Supporting the Design of Sequences of Cumulative Activities Impacting on Multiple Areas through a Data Mining Approach: Application to Design of Cognitive Rehabilitation Programs for Traumatic Brain Injury Patients

Alejandro García Rudolph, B.Sc., DEA

Thesis Submitted to Universitat Politècnica de Catalunya BarcelonaTech (UPC) for the Degree of Doctor of Philosophy (Ph.D)

Research was carried out in the Statistics and Operation Research Department, UPC under the supervision of Professor Karina Gibert.

Statistics and Operation Research Department,
Universitat Politécnica de Catalunya - BarcelonaTech

December 2015

Declaration

I hereby certify that this material, which I now submit for assessment on the program of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signature:_____ **I.D._____**

Date:_____

Acknowledgements

My most sincere gratitude and appreciation go to Dr. Karina Gibert for allowing me to carry out this research under her guidance, supervision and encouragement.

Thanks to Institute Guttmann Neurorehabilitation Hospital for receiving me since 2004, particularly to Josep María Tormos for his continuous support and advice, to my colleagues Eloy Opisso, Raquel López, Marc Morell, Marta Rudilla, David Hurtado, David Sánchez, Jaume López and María Almenara at the Research Department with whom we share day to day. Thanks also to Teresa Roig Rovira, Head of the Neuropsychology Department and her team of neuropsychologists specially Alberto García-Molina, Rocío Sánchez-Carrión, Antonia Enseñat, Pablo Rodriguez-Rajo, Celeste Aparicio, Vega Muriel, Juan Luis García Fernandez, Montserrat Martinell and Beatriz Gonzalez. Thanks to Montserrat Bernabeu Guitart, Head of the Acquired Brain Injury Unit. Special thanks to Olga Araujo from the Santi Beso Arnalot Documentation Center at Documentation Department and to Mercè Solans, Arantxa Cabrera and Maria Jurado from Institute Guttmann Teaching Department. I also want to especially acknowledge Angel Gil, Lluisa Curcoll. Joan Saurí, Dolors Soler, Carola Carbonell and Laura Pla for their participation and contributions in QVidLab project.

This research was partially supported by several national and international grants which I also acknowledge: Ministry of Industry, Tourism and Trade (Spain) AVANZA PLAN-Digital Citizen Subprogram (PT: NEUROLEARNING Grant Nr: TSI-020501-2008-0154). Institute of Health Carlos III (Spain) Strategic Action Health's Call (PT: Clinical implantation of PREVIRNEC platform in TBI and stroke patients / Grant Nr: PI08/900525). Ministry of Science and Innovation (Spain) INNPACTO Program (PT NEUROCONTENT - Grant Nr 300000-2010-30). Ministry of Education Social Policy and Social Services (Spain) IMSERSO Program (PT COGNIDAC - Grant Nr 41/2008). MARATÓ TV3 Foundation (PT: Improving Social Cognition and meta-cognition in schizophrenia: A tele-rehabilitation project - Grant Nr 091330) Spanish Ministry of Economy and Finance (PT COGNITIO – Grant Nr TIN2012 38450. EU CIP-ICT-PSP-2007-1 (PT: CLEAR - Grant No.: 224985) and EU-FP7-ICT (PT PERSSILAA Grant Nr

610359). Among projects' partners special thanks to Grupo de Ingeniería y Telemedicina de la Universidad Politécnica de Madrid, particularly to Enrique Gómez Aguilera, Javier Solana, César Cáceres, Paloma Chausa and Alexis Marcano. Also from Universitat Rovira i Virgili special acknowledgement to María Ferré Bergadà. And thanks to Xavier Monzó and Jordi Ceballos from Grupo ICA.

Special thanks to Santiago Seminario and Gareth George for their support on language editing of this work and previous publications.

I would like to thank all my family for their continuous support over the past ten years, especially to my brothers Walter and Miguel Angel, my sister Leonor and my sister in law Valeria, this work is dedicated to my parents: Alessandra Rudolph Krause and Jorge García Posas.

Dedication

To my parents

-

Table of Contents

Abstract

Traumatic brain injury (TBI) is a leading cause of morbidity and disability worldwide. It is the most common cause of death and disability during the first three decades of life and accounts for more productive years of life lost than cancer, cardiovascular disease, and HIV/AIDS combined.

Disturbances of attention, memory, and executive functioning are the most common neurocognitive consequences of TBI at all levels of severity and have a major impact on daily living activities. Despite new techniques for early intervention and intensive care units, both increasing the survival rate, there is still no surgical or pharmacological treatment for the re-establishment of lost functions following brain injury. Dating back to Luria's theory from 1978, there is a common belief that direct retraining of damaged cognitive processes through repeated stimulation and activation of the targeted brain areas can help patient recovery. Neurorehabilitation is the process of exploiting cerebral plasticity to reduce brain deficit. Cognitive Rehabilitation (CR), as part of Neurorehabilitation, aims to reduce the impact of disabling conditions and to improve the cognitive deficits caused by TBI. CR treatment consists of hierarchically organized tasks that require repetitive use of impaired cognitive functions.

While task repetition is not the only important feature, it is becoming clear that neuroplastic change and functional improvement only occur after a number of specific tasks are performed in a certain order and repetitions and does not occur otherwise. Until now, there has been an important lack of well-established criteria and on-field experience by which to identify the right number and order of tasks to propose to each individual patient.

Finding recommendations for the sequence of tasks and repetitions that will induce better improvement in a single patient is a difficult problem because: tasks show high order of interactions among them and cummulative effects, and also treatment lengths and sequential task configuration is open.

This thesis proposes the CMIS methodology to support health professionals to compose CR programs by selecting the most promising tasks in the right order.

Two contributions to this topic were developed for specific steps of CMIS through innovative data mining techniques: SAIMAP and NRRMR methodologies.

SAIMAP (Sequence of Activities Improving Multi-Area Performance) proposes an innovative combination of data mining techniques in a hybrid generic methodological framework to find sequential patterns of a predefined set of activities and to associate them with multi-criteria improvement indicators regarding a predefined set of areas targeted by the activities. SAIMAP is introduced as an integrative methodology that uses both data and prior knowledge with preprocessing, clustering, motif discovery and classes` post-processing to understand the effects of a sequence of activities on targeted areas, provided that these activities have high interactions and cumulative effects.

Furthermore, this work introduces and defines the Neurorehabilitation Range (NRR) concept to determine the degree of performance expected for a CR task and the number of repetitions required to produce maximum rehabilitation effects on the individual. An operationalization of NRR is proposed by means of a visualization tool called SAP. SAP (Sectorized and Annotated Plane) is introduced as a visualization tool to identify areas where there is a high probability of a target event occurring. Three approaches to SAP are defined, implemented, applied, and validated to a real case: Vis-SAP, DT-SAP and FT-SAP, the parametric heatmap-based visualization proposed to overcome the limitations detected in Vis-SAP.

Finally, the NRRMR (Neurorehabilitation Range Maximal Regions) problem is introduced as a generalization of the Maximal Empty Rectangle problem (MER) to identify maximal NRR over a FT-SAP.

These contributions combined together in the CMIS methodology permit to identify a convenient pattern for a CR program (by means of a regular expression) and to instantiate by a real sequence of tasks in NRR by maximizing expected improvement of patients, thus provide support for the creation of CR plans. First of all, SAIMAP is intended to provide the general structure of successful CR sequences for a single patient providing the length of the sequence and the kind of task recommended at every position (attention tasks, memory task or executive function task). Next, NRRMR aims to provide specific tasks information to help decide which particular task is placed at each position in the sequence, the number of times it needs to be repeated, and the expected range of results to maximize improvement along the treatment.

From the Artificial Intelligence point of view the two methodologies proposed are general enough to be applied to other problems matching the same structure where a sequence of interconnected activities with cumulative effects are used to impact on a set of areas of interest, for example spinal cord injury patients following a physical rehabilitation program or elderly patients facing cognitive decline due to aging who make use of cognitive stimulation programs or also on educational settings, to find the best way to combine mathematical drills in a program for a specific Mathematics course.

.

.

Abbreviations

ABI	Acquired Brain Injury
ANN	Artificial Neural Network
ARM	Association Rules Mining
BPNN	Backpropagation Neural Network
CVA	Cerebrovascular accident
ClBR	Clustering Based on Rules
CLS	Concept Learning System
CR	Cognitive Rehabilitation
DNA	Deoxyribonucleic Acid
DM	Data Mining
DT	Decision Tree
DT-SAP	Decision Tree Sectorized and Annotated Plane
EHR	Electronic Health Record
EClBR	Exogenous Clustering based on rules
FP-Tree	Frequent Pattern Tree Algorithm
GCS	Glasgow Comma Scale
ID3	Iterative Dichotomizer 3
KB	Knowledge Base
KDD	Knowledge Discovery in Databases
k-NN	k-Nearest Neighbour

MEME	Multiple Expectation-Maximization for Motif Elicitation
MER	Maximal Empty Rectangle
MRI	Magnetic Resonance Imaging
NAB	Neuropsychological Assessment Battery
NB	Naive Bayes
NRR	NeuroRehabilitation Range
NRRMR	NeuroRehabilitation Range Maximal Regions
PAT	Person Artifact Task
PTA	Post Traumatic Amnesia
QVidLab	Laboratory for Enhancing Measures of Autonomy, Personal Satisfaction and Quality of Life of People with Neurological Disabilities
RBF	Radial Basis Function
RCT	Randomized Controlled Trials
RN	Reciprocal Neighbours
SAIMAP	Sequence of Activities Improving Multi-Area Performance
SPMF	Sequential Pattern Mining Framework
SAP	Sectorized and Annotated Plane
SPAM	Sequential Pattern Mining
SPADE	Sequential Pattern Discovery using Equivalence classes

SVM	Support Vector Machine
TBI	Traumatic Brain Injury
TFBS	Transcription Factor Binding Sites
TMOD	Toolbox for Motif Discovery
TMT	Trial Making Test
Vis-SAP	Visualization based Sectorized and Annotated Plane
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
ZPD	Zone of Proximal Develpment
ZRP	Zone of Rehabilitation Potential

Index of Figures and Tables

Figures

Tables

Chapter 1. Introduction

1.1 Introduction

Traumatic brain injury (TBI) – defined as an alteration in brain function, or other evidence of brain pathology due to an external cause – is a leading cause of morbidity and disability worldwide (Scholten, et al.,2014). There is one case of TBI every 15 seconds and every 5 minutes someone becomes permanently disabled due to a head injury (Kouroupetroglou, 2013).

In Europe, brain injuries from trauma are responsible for more years of disability than any other cause (Nimmo, 2011). It is the most common cause of death and disability during the first three decades of life and accounts for more productive years of life lost than cancer, cardiovascular disease, and HIV/AIDS combined (Zitnay, et al., 2008). The incidence is increasing in lower income countries and the World Health Organization predicts that TBI will be the third major cause of disease and injury worldwide by 2020 (Dinsmore, 2013). Furthermore, TBI is considered a *silent epidemic*, because society is largely unaware of the magnitude of the problem (Rusnak, 2013).

The consequences of TBI vary from case to case but can include motor, cognitive, and behavioral deficits in the patient, disrupting their daily life activities at personal, social and professional levels. The most important cognitive deficits after suffering a TBI are those related to attention, decrease in memory and learning capacity, worsening of the capacity to schedule and to solve problems, a reduction in abstract thinking, communication problems, and a lack of awareness of one's own limitations. These cognitive impairments hamper the path to functional independence and a productive lifestyle for the person with TBI.

Despite new techniques for early intervention which increase the survival rate, there is still no surgical or pharmacological treatment for the re-establishment of lost functions following brain injury. Cognitive rehabilitation (CR) (Pascual-Leone, Amedi, Fregni, & Merabet, 2005) is currently considered the therapeutic process for re-establishing functioning in everyday life. A typical CR program mainly provides exercises which require repetitive use of the impaired cognitive system in a progressively more demanding

sequence of tasks (Sohlberg & Mateer, 2001). The rehabilitating impact of a task or exercise depends on the ratio between the skills of the treated patient and the challenges involved in the execution of the task itself. Thus, determining the correct training schedule requires a very precise trade-off between sufficient stimulation and sufficiently achievable tasks. This is far from intuitive and is still an open problem, both empirically and theoretically (Green & Bavelier, The cognitive neuroscience of video games, 2006). It is difficult to identify this maximum effective level of stimulation; therapists use their expertise in daily practice without precise guidelines on these issues.

1.2 Motivation

There is a common belief that CR is effective for TBI patients, based on a large number of studies and extensive clinical experience. Different statistical methodologies and predictive data mining methods have been applied to predict the clinical outcomes of the rehabilitation of patients with TBI (Rughani, y otros, 2010) (Ji, Smith, Huynh, & Najarian, 2009); (Pang, y otros, 2007); (Segal, y otros, 2006); (Brown, McClelland, Diehl, Englander, & Cifu, 2006); (Rovlias & Kotsou, 2004); (Andrews, y otros, 2002). Most of these studies focus on determining survival, predicting disability or the recovery of patients, and looking for the factors that better predict the patient's condition after TBI.

However, current knowledge about the factors that determine a favorable outcome is mainly empirical and the benefit of such interventions is still controversial (Rohling, Faust, Beverly, & Demakis, 2009). (ECRI, 2011). The development of new tools to evaluate scientific evidence of such effectiveness will contribute to a better understanding of CR.

Several meta-analyses (Cicerone, y otros, 2011) identify structural limitations that make it difficult to find scientific evidence under classical approaches, related mainly to the existence of uncontrolled factors and the intrinsic difficulty of guaranteeing the sample heterogeneity. Classical approaches tend to generate evidence about effectiveness by comparing two or more interventions in selected and comparable groups. Determining the comparable groups relies on identifying the factors that influence recovery or chronicity, which should be controlled during the study, and these factors are unknown in

neurorehabilitation. It seems that patient improvement might depend inter alia on the location of the injuries, cognitive profile, the duration and intensity of the proposed treatments and their level of completion (Whyte & Hart, 2003), (Cicerone, y otros, 2011), (de Noreña, y otros, 2010).

However, these seem to be only some of the determining factors and they cannot by themselves explain the overall phenomenon. Although these factors are considered in the design of rehabilitation treatments, other relevant factors exist that are much more difficult to control, and which are related to the high variability of the lesions, the complexity of cognitive functions, and the lack of proper instrumentation by which to systematize interventions. This produces intrinsic group heterogeneity and classical comparative studies do not perform well (Gibert & García-Rudolph, 2006). In turn, this makes it difficult to advance knowledge on the pathophysiology of cognitive neurorehabilitation.

For these reasons, other approaches have to be found to better understand the CR process, with the aim of obtaining scientific evidence about its effectiveness and providing relevant information for the establishment of general guidelines for CR program design that can assist CR therapists in clinical practice.

Analyzing data from new perspectives can contribute to this field (Jagaroo, 2009). Our proposal in this thesis is to approach the problem from a data-driven perspective, by developing new tools that can reduce uncertainty in the field.

1.3 Research framework

This work has been performed within a collaboration framework between Dr. Karina Gibert from Universitat Politècnica de Catalunya-BarcelonaTech and Institut Guttmann-Neurorehabilitation Hospital.

1.3.1 Institut Guttmann-Neurorehabilitation Hospital

Institut Guttmann (founded on 27 November 1965, Barcelona) is a specialized hospital in the medical and surgical treatment and comprehensive rehabilitation of people with spinal cord injury, acquired brain injury or other neurological disabilities (Institut Guttmann - Hospital de Neurorehabilitació, 2015). As stated in Article 7 of its Statutes, the Institut

Guttmann's main objective is to promote, encourage, and ensure the full rehabilitation of people affected by spinal cord injury, acquired brain injury or other neurological disabilities, and to provide the most appropriate support and services to achieve a satisfactory social reintegration while contributing to the full recognition of their rights and effective equalizing of opportunities. The scientific activity of Institut Guttmann, according to the organization's Strategic Plan, aims to:

- Promote the development and incorporation of new knowledge from the field of neuroscience, bioengineering, and medical technology.
- Integrate the advances made in clinical and translational research to promote better care alternatives.
- Promote the establishment of alliances, agreements or cooperation agreements with institutions and entities similar to our organization.

As shown in Figure 1.1, the Institut Guttmann has established three research lines: Neurorehabilitation of spinal cord injury (L1), Neurorehabilitation of acquired brain damage (L2) and Pediatric neurorehabilitation (L3). Each research line incorporates seven strategic translational research programs: Bioengineering applied to human functional autonomy (P1), Neurostimulation, neuromodulation and noninvasive stimulation (P2), ICT (Information and Communications Technologies) applied to neuropsychological rehabilitation and cognitive stimulation (P3), Neurorehabilitation in digestive function (P4), Regenerative Medicine applied to Spinal Cord Injury and Acquired Brain Injury (P5), Analysis of results applied to the generation of knowledge – QvidLab (P6), Pharmacological interventions in neurorehabilitation (P7). This thesis takes place in the intersection of Line 2 and Program 6: QvidLab.

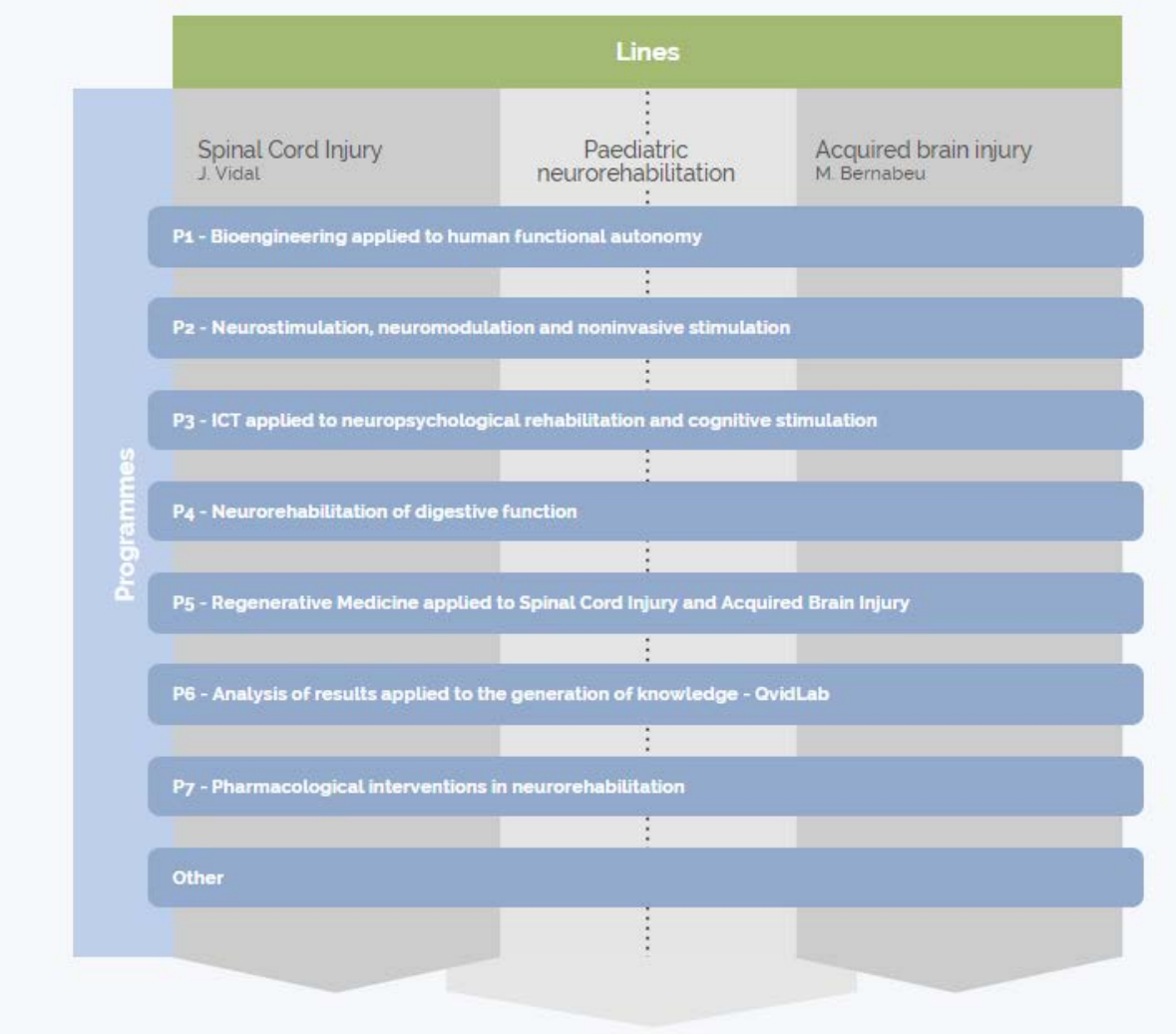


Figure 1.1: Institut Guttmann Research lines and programs. This illustration shows the three main research lines and their respective transversal programs. This thesis is in the intersection of Acquired Brain Injury research line and Analysis of results applied to the generation of knowledge-QVidLab program.

1.3.2 QVidLab

The Laboratory for Enhancing Measures of Autonomy, Personal Satisfaction and Quality of Life of People with Neurological Disabilities (QVidLab) was set up in 2006, as a result of the collaboration agreement between Ministerio de Trabajo y Asuntos Sociales-Secretaría de Estado de Servicios Sociales, Familias y Discapacidad, and Institut Guttmann. It is intended as an instrument of applied clinical research in order to study a set of biological, psychological, and social factors that come together in people with spinal cord injury or

acquired brain damage, either in the short, medium and long term, by the application of advanced data analysis methodologies (Gil Origüén, 2009). It is built upon Institut Guttmann's Electronical Health Records (EHR), which integrates relevant patient information, including structural damage, functional impairment, limitation of participation and impact of environmental factors (barriers, facilitators etc.) that might interfere in the social inclusion and independent living of disabled persons. The QvidLab is a higher level layer to EHR containing multidimensional data suitable for creating a better understanding of neurological disabilities and for generating new knowledge (Gibert & Tormos, 2014).

As shown in Figure 1.1, QVidLab spans transversally along the three main research lines. In this work we will focus on the Acquired Brain Injury (ABI) line. ABI is defined as a brain injury that has occurred after birth (Brain Injury Association of America, 2015). ABI provides a broad umbrella definition and, depending on the severity of the injury, can include etiologies such as cerebrovascular accidents (CVAs) or strokes, and encephalitis. ABI does not include brain injury that is congenital, hereditary, degenerative or induced by birth trauma.

ABI includes Traumatic brain injury (TBI), a type of ABI that refers to structural injury that has been induced traumatically and/or a physiological disruption of brain function resulting from an external force (Vincent, Roebuck-Spencer, & Cernich, 2014). Together with stroke, TBI is one of the two main causes of ABI worldwide (Kamalakannan, Gudlavalleti , Murthy Gudlavalleti, Goenka, & Kuper, 2015). Without losing generality, the methods presented in the following chapters are applied to TBI patients but can be addressed to other ABI patients undergoing CR treatment under the conditions defined in the upcoming chapters.

1.4 Clinical background

This section introduces, defines and highlights specific clinical foundations for the forthcoming chapters.

1.4.1 Traumatic Brain Injury

TBI is defined as an alteration in brain function, or other evidence of brain pathology, caused by an external force (Menon, Schwab, Wright, & Maas, 2010). Alteration in brain

function is defined as one of the following clinical signs: any period of loss or decreased consciousness; any loss of memory for events immediately before (retrograde amnesia) or after the injury (post-traumatic amnesia, PTA); neurologic deficits (weakness, loss of balance, change in vision, dyspraxia paresis/plegia, sensory loss, aphasia, etc.); any alteration in mental state at the time of the injury (confusion, disorientation, slowed thinking, etc.). Other evidence of brain pathology includes visual, neuroradiological, or laboratory confirmation of damage to the brain (Menon, Schwab, Wright, & Maas, 2010).

The central factor is that brain damage results from external forces as a consequence of direct impact, rapid acceleration or deceleration, a penetrating object (e.g. gunshot) or blast waves from an explosion. The nature, intensity, direction, and duration of these forces determine the pattern and extent of damage.

1.4.2 Cognitive Rehabilitation

Cognitive rehabilitation (CR), tries to improve the deficits caused by TBI in daily living activities (Bernabeu & Roig, 1999) by retraining attention, memory, reasoning/problem solving, and executive functions. The plasticity of the central nervous system plays a central role (Pascual-Leone, Amedi, Fregni, & Merabet, 2005) in CR, based on therapeutic plans to stimulate non-damaged neurons that can modify their structure by learning from experience, through repetition (Luria, 1976). Plasticity may represent a surrogate marker of functional recovery indicative of behavioral change that is resistant to decay. It is suggested (Kleim & Jones, 2008) that a sufficient level of rehabilitation is likely to be required in order to get the subject “over the hump,” i.e. repetition may be needed to obtain a level of improvement and brain reorganization sufficient for the patient to continue to use the affected function outside of therapy and to achieve and maintain further functional gains. A great deal of research indicates that behavioral experience can enhance behavioral performance and optimize restorative brain plasticity after brain damage (Kleim & Jones, 2008). Simply engaging a neural circuit in task performance is not sufficient to drive plasticity. Repetition of a newly learned (or relearned) behavior may be required to induce lasting neural changes. In fact, from the expert’s point of view, there is a clear perception that the effectiveness of the task also depends on the replication, as Luria also asserts.

A typical CR program mainly provides exercises that require repetitive use of the impaired cognitive system in a progressively more demanding (Sohlberg & Mateer, 2001) sequence of tasks. Each task targets a principal cognitive function and can be performed at different levels of difficulty, according to the response of the patient. The design of a CR program has become an essential issue for patient recovery. The rehabilitating effect of a task or exercise depends on the ratio between the skills of the treated patient and the challenges involved in the execution of the task itself. The difficulty is related to the level of stimulation of cognitively involved functions; maximum activation occurs when the task is "just barely too difficult" (Green & Bavelier, 2006). If the task is either too easy or too hard for the patient, it appears to be less effective. Active monitoring of the subject's progress is therefore required to adapt the difficulty of the tasks to the potential capacities and progress of the subject, always pushing them to reach a goal just beyond what they can attain, but not too far. Thus, determining the correct training schedule requires a very precise trade-off between sufficiently stimulating and sufficiently achievable tasks. This is far from intuitive and is still an open problem, both empirically and theoretically.

1.4.3 Assessment of the effects of CR

Before starting the CR program every patient undergoes a Neuropsychological Assessment Battery (NAB). This battery includes 28 items covering the major cognitive domains (attention, memory and executive functions) measured using standardized cognitive tests. NAB consists of a selection of some items from 7 assessment instruments, associated with the different cognitive functions, which in turn are evaluated under some specific sub-functions. In view of the fact that conventional neuropsychological instruments are notorious for amalgamating cognitive operations (Jagaroo, 2009); (Sabb, y otros, 2009), in the proposed approach a subset of NAB items with the highest levels of specificity has been selected in collaboration with domain experts. The final items considered in this work are the following 14 non-redundant items:

- Memory:
    - Visual and Verbal Memory:
        - The Rey Auditory Verbal Learning Test (Rey, 1964) (RAV075, RAV015 and RAV015R items)
- Attention:
    - Sustained Attention:
        - Continuous Performance Task Test (Conners & Sitarenios, 2011) (OMI, COMI and CPT items)
        - Trail Making Test-A (Reitan & Wolfson , 1993) (TMTA item)
    - Selective Attention:
        - WAIS-III Selective attention (Wechsler, 1997) (VWAIS item)
    - Divided Attention:
        - Trail Making Test-B (Reitan & Wolfson , 1993) (TMTB item).
- Executive Functions:
    - Planification:
        - WAIS-III Visuo Construction (Wechsler, 1997) (CUBES item)
    - Inhibition:
        - Stroop Test (Golden, 1994) (INTER item)
    - Flexibility
        - Wisconsin Card Sorting Test (Heaton, Chelune, Talley, Kay, & Curtiss, 1997) (TERR item)
        - Letter Fluency Test (Artiola i Fortuny, Hermosillo Romo, Heaton, & Pardee III, 1999) (PMR item)
    - Categorization:
        - The Wisconsin Card Sorting Test (Heaton, Chelune, Talley, Kay, & Curtiss, 1997) (CAT item)

All NAB items are normalized to a 0 to 4 scale (where 0 = No affectation, 1 = mild affectation, 2 = moderate affectation, 3 = severe affectation and 4 = acute affectation) (Gil Origüén, 2009).

After this initial evaluation, patients start CR program (for 2 to 5 months, depending on the patient) using a specific software specifically developed in the hospital (PREVIRNEC© platform, at the moment of the submission of this thesis' initial

publications PREVIRNEC© was the name of the platform, now at the moment of thesis submission it is Guttmann, NeuroPersonalTrainer®, but PREVIRNEC© was kept along this thesis for consistency) which is described in the next section (1.4.4). After treatment every patient undergoes the same NAB to evaluate the cognitive outcome status.

Information obtained in the NAB before and after treatment is the basis on which to understand the improvement of the patient and, in consequence, the response level to the treatment itself. Measuring global improvement in a specific cognitive function (e.g. Attention) implies studying response to treatment in each of the subfunctions involved in NAB tests (e.g. Sustained, Selective, Divided Attention). Different criteria can be adopted (take the subfunctions' average; take the maximum difference, etc). To the best of our knowledge, within the clinical CR therapists community no standardized approach is universally accepted to determine the improvement of the patient from a systematic point of view.

1.4.4 Cognitive Rehabilitation Platform

The Information Technology framework for CR treatments in our clinical setting is the PREVIRNEC© platform (Tormos, Garcia-Molina, Garcia-Rudolph, & Roig, 2009). A J2EE client-server architecture specifically designed and developed to manage CR plans assigned by therapists to patients and the follow up information about the process.

It is conceived as a tool for the enhancement of cognitive rehabilitation, the strengthening of the relationship between the neuropsychologist and the patient, the personalization of treatment, the monitoring of results, and the performance of tasks. The platform architecture consists of four main modules that group related functionalities vertically, sharing the user interface that is personalized depending on the user's role. This interface is also multi-language, with Catalan, Spanish and English already implemented, but being open to support any other language. The system also has a Help module, which guides the user in order to complete each action. Security aspects are transversal and have to be taken into account in every module, in order to keep information and all connections safe due to the confidentiality concerns of medical applications. The security module is responsible for controlling every access, including the ones related to the patient's Electronic Health Record (EHR). The four modules are briefly described below (Solana, y otros, 2011):

Information management: this module groups functionalities related to the generation and editing of information that depends on the patient's EHR, as well as the tests used to determine the grade of affection of each cognitive function. These tests are used to define the affection profile of the patient. In addition, this module controls the assignation of therapies to patients, determining which computerized tasks a patient has to do on a certain day. The results of the execution of these tasks are recorded in the system. They can then be used by the clinicians to view the evolution of the therapy, to display graphics and reports relating to the completion of the sessions, the tasks that have been used, global and individual results, and much more.

Monitoring: To comply with data protection laws, every action carried out by a user is stored in both the database and also in a log file, so that the administrator can track every action related to a patient and their data. The system also offers a module for monitoring the execution of the tasks, so the therapist can then reproduce a task as it was done by the patient. This allows the therapist to see exactly what a patient did in the monitored task. This is very useful because sometimes merely seeing the numeric results is not enough.

Administration: This module, although it has the fewest functionalities and users, includes very important functionalities such as user management, user profiles, and system monitoring (using logs).

Communication: The main element of this module is the video conference function that allows users to communicate using video, audio, and chat. By means of the videoconference function, therapists can hold tele-appointments with patients or other therapists, removing the distance barriers between users, and helping the patients to feel closer to the clinical team. In addition to the videoconference, this module has a mailing service for the exchange of internal asynchronous messages and an alert service that lets users know what tasks they have to accomplish.

The platform is based on open source web 2.0 technologies. The main architecture of the platform is based on client-server communication using HTTP and XML-RPC, as shown in Figure 1.2.

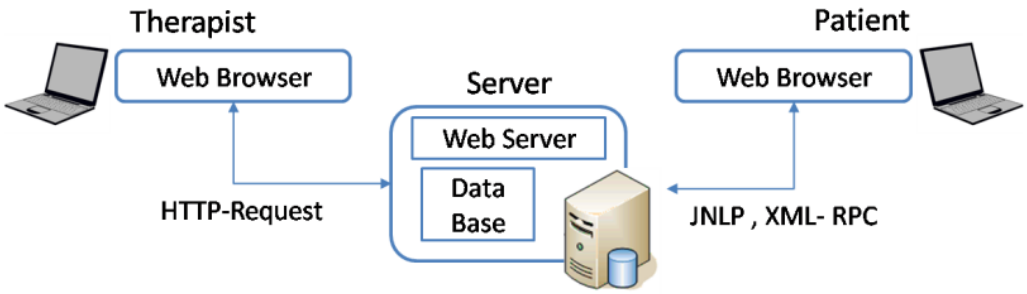


Figure 1.2 PREVIRNEC© Client / Server communication schema

A Model-View-Controller pattern was followed during the development phase. As a result, the view and the logic to access and process data are separated.

The new web application requires Java (jdk 1.6, jre 6.x) and it runs over Apache Tomcat 6.X, as it is based on Servlet/JSP. The database used is MySQL Server 5.X and MySQL Java Connector 5.X (JDBC).

With regard to the programming languages used, all of the environment is a Java 2 platform (J2EE, Enterprise Edition) that uses JavaScript and AJAX (SACK library) to dynamically change the data displayed on the HTML pages, thereby avoiding the need to reload the page every time the user wants to show or edit content.

For the videoconference module, OpenMeetings have been used. This implements the Real Time Multimedia Protocol (RTMP) using a red5 server for audio and video streaming.

For each person or entity using the system in a determined context and for a specific goal, four different user profiles have been defined:

Patients: man or woman of any age with one or some cognitive functions affected, as a consequence of suffering ABI. The caregiver role appears here, considered a secondary actor that will help the patient use the system when necessary.

Therapist: a neuropsychologist who specializes in cognitive rehabilitation in patients with ABI, who will have a number of assigned patients and be responsible for their treatment, scheduling and the monitoring of personalized and individualized therapies.

Supervisor: person in charge of the user management for each center, both patients and therapists and their assignments, apart from other management and control functions applied to the supervising center/s.

Administrator: apart from all the typical administration tasks for every informatics system, the administrator will be the person responsible for managing the categories, functions, and tasks defined in the system, which is the content the therapist will use to schedule therapy sessions for their patients.

1.4.5 Cognitive Rehabilitation Tasks

The therapeutic content used in PREVIRNEC© tele-rehabilitation sessions consists of computerized tasks grouped by cognitive functions. The neuropsychologist creates a tele-rehabilitation session by assigning a set of tasks to a certain session day. He or she is able to configure the difficulty of each task because they all have a set of input parameters.

At the time of this analysis, PREVIRNEC© includes one hundred and fifteen rehabilitations tasks. Each task is defined by a series of parameters that determine its level of difficulty. The therapist selects for each task the parameter to be used for the automatic adjustment of the difficulty level described above. This dynamic adjustment of the difficulty level is performed twice for each task as necessary. This means that if the patient does not obtain a task result in NRR in the first execution, PREVIRNEC© automatically generates the task with the adjusted difficulty level once; if again the obtained result is not in TR, PREVIRNEC© likewise generates a second version of the task.

For illustrative purposes one such task designed for visual memory treatment is described below in more detail. This task (identified as idTask=151) has been one of the most extensively administrated by neuropsychologists and executed by participants during the analyzed period (described in Chapter 9).

The objective of the task is to recall the position of pairs of identical images in a grid. A grid of fixed size (e.g. 5x5 dark colored cells) is presented to the participant at the start. When the participant left-clicks on a cell in the grid, an image of an object on a white background appears in the cell. This image remains until a second cell is clicked, then both images are shown for a period of time (e.g. 1500 ms) for the participant to remember them; afterwards both images are covered. Only two cells can be simultaneously discovered in one go. When two identical images are discovered, both of them remain visible in their cells. The aim of the task is to discover all the images in the grid with the minimum number

of clicks. The parameters that determine the different difficulty levels are shown in Table 1.1

Number of cells	Stimulus type	Proximity of the second image	Presentation time
4x4	abstract objects	2 cells	1500 ms
5x5	numbers	3 cells	3000 ms
6x6	animals	4 cells	4000 ms
8x8	colors	Random	

Table 1.1 Parameters that determine idTask= 151 level of difficulty

The quantified result parameters for the evaluation of task completion are: the total execution time, the total number of discovering clicks, the total number of wrong clicks (this number increases if the participant clicks on an image already discovered before, meaning that errors are computed after an initial exploration phase), the total number of correct clicks (in this case, although it is computed for homogeneity with other tasks, the number of clicks for all participants is constant because the task is considered unfinished until all the images are discovered; this also means that task151 does not produce omissions, and they are presumed to be zero). The task result is computed as:

Task result = [correct / (correct + wrong + omissions)]*100

In this work, the underlying structure of the CR phenomenon has been analyzed in depth and it has been seen that the CR field has some specific characteristics that make the successful application of traditional methods difficult:

- Patients following a CR program are performing neither a single task nor a single type of tasks, but rather a certain complex combination of them that is likely to be interrelated or synergistics. A single task approach cannot take into account the complex interactions between tasks.
- Cognitive tasks, even when specifically designed to target a particular cognitive function, might also have side-effects on other cognitive functions (Cicerone et al., 2011). This makes it difficult to examine the isolated effect of a single task in a specific cognitive function, and no clear evidences appear when all tasks are integrated into a classical model.
- The additional effect of a single task might be affected by the cumulative effect of the sequence of previous tasks executed under the treatment. This might determine that order of execution is relevant in the treatment.
- The effect of a single task may be too subtle to be detected, whereas the effects of the whole CR treatment may be sufficient to be detectable, given the cumulative effect of rehabilitation already mentioned.

From a structural point of view, these characteristics parallel those holding in nutritional epidemiology, where a global approach has been adopted in recent years, and all nutrients are analyzed together due to the high degree of interaction between them. This suggests we should analyze the overall CR treatment as a whole, by considering all kinds of interactions among tasks together, instead of using the traditional single task approach. Thus, in this work, CR treatment will be considered as a sequence of cognitive tasks and data mining methods will be used to determine the multivariate associations between a CR treatment (or relevant subsequences) and the degree of response of the patient. Analyzing CR tasks as treatment patterns offers an innovative perspective in neurorehabilitation, and describing their relationship with their clinical outcome provides a practical approach for evaluating the effects of rehabilitation treatments. It can also enhance our conceptual understanding of CR treatments practice, and might be useful in providing guidance for cognitive treatment interventions.

In this particular field, because of the cumulative effect of the tasks mentioned above, it is reasonable to think that the effect of a certain sequence of CR tasks can respond robustly to slight variations of the sequence. Thus, small variations in the sequence of tasks performed might keep the global effect of the treatment unaltered. This means that the model to be built should admit a certain level of variations around every relevant pattern. These characteristics have already been encountered in the bioinformatics fields, particularly in the transcription factor binding sites (TFBS) field, where slightly different sequences of DNA are associated with a certain biological function. In this field, motif discovery or motif finding methods are used to represent these weak patterns. Analogously, motif discovery methods will be introduced in our proposed methodology to identify patterns of CR treatments, where slight variations in the treatment program might be packed in a single CR motif with a similar therapeutic effect, and might be associated with a certain response level.

1.4.6 Zone of Proximal Development

In the early 1930s, Vygotsky introduced the concept of Zone of Proximal Development (ZPD) in the field of child learning, being the distance between the actual capacities of the child by himself and their potential capacities when being guided (Vygotsky, 1978). In 1986, Cicerone and Tupper transferred ZPD ideas to the neurorehabilitation field by introducing the zone of rehabilitation potential (ZRP) (Cicerone & Tupper, 1986), i.e. the zone in which maximum recovery of cognitive functions might occur, provided that the proper help is given to the subject. They propose the use of ZPD as a guiding principle in CR. This zone is supposed to reflect the patient's region of potential restoration thanks to cognitive plasticity (Calero & Navarro, 2007). Current neurorehabilitation practice tries to design therapeutic plans that keep the subject working in this area during treatment. However, determining when the patient works in ZPD or not is still an open issue. Thus in most cases CR therapists design CR plans from scratch, determining clinical settings for specific patients based mainly on their own expertise. Each specific plan evolves according to each therapist's own criteria and evaluation of the patient's follow-up. There is as yet not enough in-field knowledge regarding which specific intervention (task or exercise

assignation) is more appropriate to help CR therapists design their clinical therapeutic plans.

Learning is enhanced when the match between the skills of the learner and the challenges of the subject matter are optimized (Whalen, 1998). Csikszentmihalyi's Flow Theory (Csikszentmihalyi, 1991) provides a framework and vocabulary for understanding the experiential nexus between the active person and the facilitative environment. The experience of Flow creates information that melds actor and activity into one transactive system. In this sense, Flow may be seen as the experiential dimension of the ZPD (Whalen, 1998).

Flow or optimal experiences, also referred to as the zone (Csikszentmihalyi, 1991) represents a state of consciousness where a person is so absorbed in an activity that he or she excels in performance without consciously being aware of his or her every movement. According to Luria's theory outlined above, repeated taxing of the same neurological system facilitates and guides the reorganization of the targeted cognitive function. This approach requires implementation of repetitive exercises within the planned program which require patients to use their impaired cognitive skills at a productive level.

The emergence of serious games broadens the discipline of entertainment-education in numerous dimensions. Serious games have recently been applied in diverse areas, e.g. military training, health, higher education, city planning (Rego, Moreira, & Reis, 2010). Prior research demonstrates that videogame attributes, such as task difficulty, realism, and interactivity, affect learning outcomes in game-based learning environments (Orvis, Horn, & Belanich, 2008). These prior works suggest that in order to be most effective, instructional games should present an optimal level of difficulty to learners. This optimal range of difficulty is aligned with the Vygotsky's concept of ZPD, where training should be difficult for the learner, but not beyond his or her capabilities.

1.5 Thesis Structure

Chapter 1. Introduces and motivates the general problem, current approaches, the actual limitations, and the specific context and background of this research and the thesis structure.

Chapter 2. Presents the problem formal definition, the general objectives and the CMIS methodology, as a high level umbrella of 5 steps that combine several contributions of this thesis.

Chapter 3. Addresses the bibliography review, initially underlining the limitations of traditional data-driven methods in our context of application, reviews state of the art in the context of repeated activities search patterns as well as optimization problems related to our proposed methods.

Chapter 4. Proposes a new data mining methodology (SAIMAP) which combines tools from pre-processing, clustering, patterns identification, visualization, and post-processing.

Chapter 5. Addresses the problem of identification of the general pattern associated with a motif, which at different lengths are analyzed leading to a general treatment pattern represented as a regular expression.

Chapter 6. The NeuroRehabilitation Range (NRR) is introduced in this chapter. Data mining techniques are used to build data-driven models for NRR. The Sectorized and Annotated Plane (SAP) is proposed as a visual tool by which to identify NRR, and two data-driven methods to build the SAP are introduced. Limitations of proposed methods are analyzed and we then build on the concept of NRR and SAP tools to present and solve a new problem, i.e. the NeuroRehabilitation Range Maximal Regions problem (NRRMR).

Chapter 7. This chapter formalizes the process of determining the improvement of an individual in the several areas of impact after execution of a given task in NRR.

Chapter 8. The final treatment design proposal is presented. The regular expression associated to the recommended treatment is used as a general frame to be instantiated by specific tasks maximizing improvements.

Chapter 9. This chapter presents the application of the proposed methods in a clinical context: the Neuropsychology Department of the Acquired Brain Injury Unit at Institut Guttmann Neurorehabilitation Hospital (IG) where TBI patients undergo CR treatments

Chapter 10. Conclusions, discussion, and limitations of the proposed approaches are outlined as well as future lines of research.

1.6 Summary

This chapter introduces, contextualizes and motivates the general medical problem that initially motivated the thesis from both medical and technical points of view. The relevance of the addressed medical problem is highlighted, being TBI responsible for more years of disability than any other cause. TBI is currently treated through Cognitive rehabilitation (CR). While cognitive rehabilitation task repetition is not the only important factor, it is becoming clear that neuroplastic change and functional improvement only occur after a number of specific tasks are performed in a certain order and number of repetitions and does not occur otherwise. However there are not enough scientific evidences yet to establish clear guidelines to design the specific sequences of neurorehabilitation tasks to be prescribed to each single patient. This chapter remarks the current limitations in the area and highlights the need of new data driven approaches to better understand patient´s rehabilitation process, with the aim of obtaining scientific evidence about its effectiveness and improve prescription of individual treatments. The Information Technology framework supporting CR treatments in our clinical setting is described (the PREVIRNEC© platform). Finally this chapter presents the specific research framework where this thesis took place: a collaboration framework between Dr. Karina Gibert from Universitat Politècnica de Catalunya-BarcelonaTech and Institut Guttmann-Neurorehabilitation Hospital, at the intersection of the Neurorehabilitation of Acquired Brain Damage Research Line and the QVidLab (Laboratory for Enhancing Measures of Autonomy, Personal Satisfaction and Quality of Life of People with Neurological Disabilities) Program of Institut Guttmann.

.

Chapter 2. Problem Formulation, Objectives and Methodological Proposal

As presented in Chapter 1 the general clinical motivation of this thesis is to support the design of CR programs. This requires an understanding of the general process and particularly the relationship between CR tasks and cognitive improvements of the targeted functions measured by standardized tools (such as NAB introduced in section 1.4.3).

CR are a particular case of finding patterned sequences of activities that interact among them. Thus, from the methodological point of view, the main objective of this thesis is

> *to provide a formal frame to find the right sequence of interacting activities to be followed for a maximum improvement in multiple areas.*

To this purpose a first effort to analyze the structure of the problem and formalize it, has been done and some specific objectives have been faced to approach several aspects of the problem as pieces of a general methodology presented in Chapter 4 as the thesis proposal.

2.1 Problem Formulation

Given

$I = \{i_1 \dots i_n\}$ a set of individuals

$T = \{T_s \ s=1:\mathcal{T}\}$ a set of activities (or tasks) that can be executed by any individual

$\mathcal{A} = \{A_1, A_2, \dots, A_a\}$ set of impact areas such that each task impacts in a certain area of $\mathcal{A}$

$f: T \rightarrow \mathcal{A}$ a function that relates an activity with its area of impact: $f(T_s) = A_j$, $s=1:\mathcal{T}$ j=1:a, being A_j the area of impact of activity T_s

Given an scenario in which each individual *i* executes a sequence of t_{f_i}activities, one at each time $t = 1..t_{f_i}$. as shown in Figure 2.1

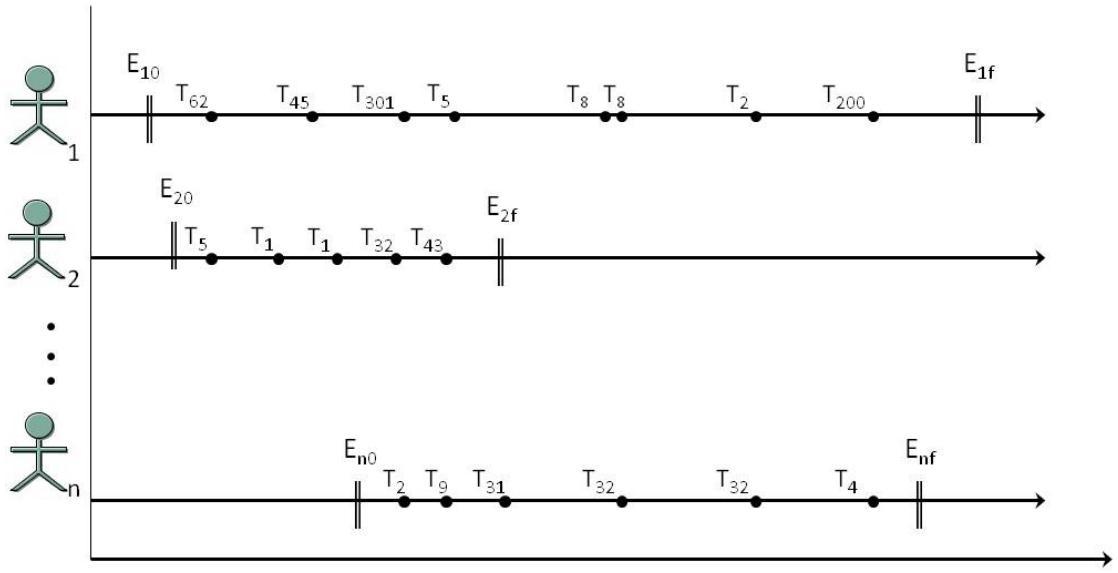


Figure 2.1. Representation of individuals executing sequences of activities impacting areas which are evaluated before and after the period of executions.

Given i, the matrix R_i provides the list of all his executions (runs):

$R_i = [i, T, t]_{(t_{f_i}, 3)}$

Matrix R represents the total set of activities executed by all individuals

$$R = \begin{bmatrix} [R_1] \\ [R_2] \\ \vdots \\ [R_i] \\ \vdots \\ [R_n] \end{bmatrix}_{(\rho, 3)}$$

being $\rho = \sum_{i=1}^{n} t_{f_i}$ the total number of activities' executions performed by all individuals.

On the other hand, the sequence of activities executed by an individual *i* on time $t = 1..t_{f_i}$is $s_i = (T_{i1}, \dots, T_{it}, \dots, T_{if_i})$. In fact, s_i is $s_i = R_i[2]^T$. The longest sequence having length $M_t = max_{i=1..n}\, t_{f_i}$

$\forall\, s_i$ the vector v_i can be defined: $v_i = (n_{i1} \dots n_{iT})$, n_{iT} = number of repetitions of task *T* performed by individual *i* in the total sequence s_i

$\chi = [T_{it}]_{(n,M_t)}$ with $i = \{1..n\}, t = \{1..t_{f_i}\}$, is a matrix where each row indicates the sequence of activities performed by individual *i*. Note that this might not be a rectangular table, as each row has length $t_{f_i} \leq M_t, \forall i = 1..n$.

$N = \begin{bmatrix} [v_1] \\ \vdots \\ [v_n] \end{bmatrix}$ contains the profile of the sequence performed by *i* in terms of number of tasks of each type

$A_{it} = f(T_{it})$ is the area of impact of activity T_{it} executed by individual *i* in time *t*

In the general scenario

$$F = \begin{array}{c} \\ 1 \\ \dots \\ \mathcal{T} \end{array} \begin{array}{c} \begin{array}{ccc} A_1 & \dots & A_a \end{array} \\ \begin{bmatrix} F_{SA} & \dots & F_{SA} \\ \dots & \dots & \dots \\ F_{SA} & \dots & F_{SA} \end{bmatrix} \end{array}$$

$$F(s, A) = \begin{cases} 1 \; if \; task \; s \; impacts \; area \; A \\ 0 \; \; otherwise \end{cases}$$

In the case of tasks impacting multi areas f is the main or primary area impacted by each task

$s_i^a = (A_{i1}, \dots, A_{it}, \dots, A_{tf_i})$ is the sequence of areas impacted by the activities executed by individual *i* in the period [1, t_{f_i}], being $A_{it} \in \mathcal{A} \; \forall \; t = 1..\; t_{f_i}$.

$\chi^a = [A_{it}]_{(n,M_t)}$ with $i = \{1..n\}, t = \{1..t_{f_i}\}$, is a matrix where each row indicates the areas of impact of the sequence of activities performed by individual *i*.

$\chi^p = [p_{it}]_{(n,M_t)}$ with $i = \{1..n\}, t = \{1..t_{f_i}\}$, p_{it} performance of execution of task *t* by individual *i* in sequence s_i

$Y_{1t}, \dots Y_{at}$ a set of numerical indicators of performance for individuals in each area of impact

Y_{jt} measures the global performance obtained of individual *i* in the Area of impact $A_j \in \mathcal{A} \;\; j = 1..a$ in a certain time point *t.*
$E_0 = (Y_{10} \dots Y_{a0})$ evaluates the performance levels of individuals in the different areas of impact before executing their sequence of activities.

$E_f = (Y_{1f} \dots Y_{af})$ evaluates the performance levels of individuals in *I*, in the areas of impact in $\mathcal{A}$ after executing their corresponding sequence of activities described in χ.

$D_j = Y_{j0} - Y_{jf}$ evaluates the effect of χ in the performance levels of A_j. Note that a global effect of the whole sequences is measured, taken into account that several activities in the sequence might impact on the same area. Ideally Y_{jf} will be an implicit or explicit function of all those activities impacting A_j independently of their position in the particular sequence, due to the cumulative effect of activities discussed above.

Assuming that 0 indicates best performance,

$D_j > 0$ indicates improvement

$D_j \leq 0$ indicates non-improvement

Depending on the particular application, other semantics might also be assigned to the values of the performance indicators as well, and this will require reinterpretation of values of the D_j variables accordingly.

$\Delta = (D_1 \ldots\ldots D_a)$ provides the effect of χ over each area of impact.

$X = (X_1 \ldots X_K)$ additional information over individuals X_K might be either numerical or qualitative

Being $\mathcal{B}$: Boolean expression build over χ^a , $\mathcal{L}$: Label; $KB = \{ r: \mathcal{B} \rightarrow \mathcal{L} \}$ is a Knowledge base composed by a set of rules partially expressing the a priori knowledge in the domain. It is important to note here that no assumption of completeness is imposed over *KB*.

Eventually, a binary variable Z might be available for model assessment, indicating the success of an individual performing its sequence of activities under a certain criterion of performance,

$$Z = \begin{cases} YES, successful\ performance \\ \\ NO, unsuccessful\ performance \end{cases}$$

Eventually Z might be a multidimensional vector and each component might be a function of some Δ component.

Also, h_i provides the scoring *p* and improvement z associated to the execution of *T* by individual *i*
$h_i = [i,T,p,z]$ and V provides this information of all executed tasks

$$V = \begin{bmatrix} [h_1] \\ \cdots \\ [h_i] \\ \cdots \\ [h_n] \end{bmatrix}$$

Under all these premises, it is desired to find:

> *a sequence of activities* $\left[a_1, a_2, \ldots, a_{n_\mu}\right]^n$ *with* $a_l \in \mathcal{A}$, *l: 1... n_μ such that the global effect of the sequence over the whole set of impact areas* $\mathcal{A}$ *leads to a successful performance.*

2.2 Possible Instances of the Thesis Problem

This general problem responds to the structure of many different real scenarios where the proposed methodology might help. Here some examples are presented. All of them can be treated as particular cases of the stated problem. See Table 2.1 on how to instantiate the proposed methodology to the different cases.

CR program. Find the best way to combine CR tasks in a sequence to configure the CR program of a TBI patient. Such tasks might stimulate different cognitive functions like Attention, Memory or Executive functions or several of them simultaneously. Repetition of tasks produce cumulative therapeutic effects even if there are other tasks in the middle. Improvement of the patient is assessed using a standard battery of assessment tests (NAB battery presented in section 1.4.3).

Primary Education. Find the best way to combine mathematical drills in an educational program for a specific math course. Each drill might work different students' skills like Logics, abstraction, mental calculus, algebraic structures, … or several of them simultaneously. Repetition of drills produce cumulative formative effects even if there are other drills in the middle (in this particular scenario it might be considered that repetition of a drill means several instances of the same problem with different numbers). Improvements of students' skills is evaluated through exams.

Physical training to prevent frailty. Find best way to combine physical training exercises that can be performed following assisted videotutorials such as Condition Coach (CoCo) platform (Roessing, 2014) addressing different activities (e.g. aimed for coordination, strength, endurance, balance) to prevent frailty in elderly population. Standard physical assessment batteries are applied to evaluate improvements in the different muscular groups targeted by exercises (e.g. grip strength, arm-hand coordination, knee).

	CR program	**Primary Education**	**Physical training**
I	TBI patients in CR	Children at primary school	Elderly population
T	CR tasks patients execute during treatment	Educational problems based learning program for mathematics course	Physical exercises for frailty prevention
$\mathcal{A}$	cognitive functions targeted by tasks (attention, memory, executive functions PREVIRNEC platform)	Students skills in Logics, Numerical , Algebra, Calculus, students skills in logics	Muscular groups targeted exercises (e.g. grip strength, arm-hand coordination, knee, CoCo platform)
R_i	Sequence of CR tasks executed by patient	Sequence of math drills executed by children in class or at home	Sequence physical exercises executed by old adults
Y	NAB battery	Examinations	Jamar dynamometer
X	numerical (age, GCS, days in PTA) or qualitative (gender or Educational level).	Characteristics of students, age, family characteristics, other related courses scores	Age, blood pressure
Z	Global cognitive improvement criteria	Evaluation of marks along course examinations	Global physical assessments scales

Table 2.1 Instantiation of proposed methodology to different application domains

2.3 The CMIS Methodology for Designing Cumulative Multiple Impact Sequences

In order to identify sequences of activities such that the global effect of the sequence over the set of impact areas leads to successful performance, the general method CMIS is proposed as a sequence of steps where each one is detailed in the specific section. The notation introduced in Section 2.1 is assumed.

1. SAIMAP methodology (see details in Chapter 4)

First of all, given the sequences of activities performed by each individual R, a reduced set of characteristics motifs $\mathcal{M}$ are found to profile the sequences followed by a small number of groups of individuals who behave similarly.

Input: R, f, E_0, E_f, F

Output:
a set of patterns $\mathcal{M}$ describing the behavior of the individuals when executing activities
$\mathcal{M}=\{\mu_1, \mu_2, \ldots \mu_m\}$ $\forall$ μ $\in \mathcal{M}$ μ is a sequence of impact areas of variable length (always lower than Mt). Thus, each pattern μ is expressed as:
$\mu = (a_1, a_2, \ldots, a_{n_\mu})$ with $a_l \in \mathcal{A}$ *l:* 1... n_μ
Such that

- $\forall$ μ , μ′ $\in \mathcal{M} : \mu \neq \mu'$
- $\forall$ i $\in$ I, $\exists$ μ $\in \mathcal{M}$: μ is a subsequence of s_i
- $\forall \mu' \in \mathcal{M}$, $\mu' \neq \mu$, μ is a not subsequence of s_i
- Thus, $\mathcal{M}$ inducing a partition P over I. Being P = $\{I_{\mu_1} \ldots . I_{\mu_m}\}$,

$I_\mu = \{i$: μ is a subsequence of $s_i\}$

2. Identification of the general patterns execution global pattern (See details in Chapter 5)

Provides a scheme of contiguous positions of the sequences where a particular area has to be impacted

Input: $\mathcal{M}$

Output: $S=\{\mathcal{S}_1,..,\mathcal{S}_{\mathcal{M}}\}$ such that $\forall$ μ $\in \mathcal{M}$ $\mathcal{S} \in S$ is the general pattern associated to the motif μ $\in \mathcal{M}$ in the form of a regular expression (see Chapter 5) $([A]^r,)^* A \in \mathcal{P}(\mathcal{A})$

3. Identification of NRRs for each task (See details in Chapter 6)

For the whole set of tasks a Neurorehabilitation range is induced from data indicating the conditions in which the task need to be executed to be therapeutic

Input: Vmatrix

Output: $\mathcal{K}_{NRR}$ knowledge base and NRR matrix

4. Evaluate Improvements on each area of impact (See details in Chapter 7)

Find the impact of each task over an area according to the number of times executed by each patient, the areas impacted by each task (*F*) and the improvement of each patient in each area of impact (Δ)

Input: χ, Δ, *F*

Output: Υ

5. Treatment design (See details in Chapter 8)

Give the general pattern of the sequences and the observed impact of tasks executed under NRR in the different areas, compose a sequence of tasks that fits the patterns in NRR conditions

Input: Υ, S, NRR

Output: The list of programs

2.4 Summary

This chapter presents the formal definition of the thesis problem as well as the general and the specific objectives of this work. Although understanding CR treatment patterns was the medical problem motivation of this research, after a deep analysis of the structure of the problem behind, a generic methodological problem was identified which constitutes the main problem of this thesis. Thus, from the methodological point of view, this thesis pretends to provide a formal frame to find the right sequence of interacting activities to be followed for a maximum improvement in multiple areas of impact, provided that tasks can interact among them, produce cumulative effects by repetition (even under a discontinuous pattern) and simultaneously impact, each of them, several impact areas. These premises are much more general than classical approaches where independence, contiguity, non cumulative effects of pure tasks impacting a single area at a time use to be assumed. This

opens a broad scope of complex real problems that can be analyzed under the proposed CMIS methodology which previously were difficult to model, from a global perspective, as illustrated in section 2.2.

The design of CR programs appears therefore as a particular case of the general problem faced in the thesis about finding patterned sequences of activities in front of high order interactions either among tasks and impaired areas.

Section 2.3 presents the general Cumulative Multiple Impact Sequences (CMIS) methodology to solve the described thesis problem. The CMIS methodology is presented as a high level umbrella of 5 innovative steps that combine several contributions of this PhD thesis to identify the best sequence of tasks to be recommended to a specific individual, according to his profile and what performed better in similar situations. The CMIS relies on two main contributions described in later chapters. SAIMAP (section 4.1) and NRR (Chapter 6), both introduced in this research for the first time.

SAIMAP takes as input the sequences of activities performed by each individual and a reduced set of characteristic treatment motifs are found to profile the sequences followed by groups of individuals who behave similarly. The second phase takes as input the previous set of motifs and returns a regular expression representing them (section 5.1). At the third phase Neurorehabilitation Range (NRR) is induced from data indicating the conditions in which the task need to be executed to be therapeutic (or effective), for the whole set of tasks. NRR concept is introduced, to determine the degree of performance expected for a CR task and the number of repetitions required to produce maximum rehabilitation effects on the individual. The fourth phase (section 7.1) aims to evaluate improvements on each area of impact by means of standardized assessment tools. Last phase (section 8.1) composes a specific sequence of tasks that fits the given general pattern obtained from second phase and maximizes expected impact of the total sequence (according to an optimization of the expected improvement function that takes into account NRR).

Chapter 3. State of the Art

As presented in section 1.3.1, this thesis is framed within Institut Guttmann's Analysis of Results applied to the generation of knowledge research Program. Before our attempts to extract useful knowledge from data, it is important to present the overall approach, the Knowledge Discovery in Databases (KDD) process and its elements, such as Data Mining (DM). Afterwards some particularities of the application of DM in the field of neurology are addressed and relevant applications on TBI are reviewed. Related problems such as finding patterns in sequential data and specific applications, as well as traditional approaches on our field of application are reviewed in this chapter.

3.1 Knowledge Discovery in Databases

This section focuses on describing and explaining the process that leads to discovering new knowledge. It defines a sequence of steps (with eventual feedback loops) that should be followed to discover knowledge (e.g. patterns) in data. Each step is usually realized with the help of available commercial or open-source software tools as will be shown in the application chapters.

Since the 1990s, several KDD processes have been developed. The initial efforts were led by academic research and were quickly followed by industry. The basic structure of the model proposed by Fayyad et al (Fayyad, Piatetsky-Shapiro, & Smyth , 1996) is the one proposed in this thesis. The process consists of multiple steps that are executed in a sequence. Each subsequent step is initiated upon successful completion of the previous step and requires the result generated by the previous step as its input.

KDD is defined (Fayyad, Piatetsky-Shapiro, & Smyth , 1996) as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Here, data are a set of facts (for example, cases in a database) and pattern is an expression in some language describing a subset of the data or a model applicable to the subset. Hence, in our usage here, extracting a pattern also designates fitting a model to data;

finding structure from data or, in general, making any high-level description of a set of data.

The Fayyad et al. KDD model consists of nine steps, which are outlined as follows:

1. Developing and understanding the application domain. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.

2. Creating a target data set. Here the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.

3. Data cleaning and preprocessing. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.

4. Data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

5. Choosing the data mining task. Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.

6. Choosing the data mining algorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.

7. Data mining. This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, etc.

8. Interpreting mined patterns. Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.

9. Consolidating discovered knowledge. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to

the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge.

A very important consideration in the Knowledge Discovery Process is the relative time required to complete each step. It includes reviews of partial results, possibly several iterations, and interactions with the data owners. In general, we acknowledge that the data preparation step is by far the most time-consuming part of it.

Given these notions, we can consider a pattern to be knowledge if it exceeds some interestingness threshold, which is by no means an attempt to define knowledge in the philosophical or even the popular view. As a matter of fact, knowledge in this definition is purely user-oriented and domain-specific and is determined by whatever functions and thresholds the user chooses.

3.1.1 Data Mining

While KDD refers to the overall process of discovering useful knowledge from data, data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. The term *Data Mining* was introduced relatively recently, in the mid-1990s, although data mining concepts have an extensive history. Data mining covers areas of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas. All of these are concerned with certain aspects of data analysis. As a result, they have much in common but each also has its own distinct problems and types of solution. The fundamental motivation behind data mining is autonomously extracting useful information or knowledge from data stores or sets. The goal of building computer systems that can adapt to special situations and learn from their experience has attracted researchers from many fields, including computer science, engineering, mathematics, physics, neuroscience, and cognitive science.

Unlike the majority of statistics, data mining typically deals with data that have already been collected for some purpose other than the data mining analysis. The majority of the applications presented in this chapter use data formerly collected for any other purposes. Data mining research has led to a wide variety of learning techniques that have the potential to renew many scientific and industrial fields.

The knowledge discovery goals are defined by the intended use of the system. We can distinguish two types of goal: (1) *verification* and (2) *discovery*. With *verification*, the system is limited to verifying the user's hypothesis (Piatetsky-Shapiro , Brachman , Khabaza , Kloesgen , & Simoudis , 1996). With *discovery*, the system finds new patterns autonomously. We further subdivide the discovery goal into *prediction*, where the system finds patterns for predicting the future behavior of some entities, and *description*, where the system finds patterns for presentation to a user in a human-understandable form. In this thesis, we are primarily concerned with discovery-oriented data mining.

3.2 Data Mining in Neurology

DM can be limited in a medical context in general and particularly in neurology by several factors. One of them is the accessibility to data that often is distributed in different settings (clinical, administration, insurers, labs, etc.). Besides, data may be incomplete, corrupt, noisy, or inconsistent. Ethical, legal, and social issues (data ownership, privacy concerns) may also arise. Many patterns found in DM may be the result of random fluctuations and therefore many such patterns may be useless. DM of medical data requires specific medical knowledge as well as knowledge of DM technology. DM requires institutional commitment and funding. Another unique feature of medical data mining is that the underlying data structures of medicine are poorly characterized mathematically, as compared to many areas of the physical sciences. Physical scientists collect data which they can put into formulas, equations, and models that reasonably reflect the relationships among their data. On the other hand, the conceptual structure of medicine consists of word descriptions and images, with very few formal constraints on the vocabulary, the composition of images or the allowable relationships among basic concepts. The fundamental entities of medicine, such as inflammation, ischemia or neoplasia, are just as real to a physician as entities such as mass, length or force are to a physical scientist; but medicine has no comparable formal structure into which a data miner can organize information, such as might be modeled by clustering, regression models or sequence analysis.

Another related issue is the lack of canonical form in medical terms, in mathematics, a canonical form is a preferred notation that encapsulates all equivalent forms

of the same concept. For example, the canonical form for one-half is 1/2, and there is an algorithm for reducing the infinity of equivalent expressions or aliases, namely 2/4, 3/6, 4/8, 5/10, ... , down to 1/2. Agreement upon a canonical form is one of the features of any mature intellectual discipline.

Unfortunately, in biomedicine even elementary concepts have no canonical form. For example, the canonical form for even a simple idea such as: *adenocarcinoma of colon, metastatic to liv*er has no consistent form of expression. Naturally the individual medical words all have a unique spelling and meaning; but there are a number of distinct expressions (and many others, easy to imagine) such as the following that are all medically equivalent: Colon adenocarcinoma, metastatic to liver; Colonic adenocarcinoma, metastatic to liver; etc.

In our previous research (Gibert K. , García-Rudolph, Curcoll, Pla, & Tormos, 2009) an integral Knowledge Discovery Methodology (clustering based on rules by states) which incorporates artificial intelligence (AI) and statistical methods as well as interpretation-oriented tools, is used to extract knowledge patterns about the evolution over time of the Quality of Life (QoL) of patients with Spinal Cord Injury. The methodology incorporates the interaction with experts as a crucial element with the clustering methodology to guarantee the usefulness and interpretability of the results.

3.3 Mining in Traumatic Brain Injury

A number of studies employ traditional DM techniques in TBI such as Classification (K-Nearest Neighbor, Decision Tree, Support Vector Machines, Neural Networks, and Bayesian Methods), Regression, Clustering, Association Rules, and Sequential Patterns. Most are used to anticipate the treatment's outcome from the usual course of the disease and/or the peculiarities of each individual case. Overall, these studies focus on determining survival, predicting gross outcome, and/or identifying predictive factors of a patient's condition after TBI (usually acute TBI). As yet there is no consensus on an optimal method. Thus, different approaches have been explored (Theodoraki, Katsaragakis, Koukouvinos, & Parpoula, 2010). The sections below review some of them.

3.3.1 Classification Techniques

The problem of data classification attempts to learn the relationship between a set of feature variables and a target variable of interest. Many practical problems can be expressed as such relationships involving features and target variables, thus providing a broad range of applications (Aggarwal C. , 2014) as shown in the following sections for our specific domain. Classification is learning a function that maps (classifies) a data item to one of several predefined classes (Weiss & Kulikowski, 1990). A wide variety of classification methods exist, but below we present a subset of popular techniques applied in subsequent chapters of this work.

Decision Trees: Most decision tree learning algorithms are variations on a top-down greedy search algorithm, with the most notable example being ID3 (Iterative Dichotomizer 3) by Quinlan (Quinlan, 1986). Quinlan references Hunt's Concept Learning System (CLS) (Hunt, 1962) as inspiration and a precursor to ID3. Hunt's Concept Learning System was a divide and conquer scheme that could handle binary (positive and negative) target values, with the decision attribute being decided by a heuristic based on the largest number of positive cases. DTs offer a series of advantages: they are self-explanatory and when compact, they are also easy to follow. In other words if the decision tree has a reasonable number of leaves, it can be grasped by non-professional users. Furthermore decision trees can be converted to a set of rules. Thus, this representation is considered as comprehensible. Among traditional classification techniques, decision trees are the most common choice mostly because of these advantages.

Since 1993 a number of publications considered different variables to address the problem of outcome prediction. In one of the first studies (Pilih, Mladenić, Lavrač, & Prevec, 1997) DTs are considered to be useful for the analysis of the importance of clinical parameters and of their combinations for the evaluation of the severity of brain injury and for outcome prediction. Due to a small number of patient data available for this study the induced DTs cannot yet be considered as a reliable prognostic tool.

Presented as an alternative to RCT, the analysis of existing patient data is proposed in (McQuatt, Sleeman, Andrews, Corruble, & Jones, 2001) as an attempt to predict the several outcomes and to suggest therapies. It uses DT techniques to predict the outcome of

head injury patients. The work is based on patient data from the Edinburgh Royal Infirmary which contains both background (demographic) data and temporal (physiological) data.

In (Brown, 2006) the primary goal was to consider all clinical elements available concerning a survivor of TBI admitted for inpatient rehabilitation, and identify those factors that predict disability (n = 3463). Predictor variables included all physical examination elements, measures of injury severity (initial Glasgow Coma Scale score, duration of post-traumatic amnesia (PTA), length of coma, CT scan pathology), gender, age, and years of education. The duration of PTA, age, and most elements of the physical examination were predictive of early disability. The duration of PTA alone was selected to predict late disability and independent living.

Similar analysis has been performed in (Chesney, y otros, 2009) where trauma injury data collected over 10 years at a UK hospital are analyzed. The data include injury details such as patient age and gender, the mechanism of injury, various measures of injury severity, management interventions, and treatment outcome. The data mining algorithm C5.0 was also used to determine those factors in the data that can be used to predict whether a patient will live or die. In this case, C5.0 shows with 77% accuracy that gender and whether the patient was referred from another hospital is important for outcome prediction.

In (Garcia, Martins, & Azevedo, 2013) a C4.5 algorithm was used to analyze severe TBI, for the purposes of identifying a model of death prediction. The database consisted of 748 records, each of which has 18 attributes that represent the characteristics related to TBI (e.g. Glasgow Comma Scale, Marshall Classification, Type of Associated Trauma, Age, Cause of TBI, Sex). Outcome prediction was classified by C4.5 algorithm with an accuracy of 87%, including combinations of indicators that lead to survival and death.

Artificial Neural Networks: An NN when used for classification is typically a collection of neuron-like processing units with weighted connections between them. To solve a particular problem, NN uses neurons which are organized processing elements (Dunham, 2003). An NN is adaptive in nature because it changes its structure and adjusts its weight in order to minimize the error network (Silver, y otros, 2001). Adjustment of weight is based on the information that flows internally and externally through the network during the

learning phase. In NN multiclass, problems may be addressed by using a multilayer feed forward technique in which neurons are employed in the output layer rather than using one neuron.

In one of the first attempts to apply ANN (Lang, Pitts, Damron, & Rutledge, 1997) conclude that outcome (dead versus alive) at 6 months after severe head injury can be predicted with logistic regression or ANN models based on data available 24 hours after injury.

Most subsequent attempts focused on dichotomous result, such as alive vs. dead. In order to predict more specific levels of outcome, (Min-Huei, Yu-Chuan, Wen-Ta, & Ju-Chuan, 2005) was conducted to determine if ANN modeling would predict outcome in five levels of Glasgow Outcome Scale (death, persistent vegetative state, severe disability, moderate disability, and good recovery) after moderate to severe head injury.

One approach to predicting an ICU patient's severity level at ICU discharge (or death, in some cases) could be to use both admission data and additional clinical data as they become available on subsequent days in the ICU to serve as inputs into ANN for developing a prediction model (Crump, Silvers, Wilson, Schlachta-Fairchild, & Ashley, 2014).

In (Güler, Gökçil, & Gülbandilar, 2009) a diagnostic system to detect the severity of traumatic brain injuries was developed using artificial neural networks. Three layered back propagation neural networks were used, with an input layer of 10 nodes whose output provided the inputs to a hidden layer. Thirty-two patients with TBI of different age and gender were taken in the study. Electroencephalography, Trauma and Glasgow coma scores were used for evaluating the data. The results obtained from the system were compared with the findings of neurologists. A significant relationship was found between the findings of neurologists and systems output for normal, mild, moderate, and severe electroencephalography tracing data.

In (Rughani, y otros, 2010) authors designed an ANN to predict in-hospital survival following traumatic brain injury. For comparison with traditional forms of modeling, 2 regression models were developed using the same training set and were evaluated on the same testing set. The ANN was compared with the clinicians and the regression models in terms of accuracy, sensitivity, specificity, and discrimination. When given the same limited

clinical information, the ANN significantly outperformed regression models and clinicians on multiple performance measures.

In (Gholipour, Rahim, Fakhree, & Ziapour, 2015) ANN were used to predict survival and length of stay of patients in the ward and the intensive care unit (ICU) of trauma patients and to obtain predictive power of the current method. This ANN model was used based on back-propagation, feed forward, and fed by Trauma and Injury Severity Score (TRISS) components, biochemical findings, risk factors and outcome of 95 patients. In the next step a trained ANN was used to predict outcome, ICU and ward length of stay for 30 test group patients by processing primary data. The sensitivity and specificity of an ANN for predicting the outcome of traumatic patients in this study calculated 75% and 96.26% respectively. 93.33% of outcome predictions obtained by ANN were correct.

Support Vector Machines: The concept of SVM introduced by Vapnik (Vapnik, 1998) is based on statistical learning theory. The SVM classifier creates a hyper plane or multiple hyper planes in a high-dimensional space that is useful for classification, regression, and other efficient tasks. In some cases it is difficult to perform separation of data points in the original input space; to make separation easier the original finite dimensional space is mapped into a new, higher dimensional space. Kernel functions are used for non-linear mapping of training samples to a high-dimensional space. Various kernel function such as polynomial, Gaussian, sigmoid etc., are used for this purpose (Cristianini & Shawe-Taylor , 2000).

In (McBride, Zhao, Nichols, & Abdul-Ahad, 2011) Support Vector Machine (SVM) analyses are employed to classify 15 TBI and 15 normal individuals' EEG recordings taken during a working memory test. The features used by the SVM analyses include different sets of event-related Tsallis entropy functionals. The analyses demonstrate a strong correlation between the Event-Related Functionals (ERFs) and the presence of TBI, attaining classification accuracies as high as 90%.

In (Aribisala, y otros, 2010) SVM was applied to classify the quantitative MRI data, in particular quantitative MRI techniques (T1, T2 mapping and diffusion tensor MRI) in 24 mild TBI patients and 20 matched controls. Quantitative MRI data can be used to separate mild TBI patients from the control group. Our results show that SVM can detect changes in

normal appearing tissues in some patients suffering mild TBI as compared with the control group. These changes may represent damage to neuronal tissue and further work is needed to determine whether this is responsible for the cognitive and affective symptoms commonly seen following mild head injury, which include memory loss, inability to concentrate, irritability, and depression.

Bayesian methods: Bayesian approaches employ probabilistic concept representations and range from the Naïve Bayes to Bayesian Networks (Domingos & Pazzani, 1997). The basic assumption of Bayesian reasoning is that the relation between attributes can be represented as a probability distribution (Maimon & Last, 2001). It is usually considered (Maimon & Rokach, 2005) that the most straightforward Bayesian learning method is the Naïve Bayes classifier (Duda & Hart, 1973). This uses a set of discriminant functions for estimating the probability of a given instance belonging to a certain class. More specifically it uses Bayes rule to compute the probability of each possible value of the target attribute given the instance, assuming the input values are conditionally independent given the target attribute. Surprisingly, a variety of empirical research shows that the Naive Bayes classifier can perform quite well compared to other methods, even in domains where clear feature dependencies exist (Domingos & Pazzani, 1997). Furthermore, Naive Bayes classifiers are also very simple and easy to understand (Kononenko, 1990).

This study (Sakellaropoulos & Nikiforidis, 1999) concerns the development and validation of Bayesian Networks for the assessment of prognosis after 24 hours for head-injured patients of the outpatients department in the University Hospital of Patras, Greece. Different selection strategies resulted in BNs with varying structures and prognostic performance.

In (Klement, y otros, 2012) when predicting the need for computed tomography (CT) imaging of children after a minor head injury, an ensemble of multiple Naive Bayes (NB) classifiers was derived as the prediction model for CT imaging decisions in imbalanced data. Naïve Bayes classifiers are specially suited to overcoming the imbalance problem. Imbalance is commonly encountered when analyzing clinical data where the population of patients with a health condition is usually significantly smaller than the population of relatively healthy ones. (Klement, Wilk, Michalowski, & Matwin, 2011)

demonstrates that a model that performs well can be developed by utilizing data undersampling when constructing an ensemble prediction classifier composed of multiple NB classifiers.

Comparative studies: A number of papers propose the comparison of several of the classifiers presented in the previous section in several TBI applications.

Lang and colleagues were the first group to demonstrate that an ANN was as accurate, sensitive, and specific as standard logistic regression in predicting 6-month survival following severe head injury (Lang, Pitts, Damron, & Rutledge, 1997).

Yin and colleagues (Yin et al., 2006) carried out a pilot study on the effectiveness of Bayesian Networks, Decision Trees, Logistic Regression, Support Vector Machines and Artificial Neural Networks on outcome after severe brain injury. The dataset consisted of over seven hundred patients with severe brain injury and estimated the model performance using 10-fold cross validation.

A reliable model predicting the outcome proved to be impracticable, but several aspects to be taken into account for this kind of study were outlined. In particular, the validation techniques for evaluating the realistic prediction reliability of extracted models and then the significant influence of outcome classes aggregation on prediction performance proved crucial. No individual algorithm outperformed the others, and authors suggested the application of multiple algorithms in parallel to reduce errors.

More recently, (Segal, y otros, 2006) have compared an ANN with a multiple regression model in predicting several different functional outcome scores at 1 year after TBI. They showed that linear models performed with the same accuracy as an ANN.

Another group (Eftekhar, Mohammad, Ardebili, Ghodsi, & Ketabchi, 2005) actually showed that an ANN was less accurate than linear regression in predicting survival, although it was marginally better at discriminating outcomes.

In (Pignolo & Lagani, 2011) compared four different machine learning methods (C4.5, SVM, Naïve Bayes and K-NN) to identify the most suitable algorithm in the prognostic evaluation of subjects in a vegetative state. They concluded that all tested algorithms are usable in this respect. SVM models may be a useful clinical tool to exclude a positive outcome. K-NN and C4.5 could be used for the same purpose, but their sensitivity

and specificity are inferior to SVM. The Naïve Bayes classifiers do not appear usable for differential prognosis, due to poor efficiency in recognizing a specific class of subjects, but have limited classification errors and can still be considered as a valid (ancillary) prognostic tool. It may be worth noting that C4.5 remains a tool with potential clinical application in spite of poor performances; it is the only algorithm among those studied to be able to provide graphical models that are user-friendly for the clinician.

The study by (Sujin, Woojae, & Rae, 2011) compared mortality prediction. The authors of this study compared the artificial neural networks, support vector machines, DT, and conventional logistic regression models. The best performance was achieved with the DT model.

In our previous research (Serra, y otros, 2013) based on a set of pre-treatment assessments, distinct classifiers (C4.5, SVM, Naïve Bayes and K-NN) are trained to predict whether the patient will improve in one or any of three cognitive areas: attention, memory, and executive functioning. Results show that variables such as the age at the time of injury, the patient's etiology or the neuropsychological evaluation scores obtained before the treatment are relevant for prognosis and easily yield statistically significant accuracies.

In (Marcano-Cedeño, y otros, 2013) in order to analyze treatment outcome prediction, we also applied and compared three different data mining techniques: the AMMLP model, a backpropagation neural network (BPNN) and a C4.5 decision tree. The prediction performance of the models was measured by ten-fold cross validation and several architectures were tested. The results obtained by the AMMLP model are clearly superior, with an average predictive performance of 91.56%. BPNN and C4.5 models have a prediction average accuracy of 80.18% and 89.91% respectively. The best single AMMLP model provided a specificity of 92.38%, a sensitivity of 91.76%, and a prediction accuracy of 92.07%.

3.3.2 Regression Models

Regression-based methods attempt to explicitly model the relationship between inputs or independent variables and the outputs, typically in the form of parametric equations in which the parameters are estimated from the data. These methods often provide explicit estimates of measures of association between individual inputs and the outcome, adjusted

for other inputs, with standard error estimates provided from the modeling paradigm used (Dasgupta, Sun, König, Bailey-Wilson, & Malley, 2011). The most common class of regression methods in the literature comes from the class of generalized linear models (McCullagh & Nelder, 1989) which includes linear regression, logistic regression, and Poisson regression. Based on a number of independent variables regression is of two types: linear and non-linear. Linear regression identifies the relation of a dependent variable and one or more independent variables. It is based on a model which utilizes linear function for its construction. Linear regression finds out a line and calculates vertical distances of points from the line and minimizes the sum of the square of the vertical distance. In this approach, dependent and independent variables are already known and the purpose is to spot a line that correlates between these variables (Fox, 1997).

Because of the strong association with the initial GCS score and outcomes, a number of investigators have studied the predictive value of the initial GCS score using various logistic regression techniques (Benzer, Traweger, & Ofner, 1995).

The objective of (Hukkelhoven, y otros, 2005) was to develop and validate prognostic models that use information available at admission to estimate a 6-month outcome after severe or moderate TBI. This study evaluated mortality and unfavorable outcome, that is, death, and vegetative or severe disability on the Glasgow Outcome Scale (GOS), at 6 months post-injury. They included seven predictive characteristics: age, motor score, pupillary reactivity, hypoxia, hypotension, computed tomography classification, and traumatic subarachnoid hemorrhage.

(Martins, y otros, 2009) investigated the mortality of Brazilian patients with severe TBI at the time of discharge, using a multiple logistical regression analysis. They analyzed clinical, demographic, radiologic, and neurosurgical variables, and mortality at time of discharge of all consecutive patients (n = 748) with severe TBI.

The aim of this work (Larsson, Björkdahl, Esbjörnsson, & Sunnerhagen , 2013) was to explore the extent to which social, cognitive, emotional, and physical aspects influence participation after TBI. Data were analyzed with logistic regression. As most data were ordinal, non-parametric statistics were used and the logistic regression was chosen as suitable for this kind of data. The analyses gave 5 predictors reflecting emotional and social aspects, which could explain up to 70% of the variation in participation. The study also tells

us that a great deal of the explanation should also be seen as being connected to an interaction between several aspects. The findings will contribute to the body of knowledge, but further studies are needed to be able to improve participation for persons with disability after a TBI.

.

3.3.3 Cluster Analysis

Clustering is different than classification since it has no predefined classes. Clustering has traditionally been studied as a branch of statistics (Arabie & Hubert, 1996) and in natural sciences (Massart & Kaufman, 1983). The general problem of clustering is stated as follows: given a set of data points, partition them into a set of groups which are as similar as possible (Aggarwal & Reddy, 2013).

When facing complex questions, our natural tendency as human beings is to break the subject into smaller pieces each of which can be explained more simply. Therefore clustering can be seen as a preliminary step and once proper clusters have been indentified it is often possible to find patterns within each one (Berry & Linoff, 2004).

In general the major clustering methods fall into one of the following categories, although it is difficult to provide a crisp categorization because a given method might have features from several categories:

Partitioned Clustering: The datasets having *n* data points partitioned into *k* groups or clusters. Each cluster has at least one data point and each data point must belong to only one cluster. Based on the choice of cluster centroid and similarity measure, the partition clustering method is divided into two categories: K-Means (Hartigan, 1975) and K-Mediods (Kaufman & Rousseeuw, 1990). K-means first selects the k-centroid randomly and then assigns the data points to these ‘k’ centroids based on some similarity measure. For every iteration, a data point is handed over to the cluster based on similarity of cluster mean (the distance between the data points) (Hamerly & Elkan, 2003). Unlike K-means, K-medoids use medoids instead of mean to group the cluster. Medoid is one of the most centrally located data point in the database. Initially, the medoids for each cluster are arbitrarily selected and after that data point is grouped with that medoid to which it is most similar.

Hierarchical Clustering: Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a *dendrogram*. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) (Jain & Dubes, 1988). An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved.

Among hierarchical methods here we particularly highlight **Clustering Based on Rules** (ClBR) which combines inductive learning elements with statistical methods to enhance clustering results (Gibert, Aluja, & Cortés, 1998). The main idea of ClBR is to allow the user to introduce constraints on the formation of clusters (classes), providing them in a declarative way. These conditions imposed by experts induce a sort of super-structure on the domain; clustering is performed within this structure respecting the user constraints. It uses an adaptation of the chained reciprocal neighbors algorithm (de Rham, 1980) which is based on the concept of reciprocal neighbors (RN). At every step, a pair of RN is aggregated in a new class.

Density Based Clustering. Most partitioning and hierarchical methods cluster objects based on the distance between objects. Therefore they can handle only spherical clusters and are not suitable for discovering clusters of arbitrary shapes. Density-based methods continue growing a given cluster as long as the density (number of objects or data points) in the neighborhood exceeds a given threshold. There are two major approaches for density-based methods. The first one pins density to a training data point (Density-Based Connectivity) (Ester, Kriegel, Sander, & Xu, 1996) and the second approach pins density to a point in the attribute space (Density Functions) (Hinneburg & Keim, 1998).

One of the first contributions (Crosson, Greene, Roth, Farr, & Adams, 1990) analyses WAIS-R performance in 93 TBI adults. In (Maleca, Machuldaa, & Smigielskia, 1993) neuropsychological test data and educational attainment for 47 patients were also studied. Relationships of cluster membership to injury severity (coma days) and disabilities

(Portland Adaptability Inventory; PAI) were examined. Results of this cluster analysis suggest a model of TBI-related disability which predicts level of chronic disability from pre-injury functioning, injury severity, and impairments in remote memory and adaptive abilities.

In (Fleming, Strong, & Ashton, 1998) the purpose was to investigate the relationship between self-awareness, emotional distress, motivation, and outcome in adults with severe traumatic brain injury. A sample of 55 patients was selected from 120 consecutive patients with severe TBI. A three-cluster solution was selected, with groups labeled as high self-awareness (n = 23), low self-awareness (n = 23), and good recovery (n = 8). Rehabilitation timing and approach may need to be tailored to match the individual's level of self-awareness, motivation, and emotional distress.

Cluster analysis can be particularly useful in identifying profiles of performance on neuropsychological testing that may be related to important disorder-related variables such as treatment outcomes, medication response, and longer-term prognosis. Illustrative of this, (Allen, y otros, 2010) investigated attention and memory heterogeneity in 150 children and adolescents with TBI using the Test of Memory and Learning (TOMAL). Clusters derived from this sample were compared to clusters derived from 150 age- and sex-matched normal controls to determine whether differing patterns of learning, memory, and attention/concentration would be evident among the groups. Also, the TBI clusters were compared on a number of important clinical, cognitive, and behavioral variables, to determine whether cluster membership might be associated with unique patterns of cognitive and behavioral disturbances.

(Thaler, y otros, 2010) also examined WISC-III clusters in 123 children with TBI. Cluster analysis of the WISC-III scores also identified four clusters that were similar in many respects to those identified by (Donders & Warschausky, 1997). Comparisons between the clusters on behavioral ratings generally indicated that the most severely impaired cluster typically exhibited the most severe behavioral disturbances. The samples for these two studies were comparable in many respects. Both studies identified average and low average clusters, as well as a more severely impaired cluster with selective impairment on perceptual organization and processing speed.

In their investigation, (Allen, Thaler, Cross, & Mayfield, 2013) in order to develop severity classifications based on TMT performance, Part A and Part B raw scores (time in seconds) were submitted to hierarchical cluster analysis using Ward's method with squared Euclidean distance as the distance measure. Ward's method of cluster analysis was selected because it is consistent with the cluster analytic methodology of previous studies of neuropsychological variables in children with TBI that examined TMT performance as an indicator of brain injury severity approximately one year following injury in children who sustained a TBI. The TMT clusters correspond in a general way with mild, moderate, and severe classifications, although the best-performing cluster obtained scores that were in the average range.

In a recent work, (Snell, Surgenor, Hay-Smith, Williman, & Siegert, 2015) examined associations between baseline demographic, clinical, psychological variables (distress, injury beliefs and symptom burden) and outcome 6 months later. A two-step approach to cluster analysis was applied (Ward's method to identify clusters, K-means to refine results). Three meaningful clusters emerged (high-adapters, medium-adapters, low-adapters). Baseline cluster-group membership was significantly associated with outcomes over time. Cluster analysis supported the notion that groups could be identified early post-injury based on psychological factors, with group membership associated with differing outcomes over time.

In our previous research (Gibert K. , y otros, 2008) a KDD framework is proposed where first, descriptive statistics of every variable was done, data cleaning and selection of relevant variables. Data was then mined using a generalization (Exogenous Clustering based on rules, EClBR) allowing the KB to be defined in terms of variables that will not be considered in the clustering process itself, to get more flexibility. Several tools as Class panel graph are introduced in the methodology to assist final interpretation. A set of 5 classes was recommended by the system and interpretation permitted profiles labelling. From the medical point of view, composition of classes corresponds closely with different patterns of increasing level of response to rehabilitation treatments.

3.3.4 Association Rules (AR)

Association Rule Mining (ARM) is the process of discovering collections of data attributes that are statistically associated in the underlying data. Association rules "aim to extract interesting correlations, frequent patterns, associations or causal structures among sets of items in the transaction databases or other repositories" (Agrawal & Srikant, 1994)

Apriori Algorithm: This algorithm is based on the principle that if an item does not fulfil a minimum support constraint or is not frequent then its descendants are also not frequent. Therefore this item must be removed from the transaction database because it does not contribute to the construction of association rules. Unlike classification and clustering, efficiency is the evaluation factor of association mining. Various methods are used to improve the efficiency of Apriori algorithms such as Hash table, transaction reduction, partitioning etc., (Agrawal & Srikant, 1994) (Agrawal, Imielinski, & Swami, 1993).

Frequent Pattern Tree Algorithm (FP-Tree): FP-tree algorithm identifies the frequent item sets without generating candidate item set. This algorithm has two steps: in the first step, FP tree data structure is constructed and in the second step frequent item set is fetched from this data structure (Han, Pei, & Yin, 2000).

AR attracts researchers attention mostly in the field of brain imaging. For content-based retrieval, association rules are employed to reduce the dimensionality of the feature vectors that represent the images and to improve the precision of the similarity queries (Ribeiro, y otros, 2009). The method proposed in (Chaves, Ramírez, Górriz, & Illán, 2012) evaluates the reliability of ARs aiming to discover interesting associations between attributes in functional brain imaging, i.e. single photon emission computed tomography (SPECT) and positron emission tomography (PET).

3.3.5 Sequential Pattern Mining

Sequential pattern mining, which discovers frequent subsequences as patterns in a sequence database, is an important data mining problem with broad applications, including the analysis of customer purchase patterns or Web access patterns, the analysis of sequencing

or time-related processes such as scientific experiments, natural disasters, and disease treatments, the analysis of DNA sequences, and so on.

The sequential pattern mining problem was first introduced by Agrawal and Srikant (Agrawal & Srikant, 1995) based on their study of customer purchase sequences, as follows*: Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified min support threshold, sequential pattern mining is to find all frequent subsequences, i.e. the subsequences whose occurrence frequency in the set of sequences is no less than min support.*

Several efficient algorithms have been proposed for sequential pattern mining such as ClaSP (Gomariz, Campos, Marin, & Goethals, 2013) CloSpan (Yan, Han, & Afshar, 2003) GSP. (Srikant & Agrawal, 1996) PrefixSpan (Pei, y otros, 2004) SPADE (Zaki, 2001) and SPAM (Ayres, Flannick, Gehrke, & Yiu, 2002). Sequential pattern mining algorithms can be categorized as using a horizontal database format (e.g. CloSpan, GSP and PrefixSpan) or a vertical database format (e.g. ClaSP, SPADE, SPAM). The vertical format has the advantage of generating patterns and computing their support without performing costly database scans. This allows vertical algorithms (CM_SPADE, CM-SPAM) to perform better on datasets with dense or long sequences than algorithms that use the horizontal format, and to have excellent overall performance (Fournier-Viger, Gomariz, Campos, & Thomas, 2014).

Although sequential pattern mining methods are suitable for our problem, we will see that they do not provide useful results from a clinical point of view. Indeed, sequential pattern mining methods can provide most frequent subsequences in a dataset, and subsequences do not require continguity of elements. Therefore this seems to be a suitable framework to model the slight variations of the patterns required in our problem.

However, the complexity of the solutions space provided by these kind of methods seems to be higher than the one in the original dataset itself and this seems to increase complexity instead of improving understanding about the underlying structure of the problem as will be seen in the application presented below.

Patterns in healthcare domain include the common patterns in paths followed by patients in hospitals, patterns observed in symptoms of a particular disease, patterns in daily activity, and health data (Gupta, 2011).

A recent example of mining in a medical context is the application of the sequential pattern mining algorithms on a database known as the RSU Dr. Soetomo medical database to find sequential disease patterns (Yuliana, Rostianingsih, & Budhi, 2009). However, age and gender were not included into the sequential rules and the author only displayed a selection of rules.

Other existing works aimed at detecting medical sequential patterns tended to focus on time series data (Pradhan & Prabhakaran, 2009) or specific illnesses, such as investigating patterns that predict the onset of thrombosis and identifying traits leading to atherosclerosis in a database of approximately 1400 middle-aged men (Klema, Novakova, Karel, Stepankova, & Zelezny, 2008).

To the best of our knowledge the identification of sequential patterns where a TBI rehabilitation treatment is considered as a sequence of CR tasks has not yet been addressed. In addition, and as stated in the introduction, the methodologies used in related works previously mentioned do not resist sets of variables with cumulative effects among them and a high degree of interaction.

3.4 Motif Discovery in Sequential Data

A motif is a short distinctive sequence pattern shared by a number of related sequences. The distinctiveness of a motif is mainly reflected in the overrepresentation of the motif pattern at certain locations in the related sequences and the underrepresentation elsewhere.

One of the early origins of motif discovery in the context of DNA analysis is the computer program written in 1977 by Korn (Korn, Queen, & Wegman, 1977). Especially relevant to gene activities are regulatory elements bound by proteins such as TFs identification (D'Haeseleer, 2006). Because a single protein often recognizes a variety of similar sequences, motifs are subject to some degree of sequence variation at each motif position without losing their function.

More than a hundred methods (Klepper & Drabløs, 2010) have been proposed for motif discovery in recent years, representing a large variation with respect to both algorithmic approaches as well as the underlying models of regulatory regions. Among them, MEME (Multiple Expectation-Maximization for Motif Elicitation) (Bailey & Elkan,

1995) is one of the best-established motif-finding tools, being quick and accurate enough and with suitable implementations available (Das et al., 2007). MEME searches for motifs by performing Expectation Maximization (EM) on a motif model of a fixed width and using an initial estimate of the number of sites.

A few existing applications can be found of motif discovery methods used to find relevant patterns in non-genetic sequences. In (Burred, 2012) they are applied to acoustic analysis; sounds are first transformed into a sequence of discrete states, and these subjected to the MEME algorithm for motif discovery, searching for repetitive patterns in sounds. In (Jawad, Kersting, & Andrienko, 2011) the relationship between biological sequences and mobility mining are revisited, searching for patterns in traffic sequence data. In (Syed, Stultz, & Guttag, 2010), motifs search is applied to find precursors of acute clinical events regarding electrocardiographic activity.

However, to the best of our knowledge, no works applying motifs to the identification of patterns in CR treatments have been conducted.

3.5 Maximal Empty Rectangle (MER)

Computational Geometry is a subfield of algorithm theory that involves the design and analysis of efficient algorithms for problems involving geometric input and output (Mount, 2002). In this thesis, Computational Geometry is required for automatic identification of some patterns of task performance and an adaptation of a MER algorithm will be used for the development of the NRRMR model. The MER problem consists of recognizing all maximal empty axes-parallel (isothetic) rectangles in a rectangular space region where some points are located. It was first introduced in 1984 (Naamad, Lee, & Hsu, 1984) as follows:

> *Given a rectilinearly oriented rectangle A in the Cartesian plane and a set S ={P1, P2…Pn} of n> 1 points in the interior of A, where each point Pi is specified by its X and Y coordinates (Xi, Yi), i = 1,2, …,n and A specified by its left boundary Al, right boundary A_r, top boundary At and bottom boundary Ab. The maximum empty rectangle (MER) problem is to find a maximum area rectangle whose sides are*

parallel with those of A and which is contained in A such that no point of S lies in its interior.

Several algorithms have been proposed for the planar problem over the years (Dumitrescu & Jiang, 2013). For instance, an early algorithm by Chazelle, Drysdale and Lee (Chazelle, Drysdale, & Lee, 1986) runs in O(nlog3n) time and O(nlogn) space. The fastest known algorithm, proposed by Aggarwal and Suri in 1987 (Aggarwal & Suri, 1987), runs in O(nlog2 n) time and O(n) space. A lower bound of Ω(nlogn) in the algebraic decision tree model for this problem has been shown by Mckenna et al. (Mckenna, O'Rourke, & Suri, 1985).

This problem arises in situations where a rectangular plant is to be located within a similar region that has a number of forbidden areas or when a 'perfect' rectangular piece from a large, similarly shaped metal sheet with some defective spots is to be cut. The problem could also be further modified so that the length and width of the sought-after rectangle have a certain ratio or a certain minimum length.

Maximal empty rectangles also arose in the enumeration of maximal white rectangles in image segmentation (Baird, Jones, & Fortune, 1990).

More recently, applications can be found in data mining (Edmonds, Gryz, Liang, & Miller, 2003) geographical information systems (GIS), and very large-scale integration design (Augustine, y otros, 2010).

3.6 Limitations of Traditional Methods

There are inherent difficulties in appraising the quality of rehabilitation studies by traditional evidence-based methods. This is particularly true of the complex, experience-based treatments that predominate in rehabilitation over medically-oriented treatments such as pharmacotherapy and surgery (Johnston, Sherer, & Whyte, 2006). Interventions that involve explicit teaching, behavior change, and/or environmental manipulations cannot typically be hidden from the patient or the therapist. Thus the removal of bias by using standard blinding procedures, such as placebo treatment, is not straightforward. Unlike medical treatments which may be aimed at specific symptoms, rehabilitation interventions usually target multiple or complex outcomes at the levels of activity and participation.

Identification of a primary outcome for such treatments may be impossible and even inappropriate. Goals associated with successful treatment will vary across participants, meaning that simple outcome measures may not provide universal and objective metrics of improvement. Moreover, a highly meaningful intervention may appear meaningless if the wrong outcome measure is selected. Rehabilitation interventions are often delivered by members of multiple disciplines working synergistically, complicating the application of quality appraisal standards that do not incorporate them (Fann, Hart, & Schomer, 2009).

Randomized Controlled Trials: Randomized controlled trials (RCTs) where patients are randomly assigned to at least two comparison groups are best able to control for threats to the internal validity of studies and ensure pre-treatment equivalence of experimental and control groups, strengthening the basis for statistical inference. Completing an RCT with an adequate sample size, appropriate randomisation techniques to account for variability in the diagnostic conditions and a combination of patient, service and/or system level outcome measures is difficult due to competition for rehabilitation research funding and the individual nature of brain injuries.

The majority of published studies that describe patients with brain injury use single-case design or are small case series. This reflects the individual nature of rehabilitation interventions, the challenges of using more complex designs, and the relative simplicity of conducting single case studies. While RCTs may suffer from problems with applicability of results or the heterogeneity of included patients or wider population, “studies of individuals and small case series can be optimal for exploring a new treatment, for titrating therapies, for documenting a promising variation in behavioural therapies, for enhancing knowledge of generalisation of treatment to a new group, and to enhance understanding of why some patients respond to a treatment of known (average) effectiveness whereas others do not, that is, for extending results of an RCT” (Johnston, Sherer, & Whyte, 2006).

The disadvantages associated with single case design studies are well reported. These include the difficulty in drawing cause-and-effect conclusions (limited internal validity), possible biases when interpreting outcomes due to observer bias and bias in data collection, and crucially, the problem of generalising findings from a single individual to a group or wider population (limited external validity). While researchers can take steps to attempt to limit the biases associated with this design there remain difficulties in assessing

behaviours which do not reverse back to baseline after withdrawal of treatment, indicating that the treatment may not have been the key variable affecting change. Single case studies are usually ranked at the bottom of the traditional hierarchy of evidence (Greenhalgh, 2006).

Ehical Issues: There are also ethical constraints in using RCTs, particularly with severely affected patients for whom clinicians believe there are no realistic alternative interventions to specialised care. Notably for conditions in which multidisciplinary rehabilitation has become the standard of care without systematic evidence to support it in practice, denying services randomly in order to conduct an RCT could be considered unethical (Prvu Bettger & Stineman, 2007).

3.7 Serious Games in Cognitive Rehabilitation

Videogames involving the sensory-motor system and problem-solving skills are serious candidates for neuro-rehabilitation and motor or cognitive training. In (Green & Bavelier, 2007) several improvements in gaming activity were identified, from reaction times to spatial skills. The opportunities for using this kind of media to improve cognitive functions in individuals with particular needs (as reviewed for surgeons and soldiers) or for training and retraining of individuals with special health-related problems (such as young disabled or the elderly) involving the nervous system were also highlighted. An improvement in the spatial resolution of attention in videogame players has been observed (Green & Bavelier, 2007). A persistent difficulty is that training can be more or less efficient depending on how it is administered and this is directly related with tasks difficulty management (Linkenhoker & Knudsen , 2002).

3.8 Flow in Computer Mediated Environments

Learning is enhanced when the match between the skills of the learner and the challenges of the subject matter are optimized (Whalen, 1998). Csikszentmihalyi's Flow Theory (Csikszentmihalyi, 1991) provides a framework and vocabulary for understanding the experiential nexus between the active person and the facilitative environment. The

experience of Flow creates information that melds actor and activity into one transactive system. In this sense, Flow may be seen as the experiential dimension of the ZPD (Whalen, 1998).

Flow or optimal experiences, also referred to as “the zone” (Csikszentmihalyi, 1991) represents a state of consciousness where a person is so absorbed in an activity that he or she excels in performance without consciously being aware of his or her every movement.

Within a Computer Mediated Environment (CME), the experience of flow in the past 20 years has demonstrated an increase in communication, office productivity software on desktop computers, learning, general web activity, online consumer settings, and online search experiences among others (Finneran & Zhang, 2005).

The practical implications of the consequences of flow experiences are clear, important, and promising. It is expected that a good understanding of the flow phenomenon would guide ICTs designers to build products that lead users to flow experiences. Little research is available concerning the application of data mining techniques in Flow. In (Mathwick & Rigdon, 2004) cluster analysis is used to identify a “flow cluster” comprised of individuals with high Internet search skills and a search task that presents a high navigational challenge.

From a research perspective however, flow is poorly defined in CME because of the numerous ways it is conceptualized, operationalized, and measured. Flow experience is associated with a person doing an activity. In traditional flow studies, the activities tend to be very clear: playing music, climbing a cliff, playing chess or reading a book. Most existing flow studies in CME do not clearly differentiate between factors that are related to the task and those that are related to the artifact.

Thus, there is a need to re-conceptualize flow in CME to consider the uniqueness of the artifacts and the complexity they add to the flow phenomenon. Indeed one of the aims of this work is to use PREVIRNEC© to produce flow experiences in the subject, thus incrementing the benefits of the neurorehabilitation process.

3.9 Summary

This chapter addresses the bibliography review of the different areas involved in the thesis. Being a multidisciplinary research, several areas have been reviewed, initially underlining the limitations of traditional Data Mining (DM) methods in our context of application. DM applications in the field of neurology have been analyzed in section 3.2. No works addressing high interactions among factors by considering cummulative effects nor multi-impact areas were found, neither in general, nor in the particular medical fields. A common belief in the literature is that DM on medical data requires specific medical knowledge as well as knowledge of DM technology. Therefore methodological approaches addressed in this field should and prior domain knowledge in the DM process.

This chapter presents a number of studies that employ traditional Data Mining techniques in TBI such as Classification (K-Nearest Neighbor, Decision Trees, Support Vector Machines, Neural Networks, Bayesian Methods) as well as Regression, Association Rules and Clustering methods. Among hierarchical clustering methods we particularly highlight Clustering Based on Rules (it is proposed as part of the SAIMAP methodology introduced in Chapter 2 and detailed in Chapter 4) which combines inductive learning elements with statistical methods to enhance clustering results. An important property of the method is that it permits to incorporate the interaction with clinical experts and prior domain knowledge in the DM process, increasing interpretability of the resulting classes.

Then the state of the art is reviewed in the context of repeated activities search patterns search (e.g. Sequential Pattern Mining methods). Some of them will be applied to our application case and results compared. Computational Geometry is reviewed, as in the step of NRR identification it is reduced by matrix algebra manipulations into a Computational Geometry problem. Motifs discovery techniques in sequential data is also reviewed as it is also introduced as part of SAIMAP methodology to describe treatment patterns.

Chapter 4. Sequence of Activities Improving Multi-Area Performance (SAIMAP) Methodology

In this chapter, the combination of pre-processing tools, clustering, motif discovery and post-processing techniques is proposed in a hybrid methodological frame, where sequential patterns of a predefined set of activities with high order interactions and cumulative effects among them are associated with multi-criteria improvement in a predefined set of areas of impact. The use of motifs is relevant because the cumulative effect of performed activities is robust to the time period intervals occurring between them and small interferences in a certain sequence do not decrease their effect on individuals performing them.

In Chapter 9 section 9.4 the results of applying this method on a real case study are presented. In section 9.10 it is discussed why the use of motif discovery is preferred for our problem to a classical supervised approach based on learning performance on the basis of sequences of tasks.

- The relationship among the patterns in $\mathcal{M}$ and the improvements in global or/and individual areas of impact in $\mathcal{A}$ due to execution of activities in T and the characteristics of individuals associated with the pattern (associations between $\mathcal{M}$ and X) This means finding associations between $\mathcal{M}$ and Z. In particular, given a threshold γ it is searched the subset $\mathcal{N} \subseteq \mathcal{M}$: $\forall\, \mu \in \mathcal{N}\ \ Prob(Z \mid I_{\mu_1} = YES) \geq \gamma$

- As described in Chapter 2, a set of distinct patterns (with no intersecting subsequences) $\mathcal{M}=\{\mu_1, \mu_2, \ldots\ \mu_m\}$, $\mu = (a_1, a_2, \ldots, a_{n_\mu})$ with $a_l \in \mathcal{A}$, inducing a partition over I is discovered over data to characterize groups of individuals following a similar sequence of activities.

4.1 The Sequence of Activities Improving Multi-Area Performance (SAIMAP) Methodology

Given the R matrix

1. Preprocessing

1. Build $s_i \ \forall i = \{1, \dots, n\}$ as $s_i = (R_i[2])^t$ $s_i$contains the sequences of tasks performed by i
2. Build $\chi = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix}$ as the matrix containing the sequence of tasks performed by each individual
3. Identify the frequency threshold ℓ to retain a task
4. Recategorize T by using a new category OTHERS grouping all infrequent tasks
5. Determine ℓ the threshold task length to be considered (percentile -95 of length of treatments distribution).
   Use only first ℓ columns of χ for the whole study and complete shorter sequences by "NULL" values
6. Build
   $\Delta = (D_1 \dots\dots D_a)$ effect of χ over each area of impact
   Z as a function of a subset of Δ

2. Descriptive Analysis

1. Build frequency plot of first ℓ columns and ℓ tasks of χ
2. Build heatmap of first ℓ columns and ℓ tasks of χ
3. Build heatmap of first ℓ columns and ℓ tasks of χ^a

3. Prior expert knowledge acquisition

Knowledge is represented by means of lf-Then rules in order to provide maximum flexibility and expressiveness to the expert. Only available knowledge is collected even if it is a partial description of the domain.

Build *KB* = { r: $\mathcal{B} \rightarrow \mathcal{L}$ } from a priori experts knowledge in target domain

4. Clustering of matrix χ:

The idea is to obtain standard patterns of sequences in terms of the areas impacted by the tasks performed by the users.

The methodology might accept any clustering method, but in our approach *Clustering based on rules* (Gibert and Zonicki, 1999) is strongly recommended, as it will be justified in the section 4.1.1 below.

Let P = $\{C_1 .. C_\xi\}$ be the set of classes found, being P a partition of I ($\forall C \in P, C \subseteq I$)

5. Split of χ per class:

Divide the data matrix in submatrices according to the different classes found in previous step.

$\forall C \in P$ build $\chi^a_C = \chi^a | C = [A_{it}]_{n_c M_t}$, i={i:i:1..n and i $\in$ C}, n_c =card{C}

χ^a_C contains only the rows corresponding to individuals in class C

6. Visualization of classes

$\forall C \in P$ build a heatmap of χ^a_C

7. Find motifs per class

1.Define an alphabet ζ of *a* single letters associated to the areas of impact in $\mathcal{A}$ such that $\forall$ A $\in \mathcal{A}$ is represented by $w \in \zeta$

2.$\forall$ C $\in$ P build χ_C^{ζ} by replacing the activities in χ_C^{a} by their corresponding initial in ζ

3.$\forall$ C $\in$ P find motifs of χ_C^{ζ} of length l (other methods can be used as well but MEME method is recommended; in any case a motif discovery method that does not use secondary and tertiary sequences must be used)

Let $e^l = \{e_{C1}^l \ldots e_{cM}^l\}$ be the vector with the E-values for all motifs found

$\forall$ motif m_C^l, C $\in$ P, l $\in$ $[l_{min}, l_{max}]$

Let Π_C^l be the *letter probability matrix* indicating the presence of the letter of alphabet in each position of the motif.

Eventually l might range in a certain interval $[l_{min}, l_{max}]$

8. Determine a level of minimum quality for motifs (α)

Usually $\alpha = 0.05$ is considered but other values can be considered as well

9. Prunning motifs: Retain the more frequent motifs for interpretation

1.$\forall$ C $\in$ P build $M_C^* = \{m \; in \; M_C \mid e_{c_m} \leq \alpha\}$

10. Visualize motifs

1. $\forall$ C **visualize $\mathbf{M_C}$** on the basis of π_C^l using the Seq_{Logo} representation, and interpret the motifs.
2. The characteristics of the sequences associated to each class might be easily identified over the motifs visualization.
3. Describe which areas of impact are addressed at which points of the sequences in each class

11. Analyze the effect of executing activities over the different areas of impact

1.Build multiple boxplot of D_j vs P**,** $\forall$ $D_j \in \Delta$

2. Kruskal-Wallis between D_j and P

Identify which areas improve the most in which classes.

12. Project all other illustrative variables over the clusters:

1. $\forall$ X_k in X

If X_k is numerical

Build the multiple boxplot X_k vs P

If $X_k \mid P \sim \mathcal{N}$ then

ANOVA

else

Kruskal-Wallis test

If X_k is qualitative

Build the Stacked Barchart of X_k vs P

If X_k vs P cross table all cells have more than 5 elements then

χ^2 independent test

else 2-tailored Fisher exact test

2. Retain all significant variables in X and build the description of additional characteristics of each cluster

13. Build final interpretation.

Associate the descriptions of motifs with the profile of performance and the characteristics of the individuals in each class, and constitute the final characterization of P

4.1.1. Proposed techniques

Although the proposed methodology is available for any clustering or motif discovery method, in this work a particular implementation using Clustering Based on Rules (ClBR) and MEME method is proposed. Brief description for this methods is provided below, together with the specific approach proposed for pattern interpretation.

Clustering phase (*ClBR*): Clustering Based on Rules (*ClBR*) combines inductive learning elements with statistical methods to enhance clustering results (Gibert et al., 1998). In our previous research *ClBR* was applied for knowledge discovery on the response to neurorehabilitation treatment of TBI patients where CR tasks have not been considered (Gibert et al., 2008). The main idea of *ClBR* is to allow the user to introduce semantic constraints on the formation of clusters (classes), providing them in a declarative way, in particular a rules knowledge base is used. This conditions provided by experts, formalize the apriori domain knowledge and induce a sort of *super-structure* on the domain; clustering is then performed within this structure by respecting the user constraints, and better approaching the clinical meaning of the resulting classes.

In the present analysis *ClBR* is applied to sequential data to identify meaningful classes of patients following similar sequences. Prior domain knowledge is considered, like the length of the prescribed treatment.

Motif discovery (MEME): The resulting clusters are then subjected to the MEME algorithm for motif discovery. MEME takes as input a group of sequences and the length of the searched motifs and outputs a motif for the group under different conditions. In our proposal we suggest not to restrict the number pof motifs in every single sequence as it makes sense that the motif might repeat several times along the treatment.

MEME then calculates the E-values of the discovered motifs, similar to a p-value for the log-likelihood of the motif. The motif with the smallest E-value in the searching space is proposed as the best motif characterizing the input dataset.

MEME also provides the position-specific probability matrix (PSPM, denoted as Π_C^l in the mehtodology) for the discovered motifs, representing the importance of each letter in each position of the motif. The PSPM matrix is the input to the sequence logos (Schneider and Stephens, 1990) (S*EQ_LOGOS tool)* providing the graphical representation for the discovered motif. The most representative motif for each of the classes is obtained together with its logo by using different motif lengths.

Patterns interpretation: Logos of all classes from different increasing lengths are used to understand regularities in the treatments of different classes.

The logos permit to sinthetyse the characteristics of treatments followed in each class. Then the relationships between those typical treatments and evaluations of patients performance might be analyzed. In our application, performance is evaluated through standardized neuropsychological assessment battery (NAB presented in section 1.4.3) and effect of treatment might be computed as pre-post differences over these batteries.

Statistical tests and multiple boxplots (Tukey, 1977) are used to relate the discovered groups with patient' characteristics, level of impairment and associated with specific treatment patterns. The proposal includes ANOVA or Kruskal-Wallis test (denoted as K_W in the proposed algorithm) for numerical variables depending on the characteristics of the variable itself and χ^2 independence test (Tukey, 1977) or two-tailored exact Fisher test (Agresti, 2012).

4.2 Summary

This chapter introduces the Sequence of Activities Improving Multi-Area Performance (SAIMAP) which is used in the first step of the CMIS methodology. SAIMAP uses a combination of pre-processing tools, clustering, motif discovery and post-processing techniques. SAIMAP takes as input the sequences of activities performed by each individual to find a reduced set of characteristics motifs for profiling the sequences followed by groups of individuals who behave similarly. It is proposed as an hybrid methodological frame in 13 formal steps, where sequential patterns of a predefined set of activities with high order interactions and cummulative effects among them are associated

with multi-criteria improvement in a predefined set of impact areas that might be targeted in parallel by a single task (or not).

The SAIMAP basically finds groups of similar treatments first, then characterizes them by using motif discovery methods local to each group.

Although the proposed methodology is available for any clustering or motif discovery method, in this work a particular implementation using Clustering Based on Rules and MEME method is proposed. Sequence logos of all classes from different increasing lengths are used to understand regularities in the treatments of different classes. Then the relationships between those typical treatments and evaluations of patients performance is analyzed. Statistical tests and multiple boxplots are used to relate the discovered groups with patient' characteristics, level of impairment and associated with specific treatment patterns. The proposal includes ANOVA or Kruskal-Wallis test for numerical variables and χ^2 independence test or two-tailored exact Fisher test, depending on the characteristics of the variable itself.

.

Chapter 5. Identification of the general pattern associated to a motif

5.1 General Pattern Identification

In this step each discovered motif included in $\mathcal{M}$ is analyzed and a general pattern of the form $([A]^{r},)^{*}$ $A \in \mathcal{A}$. is produced, giving a fix structure for the sequence design. For example the motif shown in Figure 5.1

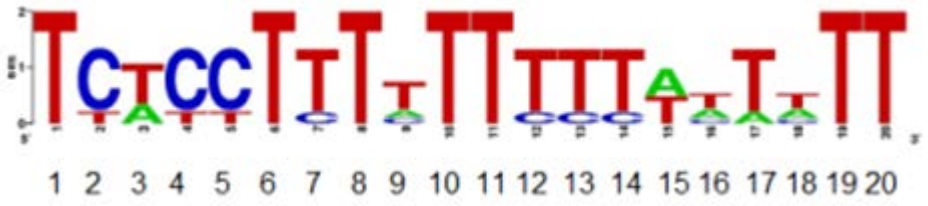


Figure 5.1 Graphical representation of motif

Will produce the global pattern (making a simplification of details and assigning a major area to each pattern):

$$[TCTC^{2}T^{9}A^{3}T^{2}]$$

The main principles are to use regular expressions built over the Π_{C}^{l} found by SAIMAP for every motif.

As a basic reference, the structure of the pattern will follow the regular expression $([A]^{r})^{*}$ where $A \in \mathcal{A}$.

However, as it is known that the motif might contain positions with uncertainty in which several letters appear (as it can be seen in the example and considering that in our context some tasks might address several impact areas simultaneously) we are working now in generalizing the pattern to an alphabet

$\mathcal{B} = \mathcal{P}(\mathcal{A})$ to produce patterns of the form $([\mathcal{B}]^{r})^{*}$

where some combination of different areas is also considered, and a more realistic approach is reached.

This will orient in the design of the final sequence of activities to select pure tasks for pure positions of the sequence where a single area requires impact according to the pattern, or to select tasks simultaneously addressing two or more areas in other positions indicated by the pattern

$$[T(C|T)(T|A)(C|T)^2(T(T|C)T(T|A)T^2(T|C)^3(T|A)^4T^2]$$

Given *l* the length of the motif, the matrix

$$\Pi_C^l = \begin{bmatrix} \pi^1_{C\,A_1} & \cdots & \pi^1_{C\,A_a} \\ \vdots & \ddots & \vdots \\ \pi^l_{C\,A_1} & \cdots & \pi^l_{C\,A_a} \end{bmatrix} \; C \in \mathrm{P},\; A_1 \ldots A_a \in \mathcal{A}$$

where $\pi^p_{C\,a}$ is the probability that area $a \in \mathcal{A}$ is impacted by the task executed in position p, p ∈ 1.. *l*. It holds that $\sum_{(\forall a \in A)} \pi^l_{C\,a} = 1$

Then

1. Determine the threshold $\gamma \in [0,1)$ such that a minimal probability of impact is retained in Π_C^l

2. Build $\Pi_C^{l^*}$ where

$$\Pi_C^{l^*}[p,a] = \begin{cases} \pi_C^l[p,a], & if\ \pi_C^l[p,a] \geq \gamma \\ 0, & otherwise \end{cases} \quad \mathrm{p} = 1..l,\ a \in w$$

3. Extract the suitable letters for a position of the motif: Process the matrix $\Pi_C^{l^*}$ by rows to build W_C^l such that $W_C^l = (w_c^1, \ldots, w_c^l)$, with

$w_c^p = \{ a \in \mathcal{A}: \pi^p_{C\,a} > 0\}$

4. Build regular expression from W_C^l

Find a regular expression $\mathcal{S}$ collapsing the repeated contiguous words in a single powered expression in such a way that

$if\ card\ (w_c^p) = 1$ then $\omega = a, a \in w_C$

$if\ card\ (w_c^p) = card(w_c^{P+1}) = \cdots card(w_c^{P+\delta})$ a single term $(w_c^p)^{\delta-1}$ is generated

$if\ card\ (w_c^p) > 1$, collapse all elements in w_c^p in a single string separating every element with "|"

Example:

$w = \{C|T\},\ \ \omega = C|T;\ \ w = \{\text{A}\}$ then ω=A

Example: Assume C = LONG6 and l = 15 and that the resulting motif gives Π^{15}_{LONG6}

Π^{15}_{LONG6}=

A	C	G	T	p
0.000000	1.000000	0.000000	0.000000	1
0.000000	0.250000	0.000000	0.750000	2
0.000000	0.750000	0.000000	0.250000	.
0.000000	0.000000	0.000000	1.000000	.
0.000000	1.000000	0.000000	0.000000	.
0.000000	1.000000	0.000000	0.000000	
0.000000	0.750000	0.000000	0.250000	
0.000000	1.000000	0.000000	0.000000	
0.000000	0.250000	0.000000	0.750000	
0.000000	0.750000	0.000000	0.250000	
0.000000	1.000000	0.000000	0.000000	
0.750000	0.250000	0.000000	0.000000	
0.000000	0.000000	0.000000	1.000000	
0.000000	0.750000	0.000000	0.250000	
0.000000	1.000000	0.000000	0.000000	15

The resulting logo is shown in Figure 5.2

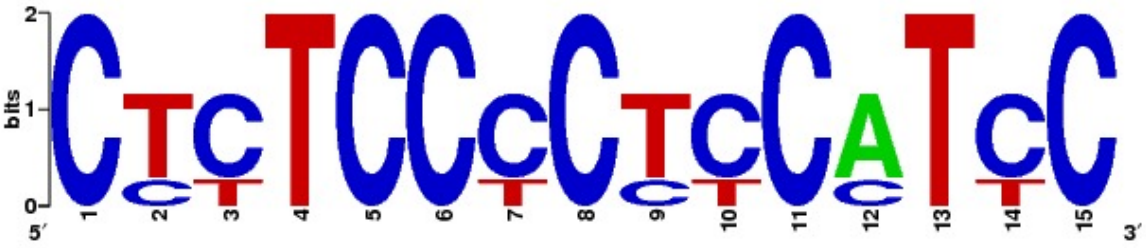


Figure 5.2 Graphical representation of LONG6 class motif l =15

Selecting γ = 0.25 it happens that $\Pi^{l^*}_{\text{C}} = \Pi^{\text{l}}_{\text{C}}$

The resulting w^{l}_{C} is:

w^{l}_{C}= ({C}, {C,T}, {C,T},{T},{C},{C}, {C,T},{C},{C,T},{C,T},{C},{A,C},{T},{C,T},{C})

The corresponding set of words is

w^1_{C}=C, w^2_{C}=C|T, w^3_{C} = C|T, w^4_{C}=T, w^5_{C}=C, w^6_{C}=C, w^7_{C}=C|T, w^8_{C}=C, w^9_{C}=C|T, w^{10}_{C}=C|T,

w^{11}_{C}=C, w^{12}_{C}=A|C, w^{13}_{C}=T, w^{14}_{C}=C|T, w^{15}_{C}=C

giving the sequence: C(C|T)(C|T)TCC(C|T)C(C|T)(C|T)C(A|C)T(C|T)C

where contiguous repeated patterns appear in some positions

C(C|T)(C|T)TCC(C|T)C(C|T)(C|T)C(A|C)T(C|T)C

2 2 2

Thus the resulting regular expression for this motif is:

$\mathcal{S}$= [C(C|T)^{2}TC^2(C|T)C(C|T)^{2}C(A|C)T(C|T)C]

Taking γ = 0.4

$\Pi_C^{l^*}$ =

A	C	G	T	p
0.000000	1.000000	0.000000	0.000000	1
0.000000	0.000000	0.000000	0.750000	2
0.000000	0.750000	0.000000	0.000000	.
0.000000	0.000000	0.000000	1.000000	.
0.000000	1.000000	0.000000	0.000000	.
0.000000	1.000000	0.000000	0.000000	
0.000000	0.750000	0.000000	0.000000	
0.000000	1.000000	0.000000	0.000000	
0.000000	0.000000	0.000000	0.750000	
0.000000	0.750000	0.000000	0.000000	
0.000000	1.000000	0.000000	0.000000	
0.750000	0.000000	0.000000	0.000000	
0.000000	0.000000	0.000000	1.000000	
0.000000	0.750000	0.000000	0.000000	
0.000000	1.000000	0.000000	0.000000	15

the resulting sequence of w_c^P is CTCTCCCCTCCATCC that results in the following expression $\mathcal{S}$ = [CTCTC^{4}TC^{2}ATC^2]

Thus, depending on γ the more or less flexibility is given to the final pattern to be considered for the treatment designs.

5.2 Summary

Taking as starting point the motifs identified by means of SAIMAP methodology in previous chapter, this chapter addresses the problem of identification of the general pattern associated with a given motif. The discovered motifs for each class at different lengths are analyzed leading to a general treatment pattern represented as a regular expression where the areas targeted at every position of the treatment are specified. This gives a fix structure

for the sequence design and will orient the health professional in the composition of the final sequence of activities provided that the set of areas impacted by each task are known.

Chapter 6. Neurorehabilitation Range (NRR) Sectorized and Anotated Plane (SAP) and NRR Maximal Regions (NRRMR) methods

- In this chapter we build on the concept of NRR and SAP tools to solve what we refer to as the NeuroRehabilitation Range Maximal Regions problem (NRRMR).
- Automatize the identification of NRRs with data-driven models that are able to avoid the limitations observed in the SAP performance.
- Overcome the problem of occlusions that appeared in the SAP, being a pure visualization tool, is, and
- Find a variable number of NRRs for a given CR task, according to different user-defined conditions concerning the acceptable degree of uncertainty.

6.1 Neurorehabilitation Range

In Clinical Pharmacokinetics, therapeutic range is defined as a range of drug concentrations within which the probability of the desired clinical response is relatively high and the probability of unacceptable toxicity is relatively low. Within this therapeutic range the desired effects of the drug are observed. Below it there is a greater probability that the therapeutic benefits are not realized (non-response or treatment-resistance); above it, toxic effects may occur (DiPiro & Spruill, 2010).

In this chapter, the concept of *NeuroRehabilitation Range* (NRR) is introduced as a translation of the classical therapeutic range from pharmacology to the field of neurorehabilitation (García-Rudolph & Gibert, 2014). The role of pharmacs in disease treatment is assumed in neurology by the role of neurorehabilitation tasks. The effect of treatment corresponds here to the restoration of cognitive functions.

Using this analogy, we will consider that a cognitive rehabilitation treatment task behaves in NRR if the desired clinical response is obtained i.e. if an observable

improvement in the targeted cognitive function is registered for the patient. As finding therapeutic range in pharmacokinetics consists of determining the proper drug concentration to be administered to a patient, finding NRR of a cognitive rehabilitation task is defined as determining the proper level of task difficulty to be proposed to the patient to obtain an optimal cognitive improvement of the targeted cognitive function.

We presume that being able to determine the NRR will provide a model that can help us to know better the relevant factors determining the ZRP proposed by Cicerone and Tupper (Cicerone & Tupper, 1986).

In PREVIRNEC©, following the execution of a given task T the subject gets a result RT ranging from 0 to 100. Section 1.4.5 details how this result is obtained, in this section we merely remark that a 0 result denotes the lowest level of task completion and a 100 the highest. Being the NRR of task T defined as NRR(T)=[r-,r+], and being r-,r+ in [0, 100], using a simple test it is easy to determine whether or not the patient performed the task in NRR:

$$\text{in NRR(RT)} \ \text{ iff } \ RT \in NRR(T) \equiv r- \leq RT \leq r+$$

- Tasks that are too easy will produce results higher than r+ and are probably out of ZRP because they only involve undamaged brain areas and do not demand impaired cognitive functions to be activated. In this case, we say the task has been executed in SupraNeuroRehabilitation Range (SNRR).
- Tasks that are too difficult will produce results lower than r- and are also likely to be out of ZRP. This is because they intensively required the implication of the impaired brain areas that cannot react to the excessively difficult cognitive stimulus. In this case we talk about InfraNeuroRehabilitationRange (INRR).

Currently, some hypotheses to determine a proper model for NRR(T), are being tested for the values of r-,r+. The aim of this chaper is to define a method by using data-driven models with PREVIRNEC© database to extract useful knowledge.

As a first attempt to find the neurorehabilitation range of a cognitive task executed by means of PREVIRNEC©, the result obtained by the patient in the execution of the task is used. The number of executions of the task performed by the patient is considered, as it is

known that repetition is highly related to activation of brain plasticity, which is in the core of cognitive functions re-establishment, as introduced in Chapter 1.

The proposed methodology presents two strategies for the analytical and graphical identification/visualization of neurorehabilitation and non-neurorehabilitation ranges based on the notion of Sectorized and Annotated Plane introduced below. The two models for NRR obtained are compared and discussed for the specific case of task151, related to visual memory cognitive function.

6.2 Sectorized and Annotated Plane (SAP)

Given three variables Y, X1, X2, where Y is a qualitative response variable, with values {y1,y2,...}, and X1, X2 numerical explanatory variables, the SAP is a 2-dimensional plot with X1 in the x axis, X2 in the y axis and rectangular regions with constant Y displayed and labeled with Y values as outlined in Figure 6.1. An SAP is therefore a graphical support tool aimed at visualization, where the response variable is constant in certain regions of the X1xX2 space. Eventually, allowing a relaxation of strict constant Y in the marked regions, the SAP might include an indicator of region purity, adding the probability of occurrence of the labeling value.

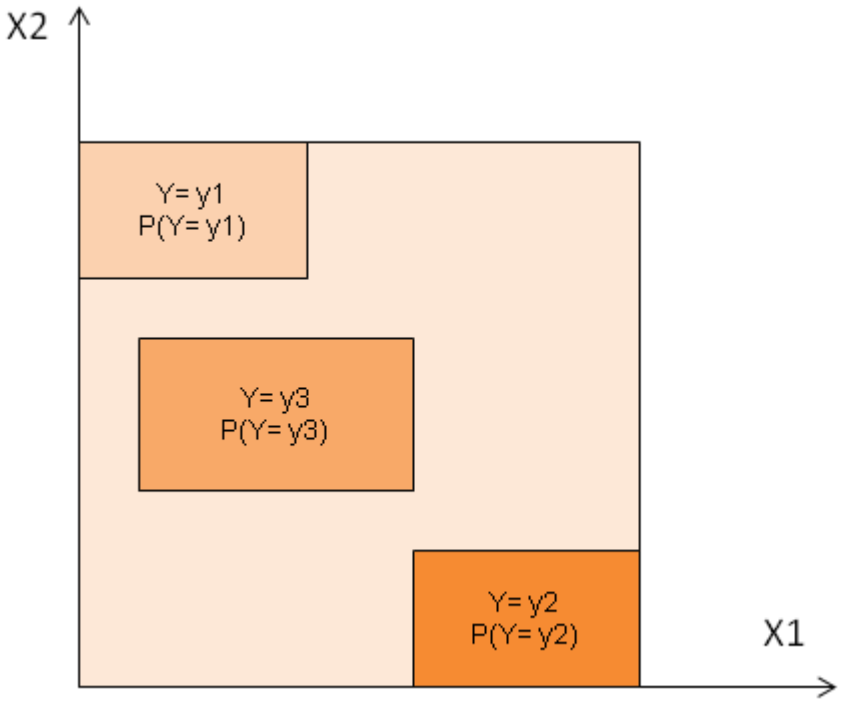


Figure 6.1. General Sectorized Annotated Plane (SAP) description

Given a particular CR task, and assuming Y as a binary variable reporting improvement of the patient in the cognitive function targeted by the task (yes, no), the SAP leads to response zones where participants show similar response to treatment. The SAP shows a plane sectorization directly related to treatment response. This allows identification of logical restrictions (rules) determining the different outcomes of treatment.

The SAP is built following two methodologies that implement two different strategies, the first one based on graphical visualization and the second based on the plane partitions induced by decision trees, as introduced below.

In our context V matrix provides the performance obtained by each individual on each task execution. A new column is added to V, giving the repetition number of each execution in the sequence performed by each patient.

Example.: Let us suppose patient 1002 following a treatement with

$$S_{1002}=[T_{80}, T_{83}, T_{65}, T_{80}, T_{82}, T_{145}, T_{68}, T_{81}, T_{82}, T_{145}, T_{66}, T_{79}, T_{79}, T_{46},$$
$$T_{79}, T_{79}, T_{148}, T_{148}, T_{151}, T_{79}, T_{85}, T_{148}, T_{148}, T_{15}]$$

The scorings achieved for each execution are also available and idPatient 1002 improves after treatment, V’ matrix will contain his executions as shown in Table 6.1

V’=

i	*t*	p	z	patientIndex
1002	T80	75	YES	1
…				
1002	T83	37	YES	1
…				
1002	T65	0	YES	1
…				
1002	T80	90	YES	2
…				
1002	T82	50	YES	1
…				
1002	T145	76	YES	1
…				
1002	T68	0	YES	1
…				
1002	T81	81	YES	1
…				
1002	T82	69	YES	2
…				
1002	T145	80	YES	2
…				
1002	T66	0	YES	1
…				
1002	T79	64	YES	1
…				
1002	T79	75	YES	2
…				
1002	T46	94	YES	1
…				
1002	T79	64	YES	3
…				
1002	T79	81	YES	4
…				
1002	T148	20	YES	1
…				
1002	T148	60	YES	2
…				

1002	T151	91	YES	1
…				
1002	T79	50	YES	5
…				
1002	T85	64	YES	1
…				
1002	T148	20	YES	3
…				
1002	T148	40	YES	4

Table 6.1. Tasks executions for idPatient 1002

The SAP is performed by using columns < P,patient index, Z >

6.3 The NRR

The NRR defined in section 6.1 extends to a bi-variate expression where:

K_{NRR} = {
if performance $p \in$ [p-,p+] & number of repetitions $r \in$ [r-,r+] *then* T is in NRR,
$\forall t \in T$
}

The intervals [p-,p+] and [r-,r+] are determined by SAP.

Alternatively a representation in form of matrix is also suitable.

The NRR matrix provides NRR found in all tasks in T as shown in Table 6.2.

IdTask	Results		Number of Executions	
	Lower bound	Upper bound	Lower bound	Upper bound
T_1	p-	p+	r-	r+
..	..		..	
$T_{\mathcal{T}}$	p-	p+	r-	r+

Table 6.2. The NRR matrix for all T

6.3.1 Visualization-Based SAP (Vis-SAP)

Data is plotted regarding X1 and X2, and each point is marked with different colors according to the values of Y. This categorized scatterplot (sometimes known as letterplot) is an exploratory technique for investigating relationships between X1 and X2 within the sub-groups determined by Y. For the particular application presented here, X1 is the number of executions of the task performed by the subject (Tasks repetitions), X2 is the result obtained at every single execution (Results), while Y is the effect of the neurorehabilitation process (improvement/non-improvement).

This exploratory analysis is used to identify systematic relationships between variables when there is no previous knowledge about the nature of those relationships. The constant-Y regions detected in the plot can be expressed in the form of logical rules involving the implied variables. The SAP is built on the basis of these rules.

6.3.2 Decision Tree-Based SAP

The tuple (X1,X2,Y) = (Execs151, Results, Improvement) can be treated as a classical classification problem, where Improvement has to be recognized on the basis of Execs151 and Results. The training dataset is used to induce a decision tree classifier which is later evaluated with the test set in the usual way. For the testing, the class label is ignored and predicted by the classifier. Performance of classifier is evaluated by comparing both predicted and real class. The confusion matrix and taxes of misclassification can be provided.

Here, the Weka (Hall, y otros, 2009) software has been used to apply the J48 (Witten & Frank, 2005) decision tree algorithm which implements Quinlan's C4.5 algorithm (Quinlan, 1986) building an unpruned tree.

Once the decision tree has been constructed, it is converted into an equivalent set of rules in the usual way.

6.3.3 Frequency table SAP (FT-SAP)

The main problem with Vis-SAP is that in every pixel in the image several points might be overlapped and not always labeled with the same response value. Detection of NRR regions is performed by labeling each pixel with the majority label, using a simple voting scheme and without taking into account the balance between improvement and non-improvement pixels overlapped.

In this section, a numerical representation of the Vis-SAP is used, based on a two-way matrix, precisely indicating how many points of each class are overlapped at any pixel in the graph (García-Rudolph & Gibert, 2015).

As in Vis-SAP, in this approach X1 is the result obtained at every single execution (Result), X2 is the number of executions of the task performed by the subject (Executions), while Y is the effect of the neurorehabilitation process: improvement/non-improvement (YES,NO) assessed by standardized neuropsychological tests.

Given mExec the maximum number of Executions of a task and mResults the maximum scoring of a task i=(1:mExec), j=(1:mResults), we define

m_{ij} = number of subjects such that *(X2= i) & (X1 = j) & (Y= YES)*

n_{ij} = number of subjects such that *(X2= i) & (X1 = j)*

$p_{ij} = m_{ij} / n_{ij}$ = percentage of subjects such that *(X2= i) & (X1 = j) & (Y = YES)*

For each *(i,j)* the matrix $P= (p_{ij})$ is built as shown in Table 6.3:

Execs/Results	*0*	*1*	*…*	*i*
1				
2				
.				
.				
j				$p_{ij} = {}^{m_{ij}}/_{n_{ij}}$
..				

Table 6.3. p_{ij} is the proportion of improving patients in pixel *(i,j)*

An FT-SAP is a graphical visualization where a gradient color from red to green can be assigned to pixel *(i,j)* according to its p_{ij} as shown in Figure 6.2

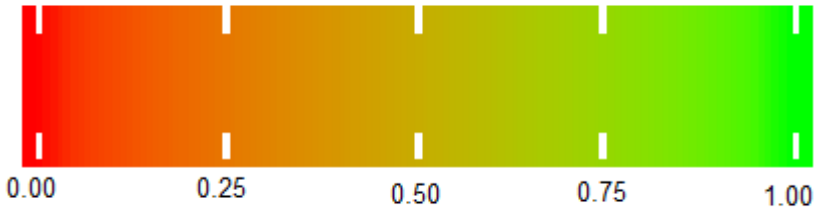


Figure 6.2 Color gradient for p_{ij} values in quartiles

Given a threshold $\gamma \epsilon [0,1]$ the NRRMR regions can be found over the FT-SAP (γ) as the set of regions *(r,s)×(t,u)* such that

$$\forall (i,j)\, r \leq i \leq s \; and \; t \leq j \leq u, it \; holds \;\; p_{ij} \geq \gamma$$

Given γ, a 2-color gradient can be defined providing a neat heatmap of the FT-SAP (γ) (as shown in Table 6.4). Being q_{ij} defined as:

$$q_{ij} = \begin{cases} 1 & p_{ij} \geq \gamma \\ 0 & p_{ij} < \gamma \end{cases}$$

Therefore a binary matrix Q is obtained, being $Q=(q_{ij})$. Q is a mask over FT-SAP filtering pixels according to γ (for empty cells, no color is provided for the pixel).

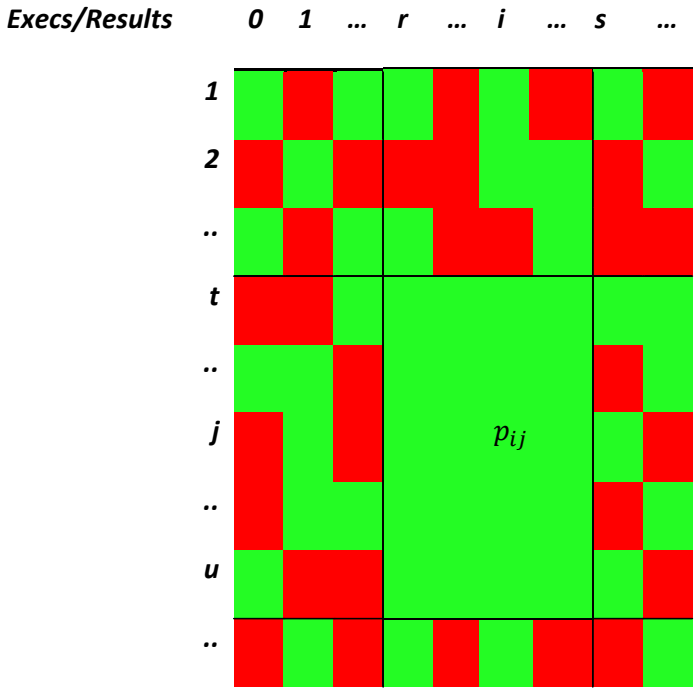


Table 6.4. A 2-color heatmap defining a two dimensional NRR for $p_{ij} \geq \gamma$

6.3.4 Analytical identification of NRR

Taking as input parameter the Q matrix resulting from filtering FT-SAP over γ, a method to automatically identify NRR (given maximum width and length of the surface to be

searched as user defined parameters) is described below. The idea is to find all rectangular groups of 1s cells equal to or greater than the minimum width and length provided by the user (García-Rudolph & Gibert, 2015).

It is solved by two pass linear O(n) time algorithm (n being the number of cells in the input matrix). As shown in Figure 6.3 below, first pass scans the matrix by columns, numbering cells consecutively until a red element (a 0 cell) is found and second pass scans by rows, searching for elements matching the length and width provided as parameters.

As is shown in the R code after Figure 6.3 the method allows for the simultaneous identification of the NRRs satisfying the user-defined conditions. The MAXRES and MAXEXEC values in the R code correspond to mResults and mExec respectively, as defined above. The proposed pseudo code is introduced below:

```
Input
Anxm matrix of red(0)/green(1) elements obtained after FT-SAP (γ):
MAXROW maximum number of rows
MAXCOL maximum number of columns

Output
NRR maxrowxmaxcolumn
First pass
For each column from bottom to top
  Repeat
    Number green element incrementally
  Until a red element is found → Restart numbering

Second pass
For each row from left to right
  NRRrows = 0 #Number of rows of the NRR solution so far
  Repeat
  If element >= MAXCOL
      Increment NRRrows
      NRR=NRR+NRR[element]
  Else NRRrows = 0
  Until NRRrows = MAXROW
Return NRR
```

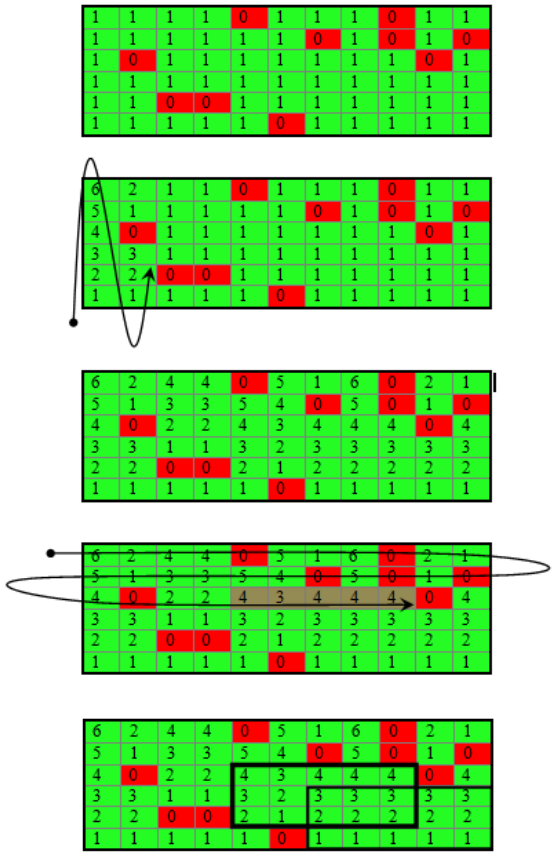


Figure 6.3 Example of the two-pass algorithm

```
#First pass
for (j in 1:MAXRES){
cont ← 1
 for (i in MAXEXEC:1) {
 if (y[i,j]==0){
cont ← 1
b[i,j] ← NA}
 else {b[i,j] ← cont
cont ← cont+1}
}}
#Second pass
Mdat ← b
apply(mdat, 1, function(x) {
  r ← rle(x >= MAXROW)
  w ← which(!is.na(r$values) & r$values & r$lengths >=MAXCOL)
  if (length(w) > 0) {
    lapply(w,FUN=function(w1){ before ← sum(r$lengths[1:w1]) - r$lengths[w1];
                    c(before+1,before+r$lengths[w1]) })
  } else
    NULL
})
```

With this algorithm the green rectangles as specified by the user in the FT-SAP for a given threshold γ can be identified and NRR established accordingly.

As will be seen in Chapter 10, some real cases provide large green areas contaminated by a small percentage of isolated red points that could be assumed as part of the NRR, provided that an uncertainty tax becomes associated with it. This implies modification of the previous algorithm to find regions with a certain degree of contamination. But the generalization about the provided implementation is not evident. Thus, a classical version of the MER algorithm has been used instead and properly modified. Section 6.4 provides our implementation of the classical MER and Section 6.4.1 provides the proposed generalization to permit a certain degree of contamination in the regions.

6.4 Maximal Empty Rectangle (MER) method

In this section, the SAP is transformed into a masked binary matrix and a geometric optimization algorithm (the Maximal Empty Rectangle problem (MER) (Naamad, Lee, & Hsu, 1984) is generalized to the NRRMR, allowing for the identification of regions satisfying user-defined conditions. Proposed methods are extended to any number of tasks grouped in cognitive functions, allowing for the identification of NRR of not only a single task, as in (García-Rudolph & Gibert, 2014) but a group of them. The key idea of the present work is to transform the Vis-Sap method from into a geometric optimization algorithm that avoids the visual effect of occlusions while permitting some degree of impurity in the detected areas of the NRR to be taken into account.

For this purpose, a generalization of the MER problem will be introduced (García-Rudolph & Gibert, 2015).

As a first attempt the direct approach to the MER problem is followed: Scan through the matrix, stopping at each element. Treat each element as a potential top-left corner of the MER rectangle. For each such top-left corner, try all other elements as a potential bottom-right corner of the MER rectangle.

Input
Amxn matrix of red/green elements obtained after FT-SAP(γ)
Output
MER submatrix of A

```
findMaxRectangleArea ← function (A) {
# 1. Initialize.
maxArea ← 0;
area ← 0;
# 2. Outer double-for-loop to consider all possible positions for top-left corner.
for (i in 1:m){
   for (j in 1:n){
  # 2.1 With (i,j) as top-left, consider all possible bottom-right corners.
   for (a in i:m){
           for (b in j:n){
   # 2.1.2 See if rectangle(i,j,a,b) is filled.
        Filled ← checkFilled (i, j, a, b);
 # 2.1.3 If so, compute it's area.
          if (filled){area ← computeArea (i, j, a, b)}
              # If the area is largest, adjust maximum and update coordinates.
           if (area > maxArea){
                      maxArea ← area;
                      topLeftx ← i;
                      topLefty ← j;
                      botRightX ← a;
                      botRightY ← b
       }
    }
  }
 }
}
maxR ← c(topLeftx,topLefty,botRightX,botRightY)
return (list(area=maxArea,rect=maxR));
}
computeArea ← function (i, j, a, b) {
if (a<i) {return(-1)}
if (b<j) {return(-1)}
return ((a-i+1)*(b-j+1))
}
checkFilled ← function (i, j, a, b) {
for (k1 in i:a){
           for (k2 in j:b){
      if (A[k1,k2]==0){return (FALSE)}
     }
}
return (TRUE)
}
```

Regarding the performance, in this approach each top-left corner visits about O(mn) locations. For each such top-left corner, the bottom-right corner visits no more than O(mn)

positions. An evaluation (checking for 1s) takes O(mn) in the worst-case for each rectangle checked. Total: O(m3n3) (worst-case). This means that finding pure regions in the Q matrix performs better over time when our implementation proposed is used.

Some improvements to the classical MER approach have been identified which improve performance: checking the area first before scanning for 1s and prune, ignoring the rectangle when the area is too small; also, eliminating as many size-1 rectangles from the search as possible and checking corners for 0s before proceeding.

6.4.1. Neuro Rehabilitation Range Maximal Regions problem (NRRMR)

To allow for the identification of non-empty regions (i.e. regions containing some degree of 0 values) a modification of the checkFilled function is introduced. A user-defined TOLERANCE is included as input to the function and only when that value is exceeded is the area considered as not filled. Figure 6.4 shows the identification of the maximal rectangle containing one non-empty element as output ([topLeftx, topLefty, botRightX, botRightY] = [5, 1, 7, 8] Area = 24), instead of the bottom-right rectangle that would be the output if no TOLERANCE parameter is introduced ([topLeftx, topLefty, botRightX, botRightY]=[13,8,16,12] Area = 20).

```
checkFilled <- function (i, j, a, b, TOLERANCE) {
tol<-0;
for (k1 in i:a){
  for (k2 in j:b){
     if (A[k1,k2]==0){
  tol<-tol+1;
      if (tol > TOLERANCE) {return (FALSE)}}
    }
}
return (TRUE)
}
```

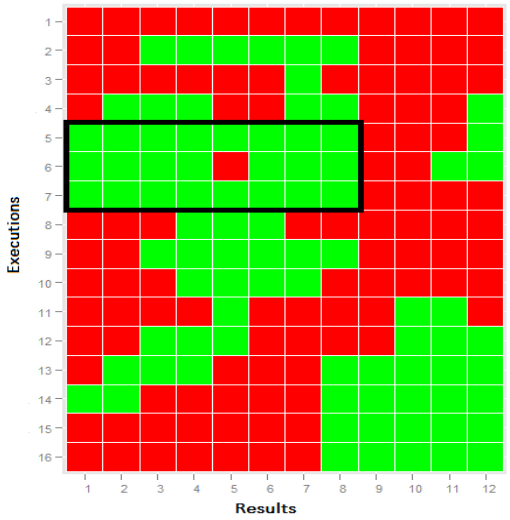


Figure 6.4 MER with user-defined TOLERANCE = 1

6.5 Quality Indicators

6.5.1 Sector Confidence

Given a sector S from the SAP graph, labeled with Y=y, the sector confidence corresponds to the empirical probability of occurrence of event y inside the sector. P(Y=y|S) is computed as the ratio between the number of positive cases and the sector size. This is in one sense a measurement of the purity of the sector and provides the quality of the assignment of class y to all elements in the sector. The higher the confidences of the SAP sectors, the better the model is considered. When Y is a binary variable P(Y= yes | S) it provides the sensitivity of S, while P(Y = No| ¬S) provides the specificity. As usual, the higher the sensibility and specificity, the higher the quality of S.

We define the *global quality of the SAP* as the pooled confidence of all sectors. Additionally for the SAP of binary variables a pooled specificity and a pooled sensitivity can be used as quality indicators.

6.5.2 Hypothesis Testing

Thus, the range of results determining those sectors where S is labeled as Y=yes (Improvement) determine the NRR of the task.

The sensitivity of the NRR is related to the fact that patients in NRR improve, or in a more relaxed formulation, that there is a high proportion of patient improvement within NRR.

The specificity is related to the fact that patients out of NRR do not improve. This can be measured by the high proportion of non-improving patients in INRR or SNRR or equivalently by the low proportion of improving patients in INRR or SNRR.

A classic 2-sample probability test is used to see whether the response to the CR therapy is significantly different for those executing tasks in NRR than those obtaining results out of NRR. It is expected that the probability of improvement is significantly higher for those in NRR. SAP models that provide sectors without significant differences should be disregarded, as they provide NRR with poor identification of the improving population. Thus, being

π_{MR} = Probability of improving being in NRR

$\pi_{M\bar{R}}$ = Probability of improving being out of NRR

the hypothesis tested is

$$H_0 : \pi_{MR} = \pi_{M\bar{R}}$$

$$H_1 : \pi_{MR} > \pi_{M\bar{R}}$$

$$e = \frac{p_{MR} - p_{M\bar{R}}}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_{MR}} + \frac{1}{n_{MNoR}}\right)}} \cong_{H_0} z$$

where p0 is the weighted common estimator of π_{MR} and $\pi_{M\bar{R}}$ under the Ho,

$$p_0 = \frac{n_{MR} p_{MR} + n_{M\bar{R}} p_{M\bar{R}}}{n_{MR} + n_{M\bar{R}}}$$

$$p_{MR} = \hat{\pi}_{MR} = \frac{number \quad of \quad patients \quad in \quad NRR \quad that \quad improve}{n_{MR}}$$

$$p_{M\bar{R}} = \hat{\pi}_{M\bar{R}} = \frac{number \quad of \quad patients \quad out \quad of \quad NRR \quad that \quad improve}{n_{M\bar{R}}}$$

The test is solved under the z-distribution, with $\boldsymbol{\alpha}$ = 0.05. The greater the difference between π_{MR} and $\pi_{M\bar{R}}$ ($\pi_{MR} > \pi_{M\bar{R}}$) the more sensitive and specific is the NRR criterion tested, the lower the p-value of the test, and better performs over real patients.

6.6 Summary

The NeuroRehabilitation Range (NRR) is introduced in this chapter as a translation of the classical therapeutic range from pharmacology to the field of neurorehabilitation. The role of medications in disease treatment is assumed in neurology by the role of neurorehabilitation tasks. The effect of treatment corresponds here to the restoration of cognitive functions. The idea is that a task is executed by a patient in NRR if a desired clinical response is obtained i.e. if an observable improvement in the targeted cognitive function (s) is registered for the patient. The NRR specifies how many times the task must be executed and the required performance to be obtained to result therapeutic. The proposed NRR model assumes uncertainty and both number of repetitions and expected performance degree are expressed by means of intervals.

Data mining techniques are used in this chapter to build data-driven models for NRR given past experiences of real CR treatments where improvements of patients is known. The expected degree of performance for a CR task and the required number of repetitions to produce maximum rehabilitation effects on the individual are determined. An operationalization of NRR is proposed by means of SAP (Sectorized and Annotated Plane) which is introduced as a visualization tool to identify areas where there is a high probability of improvement occurring. Three approaches to SAP are formally defined in this chapter

Vis-SAP, DT-SAP and FT-SAP; the parametric heatmap-based visualization proposed to overcome the limitations detected in Vis-SAP.

A classical 2-sample probability test is used to see whether the improvement is significantly different for those executing tasks in NRR than those obtaining results out of NRR. It is expected that the probability of improvement is significantly higher for those in NRR. Finally, the automatic identification of NRR is reduced to a Computational Geometry problem by algebraic manipulation of original data. The NRRMR (Neurorehabilitation Range Maximal Regions) problem is introduced as a generalization of the Maximal Empty Rectangle problem (MER), to identify maximal NRR over a FT-SAP. This permits to automatically detect NRR of a task based on a sample of executions by patients from which improvements after the global treatment are known.

Chapter 7. Evaluate improvements on each area of impact

7.1 Improvements evaluation

The inputs to this step is the χ matrix and the Δ impact evaluation

Till now we have been making an implicit assumption, assuming that function *f* introduced in the problem formulation is a function and this implicitly means that a single area of impact is targeted by an activity. This works for the identification of motifs as associating to each task the main impact area.

However, in the case where an activity might impact simoultaneously with more than a single area, *f* is not a bijective function and this is also well aligned with the motif structures where in a certain position of the motif, two letters or even more (the areas of impact) might be involved.

This might be taken into account in the sequence design, by choosing tasks impacting several areas together (motifs with 2 or more letters).

To this purpose a more realistic extension of f function is required indicating all areas targeted by a task. Thus, matrix F generalizes *f* function to a framework of activities impacting to multiple areas simultaneously.

F

Areas / Tasks	A_1	…	A_a
T_1	1	0	0
..	..	…	…
$T_{\mathcal{T}}$	1	1	1

Table 7.1. F matrix

The set of impact areas $\mathcal{A}$ is expressed in complete and disjunctive form in matrix F. This permits a framework where tasks in *T* can be multi-area, i.e. impacting simultaneously more than one area. Thus, in this more general and realistic scenario where a task might simultaneously impact several areas, *f(*A*)* indicates de primary area of impact and F(t,) gives the total list of areas impacted by the task. For scenarios with pure tasks impacting a single area each, F is a matrix with a single non null cell per row.

$$F(t,A) = \begin{cases} 1 \;\; if \; f(t) = \text{ A or t marginally impacts A} \\ \\ 0 \;\; if \; f(t) \neq \text{ A} \end{cases} \qquad \forall t \in T; A \in \mathcal{A}$$

Also, from χ matrix it is possible to derive N matrix (giving the number of executions of each task t executed by *i*)

N:

Tasks / Individuals	T_1	…	T	…	$T_{\mathcal{T}}$
i_1	r_{11}				$r_{1\mathcal{T}}$
..	..	…			…
i			$r[i,t]$		
..	..	…			…
i_n					

Table 7.2 N matrix

being *r* [*i*,*t*] the number of repetitions of task *t* in the treatment followed by patient *i*

With matrices F, Δ, N and NRR, the matrix Υ might be computed, accounting for average improvement of patients executing a certain task a certain number of times along treatment, in the area impacted by that task (or areas).

Υ

Repetitions / Tasks	rmin			rmax
T_1			$\Upsilon_{t,r}$	…
..	..	…		…
$T_{\mathcal{T}}$				…

Table 7.3. Υ matrix

Where

$$\Upsilon_{t\,r} = \frac{\sum_{a \in \mathcal{A}:F[t,a]=1} \frac{\sum_{\forall i \in I:N[i,t]=r} \Delta[i,a]}{card\,\{i: N[i,t] = r\}}}{card\{a \in \mathcal{A}: F[t,a] = 1\}}$$

The improvement is the average improvement obtained by involved patients in the area impacted by the task (according to F) or the mean of average improvements in the several areas targeted, if the task impacts more than a single area. The involved patients are those executing the task exactly *r* times along their treatment.

Moreover, as Υ is intended to be used for helping composition of individual treatments and therapists will have preference for proposing execution of tasks in NRR it is expected that task t is never recommended for a therapy a number of times out of its corresponding NRR.

Thus, having matrix NRR which provides the NRR of a given task

NRR

Ranges / Tasks	minrep	maxrep	minscore	maxscore
T_1	r_1^-	r_1^+		…
..	..	…		…
$T_{\mathcal{T}}$	$r_{\mathcal{T}}^-$	$r_{\mathcal{T}}^+$		

Table 7.4. NRR matrix

an interval of times to be repeating task t to get therapeutic results might be obtained. Thus, it is expected that given $[r_t^-, r_t^+]$ for task t, the columns $\{0, \dots r_t^- - 1\}$ and $\{r_t^+ + 1, \dots\ rmax\}$ of row t in matrix Υ are never considered by the therapist.
Thus matrix Υ^* is built as a mask of Υ masked by non NRR positions

Υ*

Reps / Tasks	rmin	$r_{t_1}^-$	…	…	$r_{t_2}^-$	…	…	…	$r_{t_1}^+$	…	$r_{t_2}^+$	…	rmax
T_1		Υ_{1,r_{t1}^-}											
T_2													
…													
…													
…													
$T_{\mathcal{T}}$													

Table 7.5. Υ* matrix

$$\Upsilon^*_{[t,r]} = \begin{cases} \Upsilon\,[t,r], if\ \text{r} \in NRR(t) = [r_t^- r_t^+] \\ NAN, otherwise \end{cases}$$

7.2 Summary

This chapter formalizes the process of determining the improvement of an individual in the several areas of impact after execution of a given task in NRR.

Here, an important generalization of the general frame is proposed in such a way that hybrid tasks are considered in the sense that they are allowed to simultaneously impact several cognitive areas (this points to a more realistic scenario like the one happening in CR treatments where some CR task simultaneously target for example attention and memory functions. A sequence of algebraic matrix operations is defined to efficiently compute the average improvement of patients executing a given task a certain number of times. As in Chapter 6 the NRR of tasks is obtained, and this establishes an interval in which the task must be repeated to be therapeutic, in this chapter this information can be used to constraint the feasible solutions space to be considered. Thus Υ^* is a matrix providing the average improvement of patients executing each task a number of times included in the respective NRR.

Chapter 8. Treatment design

8.1 Composing the treatment

Given the set of patterns $\mathcal{S}$ and F, the Υ^* matrix is used to find feasible solutions. For the particular case of finding a single task for each token of the pattern $\mathcal{S}$, the problem trivializes to maximize the corresponding column *r*, being *r* the number of times indicated by the pattern $\mathcal{S}$, of the matrix $\Upsilon^*|_A$, being A the area indicated by the pattern.

So, given a slot of the pattern of the form $\mathcal{B}^{\,r}$ $\mathcal{B} \in \mathcal{P}(\mathcal{A})$ the designer needs to identify an activity that impacts to $\mathcal{B}$ areas in Neurorehabilitation range when executed about *r* times, thus, the task choice will be

$$t^* = \arg\, max_{t \in T:F[t,]=B}\ \Upsilon^*[\mathrm{t}, \mathrm{r}]$$

As an example, given a pattern $T^2(C|T)C^3(A|T)T^4$ it contains 5 tokens. For each token two types of information appears:

1) the area to be impacted on the set of areas: $\mathcal{B}$

2) the number of times to impact them in a contiguous sequence of tasks: *r*

r is indicating the column of Υ^* to be used for optimizing. $\mathcal{B}$ is indicating the subset of impact areas to be targeted by the task. Matrix F will identify the subset of tasks targeting this particular set of areas. Optimization should occur restricted to those rows in Υ^*. As Υ^* has NAN for those tasks where *r* is out of NRR the solution is guaranteed to be in NRR.

Thus for first token in the example $\mathcal{B}$=T (means a task impacting on executive functions), *r*=2 (means that it will be repeated twice), set of possible tasks is that of tasks impacting only executive functions, that is, those with F rows of the form (0,0,1) these are 10 of the total considered tasks: $\{T_{23}, T_{34}, T_{48}, T_{53}, T_{54.}, T_{55}, T_{56}, T_{62}, T_{63}, T_{64.}\}$. Thus column 2 of Υ^* is optimized over the same subset of tasks. The maximum improvement in this column

corresponds to the more negative value (as explained before in section 1.4.3) optimum is -2.45 and corresponds to task T_{23} that therefore is selected for the CR proposal.

For the second token $\mathcal{B}$=C|T, r=1. In a similar way F helps to determine the set of possible tasks impacting simultaneously on memory (C) and executive functions (T) which is a set of the following 6 tasks: {$T_{47,}$ $T_{58,}$ $T_{72,}$ $T_{89,}$ $T_{94,}$ T_{135}}. Thus column 1 of Υ^* is optimized. Optimal delta value = -1.25 which corresponds to task T_{135}.

For the third token $\mathcal{B}$=C, r=3; set of possible tasks impacting only memory: {$T_{4,}$ $T_{11,}$ $T_{15,}$ $T_{18,}$ $T_{22,}$ $T_{29,}$ $T_{34,}$ $T_{38,}$ $T_{41,}$ $T_{42,}$ $T_{43,}$ $T_{44....}$ }.Thus column 3 of Υ^* indicates T_{22} (with optimal delta value = -0.45).

For the fourth token $\mathcal{B}$=A|T, r=1; set of possible tasks impacting simultaneously on attention and executive functions: { $T_{78,}$ $T_{80.}T_{99,}$ T_{102}}. Thus column 1 of Υ^* indicates T_{102} (optimal delta value = 0.34).

For fifth token $\mathcal{B}$=T, *r*=4; set of possible tasks impacting only executive functions: {$T_{23,}$ $T_{34,}$ $T_{48,}$ $T_{53,}$ $T_{54.,}$ $T_{55,}$ $T_{56,}$ $T_{62,}$ $T_{63,}$ $T_{64.}$}. Column 4 of Υ^* indicates T_{54} (optimal delta value = 1.45).

And following this process the recommendation for the CR program is:

T_{23} T_{23} T_{135} T_{22} T_{22} T_{22} T_{102} T_{54} T_{54} T_{54} T_{54}

8.2 Summary

This is the culmination of the whole process where all elements developed in previous chapters integrate together to provide decision support to the therapist that has to compose a CR plan for a given person. The main idea is to use the regular expression associated to the recommended treatment of a certain type of patient as a general frame to be instantiated by specific tasks. Matrix Υ* is used to solve an optimization problem where NRR of tasks are taken into account and the regular expression of the CR pattern is used as the structure to be optimized.

Chapter 9. Application to TBI CR Programs

9.1. Effects of Cognitive Rehabilitation on Traumatic Brain Injury Patients

This chapter presents the application of the proposed methods in a clinical context: the Neuropsychology Department of the Acquired Brain Injury Unit at Institut Guttmann Neurorehabilitation Hospital (IG) where TBI patients undergo CR treatments.

The Information Technology framework for CR treatments in this clinical setting is the PREVIRNEC© platform (introduced in section 1.4.4). This is a J2EE client-server architecture specifically designed and developed to manage CR plans assigned by therapists to patients and to follow up information about the process (i.e. CR session dates, task execution in each session, performance, involved therapists, patients, tasks results, and task time, as detailed in section 1.4.4).

As presented in section 1.1 there are three main cognitive functions to be rehabilitated in a CR program: *attention, memory, and executive functions;* all of them can profoundly affect individuals' daily functioning. Even mild changes in the ability to attend, process, recall and act upon information can significantly affect the patient's quality of life. Consider, for example, the cognitive skills required for successful meal preparation. The individual must *plan* a menu, identify the required ingredients, develop a shopping list for the required items, and schedule sufficient time for shopping and preparing the meal. Then the individual must *sequence* many food preparation activities in an organized way so that everything is ready at dinner time. Even a mild attention or executive function deficit can render this difficult, ineffective or even impossible.

The main hypothesis framing our proposal is:

1) Some CR rehabilitation tasks are designed to improve particular cognitive functions, although attention, memory, and executive functions are related and interdependent (Sohlberg and Mateer, 2001). Their close interdependence stems from both a functional association and their shared neurocircuitry. This means that

performing a task that targets memory can also have collateral effects on other cognitive functions like attention or executive functions.

2) The additional effect of a single task might be affected by the cumulated effect of the sequence of previous tasks executed under the treatment; this might determine that order of execution is relevant in the treatment outcome.

9.2 The Dataset

One hundred and twenty-three TBI adults following a 3-5 months CR treatment at IG Neuropsychological Rehabilitation Unit are analyzed in this study. For every patient the following demographic and clinical variables are considered: age, gender, educational level, Glasgow Comma Scale (GCS) and Post Traumatic Amnesia (PTA) duration. Table 9.1 shows the basic statistics for numerical variables while frequency distribution of qualitative ones are shown in Table 9.2.

Variable	N	N*	Mean	Std Dev	Min	Q1	Median	Q3	Max
AGE	123	0	36.56	6.50	18	25	32	40	68
GCS	89	34	6.45	3.15	0	4	6,5	40	14
PTA	40	83	131.6	140.5	34	79	103	136	947

Table 9.1 Basic descriptive statistics for numerical variables

Initial assessment of TBI severity is reported according to GCS levels. A GCS score of eight or less after resuscitation from the initial injury is classified as a *severe* brain injury. The GCS score for *a moderate* brain injury ranges between nine and thirteen and a score of thirteen or greater indicates a *mild* brain injury or concussion. As detailed in Figure 9.1, most GCS scores (86.17%) show *severe* brain injury level (mean value 6.45±3.15).

It is known that those whose length of PTA is less than two months have a very good chance of at least being able to live on their own (even if they can't return to work or school). On the other hand, patients whose length of PTA is longer than three months are unlikely to be able to return to work or school (although they might be able to live on their own). As N* shows in Table 10.1, PTA measures were not available for 67% of the participants. Considered values show very severe conditions, as indicated by the median

(103) being more reliable than the mean because of the outlier visualized in Figure 9.1 (right).

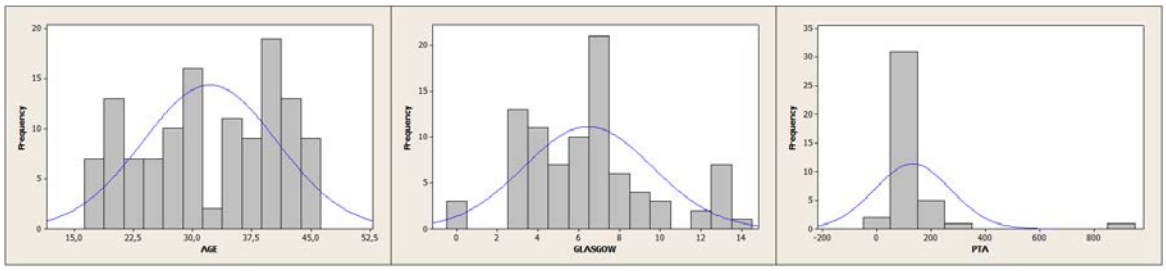


Figure 9.1. Numerical variables histograms: Age (left), Glasgow Comma Scale scores (center) and Post Traumatic Amnesia days (right).

Demographic qualitative data are shown in Table 9.2, 91 men (73.98%) and 32 women (26.02%) participated in the analysis. The educational background level of each participant is categorized into 3 groups: Group 1 (Elementary School, Group 2 (Medium) and Group 3 (third level education e.g. University degree).

GENDER	Count	Percentage	EDU	Count	Percentage
Female	32	26.02	**Elementary**	60	48.78
Male	91	73.98	**Intermediate**	40	32.52
			High	23	18.70

Table 9.2. Basic descriptive statistics of gender and educational level

All participants signed to notify their informed consent to the neuropsychological procedure, which was approved by IG's Ethical Committee. All met the criteria for initiating IG neuropsychological rehabilitation treatment.

Following NAB initial evaluation, all patients initiated a 3 to 5-month program (November 2007 to November 2009) based on personalized interventions in the PREVIRNEC© platform where patients worked in every one of the specific cognitive domains, considering the degree of the deficit and the residual functional capacity. All patients were administered the same NAB neuropsychological assessment at the end of the rehabilitation program. A total of 39412 task executions have initially been included in this analysis, involving the 96 different CR tasks included in the PREVIRNEC© platform.

9.2.1 Structure of Database

Originally, the system records the execution of every task as a single row in a log file in which the following information is also recorded:

Date is the date on which the T_s task is executed (date *yyyymmdd*)

TaskName is a descriptive name assigned to identify the task T_s

Score is the result obtained in that execution (0 to 100 real number)

NumTask is the automatic task generation number assigned to the task (0,1,2)

Difficulty is the difficulty level of the task (0,1,2,3,4)

Function is the cognitive function addressed by the task (Attention, Memory, Executive functions)

Subfunction is the specific cognitive subfunction addressed by the task (as described below, for the Attention function the addressed subfunctions are visual attention, sustained attention, selective attention, etc).

Original data structure (*S1*):

$$\begin{bmatrix} i & T & Date & TaskName & Score & NumTask & Dificulty & Function & Subfunction \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \end{bmatrix}$$

9.3 Instantiation of the Formal Problem

The dataset presented approaches the formal problem presented in Chapter 5 as a particular case where a CR treatment is the scenario in which each patient *i* executes a sequence of activities, one at a time:

- I is the set of TBI patients undergoing CR treatment at IG
- $T = \{T_s \ \ s=1:\mathcal{T}\}$ is a set of CR tasks that patients execute along treatment, where $\mathcal{T}$ is the total number of different CR tasks as presented in section 9.4.1.

- $\mathcal{A}$ is the set of areas of impact. In this particular case, it matches the family of different cognitive functions targeted by the CR tasks. According to (Sohlberg and Mateer, 2001) $\mathcal{A}$={Attention, Memory, Executive functions} are the main cognitive functions involved in daily living activities, thus these are also the cognitive functions treated on the PREVIRNEC platform.
- $f(T)$=a provides the main cognitive function a, (a $\in\mathcal{A}$) targeted by task T
- Given a patient i, the matrix R_i provides the list of all tasks executed by the patient i with its corresponding execution times throughout his CR treatment
- Matrix χ gives for each row i the sequence of CR tasks performed by patient i during treatment
- The set Y_{jt}, t=1..14 of numerical indicators of performance is, in this case, a battery of assessment tools for evaluating the degree of impairment of each cognitive function. In IG the NAB battery introduced in section 1.4.3 is applied and in this work, 14 relevant and non-redundant items from 7 assessment scales in NAB are selected as detailed in section 9.4.1.
- D_j, for each item selected in NAB, D_j, is the difference between the scores obtained by the patient before and after the prescribed CR treatment
- Δ= (D_1……D_a) represents the effect of CR treatment in all cognitive functions
- X =(X_1…X_K) additional information over patients. X_K might be either numerical (like age or GCS) or qualitative (like Sex or Educational level).
- Z indicates a global improvement of the patient after treatment.

The execution of each task by a patient occurs at different periodicities for each patient; the length of treatment varies according to both the number of tasks executions and total treatment time for the different patients; the sequence of task executions differs from one patient to the next; the result obtained in an execution determines both the task and difficulty of the next task proposed by the system; the effect of a task over cognitive functions of the patient is accumulative and the effect of a certain sequence of tasks might not be affected by small variations in the sequence itself, i.e. by the introduction of small additional tasks in intermediate positions of the sequence.

For all these reasons, our problem is suitable to be treated under SAIMAP methodology.

9.4 The Sequence of Activities Improving Multi-Area Performance (SAIMAP) Methodology

9.4.1 Preprocessing

As an initial hypothesis it is assumed (after consulting with experts) that the time interval (delay) between the execution of two consecutive tasks is irrelevant for rehabilitation purposes, since the cognitive functions of each patient are sensitive to the task execution and not so much to the time period between consecutive tasks. Thus, the *sequence* of tasks followed by each patient is to be focused on as the main target, independently of the time interval in which they have been performed. This permits a simplification of the problem to a new structure in which order of tasks is maintained, but dates are omitted.

First step of preprocessing is building the s_i sequence of tasks performed by each patient i= {1,…123} on the basis of R matrix by building $s_i = (R_i[2])^t$

Being the set of all possible tasks to be executed:

T={GlobalLocal, MathMazeComp, MathMazeExer, ConcOps, Submarine, Matching,BagOfCoins, Differences, Figures,PuzzComp, PuzzExer, LetterSoup,Bingo, DiffDirection,StraightLine, SameDirection, GroupWords, CategorizationTwo, CategorizationThree, SameCatWords, Circle, Platforms, Zigurat, GoNoGoEst, GoNoGoGame, GoNoGoPos, Hanging, SinkFleet, Maze, FourInRow, Fourth, JigSaw, BuildSentence, Fragments, Serie, CyclicSerie, SameCat, TempOrder, Position, Sequential, Simoultaneous, WordSeqDec, WordSeqSel, WordSeqDifCat, WordSeqSameCat, WordSimDec, WordSimSel, WordSimDifCat, WordSimSameCat, WordTempOrder, PairsSeqDec, PairsSeqRel, PairsSeqSel, PairsSeqSameOrder, PairsSeqRandOrder, PairsSimDec, PairsSimRel, PairsSimSel, PairsSimSameOrder, PairsSimRandOrder, SentSecOrder, SentSecTest, SentSecWrite, SentSecQuestion, SentSecTrueFalse, SentSimOrder, SentSimTest, SentSimWrite, SentSimQuestion, SentSimTrueFalse, RecSeqNumbers, RecSimNumbers, RemSecNumbers, RemSimNumbers, TextSort, TextQuestion, TextWrite, TextTrueFalse, ImgWordTempOrder, ImgWordSeqDecide, ImgWordSeqRel, ImgWordSeqSel, ImgWordSeqSameOrder, ImgWordSeqRandOrder, ImgWordSimDecide, ImgWordSimRel, ImgWordSimSel, ImgWordSimSameOrder, ImgWordSimRandOrder, DrawTemporalOrder, DrawRecognition, SceneRecognition, SceneRecall, VisualMemory, VisualSimon}

$\mathcal{T}$= card (T) = 96

and R being a matrix that for every task executed by patient provides the

$$\langle patientid, Name\ of\ task, Time\ stamp\ of\ execution \rangle$$

All tasks performed by patient *i* are collected in R_i

$$R_1 = \begin{bmatrix} 1 & TemporalOrder & 1 \\ 1 & StraightLine & 2 \\ 1 & DiffDirection & 3 \\ & \vdots & \\ 1 & NumMaze & 632 \\ 1 & MatchMaking & 633 \\ 1 & FourInRow & 634 \end{bmatrix} \ldots\ldots R_{123} = \begin{bmatrix} 123 & PositionalStimuli & 1 \\ 123 & FourInRow & 2 \\ 123 & MatchMaking & 3 \\ & \vdots & \\ & & \\ 123 & GoNoGo & 622 \\ 123 & Platforms & 623 \end{bmatrix}$$

Thus:

$s_1 = (TemporalOrder \quad StraightLine \quad DiffDirection \quad \ldots \quad NumMaze \quad MatchMaking \quad FourInRow)$

$s_{123} = (PositonalStimuli \quad FourInRow \quad MatchMaking \quad \ldots \quad GoNoGo \quad Platforms \quad TemporalOrder)$

$t_{f_1} = 634$

$t_{f_{123}} = 623$

Next, χ matrix is built by combining all s_i in the rows

$$\chi = \begin{bmatrix} TemporalOrder & StraightLine & DiffDirection & \ldots & NumMaze & MatchMaking & FourInRow \\ & & \ldots \quad \ldots \quad \ldots & & & & \\ PositonalStim & FourInRow & MatchMaking & \ldots & GoNoGo & Platforms & TempOrder \end{bmatrix}$$

Eventually the tasks are identified by a shorter alias, for simplicity:

$$\chi = \begin{bmatrix} T127 & T145 & T034 & \ldots & T256 & T045 & T145 \\ & & \ldots \quad \ldots \quad \ldots & & & & \\ T123 & T065 & T134 & \ldots & T011 & T032 & \end{bmatrix}$$

The next step is to determine the minimum ℓ to retain a task. Regarding the number of task executions, as detailed in Figure 9.2, for each IdTask (represented in *x* exe) the number of executions in the *y* exe clearly shows that there is a pack of 12 tasks from left to right as idTask 151 to IdTask 210 that are much more frequently executed than the rest. There are also a high number of available tasks, only exceptionally included in CR treatment programs.

For our purposes, the subset of the most frequently executed tasks will be targeted and all remaining tasks will be recoded into an OTHERS category.

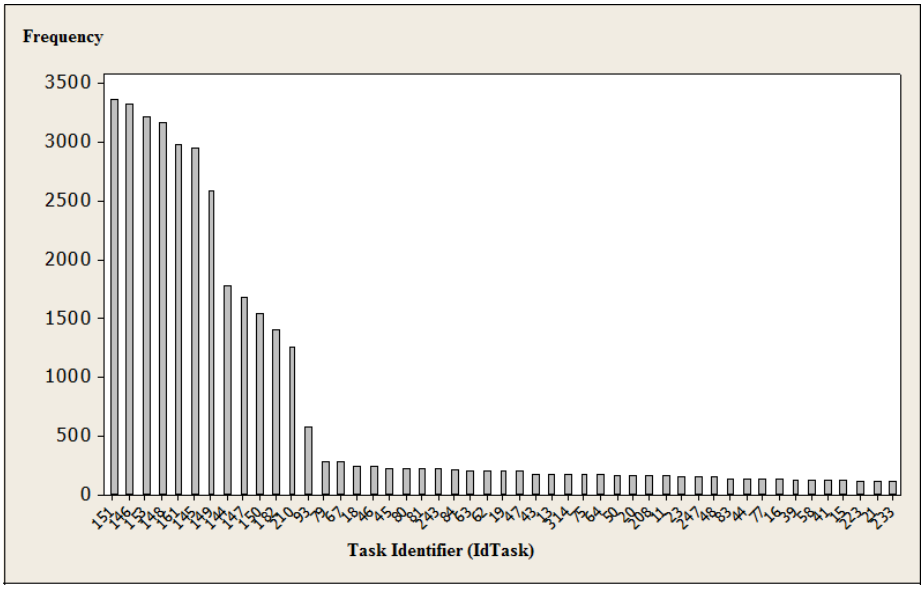


Figure 9.2. Frequencies of task executions. X axe shows the identifier of the task and tasks are ordered by increasing frequency to the left. Only tasks registering more than 100 repetitions are shown. The first 12 tasks on the left-hand side are the most frequently executed.

Table 9.3 details the percentage of the total number of executed tasks for each idTask. It shows that the 12 more frequent tasks exceed 70% of the total activity in the period considered. Thus taking f = 1000 means to retain only tasks executed more than 1000 times, and this points to the 12 more frequent tasks identified above. Given that this percentage is close to the Pareto Principle, we will focus on this pack of tasks.

IdTask	Task name	Number of executions	Percentage (all)	Percentage (selection)
151	Memory	3369	8.20	11.51
146	StraightLine	3329	8.10	11.37
153	TemporalOrder	3226	7.85	11.02
148	FourInRow	3170	7.71	10.83
161	Matching	2978	7.25	10.17
145	Exercise	2951	7.18	10.08
149	Competition	2589	6.30	8.84
144	Circles	1780	4.33	6.08
147	Series	1671	4.06	5.71
150	DiffDirection	1531	3.72	5.23
182	GoNoGoGame	1411	3.43	4.82
210	Platforms	1251	3.04	4.27
		29256	**71.17**	**99.99**

Table 9.3. Number of executions for the 12 most frequent tasks

In addition, only exceptionally do patients perform very large sequences of tasks. It will often be possible to identify a threshold length to be considered as most usual. This length will be denoted ℓ Patterns of sequences will be searched only in the first ℓ tasks executions of the patient's sequences, to avoid dealing with the sparseness of the final part of the data matrix. According to the Pareto principle ℓ threshold will be determined in such a way that no more than 20% of patients perform larger tasks. These data rows are completed with a special idTask label (e.g. "NULL"). This transforms an originally variable length matrix into a rectangular matrix χ, to be treated. As each patient's activities are different, each sequence of tasks shows a different length, the shortest one being of length 9 and the longest one of length 1391.

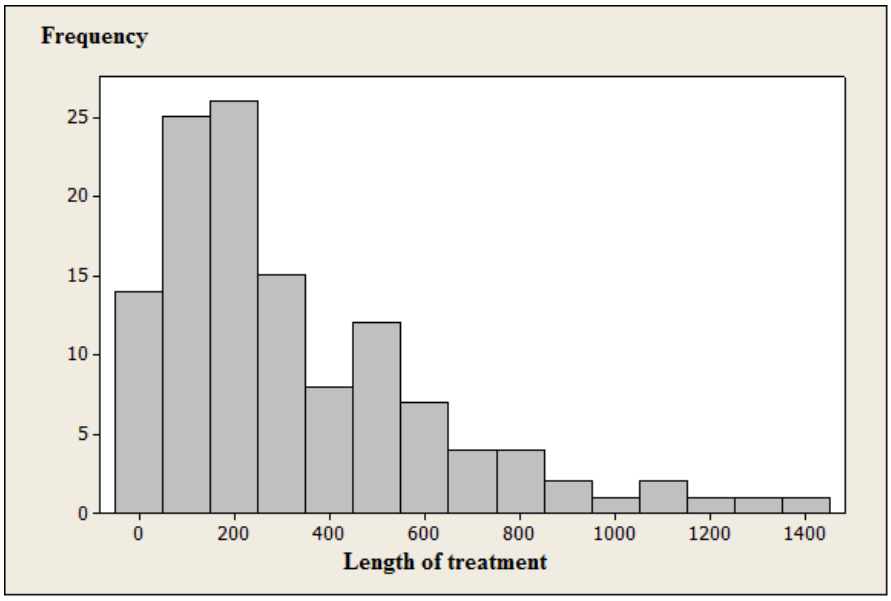


Figure 9.3. Histogram of treatment length. The X axis shows the length of treatment. The Y axis shows the observed frequency of treatments of a certain range.

As shown in Figure 9.3, most of the execution lengths are less than 600. Longer sequences represent fewer than 17% of patients. 83% of patients have followed CR treatment programs shorter than 600 task executions per patient. This includes 103 of the 123 initial patients. Therefore in our model we propose an equal-sized rectangular data matrix considering ℓ = 600 executions, 83% of those patients followed shorter sequences of CR treatments. This transforms our original variable length matrix into a rectangular matrix χ for easier treatment.

Next, the matrices to evaluate the effect of the treatment are built:

In our particular application, $\Delta = (\Delta_A, \Delta_M, \Delta_{EF})$ is composed of three normalized effect indexes, each one evaluating improvement in one cognitive function. As mentioned in Section 1.4.3. the hospital uses a specific battery of tests to evaluate the state of the patient before and after treatment. The NAB battery includes a total of 28 items. Together with the experts, the items most specifically focused on for evaluation of each of the cognitive functions were identified, as shown in Table 9.4.

Test	Item	Cognitive Function
Continuous Performance Test	OMI	A
	COMI	
	CPT	
Trial Making Test	TMTA	
WAIS-III Selective	VWAIS	
Trial Making Test	TMTB	
Rey Auditory Verbal Learning Test	RAV075	M
	RAV015	
	RAV015R	
WAIS-III-Visuo-spatial	CUBES	EF
Stroop Test	INTER	
Wisconsing Card Sorting Test	TERR	
Letter Fluency Test	PMR	
Wisconsing Card Sorting Test	CAT	

Table 9.4. Selected tests and items targeting specific cognitive functions

As all items evaluate between [0..4] the simple mean is used as a measure of the cognitive function performance of the patient either before or after treatment. Thus the Δ components are built as the pre – post difference, using those indicators.

According to (Hart et al., 2005) a global index for each cognitive function is created as the average scoring in all items that refer to that cognitive function. As all items indicate higher impairment with higher values and it is expected that patients improve along treatment, differences between scoring after and before the treatment are expected to be positive. For this reason the components of Δ are defined as:

$$\Delta_A = \frac{OMI_f + COMI_f + CPT_f + TMTA_f + VWAIS_f + TMTB_f}{6} - \frac{(OMI_0 + COMI_0 + CPT_0 + TMTA_0 + VWAIS_0 + TMTB_0)}{6}$$

$$\Delta_M = \frac{RAV075_f + RAV015_f + RAV015R_f}{3} - \frac{(RAV075_0 + RAV015_0 + RAV15R_0)}{3}$$

$$\Delta_{EF} = \frac{CUBES_f + INTER_f + TERR_f + PMR_f + CAT_f}{5} - \frac{(CUBES_0 + INTER_0 + TERR_0 + PMR_0 + CAT_0)}{5}$$

9.4.2 Descriptive Analysis

As a first step in this phase the construction of the frequency plot of the first ℓ columns and ℓ frequency of tasks for χ is performed.

In Figure 9.4 below it can be seen that less frequent tasks (shown in gray and labeled as OTHER) are more frequently executed at the beginning of the treatments, but as the treatment is longer their frequency decreases.

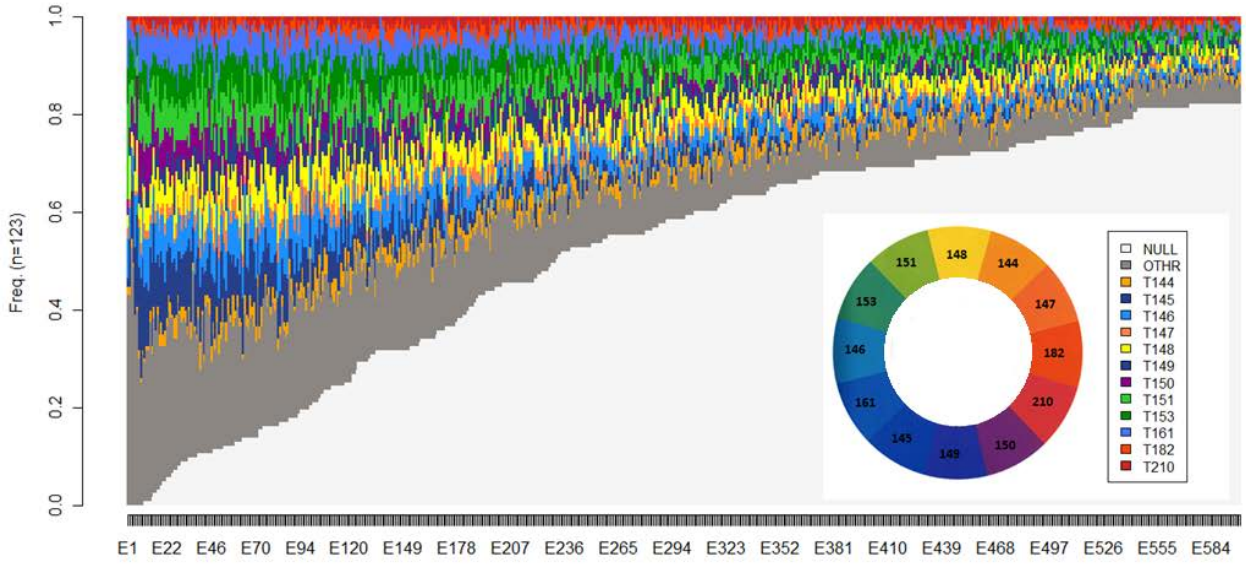


Figure 9.4. Frequency of the 12 selected tasks along the treatments. The X axis provides the time in the CR program where the task was executed. The tasks are identified by colors according to legend and circular pantone.

Next, the construction of the heatmap of the first ℓ columns and f frequency of tasks for χ is performed.

Figure 9.5 below suggests the need for a method of grouping tasks that enables execution patterns to be identified. Each task is represented with the colour gradient as in Figure 9.4 but no structure can be identified either in the figure displayed or by performing permutations of patients along the vertical axis.

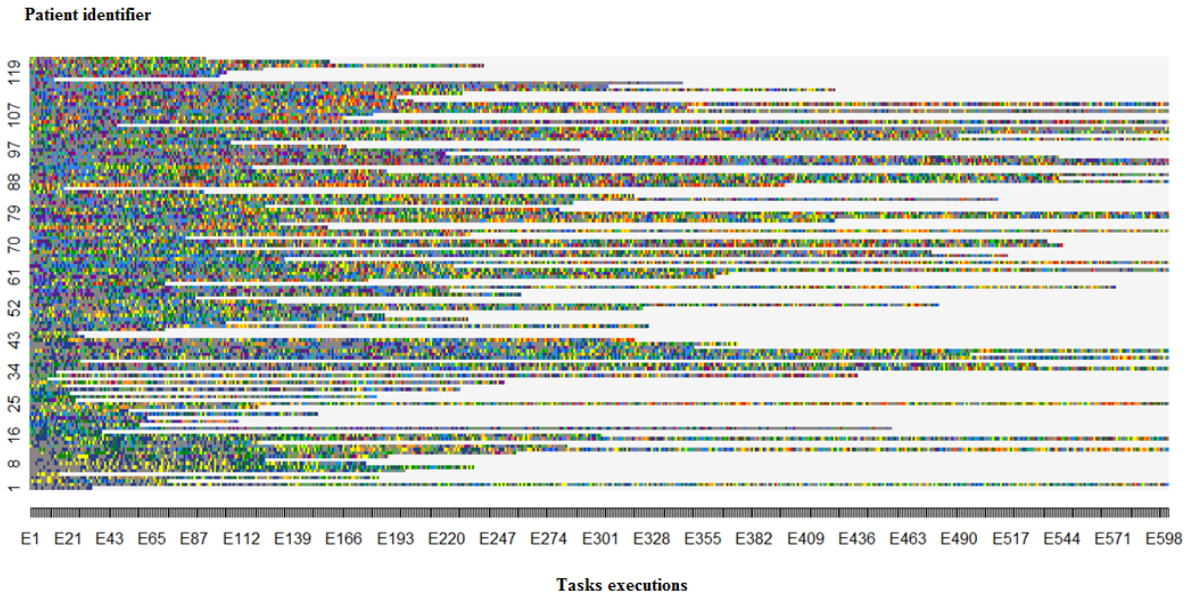


Figure 9.5. Heatmap for individual treatments representing the 12 selected tasks. The Y axis represents individual patients (identified 1 to 123) and the X axis shows the position of the task along the treatment. The first 600 tasks executions are represented.
Task colors are the same as in Figure 9.4. The heatmap shows the different treatment lengths (e.g. patient *i*=1 at the bottom of the heatmap followed a treatment comprising approximately 30 tasks, whereas patient *i*=2 immediately above followed a treatment longer than 600 tasks). This heatmap lacks any recognizable structure or pattern in the task sequences (e.g. *i*=1 patient shows only blue and gray tasks while *i*=2 shows all 14 colors in apparently random order).

Next, the construction of the heatmap of the first ℓ columns and f frequency of tasks for χ^a is performed.

Figure 9.6 below provides a heatmeap with a lower granularity of information. Tasks are grouped per cognitive function addressed and a color is assigned to every group. Instead of

representing the specific tasks executed at every step of the CR program, the cognitive function addressed is displayed (green represents memory tasks, red represents attention tasks, and yellow represents executive function tasks; gray points to other non-frequent tasks not considered at this stage of the analysis). As in Figure 9.4, execution patterns cannot be identified from this figure, even when grouped by targeted function .

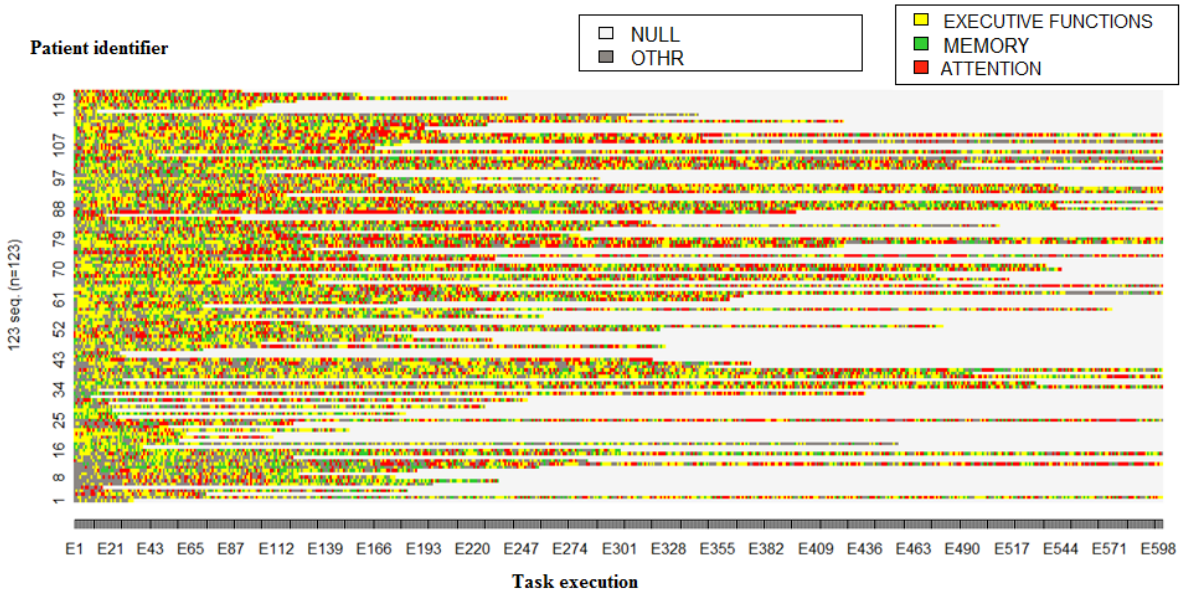


Figure 9.6. Heatmap of individual treatments representing the cognitive function addressed by each executed task. X axis represents the position of the task along the treatment sequence. Y axis represents the individual patients.

9.4.3 Prior Expert Knowledge Acquisition

Domain knowledge is represented by means of IF-THEN rules. A team of licensed and doctoral level staff with extensive education and experience has been participating in this Knowledge Acquisition stept. The team was made up of 4 members from the Acquired Brain Injury Unit at Institut Guttmann Neurorehabilitation Hospital. One of them is the medical doctor (in charge of the medical leading of the team) and three neuropsychologists as specialized consultants in diagnosis and treatments of the three main cognitive functions addressed during CR programs (i.e. attention, memory and executive functions). Experts

expressed knowledge regarding what is considered a long or short treatment according to their clinical experience in terms of the number of tasks it comprises.

KB = { r1: *if* SeqLength < 450 *then* SHORT,
r2: *if* SeqLength >480 *then* LONG
}

9.4.4 Clustering Phase

The software **KLASS v86** was the data mining platform for the ***ClBR*** algorithm executions (Gibert et al., 1998).

ClBR was run with the Ward method, Gibert's mixed distance (Gibert et al., 1998) and KB as referred knowledge base. The resulting dendrogram is shown in Figure 9.7.

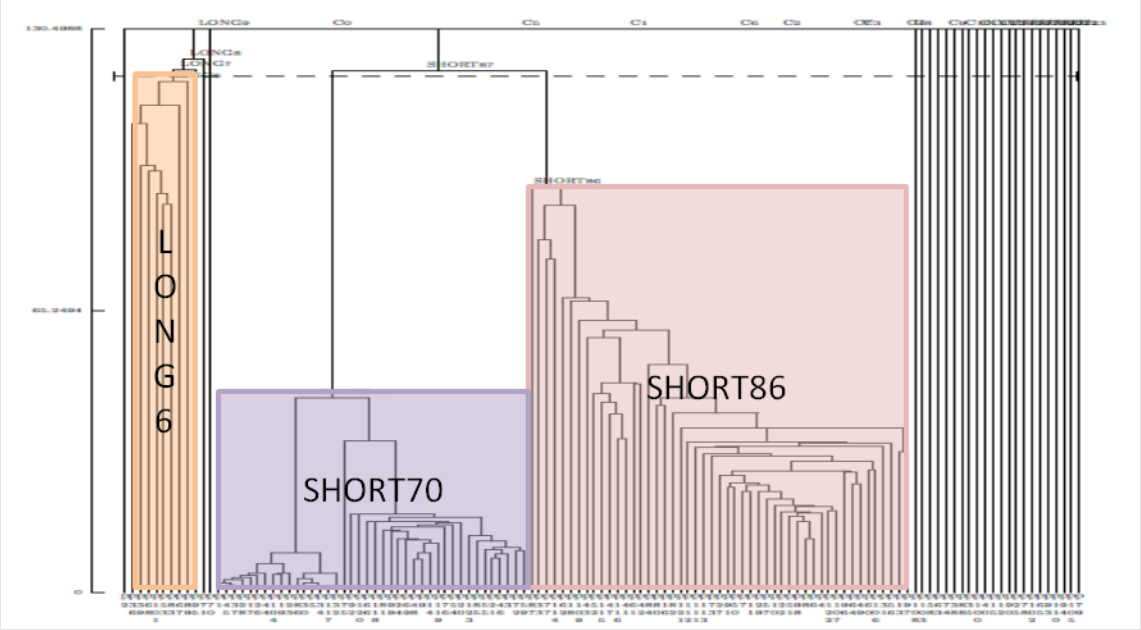


Figure 9.7. Dendrogram obtained by the *ClBR*. Leaves represent individual patients; internal nodes of the tree represent the intermediate clusters sequentially built along the Clustering process; the height of each node is proportional to the homogeneity of the class. The horizontal dashed cut of the tree represents a partition of the dataset in a set of classes (highlighted in different colors). Patients on the right-hand side are singletons that do not group within any class and remain isolated. They are not considered at this stage of the analysis.

The Calisnki-Harabasz method (Calinsky-Harabasz, 1974) suggests a cut in 29 classes for which 26 are singleton and 3 main groups are conformed. One contains most of the patients

satisfying r2 and the other two subdivide patients satisfying r1 into two subgroups. The classes obtained are shown in Table 9.5.

Class label	nc
SHORT70	40
SHOT86	49
LONG6	8

Table 9.5. Number of patients in each identified class

9.4.5 Split into Classes

According to experts, singletons were disregarded as exceptional cases to be carefully analyzed one by one.

The data matrix is then divided into 3 submatrices according to the three identified classes: SHORT70, SHORT86, and LONG6.

9.4.6 Visualization per Classes

Local heatmaps are built for every class and displayed in Figure 9.8 to Figure 9.10. It can be seen that SHORT70 (Figure 9.8) class contains patients with shorter treatments, fewer than 150 executions; SHORT86 class (Figure 9.9) contains patients with intermediate length treatments of between 150 and 460 executions; and LONG6 catches all those patients following the longest treatments, with more than 460 executions.

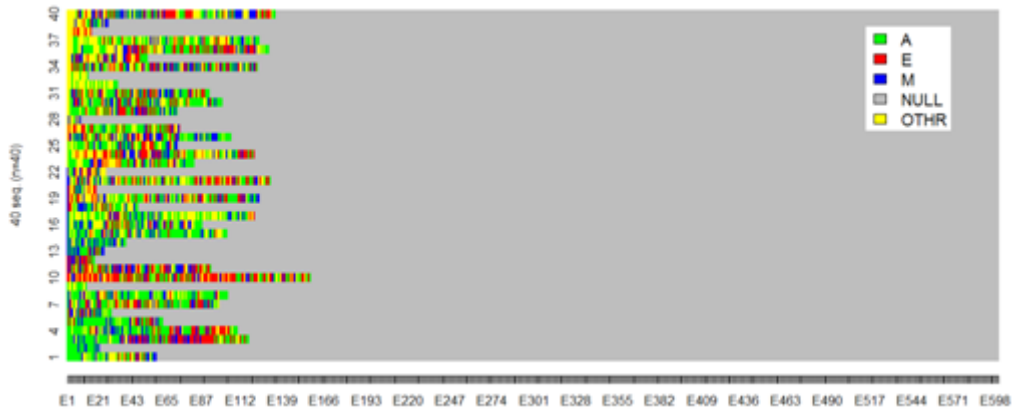


Figure 9.8. Heatmap representing SHORT70 Class executions

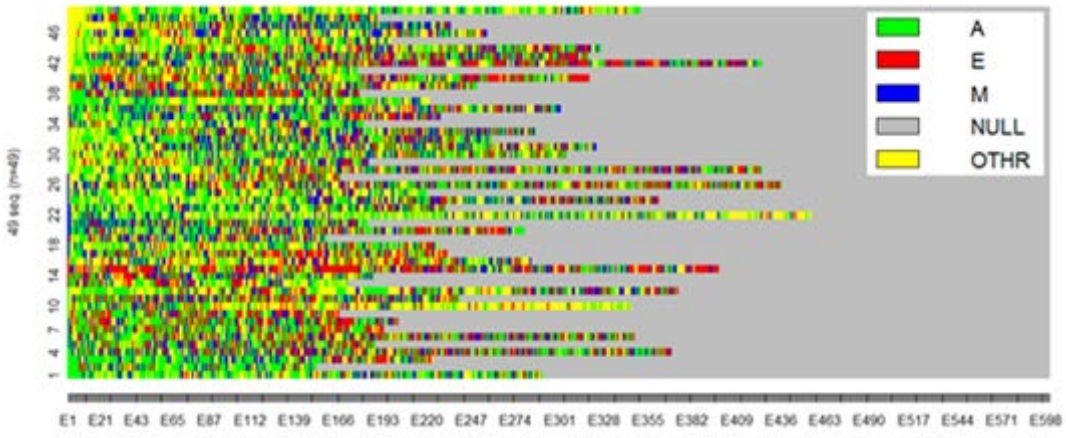


Figure 9.9 Heatmap representing SHORT86 Class executions

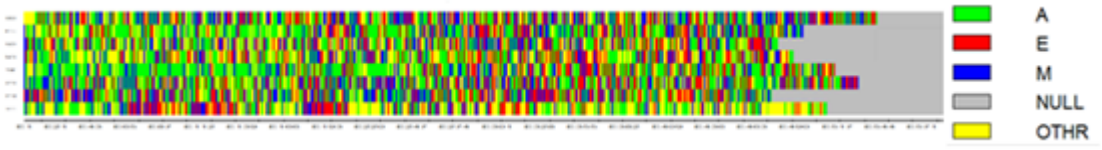


Figure 9.10 Heatmap representing LONG6 Class executions

9.4.7. Find Motifs per Class

Toolbox for Motif Discovery (TMOD) version 1.1.1 is the framework for motifs discovery (Hanchang et alt 2010) run on a 3.4 GHz Pentium IV computer with 2 GB of RAM.

MEME (Bailey and. Elkan, 1995) implementation was executed in TMOD with the following input parameters:

Sequences (in FASTA format) of cognitive functions targeting the different task executions. Therefore input data to MEME is the same as for sequential pattern mining.

Alphabet: DNA sequences or protein sequences. Run with DNA sequences which must contain only ACGT letters. For our particular context we adopt the following representation:

- A: represents attention tasks,
- C: memory tasks,
- T: executive functions tasks.

Distribution: how the occurrences of motifs are distributed in the input sequences. Run with Any Number of Repetitions (ANR) in this case MEME assumes each sequence may contain any number of non-overlapping occurrences of each motif. This option is useful because we suspect that motifs repeat multiple times within a single sequence.

Motif width: run with motifs' length parameter ranging from l= [6.20] (larger motifs are visually difficult to analyze and those shorter than 6 were discarded by domain experts because a CR session rarely includes fewer than 6 task executions).

EM algorithm: The number of iterations of EM to run from any starting point (run with default value= 50)

Performance measure: MEME searches for the motif with the smallest E-value. The E-value of the motif is an estimate of the number of motifs (with the same width and number of occurrences) that would have an equal or higher log likelihood ratio if the training set sequences had been generated randomly according to the (0-order portion of the) background model. An accepted threshold for E-value is 0.005 (Bailey and. Elkan, 1995).

MEME is then run with the parameters specified above for each identified cluster sequences. Figure 9.11 shows the obtained sequence of logos.

$$M = \left\{ \mathrm{M}_{\mathrm{C}_1} \ldots \mathrm{M}_{\mathrm{C}_\zeta} \right\}$$

$$\forall\, \mathrm{C} \in \mathrm{P}\ \mathrm{M}_{\mathrm{C}} = \{ \mathrm{m}^{6}_{\mathrm{SHORT70}} \ldots\ \mathrm{m}^{20}_{\mathrm{SHORT70,}} \mathrm{m}^{6}_{\mathrm{SHORT86}} \ldots\ \mathrm{m}^{20}_{\mathrm{SHORT86,}} \mathrm{m}^{6}_{\mathrm{LONG6}} \ldots\ \mathrm{m}^{20}_{\mathrm{LONG6}} \}$$

With a total of 14 motifs for each of the 3 analyzed classes.

The matrix with the E-values is:

e_C^l	Length/Class	SHORT70	SHORT86	LONG6
e_C^6	6	4.3e+002	2.5e-018	2.4e+004
e_C^7	7	7.5e-003	2.7e-026	1.0e+004
e_C^8	8	1.0e-002	1.8e-028	4.6e+003
e_C^9	9	5.4e-004	2.3e-032	4.5e+002
e_C^{10}	10	1.3e-003	7.7e-039	9.0e+001
e_C^{11}	11	1.7e-003	3.2e-046	1.6e+002
e_C^{12}	12	2.9e-006	4.3e-049	5.2e+001
e_C^{13}	13	3.2e-004	2.4e-048	9.0e+001
e_C^{14}	14	8.6e-007	2.8e-055	8.4e+001
e_C^{15}	15	6.8e-005	5.6e-051	6.8e+002
e_C^{16}	16	9.4e-007	2.6e-049	9.5e+001
e_C^{17}	17	1.5e-005	2.5e-045	5.2e+002
e_C^{18}	18	2.3e-007	3.0e-045	2.0e+001
e_C^{19}	19	6.0e-006	3.1e-034	1.3e+002
e_C^{20}	20	6.0e-006	**1.9e-039**	2.5e+002

For each motif a π_C^l matrix is given which will be inputed to the motif viewer. Here the $\pi_{\text{SHORT86}}^{20}$ is shown:

A	C	G	T
0.435897	0.000000	0.000000	0.564103
0.769231	0.025641	0.000000	0.205128
0.384615	0.051282	0.000000	0.564103
0.000000	0.000000	0.000000	1.000000
0.000000	0.410256	0.000000	0.589744
0.000000	0.794872	0.000000	0.205128
0.102564	0.128205	0.000000	0.769231
0.102564	0.230769	0.000000	0.666667
0.025641	0.230769	0.000000	0.743590
0.025641	0.179487	0.000000	0.794872
0.000000	0.000000	0.000000	1.000000
0.000000	0.153846	0.000000	0.846154
0.000000	0.256410	0.000000	0.743590
0.025641	0.179487	0.000000	0.794872
0.000000	0.384615	0.000000	0.615385
0.153846	0.358974	0.000000	0.487179
0.256410	0.282051	0.000000	0.461538
0.102564	0.512821	0.000000	0.384615
0.230769	0.461538	0.000000	0.307692
0.307692	0.205128	0.000000	0.487179

9.4.8 Determine a Level of Minimum Quality for Motifs

For convention $\alpha = 0.05$ is used.

9.4.9 Pruning Motifs: Retain more Frequent Motifs for Interpretation

For each motif π_C^l and l length, a weighted median of the E-values of the three classes is calculated, being the weighting factor the number of patients n_c of each class.

9.4.10 Visualize Motifs per Class

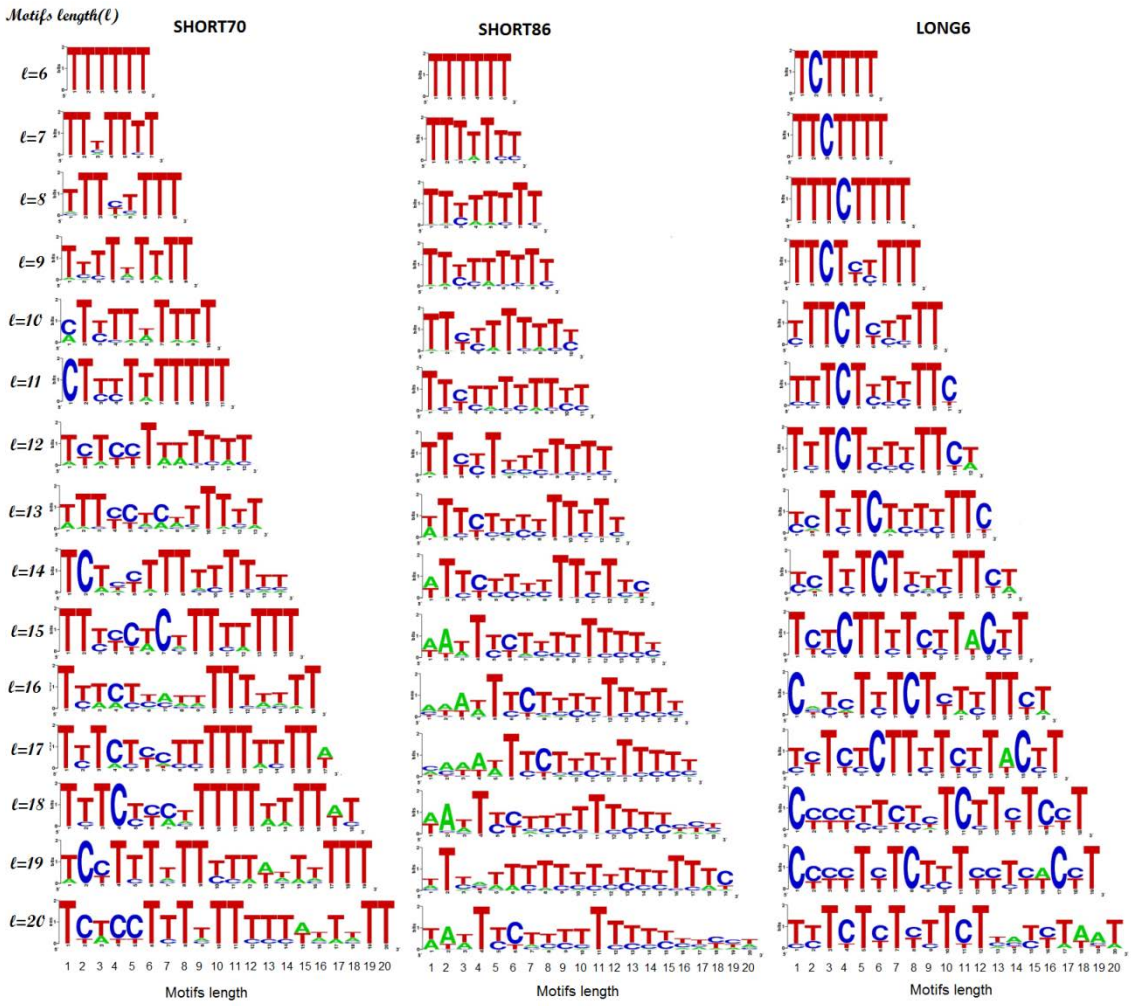


Figure 9.11 Sequence of logos by motif lengths per class. Motif length ranges from 6 to 20 tasks on the vertical axis.

Stability analysis.

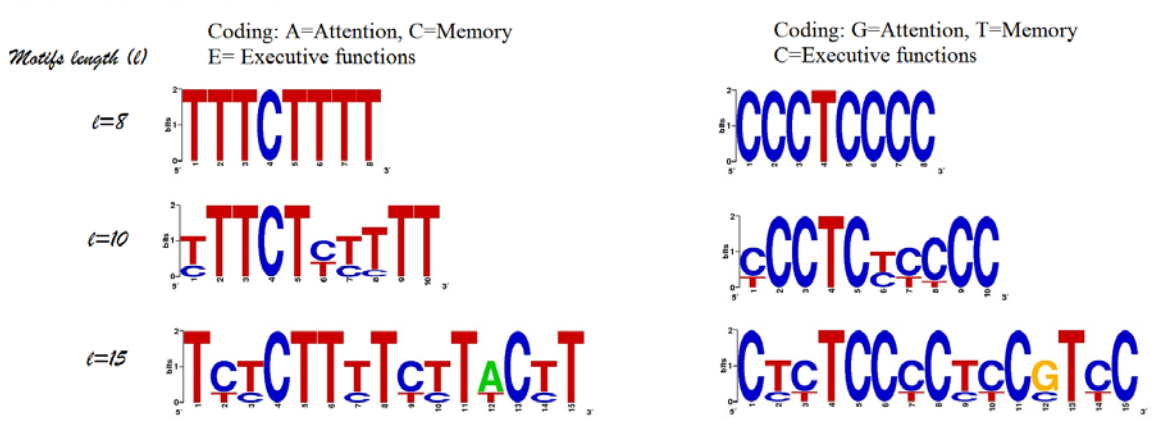


Figure 9.11b Sequence of logos by motif lengths per class. Motif length ranges from 6 to 20 tasks on the vertical axis.

Some motif discovery programs consider secondary and tertiary sequences, influencing the motif discovery, as ACGT is associated with nucleotid bases. This would be a problem in our context as the motifs could change if a different association was used between letters and cognitive functions. MEME is a basic motif discovery method which is not using this specific genetic knowledge. Thus, under another codifications, like G for attention, T for memory and C for executive functions the discovered motifs do not change as shown in Figure 9.11b for three motifs of different lengths, all from LONG6 class. Similar results are obtained for other motifs lenghts, and classes (SHORT70, SHORT86). Using MEME as underlying motif discovery method, different coding of activities do not affect identified motifs.

Motifs analysis leads to the following descriptions:

- SHORT70 class shows mainly task executions oriented to executive functions (represented as T) and some memory tasks (represented as C), mainly in the first part of the sequences.
- SHORT86 class includes fewer executive functions and memory tasks than the other classes but shows a higher number of attention tasks (represented as A) executed mainly at the begining of the identified motifs.

- In class LONG6 the number of memory tasks clearly increases and is often combined with executive function tasks. Eventually some attention tasks are performed at the end of the identified motifs.

9.4.11 Project all Other Illustrative Variables over the Clusters.

There are no significant differences in the characteristics of the patients for the three identified classes (GCS, age, PTA, gender and educational level), see Table 9.6 and Table 9.7 below where p-values for numerical and categorical variables are shown. Therefore possible differences in response to the treatments might be attributable to the task patterns performed along treatment.

GCS						
	Mean	**StD**	**Median**	**Q1**	**Q3**	**IQR**
SHORT70	6.27	2.91	6.00	5.00	7.00	2.00
SHORT86	6.04	2.63	6.00	4.00	7.00	3.00
LONG6	6.88	3.09	7.00	4.25	8.00	3.75
KW p-value	**0.667**					

AGE						
	Mean	**StD**	**Median**	**Q1**	**Q3**	**IQR**
SHORT70	32.80	8.20	33.00	27.00	40.50	13.50
SHORT86	31.65	7.99	31.00	26.00	39.00	13.00
LONG6	35.13	9.57	38.50	27.25	42.00	14.75
KW p-value 0.433						

PTA						
	Mean	**StD**	**Median**	**Q1**	**Q3**	**IQR**
SHORT70	84.3	36.9	74.00	54.00	123.00	69.00
SHORT86	156.7	209.7	89.00	78.50	149.00	70.50
LONG6	117.67	13.,65	124.00	102.00	127.00	25.00
KW p-value 0.176						

Table 9.6. Numerical variables Mean, Standard deviation, median, Q1, Q3, IQR and p-values (Kruskal-Wallis test) per class

	GENDER		EDU LEVEL			
	Female	Male	Elemen	Interm.	High	Total
SHORT70	13	27	19	14	7	40
SHORT86	12	37	27	14	8	49
LONG6	2	6	4	3	1	8
χ^2 p-value	0.691		0.949			
Fisher Exact test	0.7448		0.977			

Table 9.7. Categorical variables number of occurrences and p-values (χ^2 test) per class

9.4.12 Analyze the Effect of Executing Activities over the Different Areas of Impact

According to (Hart et al., 2005) a global index for each cognitive function is created as the average scoring in all items that refer to that cognitive function. The effect of the treatment over a certain cognitive function is measured as the value observed in the corresponding index when the difference between the score after treatment and the score before treatment is computed.

As introduced in Section 1.4.3, all NAB items are normalized to a 0 to 4 scale (where 0 = no affectation, 1 = mild affectation, 2 = moderate affectation, 3 = severe affectation and 4 = acute affectation). This normalization has been developed as a standardization method proposed at Institut Guttmann in QVidLab framework (presented in section 1.3.2). It is intended as an instrument of applied clinical research allowing for the standardization of different assessment tools for biological, psychological, and social factors (Gil Origüén, 2009). Thus, the post-pre difference measures a deficit reduction such that lower values in differences indicate a higher deficit reduction or, in other words, a positive response to the CR treatment.

Figure 9.12 shows multiple boxplots with conditional distributions in the three indexes of cognitive functions before and after treatment

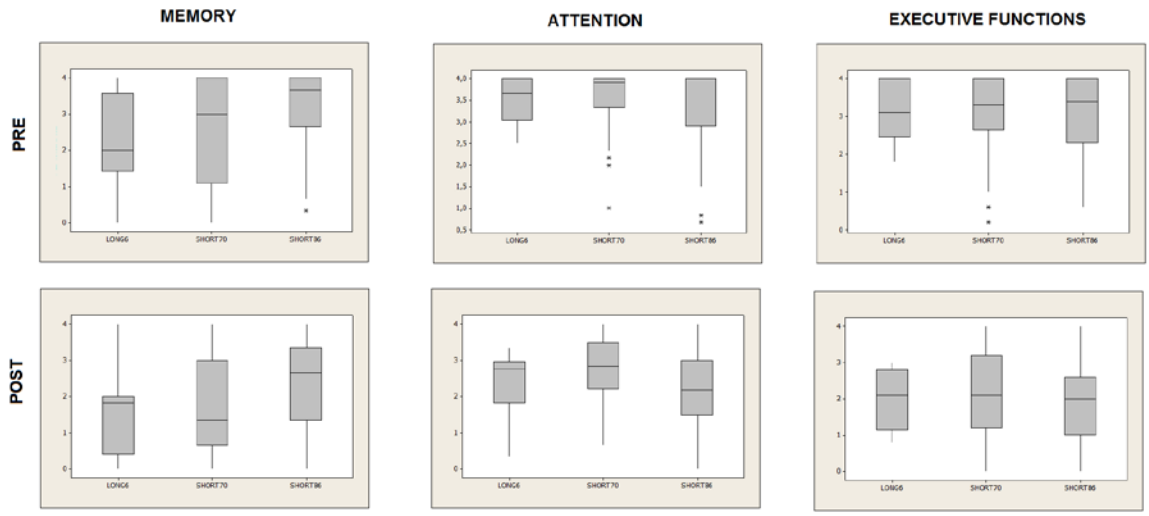


Figure 9.12 Multiple boxplots of Improvement versus class and cognitive function.

Figure 9.13 shows the multiple boxplots with the conditional distributions of effect indexes (Δ_A for attention, Δ_M for memory and Δ_{EF} for executive functions) versus the classes. Each graph represents the different effects on each class of treatment over a certain cognitive function. The first interesting observation is that all groups improve (deficit decreases) after treatment and the effects are all below 0 on average. The dimension which is placed around more negative values is attention, while memory seems to be the one with the least improvement for all groups. On the other hand, it can be seen that SHORT86 class is the one with better treatment results regarding attention, while behaving very closely to class SHORT70 regarding memory and executive functions. Also, it seems that class LONG6 is more resistant to treatment than others, especially regarding memory.

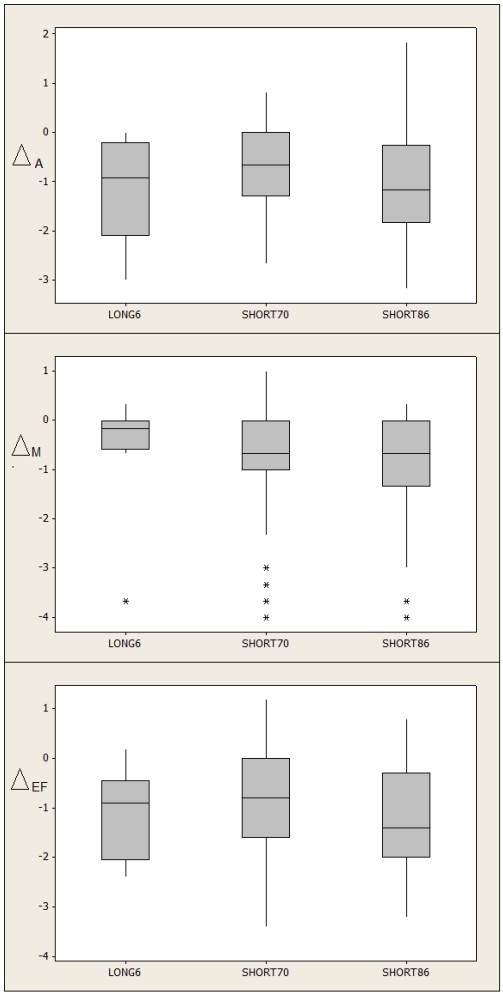


Figure 9.13 Multiple boxplots of Improvement versus class and cognitive function. For each cognitive function, a multiple boxplot of the corresponding improvement index Δ versus classes is visualized. Every boxplot displays between the minimum and maximum value of each Δ, the box indicates the interval between first and third quartile, whereas the horizontal line through the box indicates the median.

9.4.13 Build Final Interpretation.

Crossing the obtained profiles with the motifs and the effects of therapy it appears that:

SHORT70 represents short-term treatments, no longer than 150 task executions mainly oriented to executive functions preceded in some cases by memory tasks, mainly in the first part of the sequences. These persons show better response to treatment mainly in attention and executive functions rather than in memory, experiencing an intermediate improvement in level of attention compared with other classes.

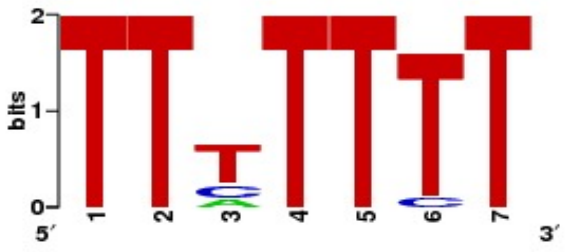
SHORT86 represents intermediate duration treatments, with no more than 460 task executions including a higher number of attention tasks executed mainly at the beginning of the sequences. Persons in this class show a higher recovery in attention than in other functions, being the group with better results for the treatment regarding attention.

LONG6 represents long-term programs including more than 460 task executions with a higher proportion of memory tasks, often combined with executive function tasks and eventually some attention tasks at the end of the sequences. However, the persons in this class are more resistant to treatment than other classes in memory and attention.

9.5 Identification of the General Pattern of the Motifs per Class

Following the indications provided in section 5.1 the regular expressions associated to all logos is computed for $\gamma = 0.25$. In Annex detailed intermmediate steps are enclosed. Here a couple of examples are detailed and final results displayed as regular expressions.

SHORT70 l=7 $\gamma = 0.25$



$\pi^{7}_{\text{SHORT70}} =$

A	C	G	T	p
0.000000	0.000000	0.000000	1.000000	1
0.000000	0.000000	0.000000	1.000000	.
0.153846	0.230769	0.000000	0.615385	.
0.000000	0.000000	0.000000	1.000000	.
0.000000	0.000000	0.000000	1.000000	
0.000000	0.076923	0.000000	0.923077	
0.000000	0.000000	0.000000	1.000000	7

$\pi^{7^*}_{\text{SHORT70}} =$

A	C	G	T	p
0.000000	0.000000	0.000000	1.000000	1
0.000000	0.000000	0.000000	1.000000	.
0.000000	0.000000	0.000000	0.615385	.
0.000000	0.000000	0.000000	1.000000	.
0.000000	0.000000	0.000000	1.000000	
0.000000	0.000000	0.000000	0.923077	
0.000000	0.000000	0.000000	1.000000	7

$W^7_{\text{SHORT70}} ==$ ({T},{T},{T},{T},{T},{T}{T})

$S = T^7$

Table 9.8 below summarizes the different expressions for 3 classes varying l (detalis in Annex)

l	***SHORT70***	***SHORT86***	***LONG6***
7	T^7	T^7	T^2CT^4
8	$T^3(C\|T)T^4$	T2(C|T)T5	T^3CT^4
9	$T(C\|T)T^2(A\|T)T^4$	$T^2(C\|T)T^6$	$T^2CT(T\|C)^2T^3$
10	$(A\|C)T(C\|T)T^2(A\|T)T^4$	$T^2(C\|T)^2T^5(C\|T)$	$(C\|T)T^2CT(C\|T)^2T^3$
11	$CT(C\|T)^2T^7$	$T^2(C\|T)^2T^5(C\|T)^2$	$T^3CT(C\|T)T(C\|T)T^2C$
12	$T(C\|T)T(C\|T)^2T(A\|T)^2T^4$	$T^2(C\|T\ ^2T(C\|T)^3T^4$	$T^3CT(C\|T)T(C\|T)T^2C(A\|T)$
13	$(A\|T)T^2(C\|T)^2T(A\|C)(A\|T)T^6$	$T^2(C\|T\ ^2T(C\|T)^3T^5$	$(C\|T)CT(C\|T)TCT(C\|T)T(C\|T)T^2C$
14	$TCT(C\|T)^2T^8(C\|T)$	$(A\|T)T(C\|T)^6T^2(C\|T)T(C\|T)^2$	$(C\|T)CT(C\|T)TCT(C\|T)T(C\|T)T^2(C\|T)(A\|T)$
15	$T^2(C\|T)^2C(A\|T)CT^8$	$(A\|T)A(A\|T)T(C\|T)^2T(C\|T)^3T^2(C\|T)^3$	$T(C\|T)^2CT^2(C\|T)T(C\|T)^2T(A\|T)C(C\|T)T$
16	$T(C\|T)(A\|T)(A\|C)(C\|T)^2(A\|C)(A\|T)^2$ $T^3(A\|T)T^3$	$(A\|C)(A\|T)^3T(C\|T)^2T(C\|T)^3T^2(C\|T)^3$	$C(A\|C\|T)(C\|T)CT(C\|T)TCT(C\|T)T(C\|T)T^2(C\|T$ $)T$
17	$T(C\|T)TCT(C\|T)CT^9(A\|T)$	$(A\|C)^2(A\|T)^3T^2CT(C\|T)^2T^4(C\|T)^2$	$(C\|T)^2T(C\|T)^2CT^2(C\|T)T(C\|T)^2T(A\|T)C(C\|T)T$
18	$T^3C(C\|T)(A\|C)T^5(A\|T)T^3(A\|T)T$	$(A\|T)A(A\|T)T(C\|T)^5T^3(C\|T)^6$	$C(C\|T)^4T(C\|T)^3TCT^2(C\|T)TC(C\|T)T$
19	$TC(C\|T)T^6(C\|T)T^2(A\|T)^2T^5$	$(A\|T)^2(C\|T)^2(A\|T)^2T^2(C\|T)T(C\|T)^5T^3(C$ $\|T)$	$C(C\|T)^3T(C\|T)TCT(C\|T)T(C\|T)^2T(C\|T)(A\|T)C$ $(C\|T)T$
20	$TC(A\|T)C^2T\ ^9(A\|T)^2T(A\|T)T^2$	$(A\|T)^3T(C\|T)^2T^6(C\|T)T(C\|T)^2(A\|C\|T)($ $C\|T)^2(A\|T)$	$T(C\|T)TCT(C\|T)T(C\|T)T^2CT(C\|T)(A\|C)(C\|T)^2$ $T(A\|T)AT$

Table 9.8. Obtained expressions for each identified class with lengths varying from 7 to 20

9.6 Identification of NRRs for each task

9.6.1. Tasks Executions Targeting the Same Cognitive Function: NRRMR Method Application

9.6.1.1 Visual Identification of NRR

PREVIRNEC© platform includes 17 tasks addressing primarily the attention function, 59 addressing primarily memory, and 20 addressing primarily executive functions. During this CR treatment, the total number of task executions is 41010 (15475 targeting attention, 14557 memory, and 10978 executive functions). Figure 9.14 shows FT-SAP (γ=0.8 left column and γ=0.9 right column) for every execution of tasks grouped by CR functions. The top pair of plots corresponds to the execution of attention tasks, the middle pair to memory tasks, and the bottom pair to executive functions. Three different responses to CR treatment patterns can be identified according to how improvement points are distributed. Attention tasks are grouped on medium to high values of Results and medium to low values of number of executions. Memory is more uniformly spread from low to high values of results; executions are all over the plot and executive functions are a mix of the above patterns with concentration on high values and also for specific lower values of results and executions.

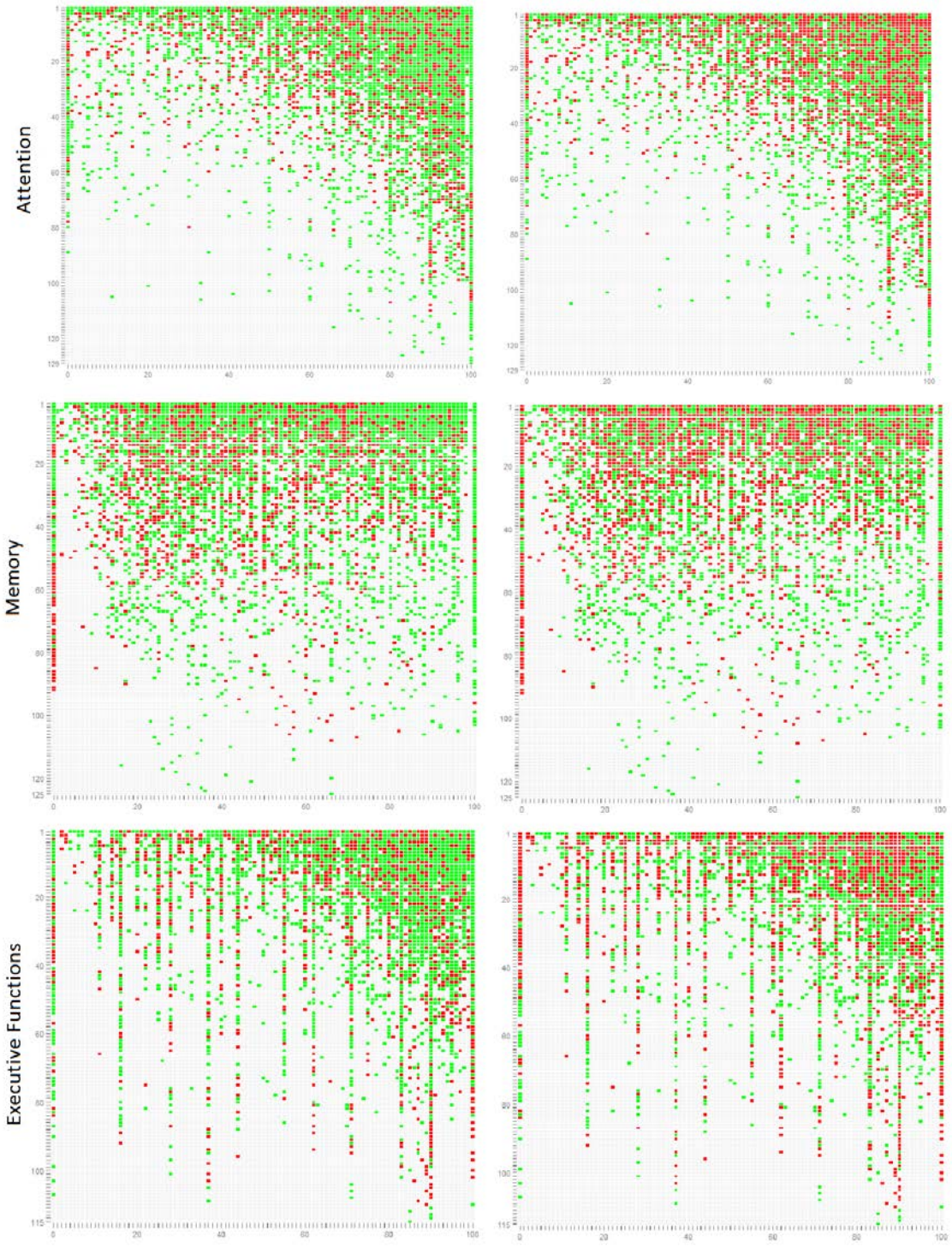


Figure 9.14. FT-SAP for each cognitive function: attention (top), memory (middle) and executive functions (bottom), γ=0.8 left column and γ=0.9 right column

9.6.1.2 Analytical Identification of NRR

Task Execution Targeting Attention Cognitive Function.

Methods presented in section 6.3.4 and 6.4.1 are applied for the analytical identification of NRRs. The first plot in Figure 9.14 (attention tasks with γ=0.8) is now analyzed using the method presented in section 6.4.1 to identify maximum zones of improvement for every execution of attention tasks, allowing for a tolerance of 2 elements. The results obtained (graphically represented in Figure 9.15) are as follows:

γ=0.8
[topLeftx,topLefty,botRightX,botRightY] = [11, 87, 20 ,88]
Area=20

Tolerance=2
γ=0.8
[topLeftx,topLefty,botRightX,botRightY] = [16, 98, 25, 100]
Area = 30

Leading to the following NRRs:

If (Results in [87,88] and Repetitions in [11,20] then P(Improvement)>=0.8
If (Results in [98,100] and Repetitions in [16,25] then P(Improvement)>=0.8

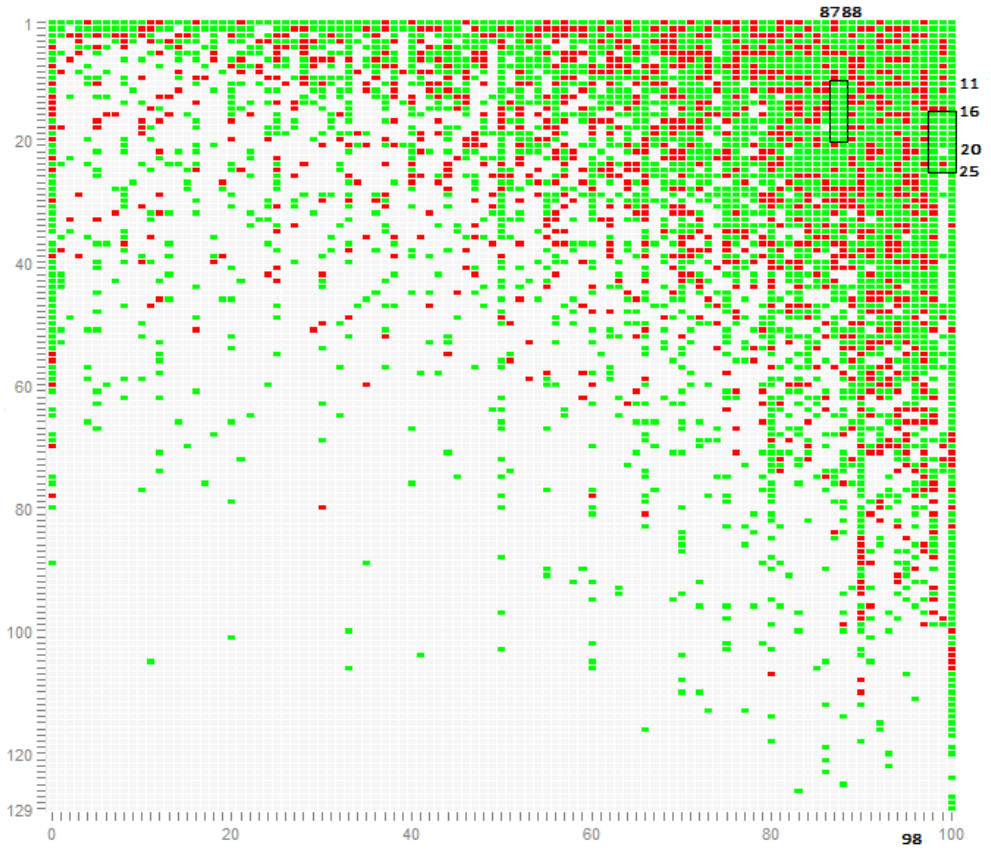


Figure 9.15 Analytical identification of NRR with and without user-defined tolerance

9.6.2 Individual Task Executions: SAP and NRRMR

9.6.2.1 Analysis of PREVIRNEC© Visual Memory Task Using [65,85] Basic Criterion

Nowadays in PREVIRNEC© a first hypothesis is being tested which considers that any participant patient has completed any task within NRR if the obtained result is in the 65 to 85 range, in INRR if it is less than 65, and in SNRR if it is higher than 85. As a reference, the current NRR used (Result ∈[65,85]) is visualized in a manually built SAP shown in Figure 9.16 with an overall sensitivity = 0.5660, overall specificity = 0.5012 and overall quality = 0.5022. The percentages represented in the SAP provide the empirical proportion of patients who improved following treatment in every area. About 60% of patients performing idTask=151 in NRR really improved. This is far from a random improvement. However, to evaluate the quality of the basic NRR used as a reference, a 2-sample probability test is used as described in section 6.5.2.

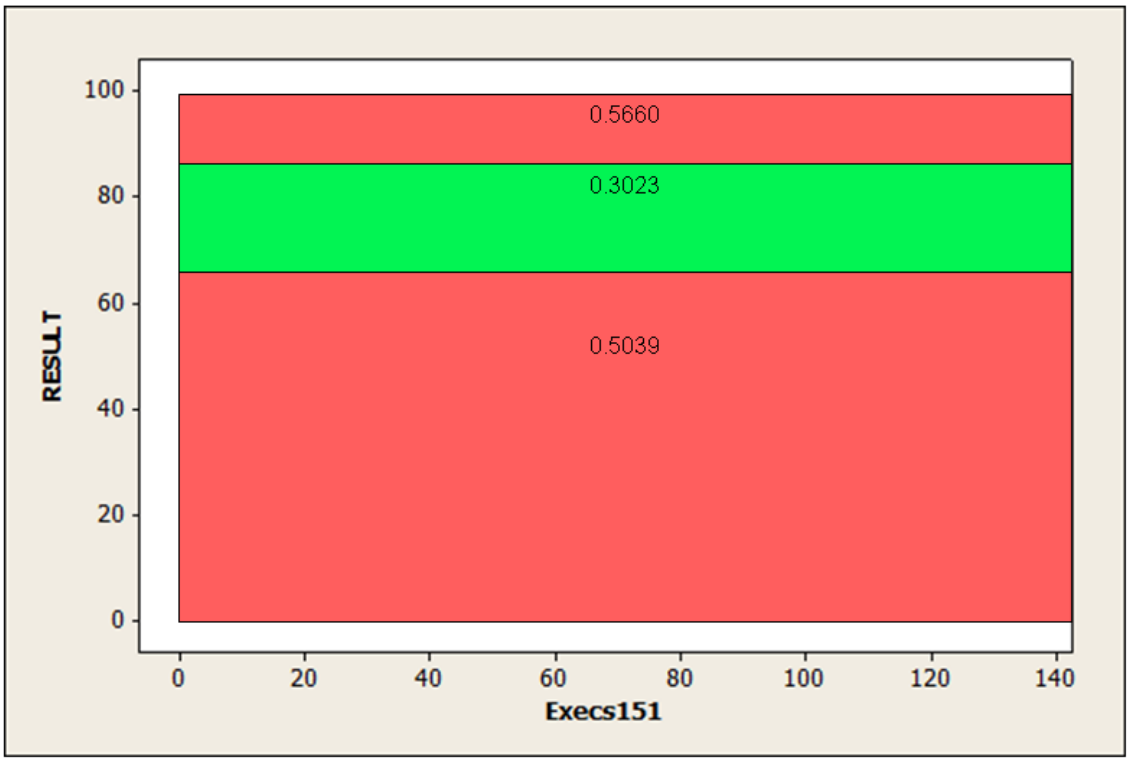


Figure 9.16 Vis-SAP for [68,85] NRR for idTask=151.

This enables verification of whether being in NRR really implies a significantly higher probability of improvement.

Table 9.9 contains the relevant information to compute the test by crossing the classification of the patients regarding two factors. Improving / Not improving and performing the idTask=151 within NRR or not (according to the basic criterion currently used).

In NRR / Improvement	Yes	No	Total
Yes	30	1652	1682
No	23	1661	1684
Total	53	3313	3366
^p(YES)	0.5660	0.4986	0.4997

Table 9.9 Contingency table for NRR with Result $\in$ [65,85] for idTask= 151

The test provided a result that is not statistically significant (z = 0.7316, p = 0.2323). This appears to be evidence that using the single result of the task is not enough to detect either the NRR or the ZRP zone. In the next sections, a model including the number of executions per task is tested.

9.6.2.2. Analysis of PREVIRNEC© Visual Memory Task Using Visualization-Based SAP (Vis-SAP)

The relationship between Results and Number of executions of patients is shown in Figure 9.17. Improving patients are shown in green and non-improving in red. Areas with a single category of patients (improving or not improving) are visually identified.

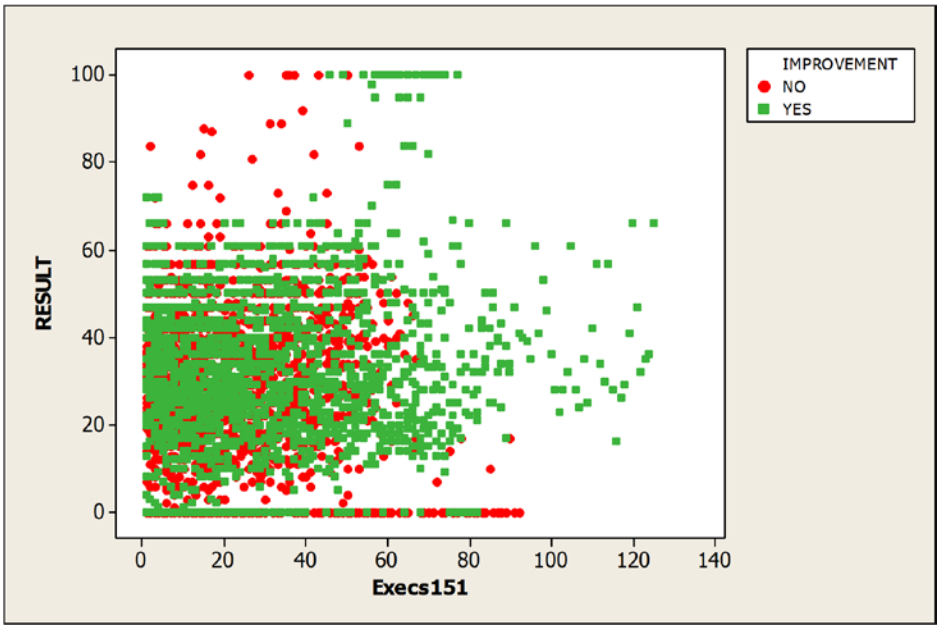


Figure 9.17. Letterplot of TaskExecs vs Result vs Improving/not for idTask=151

Looking at Figure 9.17, an area with no improvement is obtained for participants obtaining results higher than 70 and number of executions lower than 40 at the top left of the scatter plot. Furthermore, if the number of executions is higher than 60 and results higher than 20, a large region can be easily identified in which every participant is labeled in the improvement group, leading to the identification of a therapeutic range of results depending on the number of executions.

Two neat regions emerge from the above rules which can be expressed in the form of logical restriction rules, and visualized in an SAP diagram (shown in Figure 9.18 with an overall sensitivity = 0.939, overall specificity = 0.5523 and overall quality = 0.994).

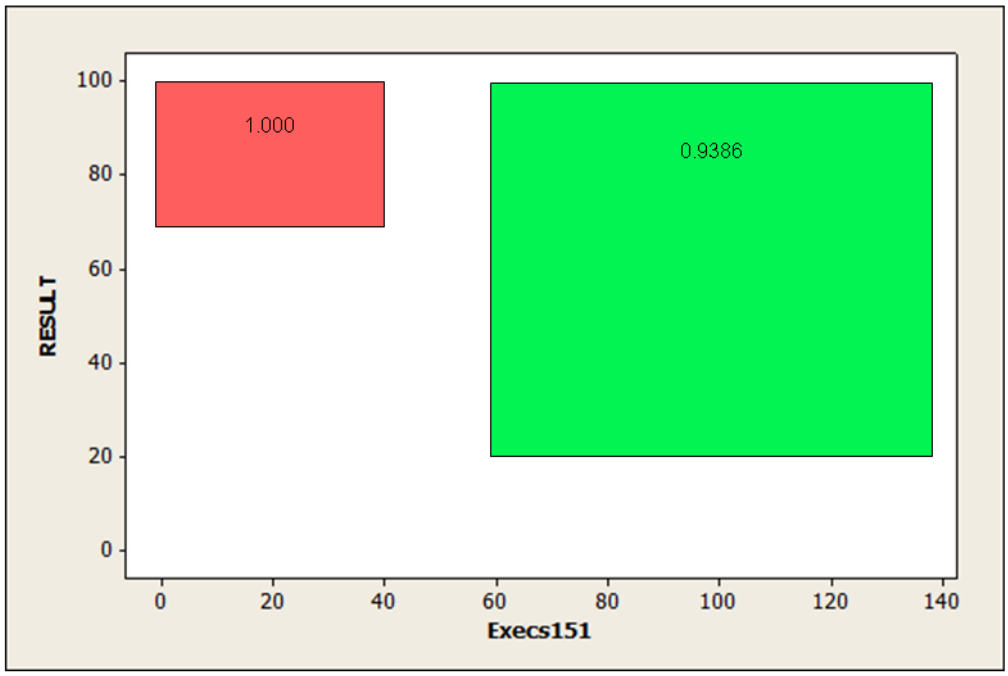


Figure 9.18: Vis-SAP for idTask=151

Identified rules are:

(Execs151 <= 40) AND (Res>70) → Not NRR

(Execs151 > 60) AND (Res>20) → NRR

The quality of the induced definition for NRR is assessed by means of the 2-sample proportion test (described in 6.5.2). Table 9.10 contains the relevant information to compute the test by crossing the classification of the patients regarding two factors. Improving / Not improving and performing the idTask=151 within NRR or not (according to the rules directly induced over the SAP). The results are z=12.42, p<<< 0.00001 is statistically significant.

In NRR / Improvement	Yes	No	Total
Yes	199	1483	1682
No	13	1671	1684
Total	212	3154	3366
^p(YES)	0.9386	0.4701	0.4997

Table 9.10 Contingency table for VIS-SAP TR for idTask= 151

In this case we can ensure that repeating idTask = 151 more than 60 times and getting results higher than 20 in all executions is associated with a group of patients with a sensibly higher probability of improving visual memory function. Thus, upon the Vis-SAP criterion, NRR (task151) = Execs151 > 60 and Results > 20.

9.6.2.3. Analysis of PREVIRNEC© Visual Memory Task Using DT-SAP

Although the visual-based SAP method seems to produce good results, the NRR has been defined on the basis of the visual expertise of the data miner, and this approach is totally dependent on the ability of the data miner itself. This is why an attempt to find the NRR automatically from data is presented. A supervised classification method, the J48 decision tree is applied to the target dataset (3366 instances) for SAP generation. Experiments were conducted in Weka with J48 decision tree for default configuration input parameters

(confidence factor = 0.25), since default values produced the best accuracy (57.45%), precision and recall. As usual, 10-fold cross validation was used to evaluate the goodness of results. As shown in Figure 9.19, three leaves of the resulting tree are labeled as an improvement.

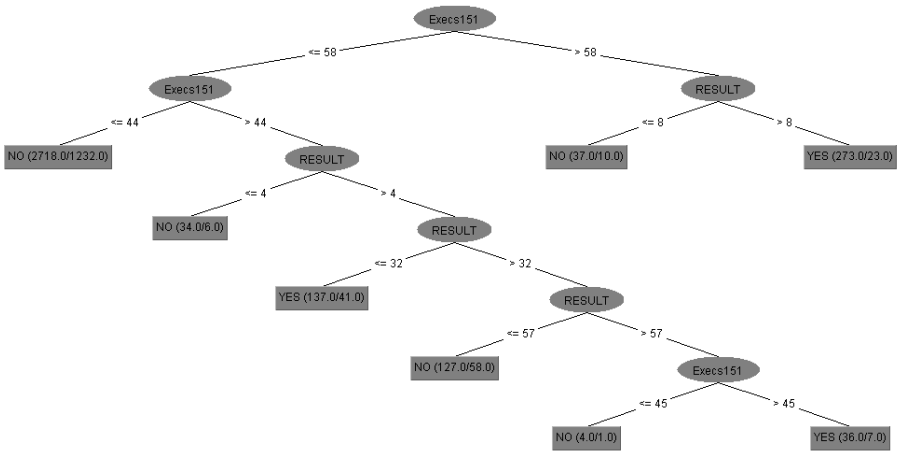


Figure 9.19 DT for idTask =151

Following the path from the root to those leaves, a condition for a patient's improvement can be induced, and the NRR defined as:

(Execs151<=58) AND (Execs151>44) AND (Res>4) AND (Res<=32) OR

(Execs151<=58) AND (Execs151>45) AND (Res>57) OR

(Execs151 > 58) AND (Res>8) → NRR

The quality of this induced criterion is assessed through the two samples proportion test presented in section 6.5.2 The relevant information for assessing the goodness of the NRR found for Task151 is presented in Table 9.11.

In NRR / Improvement	YES	NO	Total
Yes	375	1307	1682
NO	71	1613	1684
Total	446	2920	3366
^p(YES)	0.8408	0.4476	0.4997

Table 9.11 Contingency table for DT-SAP TR for idTask= 151

With this information, the test is statistically significant (z=13.45 p<<<<0.00001).

In fact, J48 identifies linear partitions of the space by means of linear separators that can be represented in an SAP diagram and compared with the zones identified in Figure 9.20 with an overall sensitivity=0.87, overall specificity=0.689, and overall quality= 0.716.

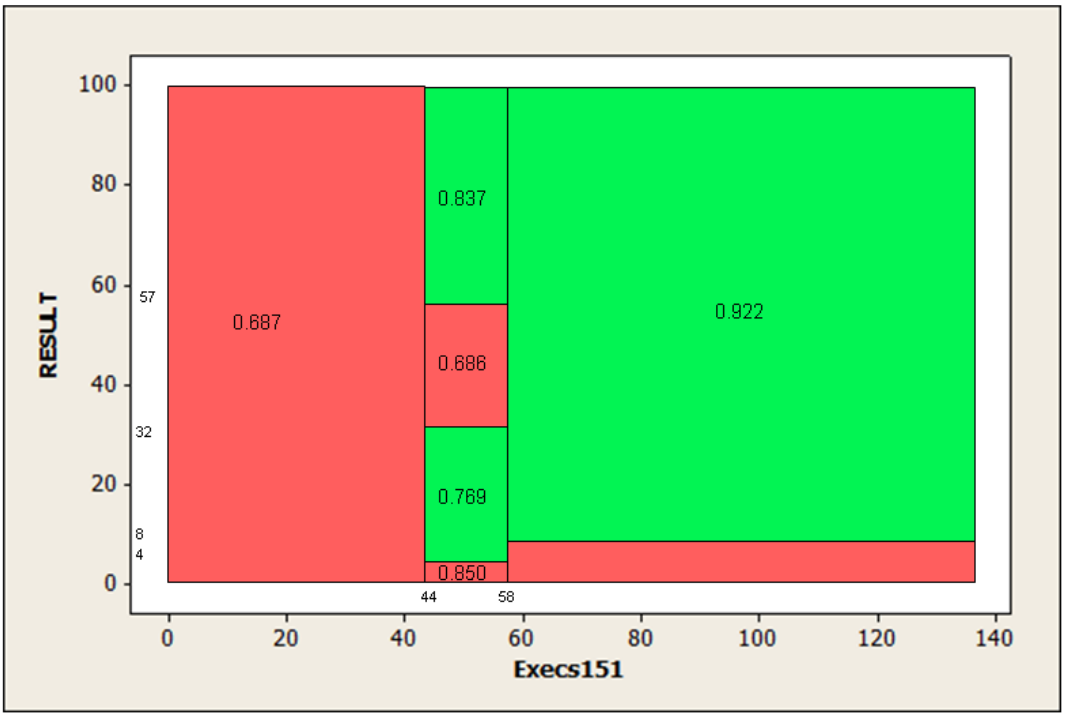


Figure 9.20 DT-SAP for idTask=151

Figure 9.21 shows the ROC curve of the three models. It can be seen that both Vis-Sap and DT-SAP perform significantly better than the experts-based current criterion. The Vis-SAP method provides a slightly better, also giving the higher global quality.

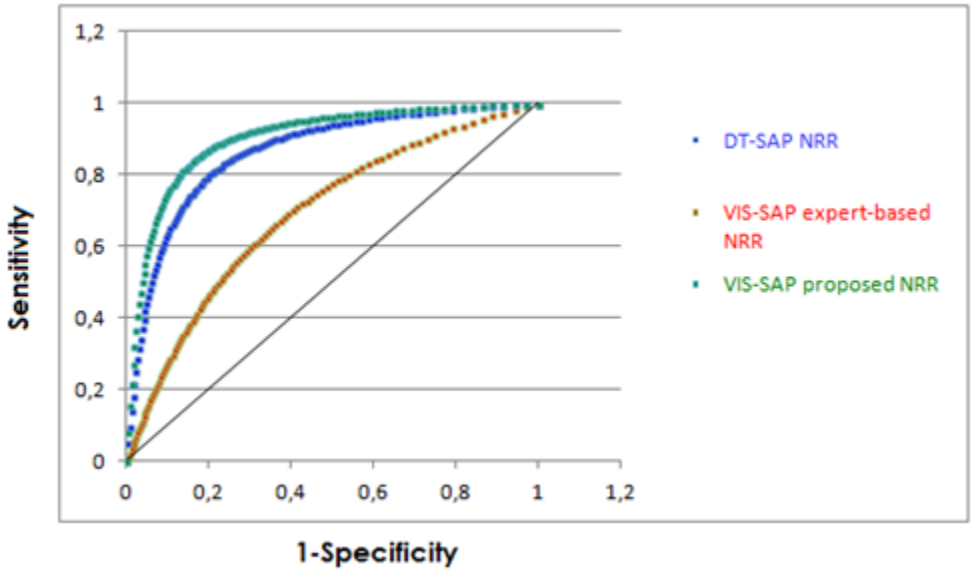


Figure 9.21. ROC curves comparison for VIS-SAP current hypothesis, VIS-SAP proposed NRR and DT - SAP

9.6.2.4. Clinical Validation

Following the NRR identification phase, 327 patients not included so far in this study were considered for participation in order to validate the results. The clinical staff of the hospital randomly selected 10 of them - after participants provided consent in the usual way for these interventions - to test the validity of the clinical hypothesis about NRR of task 151 arising from Section 9.6.2.2. Patients were evaluated before treatment according to the standard clinical protocol (NAB). The neuropsychologists in charge of the NR program of each patient included in the program the execution of task 151 a minimum of 60 times in such a difficulty configuration as to guarantee that the patient obtained a result higher than 20. In this validation phase, the tasks were manually configured for each patient by the specialist, according to the performance shown in previous executions and the specific clinical condition of each participating patient.

All patients were evaluated after treatment following the same standard protocol and the improvement of the patient was assessed in the usual way by comparing scores before the treatment with scores at the end of it.

Of the 317 patients following the classical NR program, 189 showed improvement and 128 did not. Meanwhile, we were able to verify that all of the participating patients under the NRR recommendations improved in the targeted cognitive function. A twofold impact was observed: SAP recommendations can support cognitive therapies with new (previously unknown and specific) configurations of tasks and those recommendations show a higher probability of obtaining measurable improvements in the participating patients.

The authors are aware that the sample size is small to guarantee improvement, but it can be claimed a guarantee of increasing the probabilities of improvement of the participating patients. For the standard treatment group, an improvement tax of 59% was found with a 95% CI: [0.536, 0.644]. This means that the improvement tax would rarely be higher than the 64% of patients. The whole set of 10 patients submitted to the NRR recommendations improved. The 95% CI with a 100% improvement tax is [1,1], since the length of the CI is computed as a function of p(1-p). However, even in the hypothetical case of having one patient without improvement in this small size set, this would lead to a 90% of improvement with a 95% CI of [0.714, 1], meaning that the improvement tax rarely drops

to 71.4%. This lower bound is much higher than the upper bound of the general group, thus proving the efficacy of the recommendation. We can also confirm this issue by the standard test of comparing two proportions as shown in Table 9.12.

In NRR / Improvement	YES	NO	Total
Yes	10	189	199
NO	0	128	128
Total	10	317	327
^p(YES)	1.000	0.5962	0.6100

Table 9.12 Contingency table for validation of idTask= 151

9.6.2.5 Visual Identification of NRR FT-SAP

The first application is the FT-SAP for a CR task (idTask =146 targeting the Attention cognitive function) with $\gamma = 0.8$. The 2-color heat map shown in Figure 9.22 is obtained. "Results" are plotted along the x axis ranging from 0 to 100 and "Number of executions" along the y axis, also ranging from 0 to 100. Two neat NRR regions can be visually identified for high values of Result and mid to high values of number of executions. The identified NRR might indicate that other tasks of the same type (e.g. targeting the same function or subfunction) could behave in a similar way as confirmed in section 9.6.1.1 when tasks are grouped by cognitive function.

CR treatment for this task comprises 3329 executions in total where 1950 of them correspond to patients with improvement = YES and 1379 to improvement = NO.

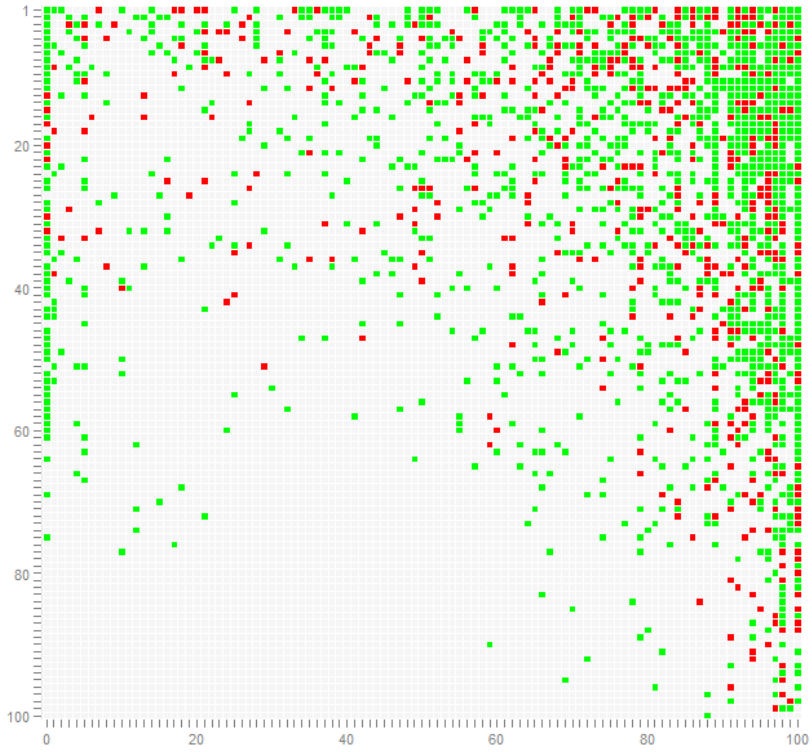


Figure 9.22 FT-SAP(0.8) for idTask= 146 for a total number of 3329 executions

9.6.2.6. Analytical Identification of NRRMR

NRRMR method is applied for the analytical identification of the NRRs. The results obtained are shown in Figure 9.23 below with input parameter values MAXROW=4 and MAXCOLUM=3.

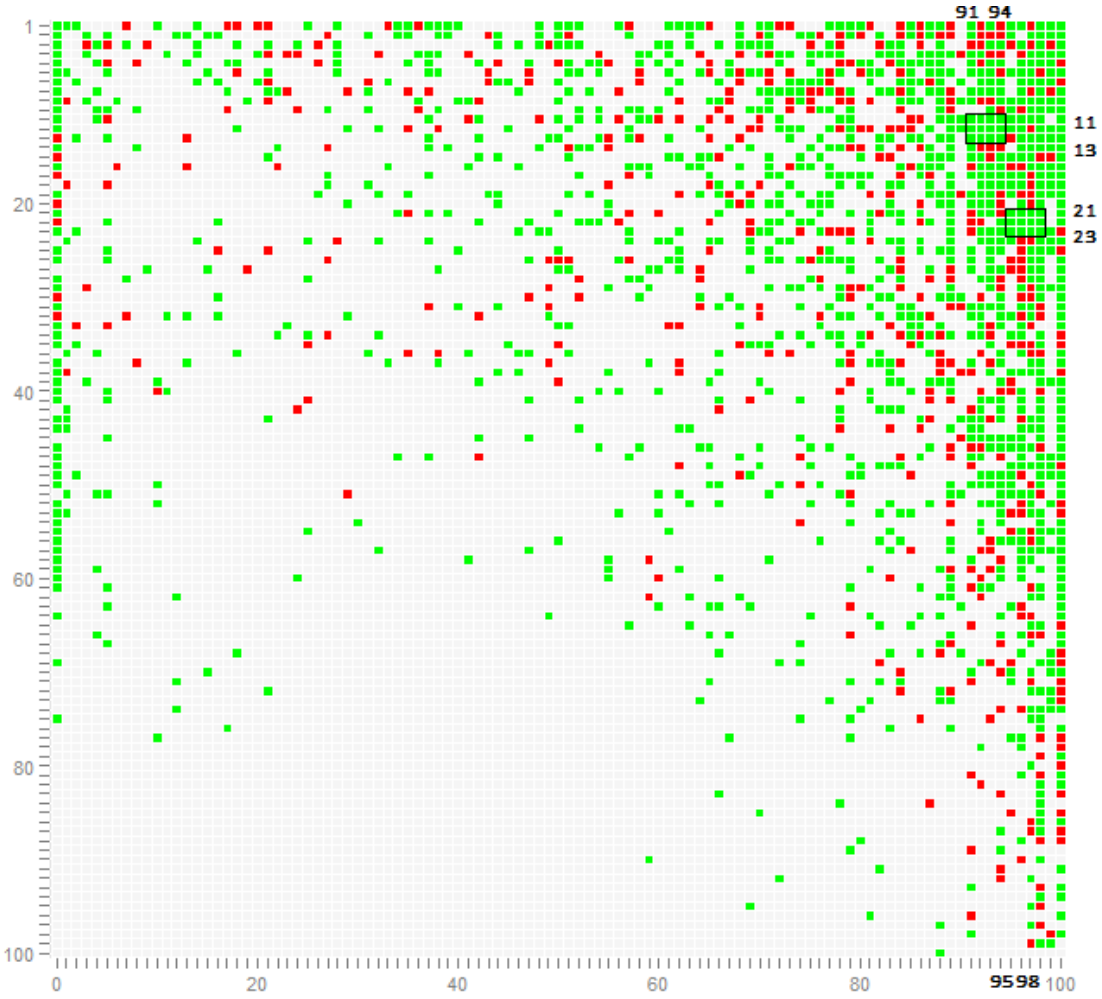


Figure 9.23 NRR coordinates identified by proposed method

The resulting NRRs is the following:

If (Results in [91,94] and Repetitions in [11,13]) then P(Improvement)>=0.8
If (Results in [95,98] and Repetitions in [21,23]) then P(Improvement)>=0.8

9.6.2.7 Vis-SAP and FT-SAP comparison

Figure 9.24 left presents FT-SAP proposed in this work for idTask=151 and γ =1 and Figure 9.24 right shows Vis-SAP obtained in (García-Rudolph & Gibert, 2014). FT-SAP(1) represents a green point at position (i,j) if pij $\geq \gamma$, where pij = 1, i.e. all patients executing i times Task 151 and obtaining score j, improve after treatment.

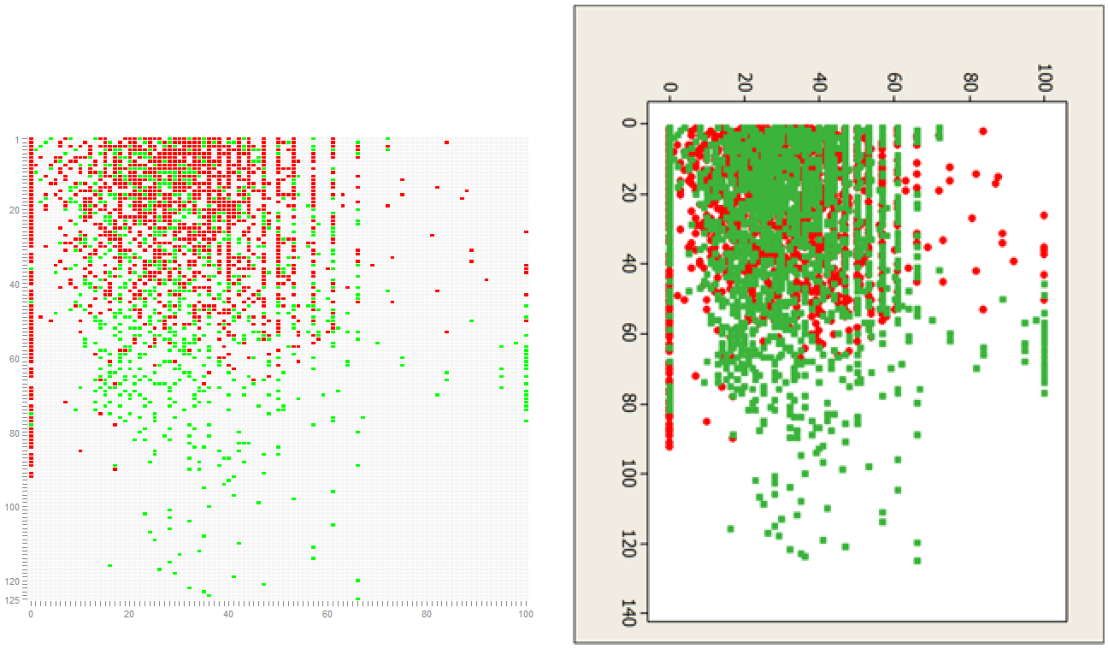


Figure 9.24 FT-SAP (left) and Vis-SAP (right) for idTask=151

In Figure 9.24 left, gray cells do not register observations. As shown in Figure 9.24 (right), no subject with Y=NO who executed the task more than 60 times obtained results other than zero, leading to the identified rule: (NRR(151) = Execs151 > 65 and Res > 20).

However, it can be seen that the area [22,35]×[15,30] appears as a totally green area in the Vis-SAP, whereas there are plenty of red points in the FT-SAP. This indicates that most of the points in this area do not have 100% of patients improving. This is the major contribution of the FT-SAP. One can evaluate the degree of certainty of the induced NRR as the points occlusion occurring in the VIS-SAP is overcome. On the other hand, in the areas of the plot with high concentrations of executions and results (as shown for results lower than 40 and number of executions lower than 60 in the Figure 9.24 right plot) Vis-

SAP does not provide a neat visualization. By construction, FT-SAP avoids the confusion produced by overlapping points. Decreasing γ to 0.5, i.e. admitting half of the patients in a non-improving point after the treatment, produces an FT-SAP(0.5) as shown in Figure 9.25 with many more green points, but it is still difficult to identify an interesting rectangular green region to establish a second area of NRR for Task 151. In conclusion, the FT-SAP provides a refinement of the Vis-SAP that enables uncertainty to be dealt with and, by construction, avoids confusions produced by several patients overlapping in the same point.

Figure 9.25 FT-SAP (0.5) for idTask=151

When longer periods of CR treatments are considered, including therefore an increasing number of subjects, the areas of the plot where no task executions can be found tend to decrease. Also, when a group of tasks targeting the same cognitive function is considered instead of a single one, more robust NRR can be induced from the proposed FT-SAP representation.

In addition, both plots in Figure 9.24 agree on the identification of a zone where NRR is not achieved, as shown by the high values for results and the low number of executions. This seems to suggest that for this type of task, the therapist might expect to achieve NRR for lower results. As shown in section 9.6.1.1, where tasks are analyzed grouped by cognitive function, a different pattern can be identified for groups of tasks. Executive functions (at the bottom of Figure 9.14) seem to be a combination of the attention and memory plots. An explanation for this might be that executive functions are those abilities that allow individuals to efficiently and effectively engage in complex goal-directed behaviors such as planning, sequencing, categorization, flexibility, and inhibition. According to Lezak (Lezak, 1995) this includes the capacity to set goals, to form plans, to initiate actions, and to regulate and evaluate behavior according to the plan and to situational constraints. Therefore executive functions are considered higher level functions which control the more basic cognitive functions such as attention and memory. This implies that preservation of the executive functions might determine whether a brain-damaged individual with lower level cognitive deficits, e.g. selective or divided attention processing or memory deficits, is able to compensate for these deficits and to adapt to the altered situations by restructuring activities (Trettin, 2007).

This suggests that the current NRR considered (the scoring interval [65,85]) might be enlarged to include the number of executions as introduced in (García-Rudolph & Gibert, 2014) and also could be addressed by cognitive function, possibly leading to a different NRR for each function, as shown in section 9.6.1.1.

Methods presented in section 9.6 were tested on a Windows 7 Professional SP 1 PC, Intel Core i3 2.40 GHz (2 GB RAM) 64 bits OS.

The algorithm presented in section 6.3.4 takes a few seconds to run. The MER method described in section 6.4 and 6.4.1. took about 15 minutes to execute. Though inefficient, it provides a good basis upon which to build. To improve its performance, direction to the search needs to be introduced. The proposed algorithm could enumerate the sub-rectangles in any random order and still find the correct solution. Instead, we might take advantage of the fact that if a small rectangle contains a zero, so will each of its surrounding rectangles. Therefore, rectangles will be grown for each possible lower-left corner. This growing

process will only produce upper-right corners defining rectangles which contain only successes (ones, i.e. improvements).

As presented in (García-Rudolph & Gibert, 2014) the main drawbacks of the Vis-SAP proposal are twofold: on the one hand, the lack of completeness of the Vis-SAP criterion proposed. Indeed, looking at the SAP diagram, VIS-SAP is not assigning improvement or non-improvement to the whole area, but only to small parts of the diagram corresponding to concrete and reduced areas where either improvement or non-response can be ensured. Therefore it could be said that Vis-SAP provides a semi-deterministic procedure where a particular configuration for both results and repetitions ensures improvement, a second configuration where the task does not produce patient improvement, and out of these regions the outcome is undetermined. On the other hand, the proposed analysis considers each task individually, being NRR defined for every single task. FT-SAP and the proposed NRRMR methods overcome both these drawbacks.

9.7 Evaluate Improvements on Each Area of Impact

9.7.1 Build F Matrix

As introduced in section 9.4.1 T is the set of all executed tasks, $\mathcal{T}$ = card (T) = 96

T={GlobalLocal, MathMazeComp, MathMazeExer, ConcOps, Submarine, Matching, BagOfCoins, Differences, Figures, PuzzComp, PuzzExer, LetterSoup, Bingo, DiffDirection, StraightLine, SameDirection, GroupWords, CategorizationTwo, CategorizationThree, SameCatWords, Circle, Platforms, Zigurat, GoNoGoEst, GoNoGoGame, GoNoGoPos, Hanging, SinkFleet, Maze, FourInRow, Fourth, JigSaw, BuildSentence, Fragments, Serie, CyclicSerie, SameCat, TempOrder, Position, Sequential, Simoultaneous, WordSeqDec, WordSeqSel, WordSeqDifCat, WordSeqSameCat, WordSimDec, WordSimSel, WordSimDifCat, WordSimSameCat, WordTempOrder, PairsSeqDec, PairsSeqRel, PairsSeqSel, PairsSeqSameOrder, PairsSeqRandOrder, PairsSimDec, PairsSimRel, PairsSimSel, PairsSimSameOrder, PairsSimRandOrder, SentSecOrder, SentSecTest, SentSecWrite, SentSecQuestion, SentSecTrueFalse, SentSimOrder, SentSimTest, SentSimWrite, SentSimQuestion, SentSimTrueFalse, RecSeqNumbers, RecSimNumbers, RemSecNumbers, RemSimNumbers, TextSort, TextQuestion, TextWrite, TextTrueFalse, ImgWordTempOrder, ImgWordSeqDecide, ImgWordSeqRel, ImgWordSeqSel, ImgWordSeqSameOrder, ImgWordSeqRandOrder, ImgWordSimDecide, ImgWordSimRel, ImgWordSimSel, ImgWordSimSameOrder, ImgWordSimRandOrder, DrawTemporalOrder, DrawRecognition, SceneRecognition, SceneRecall, VisualMemory, VisualSimon}

Matrix F is shown below, split in two columns:

F

Task Name	A	M	E
GlobalLocal	1		
MathMazeComp	1		
MathMazeExer	1		
ConcOps	1		
Submarine	1		
Matching	1		
BagOfCoins	1		
Differences	1		
Figures	1		
PuzzComp	1		
PuzzExer	1		
LetterSoup	1		
Bingo	1		
DiffDirection	1		
StraightLine	1		
SameDirection	1		
GroupWords	1		
CategorizationTwo			1
CategorizationThree			1
SameCatWords			1
Circle			1
Platforms,			1
Zigurat			1
GoNoGoEst			1
GoNoGoGame			1
GoNoGoPos			1
Hanging			1
SinkFleet		1	
Maze			1
FourInRow	1		1
Fourth	1		1
JigSaw	1		1
BuildSentence	1		1
Fragments		1	
Serie		1	
CyclicSerie		1	
SameCat		1	
TempOrder		1	
Position		1	
Sequential		1	
Simoultaneous		1	
WordSeqDec,		1	
WordSeqSel,		1	
WordSeqDifCat		1	
WordSeqSameCat		1	
WordSimDec		1	
WordSimSel,		1	
WordSimDifCat		1	

Task Name	A	M	E
WordSimSameCat		1	
WordTempOrder		1	
PairsSeqDec		1	1
PairsSeqRel		1	1
PairsSeqSel		1	1
PairsSeqSameOrder		1	1
PairsSeqRandOrder,		1	1
PairsSimDec		1	1
PairsSimRel		1	
PairsSimSel,		1	
PairsSimSameOrder		1	
PairsSimRandOrder		1	
SentSecOrder		1	
SentSecTest		1	
SentSecWrite		1	
SentSecQuestion,		1	
SentSecTrueFalse		1	
SentSimOrder,		1	
SentSimTest		1	
SentSimWrite		1	
SentSimQuestion		1	
SentSimTrueFalse		1	
RecSeqNumbers		1	
RecSimNumbers		1	
RemSecNumbers		1	
RemSimNumbers,		1	
TextSort		1	
TextQuestion		1	
TextWrite		1	
TextTrueFalse		1	
ImgWordTempOrder,		1	
ImgWordSeqDecide		1	
ImgWordSeqRel		1	
ImgWordSeqSel		1	
ImgWordSeqSameOrd		1	
ImgWordSeqRandOrd		1	
ImgWordSimDecide		1	
ImgWordSeqRandOrd		1	
ImgWordSimRel,		1	
ImgWordSimSel,		1	
ImgWordSimSameOrd		1	
ImgWordSimRandOrd		1	
DrawTemporalOrder		1	
DrawRecognition		1	
SceneRecognition		1	
SceneRecall,		1	
VisualMemory		1	
VisualSimon		1	

9.7.2 Build N Matrix

A snapshot of N is shown below for some representative tasks (T56..T64)

Id	*T1*	*…*	*T56*	*T57*	*T58*	*T59*	*T60*	*T61*	*T62*	*T63*	*T64*	*…*	*T96*
i_1	60		0	60	10	0	22	1		9	4		60
i_2	25		0	25	16	0	39	22		10	50		25
i_3	49		0	49	10	0	36	1		5	5		49
i_4	48		0	48	1	0	12	0		1	8		48
i_5	23		3	23	1	11	15	13	16	16	24		23
i_6	16		4	16	9	7	20	13	8	17	45		16
i_7	45		52	45	77	60	110	75		34	77		45
i_8	14		15	14	23	8	29	17	7	25	14		14
i_9	20		0	20	10	0	6	17		4	10		20
…													
…													
	17		17	17	53	42	58	79	27	70	84		17
	10			10	6		4	5	5	10	20		10
	64		44	64	99	43	103	74	33	92	55		64
	29		24	29	35	15	23	22	33	55	19		29
	63		25	63	48	30	76	33	25	70	90		63
	10		22	10	62	19	36	28	1	58	43		10
	8			8	3		5	2	3	9	1		8
	64		87	64	100	85	77	82	63	125	104		64
	16		9	16	8	4	17	6	4	9	13		16
	23		3	23	10	6	3	3	14	13	13		23
	35		22	35	58	33	70	35	7	53	27		35
i_{122}	11		14	11	35	17	34	26	11	39	17		11
i_{123}	40		20	40	52	22	48	56	34	45	30		40

9.7.3 Build Δ Matrix

id	A	M	F		Id	A	M	F
1	0,0	-1,0	0		62	-2,8	-0,3	-2,4
2	-1,7	-0,7	-0,4		63	-1,8	-1,0	-2,6
3	-0,3	-1,7	0		64	-1,8	-0,3	0,2
4	-2,3	-2,0	-2,2		65	-1,8	-1,0	-2,2
5	-1,0	-0,3	-3,2		66	-2,7	-0,3	-1,6
6	-0,2	-0,3	-1		67	-1,2	0,3	0,2
7	0,0	-1,3	-2		68	-3,0	0,0	-2,2
8	-1,8	-0,7	-1,6		69	0,0	-0,7	-0,2
9	-1,0	-0,7	-1,8		70	-2,2	0,0	-1,4
10	0,0	-1,7	0		71	-0,3	0,0	-2,4
11	-1,0	-3,7	-2,8		72	-0,7	-3,7	-0,8
12	-1,3	-0,7	-1,6		73	-1,8	-3,0	-2,8
13	-0,5	-0,7	-1,4		74	-1,3	-0,3	-1,6
14	-1,2	-0,3	0		75	-2,0	0,0	0

15	-2,0	-0,7	-2,6		76	-1,5	-0,3	-2
16	-3,2	-0,7	-0,8		77	-2,5	-4,0	-1
17	-0,8	-3,0	-1,6		78	-2,5	-1,0	-2,8
18	-0,7	-0,3	0		79	-1,8	-0,3	-1
19	0,0	0,3	-0,8		80	-1,2	-2,0	0,2
20	0,3	-3,3	-0,6		81	-0,8	0,0	-0,2
21	0,0	-0,3	-1		82	-2,3	-1,0	-1,4
22	-1,2	-0,3	-1,4		83	0,0	0,0	-0,6
23	-1,0	-0,7	-1		84	1,3	-0,7	-0,8
24	-0,3	-0,3	-2,4		85	0,8	0,3	1,2
25	-1,0	-0,3	-0,6		86	-0,8	-0,7	1
26	0,0	-0,7	-2,8		87	-1,7	-1,3	-0,4
27	-2,7	-2,3	-0,6		88	0,0	-2,3	-0,8
28	-0,3	-1,7	-0,8		89	-2,3	-3,7	-1
29	0,0	0,0	-1		90	-0,7	-0,3	0
30	-1,5	0,0	0		91	-1,0	0,3	-1,4
31	-2,3	-1,7	-3		92	-0,3	0,0	-2
32	-1,8	-1,0	-0,4		93	-2,0	-3,0	-2,4
33	-0,7	-0,7	0		94	-1,3	-1,0	0
34	0,0	0,0	0		95	-2,5	-0,3	-3,2
35	0,0	0,0	0		96	-0,8	-0,3	-1,6
36	-0,3	-0,3	-0,4		97	0,0	0,0	-0,2
37	0,3	0,0	-0,4		98	-1,2	-0,3	-0,8
38	-2,2	-0,3	-3		99	0,0	-0,3	-1,6
39	-1,3	0,0	-2,4		100	-0,7	0,7	0
40	0,3	-2,0	0		101	-0,7	-0,7	-1,6
41	-1,8	-0,7	-0,2		102	-1,3	-0,7	-2,4
42	-1,7	0,0	-2,6		103	-0,5	0,0	-2
43	-0,3	-1,0	0,8		104	-1,7	-1,0	-0,4
44	-1,2	-4,0	-3,4		105	-1,2	-0,7	-1,6
45	0,0	-0,3	-0,4		106	-0,2	0,0	0,2
46	-1,8	-2,3	-1,4		107	-0,8	-1,0	-1,4
47	-1,0	0,0	0,2		108	-2,0	0,0	-2,4
48	0,0	-1,0	-1		109	1,8	-0,7	-1,8
49	0,7	-0,3	-2,2		110	1,7	0,0	-0,8
50	-2,2	-1,0	-3,2		111	0,0	0,0	-0,2
51	-1,3	-0,3	-3,2		112	-2,5	0,0	-1,4
52	-3,2	-0,7	-2		113	-1,7	-0,3	-2,4
53	-0,2	-0,3	-0,8		114	-2,5	-1,0	-0,8
54	-2,5	-1,0	-1		115	-0,7	0,0	-0,6
55	-1,2	-1,0	-2		116	-1,5	0,0	-1,4
56	-1,2	0,0	-0,6		117	-0,7	0,0	0
57	-1,0	-2,3	-1,6		118	-1,3	-1,3	0
58	-2,5	0,0	-2,8		119	0,0	-0,7	-0,8
59	0,0	-1,0	0		120	-0,5	1,0	-1,2
60	-1,7	0,0	-2		121	0,0	-3,0	-0,6
61	-1,2	-1,7	-1,8		122	0,0	0,0	0
					123	-2,7	0,0	-2,6

9.7.4 Build ϒ Matrix*

ϒ matrix is built from N, Δ, F, and NRR (as shown in Chapter 6). Figure 9.26 shows a heatmap representation of ϒ* matrix, where green color represents negative values (higher levels of improvement) x exe shows the number of executions (from N matrix) and y exe the tasks identifiers. For example for T62 the executions interval is [5,15].

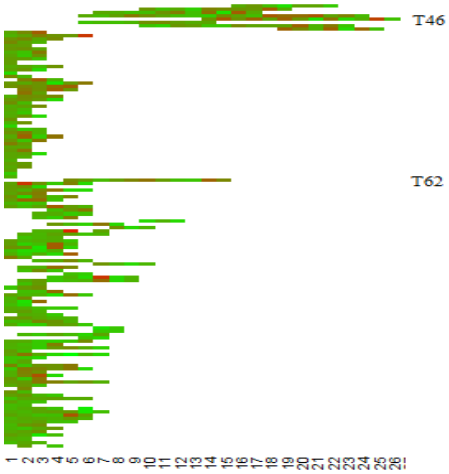


Figure 9.26 Heatmap representation of ϒ* matrix

9.8 Treatment Design

Given a pattern $\mathcal{S}$, matrix F, N, Υ^* and NRR obtained as shown in Chapter 6, a treatment program is built as in the example shown below, for class SHORT70 and l=8 as shown in Figure 9.23.

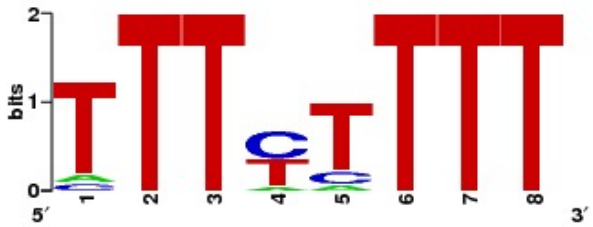


Figure 9.23 Sequence logo for classs SHORT70 and *l*=8

W= ({T},{T},{T},{C,T},{T},{T}{T},{T})
$\mathcal{S}$= $T^3(C|T)T^4$

For the first token $\mathcal{B}$=T, r=3 set of possible tasks impacting only executive functions: {$T_{23,}$ $T_{34,}$ $T_{48,}$ $T_{53,}$ $T_{54.}$ $T_{55,}$ $T_{56,}$ $T_{62,}$ $T_{63,}$ $T_{64.}$}. Column 3 of Υ^* indicates T_{62} (optimal delta value = -1.23).

For the second token $\mathcal{B}$=C|T, r=1. In a similar way F helps to determine the set of possible tasks impacting simultaneously on memory (C) and executive functions (T) which is a set of the following 6 tasks: {$T_{47,}$ $T_{58,}$ $T_{72,}$ $T_{89,}$ $T_{94,}$ T_{135}}. Thus column 1 of Υ^* is optimized. Optimal delta value = -1.25 which corresponds to task T_{135}.

For third token $\mathcal{B}$=T, *r*=4; set of possible tasks impacting only executive functions: {$T_{23,}$ $T_{34,}$ $T_{48,}$ $T_{53,}$ $T_{54.}$ $T_{55,}$ $T_{56,}$ $T_{62,}$ $T_{63,}$ $T_{64.}$}. Column 4 of Υ^* indicates T_{54} (optimal delta value = 1.45).

And following this process the recommendation for the CR program is:

T_{62} T_{62} T_{62} T_{135} T_{54} T_{54} T_{54} T_{54}

9.10 Comparison between Motif Discovery and Classical Supervised Approaches and Sequential Patterns to Find CR General Patterns

To show that the predictive power of the features considered (i.e. CR task executions) is generic and not biased towards a specific classification scheme, we employed traditional classification algorithms that exploit four different machine-learning principles: decision-tree learning (j48), instance-based learning (IBk), probabilistic learning (Naïve Bayes), and RBF neural networks. The prediction performance of the models was measured by ten-fold cross validation and several parameter configurations were tested. Table 9.13 shows the results obtained with by-default parameters but j48 was also tested with 3 different confidence factors (0.25, 0.30, 0.40) decreasing post-pruning and varying the minimum number of objects per leaf. For IBk we used the Euclidean distance (with and without weighting) and different window sizes were tested, also varying the k parameter for the number of neighbors.

Results in Table 9.13 show that for almost 80% of the executions the performance obtained is below 60% and none of them reached a 65% level of accuracy after 10-fold cross validation. These results persist no matter the number of features (CR tasks executions) introduced in the different models. Table 9.13 shows results for models including only initial CR sessions (10 or 20 attributes) up to 600 or 1300 executions. Intermediate values (e.g. 700, 800, 900 number of CR tasks executions) were also tested with similar results in performance. Regarding sequential pattern mining, tested methods (CM-SPAM and CM-PREFIXSPAM besides CM-SPADE results presented in Table 9.16) show acceptable performance regarding support (e.g. 0.8 or 0.9) and execution time and space. But as detailed in Table 9.16 for example when support >= 0.8 CM-SPADE identifies 44690 frequent sequences of length 9, 101685 sequences of length 10 and 2415 of length 11. Frequent sequential pattern algorithms were also tested after performing a ClBR clustering phase, but the results obtained were consistent regardless of sequence lengths, i.e. shorter sequence clusters did not decrease the number of identified frequent sequential patterns and therefore did not lead to a set of frequent patterns that were easier to process.

Pattern discovery with traditional classifiers. Preliminary analysis and problem representation provided appropriate data structures, data transformations, and domain knowledge to undergo patterns discovery. Classical classification techniques (section 3.3.1) are proposed to study response to CR treatment.

Matrix χ is used for the classifier with a response variable Z such as

$$Z = \begin{cases} YES, patient\ improved\ after\ treatment \\ NO, patient\ didn't\ improve\ after\ treatment \end{cases}$$

Z being composed as indicated in Section 2.1.

Waikato Environment for Knowledge Analysis (WEKA) (Hall et al 2009), v 3.6.5 was the data mining platform for running classifiers. All of them were run with default parameters on a 3.4 GHz Pentium IV PC with 2 GB of RAM. The classifiers run in this application were:

- J48 is the WEKA implementation of the C4.5 decision tree (Quinlan 1993).
- NaiveBayes implements the probabilistic Naïve Bayes classifier (John, 1995).
- IBk is the implementation of KNN (Aha 1991) the k-nearest-neighbor classifier IBk has a parameter (k set in our tests to 1,2,3, and 5) that sets the neighborhood size.
- RBFNetworks implements a popular type of feed-forward network, radial basis function (RBF) network (Witten, 2011).

Table 9.15 shows prediction accuracy for each classifier after 10-fold cross validation. In this study the data set was split into 9 subsets with 12 records and 1 subset with 15.

Each classifier is trained 10 times, each time using a version of the data in which one of the subsets is omitted (testing data). Each trained classifier is then tested on the data from the subset which was not used during training. The results are averaged over the 10 classifiers to obtain an overall accuracy shown in Table 9.13.

Attributes	C4.5 (J48)	Naive Bayes	KNN (IBk)				RBFNetworks
			IB1	*IB2*	*IB3*	*IB5*	
10	57.72	54.47	52.84	50.40	53.65	57.62	56.91
20	56.09	52.84	56.09	54.47	58.53	56.91	47.15
30	57.72	58.53	56.91	58.53	64.22	58.53	56.09
40	57.72	56.91	53.65	51.21	56.91	61.78	59.34
50	53.65	58.53	57.72	57.72	62.60	58.53	59.34
60	51.21	57.72	57.72	56.91	57.72	56.09	58.53
70	60.97	57.72	57.72	56.91	60.16	59.34	52.84
80	59.34	59.34	56.09	54.47	62.60	63.41	54.47
100	63.41	55.28	56.09	52.03	60.97	61.78	49.59
600	64.41	61.78	59.34	55.28	60.16	60.16	46.34
1391	60.16	58.53	60.16	56.09	60.16	62.60	46.22

Table 9.13. Accuracy for each classifier after 10-fold cross validation, first column shows the number of CR tasks executions considered, therefore first line (10 attributes) represents an 11-tuple as in *Example 1* above.

Sequential patterns analysis. As presented in section 3.3.5, sequential pattern mining techniques might be applied to find patterns of execution of CR tasks targeting cognitive functions, identified patterns might help to understand responses to treatment. The input is matrix χ.

Sequential Pattern Mining Framework (SPMF) version **v0.96q** was the data mining platform for the Sequential Pattern Mining algorithm executions. (Fournier-Viger, 2014) All of them were run with default parameters on a 3.4 GHz Pentium IV computer with 2 GB of RAM.

SPADE and SPAM are very efficient for datasets with dense or long sequences and have excellent overall performance. This is because unlike algorithms using the horizontal format, performing join operations to calculate the support of candidates does not require scanning of the original database. For example, in a worst-case scenario the well-known PrefixSpan algorithm, which uses the horizontal format, performs a database projection for each item of each frequent sequential pattern, which is extremely costly.

CM-SPADE is the SPMF implementation of SPADE algorithm (Fournier-Viger, 2014). As presented in section 3.3.5, the support of a sequential pattern is the number of sequences where the pattern occurs divided by the total number of sequences in the database.

Table 9.14 shows for support >= 0.88 identified frequent patterns. The first column shows the length of the patterns, the second column the number N of identified sequences for each

pattern length, and the support Median, mean, Standard deviation, minimum and maximum values as well as Q1 and Q3 statistics. Therefore 31501 patterns of length 15 are found.

CM-SPADE

	LEN	N	Mean	St Dev	Min	Max	Q1	Median	Q3
0.88	1	3	0.9675	0.0422	0.9187	0.9919	0.9187	0.9919	0.9919
	2	9	0.9431	0.0407	0.9024	0.9919	0.9106	0.9106	0.9837
	3	27	0.92442	0.03785	0.88618	0.99187	0.89431	0.90244	0.97561
	4	64	0.91527	0.03536	0.88618	0.99187	0.89431	0.90244	0.95325
	5	118	0.91264	0.03440	0.88618	0.99187	0.88618	0.89431	0.95325
	6	172	0.91591	0.03373	0.88618	0.98374	0.88618	0.89431	0.95122
	7	238	0.92150	0.03073	0.88618	0.96748	0.88618	0.93496	0.95122
	8	354	0.92609	0.02451	0.88618	0.96748	0.89431	0.93496	0.94309
	9	589	0.92715	0.01716	0.88618	0.95935	0.92683	0.93496	0.93496
	10	1064	0.92408	0.01164	0.88618	0.95122	0.91870	0.92683	0.93496
	11	2060	0.91756	0.00976	0.88618	0.94309	0.91057	0.91870	0.92683
	12	4097	0.91035	0.00907	0.88618	0.93496	0.90244	0.91057	0.91870
	13	8192	0.90385	0.00834	0.88618	0.93496	0.89431	0.90244	0.91057
	14	16339	0.89834	0.00760	0.88618	0.92683	0.89431	0.89431	0.90244
	15	31501	0.89391	0.00665	0.88618	0.91870	0.88618	0.89431	0.89431
	16	53247	0.89084	0.00549	0.88618	0.91870	0.88618	0.88618	0.89431
	17	69573	0.88902	0.00438	0.88618	0.91057	0.88618	0.88618	0.89431
	18	64130	0.88794	0.00348	0.88618	0.91057	0.88618	0.88618	0.88618
	19	38554	0.88730	0.00281	0.88618	0.90244	0.88618	0.88618	0.88618
	20	14341	0.88689	0.00230	0.88618	0.89431	0.88618	0.88618	0.88618
	21	3159	0.88651	0.00161	0.88618	0.89431	0.88618	0.88618	0.88618
	22	401	0.88618	0.000000	0.88618	0.88618	0.88618	0.88618	0.88618
	23	32	0.88618	0.000000	0.88618	0.88618	0.88618	0.88618	0.88618
	24	1	0.88618	*	0.88618	0.88618	*	0.88618	*

Table 9.14 Sequential patterns identified by CM-SPADE for a support of 0.88. First column shows the different lengths of the discovered patterns and N column the number of patterns, Media column shows the Mean support value.

Sequential pattern mining on each class. SM-SPADE is applied on each of the identified classes but as shown in the Table 9.15 below, the problems identified in section 3.3.5 are not overcome: e.g. number (N) of identified patterns in each class.

CLASS SHORT70

	LEN	N	Mean	St Dev	Min	Max	Q1	Median	Q3
0.5	1	8	0.7562	0.1223	0.5500	0.8750	0.6312	0.8000	0.8500
	2	52	0.6418	0.0977	0.5000	0.8250	0.5500	0.6500	0.7250
	3	220	0.58136	0.06786	0.50000	0.77500	0.52500	0.57500	0.62500
	4	540	0.54819	0.04859	0.50000	0.72500	0.50000	0.52500	0.57500
	5	658	0.53533	0.03937	0.50000	0.67500	0.50000	0.52500	0.55000
	6	481	0.52651	0.03073	0.50000	0.62500	0.50000	0.52500	0.55000
	7	214	0.51636	0.02314	0.50000	0.60000	0.50000	0.50000	0.52500
	8	42	0.51250	0.02084	0.50000	0.57500	0.50000	0.50000	0.52500
	9	2	0.50000	0.000000	0.50000	0.50000	*	0.50000	*
0.7	1	6	0.8167	0.0563	0.7250	0.8750	0.7625	0.8375	0.8562
	2	19	0.75000	0.04330	0.70000	0.82500	0.70000	0.75000	0.77500
	3	17	0.72794	0.02319	0.70000	0.77500	0.70000	0.72500	0.75000
	4	6	0.70833	0.01291	0.70000	0.72500	0.70000	0.70000	0.72500
0.8	1	4	0.8500	0.0204	0.8250	0.8750	0.8312	0.8500	0.8688
	2	4	0.81250	0.01443	0.80000	0.82500	0.80000	0.81250	0.82500

CLASS SHORT86

	LEN	N	Mean	St Dev	Min	Max	Q1	Median	Q3
0.8	1	16	0.8200	0.1455	0.5800	0.9800	0.6800	0.8400	0.9750
	2	87	0.8480	0.1120	0.5600	0.9800	0.8200	0.8400	0.9600
	3	444	0.86032	0.06557	0.56000	0.9800	0.80000	0.86000	0.92000
	4	1827	0.85606	0.05152	0.56000	0.98000	0.82000	0.84000	0.90000
	5	6694	0.84476	0.04158	0.80000	0.98000	0.80000	0.84000	0.88000
	6	19163	0.83284	0.03304	0.80000	0.96000	0.80000	0.82000	0.86000
	7	38639	0.82293	0.02527	0.80000	0.94000	0.80000	0.82000	0.84000
	8	53869	0.81440	0.01874	0.80000	0.92000	0.80000	0.80000	0.82000
	9	44690	0.80805	0.01362	0.80000	0.90000	0.80000	0.80000	0.82000
	10	16853	0.80442	0.00991	0.80000	0.88000	0.80000	0.80000	0.80000
	11	2415	0.80206	0.00678	0.80000	0.86000	0.80000	0.80000	0.80000
	12	83	0.80289	0.00834	0.80000	0.84000	0.80000	0.80000	0.80000
	13	5	0.80000	0.00000	0.80000	0.80000	0.80000	0.80000	0.80000

CLASS LONG6

	LEN	N	Mean	St Dev	Min	Max	Q1	Median	Q3
0.8	1	9	0.9583	0.0625	0.8750	1.0000	0.8750	1.0000	1.0000
	2	60	0.94375	0.06271	0.87500	1.00000	0.87500	1.00000	1.00000
	3	320	0.92852	0.06195	0.87500	1.00000	0.87500	0.87500	1.00000
	4	1355	0.91504	0.05835	0.87500	1.00000	0.87500	0.87500	1.00000
	5	4489	0.90541	0.05364	0.87500	1.00000	0.87500	0.87500	0.87500
	6	11659	0.89830	0.04868	0.87500	1.00000	0.87500	0.87500	0.87500
	7	23216	0.89364	0.04453	0.87500	1.00000	0.87500	0.87500	0.87500
	8	35347	0.88998	0.04059	0.87500	1.00000	0.87500	0.87500	0.87500
	9	39390	0.88649	0.03611	0.87500	1.00000	0.87500	0.87500	0.87500
	10	29588	0.88307	0.03072	0.87500	1.00000	0.87500	0.87500	0.87500
	11	13003	0.88003	0.02456	0.87500	1.00000	0.87500	0.87500	0.87500
	12	2860	0.87732	0.01686	0.87500	1.00000	0.87500	0.87500	0.87500
	13	250	0.87600	0.01116	0.87500	1.00000	0.87500	0.87500	0.87500
	14	7	0.87500	0.000000	0.87500	0.87500	0.87500	0.87500	0.87500

Table 9.15. Identified sequential patterns on each class

9.11 Summary

This chapter presents the application of the proposed methods in a real clinical context: the Neuropsychology Department of the Acquired Brain Injury Unit at Institut Guttmann Neurorehabilitation Hospital (IG) where TBI patients undergo CR treatments . One hundred and twenty-three TBI adults following a 3-5 months CR treatment at IG Neuropsychological Rehabilitation Unit are analyzed in this study. A total of 39412 task executions have been included in this analysis, involving the 96 different CR tasks included in the PREVIRNEC© platform.

The CMIS methodology is presented as a high level umbrella of 5 steps applied in this Chapter.The CMIS relies on two main contributions: SAIMAP (section 4.1) and NRR (Chapter 6). The results of identifying NRR by means of Vis-SAP and DT-SAP are presented. A clinical validation of the obtained results was performed on 327 patients not included in the study, this chapter shows that the identified sectors have a clear positive effect on patient recovery and also the ones where no recovery effect is shown can be considered as useful clinical hypothesis. The NRRMR method is applied to any number of tasks allowing for the identification of new NRRs. When grouped by cognitive function three clear different patterns are identified. Afterwards SAIMAP methodology was applied to the same dataset, , a previous clustering process is performed in such a way that three program profiles are identified. Later, local motif discovery to each profile is performed to understand the structure of the tasks sequences associated with the classes.

Statistical tests seem to indicate that basic demographic and clinical characteristics of the patients (GCS, PTA, gender, educational level, age) do not show significant differences along the classes thus indicating that differences among groups may be due to the structure of the treatment itself. Afterwards, improvements of the patients for the different classes have been studied by means of conditional distributions of improvement indicators (effect indexes) versus the classes and this seems to confirm that different responses to the treatments are associated to the classes.

List of Contributions

The following are the main contributions of this work from both clinical and technical points of view:

- In order to identify sequences of activities such that the global effect of the sequence over a set of impact areas leads to successful performance, the general CMIS (Cumulative Multiple Impact Sequences) methodology is introduced as a sequence of steps.
- CMIS relies on two main contributions: SAIMAP and NRR both introduced in this research for the first time.
- SAIMAP (Sequence of Activities Improving Multi-Area Performance) is an innovative combination of pre-processing tools, clustering, motif discovery and post-processing techniques in a hybrid methodological frame, where sequential patterns of a predefined set of events with high order interactions and cumulative effects are associated with multi-criteria improvement in a predefined set of areas.
- Definition of the NRR (Neurorehabilitation Range) concept to determine the degree of performance expected for a CR task and the number of repetitions required to produce maximum rehabilitation effects.
- Operationalization of NRR by means of SAP
- Introduce, define, implement, apply, and validate to a real case the Sectorized and Annotated Plane (SAP): visualization tool to identify areas with high probability of occurrence of a target event.
- Vis-SAP method: data mining methodologies for building the SAP
- DT-SAP method: decision-tree based method to automatically build SAP
- Introduction of a quality measure for SAP based on pooled confidence of all labeled sectors.
- For the SAP of binary variables a pooled specificity and a pooled sensitivity are used as quality indicators.
- Application of SAP to identify the NRR of a given neurorehabilitation cognitive task.

- The definition of a quality criterion to assess NRR models, based on quality indicators introduced for the SAP. The defined criterion quantifies the probability of improvement with the execution of a task under certain conditions.
- FT-SAP has been introduced as a parametric heatmap-based visualization tool to overcome the limitation of Vis-SAP method produced by occlusions.
- Introduction of the NRRMR (Neurorehabilitation Range Maximal Regions): Generalization of the Maximal Empty Rectangle problem (MER) to identify maximal NRR over a FT-SAP.
- Innovative combination of Clustering Based on Rules and Motif Discovery in SAIMAP methodology
- Results are compared to state-of-the-art sequential pattern mining techniques and to traditional classification algorithms that exploit different machine learning principles.

Chapter 10. Conclusions and Future Plans

This thesis aims to support the design of sequences of cumulative activities impacting on multiple areas in order to obtain a better response to standard assessments performed on the areas after the execution of activities. Our approach is twofold: theoretical and practical in order to provide recommendations for the selection of the sequence of activities and repetitions that will induce a better response of individuals on the activities' impacted areas. It is a difficult problem because activities demonstrate a high level of interaction between them and cumulative effects, and also because the length and configuration of activity sequences is open.

From the theoretical point of view, this work proposes two contributions to this topic through innovative data mining techniques: SAIMAP (Sequence of Activities Improving Multi-Area Performance) and NRRMR (Neurorehabilitation Range Maximal Regions) methods, integrated in a general CMIS methodology. SAP has been introduced as a general visualization tool to identify areas with a high probability of occurrence of a target event. Three approaches to SAP are defined, implemented, applied, and validated to a real case: Vis-SAP, DT-SAP and FT-SAP the parametric heatmap-based visualization proposed to overcome the limitations detected in Vis-SAP.

From a practical point of view this work introduces a new concept, the NeuroRehabilitation Range (NRR) as the framework to describe the degree of performance of a CR task which produces maximum rehabilitation effects. The NRR helps provide an operational definition for the zone of maximum rehabilitation potential and represents an operationalization of the Zone of Proximal Development (ZPD). Analytical and visual tools are also proposed, defined and validated in this work, in order to find an operational definition of an NRR from a data-driven approach. For this particular application, the SAP identifies areas with a high probability of cognitive improvement. Although SAP is not a complex concept, it has shown great potential for finding the NRR region of a cognitive rehabilitation task in quickly, simply and very intuitively. This has proved to be highly useful at clinical practice

level. Also, for the first time, the NRR is defined as a bivariate structure involving conditions in both results and repetitions of the tasks.

A quality criterion to assess NRR models, based on pooled confidence and pooled specificity is also introduced in this work. The defined criterion is based on the capacity of an NRR model to detect the patients improving with the execution of a task. This provides some form of global performance indicator, although ROC curves have also been used to test the quality of obtained models. It confirms that both proposed methods outperform the univariate and static NRR [65,85] currently used by the experts, and also that Vis-SAP performs slightly better than DT-SAP.

Clinicians established an initial hypothesis about the NRR, assuming it to be fixed and task-independent (NRR(T)= [65, 85]); these bounds have been defined according to CR therapists' expertise. PREVIRNEC© allows systematic pre- and post-evaluation of participants covering the major cognitive domains. This provides empirical data useful to validate or clarify clinical hypothesis. For the first time, data collected through the PREVIRNEC© platform has been used to learn more about the NRR. Although the ratio of improvement of patients in that initial NRR was not low, this work provided evidence that a formulation for NRR regarding only the Results obtained is insufficient to identify the group of patients with better response to CR treatment. According to our results NRR cannot be defined by means of univariant analysis (considering only the Result of performing a task). A predictive model considering other implied co-variables needs to be developed. This thesis is a first attempt in that direction. It has been shown that the number of repetitions that a patient performs of a certain task is also relevant for the patient's outcome, according to literature. Bidimensional NRR, depending not only on performances, but also on repetition, significantly improves the CR treatment design. On the other hand, the range of therapeutic performances might change from task to task. This work proposes to target a specific performance-range for each task (or cognitive function) instead of the current [65,85] range used for the whole set of cognitive tasks available.

This work provides objective criteria for NRR that can be integrated into the daily clinical practice of the institution, as well as operationalized for the PREVIRNEC© platform, and which provides the support required to verify clinical hypothesis.

Furthermore, SAP and MER (Maximal Empty Rectangle) solutions were adapted and applied to automatically generate data-driven models in order to identify bi-dimensional NRRs, taking into account the proper combinations of repetition of tasks and performance. A method is introduced to identify a variable number of NRRs satisfying a certain degree of reliability (γ) for a given task. A direct MER algorithm is implemented and modified to identify a region's minimum (γ) probability of improvement, in order to solve the NRRMR. Proposed methods are also applied to any number of CR tasks grouped in cognitive functions. This allows for the identification of NRR, not only for a single task but for a group of them stimulating the same cognitive function.

When grouped by cognitive functions, a different response pattern has been identified for memory skills, attention or executive functions, suggesting that NRR might also depend on the targeted function. Further analyses, including subfunctions of each cognitive function, are currently underway.

Until now, CR plans have been mainly built from scratch for every patient, on the basis of the expertise of the therapist and the follow-up of the patient, because no standard guidelines were available in this domain. The findings from the present study have led to new actionable knowledge in the field of rehabilitation practice, opening the door towards more precise, predictable, and powerful CR treatments that are customized for the individual patient. Clinical hypotheses are being formulated by specialists on the basis of these results and are currently under validation, as a previous step to the establishment of a methodology for personalized therapeutic interventions based on clinical evidence.

As future research lines, the automatic construction of SAP still requires more work since decision trees imply, by construction, some intrinsic error taxes in every branch that will always be propagated to the NRR performance and automation from Vis-SAP has to be started from scratch.

This work is currently being enriched by analyzing how patients walk through the SAP areas (or sectors) during their rehabilitation process. This can be analyzed by connecting the points corresponding to the same patient in the SAP and finding prototypical patterns according to the form of the paths designed on the SAP. Later on, this dynamic analysis can be generalized to find dynamic patterns on the global treatment of the patient involving the

whole sequence of tasks performed during the treatment, and providing information about the possible positive interactions between tasks that empower the improvement capacity.

Although the NRR models that use number of executions and results seem to provide quite a high level of sensitivity and specificity, other factors, such as task difficulty, may be supposed to be highly determinant of cognitive improvement. Extension of the current proposals to include such other factors is currently being explored.

Finally, better interpretation of results obtained results by clinicians can be expected when other demographic and clinical variables are included in the model, e.g. participants' educational level, age, time since injury, obtained results in pre-treatment evaluation.

As presented in Chapter 3, other factors may be supposed to be highly determinant of response to treatment, such as the TBI severity reported by GCS, the time since injury, age, and educational level (Cicerone, 2011). Extension of the current proposals to include such other factors is currently being explored, provided that the formal framework is easily extendable to hypercubes instead of two-way tables as shown in this work.

List of Publications

Journal papers

- Alejandro García-Rudolph and Karina Gibert. Understanding Effects of Cognitive Rehabilitation Under a Knowledge Discovery Approach. Engineering Applications of Artificial Intelligence (Submitted 2015).
  SCI IF(2015): 2.207 Q1 (COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE)

- Alejandro García-Rudolph and Karina Gibert. Data Mining Approach for Visual and Analytical Identification of Neurorehabilitation Ranges in Traumatic Brain Injury Cognitive Rehabilitation. Abstract and Applied Analysis vol. 2015.
  Article ID 823562, 14 pages, 2015.
  doi:10.1155/2015/823562
  SCI IF(2013):1.274 Q1 (MATHEMATICS, APPLIED)

- Alejandro García-Rudolph and Karina Gibert. A data mining approach to identify cognitive NeuroRehabilitation Range in Traumatic Brain Injury patients. Expert Systems with Applications. 41 - 11, pp. 5238 - 5251. 09/2014.
  doi:10.1016/j.eswa.2014.03.001
  SCI: IF(2014): 2.240 Q1 (COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE)

- Alexis Marcano-Cedeño, Paloma Chausa, Alejandro García-Rudolph, César Cáceres, Josep M. Tormos, Enrique J. Gómez: Data mining applied to the cognitive rehabilitation of patients with acquired brain injury. Expert Systems with. Applications. 40(4): 1054-1060 (2013).
  doi:10.1016/j.eswa.2012.08.034
  SCI IF(2013): 1.965 Q1 (COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE)

- Alexis Marcano-Cedeño, Paloma Chausa, Alejandro García-Rudolph, César Cáceres, Josep M. Tormos, Enrique J. Gómez: Artificial metaplasticity prediction model for cognitive rehabilitation outcome in acquired brain injury patients. Artificial Intelligence in Medicine 58(2): 91-99 (2013)
  doi:10.1016/j.artmed.2013.03.005
  SCI IF(2013): 2.019 Q2 (COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE)

- Karina Gibert; Alejandro García-Rudolph; Alberto García-Molina; Teresa Roig-Rovira; Montse Bernabeu; José María Tormos. Response to traumatic brain injury neurorehabilitation through an artificial intelligence and statistics hybrid knowledge

discovery from databases methodology. Medicinski Arhiv. 62 (3), pp. 132 - 135. 2008. ISSN0350-199X PMID: 18822937 [PubMed - indexed for MEDLINE]

- Karina Gibert, Alejandro García-Rudolph, Gustavo Rodríguez-Silva. The Role of KDD Support-Interpretation Tools in the Conceptualization of Medical Profiles: An Application to Neurorehabilitation. Acta Inform Med. 2008; 16(4): 178-182, ISSN 0353-8109

Chapters in Books

- Karina Gibert and Alejandro García-Rudolph. Posibilidades de aplicación de minería de datos para el descubrimiento de conocimiento a partir de la práctica clínica. Tecnologías Aplicadas al Proceso Neurorrehabilitador. Capítulo 6 pp 57-63 Desarrollo de herramientas para evaluar el resultado de las tecnologías aplicadas al proceso rehabilitador. Estudio a partir de dos modelos concretos: Lesión Medular y Daño Cerebral Adquirido. Madrid: Plan de Calidad para el Sistema Nacional de Salud. Ministerio de Sanidad y Consumo. Agència d'Avaluació de Tecnologia i Recerca Mèdiques de Cataluña, 2007. Informes de Evaluación de Tecnologías Sanitarias, AATRM .Edita: Agència d'Avaluació de Tecnologia i Recerca Mèdiques de Cataluña ISBN: 978-84-393-7890-7 2008

Chapters in Collections

- Alejandro García-Rudolph, Karina Gibert Finding Patterns in Cognitive Rehabilitation. Frontiers in Artificial Intelligence and Applications Vol 256 pp 193 – 202 Artificial Intelligence Research and Development 2013 IOSPress doi 10.3233/978-1-61499-320-9-193

- Karina Gibert; Alejandro García-Rudolph; Lluïsa Curcoll; Dolors Soler; Laura Pla; José María Tormos. Knowledge discovery about quality of life changes of spinal cord injury patients: clustering based on rules by states. Studies in Health Technology and Informatics. 150, pp. 579 - 583. 2009. ISSN 0926-9630

- Karina Gibert, Alejandro García Rudolph, Alberto García-Molina, Teresa Roig-Rovira, Montserrat Bernabeu, Josep Maria Tormos. Knowledge Discovery on the Response to Neurorehabilitation Treatment of Patients with Traumatic Brain Injury

through an AI&Stats and Graphical Hybrid Methodology. Artificial Intelligence Research and Development. Frontiers in Artificial Intelligence and Applications. Vol 184 pp 170-177 October 2008. IOS Press, ISSN 0922-6389.

Conference papers

- Joan Serra, Josep Lluis Arcos, Alejandro García-Rudolph**,** Alberto García-Molina, Teresa Roig, Josep Maria Tormos. Cognitive prognosis of acquired brain injury patients using machine learning techniques. Int. Conf. on Advanced Cognitive Technologies and Applications (COGNITIVE), IARIA, Valencia, Spain, p.108-113 (2013)

QVidLab and other related projects Publications

- Laia Subirats, Raquel Lopez-Blazquez, Luigi Ceccaroni, Mariona Gifre, Felip Miralles, **Alejandro García-Rudolph**, Josep Maria Tormos. Monitoring and Prognosis System Based on the ICF for People with Traumatic Brain Injury. Internatinal. Journal of Environmental Research. Public Health 2015, 12, 9832-9847.

- Pedro Antonio Moreno ; Paloma Chausa, **Alejandro García Rudolph**; Raquel López Blázquez; Jordi Ceballos; Josep Manel Saperas; Patricia Sánchez Gónzalez, José María Tormos; Enrique J.Gómez CareCloud: Plataforma de soporte y asistencia a cuidadores informales de personas en situación de dependencia por una discapacidad de origen neurológico, CASEIB 2014

- Solana Sánchez, Javier; García Molina, **A.; García Rudolph**, A.; Cáceres Taladriz,César; Chausa Fernández, Paloma; Roig Rovira, Teresa; Tormos Muñoz, Josep M. y Gómez Aguilera, Enrique J. Clustering techniques for patients suffering acquired brain injury inneuro personal trainer. En: "International Conference on

Recent Advances in Neurorehabilitation (ICRAN2013)", 07/03/2013 - 08/03/2013, Valencia, Spain.

- Luna Serrano, Marta; Caballero Hernandez, Ruth; González Rivas, Luis Miguel; García Molina, **A.; García Rudolph**, A.; Cáceres Taladriz, César; Sanchez Carrion, R.; Roig Rovira, Teresa; Tormos Muñoz, Josep M. y Gómez Aguilera, Enrique J. (2013). Dysfunctional 3D model based on structural and neuropsychological information.En: "International Conference on Recent Advances in Neurorehabilitation (ICRAN 2013)", 07/03/2013 - 08/03/2013, Valencia, Spain.

- Laia Subirats, Luigi Ceccaroni, Raquel Lopez-Blazquez, Felip Miralles, **Alejandro García-Rudolph**, José María Tormos: Circles of Health: Towards an advanced social network about disabilities of neurological origin. Journal of Biomedical Informatics 46(6): 1006-1029 (2013)

- L Subirats, S Orte, R López, S Torrellas, **A García-Rudolph**, L Ceccaroni Uso de Estándares Bio-Psico-Sociales y Experiencia de Usuario en una Red Social Orientada a Personas con Discapacidad de Origen Neurológico DRT4All: IV Congreso Internacional de Diseño, Redes de Investigación 2013

- Caballero Hernandez, Ruth; Luna Serrano, Marta; Cáceres Taladriz, César; García Molina, A.;**García Rudolph, A**.; López Blázquez, R.; Toig Rovira, T.;Tormos Muñoz, Josep M. y Gómez Aguilera, Enrique J. (2013). Knowledge representation tool for cognitiveprocesses modeling. En: "International Conference on Recent Advances in Neurorehabilitation (ICRAN 2013)",07/03/2013 - 08/03/2013, Valencia, Spain.

- J.M. Martínez-Moreno, P. Sánchez-González, **A. García Rudolph**, S. González, C. Cáceres, R. Sánchez-Carrión, T. Roig, J.M. Tormos, E.J. Gómez A Graphical Tool for Designing Interactive Video Cognitive Rehabilitation Therapies XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, MEDICON 2013

- Solana Sánchez, J.; García Molina, A.; Cáceres Taladriz, César; **García Rudolph, A.;** Tormos Muñoz, J.M.; Gómez Aguilera, E.J.; Neuro Personal Trainer: Plataforma de Telerrehabilitación Cognitiva. XV Congreso Nacional de Informática de la Salud (InforSalud 2012). Madrid, (Marzo, 2012).

- Solana, J.; García Molina, A.; Steblin, A.; Cáceres Taladriz, César; Lorenzo, J.A.; Roig Rovira, T.; **García Rudolph,A**.; Morell Vilaseca, M.; Pérez de la Fuente, C.; Tormos Muñoz, J. M.; and Gómez-Aguilera, E. J.; Módulo de informes para evaluación de terapias y del uso de la plataforma de telerrehabilitación PREVIRNEC.XXIX Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2011). Cáceres, (Noviembre, 2011).

- J. M. Martínez-Moreno, P. Sánchez-González, M. Morell Vilaseca, **A. García Rudolph,** S. González Palmero, A. García Molina, T. Roig Rovira, C. Cáceres Taladriz, J. M. Tormos Muñoz and E. J. Gómez Aguilera, Diseño y Entornos virtuales de vídeo interactivo para neurorrehabilitación cognitiva (2011), in: Actas del XXIX Congreso anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2011)

- Ó. Orúe, R. Caballero Hernández, A. García Molina, Jose Maria Martinez Moreno, **A. García Rudolph**, C. Cáceres, J. M. Tormos Muñoz and E. J. Gómez Aguilera, Entorno colaborativo de edición de Tareas en Neurorrehabilitación cognitiva (2011), in: Actas del XXIX Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2011)

- Solana J, Cáceres C, Gómez EJ, Ferrer-Celma S, Ferre-Bergada M, García-López P, García-Molina A, **García-Rudolph A**, Roig T, Tormos JM PREVIRNEC A new platform for cognitive tele-rehabilitation. 2011. "COGNITIVE 2011, The Third International Conference on Advanced Cognitive Technologies and Applications", ISBN 978-1-61208-155-7. pp. 59-62.

- R. Caballero Hernández, C. Gómez Pérez, C. Cáceres Taladriz, **A. García Rudolph**, J. Vidal Samsó, M. Bernabeu Guitart, J. M. Tormos Muñoz and E. J. Gómez Aguilera,Modelado de Procesos de Neurorrehabilitación, in: XXIX Congreso Anual de la Sociedad Española de Ingeniería Biomédica, 2011

- García-Molina, P. Rodríguez Rajo, R. Sánchez-Carrión, A. Gómez Pulido, A. Ensenyat, **A. García Rudolph**, J. Solana, C. Cáceres, M. Ferre, T. Roig, Clinical program of cognitive tele-rehabilitation for traumatic brain injury eChallenges, 2010; 11/2010

- Laura Lozano, César Cáceres, Almudena Gómez, **Alejandro García Rudolph**, Raquel López, Teresa Roig, José María Tormos Muñoz, Enrique J. Gómez Aguilera

Técnicas de extracción de conocimiento en procesos de rehabilitación cognitiva de pacientes afectados por Daño Cerebral Adquirido (DCA) CASEIB 2010

- Alberto García-Molina, Pablo Rodríguez-Rajo, Rocío Sánchez-Carrión, Almudena Gómez-Pulido, Antonia Ensenyat, **Alejandro García-Rudolph**, Javier Solana,Cesar Cáceres, María Ferre, Teresa Roig, Institut Guttmann, Spain A Clinical Program of Cognitive Tele-rehabilitation for Traumatic Brain Injury Published in: eChallenges e-2010 Conference Proceedings, Paul Cunningham and Miriam Cunningham (Eds), IIMC International Information Management Corporation Ltd 2010, ISBN 978-1-905824-20-5, ISBN: 978-1-905824-20-5

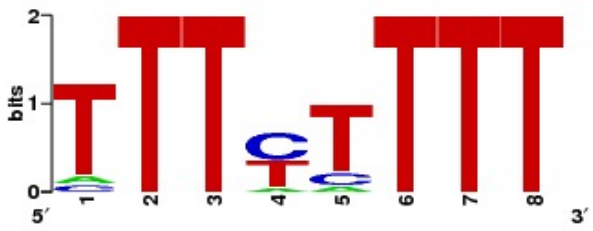
- Tormos, J. M., Garcia-Molina, A., **Garcia Rudolph, A**., & Roig, T. (2009).Information and communications technology in learning development and rehabilitation. International Journal of Integrated Care, 9(Supl), e72. 22 June, ISSN1568-4156.

- Gómez-Pulido, A. García-Molina, R. Sánchez-Carrión, A. Enseñat, **A. García Rudolph**, R. López, D. Tost, P. García, M. Ferré, M. Bernabeu, JM, Tormos, T. Roig-Rovira. Neuropsychological outcome after a computerized cognitive rehabilitation program. 5th Satellite Symposium on Neuropsychological Rehabilitation (Foz do Iguaçu, Brazil). Julio de 2008

Annex

Regular expressions generation (section 5.1)

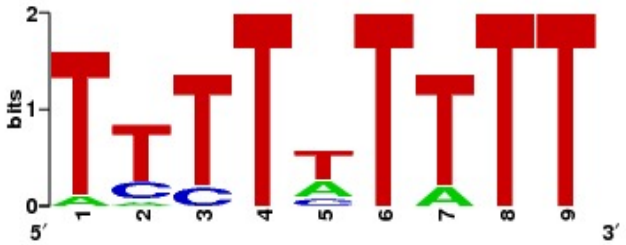
SHORT70 l=8




```
0.076923  0.076923  0.000000  0.846154
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.076923  0.461538  0.000000  0.461538
0.076923  0.153846  0.000000  0.769231
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

W= ({T},{T},{T},{C,T},{T},{T}{T},{T})
$\omega = T^3(C|T)T^4$

SHORT70 l=9

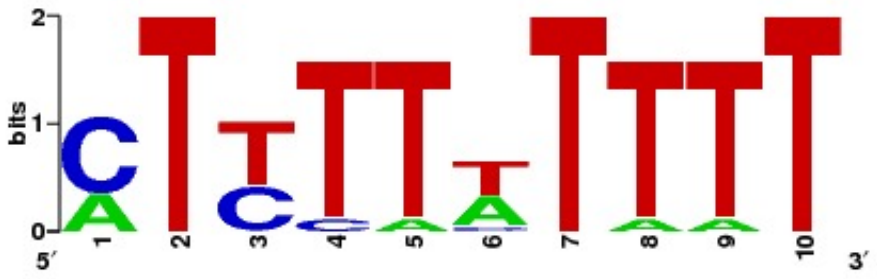



```
0.076923  0.000000  0.000000  0.923077
0.076923  0.230769  0.000000  0.692308
0.000000  0.153846  0.000000  0.846154
0.000000  0.000000  0.000000  1.000000
```

```
0.307692  0.153846  0.000000  0.538462
0.000000  0.000000  0.000000  1.000000
0.153846  0.000000  0.000000  0.846154
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

W= ({T},{C,T},{T},{T},{A,T},{T},{T}{T},{T})
$T(C|T)T^2(A|T)T^4$

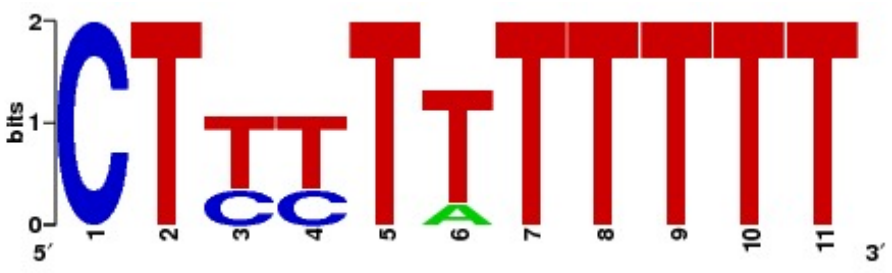
SHORT70 ***l=10***




```
0.333333  0.666667  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.416667  0.000000  0.583333
0.000000  0.083333  0.000000  0.916667
0.083333  0.000000  0.000000  0.916667
0.416667  0.083333  0.000000  0.500000
0.000000  0.000000  0.000000  1.000000
0.083333  0.000000  0.000000  0.916667
0.083333  0.000000  0.000000  0.916667
0.000000  0.000000  0.000000  1.000000
```

W= ({A,C},{T},{C,T},{T},{T},{A,T},{T}{T},{T},{T})
$(A|C)T(C|T)T^2(A|T)T^4$

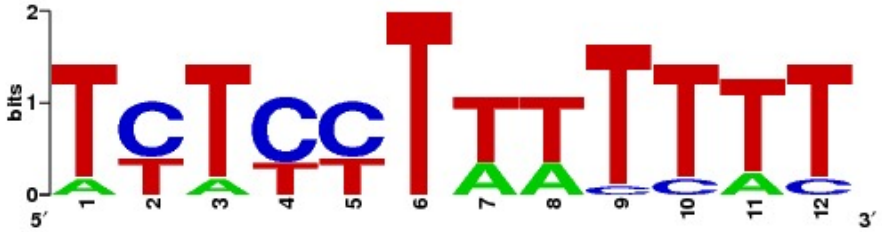
SHORT70 ***l= 11***




```
0.000000  1.000000  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.333333  0.000000  0.666667
0.000000  0.333333  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.166667  0.000000  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

W= ({C},{T},{C,T},{C,T},{T},{T},{T}{T},{T},{T},{T})
$CT(C|T)^2T^7$

SHORT70 ***l=12***

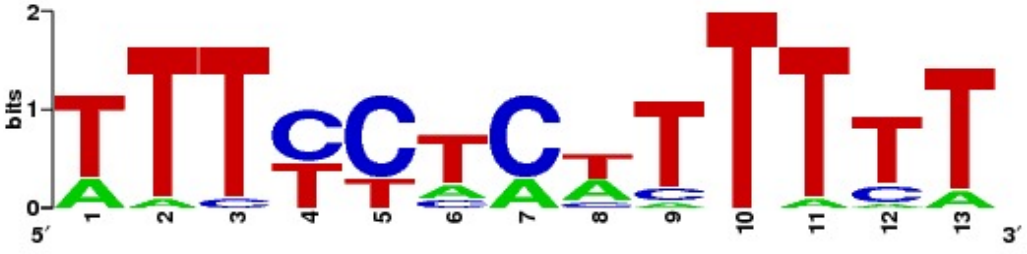



```
0.133333  0.000000  0.000000  0.866667
0.000000  0.600000  0.000000  0.400000
0.133333  0.000000  0.000000  0.866667
0.000000  0.666667  0.000000  0.333333
0.000000  0.600000  0.000000  0.400000
0.000000  0.000000  0.000000  1.000000
0.333333  0.000000  0.000000  0.666667
0.333333  0.000000  0.000000  0.666667
0.000000  0.066667  0.000000  0.933333
0.000000  0.133333  0.000000  0.866667
```

```
0.200000  0.000000  0.000000  0.800000
0.000000  0.133333  0.000000  0.866667
```

W= ({T},{C,T},{T},{C,T},{C,T},{T},{A,T}{A,T},{T},{T},{T},{T})
$T(C|T)T(C|T)^2T(A|T)^2T^4$

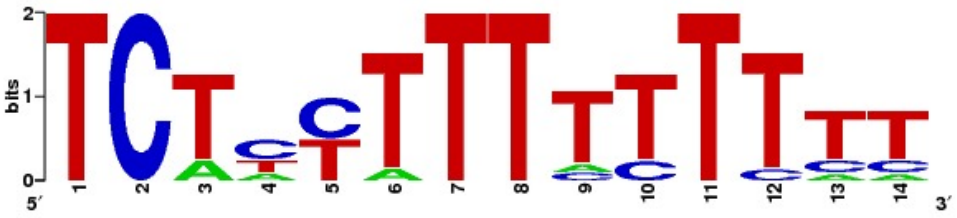
SHORT70 ***l=13***




```
0.266667  0.000000  0.000000  0.733333
0.066667  0.000000  0.000000  0.933333
0.000000  0.066667  0.000000  0.933333
0.000000  0.533333  0.000000  0.466667
0.000000  0.733333  0.000000  0.266667
0.200000  0.133333  0.000000  0.666667
0.266667  0.733333  0.000000  0.000000
0.400000  0.133333  0.000000  0.466667
0.066667  0.133333  0.000000  0.800000
0.000000  0.000000  0.000000  1.000000
0.066667  0.000000  0.000000  0.933333
0.066667  0.200000  0.000000  0.733333
0.133333  0.000000  0.000000  0.866667
```

W= ({A,T},{T},{T},{C,T},{C,T},{T},{A,C}{A,T},{T},{T},{T},{T},{T},{T})
$(A|T)T^2(C|T)^2T(A|C)(A|T)T^6$

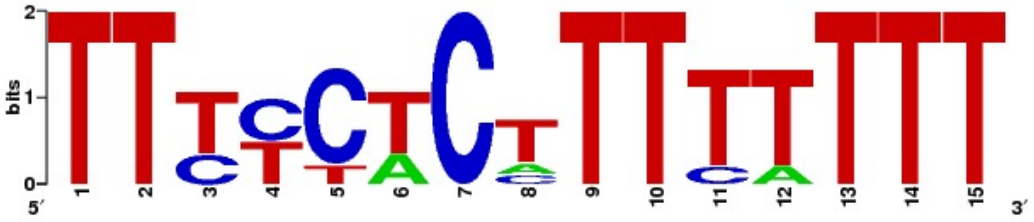
SHORT70 ***l=14***



```
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.200000  0.000000  0.000000  0.800000
0.200000  0.500000  0.000000  0.300000
0.000000  0.500000  0.000000  0.500000
0.100000  0.000000  0.000000  0.900000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.100000  0.100000  0.000000  0.800000
0.000000  0.200000  0.000000  0.800000
0.000000  0.000000  0.000000  1.000000
0.000000  0.100000  0.000000  0.900000
0.100000  0.200000  0.000000  0.700000
0.100000  0.200000  0.000000  0.700000
```

W= ({T},{C},{T},{C,T},{C,T},{T},{T}{T},{T},{T},{T},{T},{T},{C,T})
TCT(C|T)^{2}T^{8}(C|T)

SHORT70 ***l=15***



```
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.333333  0.000000  0.666667
0.000000  0.500000  0.000000  0.500000
0.000000  0.833333  0.000000  0.166667
0.333333  0.000000  0.000000  0.666667
0.000000  1.000000  0.000000  0.000000
0.166667  0.166667  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

```
0.000000  0.166667  0.000000  0.833333
0.166667  0.000000  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

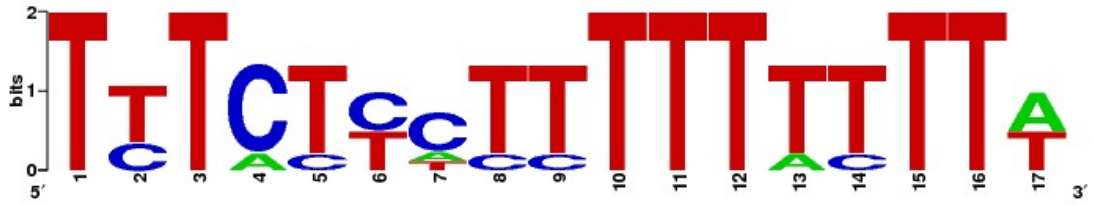
W= ({T},{T},{C,T},{C,T},{C},{A,T},{C}{T},{T},{T},{T},{T},{T},{T},{T})
$T^2(C|T)^2C(A|T)CT^8$

SHORT70 ***l=16***


```
0.000000  0.000000  0.000000  1.000000
0.000000  0.333333  0.000000  0.666667
0.250000  0.000000  0.000000  0.750000
0.250000  0.750000  0.000000  0.000000
0.000000  0.250000  0.000000  0.750000
0.083333  0.416667  0.000000  0.500000
0.583333  0.333333  0.000000  0.083333
0.250000  0.166667  0.000000  0.583333
0.250000  0.166667  0.000000  0.583333
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.083333  0.000000  0.916667
0.333333  0.083333  0.000000  0.583333
0.166667  0.166667  0.000000  0.666667
0.083333  0.000000  0.000000  0.916667
0.000000  0.000000  0.000000  1.000000
```

W=
({T},{C,T},{A,T},{A,C},{C,T},{C,T},{A,C}{A,T},{A,T},{T},{T},{T},{A,T},{T},{T},
{T})
$T(C|T)(A|T)(A|C)(C|T)^2(A|C)(A|T)^2T^3(A|T)T^3$

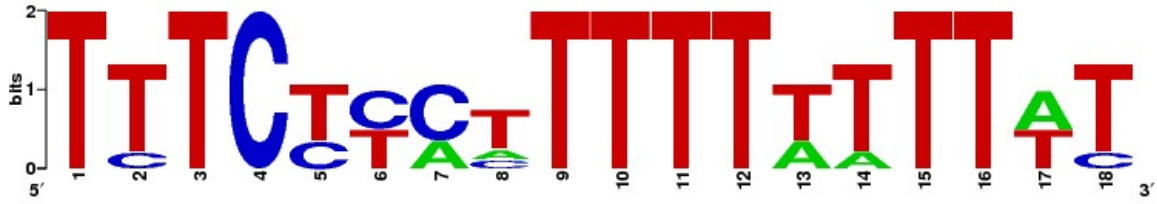
SHORT70 ***l=17***



```
0.000000  0.000000  0.000000  1.000000
0.000000  0.333333  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.166667  0.833333  0.000000  0.000000
0.000000  0.166667  0.000000  0.833333
0.000000  0.500000  0.000000  0.500000
0.166667  0.666667  0.000000  0.166667
0.000000  0.166667  0.000000  0.833333
0.000000  0.166667  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.166667  0.000000  0.000000  0.833333
0.000000  0.166667  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.500000  0.000000  0.000000  0.500000
```

W= ({T},{C,T},{T},{C},{T},{C,T},{C}{T},{T},{T},{T},{T},{T},{T},{T},{T},{A,T})
T(C|T)TCT(C|T)CT^{9}(A|T)

SHORT70 ***l=18***

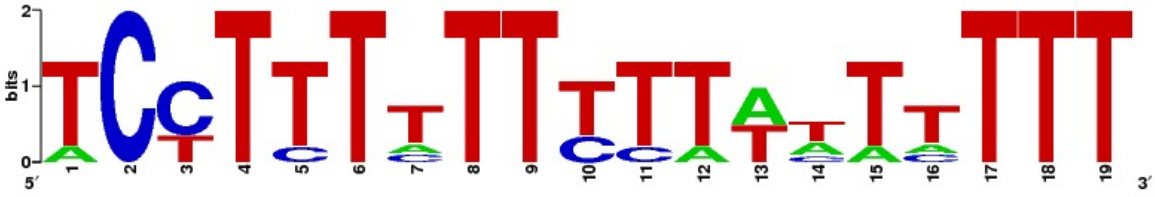


```
0.000000  0.000000  0.000000  1.000000
0.000000  0.166667  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.333333  0.000000  0.666667
0.000000  0.500000  0.000000  0.500000
0.333333  0.666667  0.000000  0.000000
0.166667  0.166667  0.000000  0.666667
```

```
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.333333  0.000000  0.000000  0.666667
0.166667  0.000000  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.500000  0.000000  0.000000  0.500000
0.000000  0.166667  0.000000  0.833333
```

W=
({T},{T},{T},{C},{C,T},{C,T},{A,C}{T},{T},{T},{T},{T},{A,T},{T},{T},{T},{A,T},{T})
$T^3C(C|T)(A|C)T^5(A|T)T^3(A|T)T$

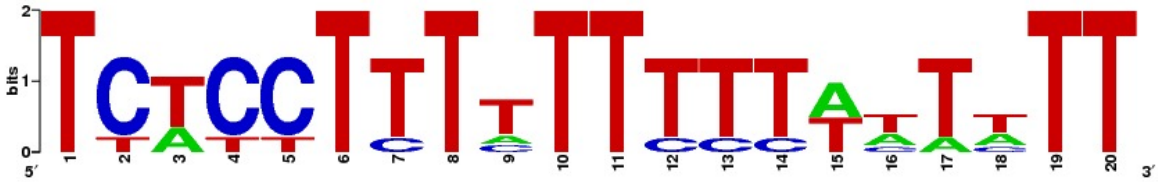
SHORT70 ***l=19***




```
0.166667  0.000000  0.000000  0.833333
0.000000  1.000000  0.000000  0.000000
0.000000  0.666667  0.000000  0.333333
0.000000  0.000000  0.000000  1.000000
0.000000  0.166667  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.166667  0.166667  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.333333  0.000000  0.666667
0.000000  0.166667  0.000000  0.833333
0.166667  0.000000  0.000000  0.833333
0.500000  0.000000  0.000000  0.500000
0.333333  0.166667  0.000000  0.500000
0.166667  0.000000  0.000000  0.833333
0.166667  0.166667  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

W=
({T},{C},{C,T},{T},{T},{T},{T},{T}{T},{T},{C,T},{T},{T},{A,T},{A,T},{T},{T},{T},{T},{T})
$TC(C|T)T^6(C|T)T^2(A|T)^2T^5$

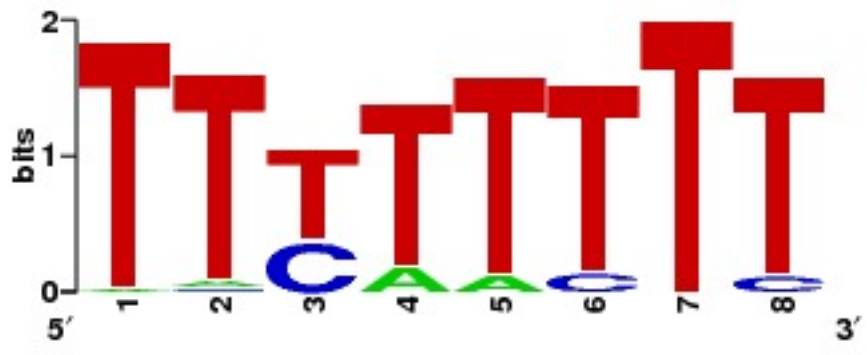
SHORT70 ***l=20***



0.000000	0.000000	0.000000	1.000000
0.000000	0.833333	0.000000	0.166667
0.333333	0.000000	0.000000	0.666667
0.000000	0.833333	0.000000	0.166667
0.000000	0.833333	0.000000	0.166667
0.000000	0.000000	0.000000	1.000000
0.000000	0.166667	0.000000	0.833333
0.000000	0.000000	0.000000	1.000000
0.166667	0.166667	0.000000	0.666667
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.166667	0.000000	0.833333
0.000000	0.166667	0.000000	0.833333
0.000000	0.166667	0.000000	0.833333
0.500000	0.000000	0.000000	0.500000
0.333333	0.166667	0.000000	0.500000
0.166667	0.000000	0.000000	0.833333
0.333333	0.166667	0.000000	0.500000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000

W=
({T},{C},{A,T},{C},{C},{T},{T}{T},{T},{T},{T},{T},{T},{T},{A,T},{A,T},{T},{A,T},{T},{T})
$TC(A|T)C^{2}T\ ^{9}(A|T)^{2}T(A|T)T^{2}$

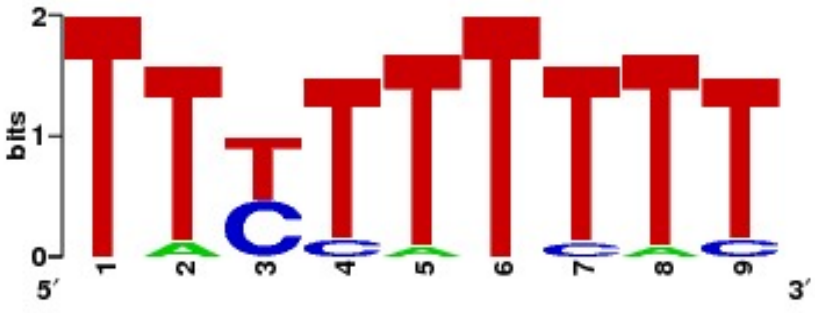
SHORT86 ***l=8***




```
0.020408  0.000000  0.000000  0.979592
0.040816  0.020408  0.000000  0.938776
0.000000  0.367347  0.000000  0.632653
0.142857  0.000000  0.000000  0.857143
0.081633  0.000000  0.000000  0.918367
0.000000  0.102041  0.000000  0.897959
0.000000  0.000000  0.000000  1.000000
0.000000  0.081633  0.000000  0.918367
```

W= ({T},{T},{C,T},{T},{T},{T},{T}{T})
$T^2(C|T)T^5$

SHORT86 ***l=9***

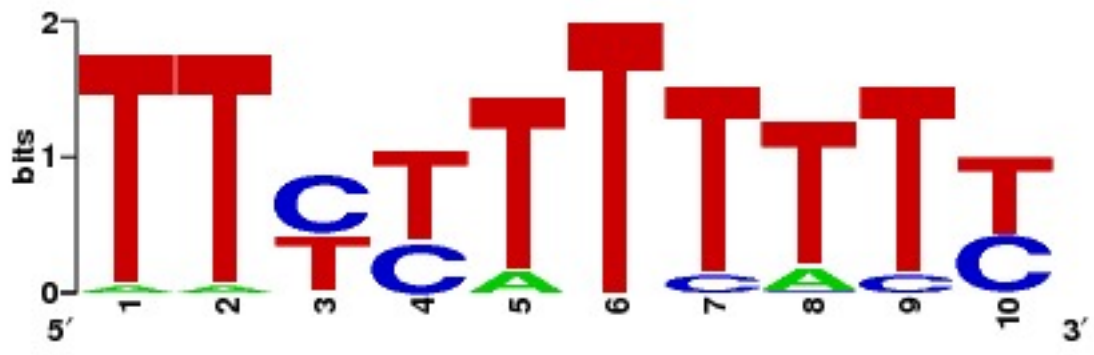



```
0.000000  0.000000  0.000000  1.000000
0.083333  0.000000  0.000000  0.916667
0.000000  0.472222  0.000000  0.527778
0.000000  0.111111  0.000000  0.888889
0.055556  0.000000  0.000000  0.944444
0.000000  0.000000  0.000000  1.000000
```

```
 0.000000  0.083333  0.000000  0.916667
 0.055556  0.000000  0.000000  0.944444
 0.000000  0.111111  0.000000  0.888889
```

W= ({T},{T},{C,T},{T},{T},{T},{T},{T},{T})
$T^2(C|T)T^6$

SHORT86 l=10




```
 0.040816  0.000000  0.000000  0.959184
 0.040816  0.000000  0.000000  0.959184
 0.020408  0.510204  0.000000  0.469388
 0.000000  0.367347  0.000000  0.632653
 0.122449  0.000000  0.000000  0.877551
 0.000000  0.000000  0.000000  1.000000
 0.000000  0.102041  0.000000  0.897959
 0.142857  0.020408  0.000000  0.836735
 0.000000  0.102041  0.000000  0.897959
 0.000000  0.428571  0.000000  0.571429
```

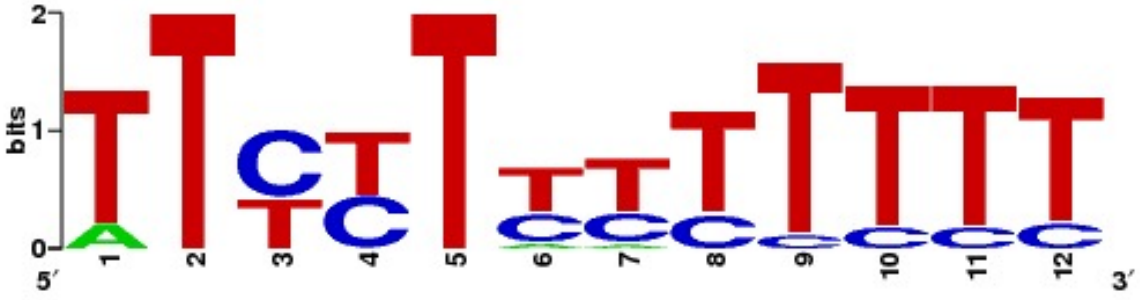
W= ({T},{T},{C,T},{C,T},{T},{T},{T},{T},{T},{C,T})
$T^2(C|T)^2T^5(C|T)$

SHORT86 ***largo11***


```
0.020000  0.000000  0.000000  0.980000
0.000000  0.120000  0.000000  0.880000
0.000000  0.560000  0.000000  0.440000
0.000000  0.260000  0.000000  0.740000
0.140000  0.040000  0.000000  0.820000
0.000000  0.080000  0.000000  0.920000
0.000000  0.220000  0.000000  0.780000
0.100000  0.000000  0.000000  0.900000
0.000000  0.100000  0.000000  0.900000
0.000000  0.340000  0.000000  0.660000
0.000000  0.220000  0.000000  0.780000
```

W= ({T},{T},{C,T},{C,T},{T},{T},{T},{T},{T},{C,T},{C,T})
$T^2(C|T)^2T^5(C|T)^2$

SHORT86 l=12

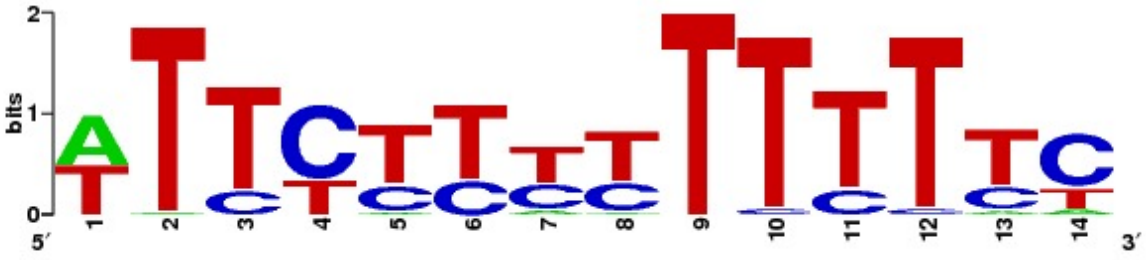



```
0.163265  0.000000  0.000000  0.836735
0.000000  0.000000  0.000000  1.000000
0.000000  0.571429  0.000000  0.428571
0.000000  0.448980  0.000000  0.551020
0.000000  0.000000  0.000000  1.000000
0.081633  0.367347  0.000000  0.551020
0.061224  0.346939  0.000000  0.591837
0.000000  0.265306  0.000000  0.734694
```

```
0.000000  0.081633  0.000000  0.918367
0.000000  0.142857  0.000000  0.857143
0.000000  0.142857  0.000000  0.857143
0.000000  0.183673  0.000000  0.816327
```

W= ({T},{T},{C,T},{C,T},{T},{C,T},{C,T},{C,T},{T},{T},{T},{T})
$T^2(C|T\ ^2T(C|T)^3T^4$

SHORT86 l=14




```
0.500000  0.000000  0.000000  0.500000
0.020000  0.000000  0.000000  0.980000
0.000000  0.200000  0.000000  0.800000
0.000000  0.680000  0.000000  0.320000
0.040000  0.300000  0.000000  0.660000
0.000000  0.320000  0.000000  0.680000
0.080000  0.380000  0.000000  0.540000
0.040000  0.360000  0.000000  0.600000
0.000000  0.000000  0.000000  1.000000
0.000000  0.040000  0.000000  0.960000
0.000000  0.220000  0.000000  0.780000
0.000000  0.040000  0.000000  0.960000
0.060000  0.280000  0.000000  0.660000
0.080000  0.660000  0.000000  0.260000
```

W=
({A,T},{T},{C,T},{C,T},{C,T},{C,T},{C,T},{C,T},{T},{T},{C,T},{T},{C,T},{C,T})

$(A|T)T(C|T)^6T^2(C|T)T(C|T)^2$

SHORT86 l=15


```
0.600000  0.020000  0.000000  0.380000
0.840000  0.000000  0.000000  0.160000
0.300000  0.060000  0.000000  0.640000
0.000000  0.000000  0.000000  1.000000
0.000000  0.380000  0.000000  0.620000
0.000000  0.660000  0.000000  0.340000
0.120000  0.060000  0.000000  0.820000
0.060000  0.340000  0.000000  0.600000
0.020000  0.220000  0.000000  0.760000
0.020000  0.200000  0.000000  0.780000
0.020000  0.000000  0.000000  0.980000
0.000000  0.160000  0.000000  0.840000
0.000000  0.240000  0.000000  0.760000
0.020000  0.260000  0.000000  0.720000
0.060000  0.460000  0.000000  0.480000
```

W=
({A,T},{A},{A,T},{T},{C,T},{C,T},{T},{C,T},{C,T},{C,T},{T},{T},{C,T},{C,T},{C,T}
)
$(A|T)A(A|T)T(C|T)^2T(C|T)^3T^2(C|T)^3$

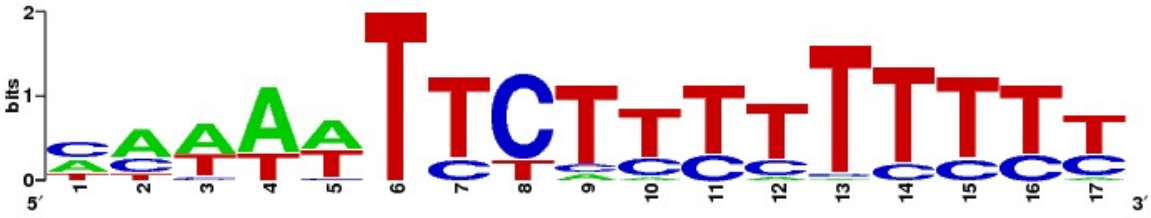
SHORT86 l=16


```
0.560000  0.260000  0.000000  0.180000
0.500000  0.100000  0.000000  0.400000
0.740000  0.000000  0.000000  0.260000
0.440000  0.060000  0.000000  0.500000
```

```
0.000000  0.000000  0.000000  1.000000
0.000000  0.240000  0.000000  0.760000
0.000000  0.800000  0.000000  0.200000
0.080000  0.100000  0.000000  0.820000
0.060000  0.260000  0.000000  0.680000
0.000000  0.280000  0.000000  0.720000
0.040000  0.260000  0.000000  0.700000
0.020000  0.020000  0.000000  0.960000
0.000000  0.140000  0.000000  0.860000
0.000000  0.240000  0.000000  0.760000
0.000000  0.200000  0.000000  0.800000
0.080000  0.340000  0.000000  0.580000
```

W=
({A,C},{A,T},{A,T},{A,T},{T},{C,T},{C,T},{T},{C,T},{C,T},{C,T},{T},{T},{C,T},{C,T},{C,T})
$(A|C)(A|T)^3T(C|T)^2T(C|T)^3T^2(C|T)^3$

SHORT86 l=17

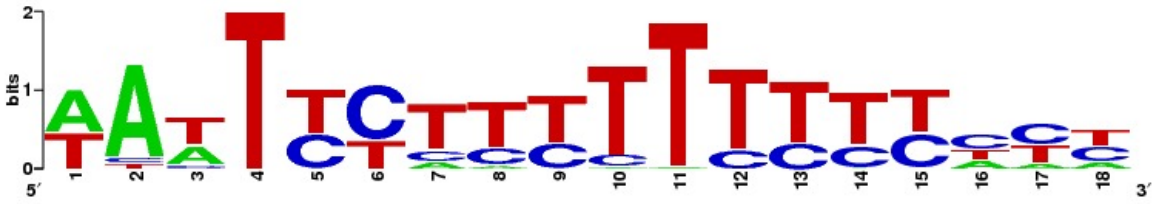



```
0.340000  0.440000  0.000000  0.220000
0.560000  0.300000  0.000000  0.140000
0.540000  0.080000  0.000000  0.380000
0.700000  0.000000  0.000000  0.300000
0.480000  0.060000  0.000000  0.460000
0.000000  0.000000  0.000000  1.000000
0.000000  0.220000  0.000000  0.780000
0.000000  0.800000  0.000000  0.200000
0.080000  0.100000  0.000000  0.820000
0.060000  0.260000  0.000000  0.680000
0.000000  0.280000  0.000000  0.720000
0.060000  0.220000  0.000000  0.720000
0.020000  0.040000  0.000000  0.940000
0.000000  0.160000  0.000000  0.840000
0.000000  0.220000  0.000000  0.780000
```

```
 0.000000  0.280000  0.000000  0.720000
 0.060000  0.340000  0.000000  0.600000
```

W=
({A,C},{A,C},{A,T},{A,T},{A,T},{T},{T},{C},{T},{C,T},{C,T},{T},{T},{T},{T},{C,T},{C,T})
$(A|C)^2(A|T)^3T^2CT(C|T)^2T^4(C|T)^2$

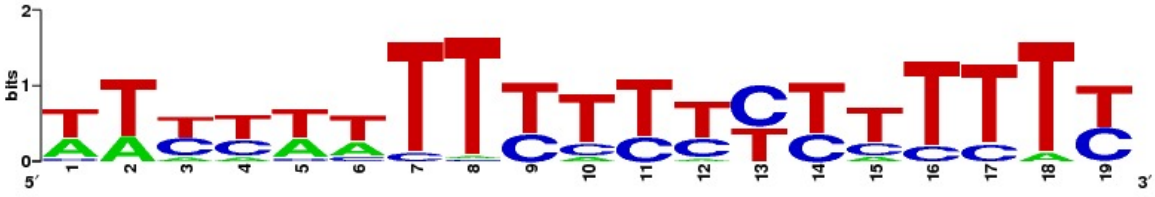
SHORT86 l=18




```
0.540000  0.000000  0.000000  0.460000
0.880000  0.060000  0.000000  0.060000
0.360000  0.100000  0.000000  0.540000
0.000000  0.000000  0.000000  1.000000
0.000000  0.440000  0.000000  0.560000
0.000000  0.660000  0.000000  0.340000
0.120000  0.180000  0.000000  0.700000
0.060000  0.260000  0.000000  0.680000
0.020000  0.340000  0.000000  0.640000
0.020000  0.120000  0.000000  0.860000
0.020000  0.000000  0.000000  0.980000
0.000000  0.200000  0.000000  0.800000
0.000000  0.300000  0.000000  0.700000
0.020000  0.300000  0.000000  0.680000
0.000000  0.460000  0.000000  0.540000
0.240000  0.440000  0.000000  0.320000
0.120000  0.480000  0.000000  0.400000
0.180000  0.400000  0.000000  0.420000
```

W=
({A,T},{A},{A,T},{T},{C,T},{C,T},{C,T},{C,T},{C,T},{T},{T},{T},{C,T},{C,T},{C,T},{C,T},{C,T},{C,T})
$(A|T)A(A|T)T(C|T)^5T^3(C|T)^6$

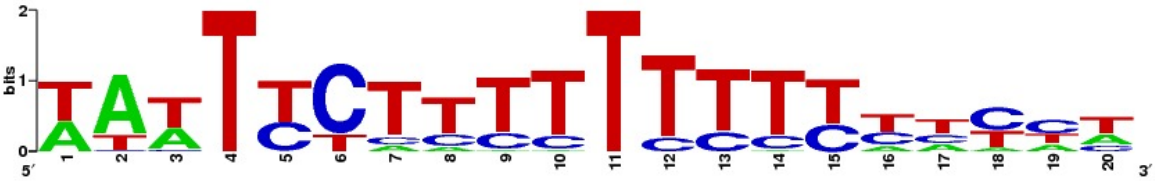
SHORT86 ***l= 19***



```
0.361111  0.083333  0.000000  0.555556
0.305556  0.000000  0.000000  0.694444
0.111111  0.416667  0.000000  0.472222
0.111111  0.361111  0.000000  0.527778
0.361111  0.083333  0.000000  0.555556
0.305556  0.138889  0.000000  0.555556
0.000000  0.083333  0.000000  0.916667
0.027778  0.027778  0.000000  0.944444
0.000000  0.361111  0.000000  0.638889
0.083333  0.194444  0.000000  0.722222
0.000000  0.305556  0.000000  0.694444
0.055556  0.333333  0.000000  0.611111
0.000000  0.555556  0.000000  0.444444
0.000000  0.361111  0.000000  0.638889
0.111111  0.250000  0.000000  0.638889
0.000000  0.166667  0.000000  0.833333
0.000000  0.194444  0.000000  0.805556
0.083333  0.000000  0.000000  0.916667
0.000000  0.444444  0.000000  0.555556
```

W=
({A,T},{A,T},{C,T},{C,T},{A,T},{A,T},{T},{T},{C,T},{T},{C,T},{C,T},{C,T},{C,T},{C,T},{T},{T},{T},{C,T})
$(A|T)^2(C|T)^2(A|T)^2T^2(C|T)T(C|T)^5T^3(C|T)$

SHORT86 ***l=20***

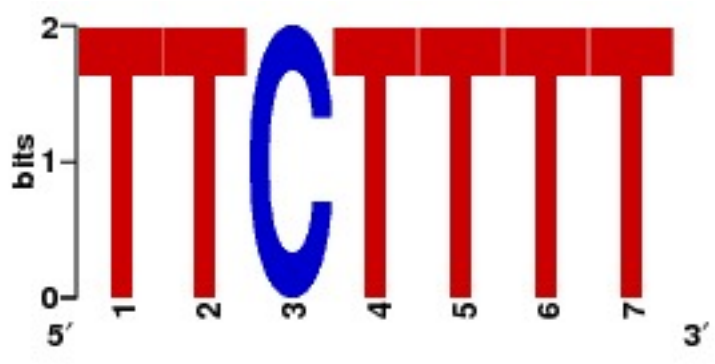


```
0.435897  0.000000  0.000000  0.564103
0.769231  0.025641  0.000000  0.205128
0.384615  0.051282  0.000000  0.564103
0.000000  0.000000  0.000000  1.000000
```

0.000000	0.410256	0.000000	0.589744
0.000000	0.794872	0.000000	0.205128
0.102564	0.128205	0.000000	0.769231
0.102564	0.230769	0.000000	0.666667
0.025641	0.230769	0.000000	0.743590
0.025641	0.179487	0.000000	0.794872
0.000000	0.000000	0.000000	1.000000
0.000000	0.153846	0.000000	0.846154
0.000000	0.256410	0.000000	0.743590
0.025641	0.179487	0.000000	0.794872
0.000000	0.384615	0.000000	0.615385
0.153846	0.358974	0.000000	0.487179
0.256410	0.282051	0.000000	0.461538
0.102564	0.512821	0.000000	0.384615
0.230769	0.461538	0.000000	0.307692
0.307692	0.205128	0.000000	0.487179

W=
({A,T},{A,T},{A,T},{T},{C,T},{C,T},{T},{T},{T},{T},{T},{T},{C,T},{T},{C,T},{C,T}
,{A,C,T},{C,T},{C,T},{A,T})
$(A|T)^3T(C|T)^2T^6(C|T)T(C|T)^2(A|C|T)(C|T)^2(A|T)$

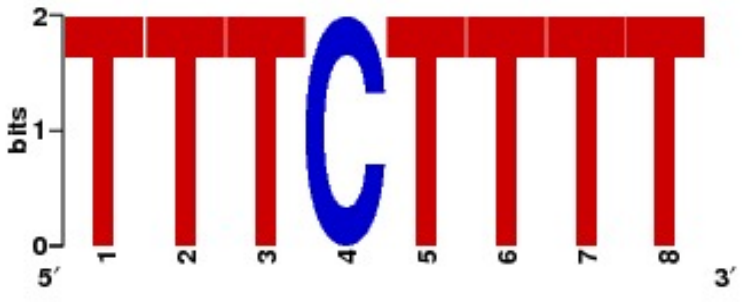
LONG6 ***l=7***



0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	1.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000

W= ({T},{T},{C},{T},{T},{T},{T})
T^2CT^3

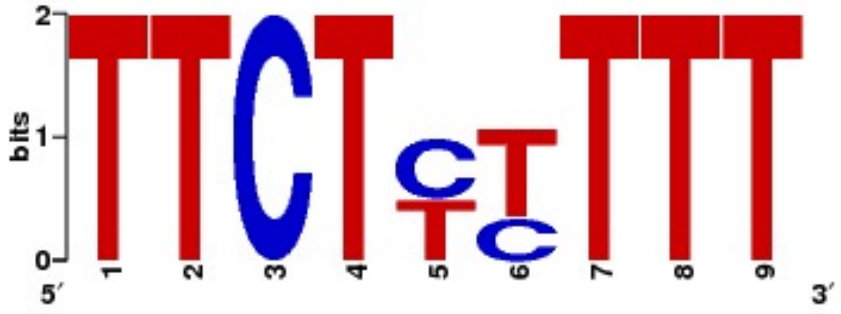
LONG6 ***l=8***




```
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

W= ({T},{T},{T},{C},{T},{T},{T},{T})
T^3CT^4

LONG6 ***l=9***

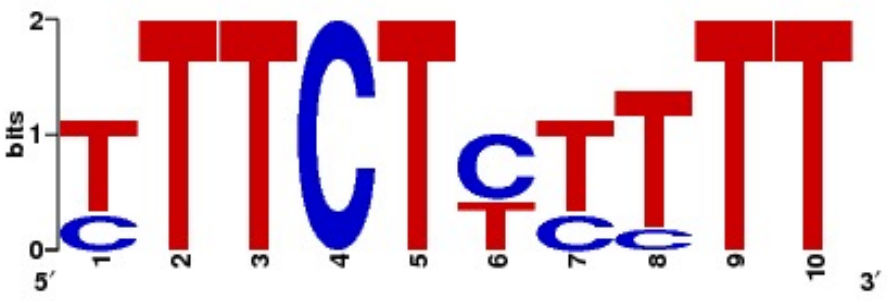



```
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.500000  0.000000  0.500000
0.000000  0.333333  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

```
 0.000000  0.000000  0.000000  1.000000
```

W= ({T},{T},{C},{T},{C,T},{C,T},{T},{T},{T})
$T^2CT(T|C)^2T^3$

LONG6 ***l=10***




```
0.000000  0.285714  0.000000  0.714286
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.571429  0.000000  0.428571
0.000000  0.285714  0.000000  0.714286
0.000000  0.142857  0.000000  0.857143
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
```

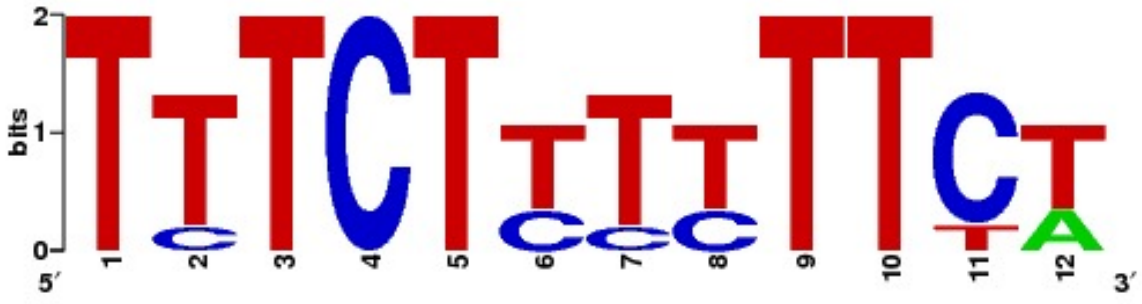
W= ({C,T},{T},{T},{C},{T},{C,T},{C,T},{T},{T},{T})
$(C|T)T^2CT(C|T)^2T^3$

LONG6 ***l=11***

```
0.000000  0.142857  0.000000  0.857143
0.000000  0.142857  0.000000  0.857143
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.428571  0.000000  0.571429
0.000000  0.142857  0.000000  0.857143
0.000000  0.285714  0.000000  0.714286
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.857143  0.000000  0.142857
```

W= ({T},{T},{T},{C},{T},{C,T},{T},{C,T},{T},{T},{C})
$T^3CT(C|T)T(C|T)T^2C$

LONG6 l=12




```
0.000000  0.000000  0.000000  1.000000
0.000000  0.166667  0.000000  0.833333
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.333333  0.000000  0.666667
0.000000  0.166667  0.000000  0.833333
0.000000  0.333333  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.833333  0.000000  0.166667
0.333333  0.000000  0.000000  0.666667
```

W= ({T},{T},{T},{C},{T},{C,T},{T},{C,T},{T},{T},{C},{A,T})
$T^3CT(C|T)T(C|T)T^2C(A|T)$

LONG6 l=13


```
0.000000  0.428571  0.000000  0.571429
0.142857  0.714286  0.000000  0.142857
0.000000  0.000000  0.000000  1.000000
0.000000  0.285714  0.000000  0.714286
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.142857  0.000000  0.000000  0.857143
0.000000  0.285714  0.000000  0.714286
0.000000  0.142857  0.000000  0.857143
0.000000  0.285714  0.000000  0.714286
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.857143  0.000000  0.142857
```

W= ({C,T},{C},{T},{C,T},{T},{C},{T},{C,T},{T},{C,T},{T},{T},{C})
(C|T)CT(C|T)TCT(C|T)T(C|T)T^2C

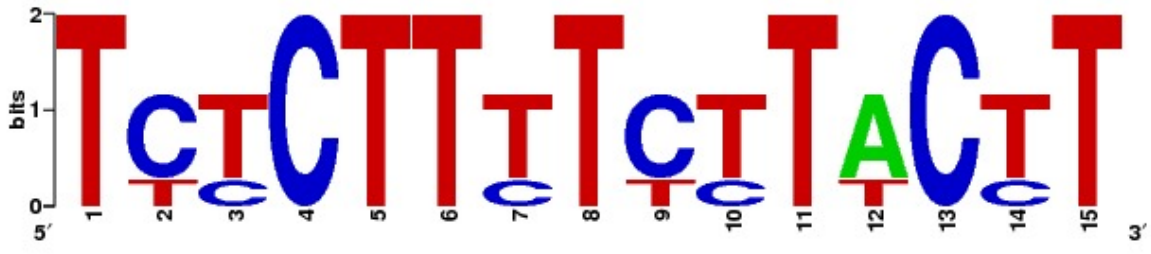
LONG6 l=14


```
0.000000  0.428571  0.000000  0.571429
0.142857  0.714286  0.000000  0.142857
0.000000  0.000000  0.000000  1.000000
0.000000  0.285714  0.000000  0.714286
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
```

0.000000	0.000000	0.000000	1.000000
0.000000	0.428571	0.000000	0.571429
0.142857	0.142857	0.000000	0.714286
0.000000	0.285714	0.000000	0.714286
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.714286	0.000000	0.285714
0.285714	0.000000	0.000000	0.714286

W= ({C,T},{C},{T},{C,T},{T},{C},{T},{C,T},{T},{C,T},{T},{T},{C,T},{A,T})
(C|T)CT(C|T)TCT(C|T)T(C|T)T^2(C|T)(A|T)

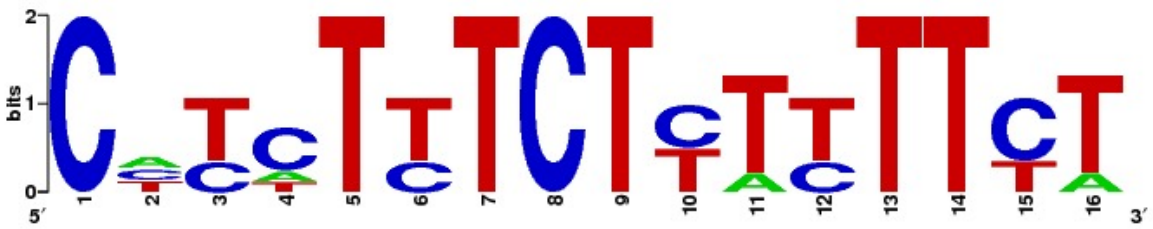
LONG6 ***l=15***



0.000000	0.000000	0.000000	1.000000
0.000000	0.750000	0.000000	0.250000
0.000000	0.250000	0.000000	0.750000
0.000000	1.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.250000	0.000000	0.750000
0.000000	0.000000	0.000000	1.000000
0.000000	0.750000	0.000000	0.250000
0.000000	0.250000	0.000000	0.750000
0.000000	0.000000	0.000000	1.000000
0.750000	0.000000	0.000000	0.250000
0.000000	1.000000	0.000000	0.000000
0.000000	0.250000	0.000000	0.750000
0.000000	0.000000	0.000000	1.000000

W= ({T},{C,T},{C,T},{C},{T},{T},{C,T},{T},{C,T},{C,T},{T},{A,T},{C},{C,T},{T})
T(C|T)^{2}CT^2(C|T)T(C|T)^{2}T(A|T)C(C|T)T

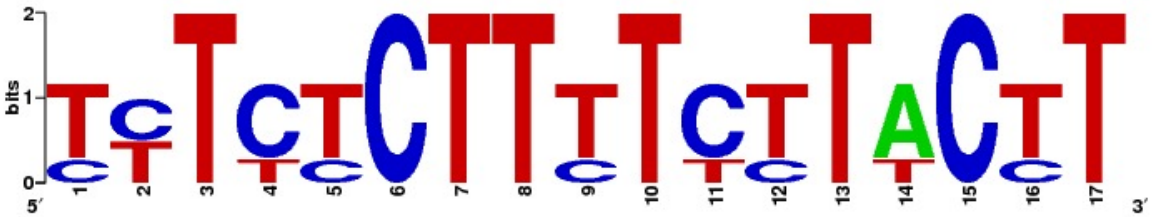
LONG6 ***l=16***



```
0.000000  1.000000  0.000000  0.000000
0.333333  0.333333  0.000000  0.333333
0.000000  0.333333  0.000000  0.666667
0.166667  0.666667  0.000000  0.166667
0.000000  0.000000  0.000000  1.000000
0.000000  0.333333  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.500000  0.000000  0.500000
0.166667  0.000000  0.000000  0.833333
0.000000  0.333333  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  0.000000  0.000000  1.000000
0.000000  0.666667  0.000000  0.333333
0.166667  0.000000  0.000000  0.833333
```

W=
({C},{A,C,T},{C,T},{C},{T},{C,T},{T},{C},{T},{C,T},{T},{C,T},{T},{T},{C,T},{T})
C(A|C|T)(C|T)CT(C|T)TCT(C|T)T(C|T)T^2(C|T)T

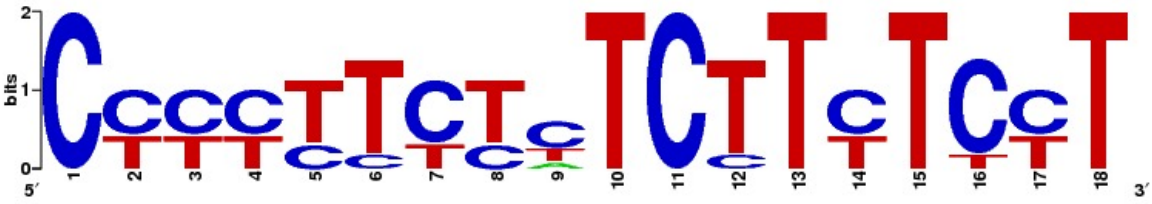
LONG6 ***l=17***



0.000000	0.250000	0.000000	0.750000
0.000000	0.500000	0.000000	0.500000
0.000000	0.000000	0.000000	1.000000
0.000000	0.750000	0.000000	0.250000
0.000000	0.250000	0.000000	0.750000
0.000000	1.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.250000	0.000000	0.750000
0.000000	0.000000	0.000000	1.000000
0.000000	0.750000	0.000000	0.250000
0.000000	0.250000	0.000000	0.750000
0.000000	0.000000	0.000000	1.000000
0.750000	0.000000	0.000000	0.250000
0.000000	1.000000	0.000000	0.000000
0.000000	0.250000	0.000000	0.750000
0.000000	0.000000	0.000000	1.000000

W=
({C,T},{C,T},{T},{C,T},{C,T},{C},{T},{T},{C,T},{T},{C,T},{C,T},{T},{A,T},{C},{C,T},{T})
$(C|T)^2T(C|T)^2CT^2(C|T)T(C|T)^2T(A|T)C(C|T)T$

LONG6 ***l=18***

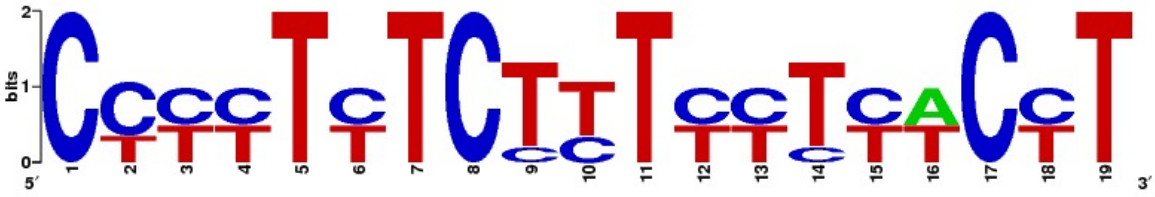


0.000000	1.000000	0.000000	0.000000
0.000000	0.571429	0.000000	0.428571
0.000000	0.571429	0.000000	0.428571
0.000000	0.571429	0.000000	0.428571
0.000000	0.285714	0.000000	0.714286
0.000000	0.142857	0.000000	0.857143
0.000000	0.714286	0.000000	0.285714
0.000000	0.285714	0.000000	0.714286
0.142857	0.571429	0.000000	0.285714
0.000000	0.000000	0.000000	1.000000
0.000000	1.000000	0.000000	0.000000
0.000000	0.142857	0.000000	0.857143
0.000000	0.000000	0.000000	1.000000

```
0.000000  0.571429  0.000000  0.428571
0.000000  0.000000  0.000000  1.000000
0.000000  0.857143  0.000000  0.142857
0.000000  0.571429  0.000000  0.428571
0.000000  0.000000  0.000000  1.000000
```

W=
({C},{C,T},{C,T},{C,T},{C,T},{T},{C,T},{C,T},{C,T},{T},{C},{T},{T},{C,T},{T},{C},{C,T},{T})
$C(C|T)^4T(C|T)^3TCT^2(C|T)TC(C|T)T$

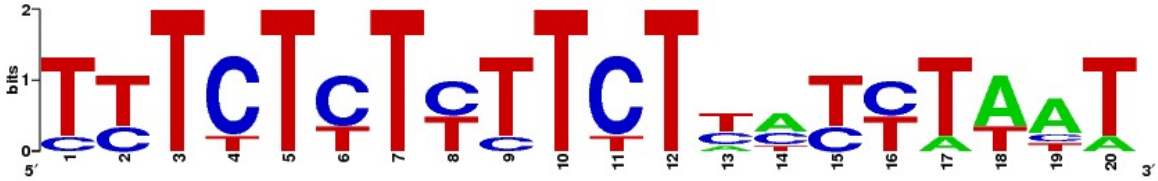
LONG6 ***l=19***




```
0.000000  1.000000  0.000000  0.000000
0.000000  0.666667  0.000000  0.333333
0.000000  0.500000  0.000000  0.500000
0.000000  0.500000  0.000000  0.500000
0.000000  0.000000  0.000000  1.000000
0.000000  0.500000  0.000000  0.500000
0.000000  0.000000  0.000000  1.000000
0.000000  1.000000  0.000000  0.000000
0.000000  0.166667  0.000000  0.833333
0.000000  0.333333  0.000000  0.666667
0.000000  0.000000  0.000000  1.000000
0.000000  0.500000  0.000000  0.500000
0.000000  0.500000  0.000000  0.500000
0.000000  0.166667  0.000000  0.833333
0.000000  0.500000  0.000000  0.500000
0.500000  0.000000  0.000000  0.500000
0.000000  1.000000  0.000000  0.000000
0.000000  0.500000  0.000000  0.500000
0.000000  0.000000  0.000000  1.000000
```

W=
({C},{C,T},{C,T},{C,T},{T},{C,T},{T},{C},{T},{C,T},{T},{C,T},{C,T},{T},{C,T},{A,T},{C},{C,T},{T})
$C(C|T)^3T(C|T)TCT(C|T)T(C|T)^2T(C|T)(A|T)C(C|T)T$

LONG6 ***l=20***



0.000000	0.166667	0.000000	0.833333
0.000000	0.333333	0.000000	0.666667
0.000000	0.000000	0.000000	1.000000
0.000000	0.833333	0.000000	0.166667
0.000000	0.000000	0.000000	1.000000
0.000000	0.666667	0.000000	0.333333
0.000000	0.000000	0.000000	1.000000
0.000000	0.500000	0.000000	0.500000
0.000000	0.166667	0.000000	0.833333
0.000000	0.000000	0.000000	1.000000
0.000000	0.833333	0.000000	0.166667
0.000000	0.000000	0.000000	1.000000
0.166667	0.333333	0.000000	0.500000
0.500000	0.333333	0.000000	0.166667
0.000000	0.333333	0.000000	0.666667
0.000000	0.500000	0.000000	0.500000
0.166667	0.000000	0.000000	0.833333
0.666667	0.000000	0.000000	0.333333
0.666667	0.166667	0.000000	0.166667
0.166667	0.000000	0.000000	0.833333

W=
({T},{C,T},{T},{C},{T},{C,T},{T},{C,T},{T},{T},{C},{T},{C,T},{A,C},{C,T},{C,T},{T},{A,T},{A},{T})

T(C|T)TCT(C|T)T(C|T)T^2CT(C|T)(A|C)(C|T)^{2}T(A|T)AT

Bibliography

Sohlberg, M. M., & Mateer, P. A. (2001). *Cognitive Rehabilitation: An Integrative Neuropsychological Approach* (2nd ed.). The Guilford Press;.

Aggarwal, A., & Suri, S. (1987). Fast algorithms for computing the largest empty rectangle. *Proceedings of the 3rd Annual Symposium on Computational Geometry*, (pp. 278-290).

Aggarwal, C. (2014). *Data Classification: Algorithms and Applications.* Publisher Chapman and Hall/CRC.

Aggarwal, C., & Reddy, C. (2013). *Data Clustering: Algorithms and Applications* (Vol. Data Mining and Knowledge Discovery Series). Chapman & Hall/CRC.

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of 20th International Conference on Very Large Data Bases*, (pp. 487-499). Santiago de Chile.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *International Conference on Data Engineering*, (pp. 3-14). Taipei.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *SIGMOD*, (pp. 207-216).

Allen, D. N., Leany, B. D., Thaler, N. S., Cross, C., Sutton, G. P., & Mayfield, J. (2010). Memory and attention profiles in pediatric traumatic brain injury. *Archives of Clinical Neuropsychology, 25*(7), 618-633.

Allen, D. N., Thaler, N. S., Cross, C., & Mayfield, J. (2013). Classification of Traumatic Brain Injury Severity: A Neuropsychological Approach. In *Cluster Analysis in Neuropsychological Research* (pp. 95-123). Springer Publishing Company.

Andrews, P. J., Sleeman, D. H., Statham, P. F., McQuatt, A., Corruble, V., Jones, P. A., . . . Macmillan, C. S. (2002). Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. *Journal of Neurosurgery, 97*(2), 326-336.

Arabie, P., & Hubert, L. (1996). An overview of combinatorial data analysis. In P. Arabie, & L. Hubert, *Clustering and Classification* (pp. 5-63). World Scientific Publishing Co.

Aribisala, B. S., Cowie, C. J., He, J., Wood, J., Mendelow, D., Mitchell, P., & Blamire, A. (2010). Multi-parametric Classification of Traumatic Brain Injury Patients Using Automatic Analysis of Quantitative MRI Scans. *Medical Imaging and Augmented Reality Lecture Notes in Computer Science, 6326*, 51-59.

Artiola i Fortuny, L., Hermosillo Romo, D., Heaton, R. K., & Pardee III, R. E. (1999). *Manual de normas y procedimientos para la batería neuropsicológica en Español.* Tucson: mPress.

Augustine, J., Das, S., Maheshwari, A., Nandy, S. C., Roy, S., & Sarvattomananda, S. (2010). Recognizing the largest empty circle and axis-parallel rectangle in a desired location. *Computing Research Repository*.

Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002). Sequential pattern mining using a bitmap representation. *Proc. 8th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, 429–435.

Bailey, T. L., & Elkan, C. P. (1995). The value of prior knowledge in discovering motifs with MEME. In T. I. Biology (Ed.). (pp. 21-29). AAAI Press.

Baird, H. S., Jones, S. E., & Fortune, S. J. (1990). Image segmentation by shape-directed covers. *Proc. 10th Internat. Conf. Pattern Recognition*, *1*, pp. 820–825.

Benzer, A., Traweger, C., & Ofner, D. (1995). Statistical modelling in analysis of outcome after trauma: Glasgow Coma Scale and Innsbruck Coma Scale. *Anasthesiol Intensivmed Notfallmed Schmerzther, 30*, 231-235.

Bernabeu, M., & Roig, T. (1999). *La rehabilitación del traumatismo craneoencefálico: un enfoque interdisciplinar.* (F. I. Guttmann, Ed.) Barcelona.

Berry, M., & Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons.

Brain Injury Association of America. (2015, September). Retrieved October 2015, from http://www.biausa.org/about-brain-injury.htm#types

Brown, A. W., McClelland, R. L., Diehl, N. N., Englander, J., & Cifu, D. X. (2006). Clinical elements that predict outcome after traumatic brain injury: a prospective multicenter recursive partitioning (decision-tree) analysis. *Journal of Neurotrauma, 22*(12).

Burred, J. J. (2012). Genetic motif discovery applied to audio analysis. *International Conference on Acoustics, Speech and Signal Processing* (pp. 361-364). IEEE.

Calero, M. D., & Navarro, E. (2007). Cognitive plasticity as a modulating variable on the effects of memory training in elderly persons. *Archives of Clinical Neuropsychology, 22*, 63–72.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*, 1–27.

Chaves, R., Ramírez, J., Górriz, J. M., & Illán, I. A. (2012). Functional brain image classification using association rules defined over discriminant regions. *Pattern Recognition Letters, 33*(12), 1666-1672.

Chazelle, B., Drysdale, R. L., & Lee, D. T. (1986). Computing the largest empty rectangle. *SIAM Journal on Computing, 15*(1), 300 - 315.

Chesney, T., Penny, K., Oakley, P., Davies, S., Chesney, D., Maffulli, N., & Templeton, J. (2009). Data mining trauma injury data using C5.0 and logistic regression to determine factors associated with death. *International Journal of Healthcare Technology and Management, 10*(1), 16 – 26.

Cicerone, K. D., & Tupper, D. E. (1986). *Cognitive assessment in the neuropsychological rehabilitation of head-injured adults*, 59–84.

Cicerone, K. D., Langenbahn, D. M., Braden, C., Malec, J. F., Kalmar, K., Fraas, M., . . . Ashman, T. (2011). Evidence-based cognitive rehabilitation: updated review of the literature from 2003 through 2008. *Archives of Physical Medicine and Rehabilitation, 92*(4), 519-530.

Conners, C. K., & Sitarenios, G. (2011). Conners' Continuous Performance Test. *Encyclopedia of Clinical Neuropsychology*, 681-683.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27.

Cristianini , N., & Shawe-Taylor , J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (1 ed.). Cambridge University Press.

Crosson, B., Greene, R., Roth, D. L., Farr, S. P., & Adams, R. L. (1990). WAIS-R cluster analysis in a heterogeneous sample of blunt head injury. *The Clinical Neuropsychologist, 4*, 253-262.

Crump, C., Silvers, C. T., Wilson, B., Schlachta-Fairchild, L., & Ashley, J. (2014). Predicting patient outcomes via neural network estimation of discharge APACHE scores for traumatic brain injury. *American Journal of Health Research, 2*(6), 361-365.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience.* Harper Perennial Modern Classics.

D'Haeseleer, P. (2006). How does DNA sequence motif discovery work? *Nature Biotechnology*, 959–961.

Dasgupta, A., Sun, Y., König, I., Bailey-Wilson, J., & Malley, J. (2011). Brief Review of Regression-Based and Machine Learning Methods in Genetic Epidemiology: The Genetic Analysis Workshop 17 Experience. *Genetic Epidemiology*, 5-11.

de Noreña, D., Ríos-Lago, M., Bombín-González, I., Sánchez-Cubillo, I., García-Molina, A., & Tirapu-Ustárroz, J. (2010). Efectividad de la rehabilitación neuropsicológica en el daño cerebral adquirido (I): atención, velocidad de procesamiento, memoria y lenguaje. *51*(11), 687-698.

de Rham, C. (1980). La classification hiérarchique ascendante selon la méthode des voisins réciproques. *Les Cahiers de l'Analyse des Données*, 135-144.

Dinsmore, J. (2013, December). Traumatic brain injury: an evidence-based review of management. *Continuing Education in Anaesthesia, Critical Care & Pain, 13*(6), 189-195.

DiPiro, J. T., & Spruill, W. J. (2010). *Concepts in clinical pharmacokinetics* (5 ed.). American Society of Health-System Pharmacists.

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Naive Bayes Classifier under Zero-One Loss. *Machine Learning, 29*(2), 103-130.

Donders, J., & Warschausky, S. (1997). WISC-III factor index score patterns after traumatic head injury in children. *Child Neuropsychology, 3*(1), 71–78.

Duda, R., & Hart, P. (1973). *Pattern Classification and Scene Analysis.* New York: John Wiley and Sons.

Dumitrescu, J. A., & Jiang, M. (2013). On the Largest Empty Axis-Parallel Box Amidst n Points. *Algorithmica, 66*(2), 225-248.

Dunham, M. H. (2003). *Data mining introductory and advanced topics.* (U. S. River, Ed.) New Jersey: Pearson Education, Inc.

ECRI. (2011). *Cognitive Rehabilitation Therapy for Traumatic Brain Injury: What We Know and Don't Know about Its Efficacy.* IOM's New Report.

Edmonds, J., Gryz, J., Liang, D., & Miller, R. J. (2003). Mining for empty spaces in large data sets. *Theoretical Computer Science, 296*, 435–452.

Eftekhar, B., Mohammad, K., Ardebili, H., Ghodsi, M., & Ketabchi, E. (2005). Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Medical Informatics Decision Making , 5*(3).

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd ACM SIGKDD*, (pp. 226-231). Portland.

Fann, J., Hart, T., & Schomer, K. (2009). Treatment for depression after traumatic brain injury: a systematic review. *Journal of Neurotrauma, 26*(12), 2383-23402.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth , P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 37-54.

Finneran, C. M., & Zhang, P. (2005). Flow in computer-mediated environments:Promises and challenges. *Communications of the Association for Information Systems, 15*, 82–101.

Fleming, J. M., Strong, J., & Ashton, R. (1998). Cluster analysis of self-awareness levels in adults with traumatic brain injury and relationshipto outcome. *Journal of Head Trauma Rehabilitation, 13*(5), 39-51.

Fournier-Viger, P., Gomariz, A., Campos, M., & Thomas, R. (2014). Fast Vertical Sequential Pattern Mining Using Co-occurrence Information. *Proc. 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining.*

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods.* SAGE Publications.

Garcia, M. C., Martins, E. T., & Azevedo, F. M. (2013). Decision Tree Induction to Prediction of Prognosis in Severe Traumatic Brain Injury of Brazilian Patients from Florianopolis City. *Bioinformatics and Bioengineering (BIBE)* (pp. 1-4). IEEE.

García-Rudolph, A., & Gibert, K. (2014). A data mining approach to identify cognitive NeuroRehabilitation Range in Traumatic Brain Injury patients. *Expert Systems with Applications*, 5238 - 5251.

García-Rudolph, A., & Gibert, K. (2015). Data Mining Approach for Visual and Analytical Identification of Neurorehabilitation Ranges in Traumatic Brain Injury Cognitive Rehabilitation. *Abstract and Applied Analysis*, 1-14.

Gholipour, C., Rahim, F., Fakhree, A., & Ziapour, B. (2015). Using an Artificial Neural Networks (ANNs) Model for Prediction of Intensive Care Unit (ICU) Outcome and Length of Stay at Hospital in Traumatic Patients. *Journal of Clinical and Diagnostic Research , 9*(4), 19-23.

Gibert, K., & García-Rudolph, A. (2006). Posibilidades de aplicación de minería de datos para el descubrimiento de. In *Desarrollo de herramientas para evaluar el resultado de las tecnologías aplicadas al proceso rehabilitador Estudio a partir de dos modelos concretos: Lesión Medular y Daño Cerebral Adquirido* (pp. 57-61). Agència d'Avaluació de Tecnología i Recerca Mèdica de Catalunya.

Gibert, K., & Tormos, J. M. (2014). *Pan European Networks - Science and Technology, 12*, 125.

Gibert, K., Aluja, T., & Cortés, U. (1998). Knowledge Discovery with Clustering Based on Rules. Interpreting Results. *Proceedings of Principals of Data Mining and Knowledge Discovery* (pp. 83-92). SPRINGER-VERLAG.

Gibert, K., García-Rudolph, A., Curcoll, L., Pla, L., & Tormos, J. M. (2009). Knowledge discovery about quality of life changes of spinal cord injury patients: clustering based on rules by states. *Studies in Health Technology and Informatics*, (pp. 579 - 583).

Gibert, K., García-Rudolph, A., García-Molina, A., Roig-Rovira, T., Bernabeu, M., & Tormos, J. M. (2008). Knowledge Discovery on the Response to Neurorehabilitation Treatment of

Patients with Traumatic Brain Injury through an AI&Stats and Graphical Hybrid Methodology. *Artificial Intelligence Research and Development. Frontiers in Artificial Intelligence and Applications, 184*, 170-177.

Gil Origüén, A. (2009). *Laboratorio de medidas potenciadoras de la autonomía, satisfacción personal y calidad de vida de las personas con lesión medular y daño cerebral;tres años de investigación en calidad de vida y discapacidad : memoria 2006/2008;* (1 ed., Vols. Informes, estudios e investigación / Ministerio de Sanidad y Política Social). Madrid: Ministerio de Sanidad y Política Social.

Golden, C. (1994). *STROOP Test de colores y palabras.* Madrid: TEA.

Gomariz, A., Campos, M., Marin, R., & Goethals, B. (2013). ClaSP: An Efficient Algorithm for Mining Frequent Closed Sequences. In J. Pei (Ed.), *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). 7818*, pp. 50–61. Springer.

Green, C. S., & Bavelier, D. (2006). The cognitive neuroscience of video games. (L. Humphreys, & P. Messaris, Eds.) *Digital Media: Transformations in human communications*, 211-223.

Green, C. S., & Bavelier, D. (2007). *Psychological Science, 18*(1), 88–94.

Greenhalgh, T. (2006). *How to read a paper. The basics of evidence based medicine.* Oxford: Blackwell Publishing Ltd.

Güler, I., Gökçil, Z., & Gülbandilar, E. (2009). Evaluating of traumatic brain injuries using artificial neural networks. *Expert Systems with Applications, 36*(7), 10424-10427.

Hall, M., Frank, E., Holmes, G., Pfahringer , B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18 .

Hamerly, G., & Elkan, C. (2003). Learning the K in K-means. *Proceedings of the 17th Annual Conference on Neural Information Processing Systems.*

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD*, *29*, pp. 1-12.

Hart, T., Whyte, J., Kim, J., & Vaccaro, M. (2005). Executive function and self-awareness of "real-world" behavior and attention deficits following traumatic brain injury. *Journal of Head Trauma Rehabilitation, 4*, 333-347.

Hartigan, J. (1975). *Clustering Algorithms.* New York: John Wiley & Sons.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1997). *WCST: Test de clasificación de tarjetas Wisconsin.* Madrid: TEA.

Heuser, A., Kourtev, H., Winter, S., Fensterheim, D., Burdea, G., Hentz, V., & Forducey, P. (2007). Telerehabilitation using the Rutgers Master II glove following carpal tunnel release

surgery: proof-of-concept. *IEEE Transactions Neural System Rehabilitation Engineering, 15*(1), 43-49.

Hinneburg, A., & Keim, D. (1998). An efficient approach to clustering large multimedia databases with noise. *Proceedings of the 4th ACM SIGKDD*, (pp. 58-65). New York.

Hu, F. (2002). Dietary pattern analysis: a new direction in nutritional epidemiology. *Current Opinion in Lipidology, 13*(1), 3-9.

Hukkelhoven, C. W., Steyerberg, E. W., Habbema, J. D., Farace, E., Marmarou, A., Murray, G. D., . . . Maas, A. I. (2005). Predicting outcome after traumatic brain injury: development and validation of a prognostic score based on admission characteristics. *Journal of Neurotrauma, 22*(10), 1025-1039.

Hunt, E. B. (1962). *Concept Learning: An Information Processing Problem.* John Wiley and Sons.

Institut Guttmann - Hospital de Neurorehabilitació. (2015, 09 01). Retrieved from http://www.guttmann.com/

Jagaroo, V. (2009). *Neuroinformatics for Neuropsychology* (1 ed.). New York: Springer-Verlag New York.

Jain, A., & Dubes, R. (1988). *Algorithms for Clustering Data.* Englewood Cliffs: Prentice-Hall.

Jawad, A., Kersting, K., & Andrienko, N. (2011). Where traffic meets DNA: mobility mining using biological sequence analysis revisited. *19th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems* (pp. 357-360). ACM-GIS.

Ji, S. Y., Smith, R., Huynh, T., & Najarian, K. (2009). A comparative analysis of multi-level computer-assisted decision making systems for traumatic injuries. *BMC Medical Informatics & Decision Making, Vol. 9 Issue 1*.

Johnston, M., Sherer, M., & Whyte, J. (2006). Applying evidence standards to rehabilitation research. *American Journal Physical Medical Rehabilitation, 85*(4), 292-309.

Kamalakannan, S. K., Gudlavalleti , A. S., Murthy Gudlavalleti, V. S., Goenka, S., & Kuper, H. (2015). Challenges in understanding the epidemiology of acquired brain injury in India. (Medknow, Ed.) *Annals of Indian Academy of Neurology, 18*(1), 66-70.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley and Sons.

Kleim, J. A., & Jones, T. A. (2008). Principles of experience-dependent neural plasticity: implications for rehabilitation after brain damage. *Journal of Speech Language and Hearing Research, 51*(1), 225-239.

Klema, J., Novakova, L., Karel, F., Stepankova, O., & Zelezny, F. (2008). Sequential data mining: A comparative case study in development of atherosclerosis risk factors. *IEEE Transactions on Systems, Man, and Cybernetics: Part C: Applications and Reviews, 38*(1), 3-15.

Klement, W., Wilk, S., Michalowski, W., & Matwin, S. (2011). Classifying Severely Imbalanced Data. *Advances in Artificial Intelligence Lecture Notes in Computer Science, 6657*, 258-264.

Klement, W., Wilk, S., Michalowski, W., Farion, K. J., Osmond, M. H., & Verter, V. (2012, March). Predicting the need for CT imaging in children with minor head injury using an ensemble of Naive Bayes classifiers. *Artificial Intelligence in Medicine, 54*(3), 163-170 .

Klepper, K., & Drabløs, F. (2010). PriorsEditor: a tool for the creation and use of positional priors in motif discovery. *Bioinformatics, 26*(17), 2195-2197.

Kononenko, I. (1990). Comparison of inductive and Naive Bayes learning approaches to automatic knowledge acquisition. In Wielinga, *Current Trends in Knowledge Acquisition.* Amsterdam: IOS Press.

Korn, L. J., Queen, C. L., & Wegman, M. N. (1977). Computer analysis of nucleic acid regulatory sequences. *Proceedings of the National Academy of Sciences, 74*(10), 4401-4405.

Kouroupetroglou, G. (2013). *Assistive Technologies and Computer Access for Motor Disabilities.* IGI Global Release.

Lang, E., Pitts, L. H., Damron, S. L., & Rutledge, R. (1997). Outcome after severe head injury: an analysis of prediction based upon comparison of neural network versus logistic regression analysis. *Neurological Research, 19*(3), 274-280.

Larsson, J., Björkdahl, A., Esbjörnsson, E., & Sunnerhagen , K. S. (2013). Factors affecting participation after traumatic brain injury. *Journal of Rehabilitation Medicine, 45*(8), 765-770.

Lezak, M. D. (1995). *Neuropsychological Assessment.* New York: Oxford University Press.

Linkenhoker, B. A., & Knudsen , E. I. (2002). Incremental training increases the plasticity of the auditory space map in adult barn owls. *Nature*, 293-296.

Luria, A. R. (1976). *The Working Brain: An Introduction to Neuropsychology.* Basic Books.

Ma, M., McNeill, m., Charles, D., McDonough, S., Crosbie, J., Oliver, L., & McGoldrick, C. (2007). Adaptive Virtual Reality Games for Rehabilitation of Motor Disorders. *Universal Access in Human-Computer Interaction. Ambient Interaction, 4555*, 681-690.

Maimon, O., & Last, M. (2001). *Knowledge Discovery and Data Mining: The Info-Fuzzy Network (IFN) Methodology.* (M. C. series, Ed.) Kluwer Academic Publishers.

Maimon, O., & Rokach, L. (2005). *Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications* (Vol. 61). (S. i. Intelligence, Ed.) World Scientific Publishing.

Maleca, J. F., Machuldaa, M. M., & Smigielskia, J. S. (1993). Cluster analysis of neuropsychological test results among patients with traumatic brain injury (TBI): Implications for a model of TBI-related disability. *Clinical Neuropsychologist, 7*(1), 48-58.

Marcano-Cedeño, A., Chausa, P., García-Rudolph, A., Cáceres, C., Tormos, J. M., & Gómez, E. (2013). Artificial metaplasticity prediction model for cognitive rehabilitation outcome in acquired brain injury patients. *Artificial Intelligence in Medicine, 58*(2), 91-99.

Martins, E. T., Linhares, M. N., Sousa, D. S., Schroeder, H. K., Meinerz, J., Rigo, L. A., . . . Walz, R. (2009). Mortality in severe traumatic brain injury: a multivariated analysis of 748 Brazilian patients from Florianópolis City. *Journal of Trauma, 67*(1), 85-90.

Massart, D., & Kaufman, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis.* New York: John Wiley & Sons.

Mathwick, C., & Rigdon, E. (2004). Play, Flow, and the online search experience. *Journal of Consumer Research, 31*(2), 324–332.

McBride, J., Zhao, X., Nichols, T., & Abdul-Ahad, T. (2011). Classification of traumatic brain injury using support vector machine analysis of event-related Tsallis entropy. *Biomedical Sciences and Engineering Conference (BSEC)* (pp. 1-4). IEEE.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models.* Chapman & Hall.

Mckenna, M., O'Rourke, S., & Suri, S. (1985). Finding the largest rectangle in an orthogonal polygon. *Proceedings of the 23rd Annual Allerton Conference on Communication.* Illinois: Control and Computing.

McQuatt, A., Sleeman, D., Andrews, P., Corruble, V., & Jones, P. (2001). Discussing anomalous situations using decision trees: a head injury case study. *Methods of Information in Medicine, 40*(5), 373-379.

Menon, D. K., Schwab, K., Wright, D. W., & Maas, A. I. (2010). Position statement: definition of traumatic brain injury. *Archives of Physical Medicine and Rehabilitation, 91*(11), 1637-1640.

Millen, B. E., Quatromoni, P. A., & Gagnon, D. R. (1996). Dietary patterns of men and women suggest targets for health promotion: the Framingham Nutrition Studies. *American Journal of Health Promotion, 11*, 42-53.

Min-Huei, H., Yu-Chuan, L., Wen-Ta, C., & Ju-Chuan, Y. (2005). Outcome Prediction after Moderate and Severe Head Injury Using an Artificial Neural Network. In Engelbrecht, & Engelbrecht (Ed.), *Connecting Medical Informatics and Bio-Informatics.*

Naamad, A., Lee, D. T., & Hsu, W. L. (1984). On the maximum empty rectangle problem. *Discrete Applied Mathematics*, 267–277.

Nimmo, G. R. (2011). *ABC of Intensive Care* (2nd ed.). (M. Singer, Ed.) BMJ Books.

Orvis, K., Horn, D., & Belanich, J. (2008). The roles of task difficulty and priorvideogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior, 24*, 2415–2433.

Pang, B. C., Kuralmani , V., Joshi, R., Hongli, Y., Lee , K. K., Ang, B. T., . . . Ng, I. (2007). Hybrid outcome prediction model for severe traumatic brain injury. *Journal of Neurotrauma, 24*(1), 136-146.

Pascual-Leone, A., Amedi, A., Fregni, F., & Merabet, L. B. (2005). The Plastic Human Brain Cortex. *Annual Review of Neuroscience, 28*, 377-401.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., . . . Hsu, M. (2004). Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Trans. Knowledge Data Engineering, 16*(11), 1424–1440.

Piatetsky-Shapiro , G., Brachman , R., Khabaza , T., Kloesgen , W., & Simoudis , E. (1996). An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. In E. Simoudis, J. Han , & U. Fayyad (Ed.), *Second International Conference on Knowledge Discovery and Data Mining* (pp. 89-95). AAAI Press.

Pignolo, L., & Lagani, V. (2011). Prediction of Outcome in the Vegetative State by Machine Learning Algorithms: A Model for Clinicians? *Journal of Software Engineering and Applications, 4*, 388-390.

Pilih, I., Mladenić, D., Lavrač, N., & Prevec, T. (1997). Data Analysis of Patients with Severe Head Injury. *Intelligent Data Analysis in Medicine and Pharmacology, 414*, 131-148.

Pradhan, G. N., & Prabhakaran, B. (2009). Association rule mining in multiple, multidimensional time series medical data. *IEEE international conference on Multimedia and Expo*, (pp. 1716-1719).

Prvu Bettger, J., & Stineman, M. (2007). Effectiveness of multidisciplinary rehabilitation services in postacute care: state-of-the-science. A review. *Archives of Physical Medicine and Rehabilitation, 88*(11), 1526-1534.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106.

Rego, P., Moreira, P. M., & Reis, L. P. (2010). Serious games for rehabilitation a survey and a classification towards a taxonomy. *Information systems and technologies*, 1–6.

Reitan , R. M., & Wolfson , D. (1993). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation* (2 nd ed.). Neuropsychology Press.

Rey, A. (1964). *L'examen clinique en psychologie.* Paris: Presses universitaires de France.

Ribeiro, M. X., Bugatti, P. H., Traina Jr, C., Marques, P., Rosa, N. A., & Traina, A. (2009). Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques. *Data and Knowledge Engineering, 68*(12), 1370–1382.

Rohling, M. L., Faust, M. E., Beverly, B., & Demakis, G. (2009). Effectiveness of cognitive rehabilitation following acquired brain injury: a meta-analytic re-examination of Cicerone et al.'s (2000, 2005) systematic reviews. *Neuropsychology, 23*(1), 20-39.

Rovlias, A., & Kotsou, S. (2004). Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables. *Journal of Neurotrauma, 21*(7), 886-893.

Rughani, A. I., Dumont, T. M., Lu, Z., Bongard, J., Horgan, M. A., Penar, P. L., & Tranmer, B. I. (2010, September). Use of an artificial neural network to predict head injury outcome. *Journal of Neurosurgery, 113*(3), 585-590.

Rusnak, M. (2013, April). Traumatic brain injury: Giving voice to a silent epidemic. *Nature Reviews Neurology*, 186-187.

Sabb, F. W., Burggren, A. C., Higier, R. G., Fox, J., He, J., Parker, D. S., . . . Bilder, R. M. (2009). Challenges in phenotype definition in the whole-genome era: multivariate models of memory and intelligence. *Neuroscience, 164*(1), 88-107.

Sakellaropoulos, G. C., & Nikiforidis, G. (1999). Development of a Bayesian Network for the prognosis of head injuries using graphical model selection techniques. *Methods of Information in Medicine, 38*, 37-42.

Schneider, T. D., & Stephens, R. M. (1990). Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Research*, 6097-6100.

Scholten, A. C., Haagsma, J. A., Panneman, M. J., van Beeck, E. F., & Polinder, S. (2014). Traumatic Brain Injury in the Netherlands, incidence, costs and disability adjusted life years. *PLoS One*, 9-10.

Segal, M. E., Goodman, P. H., Goldstein, R., Hauck, W., Whyte , J., Graham, J. W., . . . Hammond, F. M. (2006). The accuracy of artificial neural networks in predicting long-term outcome after traumatic brain injury. *Journal of Head Trauma Rehabilitation, 21*(4), 298-314.

Serra, J., Arcos, J. L., García-Rudolph, A., García-Molina, A., Roig, T., & Tormos, J. M. (2013). Cognitive prognosis of acquired brain injury patients using machine learning techniques. *Int. Conf. on Advanced Cognitive Technologies and Applications (COGNITIVE)*, (pp. 108-113). Valencia.

Silver, M., Sakara, T., Su, H. C., Herman, C., Dolins, S. B., & O'Shea, M. J. (2001). Case study: how to apply data mining techniques in a healthcare data warehouse. *Journal of Healthcare Information Management, 15*(2), 155-164.

Snell, D. L., Surgenor, L. J., Hay-Smith, E. J., Williman, J., & Siegert, R. J. (2015). The contribution of psychological factors to recovery after mild traumatic brain injury: is cluster analysis a useful approach? . *Brain Injury, 29*(3), 291-299.

Solana, J., Cáceres, C., Ferrer-Celma, S., Ferre-Bergada, M., García-López, P., García-Molina , A., . . . Tormos, J. M. (2011). PREVIRNEC A new platform for cognitive tele-rehabilitation. *COGNITIVE 2011, The Third International Conference on Advanced Cognitive Technologies and Applications*, (pp. 59-62).

Srikant, R., & Agrawal, R. (1996). Mining Sequential Patterns: Generalizations and Performance Improvements. *Extending Database Technology (EDBT) 1996. LNCS. 1057*, pp. 3–17. Springer.

Sujin, K., Woojae, K., & Rae, W. (2011). A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare Information Research, 17*, 232–243.

Syed, Z., Stultz, C., & Guttag, J. (2010). Motif discovery in physiological datasets: A methodology for inferring predictive elements. *ACM Transactions in Knowledge Discovery in Data, 4*(1).

Thaler, N. S., Bello, D. T., Randall, C., Goldstein, G., Mayfield, J., & Allen, D. N. (2010). IQ profiles are associated with differences in behavioral functioning following pediatric traumatic brain injury. *Archives of Clinical Neuropsychology, 25*(8), 781-790.

Theodoraki, E., Katsaragakis, S., Koukouvinos, C., & Parpoula, C. (2010). Innovative data mining approaches for outcome prediction of trauma patients. *Journal of Biomedical Science and Engineering, 3*, 791-798 .

Tormos, J. M., Garcia-Molina, A., Garcia-Rudolph, A., & Roig, T. (2009). Information and communications technology in learning development and rehabilitation. *International Journal of Integrated Care, 9*.

Trettin, L. J. (2007). *Executive functions following traumatic brain injury: The impact of depression upon performance.*

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Vapnik, V. (1998). The Support Vector Method of Function Estimation. *Nonlinear Modeling*, 55-85.

Vincent, A. S., Roebuck-Spencer, T. M., & Cernich, A. (2014). Cognitive changes and dementia risk after traumatic brain injury: implications for aging military personnel. *Alzheimer's & Dementia, 10*(3), 174-187.

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes.* Cambridge: Harvard University Press.

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III).* San Antonio, Texas: Psychological Corporation.

Weiss , S. M., & Kulikowski, C. A. (1990). *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems (Machine Learning Series).* Morgan Kaufmann Publishers.

Whalen, S. (1998). Revisiting the problem of match. In N. Colangelo, & S. Assouline (Ed.), *Henry B. and Jocelyn Wallace national research symposium on talent development.*

Whyte, J., & Hart, T. (2003). It's more than a black box; it's a Russian doll: defining rehabilitation treatments. *American Journal of Physical Medicine and, 82*(8), 639-652.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* (T. M. systems, Ed.) San Francisco, California: Morgan Kaufmann Publishers Inc.

Yan, X., Han, J., & Afshar, R. (2003). CloSpan: Mining closed sequential patterns in large datasets. *Proc. 3rd Society for Industrial and Applied Mathematics (SIAM) Intern. Conf. on Data Mining*, (pp. 166–177).

Yuliana, O. Y., Rostianingsih, S., & Budhi, G. S. (2009). Discovering sequential disease patterns in medical databases using FreeSpan mining approach. *International Conference on Advanced Computer Science and Information Systems ICACSIS'09.* Jakarta.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning, 42*(1), 31-60.

Zambelli , F., Pesole, G., & Pavesi, G. (2012). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics, 14*, 225-237.

Zitnay, G. A., Zitnay, K. M., Povlishock, J. T., Hall, E. D., Marion, D. W., Trudel, T., . . . Barth, J. T. (2008). Traumatic brain injury research priorities: the Conemaugh International Brain Injury Symposium. *Journal of Neurotrauma, 25*, 1135-1152.