, vol. 52, pp. 345-370, 1987.

[6] X. F. Dai, T. Erkkilä, O. Yli-Harja, and H. Lähdesmaki, *A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data*. BMC Bioinformatics, accepted.

[7] X. F. Dai, H. Lähdesmaki, and O. Yli-Harja, *BGMM: a Beta-Gaussian mixture model for clustering genes with multiple data sources*. Fifth international workshop on computational system biology (WCSB 2008), Leipzig, Germany, 11 - 13 June 2008, pp. 25-28.

[8] X. F. Dai, O. Yli-Harja, and A. S. Ribeiro, *Determing noisy attractors of delayed stochastic Gene Regulatory Networks from multiple data sources*. submitted.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*. Proceedings of the National Academy of Sciences of the United States of America, vol. 95, pp. 14863-14868, 1998.

[10] C. Fraley, *Algorithms for model-based Gaussian hierarchical clustering*, SIAM Journal on Scientific Computing, vol. 20, no. 1, pp. 270-281, 1999.

[11] C. Fraley and A. E. Raftery, *Model-based clustering, discriminant analysis, and density estimation*, Journal of the American Statistical Association, vol. 97, no. 458, pp. 611-631, 2002.

[12] D. Ghosh and A. M. Chinnaiyan, *Mixture modeling of gene expression data from microarray experiments*. Bioinformatics, vol. 18, no. 2, pp. 275-286, 2002.

[13] G. D. Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, *DAVID: Database for Annotation, Visualization, and Integrated Discovery*, Genome Biology, vol. 4, no. 9, pp. R60, 2003.

[14] H. Lähdesmäki, A. G. Rust, and I. Shmulevich, *Probabilistic Inference of Transcription Factor Binding from Multiple Data Sources*, PLoS ONE, vol. 3, no. 3, pp. e1820, 2008.

[15] R. Herwig, A. J. Poustka, C. Muller, C. Bull, H. Lehrach, and J. O'Brien *Large-scale clustering of cDNA-fingerprinting data*, Genome Research, vol. 9, no. 11, pp. 1093-1105, 1999.

[16] D. X. Jiang, C. Tang, and A. D. Zhang, *Cluster analysis for gene expression data: a survey*, IEEE Transactions on knowledge and data engineering, vol. 16, no. 11, pp. 1370-1386, 2004.

[17] Y. Ji, C. Wu, P. Liu, J. Wang, R. K. Coombes, *Applications of beta-mixture models in bioinformatics*. Bioinformatics, vol. 21, no. 9, pp. 2118-2122, 2005.

[18] H. Li and F. Hong *Cluster-rasch models for microarray gene expression data*, Genome Biology, vol. 2, no. 21, pp. research0031.1-0031.13, 2001.

[19] M. J. Herrgard, B. Lee, V. Portnoy, and B. Palsson, *Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces*, Genome Research, vol. 16, pp. 627-635, 2006.

[20] G. Mclachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.

[21] G. Mclachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, Manhattan, USA, 2000.

[22] M. Meila and D. Heckerman, *An experimental comparison of model-based clustering methods*, Machine Learning, vol. 42, pp. 9-29, 2001.

[23] L. A. O'Neill, K. A. Fitzgerald, and A. G. Bowie, *The Toll-IL-1 receptor adaptor family grows to five members*, Trends in Immunology, vol. 24, no. 6, pp. 286-290, 2003.

[24] W. Pan, J. Z. Lin, and C. T Le, *Model-based cluster analysis of gene expression data*. Genome Biology, vol. 3, no. 2, pp. research0009.1-0009.8, 2002.

[25] W. Pan, *Incorporating gene functions as priors in model-based clustering of microarray gene expression data*. Bioinformatics, vol. 22, no. 7, pp. 795-801, 2006.

[26] S. A. Ramsey, S. L. Klemm, D. E. Zak, K. A. Kennedy, V. Thorsson, B. Li, M. Gilchrist, E. S. Gold, C. D. Johnson, V. Litvak, G. Navarro, J. C. Roach, C. M. Rosenberger, A. G. Rust, N. Yudkovsky, A. Aderem, and I. Shmulevich, *Uncovering a Macrophage Transcriptional Program by Integrating Evidence from Motif Scanning and Expression Dynamics*, PLoS Computational Biology, vol. 4, no. 2, pp. e1000021, 2008.

[27] G. Schwarz, *Estimating the dimension of a model*. Annals of Statistics, vol. 6, pp. 461-464, 1978.

[28] E. Segal, H. Wang, and D. Koller *Discovering molecular pathways from protein interaction and gene expression data*, Bioinformatics, vol. 19, no. 1, pp. i264-i272, 2003.

[29] G. Sherlock *Analysis of large-scale gene expression data*, Briefings in Bioinformatics, vol. 2, no. 4, pp. 350-362, 2001.

[30] P. Smyth, *Model selection for probabilistic clustering using cross-validated likelihood*. Statistics and Computing, vol. 9, pp. 63-72, 2000.

[31] M. Symons, *Clustering criteria and multivariate normal mixtures*, Biometrics, vol. 37, pp. 35-43, 1981.

[32] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation*, Proceedings of the National Academy of Sciences of the United States of America, vol. 96, pp. 2907-2912, 1999.

[33] A. Taylor and D. J. Higham, *Contest: A controllable test matrix toolbox for MATLAB*. Genome Research, vol. 16, pp. 627-635, 2007.

[34] K. Tu, H. Yu, and Y. X. Li, *Combing gene expression profiles and protein-protein interaction data to infer gene functions*, International Journal of Biotechnology, vol. 124, no. 3, pp. 475-485, 2006.

[35] N. Tuncbag, T. Haliloglu, O. Keskin, *Correspondence between function and interaction in protein interaction network of Saccaromyces cerevisiae*. International Journal of Biomedical Sciences, vol. 1, no. 1, pp. 1306-1216, 2006.

[36] S. Vaithyanathan and B. Dom, *Model-based hierarchical clustering*, Proceedings of the 16th conference on Uncertainty in Artificial Intelligence, Stanford, California, USA, 30 June - July 3, 2000, pp. 599-608.

[37] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, *Model-based clustering and data transformation for gene expression data*, Bioinformatics, vol. 17, no. 10, pp. 977-987, 2001.

Data set 1		cluster 1				cluster 2				cluster 3			
gB	α	20	5	3	30	20	25	30	35	2	15	33	4
	β	2	15	33	4	20	25	30	35	20	5	3	30
bB	α	33	30	22	20	30	27	20	18	27	24	18	16
	β	30	33	20	22	27	30	18	20	24	27	16	18
gG	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_m	μ	3	15	5	11	2	13	6	9	1	14	7	10
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_v	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	10	20	30	25	10	20	30	25	10	20	30	25
Data set 2		cluster 1				cluster 2				cluster 3			
gB	α	20	5	3	30	20	25	30	35	2	15	33	4
	β	2	15	33	4	20	25	30	35	20	5	3	30
bB	α	33	30	22	20	30	27	20	18	27	24	18	16
	β	30	33	20	22	27	30	18	20	24	27	16	18
gG	μ	10		1		-20		0		-10	8	5	15
	σ	1		2		3		2.5		1	2	3	2.5
bG_m	μ	2		13		6		9		1	14	7	10
	σ	1		2		3		2.5		1	2	3	2.5
bG_v	μ	10		1		-20		0		-10	8	5	15
	σ	10		20		30		25		10	20	30	25
Data set 3		cluster 1				cluster 2				cluster 3			
gB	α	20		5		3		30		2	15	33	4
	β	2		15		33		4		20	5	3	30
bB	α	30		27		20		18		27	24	18	16
	β	27		30		18		20		24	27	16	18
gG	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_m	μ	3	15	5	11	2	13	6	9	1	14	7	10
	σ	1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
bG_v	μ	5	-8	20	15	10	1	-20	0	-10	8	5	15
	σ	10	20	30	25	10	20	30	25	10	20	30	25

Note: 'gB' and 'bB' each stands for 'beta' distributed data that are of 'good' and 'bad' quality respectively; 'gG', 'bG_m' and 'bG_v' each represents 'Gaussian' distributed data that are of 'good' quality and 'bad' quality with respect to close means and large variances respectively; '||' separate the parameters of different clusters, and '|' separate the parameters of different dimensions (2nd dimension) within the same cluster.

Table I
PARAMETERS OF BETA AND GAUSSIAN DISTRIBUTED DATA.

		All		F		C		P		
Model	Criterion	M1	M2	M1	M2	M1	M2	M1	M2	N
BMM	1∼4	0.2487	0.2945	0.3546	0.3484	0.3579	0.3479	0.3498	0.3423	4
GMM	1∼2	0.1889	0.1756	0.2640	0.3149	0.2924	0.3451	0.2740	0.3334	13
	3∼4	0.2112	0.1914	0.2955	0.3027	0.3239	0.3587	0.3019	0.3356	29
BGMM	1∼4	0.1351	0.1350	0.2369	0.2681	0.2847	0.3117	0.2506	0.2976	4
sBGMM$_{P1_a}$	1∼4	**0.0848**	**0.0710**	0.2128	**0.2021**	0.2503	**0.2483**	0.2290	**0.2307**	4
sBGMM$_{P1_b}$	1∼4	0.0913	0.0747	**0.1947**	0.2174	**0.2272**	0.2684	**0.2110**	0.2409	4
sBGMM$_{P2_a}$	1∼4	0.1740	0.1840	0.2911	0.3279	0.3173	0.3337	0.2963	0.3225	16
sBGMM$_{P2_b}$	1∼4	0.1506	0.1291	0.3000	0.3536	0.3218	0.3700	0.3098	0.3638	9
sBGMM$_{P2_c}$	1∼4	0.1817	0.1785	0.2429	0.2926	0.2697	0.3083	0.2556	0.3040	8
P2$_a$		0.1948	0.1810	0.2610	0.2530	0.2833	0.3035	0.2649	0.2609	8
P2$_b$		0.2055	0.2167	0.2707	0.2736	0.2970	0.3074	0.2768	0.3043	29
P2$_c$		0.2216	0.2286	0.2726	0.2577	0.2999	0.2938	0.2815	0.2862	31

Note: 'F', 'C', 'P' represent the three aspects of gene ontology, and 'All' means all three aspects are included. 'M1' and 'M2' stand for the mean and median of the p-values across all the clusters, respectively. 'Model' and 'Criterion' represent the model and model selection criteria, respectively. Subindexes of sBGMM indicate the prior that is used, e.g. sBGMM$_{P1_a}$ stands for using prior 'P1$_a$'. '1' to '4' each represents model selection criterion BIC, ICL, AIC, AIC3 respectively. 'N' means the number of clusters generated by each model. The last three lines show the corresponding statistics for the clusters given by DAVID. The smallest p-value in each column is shown in bold face. All fractions are rounded to four decimal points.

Table IV
PERFORMANCE TEST RESULTS OF SBGMM WITH REAL DATA.

		All		F		C		P		
Moldel	Crit	M1	M2	M1	M2	M1	M2	M1	M2	N
sBGMM$_{normal}$	1∼4	0.1662	0.1002	0.2356	0.2776	0.2797	0.3222	0.2441	0.2976	13
sBGMM$_{core}$	1∼4	**0.0557**	**0.0308**	**0.1418**	**0.1492**	**0.1922**	**0.1595**	**0.1617**	**0.1551**	5
BGMM$_{core}$	1∼4	0.1279	0.0714	0.2259	0.2701	0.2682	0.2833	0.2373	0.2637	8

Note: 'F', 'C', 'P' represent the three aspects of gene ontology, and 'All' means all three aspects are included. 'M1' and 'M2' stand for the mean and median of the p-values across all the clusters, respectively. 'Model' and 'Crit' represent the model and model selection criteria, respectively. 'bef' and 'aft' in the subindexes of sBGMM represent that the clustering is done before and after knowing the core TF genes, respectively, and the last digit 'i' ($i \in 1, \ldots, 3$)in the subindex represents the 'i^{th}' repetition of clustering with this model. '1' to '4' each represents model selection criterion BIC, ICL, AIC, AIC3 respectively. 'N' means the number of clusters generated by each model. The smallest p-value in each column is shown in bold face. All fractions are rounded to four decimal points.

Table V
CLUSTERING RESULTS WITH WHOLE DATA SET.

TERAPERS - Intelligent Solution for Personalized Therapy of Speech Disorders

Mirela Danubianu; Stefan-Gheorghe Pentiuc; Ovidiu Andrei Schipor; Marian Nestor ; Ioan Ungureanu; Doina Maria Schipor

"Stefan cel Mare" University
13 Universitatii Street, Suceava
Romania

mdanub@eed.usv.ro; pentiuc@eed.usv.ro; schipor@eed.usv.ro; mnestor@stud.usv.ro; ioanu32@yahoo.com; vmdoina@yahoo.com

Abstract - **The aim of this paper is to describe an intelligent system designed for assisting the personalized therapy of dyslalia for the Romanian pre-scholars children. This system is developed in the framework of the TERAPERS project which includes informational technologies in response to society challenges for health early diagnosis and personalized therapy. The Romanian language is a phonetic one that has its own special linguistic particularities, there is a real need for the development and use of audio-video systems, which can be used in the therapy of different speech problems. The system has a high degree of originality because his objective is to treat the pronunciation disorders in the Romanian language. Furthermore, the complexity of the project results from the high number of different research areas involved: artificial intelligence (expert system), virtual reality, digital signal processing, digital electronic and psychology (assessment procedures and therapeutically guide).**

Keywords - ***intelligent system; expert system; mobile device; speech disorder; personalized therapy***

I. INTRODUCTION

Individuals with disabilities have become more prominent and with the advent of new information technologies that have been applied to the diagnosis and treatment of these individuals, the implementation of more efficient systems has been realized.

A speech disorder is a problem with fluency, voice, and/or how a person says speech sounds.

Classification of speech into normal or disorder is a complex one. Statistics points out that only 5% to 10% of the population has a completely normal manner of speaking, all others suffer from one disorder or another.

The most common speech disorders are: stuttering, cluttering, voice disorders, dysartria and speech sound disorders.

Dyslalia is articulation disorder that consists of difficulties with the way sounds are formed and strung together. These are usually characterized by omitting, distorting a sound or substituting one sound for another.

Dyslalia has the greatest frequency among handicaps of language for psychological normal subjects as well as for those with deficiencies of intellect and sensory. Thus, the opinion of Sheridan (1946) is that at the age of eight years dyslalia are in proportion of 15% for girls and in proportion of 16% for boys.

Speech disorder therapy should begin as soon as possible. Children enrolled in therapy early in their development (younger than 5 years) tend to have better outcomes than those who begin therapy later.

In the process of therapy, speech therapists use a variety of strategies including: oral motor or feeding therapy, articulation therapy and language intervention activities.

In the step of oral motor/feeding therapy the therapist will use a variety of oral exercises, including facial massage and various tongue, lip, and jaw exercises, to strengthen the muscles of the mouth. It is also important to work with different food textures and temperatures to increase a child's oral awareness during eating and swallowing.

The exercises used in order to get the correct articulation, or sound production, involve to have a correct therapist model of sounds and syllables for a child and it is often used during his play activities. Child's play level is in accordance with the child's specific needs. The therapist will physically show the child how to make certain sounds, such as the "r" sound, and may demonstrate how to move the tongue to produce specific sounds.

During the language intervention activities the therapist will interact with a child by playing and talking. He may use pictures, books, objects, or ongoing events to stimulate language development. The therapist may also model correct pronunciation and use repetition exercises to build speech and language skills.

In the area of speech disorder, there are some European projects developed as part of the European Union (EU) Quality of Life and Management of Living Resources program.

II. STATE OF THE ART

The priorities on the international level are represented by the developing information systems that permits personalized therapeutically pathways. The following main directions are considered: development of expert systems that personalize therapeutic guides to the child's evolution

and the evaluation of the motivation and progresses that the child's achieves.

The most important objective is the determination of methods for the evaluation of speech impairments [3] where the data set is based on children, aged between 2 and 2 years and 11 months old and have English language skills. To date, there are only a few articles about the results obtained on this subject, although research in this area is progressing.

The potential users of the system are children affected by speech impairments and logopaed professors (speech therapists).

The OLP (Ortho-Logo-Paedia) project [2] for speech therapy began in 2002. The EU finances this complex project. It involves the Institute for Language and Speech Processing in Athens and seven other partners from academia and the medical domains. The scope of this project aims to establish a three – module system (OPTACIA, GRIFOS and TELEMACHOS) capable of interactively instructing the children suffering from dysarthria (difficulty in articulating words due to disease of the central nervous system). The proposed interactive environment is a visual one and is adapted to the subjects' age (games, animations). The audio and video interface with the human subject will be the OPTACIA module, the GRIFOS module will make pronunciation recognition and the computer-aided instruction will be integrated in the third module – TELEMACHOS.

An interesting developing project is Speech Training, Assessment, and Remediation (STAR) [4], which began in 2002. STAR members -AI. duPont Hospital for Children and The University of Delaware- aim to build a system that would initially recognize phonemes and then sentences. This research group offers a voice generation system (ModelTalker) and other open source applications for audio processing.

On the international level, is Speechviewer III developed by IBM [5] that creates an interactive visual model of speech while users practice several speech aspects (e.g. the sound voice or special aspects from current speech).

The ICATIANI device developed by TLATOA Speech Processing Group, CENTIA Universidad de las Américas, Puebla Cholula, Pue, México uses sounds and graphics in order to ensure the practice of Spanish Mexican pronunciation [6].

A recent project Articulation Tutor (ARTUR) [11] goal was to obtain an integrated speech therapy system with an intuitive graphical interface named *Wizard-of-Oz* and a virtual speech tutor named ARTUR. Based on audio (user's utterance) and video (facial data) information, the system recognizes and reproduces mispronunciations. Then, ARTUR suggests the correct pronunciation (audio data) and the correct speech elements' position (virtual articulator model).

At the national level, little research has been conducted on the therapy of speech impairments. What has been funded has been focused on traditional areas such as voice recognition, voice synthesis and voice authentication. These studies were conducted in the Psychology and Education Science Department from "Al. I. Cuza" University of Iasi. These studies have lead to development of software for aided instruction that provides feedback regarding oral fluency. Although there are a lot of children with speech disorder, the methods used today in logopaedia are mostly based on individual work with each child. The few existing computer-assisted programs in Romania don't provide any feedback.

There are expensive ($500-$1,500 USD) software applications but are not appropriate for the phonetic specifics of the Romanian language, which has its own special linguistic particularities. Therefore, we considered the real need for the development and use of audio-video systems, which can be used in the therapy of different pronunciation problems.

III. SYSTEM OBJECTIVES

The information systems with real-time feedback that address pathological speech impairments are relatively the new due to the increasing amount of processing power they require [7]. The progress in computer science allows at the moment for the development of such a system with low risk factors. A child's pronunciation is also used to enrich the existing audio database and to improve the current diagnosis system's performances.

Our system has reached some specific objectives [1]:

- initial and during therapy evaluation of children and identification of a modality of standardizing their progresses and regresses (at the level of the physiological and behavioral parameters);
- rigorous formalization of an assessing methodology and development of a pertinent database in this area;
- development of an expert system for the personalized therapy of speech impairments that allows designing a training path for pronunciation, individualized according to the speech disorder category, previous experience and the child's therapy previous evolution;
- development of a therapeutic guide that allows mixing classical methods with the adjuvant procedures of the audio-visual system ; and
- design and the achievement of a database that contains the child's dates, the set of exercises and the results obtained by the child.

The high degree of complexity of the project is due to the high number of different research areas involved: artificial intelligence (learning expert systems, pattern recognition), virtual reality, digital signal processing, digital electronic (VLSI), computer architecture (System on Chip, embedded device) and psychology (evaluation procedures, therapeutic guide, experimental design for validation).

IV. SYSTEM ARCHITECTURE

Assisted therapy is based on the interactions between six functional blocks: child, speech therapist, lab monitor program, expert system, 3D model and the child monitor program. The system's information flow is presented in Fig 1.

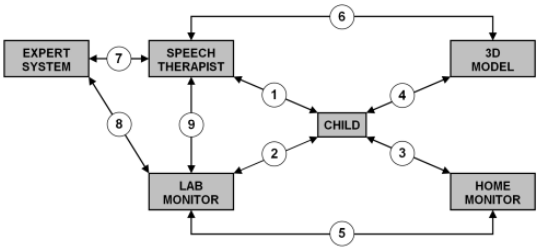


Figure 1. The system information flow

1. There is a close interpersonal relationship between the speech therapist and the child. All the other modules assist the teacher in his therapeutic action.
2. The monitor program allows the introduction of a complex examination's information and offers the possibility of making periodically records with the child's speech. The child receives an instant audio feedback and he can see the history of his audio recordings.
3. The role of home monitor program is to create a virtual interface between teacher and child (home speech therapy). This component is implemented both for personal computer (PC) and personal digital assistant (PDA). It can run exercises in a game manner, can offer feedback and can perform statistics base on current subject scores [18]
4. The 3D model provides viewing of the correct positioning of language, lips and teeth for each sound. The child may change the transparency of these items.
5. The monitor program performs homework transmission to the child PC or PDA. Later, when the child comes back, he can receive the activity report.
6. The professor will analyze the images offered by the 3D model and can correct some of the mistakes.
7. Expert system, if it is activated make suggestions regarding some training parameters like session frequency, length and content (exercises) according with some input variables. If the teacher observes erroneous conclusions, he can view the inferential route and can change the knowledge base.
8. The expert takes the data input from the monitor program and generates, upon request, sets of personalized exercises.
9. Monitor program is an interface between the speech therapist and other components like data base, expert system and child monitor program. At this level, speech therapist can collect both textual and audio information regarding each child, can administrate exercises and can manage all therapy aspects: selection of children, scheduling for therapy, offer all statistical reports that are required.

Fig. 2 presents therapeutic steps and necessary knowledge bases:

- dyslalia therapeutically guides
- speech therapy centers experience

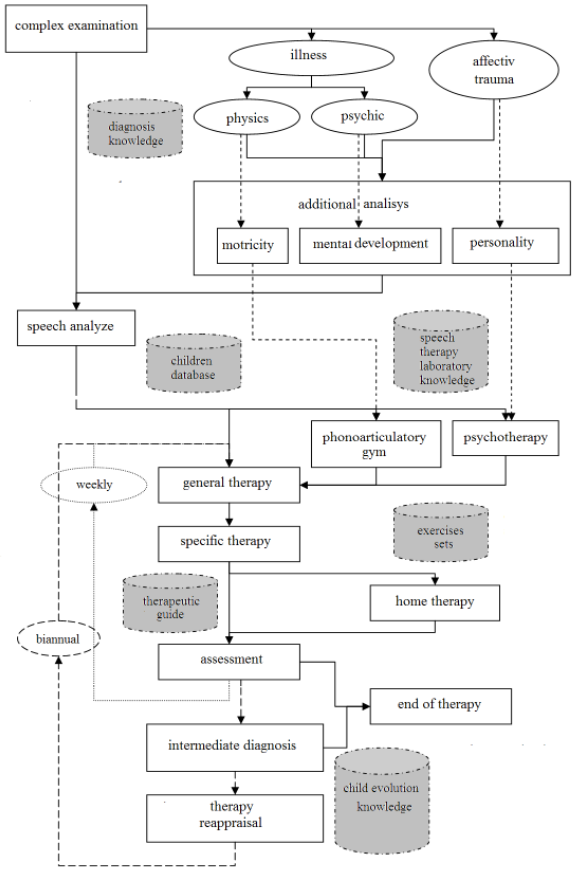


Figure 2. Therapeutically steps and knowledge bases

- dyslalia exercises sets
- historical data of therapy

According to Levitt [13], speech therapy software can help speech problem diagnostic, can offer real-time, audio-visual feedback, can improve analysis of a child's progress and can extend speech therapy at the child's home.

The architecture of TERAPERS - the personalized therapy system of dyslalia - is presented in Fig. 3. The system contains two main components: an intelligent system installed on each speech therapist's office computer and a mobile system used as a friend of child therapy. The two systems are connected. The intelligent system is the fix component of the system and it is installed on each speech therapist's office computer. This system includes the following parts:

- a child information management module,
- an expert system that will produce inferences based on the data presented by the evaluation module,
- a virtual module of the mouth, that allows the presentation of every hidden movement that occurs during speaking and
- an exercises management module that allows for the creation or modification of the exercises corresponding to various stages of therapy and grouping them in

complex issues.

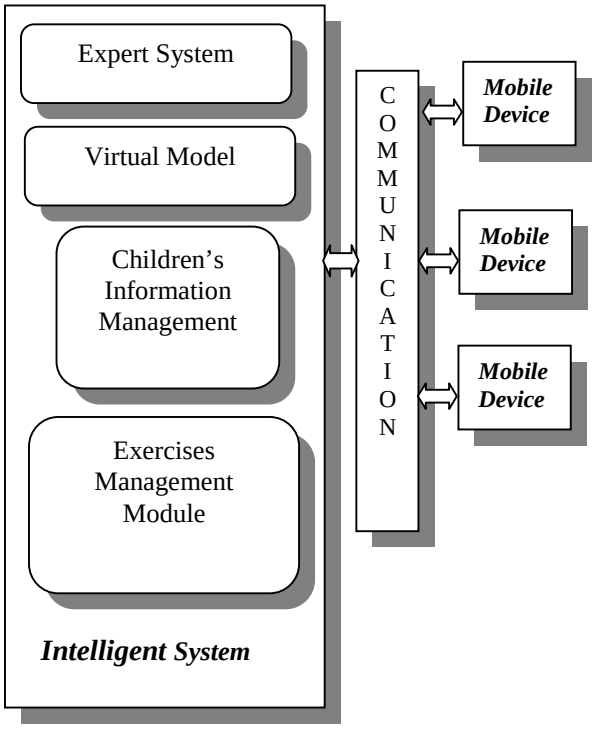


Figure 3. The architecture of TERAPERS

The mobile device of personalized therapy has two main objectives. It is used by the child in order to resolve all homework prescribed by the speech therapist and delivers to the intelligent system a personalized activity report of the child.

V. INTELLIGENT SYSTEM IMPLEMENTATION

In order to manage a child's logopaedic activity we have designed and implemented a complex software system named LOGOMON. The speech therapy teachers use this system for [10]:

- introduction and analysis of child's specific information (automatic obtain special reports);
- production of audio recordings with phonemes and scoring them (for each altered sounds);
- obtaining decision support from an integrated expert system;
- creation and evaluation of a large set of exercises for children; and
- performing homework transmission to the child's PC or PDA and receiving the activity report.

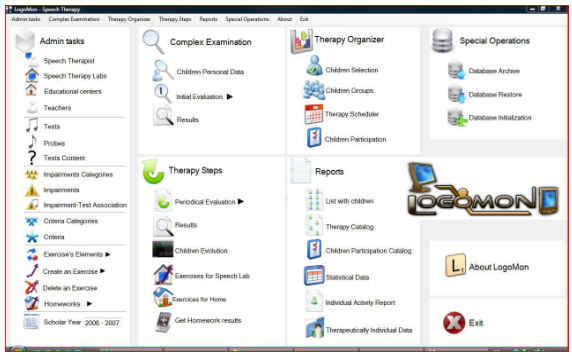


Figure 4. The LOGOMON Interface

To assess the way in which therapy evolves including the child's attendance at therapy sessions and the activity report downloaded from the mobile device. This activity report gives information regarding the exercises that have been done, how many times each exercise has been repeated, the time needed for each exercise and the results. Fig. 4 presents the interface of LOGOMON. This is a graphical user interface that allows the therapist to perform all these activities.

A. Children Information Management

LOGOMON is an intelligent system which assists the individually therapy of the speech disorders. It consists of software that performs the specific tasks of a monitor system. The aim of this system is: the initial evaluation for determining those children with speech disorders, registration of them in a database, suggestion of the diagnosis with the possibility for the expert to confirm or to modify this diagnosis, the selection of children for the therapy, the management of the therapy process and the supervision of the children's progress.

Moreover by using LOGOMON we collect and process the data such as we can remove any manual action of collecting or processing of data. Also we can eliminate the data storage on the paper.

The previous objectives are achieved using the following distinct activities:

- detection and recording data regarding children affected by dyslalia. Detection is made through an initial evaluation, based on tests applied to children. These tests are conducted in educational establishments (schools or kindergartens) by the speech therapists. It has five standardized tests for examination of pronunciation of sounds s, ş, ţ, ci, and r, each of them containing nine samples. Each sample consists of three trials, in sequence, the score obtained may vary between 0 points - when all the trials are wrong, and 3 points if all attempts are successful
- establish a presumptive diagnosis based on the sum of scores obtained as a result of the tests. A less than a minimum threshold, set by the specialist, show the existence of a deficiency.

- selecting the children that follow the speech therapy according to legal criteria. These criteria, specified in the order of priority, are: the type and the severity of the deficiency, the age of the child and the family implication. Each child is associated with a particular type of therapy (for groups of diagnosis G, or for individual activity A).
- design the personalized therapy according to the identified diagnosis. This involves specifying the stages of therapy and the choice of the right exercises for each stage of therapy.
- programming the therapy sessions and looking after the therapy progress.
- evaluation of the progress made by children and, if it is necessary, redesign of the intervention. This evaluation is based on the same tests that have enabled the detection of deficiencies, applied at the end of each phase of therapy or at fixed intervals. A comparison of test scores from the current to those achieved in previous tests can provide information regarding the evolution of a child. The result of this comparison may lead to one of the following conclusions about the current state of the child: corrected, ameliorated or stationary.
- collecting and recording all the data that allow preparation the logopaedic summary of each child. These data refer to anamnesis, complex logopaedic examination, diagnosis and various recommendations, the evolution during therapy or other final comments.

All the collected and processed data are stored in a relational database implemented in Oracle database management system.

B. Recording and Evaluation of Children Phonemes

An important part of our research refers to the automatic parsing of audio recordings. These recordings are obtained from children with dyslalia and are necessary for an accurate identification of speech problems. We have developed a software application that helps parse audio and real time recordings [10].

The main objective of this task is to record the children, using different audio environments during recording (some phonemes will be used for training a real-time recognition system). The speech therapist's voice must be ignored and after recording is necessary to split the stream into phonemes. The cost of recording devices and the children's impact must be minimized.

We utilize a digital voice recorder in high quality mode and with Variable Control Voice Actuator (VCVA) activated. The record format is IMA-ADPCM, 16 KHz and 4 bits (16 bits PCM). A microphone was placed at 10 cm from mouth in order to minimize environment noise.

A software set of classes (C#) was created for handling audio stream (read, conversion between different format, and write). We also have proposed an original solution for placing markers in audio stream. These markers are needed for correct parsing of full record.

C. Expert system

The expert system is based on a therapy guide, written in a natural language. This guide formalized in knowledge base consists of [19]:

- the muscular of phonon-articulator system development methods (e.g. setting up exercises for cheeks, lips and tong);
- the rhythm of respiration controlling methods (e.g. supervised inspiration and expiration from the temporal and intensity standpoint);
- the phonomatic hear development methods (e.g. the onomatopoeic pronunciation, rhythmic pronunciation exercises, distinguish along the paronyms);
- the method for the sound consolidation (e.g. the pronunciation sound of direct, inverse and complex syllable, of words, of paronyms, etc); and
- the sound's utilization in complex contexts (e.g. sentence, short stories, poems, riddles).

In this project we have used a fuzzy expert system for therapy of dyslalic children. With fuzzy approach we can create a better model for speech therapist decisions. A software interface was developed for validation of the system.

The main objectives of this task are:

- personalized therapy (the therapy must be appropriate for the child's problems level, context and possibilities);
- speech therapist assistant (the expert system offer some suggestion regarding what exercises are better for a specific moment and from a specific child);
- (self)teaching (when system's conclusion is different that speech therapist's conclusion the last one must have the knowledge base change possibility).

We use a rule-based expert system which has two major advantages: usually that kind of systems do not requires a large training set, and since the expert thinking is explicitly spelled out, we know how he thinks about the problem. Regarding that, it has the disadvantage that the knowledge acquisition phase may be difficult. A great advantage of fuzzy expert systems is that most rules can be written in language that the expert can directly understand, rather than in computer jargon; communication between domain expert and knowledge engineer is greatly eased. Another advantage of rule-based expert systems is the potential ability to learn by creation of new rules and addition of new data to the expert knowledge data base.

Fuzzy logic has ability to create accurate models of reality. It's not an "imprecise logic". It's a logic that can manipulate imprecise aspects of reality. Recently, many fuzzy expert systems were developed. [14][15]

In the next set of figures, we present an example of fuzzy inference. There are three input linguistic variables (speech problems level – Fig. 5, family implication – Fig. 6 and children age – Fig. 7) and one output linguistic variable (weekly session number – Fig. 8). We consider five fuzzy rules and, base on these rules, we illustrate specific fuzzy result (Fig. 9). If the system user wants a crisp value, defuzification is a good solution (Fig. 10).

To express a number in words, we need a way to translate input numbers into confidences in a fuzzy set of word descriptors, the process of *fuzzification*. In fuzzy math, that is done by *membership functions* [7].

Defuzzification is the reverse process of fuzzification. We have confidences in a fuzzy set of word descriptors, and we wish to convert these into a real number.

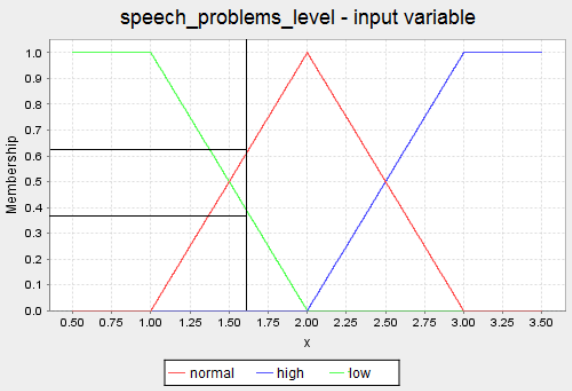


Figure 5. Speech_problems_level language variable

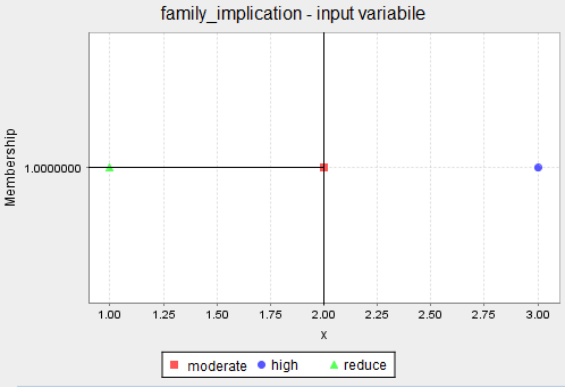


Figure 6. Family_implication language variable

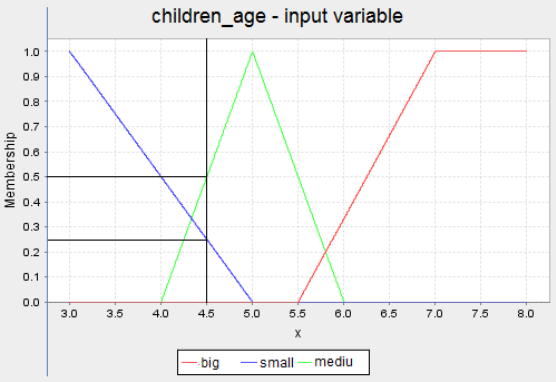


Figure 7. Children_age language variable

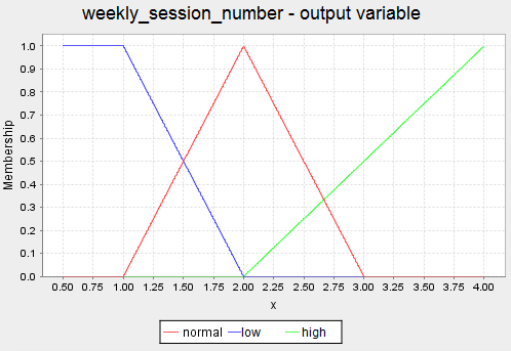


Figure 8. Weekly_session_number language variable

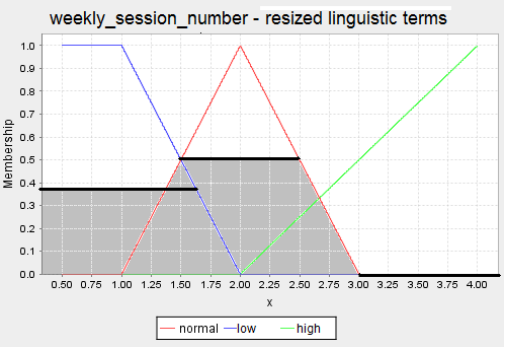


Figure 9. Obtain a fuzzy result

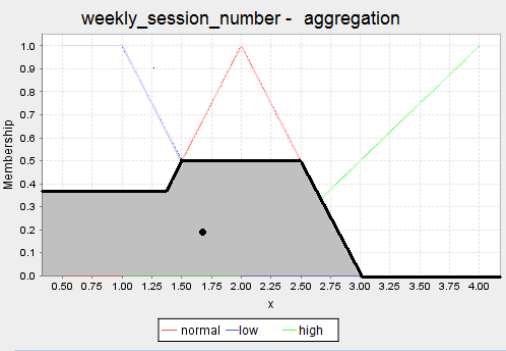


Figure 10. Obtain a crisp value

The first three variables have the following representation:

speech_problems_level (1.62)
= {“low”/0.37,”normal”/0.62,”high”/0.0}

family_implication (2.00)
= {“reduce”/0.0,”moderate”/1.0,”high”/0.0}

children_age (4.50)
= {“small”/0.25,”medium”/0.5,”big”/0.0}

We consider five rules for illustrate the inference steps:

- IF (speech_problems_level is high) and (child_age is medium) and (family_implication is reduce) THEN weekly_session_number is high;
  min (0.00, 0.50, 0.00) = **0.00** for linguistic term **high**

- IF (speech_problems_level is low) and (child_age is small) and (family_implication is moderate) THEN weekly_session_number is low;
  min (0.37, 0.25, 1.00) = **0.25** for linguistic term **low**

- IF (speech_problems_level is low) and (child_age is medium) and (family_implication is moderate) THEN weekly_session_number is low;
  min (0.37, 0.50, 1.00) = **0.37** for linguistic term **low**

- IF (speech_problems_level is normal) and (child_age is small) and (family_implication is moderate)THEN weekly_session_number is normal
  min (0.62, 0.25, 1.00) = **0.25** for linguistic term **normal**

- IF (speech_problems_level is normal) and (child_age is medium) and (family_implication is moderate)THEN weekly_session_number is normal
  min (0.62, 0.5, 1.00) = **0.50** for linguistic term **normal**

Final confidence coefficients levels are obtained using max function:

- **high** = max (0.00) = **0.00**
- **low** = max (0.25, 0.37) = **0.37**
- **normal** = max (0.25, 0.50) = **0.50**

Each linguistic term of output variable has another representation and in this manner is obtained as the final graphical representation of weekly_session_number variable. If the system user wants to get a single output value, then the area center of gravity is calculated. In our case (value 1.62), the child must participate in one to two session (but two is preferred).

We implement over 150 fuzzy rules for control various aspects of personalized therapy (19 variables presented in Fig. 11). These rules are currently validated by speech therapists and can be modified in a distributed manner.

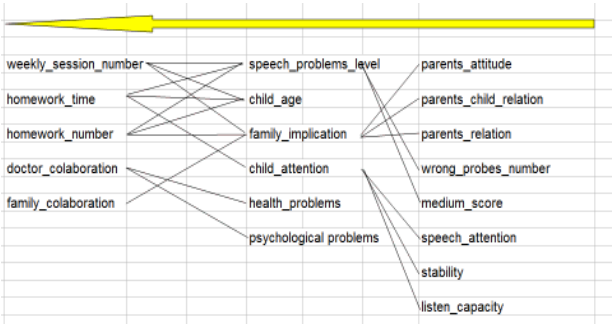


Figure 11. Fuzzy variables used for expert system

D. *Exercises generator*

All kind of exercises, part of different phases of speech therapy, are grouped in two main categories:

- general therapy (mobility development, air flow control, hear development);
- specific therapy (sound obtaining, consolidation and regular utilization).

Each therapy session contains a formative assessment and will be followed by home training. After three months, the speech therapist can finalize the treatment or can consider continuing it.

In order to help children with dyslalia we have created a consistent set of software exercises. This set has a unitary software block (data base, programming language, programming philosophy) and a big number of multimedia items for each Romanian language sound (over 5000 audio recordings and over 1000 image).

The speech therapist has the possibility to create and save exercises. They can also transmit these exercises to mobile device of children.

For example, the aim of phonematic hearing phase of therapy is to educate the ability to distinguish and differentiate sounds and words. As a result, appropriates exercises should allow including a) to identify words that contain certain phoneme, b) to identify the word that does not contain certain phoneme from a set of word, and c) to distinguish certain phoneme from pairs of paronyms (a word from pair contain certain phoneme and the other word contain a similar phoneme). This type of exercise can run in two ways: a words is represented by significant images, or the words are represented by flags, because there are many words in paronymic pairs which can not be associated with images (such as verbs, adjectives, adverbs, etc…).

Fig. 12 presents an example of an exercise which request to identify the words that contain the *r* sound.[16][17]

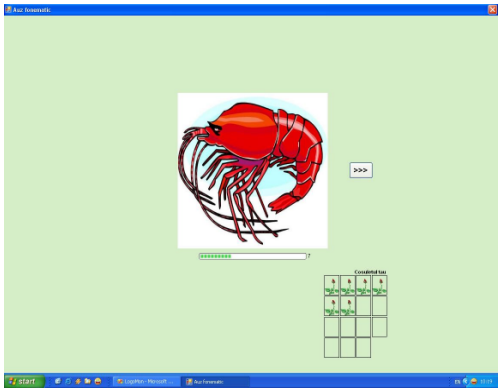


Figure 12. Example of an exercise interface

We observe that the interface is very simple. The child must only click on the image if the sound is present in the corresponding word. The feedback offered by the system is appropriate to the child's age. Every successful attempt is recompensed with a flower in the matrix from the right corner.

At the end of exercise some statistics are presented. These statistics refers: total number of words from the exercise, the number of correct answers and the number of wrong answers, the percentage of success (one to must be achieved) and the percentage obtained by the child.

Fig. 13 presents such a statistic.

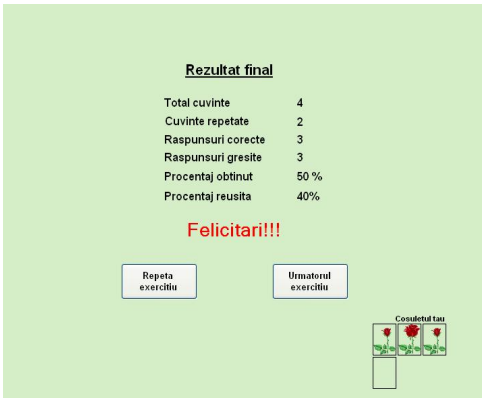


Figure 13. Final statistics of an exercise

VI. THE MOBILE DEVICE

The mobile device of personalized therapy system has two main objectives. It is used by the child in order to resolve the homework prescribed by the speech therapist and delivers to the intelligent system a personalized activity report of the child. The functions of the mobile device for assisted training are: presenting the exercises that the child should solve by himself, a personalized interaction with the human subject during therapy, the possibility of evaluating and encouraging the progresses obtained by the human subject through out an appropriate feedback, the capacity of collecting of the audio samples for learning and of the exercises solved by the child and the ability to communicate with the speech therapist's computer.

The device has two kinds of facilities: multimedia (to be able to record, to process and to play audio samples) and graphics (a friendly and accessible interface). Fig. 14 ilustrates the main page of the application implemented on the mobile device (a) and two types of exercises: (b) the child must detect if a sound is present inside a word (indicated by an image) and (c) where the child must select a word from a group of paronyms.

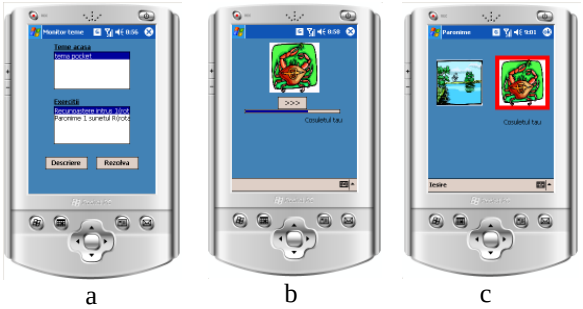


Figure 14. Three interfaces on the mobile device

VII. DISTRIBUTED FRAMEWORK

TERAPERS system was conceived as an assistant for the personalized therapy of speech disorders. In addition it provides a solid basis for improving the speech therapy at the national level through sharing experience of successful therapies from various speech therapists' offices and through collecting data that will lead to synthetic accurate information.

In Romania, the logopaedic assistance and therapy system has a hierarchical structure. Each speech therapist's office represents a node of a County Logopaedic Center network. At one superior level there is a National Center. In this context it is necessary to have access to data of each office in order to achieve a summary of the national condition. This situation refers to the number of children detected with speech impairments, grouped by age and diagnosis, the number of children included in the therapy program at the national level or synthetic results obtained through the application of different schemes of personalized therapy.

The fix part of the system is installed on each speech therapist's office computer. Because on each of these points create and maintain a database, it can be said that each of these local databases is a fragment of a distributed database and each speech therapist's office is a node in a distributed system. Communication between nodes can be done through an internal network, dedicated to this purpose, or using Internet technology.

The main advantage of this kind of system is that data are located near the place where the greatest demand is and that conduct faster data access, faster data processing and reduced operation costs.

Our project consists of a homogeneous distributed system, because all the nodes have the same operating system and the same application LOGOMON, so there is a unique database management system. Since all local schemes are identical we have used a horizontal fragmentation, in which each site has complete data regarding the children treated in the office and the exercises created by the therapist.

Regarding data replication there are two aspects. Personal data of children are not replicated. They are found only on the node where are managed. Instead, all the tables containing data regarding the exercises and themes are replicated in all nodes. This type of replication is necessary not only for security but also because it is necessary to share the experience and results, in particular the success ones, with all in the system. Thus each specialist can easily see the full set of exercises.

VIII. SYSTEM VALIDATION

The therapeutic system including the expert system has have been validated from three distinct perspectives:

- *Theoretical validation of the knowledge base*. From theories and models offered by psychology and speech therapies on tried to build a coherent set of rules

- *Practical validation of the therapeutic system*. The TERAPERS system was tested by the Interschool Regional Logopaedic Center of Suceava.
- *Experimental validation of therapeutic and expert system*. We have implemented two experiments in order to validate the system.

The subjects were 40 children, boys and girls selected by the speech therapists from Interschool Regional Logopaedic Center of Suceava, age range 5 and 6 years, with difficulties in pronunciation of R and S sounds.

They were divided into equal groups a control group and a program group. Each child attended two meetings weekly and was rated weekly. The two groups were constructed so as to be equivalent in terms of characteristics and in terms of initial tests assessment scores. Session length and course was designed to be the same for both groups. All preparatory actions, including the exercises selection, were developed before the session began.

The first group used the classical method of therapy, where the speech therapist select the exercises, while for the second group used the TERAPERS system, particularly the exercises were selected by the expert system.

Because the small number of subjects in each group (under the limit of 30), scores' distribution was not in generally normal. That is why, statistic data were processed using nonparametric tests: ManWitney test for difference between groups and Wilcoxon for difference between pretest and posttest scores [20]. The experiment conducts to the following results: a) groups were parametrically and statistically equivalent (ManWitney, session 0); b) both groups have progressed (Wilcoxon); c) both groups have arrived at the same performance (ManWitney, session 24).

We have not achieved significant differences between the two groups at the end of the 24 meetings, so we may consider that the exercises' choice can be performed either by speech therapist or expert system. This may be explained by the fact that the expert system has been tested during a six month period. All this time speech therapist have compared its decision with those suggested by the expert system and has adjusted knowledge base.

During the experiment were observed certain advantages of expert system utilization: speech therapist has possibility to be more concentrate on therapy because he don't spent time creating exercises (average time is 7 minutes per session) and rigor and predictability.

IX. CONCLUSION AND FUTURE WORK

In order to improve speech therapy activity, we have developed an integrated system that can be used for assisting the personalized therapy of dyslalia for the Romanian pre-scholars children. This system is actually tested by Interschool Regional Logopaedic Center of Suceava.

We estimate that this research will stimulate the preoccupation for the classical therapy efficiency compared with the therapy assisted by computer, at national and also European level. The after results will be compared with the similar smart systems from other countries, which use a similar phonetic language to the Romanian language.

The use of informatics applications in order to assist and to track the speech therapy has provided huge volumes of data. This has been possible due to the development of database technologies and the development of media storage, which have the capacity to keep an impressive amount of data.

However increased volume of available data does not lead immediately to a similar volume of information to support the decisions regarding therapy because classical methods of data processing are not applicable. For this reason we think that data mining, as techniques for automate detecting of relevant patterns in databases may helps therapists to build personalized therapy programs by identifying and anticipating the needs and the evolution of different types of patients. Consequently we intend to extend the capabilities of TERAPERS system by using some adequate data mining techniques.

ACKNOWLEDGMENT

The financial support was granted by the National Agency for Scientific Research, contract ref. no. 56-CEEX-II03/27.07.2006.

REFERENCES

[1] M. Danubianu, St.Gh. Pentiuc, O. Schipor, I. Ungureanu, M. Nestor, Distributed Intelligent System for Personalized Therapy of Speech Disorders, in Proc. of The Third International Multi-Conference on Computing in the Global Information Technology, ICCGI 2008, July 27- August 01, Athens, Greece.

[2] OLP (Ortho-Logo-Paedia) – Project for Speech Therapy (http://www.xanthi.ilsp.gr/olp);

[3] Diagnostic evaluation of Articulation and Phonology (http://www.harcourtuk.com);

[4] STAR Speech Training, Assessment, and Remediation (http://www.asel.udel.edu/speech);

[5] Speechviewer III – (http://www.synapseadaptive.com/edmark/prod/sv3)

[6] R. Laboissière, D. J. Ostry, and A. G. Feldman, The control of multi-muscle systems: Human jaw and hyoid movements. Biological Cybernetics, 74, pp. 373-384, 1996.

[7] C. Chivu, Applications of Speech Recognition for Romanian Language, Advances in Electrical and Computer Engineering, Suceava, Romania, ISSN 1582-7445, No 1/2007, volume 7 (14), pp. 29.

[8] L. Mititiuc, Psychotherapeutic problems of children with speech impairments, Ankarom, Iasi, Romania, 1999.

[9] E. Vrasmas, Speech problems therapy, Bucuresti, Romania.

[10] M. Danubianu, S. G. Pentiuc, O. Schipor, C. Belciug, I. Ungureanu, M. Nestor, Distributed system for therapy of dyslalia, Distributed Systems, "Stefan cel Mare" University of Suceava Press, Romania, Sept. 2007.

[11] O. Schipor, M. Nestor, Automat parsing of audio recordings. Testing children with dyslalia. Theoretical background., Distributed Systems, "Stefan cel Mare" University of Suceava Press, Romania, Sept. 2007.

[12] http://www.speech.kth.se/multimodal/ARTUR/index.html

[13] H.Levitt. The impact of technology on speech rehabilitation. In: Proceedings of an ESCA Workshop on Speech and Language Technology for Disabled Persons, Stockholm, Sweden, 1993.

[14] V. Mukhin, E. Pavlenko, Adaptative Networks Safety Control, Advances in Electrical and Computer Engineering, Suceava, Romania, ISSN 1582-7445, No 1/2007, volume 7 (14), pp. 54-58.

[15] R. Fuller, Carlsson, Fuzzy Reasoning in Decision Making and Optimization. Physica-Verlag, Heidelberg, Germany , 2002.

[16] C. E. Belciug, O. A Schipor, M. Danubianu, Exercises For Cildren With Dyslalia - Software Infrastructure, In Proceedings of The Distributed Systems Seminar, Suceava, Romania, 2007.

[17] S. G. Pentiuc, F. Giza Blciug, D. M. Schipor, Phonematic Exercises For Cildren With Dyslalia. -Software Application**,** In Proceedings of The Distributed Systems Seminar, Suceava, Romania, 2007.

[18] M. Cerlinca, A. Graur, S. G. Pentiuc,T. I. Cerlinca, Developing a Logopaedic Mobile Device Using a FPGA. In: Proceedings of SACI '07 - 4th International Symposium on Applied Computational Intelligence and Informatics, Romania, 2007, pp. 89-92.

[19] S. G. Pentiuc, O. A. Schipor, M. Danubianu, M. D. Schipor, Therapy of Dyslalia Affecting Pre-Scholars. In: Proceedings of Third European Conference on the Use of Modern Communication Technologies - ECUMICT, Gent, Belgium, 2008

[20] O. A. Schipor, S. G. Pentiuc, M. D. Schipor Improving computer based speech therapy using a fuzzy expert system, Computing and Informatics, Vol. 22, 2003, pp.1001–1016

Porphyrin – Cis-Platin drug system for HeLa cells photodynamic treatment

Rodica-Mariana ION[a,b,*], Luciana MARESCA[c], Danilo MIGONI[d], Francesco P. FANIZZI[d]

ICECHIM, Bucharest, Romania; Valahia University from Targoviste, Romania
e-mail: rodica_ion2000@yahoo.co.uk
Bari University, Dept of Chemical Pharmaceutic, Bari, Italy
e-mail: maresca@farmchim.uniba.it
Universita di Lecce, Dipartimento di Scienze e Tecnologie Biologiche ed Ambientali, Lecce, Italy
e-mail: fp.fanizzi@unile.it

Keywords: Photodynamic therapy, HeLa cells, singlet oxygen, TSPP, Cis-Pt

Abstract

Photodynamic therapy (PDT) is a clinical approach that use light-activated drugs for the treatment of different kind of tumor tissues. HeLa tumor cells, as an experimental model for study the new biomedical concept of associated PDT with 5,10,15,20-sulphonato-phenyl-porphyrin (TSPP) and an antitumor agent – cisplatin (CisPt), yielded to an enhanced effect on the suppression of tumor cells in vitro. The ultrastructure changes of cells caused by the action of the new concept and laser irradiation were analyzed, putting into evidence a linear survival curves, by the linear quadratic model. The ultrastructural morphologic aspect of the HeLa cells phototreated with TSPP reinforces the results that a great number of cells with plasma membrane damage, characteristics of irreversible cellular lesions were observed.

1. Introduction

Photodynamic therapy (PDT) is a novel treatment for cancer and certain non-cancerous diseases that are generally characterized by overgrowth of unwanted or abnormal cells [1]. By using a combination of a photosensitizer and a light source, the procedure require exposure of cells or tissues to a photosensitizing drug followed by irradiation with light of the appropriate wavelength, compatible with the absorption spectrum of the drug [1,2]. In PDT, photosensitizers are used to absorb energy from a light source after its administration to tumour cells, producing reactive oxygen species that will cause cell death.

Intense research has been devoted to understanding the molecular processes involved in apoptotic cell death and deciding about the strategies that can restore the apoptotic potential to tumor cells. PDT with most of the sensitizers tested, acts via singlet oxygen production. Singlet oxygen has been indicated as the active PDT agent causing injury to cells and tissues. Because of the short half live of this excited species in cells (<0.1 ms) and short radius of action (<0.02 mm), damage will occur mainly next to the region the sensitizer is concentrated. Incubation time of cells with photosensitizer, is another parameter that will affect the mode of cell death during exposure to light. For short period of incubation, the plasma membrane is an important site of damage [3]. Prolonged incubation with Photofrin®, the first PDT photosensitizer to win approval by regulatory agencies in several countries, tends to localize in the mitochondria membrane.

Many of the sensitizers used in the experimental or clinical PDT, localize in the plasma membrane, mitochondria, endoplasm reticulum and lysosomes [4]. In photodynamic therapy, the cellular photo-modification consists of: photosensitizer transport to the active site; possible binding or aggregation; light absorption by the site; production of energetic intermediate states; reaction with cellular bio-molecules; modifications of cellular function.

Proteins, lipids and nucleic acids are the most vulnerable cellular targets which can support their photo-oxidation.

As documented in the literature, many efforts have been made to discover new sensitizers with high singlet oxygen yield and high photodynamic activity. Among different molecules, porphyrins can be considered ideal sensitizers because of their optimal photodynamic activity [5]. Cisplatin (*cis*-diamminedichloroplatinum; *cis*Pt) is a potent inducer of growth arrest and/or apoptosis in most cell types and is among the most effective and widely used chemotherapeutic agents

employed for treatment of human cancers. Cisplatin as one of the most widely used metal-containing anticancer drug, is one of the most effective agents used to treat various types of human cancer (bladder, testicular, ovarian, and head and neck tumors) and, but its clinical effectiveness has been limited by significant undesiderable side effects, such as dose-dependent nephrotoxicity and neurotoxicity. Moreover, the use of cisplatin is limited by lack of activity against tumors with neutral or acquired resistance to this drug. Pt(II) compounds have also an extensive history exhibiting virucidal activity, including a recent report of anti HIV-1 activity [6]. In this context there is a clear evidence that the carrier ligand influences the antiviral activity and modifications of the carrier ligand in cisplatin may broaden the range of antitumor and antiviral activities, therefore, there is still a need to synthesize platinum(II) complexes with novel ligands and to test them for antitumor activity, in the hope of overcoming the above mentioned limitations. There is an urgent need for the development of new anticancer drug candidates to replace CisPt causes apoptosis by DNA fragmentation. The toxicity of platinum anticancer drugs presents a major obstacle in the effective treatment of tumours. Much of the toxicity stems from a lack of specificity of the drugs for the sites at which they are able to exert maximum anticancer activity. An improved understanding of the behaviour of the drugs in the tumour environment may assist in the rational design of future platinum anticancer agents with enhanced specificity and reduced toxicity.

Platinum complexes currently make up one of the three most widely used groups of anticancer drugs in the world. The anticancer activity of cisplatin (cis-[PtCl2(NH3)2], see Figure 1) was discovered serendipitously in the 1960s. Since 1978 it has been used in the clinic against a variety of cancers, including testicular, ovarian, head and neck, bladder, cervical, lymphoma and melanoma. Treatment with cisplatin often causes severe side effects such as nausea, vomiting, nephrotoxicity, neurotoxicity, myelotoxicity, and emetogenesis. These side effects arise mainly as a result of the limited selectivity of cisplatin for tumour cells as compared to healthy cells, and may also be due to reactions with thiol-containing species in blood plasma, such as cysteine and human serum albumin. In spite of its widespread clinical use, many tumours are unresponsive to cisplatin treatment due to intrinsic (eg. colon cancer, non-small-cell lung cancer) or acquired resistance (eg. Ovarian cancer, small-cell lung cancer). The cellular mechanisms of cisplatin resistance have been identified and recently reviewed. The main factors that modulate resistance include decreased drug accumulation, increased levels of intracellular thiols that can deactivate cisplatin, and an increased capability of cells to repair or tolerate DNA damage caused by cisplatin. Other processes have also been implicated.

Porphyrins, however, have limited photo stability, so their association with other drugs, in order to increase both light stability and photodynamic efficiency, is strongly recommended [7-10].

Photosensitizer as TSPP (TSPP = 5,10,15,20–tetra-sulphonated-phenylporphyrin) are able to decrease the fraction of single stranded circular genomic DNA by converting it to linear form [11]. Porphyrin and/or metalloporphyrin mediated cleavage of nucleic acids occurs via oxidative attack on the sugar moiety with consequent nucleobase modifications leading to strand scission; or by a photo induced mechanism involving either the porphyrin excited state or singlet oxygen.

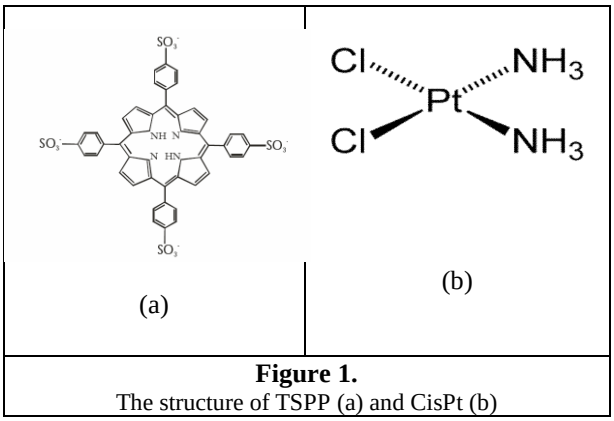
The photosensitized tumor therapy, carried on in the presence of certain chemical agents, can lead to enhanced tumor and tumor cell response. In this context, the combination between a porphyrin, TSPP, and a drug, cisplatin (CisPt), appears to be a useful and promising system, which has been studied, and it will be described hereafter.

In this work, HeLa tumor cells were used as an experimental model for studying the action of different new sensitizers in photodynamic therapy (PDT). Tumor monolayer cultures were incubated for 18 h at 37° C with TSPP and CisPt, and observed before and after photodynamic treatment with He-Ne laser. The association of a sensitizer with an antitumor agent, such as cisplatin, has an enhanced effect on the suppression of tumor cells in vitro which is more pronounced than that caused by individual components. Also, the ultrastructure changes caused by the action of two photosensitizers and laser irradiation on HeLa (neoplastic) cells were analyzed by transmission electron microscopy. The results showed induction of apoptosis.

2. Experimental part

2.1. Photosensitizers

TSPP (Figure 1a) has been prepared and purrified according to a published procedure [12].



(a) (b)

Figure 1.
The structure of TSPP (a) and CisPt (b)

Cisplatin (CisPt) (Figure 1b) has been used as pharmaceutical product without any further purification; the optimal concentrations were between 20-50 ng (105 cells) incubated for 18 h at 37° C.

2.2. Cell culture

Human HeLa cervical adenocarcinoma cells (ATCC CCL-2) were cultured at 37 °C in a humidified sterile atmosphere of 95% air and 5% CO2 at 37 °C, using DMEM supplemented with fetal calf serum (10%, v/v), glucose (4.5 g/l), L-glutamine

(292 mg/l), streptomycin sulfate (10 mg/l) and potassium penicillin (10000 U/l). HeLa cells were maintained frozen in DMEM with 10% DMSO. 1.8 ml CryoTubes™ (Nunc, Nalge Nunc International, IL, USA) were filled with the cellular suspension and then were placed in a cell Cryo 1 °C Freezing Container (Nalgene, Nalge Nunc International, IL, USA) to be slowly frozen up to -80 °C at a cooling rate of -1 °C/min for successful cell cryopreservation. Frozen cells were rapidly transferred to a liquid nitrogen container (-196 °C) and stored. HeLa cells are adherent cells which grow up to form cellular monolayers toward confluence after inoculation. Cell viability was evaluated with 0.2 % trypan blue solution. HeLa tumor cell cultures were treated with TSPP, CisPt and TSPP/CisPt.

2.3. Methods and apparatus

All spectroscopic measurements were carried out in 1-cm quartz cuvettes (Hellma, Germany) at room temperature and, in the case of cell suspensions, the samples were continuously stirred.

2.4. Cellular uptake

The cellular uptake of the photosensitizers was estimated by flow cytometry, using Foreskin cells cultured in 75 cm^2 tissue culture flasks as indicated above. Our compounds, dissolved in DMSO:water (0.05%-99.95%), were added to the flasks at a concentration of 1 μM when a confluence of 80–85% was reached and the cells were incubated with the sensitizer for different periods of time in the dark. Immediately after each time, the cells were washed with PBS to remove the non-entrapped sensitizer. Trypsinization was carried out using PBS containing 0.2% trypsin and 0.5 mM EDTA. The suspension of cells was centrifuged at 1500 rpm for 5 min. The pellet was suspended in PBS and prior to flow cytometric analysis, the new suspension was filtered through nylon filters (Nytal, 70 μm mesh, Sefar Maissa S.A., Barcelona, Spain) to exclude cellular aggregates.

2.5. Irradiation protocol

Cell suspensions were subjected to photodynamic therapy (PDT), namely pre-incubated for 1h-18h with the derivatives, and subjected to irradiation after washing them 3x in culture medium and resuspended at 3×10^5 cells/mL. The cell suspension irradiation was performed with a polychromatic lamp in a quartz cuvette, exposure time 30 minutes. Cellular photodegradation kinetics were recorded during irradiation with an UV-VIS double-beam spectrophotometer (Carl Zeiss Jena), connected to an external computer for data processing. Cell numbers were evaluated by means of maximum absorption at the band located at 275 nm, and fitted with a mathematical model. The absorption of UV in this region in a fixed volume of solubilized cells is proportional to the cell number, and therefore can be used as a simple means of obtaining a cell count. Cell counts obtained in this way can be combined with measurements of the inhibition of DNA synthesis ([3H]-thymidine incorporation) by test compounds, to produce an index of cytotoxicity [17]. A time dependency of log (No/N) was plotted, representing the decrease in the actual number of intact cells during the irradiation period, where No = initial number of cells and N = number of cells at a given time point. As cellular controls, we have used unloaded cells suspensions subjected to irradiation.

After irradiation, cell suspensions were washed twice for removal of cell debris generated during irradiation. After PDT, cell suspension was recorded in the

Counting Chamber (Roth) with 0.4% Trypan Blue Stain and the actual decrease in the number of cells subjected to PDT was presented.

2.5. Singlet oxygen test

Measurements were carried out in a quartz cell (1cm x 1cm) at 20°C. A DMSO solution (2.3 ml) containing sensitizer (4.8×10^{-5} M) and DPBF (2.7×10^{-5} M) was irradiated with light beam from a UV-Vis spectrophotometer. Solutions of sensitizers were freshly prepared and kept in the dark before measurements. The decreasing of the DPBF concentration was followed by a special programm ruled on a computer at the absorbance from 415 nm (the molar coefficient of absorption for DPBF is 23300 $M^{-1}cm^{-1}$) as function of the irradiation time (irradiation cycles 50 x 25 s). The reaction showed a zero order-kinetics in the first 100 s. The incident photon flow was 4.65×10^{-9} $M.s^{-1}$. Using the absorption spectra of the photosensitizer, the absorbed photon flow (I_{abs}) was evaluated. The quantum yield of the photooxidation of DPBF was calculated from the eq.l

$$\Phi_{DPBF} = ([DPBF]/I\ abs.V);\ V= 3\ cm^3 \quad (1)$$

The quantum yield for singlet oxygen generation was calculated from eq.2.

$$1/\Phi_{DPBF}=1/\Phi_{1O2} + (1/\Phi_{1O2}\ K_d/K_a)(1/[DPBF]) \quad (2)$$

From the intercept of the Stern-Volmer plots,we obtained the quantum yield for singlet oxygen generation.

2.6. Cell phototoxicity after irradiation

The cells were plated at a number of 106 cells/ml in each well of a 24 wells plate (Nunc, Denmark) as follows: six wells for light and photosensitizer and, six wells for light only (control). After 24 h of culture, cells were incubated with 10 μM of TSPP or Cis Pt in culture medium without serum for 60 min. Cells were washed twice with phosphate buffered saline (PBS) and 200 μl of fresh PBS was added for irradiation. Dark barriers were placed between wells to avoid scattered light during irradiation. Another dark barrier with an orifice of the diameter of the wall was placed on the top of the 24 wells plate for the same purpose. The irradiation was done in the dark with a He-Ne laser light exposure (Jena, 632.8 ms, power 50 mW) was performed for different times at 37° C in bottles with quartz windows, the cell suspensions being gently stirred during irradiation.

After He- Ne laser treatment, cell suspension were 1:10 diluted with growth medium ($1x10^5$ cells/ ml final concentration). The cellular suspensions were incubated for 24 h at 37° C in 5% CO_2 atmosphere. After irradiation, PBS was removed and culture medium with 10% FBS was added to the cells for 24, 48 and 72 h of culture in a humidified 5% CO_2 at 37ºC. After each period of incubation, the number of living cells was counted by the Trypan blue exclusion test.

2.7. Proliferation assay

The growth inhibitory effect of the studied compounds (TSPP, CisPt, or a combination of the two) towards HeLa cells was evaluated by using the spectrophotometric MTT assay [13]. The cells, grown in the culture flasks, were trypsinized and seeded at a density of 10^6 cells/well in well plates. DMEM growth medium (with 10% FBS) was used, and the cells were incubated overnight at 37°C in humidified environment containing 5% CO_2 to allow adherence. The tested compounds were diluted in FBS-free growth medium and then administered in growing doses (1, 10, 100 and 200 μM); their cytotoxic effect was evaluated using different periods of incubation (24, 48 and 72 hours). The MTT solution (5 mg/ml) was added to each well (10 μl) at the end of the incubation time with inhibitors and further incubated for 4 h at 37°C. After this period, the medium was aspirated and the purple crystals of formazan, which had been formed, were dissolved by addition of isopropanol/HCl 0.04 N (100 μl/well). Absorbance at 550 nm was measured on a spectrophotometer. The absorbance of the treated cells was then given as percentage of that of control cells and the resulting data are based on the mean value of 8 wells ± S. D.

2.8. Kinetic model

Irradiation of cells with light radiation produces linear survival curves [14]. The relationship between the surviving fraction S and the light dose D is then:

$$S = \exp(-\alpha D) \quad (3)$$

where:

S is the number of surviving cells;

$-\alpha$ is the slope

D is the radiation dose delivered.

The relationship is more commonly represented as:

$$S= \log N/No = \exp(-D/Do) \quad (4)$$

by defining Do as $1/\alpha$. When D=Do, $S = e^{-1} = 0.37$.

2.9. Electron microscopy

After irradiation, PBS was replaced by medium with serum. After 24 h of culture, cells were washed twice with PBS and fixed with 2.5% glutaraldehyde and 4% freshly prepared formaldehyde in phosphate buffer 0.1M (pH 7.2) for at least 2 h at 4ºC. Cells were detached from the dish with a cell scraper, centrifuged three times (1500g, 10 min) with 0.1M fresh phosphate buffer and post fixed in 1% osmium tetroxide in phosphate buffer for 30 min. Finally, cells were wahsed again and dehydrated in acetone and embedded in Epon. After sectioning, cells were contrasted with uranyl acetate for 30 min and lead citrate for five min. Transmission electron microscopy was performed using a Zeiss 900 and a Zeiss EM10 microscope.

2.10. Fluorescence microscopy

It was performed in order to investigate the cellular loading efficiency. Thus, in an Nikon E300 inverted fluorescence microscope with image capture, HeLa cells loaded 24h with non-toxic concentration of the investigated compounds were investigated with the V2A filter, excitation 380-420nm, emission < 450nm. After loading, cells were washed and in RPMI1640 medium without phenol red resuspended at $5x10^6$ cells/mL concentration. Live cell suspensions were laid on fluorescence slides (Marienfeld). Image was captured both in fluorescence and in phase contrast.

2.11. Confocal fluorescence microscopy

HeLa cells were viewed using a BIO RAD Radiance Plus Confocal Microscope. Images were obtained using a 100 x oil immersion objective (Nikon). The compounds CisPt, TSPP and CisPt-TSPP were excited at 543 nm, and detected at λ_{em} = 555 – 626 nm. Stains were excited using λ_{ex} = 488 nm, and detected at λem = 500-560 nm. HeLa cells were viewed using a Zeiss LSM510 confocal laser scanning microscope (Carl Zeiss, Welwyn Garden City, UK). Images were obtained using a 63x objective oil immersion objective. The compounds were excited at 543 nm, 76% laser strength, and detected at λ_{em} >560 nm. Stains were excited using λ_{ex} = 488 nm, 10-15% laser strength, and detected at λ_{em} = 505-530 nm.

2.11. Caspase-3 activity

It was measured from cell lysates by a colorimetric method (CaspACE™ Assay System, Promega Corporation), using its high sensitive substrate coupled to a cromophore (p-nitroaniline, pNA). The free pNA released from substrate upon enzyme action, is spectrophotometrically detected at 405 nm. Using a calibration curve and the actual proteic concentration in cell lysates (Bradford method) the specific caspase activity was calculated to the total proteic content of each lysate and normalized to $1x10^5$ cells. Results are presented as microMpNA/$1x10^5$ cells. Cell cultures were tested in different stages of the cell cycle, different batches of the same cell lines or primary cell cultures in order to have the statistical significant basal caspase 3 activity related to the cell type.

2.12. Apoptosis

It was evaluated by flow cytometry, using the annexin V / propidium iodide method (BD Biosciences kit). Annexin V (Ann) highlights apoptosis-associated loss of plasma membrane asymmetry, namely phosphatidylserine (PS) translocation from the inner to the outer leaflet of the plasma membrane, which is recognized by annexin V. Propidium iodide (PI) is used as vital stain. Annexin V staining identifies early apoptosis events, while cells double positive (Ann+PI+) are in late apoptosis or already dead. Positive controls were camptothecin 1microM (apoptosis inducer) treated cells. Samples were analyzed by flow cytometry (FACScalibur cytometer, Becton Dickinson) within one hour, using CellQuest software.

3. Results and discussion

Exposition of HeLa tumor cells to a combination TSPP with cisplatin showed a significant synergistic effect, which improves the efficacy of the antitumor drug, Figure 2.

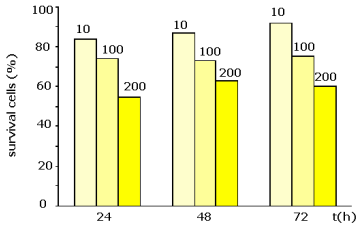


Figure 2.
Survival rate of HeLa in the presence of TSPP/CisPt

In the treatment with TSPP, at 10mM concentration of the drug, the surviving cell population was reduced to 60% after 24 h, and to less than 20% after 48 h of incubation respectively (the effect being more important for higher cisplatin concentration), Figure 3.

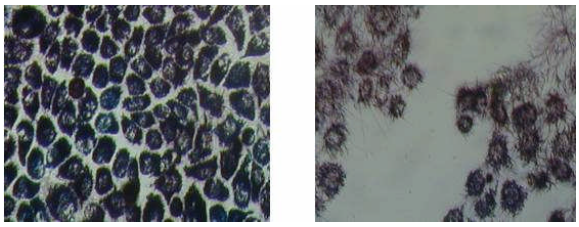


Figure 3.
MTT test for HeLa cells (control)-left and incubated with TSPP-CisPt 4 mM (right)

HeLa tumor cells were exposed to a photosensitizer (TSPP) alone or in association with an antitumor agent (cisplatin). When HeLa cells, in a stationary development stage, were treated with TSPP in concentrations ranging between 1 and 200 μM (or 5 and 500 ng/10^5 cells), the surviving rates were almost 100% after 24 h, and not less than 90% after 48 h of incubation (Figure. 4).

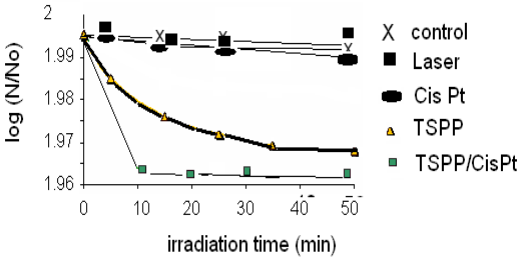


Figure 4.
Number of living cells (HeLa) after photodynamic treatment with TSPP, CisPt, TSPP/CisPt at different times.

On irradiation of HeLa tumor cells only with He-Ne laser, or Cis Pt not significant cytotoxic effect had been observed.

HeLa tumor cells were then irradiated wit He-Ne laser before and after treatment with TSPP, CisPt, or a combination of the two, and SEM examination of treated and untreated cells was also performed.

It is known that laser radiation can stimulate cell proliferation, a mechanism dependent on the fluency applied. In this work, the fluency used (0.5 J/cm^2) has an inhibitory effect of He-Ne laser (632.8 nm) in the region of 300 to 600 mJ/cm^2 in HeLa cells. It is widely accepted that non-ionized species can cross the plasma membrane more easily than charged compounds [15]. TSPP that is soluble in water is not capable of penetrating into the cell by means of passive diffusion. It is endocytosed by the cell and therefore localized in endosomes and lysosomes. Many authors consider that the cell uptake of the sensitizer is more efficient for lipophilic photosensitizers due to the better penetration through the cell membranes.

The ultrastructural morphologic aspect of the HeLa cells phototreated with TSPP reinforces the results that a great number of cells with plasma membrane damage, characteristics of irreversible cellular lesions were observed. These are characteristics of the cell necrosis process. Despite the knowledge that TSPP localizes preferentially in the mitochondrial membrane and that after laser irradiation activation of proteins leads to the apoptotic process, the initial localization of the photosensitizer in the plasma membrane and the interaction time in our experiments suggest a process of cell death with necrosis like characteristics.

PDT of HeLa cells with TSPP, shows ultrastructure features that suggest apoptotic cell death. A characteristic of this kind of death is the conservation of membrane integrity. According to Zhang [16], the relation between the mode of cell death (apoptosis or necrosis) and the dose of PDT, may be dependent mainly on the cell line and the photosensitizer used. Our results demonstrate cell apoptosis occurring in the neoplastic cell line. This may be related to the fact that lesion to the lysosomes membrane are followed by enzyme leakage to the cytoplasm. The activation of these enzymes causes enzymatic digestion of cellular components evidenced by nuclear alterations like picnotic nuclei and the characteristic ladder pattern of DNA fragmentation. TPPS as hydrophilic sulfonated porphyrin is taken up into lysosomes by endocytosis. PDT with either of the membrane- localizing photosensitizers resulted in increasing numbers of cells becoming apoptotic (TUNEL positive) during the first 12 h, but apoptotic bodies were not observed. In contrast, after photoactivation of the lysosome localized photosensitizers, apoptotic cells were not detected until after 12 h but extensive fragmentation of the cells into apoptotic bodies was found. These data provide evidence for at least two distinct pathways by which PDT can induce apoptosis. HeLa cells present a specificed morphological aspect, where only mitochondrial alteration is observed. However, maintaining the cell membrane integrity as a whole. The oxygen free radicals partially reduced are highly toxic molecules that cause lesion to cell membranes and other cell constituents [17]. Mitochondria and lysosomes have been identified as key components in the induction of apoptosis [18, 19].

SEM scan of untreated HeLa tumor cells, cultured 72 h at 37° C, showed a continuous, uniformly distributed monolayer without any cellular alteration (Figure 5). After laser irradiation the following features could be observed: (i) cyto plasmatic extensions branching out towards the periphery and contacting the neighboring preelongations and membranes, (ii) swellings on the cells surface, and finally (iii) some breaks (ruptures) at the cells periphery. Some cytoplasmatic prelongations appear, some of them having the tendency of branching with others cells.

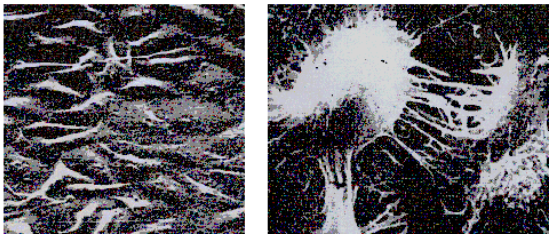


Figure 5.
SEM scan of untreated HeLa tumor cells, cultured 72 h at 37° C (left) and laser irradiated (right)

a b c

Figure 6.
HeLa tumoral cells after PDT treatment for control cells (a), TSPP treated (b) and TSPP/ CisPt (c)

The main explanation resides in the fact that cisplatin induces apoptosis in HeLa tumor cells implying attack on mitochondria (Fig. 6).

With TSPP, under He-Ne laser irradiation, presence of microvilli with or without microswellings, flattening tumoral cells, smooth surface without microvilli but with swellings on their surface, small craters in the central zone resulted from cytoplasmatic swellings, cytoplasmatic prelongations some of them having the tendency of branching with others cells.

The interpretation of the shape of the cell survival curve is still debated, as is the best way to fit the types of data mathematically.

For a linear kinetic curve (single hit, single target) the parameter Do can then be used to characterize the sensitivity in the linear region of the curve. Extrapolation of the terminal straight line portion of the curve back to the abscissa defines a value, n, the extrapolation number.

In the shoulder region of the curve the proportion of the cells killed, two interpretations are possible:

• Cell death results from the accumulation of events that are individually incapable of killing the cell, but which become lethal when added together (target models).

• Lesions are individually repairable but become irreparable and kill the cell if the efficiency of the enzymatic repair mechanisms diminishes with number of lesions and therefore the dose (repair models).

In this case, linear-quadratic model is the most used model for HeLa cells, from the following point of view:

• Do, the initial slope, due to single event killing, the dose to reduce survival to 37%, valuable for all the case except TSPP/CisPt;

• D, the final slope, interpreted as multiple-event killing, the dose to reduce survival by 67% from any point on the linear portion of the curve, which could apply for TSPP/CisPt.

The linear quadratic model assumes that a cell can be killed in two ways.

• Single lethal event

• Accumulation of sub lethal events.

From kinetic point of view, the survival fraction is evaluated as a plot of log ($N(t)/N_0$) versus irradiation time. In this case, could be applied the Theory of Dual Radiation Action, where the lesions responsible for cell photo destruction result from the interaction of sub lesions, resulted from unrepaired DNA double-strand breaks.

In the presence of the combined drugs TSPP-CisPt and under He-Ne laser irradiation could be observed flattened and covered cells with microswellings, micro craters at the cell periphery, lysis of the swellings and cells with smooth aspects.

Photodynamic therapy (PDT) is a standard treatment for various cancers (lung, esophagus, stomach, cervix, bladder, etc.) as well as for non-malignant conditions such as age-related macular degeneration, actinic keratoses and psoriasis. It is based on the selective retention of a previously administered nontoxic photosensitizer in the target cells, and irradiation of these cells with visible light at the appropriate wavelength (1). Upon illumination, the photosensitizer generates reactive oxygen species (singlet oxygen and free radicals, such as OH^-, HO_2^- and $.O_2^-$; Figure **7**).

These reactive species ultimately eliminate highly proliferating cells by damaging membranes, DNA and other cell structures, and also by affecting extracellular matrix (ECM) components.

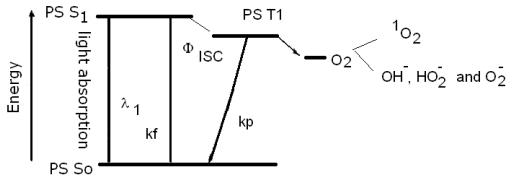


Figure 7.
Photosensitization process represented by a modified Jablonski diagram. PS S_0 = singlet ground state photosensitizer; PS S_1 = short-lived singlet excited state photosensitizer; PS T_1 = long-lived triplet state photosensitizer; $h\nu_f$ = fluorescence; $h\nu_p$ = phosphorescence; Φ_{isc} = intersystem crossing; $^1O_2$ = singlet oxygen.

Cell membranes have been identified as an important intracellular target, and many of their natural constituent macromolecules are readily susceptible to the organism, reacting with the singlet oxygen produced during the photochemical pathway, typically present in the PDT process. Such membranes include the plasma membrane surrounding the cell, the membranes of the endoplasmic reticulum distributed throughout the cytoplasm and the membranes of mitochondria and Golgi apparatus.

Cytoplasmatic residues resulted from cells desintegration as a consequence of the combined treatment. Scanning examination of He La tumor cells, exposed to photodynamic treatment (TSPP) associated with CisPt put into evidence various morphological lesions such as: microswellings, microcraters at the cell periphery, lysis of the swelling.

Apoptosis, also known as 'programmed cell death' or 'cellular suicide', is an active form of death with particular changes in cell morphology and protein activity. It is characterized by cell shrinking, surface membrane blebbing, chromatin condensation and DNA fragmentation. Apoptosis can be initiated in various manners, including PDT, and the common effector mechanism is to induce caspase-mediated cleavage of substrates. Initiator caspases are responsible for the first proteolytic events, e.g. cleavage of the cytoskeleton and related proteins including actin, and fodrin (a membrane-associated cytoskeletal protein). Amongst others, these early apoptotic events are thought to be responsible for the characteristic cell surface blebbing.

Three principal mechanisms are suggested for PDT action: cellular damage of targeting (photodamage by involving the process of apoptosis), vascular damage and immunological response.

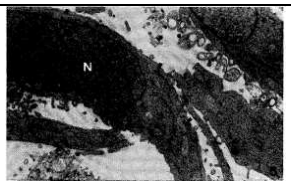


Figure 8.
Control HeLa cells. (N= nuclear chromatin; arrows = mitochondria).

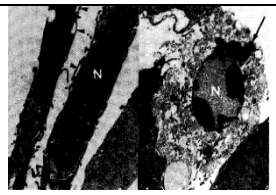


Figure 9.
HeLa cells 24 h after photodynamic treatment with TSPP-CisPt.

By transmission electron microscope structural alteration it was a found that characterize the apoptotic death mechanism for example, in the organization of the citoplasmatic membrane (figures 8,9), the condensed chromatin and aggregated chromatin in the nuclear peripheral, the condensation and presence of apoptotic bodies Our results indicate that the morphological criteria (apoptotic cell rounding and shrinkage) allow distinguishing the apoptosis as cell death mechanisms of photodynamic treatment. Thus, the morphological analysis under light microscopy and transmission electron microscope constitutes a very important and even decisive tool to identify the specific type of cell death unambiguously.

After photodynamic treatment with TSPP-CisPt, HeLa cells presented condensed chromatin, cell elongation, and cytoplasm condensation. All these are proves for apoptosis. Aspects of apoptotic bodies generated into cells are visualized in Figures 10,11.

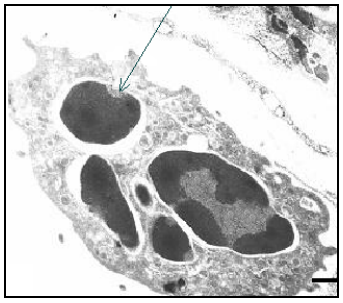


Figure 10. Apoptotic bodies inside HeLa cells (arrowhead).

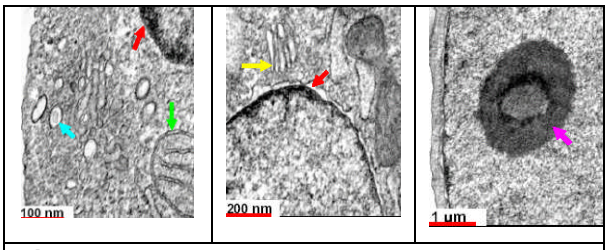


Figure 11. TEM micrographs of control HeLa cells

Figure 12 show that the CisPt-TSPP probe was confined to the nuclear compartment. It should be noted that shorter (<1 hour) and longer incubation times (>24 hours) revealed similar distribution patterns. Hence the above compound do not localise in the nuclei of the cells.

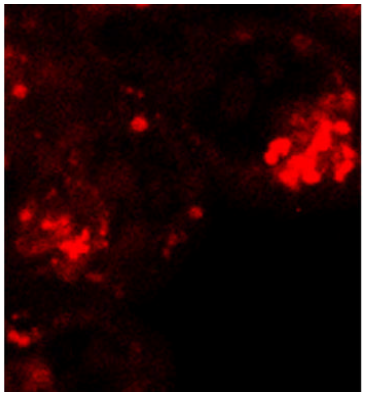


Figure 12.
Confocal image of CisPt-TSPP in HeLa cells

4. Conclusion

Photodynamic therapy (PDT) is a clinical approach that uses light-activated drugs for the treatment of different kind of tumor tissues. HeLa tumor cells, as an experimental model for study the new biomedical concept of associated PDT with 5,10,15,20-sulphonato-phenyl-porphyrin (TSPP) and an antitumor agent – cisplatin (CisPt), yielded to an enhanced effect on the suppression of tumor cells in vitro. The ultrastructure changes of cells caused by the action of the new concept and laser irradiation were analyzed, putting into evidence a linear survival curves, by the linear quadratic model.

In the presence of the combined drugs TSPP-CisPt and under He-Ne laser irradiation could be observed flattened and covered cells with microswellings, microcraters at the cell periphery, lysis of the swellings and cells with smooth aspects. Cytoplasmatic residues resulted from cells desintegration as a consequence of the combined treatment. Scanning examination of He La tumor cells, exposed to photodynamic treatment (TSPP) associated with CisPt put into evidence various morphological lesions such as: microswellings, microcraters at the cell periphery, lysis of the swelling.

5. References

[1] R.M.Ion, L. Maresca, D. Migoni, and F.P. Fanizzi, "Concept of associated photodynamic therapy with porphyrin – cis-platin drug system and applications on HeLa cells", *Biocomputation, Bioinformatics, and Biomedical Technologies, 2008, BIOTECHNO 2008*, Int. Conf. on Volume, 2008, pp. 70–75.

[2] K.R. Weishaupt, C.J.Gomer and T.J. Dougherty. "Identification of singlet oxygen as the cytotoxic agent in photoinactivation of a murine tumor", *Cancer Res.*, Vol.36, 1976, pp. 2326-2329.

[3] T.J. Dougherty, C.J. Gomer, B.W. Henderson, G. Jori, D. Kessel, M. Korbelik, J. Moan and Q. Peng, "Photodynamic Therapy: Review". *J Natl Cancer Inst.*, Vol.90, 1998, pp. 889-902.

[4] J. Moan and K. Berg, "The photodegradation of porphyrins in cells can be used to estimate the lifetime of singlet oxygen", *Photochem Photobiol.*, Vol.53, 1991, pp. 549-553.

[5] T. Christensen, L. Smedshammer, A.Wahl and J. Moan, "Photodynamic effects and hyperthermia in vitro", *Adv Exp Med Biol.*, Vol.193, 1985, pp. 69-78.

[6] A.N. Vzorov, D. Bhattacharyya, L.G. Marzilli, and R.W. Compans, "Prevention of HIV-1 infection by platinum triazines", *Antiviral Res.*, Vol. 65, 2005, pp. 57–67.

[7] N.L. Oleinick, "Apoptosis in response to photodynamic therapy". *Photodynamics News.*, Vol. 6, 1998, pp. 8-9.

[8] M.L. Agarwal, M.E. Clay, E.J. Harvey, H.H. Evan, A.R. Antunez and N.L. Oleinick, "Photodynamic therapy induces rapid cell death by apoptosis in L5178Y mouse lymphoma cells", *Cancer Res.*Vol. 5, 1991, pp. 5993-5996.

[9] K.Berg, Q. Peng, J.M. Nesland and J. Moan , "Cellular responses to photodynamic therapy", *Proc. SPIE.* Vol.2078, 1995, pp. 278-285.

[10] R.M. Ion, D.V.Brezoi, M.Neagu, G.Manda, C.Constantin, "Laser effect in photodynamic therapy of tumors", *Proc.SPIE*, Vol. 6606, 2007, pp. 66061G-660671G.

[11] R.Alexandrova, O.Sabotoniv, E.Stoykova, RM Ion, S.Shurilinkov, G.Minchev, "In vitro cytotoxicity assessment of TSPP on animal tumor and non-tumor cell lines", *Proc. SPIE*, Vol. 5449, 2004, pp. 227-232.

[12] R.Alexandrova, R.M. Ion, E.Stoykova, S.Shurilinkov, "In vitro investigation on cytotoxic activity of TS_4PP [5,10,15,20-tetra (4-sulfophenyl) porphyrin]", *Proc. SPIE*, Vol. 5226, 2003, pp. 423-427.

[13] T.R. Mosman and T.A.T. Fong, "Specific assay for cytokine production by T cells", *J Immunol Methods*, Vol.116, 1989, pp. 151–158.

[14] R.Oliver and B.J.Shepstone, "Some Practical Considerations in Determining the Parameters for Multi-

target and Multi-hit Survival Curves", *Phys. Med. Biol.* Vol.9, 1964, pp. 167-175.

[15] M.Neagu, G.Manda, C.Constantin, E.Radu, RM Ion, "Synthetic porphyrins in experimental photodynamic therapy induce a different antitumoral effect", *J.Porphyrins and Phthalocyanines*, Vol.11(01), 2007, pp. 58-65.

[16] W.W. Zhang, L.X. Zhang, R.K. Busch, J. Farrés and H Busch, "Purification and characterization of a DNA-binding heterodimer of 52 and 100 kDa from HeLa cells", *Biochem. J.*, Vol. 290, 1993, pp. 267-272.

[17] R.Alexandrova, E.Stoykova, RM Ion, E.Nedkova, K. Zdravkov, G.Minchev, "In vitro cytotoxicity assessment of phthalocyanines on virus-transformed animal cells", *Proc. SPIE*, Vol. 5830, 2005, pp.404-408.

[18] J. Piette, C. Volanti, A. Vantieghem, J.-Y. Matroule,Y. Habraken and P. Agostinis, "Cell death and growth arrest in response to photodynamic therapy with membrane-bound photosensitizers", *Biochemical Pharmacology***,** Vol.66, 2003, pp. 1651-1659.

[19] P. Castano, T. N. Demidova and M.R. Hamblin, "Mechanisms in photodynamic therapy: part two—cellular signaling, cell metabolism and modes of cell death", *Photodiagnosis and Photodynamic Therapy*, Vol.2, 2005, pp. 1-23.

Health Information Exchange and Care Integration

Kari Harno[a], Pekka Ruotsalainen[b], Pirkko Nykänen[c], Kyösti Kopra[d]

[a] *Helsinki University Central Hospital, Department of Medicine, Helsinki, Finland*
[b] *National Institute for Health and Welfare, Helsinki, Finland*
[c] *University of Tampere, Department of Computer Sciences , Finland*
[d] *Hospital District of Helsinki and Uusimaa, ICT and Medical Engineering, Finland*

Abstract

The Finnish health care system is a mixture of public and private services. Its governance and funding is highly decentralized. Contrary to governmental policies and expectations, autonomous decisions by local authorities and service providers associated with lack of information models and common technical standards have led to a broad spectrum of one-off ICT –systems with little technical and semantic interoperability.

National strategies to overcome these challenges have prompted initiatives to create sharable electronic health records (EHR) by supporting a collection of federated interoperable repositories with regional middleware services. In the Hospital District of Helsinki and Uusimaa an established regional *e*Health network (RHIN) connects 24 public hospitals, 29 municipal health centers (primary care) and two private health care clinics. There are 7.500 end-users, information from 1,4 million citizens and 40 million links to EPRs.

Migration to a national eHealth network (NHIN) providing a platform for delivery of a longitudinal view on patient's relevant health records (summary care record) will assist care integration between providers and improve the safety and quality of healthcare.

Keywords:
Electronic patient records, interoperability, health information exchange, shared EHR, regional and national information networks, core data set, summary care record

Introduction

Parts of this paper have been presented at the Second International Conference on the Digital Society and published in the proceedings of this conference [1]. In this paper policies and decisions aiming to improve coordination of care and healthcare information exchange (HIE) are first described, followed by evidence of benefits from two controlled studies.

Despite the adoption of appropriate and timely sharing of patient data within the same care settings across legally distinct organizations (RHIN) policies were newly revised upgrading regional HIE to a national infrastructure and this migration process and context are described.

Policies for coordinating care and improving health information exchange

Finland's progress in the realm of information society thus far and its opportunities for future success could be rated as mediocre or good. Finland has managed to remain in the running, but has not achieved a significant competitive advantage. The ingredients for success are there, and our capacity to exploit them will be determined over the course of the next ten to fifteen years.

A number of recommendations by the OECD for incremental changes to certain structural features of the Finnish health system underline the transition phase of the system and the continuing need for building on the strengths of the system in the future [2].

In the healthcare domain 96 % of primary care health centres, 95 % of hospital districts and 89 % of private service providers use an electronic patient record (EPR). So far, the uptake of ICT in the field of social welfare and health care has not yielded sufficient gains at the national level [3]. The level of investment in ICT has been relatively low in this sector – approximately two per cent of all annual health care expenditure.

The government led project to restructure municipalities and services aims to diminish the waste of decentralised governance and merge municipalities to form health care areas with populations 20.000 – 30.000. These mergers are supported by a national development plan, Kaste Programme, that defines the development objectives for social and healthcare services in the next few years and the main measures for achieving them. By the end of 2008, the number of municipalities has decreased by 16 % since the legislation was approved in 2007, but it is too early to distinguish effects on healthcare services or organizational developments.

Local experimentation and reform in the Finnish health system have so far involved one or more of three types of

organisational development: regional cooperation among municipalities; integration of primary and secondary care and outsourcing of services to alternative providers.

Regional cooperation has attracted most interest, because The Act on Experiments with Seamless Service Chains in 2000 has provided the legislative foundation for regional health information organizations. Initial investment funding for regional *e*Health networks was made available by The Ministry of Social Affairs and Health.

The previous National Coordinator for Health Information Technology in the USA, David Brailer argued that the acid test for regional health information organizations (RHIO) would be a governance model that brings in all stakeholders [4]. The Act on Experiments with Seamless Service Chains did not include directives, nor did the Ministry of Social Affairs and Health issue formal guidance for building the RHIO governing body responsible for its accountability, authority and oversight.

A formal application to the Ministry by joint authorities of hospital districts or municipalities was the prerequisite for acceptance of stakeholders to join. After preliminary consideration of applications three hospital districts (Hospital District of Helsinki and Uusimaa, Hospital District of Satakunta and Hospital District of Pirkanmaa) were approved in 2001 for the primary phase of the national experiment on regional cooperation through ICT.

No strategies existed previously for regional information management and from the year 2000 The Act on the Experiments on Seamless Service Chains opened a window of opportunity for improvement, a new paradigm for cooperation between organizations and transformation of healthcare. Similar solutions for cross-organizational healthcare information exchange (HIE) were pursued at the same time in other countries or regions [5-7].

A few small rural regions have advanced even further in cooperation since 2001 and integrated specialized and primary health care services into a regional service system, which applies one electronic patient record system. This integrated model removes administrative boundaries and incorporates seamless processes between the providers. These regional health service districts cover basic secondary care.

Clinically integrated systems have been indicated to be the next step in health reform [8]. The increasing prevalence of chronic disease, and the need to improve the quality of specialist services like diabetes care, will require not only closer collaboration between providers but also clinical integration between primary and secondary care and the development of clinical networks.

Hospital District of Helsinki and Uusimaa

The Hospital District of Helsinki and Uusimaa (HUS) comprises 24 hospitals in the province of Uusimaa, which includes the Finish capital of Helsinki, in Southern Finland. In order to organise the provision of specialised medical care, Finland is subdivided into 20 hospital districts. The HUS district is the largest of these. As a joint authority it was founded in 2000 to provide health services for the 1,434,513 residents in its 30 member municipalities.

Among the HUS hospitals Helsinki University Central Hospital is nationally responsible for treating special, severe and rare illnesses. In 2006, more than 430,000 different people were treated in the HUS, roughly 310,000 of them outpatients.

The record locator service is a central reference database (register) containing links to patient data stored in their legacy systems [9]. The upgrading of the legacy systems is made possible by application integration across the extended regional infrastructure. Provider access is possible by web browsers and patient information includes (primary care/hospital outpatient) visits, critical data, images and reports, laboratory results, referrals and discharge letters. All data is sorted according to social security coding, which is standard procedure in Finland.

Standard API connections between primary care information systems and the reference data have been installed. The documents are produced in CDA R1/R2 format and messages transferred in a standard pattern (HL 7/XML).

The service was launched in 2003 and presently most (29/30) municipalities, as well as all hospitals (24), are connected to the RHIN and apply the record locator service for regional exchange of information. Currently 7.500 professional end-users view over 200.000 documents (x-rays and documents) monthly.

The process of information sharing and archiving of EPR documents will be enabled in the future by the centralized *e*Archive instead of the RHIN. The RHIN locator service may be transformed to a centralized indexing service and a shift in the administration of patients' consents for sharing information within the *e*Health Network is to be expected. The role of the RHIN may be seen in the future as an enabler in healthcare delivery of services by orchestrating processes and work flow.

Established and Perceived Benefits of RHIN

In the past, patient information located in different organizations was inaccessible online for the professional. The only way to access information was to order the papers by mail. In addition to the costs and long delivery time, the difficulty was to know which systems contained relevant information. The regional RHIN Navitas service has been designed to overcome these barriers and to enable a seamless, cross-organizational access to patient records in the HUS region.

The interest in interoperability has increased with the growing number of EPR installations and HIE needs and as a result of reports on the role of information in relation to medical errors, patient safety and healthcare quality. Interoperability has been defined as the ability of two or more systems or components to exchange information and to use and understand the information that has been exchanged [10].

When different levels of interoperability for healthcare information system applications are considered, preserving the content (syntactic) and meaning (semantic) of the exchanged data must be eluded. The evolution of interoperability in the healthcare domain has mostly developed through information integration across the entire healthcare chain based on standards (e.g. CDA) and messaging (e.g. HL7) supported often by vendor specific solutions.

Navitas is a regional service designed to overcome the organizational and interoperability barriers restricting the use of clinical information between secondary and primary health care. Navitas is provided as a fully hosted ASP (application service provision) service to HUS and the municipalities in the joint authority of the Hospital District by a consortium of three vendors. The system was originally developed as part of an EU funded InterCare project together with HUS and the participating companies. It has been originally in use since 2001, but the present version was implemented in 2003.

The core of the federated model allowing participants to view and share patient information is the Navitas record locator service. It is a service which maintains a regional directory of links pointing to patient and treatment information located in any of the connected health care information repository systems in the region: each participating organization has its own patient information system in addition to the 11 presently stand-alone patient information systems in HUS. HUS has also many other clinical information systems e.g. the laboratory system and HUSpacs, which have all been integrated to the link directory. The regional health information network architecture for healthcare information exchange in HUS is described in figure 1.

At the moment there are 15 different patient information systems in some 55 organizations connected to the locator service. Specific adapter software has been installed locally into each of the systems through which links are fed into the locator service. Links are HL7 (Health level 7)/ CDA (clinical document architecture) compliant messages containing the identification of a patient and a short description of the contents of the particular patient record. No actual records or documents are stored into the locator service directory.

Navitas has a regional user database and centralized authentication and authorization services; this enables the participating organizations to have complete control over their own users. The health care professionals can access Navitas from their personal workstations using a web browser. The data transfer is encrypted and only private, dedicated networks (VPN) are used to transmit the data. Viewing of the patient data through the links requires the patient's informed consent. When clicking a link, a window will open up to display the actual clinical information. The information is queried by the Navitas locator service from the patient information system itself. The view provided by the locator service is a read-only view, structured in a user-friendly and visual way.

The Navitas locator service is available today for all health care professionals in the Hospital District. The directory contains information from 1.4 million citizens. Currently there are over 40 million links in the database. The number of links has been minimized in order to make it easier for the professional to get a holistic view on the patient's medical history. In HUS, for example, several visits are grouped into one care period.

The Regional *e*Health Network Navitas is actively used by 7.500 professionals in different organizations. The monthly access number to the directory exceeds 200.000 and over 2 million annually. More than half of the quaries that result in access, stem from the need to view images and the rest from document retrieval.

Two controlled studies are described in brief to demonstrate how shared healthcare information has revised the current practice between primary and secondary care. They not only facilitate the delivery of improved services, but create information system benefits by improving the performance of healthcare processes.

Case 1 Information exchange with *e*Referrals.
The first case represents healthcare information exchange with an *e*Referral system integrating primary and secondary care physicians by allowing interactive *e*Consultations between healthcare professionals. The focus is not only on the 5 - 10 % of patients, who are generally referred by GPs for specialized care, but also on some 30 % of primary care patients, who represent actual cases where *e*Consultation referrals to the hospital specialist were deemed necessary by GPs.

We have set up a wide-area referral network between primary care and three university hospitals [9,11]. This network was initially launched in 1990. In the university hospitals all specialties are involved. In 2002 there were 67,000 e-referrals transferred between the Helsinki University Hospitals and primary care. The solutions extend from the initial VPN use (Vantaa) to EDIFACT standard (Espoo) and HL-7 (Helsinki). A transition to standardized HL7 messages utilizing C-way message transfer systems (HUSway) through a single Network Access Point (HUSnap) has been implemented. Over 100 000 *e*referral messages (40 % of total) were transferred between health care providers in 2005 and by the end of 2007 the number of *e*Referrals has increased to 200.000 (70 % of total).

The *e*Referral between primary and secondary care not only speeds up the transfer of the referral, but also improves the access to service by offering an option for interaction in the form of *e*Consultation between general practitioners and hospital specialists. By sharing information and knowledge remote *e*Consultations create a new working environment for integrated delivery of *e*Services between the health care providers. Interactive eConsultations enable supervised care leading to the reduction of outpatient first visits (-36 % for clinical visit intended referrals and - 50 % for total referrals) in the outpatient departments for internal medicine, i.e. more timely appointments and cost containment.

The implementation of the referral system increased the number of referrals from primary care. The total number of referrals to the outpatient clinic was 7,5 vs 2,8 referrals per 1,000 inhabitants over the age of 15 for *e*referrals and paper referrals. Despite the increase in the number of

*e*Referrals, the running costs of the outpatient department were 20 % lower than with the traditional process. The direct costs for applying the *e*Referral were only one seventh of the costs for traditional outpatient visits (32 € vs 211 €). The patients needed fewer repeat visits to the outpatient clinic after being first consulted through the *e*Referral.

From the patient's viewpoint *e*Consultations provide just-in-time expert opinions from hospital outpatient department specialists, make the expertise accessible more quickly than the traditional process and reduce the need for low value or unnecessary visits to the hospital. Eight out of ten patients preferred to continue receiving medical care in this way.

*e*Referrals and *e*Consultations particularly had an effect on the care of less urgent patients in contrast to urgent *e*Referrals. Only one out of ten patients with urgent (visit within one week) *e*Referrals to the outpatient clinic received *e*Consultations whereas over 50 % of less urgent patients were managed with consultation alone. This allows the urgent patients to have access and be examined at the outpatient clinic within the set target time range.

Case 2 Image exchange with RHIN.
The second case describes HIE benefits from transferring and viewing digital images remotely with the regional *e*Health network by primary care physicians and orthopaedic surgeons. This seamless radiological chain, besides creating a tool for remote radiological *e*Consultations, also discloses some difficulties in compliance with care pathway performance and the need for better integration of processes between primary and secondary care.

The picture archiving and communication system (PACS) project (HUSpacs) was initiated in the Hospital District of Helsinki and Uusimaa in 1998 and the first two hospitals became filmless in 1999. All hospitals in the catchment area of HUS became filmless by 2004. Installing one of the largest regional PAC systems in the Hospital District of Helsinki and Uusimaa produces roughly 1 million imaging examinations with 20 terabytes of storage capacity.

HUSpacs infrastructure includes local short-term archives and a centralized long-term or back-up archive, which are connected with a wide-area network (ATM). Different modalities are integrated by applying standards (HL7, Dicom), which allow the radiological information system (RIS) and hospital information system (HIS) to share patient information.

These archives also serve the regional *e*Health network in distributing images and sharing radiological reports between hospitals and primary care, supporting HIE between healthcare professionals in the care process. The framework for the assessment of the regional HUSpacs after re-engineering of hospital and external processes has been previously described [12].

The assessment was performed in two health centers. The municipality of Vantaa contracts all its images from the HUS hospital in Vantaa (Peijas Hospital). Tikkurila, which is a suburb of Vantaa with 45.000 inhabitants, has a secure connection (VPN) to the record locator service and HUSpacs archive, i.e. in the health center GPs may view the clinical and radiological information documented in Peijas. The health center of Kerava, a municipality with 30.000 inhabitants and approximately at similar distance from Peijas Hospital, but without the availability of digital image transfer, served as control and received the images from Peijas by mail delivery.

In 2003, the radiological HUSpacs chain between Peijas Hospital and the health centers was evaluated. Based on request, previous images of the same anatomical region were pre-fetched for the reporting radiologist, films were digitized and stored in the HUSpacs database. These images are also routed to the web-server in Vantaa. The radiologist dictates the report to the RIS, which sends an HL7 message containing the report to Tikkurila.

The primary aim was to assess the effect of the regional HUSpacs process (production, archiving, viewing of images, as well as remote consultations) on improved access to or quality of care. We also inquired about the satisfaction of personnel in these institutions to the regional HUSpacs application.

The setting of the trial was comparative and included patients with conservatively treated fractures of the ankle or wrist. They had to be first diagnosed in Peijas Hospital and later controlled by their general practitioners in the Tikkurila or Kerava Health Center. The availability of images and imaging reports, as well as the impacts of these were compared in the digital health center (study center) in Tikkurila with a health center (control center) in Kerava, where GPs were lacking HIE services.

The study group consisted of 60 patients with conservatively treated ankle or wrist fractures – 41 participants in the digital and 19 in the traditional follow-up group. The clinical follow-up information on encounters in the health centers was collected from electronic patient records after the patients had given their consent to this. Activity based costing was applied for personnel cost evaluation and investments were available from the HUSpacs project.

The performance measures in the process were set to evaluate how often patients actually received the recommended therapy of care pathways designed by Peijas Hospital for the two conservatively treated fracture groups. We also examined the quality of the control visits in respect to the availability of image information and assessed the direct costs for processing images in both groups of fracture patients.

The quality of processes were superior in the digital group, since the GPs had available the primary incident images for all the 41 patients, whereas in the traditional process none of the GPs could track the primary images during the first control visit. The whole process included 122 programmed visits to Tikkurila health center and for only two visits (1,6 %) the GP was unable to apply the regional *e*Health network for sharing images or reports due to web service problems in their work stations. Instead, in these cases they acquired the image information from their health center radiological unit or booked a later control call for the patient.

Although the primary images were missing for the first visit, plain films in Kerava were later available for follow-up, because the control radiological examinations were performed in the health center. However, the time for preparing the digital process in Tikkurila by nursing and administrative staff was only 16 % of the staff time needed in the traditional work process in Kerava. After investments were included the costs for applying HUS-pacs for the regional image transfer, were nevertheless 50 % lower than in the traditional film process. However, no significant differences existed in activity based costing for clinical follow-up visits.

The orthopedic and the GP could view the same image in real-time (figure 2). These *e*Consultations were available once a week for GPs in the digital health center, but GPs from both municipalities also requested radiological consultations at a similar rate. These consultation reports were available in the digital process on the same or next day, whereas it took three days for the reports to reach Kerava by mail delivery.

Migration process to national health information network

The integration of healthcare services has been pursued in Finland by issuing a bill (The New Health Care Act) in 2007 to be considered as the basis for new healthcare legislation. There is good fit with the English integrated health service reform model to the previously described rural experiments of Finnish health care, to the proposed new health act and to application of the regional *e*Health network services. Still there would remain competition between public and private organizations, although outsourcing and private-public-partnerships have created collaboration alongside competition.

The Finnish Parliament passed in 2007 an umbrella legislation mandating the building of a centralized national eArchive and secure communication network connecting all health service providers. Another corollary act requires the creation of a National *e*Prescription service. The statutory contractor for the national eArchiving and ePrescription services is the Social Insurance Institute (SII) [13]. Public and private healthcare providers and pharmacies, responsible for documenting and managing health care information in their legacy systems, are required to use the national services for archiving of documents and prescriptions. They are also obliged to modify their ICT applications accordingly [14].

The national migration plan from regional or local EPR -systems to the NHIN and interoperable EHR have been a combination of following activities:

- the development of legal framework for shared EHRs and ePrescriptions,
- the defining of responsibilities between the national actors (e.g. the Ministry of Social Affairs and Health, SII, National Institute for Health and Welfare, National Supervisory Authority for Welfare and Health and the Association of Finnish Municipal Authorities),
- the development of a common structured core data set for EHRs (e.g. common headings, classifications and terms),
- selection of technical standards both for communication and long term archiving of EHRs,
- the definition of the technical architecture for NHIN,
- the development of use-cases both for the ePrescription and eArchiving services,
- selection of vendors for the development and implementation of national services, and
- the creation of certification requirements for the NHIN, regional and local EPR-systems and ePrescribing services.

The Ministry of Social Affairs and Health has the authority for coordinating the transition process, creation of decrees, development of use-cases and guidelines for system integration. SII is responsible for deploying and maintaining the eArchive, national ePresciption services and e-services for citizens.

The Institute for Health and Welfare (previously National R & D Centre for Welfare and Health) is liable for the content of terminological services, development and maintaining national codes and classifications. A code and terminology server for semantic interoperability has been in place for RHINs and will be scaled for national use by the National Institute for Health and Welfare.

The National Supervisory Authority for Welfare and Health (Valvira) is obliged to create national authorization and identification services for both healthcare professionals (e.g. doctors and nurses) and for service organizations. These services are based on PKI-system and the use of health professional cards (HPC). Valvira also offers on-line attribute services including information on the engagement roles of healthcare professionals and their subsequent justification to view patient information.

Health service provider organizations (e.g. hospital districts, community-wide enterprises, public and private health or medical centers and pharmacies) have a major role in this migration process. Because patient data is sent to the eArchive as documents including standardized metafile and body (HL7 CDA standard), present EPR-systems must incorporate many new features. The most demanding of them is the extraction of patient data items from present RDB-files. Extracted data must then be transformed into the harmonized document form using nationally defined terms and classifications. *e*Prescription applications need to be integrated both functionally and technically to present EPR software. Finally all service provider organizations should implement HPCs and PKI-services to enable trusted communication with ePrescribing and eArchiving services.

In spring 2007 the vendor (Fujitsu Services Ltd) for the national eServices was chosen and has initiated the operation of the project in September 2007. First pilots for ePrescriptions and eDocuments are planned in 2009.

The process to determine the migratory procedures from a regional to a national health information network application is currently being considered based on specifications

that were proposed before the summer 2008 by a cluster of hospital district and RHIN service providers [15].

Roadmap to NHIN

Although some hospital districts adopted a federated model for regional HIE, a central data repository based on a single vendor solution was chosen by several smaller hospital districts due to lack of coordinated efforts to strive for a common integration and architecture strategy. A national health information network would be achievable by allowing integration of federated or centralized regional *e*Health networks to share EPR information, but to circumvent the challenges of interoperability between hospital districts the Ministry of Social Affairs and Health had to revisit its information strategy and architecture.

The updated strategy and communication architecture for health care drawn by the Ministry of Social Affairs and Health has set the following targets:

- for semantic interoperability all EPR-systems should implement a common core data set for EHRs and use HL7 CDA.
- communication between EPR-systems and the *e*Archive shall be based on a standardized message system (e.g. HL7CDA-messages. XML-formats and SOAP envelopes)
- all patient records will be archived into a logically single national archive
- new national architecture, incl. legacy systems, RHINs and national services, must be trusted. Two new acts, EU directive 95 and Act on Patient's Rights form the basis for privacy protection and security

The core data set is a compilation of the key information relating to the health and medical care of the patient [16]. The uniform structured data in an EPR are created chronologically as a summary of the periods of medical care and/or consultations. The data are entered by either the healthcare professionals or persons carrying out the data inclusion. The purpose of the core data set is to provide a general picture of the patient's health and medical history and the related treatment and instruction.

The core data can be used as a link to the detailed medical and patient record data. On the other hand, core data can also be sunk straight into the text. Appropriate document modification may be applied to produce a view of the record which shows the core data or part of the core data. It is also possible to prepare from the structured core data summary care reports or a care plan. These can be utilized for continuum of care, quality control, decision support and research.

In the Finnish NHIN patient records are transferred to the eArchive in the form of documents. For secure long term archiving, the data structure must include a metafile with multi-faced security policy. This metafile should consist of the following information:

- used security policy
- unique identification of data producer, patient and organization
- context and purpose information of the data
- the nature of data
- information for purposes data can be disclosed
- information when patients consent is required for disclosure of data

This national EHR-archive will disclose records in the form of HL7CDA document if necessary legal and other conditions exist. The centralized eArchive forms the basis of the future collection for citizens' life-life long health history. In the future the eArchive will be the point of record sharing in Finland.

The federated data sharing in hospital districts faced decisions on how to migrate from a regional to a national HIE. Technically there were three different variations: (a) regional locator services remain separate from the national *e*Archive, (b) regional locator services and national *e*Archive functions are coordinated and they share work loads, and (c) regional locator services remain between legacy systems and national *e*Archive [11].

The two latter variations – integration and active participant – have remained theoretical, because they would not bring additional advantages, but only costs. They have therefore been discarded from further consideration and discussion.

The first option induces a set of regional operations and processes to be continued after the deployment of the national *e*Archive. In radiology this would mean that images would be stored in the national long term archive after six months, but other images could be readily applied regionally for viewing in secondary and primary care from a regional short-term PACS archive. This consideration stems from the need for availability of information intensive images (f.ex. dynamic magnetic resonance images) in clinical practice and requirements for network capacity to transfer these images across different settings. These demands may not be otherwise always met.

In case regional locator services would be more suitable than the national *e*Archive for information search, consideration for applying regional *e*Health networks in these circumstances might be substantiated. The data-sharing solution of regional *e*Health networks would also enable access to patient encounter data that is not yet stored in the national repository, but resides only in the local EPR. The data is transferred to the national *e*Archive only after the approval of the encounter.

Regional booking and *e*Referral systems which are incorporated into regional *e*Health networks would be applicable in this option. Both of these systems are presently under construction in the national architecture and need to be reconsidered later. Chronic disease and care coordination strategies are being planned and measuring performance indicators will call for data warehouse solutions in order for outcome evaluation and resource allocation to be applied regionally. These requirements have now been partly solved by applying improved mapping procedures to link effectively disparate patient information that is not in structured form (core data set).

Health information exchange in RHINs should therefore be established for those services that need to be preserved

or partly developed for regional care coordination. These services are suggested as:

- regional governance services (e.g. regional resource planning, booking and accounting, regional management)
- regional primary care accident and emergency services (f.ex. acute care portal supported by mapping of data from disparate EPRs)
- regional registries (f.ex. chronic care disease registries in diabetes)
- regional imaging (f.ex. short term archives in radiology, fundus and endoscopy image managing)
- regional consulting services (f.ex. consultation markets for public and private experts)
- regional user and use management

These regional services may be acquired and implemented with present regional *e*Health networks. The regional services may support in synergy the national infrastructure and distribute specific services that are not provided by NHIN. The road map needs to be discussed and developed in coordination with hospital districts operating federated models of HIE.

Certification is an integral part of the migration process to the NHIN. Its central role is to establish trust in all stakeholders. This means that the NHIN as a whole and all its parts fulfill regulatory and other mandatory requirements. Furthermore, the NHIN should meet functional and technical requirements. It is also mandatory for all RHIN and local EHR-systems connected to the national services to possess the same level of functionality, security and privacy protection.

In the context of this migration process certification has been defined as a process confirming that a system or component complies with its specified requirements and is acceptable for operational use [17]. Certification has been divided into two separate processes, the development of certification requirements and the practical certification by an authorized organization using previously developed certification requirements.

As a part of the migration process to the NHIN, the Ministry of Social Affairs and Health initiated a project (TJSERT-project) to develop basic certification criteria for ePrescribing and eArchiving processes. Requirements were developed during the years 2007-2008 by the National R&D Center for Welfare and Health and University of Kuopio. Certification requirements were developed for three basic areas: security and privacy protection, interoperability and functionality.

In the initial phase of the project it was necessary to develop a new method for creating certification criteria for large ICT-systems such as NHIN (LS-CeRM, Large-scale Systems Certification Requirements Methodology). This four-layer graphical method was then used to develop concrete and practical certification requirements. One integral part of the LS-CeRM method is that it allows to validate in practice that requirements are fulfilled [18].

For ePrescribing systems a total of 141 certification targets were recognized (53 for EPR-systems and 88 for pharmacies), and 457 separate criteria developed. In eArchiving 41 targets were established and 109 criteria prepared. From all targets 20 are compatible for both ePrescription and eArching systems. Because in ePrescribing many use-cases were available, it was possible to develop a detailed set of requirements. For eArching systems no use-cases were at our disposal.

Any provider aiming to use the national eArching or ePrescribing services need to comply and fulfill the certification criteria developed in the TJSERT-project. Considering the present state of art of RHIN and EPR systems this can not be achieved instantly. Therefore every requirement was classified for its urgency and three urgency classes (1, 2 and 3) are applied. Class one forms the mandatory basic level. The Ministry of Social Affairs and Health has in the end of 2008 selected class 1 requirements for ePrescribing, and first ePrescribing systems will be certified in the spring of 2009. Later this year level 1 requirements for EPRs to be connected to the eArchiving system will be selected.

The roadmap should implement a holistic view on health care information infrastructure in such a way that local, regional and national services form a comprehensive system that delivers services at all health system levels and improve the effectiveness of various subsystems. Key points in the national eHealth roadmap are application of harmonised architecture approaches and use of standards in all development efforts. The existing and working regional systems should form the basis for future national infrastructure and these current RHINs should be modified to serve also the national requirements for health data management.

Discussion

The social model in Finland resembles the Nordic type based on high taxation and public health care services provided by autonomic municipalities. This decentralization has resulted in investments on one-off electronic patient records needing networking and integration with different stakeholders. As a consequence, federated regional health IT models were developed supported by legislation and funding.

Transformation in healthcare has been slow, nevertheless. This may depend on the pace of information technology adoption, since 80 per cent of the technology in use over the period 1995-2005 is less than ten years old, but 80 percent of the workforce was trained more than ten years ago [19]. It was therefore somewhat unexpected that experienced Finnish physicians considered information technology in healthcare to have led to greater efficiency and facilitated information retrieval [20].

These results are part of an extensive physician survey in 2003 and represent comments from 480 physicians, whose age was between 40 and 55 years at the time of the survey. The physicians who expressed opinions about changes in their work represented only 18 % of the responders to the survey, which undermines the general applicability and reliability of the result.

Achieving reforms of the social model will not be easy and most of the innovations need to be introduced at the national level [21]. The role of the Ministry of Social Affairs and Health warranted the change in healthcare information technology (HIT) architecture from regional to a more centralized *e*Health network, although integrated care will remain for the most part at regional level. The information architecture model is not a pure centralized form, but rather a hybrid, that is dependent on legacy systems and deployed on a service oriented architecture (SOA).

The consequence of this regional cooperation may be envisioned applying theories and practice of the social life of information [21]. People with similar practices and similar resources develop similar identities. These common practices allow people to form social networks along which knowledge about that practice can travel rapidly and be assimilated readily.

Two types of networks may be created – networks of practice and communities of practice. In the first network, professionals may be more loosely connected than in the second network, where they are working together on the same or similar tasks. These new types of collaboration may be used to organize regional health districts into a cluster matrix organisation. Horizontal relationships make up communities of practice, whilst vertical relationships link shared practices as demonstrated by the *e*Referral and *e*Consultation applications.

These two types lend themselves to transformation of healthcare and coexist in the integration models for primary and secondary care [8]. Hospitals would be joined with medical groups or primary care in vertically integrated organizations, that remind the networks of practice and could be adopted in rural areas. An alternative would be for hospitals to remain organizationally distinct and to form long-term alliances with one or more multi-specialty primary care medical groups in a form of virtual integration suited for virtual integration in urban areas.

In healthcare interactive dialogue between management sciences and information systems science has not received much contemplation. The use of ICT systems in health care is not direct evidence of their capacity to generate actual value. ICT systems are enablers for benefits to be reaped if working methods are simultaneously revised. This was demonstrated in the *e*Referral use case.

However, benefits may accrue by just reengineering the care processes [22]. New surgery arrangements for artificial joint patients involved relocating the anaesthesia phase outside the operating theatre The reorganization of the patient care process for joint replacement surgery succeeded in achieving a 50 % increase in operations before the introduction of a new IT system planned as part of the project.

In Satakunta Hospital District a regional *e*Health network based on a federated model with record locator service use has been operating since 2004 [23]. Direct costs were calculated for the district and its four primary care units. The results showed that net savings were annually on average 6 % of the total health care costs. Savings were related to an estimated 20 % reduction in redundant examinations and repeat visits due to lack of diagnostic information. Indirect cost savings were achieved by the delivery of timely care and by avoiding prolongation of disease, absence from work and unnecessary travel costs.

Virtually all work processes are affected by imaging results and telemedicine applications in radiology have been proven to result in savings through avoidance of unnecessary patient transfer or patient travel [24]. HUS-pacs may therefore be seen not only as a technical imaging data transferring system, because it brings added value for regional services and has deep influences on the way of working as was demonstrated by the second use case.

The interrelationship between ICT systems and information and its significance for evaluation of achieved benefits of ICT system use has been stressed [25]. In the future noteworthy ICT system benefits in healthcare should be more the result of development work on the macro level. Therefore, the National Health Project has initiated attempts to improve the quality of patient record systems during 2007-2011 through the construction of a national health information network.

To promote better access to patient information in the legacy repositories of different health care providers, a national project has produced national specifications on the requirements for the content and structure of information systems concerning open interfaces, data protection, information security and construction of information system architecture. The project also concentrates on the distribution of classifications and codes necessary in ICT.

All large scale ICT-systems (the Finnish NHIN is a typical example) sharing personal health data require both careful security and privacy protection planning and implementation of practical security and privacy protection tools and services. From this perspective the Finnish NHIN with a centralized eArchive is exceptionally demanding because all EPRs are stored and disclosed by the same “governmental” organisation. Security features of the Finnish eArchive are based on a new ISO 21457 standard (Secure archiving of electronic health records) [26].

Remodelling of EPR data structure, present identification and authorization mechanisms, development of rule and role based access control services, creation of new audit records, consent management services and audit services for citizens are needed to create a trusted environment for the NHIN. In the long run we need to proceed to the policy based on data access and disclosure mechanism.

For identification of health professionals the national PKI infrastructure applying health professional cards is in the implementation phase. Later a role based access control service will be implemented. Identification of entities are based on the same technical infrastructure.

For trustfulness it is necessary to define the level of non-repudiation of events and which processes or events should be audited. The previously mentioned TJSERT-project already highlighted more than 20 processes or events that should be audited.

It is also necessary that citizens can check to whom, when and for what purpose the national eArchive has disclosed his health information. This is also one service of the Finnish eArchive. Patients and citizens can check data disclose audit-log via the Internet without restrictions.

The National Institute for Health and Welfare has started a project to build a health information portal for citizens. The objective is to make online health education and expert advice available to citizens. It has been also planned that patients will get access to view their own patient records stored in the centralized *e*Archive and unlimited right to check who, when and why, has accessed his or her data. This checking can be done also using the Internet.

Finally, this opens the option for a personalized health record (PHR) to be connected with the EHR. The citizens will be more informed and this may result in to an increasing demand for personalized health services.

Conclusions

Finnish experiments demonstrate the migration from regional to national *e*Health network to be a multidimensional and complicated process. It is not only a technological challenge, but also political, organizational, process related and human factors must be taken into account. Typically this kind of change is loaded (sometimes overloaded) with many expectations and benefits. The migration process should be carefully planned and each step forward should reap benefits to the users. The ICT technology is not the most demanding and most of the necessary solutions are already available on the market. From ICT point of view the most demanding tasks are the creation of common understanding of functionalities and processes of the NHIN, selection of harmonized standards, building better semantic interoperability between EPRs and creating trust in the NHIN.

National development work currently focuses on the specification of technological structures, the networking of actors and the construction of a legislative foundation. In addition to these an agreed and shared common information model (care ontology) would enable the semantic interoperability of information systems and their clinical integration. The governance model for data and information exchange is an essential component of a national IS infrastructure and the current RHINs have experience and implementations of these that need to be included as part of the future development. The common foundation supports the modernisation of operational processes, the broad application of operative innovations and the enforcement of the role of the citizen.

References

[1] Harno K., Ruotsalainen P., Nykänen, P., Kopra K. (2008). Migration from Regional to a National *e*Health Network, Second International Conference on the Digital Society, pp.107-110. http://doi.ieeecomputersociety.org/10.1109/ICDS.2008.28
[2] OECD Reviews of Health Systems, Finland, (2005).
[3] Towards a Networked Finland. The Information Society Council's Report, 2005;57-65.
[4] Thielst C.B., Jones L.E. Guide to Establishing a Regional Health Information Organization. HIMSS, 2007.
[5] Malmqvist G., Nerander K.G., Larsson M. Sjunet – the national IT infrastructure for healthcare in Sweden. In: Iliakovidis I., Wilson P., Healy JC. Current situation and examples of implemented and beneficial eHealth application. IOS Press, The Netherlands, 2005, 41-49.
[6] Bjerregaard Jensen H., Duedal Pedersen C. MEDCOM: Danish health care network. In: Iliakovidis I., Wilson P., Healy JC. Current situation and examples of implemented and beneficial eHealth application. IOS Press, The Netherlands, 2005, 59-65.
[7] Orphanoudakis S. HYGEIAnet: The integrated regional health information network of Crete. In: Iliakovidis I., Wilson P., Healy JC. Current situation and examples of implemented and beneficial eHealth application. IOS Press, The Netherlands, 2005, 66-78.
[8] Ham, C. Clinically Integrated Systems: The Next Step in English Health Reform. The Nuffield Trust Briefing Paper, (2007).
[9] Harno K. UUMA Regional *e*Health Services in the Hospital District of Helsinki and Uusimaa (HUS). In: Iliakovidis I., Wilson P., Healy JC. Current situation and examples of implemented and beneficial eHealth application. IOS Press, The Netherlands, 2005, 101-108.
[10] Stegwee R.A, Rukanova B.D. (2003) Identification of Different Types of Standards for Domain-Specific Interoperability. MIS Quarterly Special Issue Workshop on Standard Making: A Critical Research Frontier for Information Systems, pp 161-170.
[11] Harno, K., Paavola, T., Carlson, C., Viikinkoski, P. Improvement of health care process between secondary and primary care with telemedicine – assessment of an intranet referral system on effectiveness and cost analysis. *Journal of Telemedicine and Telecare* 2000; 6,320-329.
[12] Harno K., Roine R., Pohjonen H., Kinnunen J., Kauppinen T. Framework for systematic assessment of the regional HUSpacs after re-engineering of hospital and external processes. CARS 2002 Computer Assisted Radiology and Surgery, Proceedings of the 16th International Congress and Exhibition Paris, June 26-29, 2002. Pp 618-622. Eds Lemke HU, Vannier MW, Inamura K, Farman AG, Doi K, Reiber JHC. Springer-Verlag Berlin Heidelberg.
[13] Finnish Legislation, (2007). *Law on the processing of personal data relating to social- and healthcare.* http://www.finlex.fi/fi/laki/kokoelma/2007/20070023.pdf
[14] Ruotsalainen P., Iivari A-K, Doupi P. Finland's strategy and implementation of citizen's access to health information. ICMCC Event 2008, London, UK.
[15]Nykänen P., Ohtonen J., Seppälä A. The current role and perspectives of reference data base Regional *e*Health Networks in the context of the national architecture. University of Tampere, Department of Computer Sciences, B-2008-1. Juvenes Print, Tampere, (2008)
[16] Häyrinen, K., Porrasmaa, J., Komulainen, J., & Hartikainen, K. (2004). *Uniform Structured Core Data in an Electronic Patient Record, Final Report 3.2.2004.* University of Kuopio and Association of Finnish Local Authorities.
[17] ISO/IEC 24765, Systems and Software Engineering Vocabulary, 2006.
[18] Ruotsalainen P. Certification Criteria for the National Health Information System. IMIA WG1, WG4 and JHIFT joint WG, Brisbane, 2007.
[19] Giddens, A. (2007). *Europe in the Global Age*. Polity Press, Cambridge, UK.
[20] Haukilahti R-L., Virjo I., Halila H., Hyppölä H., Isokoski M., Kujala S., Vänskä J., Mattila K. Information Technology in Health Care Evaluated by Experienced Physicians. Journal of the Finnish Medical Association 2008;63:4223-4229.
[21] Brown, J.S., & Duguid, P..(2002) *The Social Life of Information.* Harvard Business School Publishing Corporation, USA.
[22] Paavola T., Exploiting Process Thinking in Health Care. International Journal of Healthcare Information Systems and Informatics. 2008;3 (2):12-20.
[23] Maass M., Asikainen P., Mäenpää T., Wanne O., Suominen T. Regional Health Care Network efficient and cost-saving. Journal of the Finnish Medical Association 2007;62:2673-2677.
[24] Maass M., Kosonen M, Kormano M. Transportation savings and medical benefits of a teleneurological network. Journal of Telemedicine and Telecare 2000;6(3):142-6.
[25] Paavola T. Exploring IT System Benefits in Health Care. (2008). Dissertation. Tampere University of Technology. Publication 756.
[26] ISO TS 21457, Secure Archiving of Electronic Health Records, 2008.

Address for correspondence

Kari Harno, M.D. Ph.D.
Chief Physician
Helsinki University Central Hospital
Stenbäckinkatu 9
00029 HUS
Finland

Figure 1. The Regional *e*Health Network Architecture in the Hospital District of Helsinki and Uusimaa.

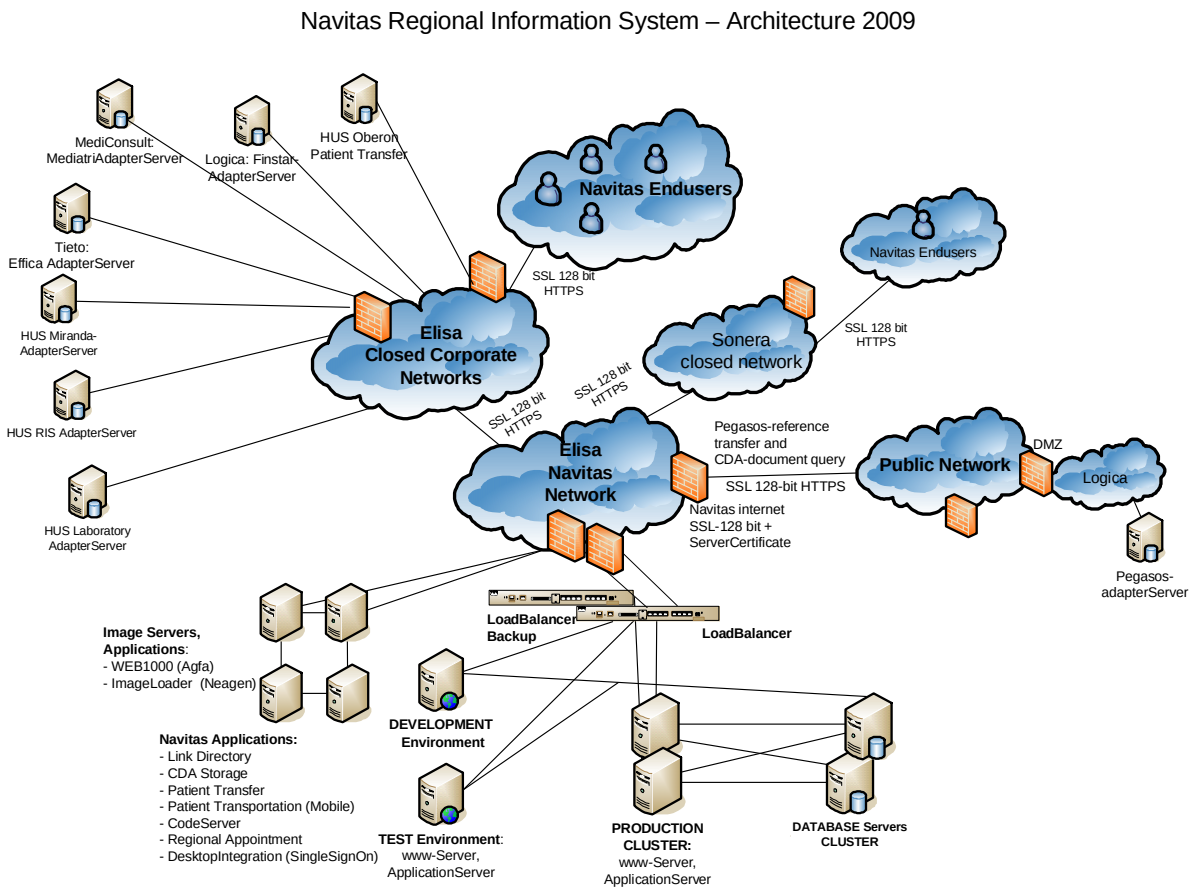
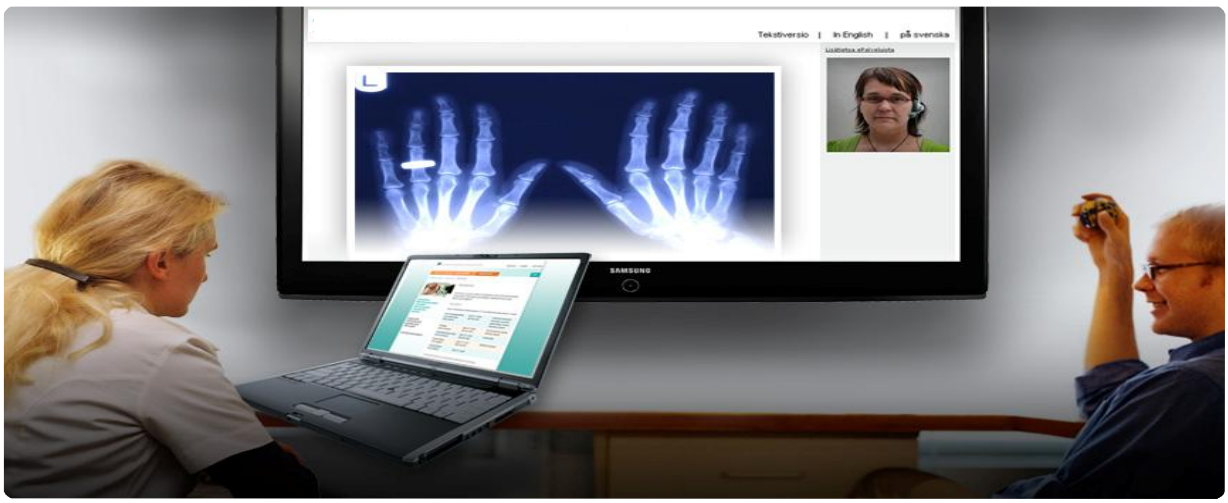


Figure 2. Regional *e*Health network provides an opportunity for GPs and orthopaedics to view the same image in real-time and exchange information as consultations.



www.iariajournals.org

International Journal On Advances in Intelligent Systems
ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
issn: 1942-2679

International Journal On Advances in Internet Technology
ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
issn: 1942-2652

International Journal On Advances in Life Sciences
eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
issn: 1942-2660

International Journal On Advances in Networks and Services
ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
issn: 1942-2644

International Journal On Advances in Security
ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
issn: 1942-2636

International Journal On Advances in Software
ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
issn: 1942-2628

International Journal On Advances in Systems and Measurements
ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
issn: 1942-261x

International Journal On Advances in Telecommunications
AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
issn: 1942-2601