

Valeria Zazzu · Maria Brigida Ferraro  
Mario R. Guarracino *Editors*

# Mathematical Models in Biology

Bringing Mathematics to Life

 Springer

# Mathematical Models in Biology



Valeria Zazzu • Maria Brigida Ferraro  
Mario R. Guarracino  
Editors

# Mathematical Models in Biology

Bringing Mathematics to Life

 Springer

*Editors*

Valeria Zazzu  
Institute of Genetics and Biophysics  
(IGB) "ABT"  
National Research Council of Italy (CNR)  
Naples, Italy

Maria Brigida Ferraro  
Department of Statistical Sciences  
Sapienza University of Rome  
Rome, Italy

Mario R. Guarracino  
High Performance Computing  
and Networking Institute (ICAR)  
National Research Council of Italy (CNR)  
Naples, Italy

ISBN 978-3-319-23496-0      ISBN 978-3-319-23497-7 (eBook)  
DOI 10.1007/978-3-319-23497-7

Library of Congress Control Number: 2015956528

Mathematics Subject Classification (2010): 92c99, 92Bxx, 97M10, 90B10, 90B15, 92C42, 92B20

Springer Cham Heidelberg New York Dordrecht London  
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface

Life scientists are not always fully aware of the powerful role that mathematical models have in both answering biological research questions and in making predictions. Scientists have a clear view of the problem; they know the questions; they have identified ways to answer; and they produce the data to be analysed. Novel high throughput technologies are utilized that give rise to an unprecedented quantity of data. However, the data is ‘noisy’, and the answer to each question can be well hidden under terabytes of incomprehensible text files.

It is here that the mathematicians can help: they know ‘how’ to do things; they love the huge, ugly text files; they foresee hundreds of statistics that could be calculated; they want to try all of them because there is always uncertainty. Mathematicians see paths, trends, connections, and correlations. Ultimately the need to identify the beautiful biological mechanisms that are hidden, must come to light. Indeed, mathematicians too, get stuck, lost among protein sticks, bubbles, helices, and sheets.

During the ‘Bringing Maths to Life’ workshop, held in Naples, Italy, October 27–29, 2014, biologists and mathematicians joined forces to address key areas in biology that face demanding mathematical challenges. A list of invited speakers and participants came from leading European universities and the international scientific community; especially computational biologists, mathematicians, and researchers in the life sciences. Interdisciplinary discussions surrounded existing cases in an effort to identify gaps or to share existing solutions. Finding the best mathematical resolution to interpret data from a biological perspective, or—inversely—understanding the biological issue and its real-life constraints from a mathematical viewpoint, required both communities to closely engage. The present volume gathers a number of chapters selected from the most interesting contributions to the workshop.

The workshop had featured three main sessions. ‘**Zoom inside the cell: microscopy images processing**’ had been the topic of the first session. Biological visualization provides the means through which to place genomic and proteomic information in a cellular or tissue context. While existing software enables particular assays for distinct cell types, high throughput image analysis has, to this point, been impractical unless an image analysis expert develops a customized solution, or

unless commercial packages are used with their built-in algorithms for limited sets of cellular features and cell types. There exists a clear need for powerful, flexible tools for high throughput cell image analysis. Computer vision researchers have contributed new algorithms to the project so that their theoretical work can be applied to practical biological problems.

The session on **‘Genetic variability and differential expression: sequence data analysis’** had addressed the recent revolution in DNA sequencing technology brought by the sequencing of an increasing number of genomes. Changes in data quantity and format (large numbers of short reads or pairs of short reads *versus* relatively long reads produced by traditional Sanger sequencing) imply changes of sequence data management, storage, and visualization, and provide a challenge for bioinformatics.

**‘Deciphering complex relationships: networks and interactions’** had dealt with biological systems composed of thousands of different types of components and the problems related to the huge networks that comprise numerous non-linearly interacting dimensions, from which, in turn, biological functions emerge. The networks are far too complex to be understood by the unassisted human mind and therefore to analyze these complex biological systems and to obtain relevant answers, biology requires quantitative models that draw from modern computer science and mathematics.

Additionally, there had been three invited sessions. The first one was on **‘Molecular Dynamics and Modelling of Protein Structure and Function via High Performance Computing Simulations’** (organized by Alessandro Grottesi from CINECA, Italy). Molecular dynamics simulations are computational tools aimed at studying protein structure and dynamics as well as protein-protein interactions at the atomic level. The high performance computing of current computer architectures, as well as the developing of valid force fields for the mathematical modelling of biochemical interactions, have provided new tools to help biologists studying and testing hypotheses to understand biochemical phenomena in a new perspective. This session has highlighted the advantages and limitations of this powerful computational technique.

In the second invited session, **‘Statistical challenges in omics research within Life Sciences’** (organized by J.J. Houwing-Duistermaat from Leiden University Medical Center, The Netherlands and Luciano Milanese from Institute of Biomedical Technologies, CNR, Italy), several statistical issues in omics datasets were addressed, from preprocessing up to building statistical models for joint interpretation of the datasets. These datasets contain information about different aspects of the same biological processes. Therefore in many studies, multiple omics datasets are nowadays available and integrated analyses of these omics datasets is the ultimate goal to understand biological mechanisms underlying traits. However integration of these datasets is not straightforward since they vary in measurement error distributions, scale, sparseness and size. In this session challenges were addressed in single omics datasets analysis as well as combined analysis of multiple omics datasets.

The third invited session had been dedicated to ‘**Artificial neurons and realistic simulation of neuronal functions**’ (organized by Angela Tino from Institute of Cybernetics, CNR, Italy). Modern neuroscience research has generated vast volumes of experimental data, and large scale initiatives launched in recent years will gather much more. Nonetheless, much of the knowledge needed to build multilevel atlases and unifying models of the brain is still missing. Brains are a large network composed of many neurons with their synaptic connections, each expressing different proteins on the cell membrane and each with its own complex internal structure. Despite huge advances, there is no technology that allows us to characterize more than a tiny part of this complexity. The session had shed light on novel solutions from neural-inspired artificial models and software, realistic neuronal function simulation, and functional and molecular neurobiology and had aimed to gather scientists from diverse disciplines to foster integrated approaches to unravel complex brain functions.

Naples, Italy  
Rome, Italy  
Naples, Italy

Valeria Zazzu  
Maria Brigida Ferraro  
Mario R. Guarracino





# Acknowledgments

The workshop has been organized by: Alessandra Rogato (Institute of Biosciences and Bioresources); Valeria Zazzu and Enza Colonna (Institute of Genetics and Biophysics); Mario Guarracino (High Performance Computing and Networking Institute and Institute for Higher Mathematics ‘F. Saveri’) from the Italian National Research Council (CNR), Italy; Maria Brigida Ferraro from Sapienza University of Rome, Italy; Martijn Moné from the VU University Amsterdam and ISBE—Infrastructure for Systems Biology Europe, The Netherlands. Gerardo Toraldo from the Department of Mathematics and Applications ‘Renato Caccioppoli’, University of Naples Federico II, contributed to the organization.

The initiative has been supported by the Italian National Research Council (CNR), the Institute for High Mathematics ‘F. Saveri’ (INDAM), the High Performance Computing and Networking Institute (ICAR), the University of Naples Federico II, Mimomics, Interomics, the Department of Bio-Agriculture Sciences (CNR), the Department of Biomedical Sciences (CNR), LABGTP and Tecnologica.



# Contents

|   |    |
|---|----|
| <b>Image Segmentation, Processing and Analysis in Microscopy and Life Science</b> .....   | 1  |
| Carolina Wählby   |    |
| <b>Image Analysis and Classification for High-Throughput Screening of Embryonic Stem Cells</b> .....  | 17 |
| Laura Casalino, Pasqua D’Ambra, Mario R. Guarracino, Antonio Irpino, Lucia Maddalena, Francesco Maiorano, Gabriella Minchiotti, and Eduardo Jorge Patriarca   |    |
| <b>Exploiting “Mental” Images in Artificial Neural Network Computation</b> .....  | 33 |
| Massimo De Gregorio and Maurizio Giordano   |    |
| <b>Applying Design of Experiments Methodology to PEI Toxicity Assay on Neural Progenitor Cells</b> .....  | 45 |
| Sara Mancinelli, Valeria Zazzu, Andrea Turcato, Giuseppina Lacerra, Filomena Anna Digilio, Anna Mascia, Marta Di Carlo, Anna Maria Cirafici, Antonella Bongiovanni, Gianni Colotti, Annamaria Kisslinger, Antonella Lanati, and Giovanna L. Liguori |    |
| <b>A Design of Experiment Approach to Optimize an Image Analysis Protocol for Drug Screening</b> .....  | 65 |
| Antonella Lanati, Cecilia Poli, Massimo Imberti, Andrea Menegon, and Fabio Grohovaz   |    |
| <b>Computational Modeling of miRNA Biogenesis</b> .....   | 85 |
| Brian Caffrey and Annalisa Marsico  |    |
| <b>Tunicate Neurogenesis: The Case of the <i>SoxB2</i> Missing CNE</b> .....  | 99 |
| Evgeniya Anishchenko and Salvatore D’Aniello  |    |

**MECP2: A Multifunctional Protein Supporting Brain Complexity** ..... 109  
Marcella Vacca, Floriana Della Ragione, Kumar Parijat Tripathi,  
Francesco Scalabrì, and Maurizio D’Esposito

**DNA Barcode Classification Using General Regression Neural  
Network with Different Distance Models** ..... 119  
Massimo La Rosa, Antonino Fiannaca, Riccardo Rizzo,  
and Alfonso Urso

**First Application of a Distance-Based Outlier Approach  
to Detect Highly Differentiated Genomic Regions  
Across Human Populations.** ..... 133  
Stefano Lodi, Fabrizio Angiulli, Stefano Basta, Donata Luiselli,  
Luca Pagani, and Claudio Sartori

**Predicting the Metagenomics Content with Multiple CART Trees** ..... 145  
Dante Trivisani, Diego Galarce, Alejandro Maass,  
and Rodrigo Assar

**A Statistical Approach to Infer 3D Chromatin Structure** ..... 161  
Claudia Caudai, Emanuele Salerno, Monica Zoppè,  
and Anna Tonazzini

**Basic Exploratory Proteins Analysis with Statistical Methods  
Applied on Structural Features** ..... 173  
Eugenio Del Prete, Serena Dotolo, Anna Marabotti, and Angelo  
Facchiano

**Modelling of Protein Surface Using Parallel Heterogeneous  
Architectures** ..... 189  
Daniele D’Agostino, Andrea Clematis, Emanuele Danovaro,  
and Ivan Merelli

# Image Segmentation, Processing and Analysis in Microscopy and Life Science

Carolina Wählby

**Abstract** Microscopes have been used for more than 400 years to understand biological and biomedical processes by visual observation. Science is the art of observing, but science also requires measuring, or quantifying, what is observed. Research based on microscopy image data therefore calls for methods for quantitative, unbiased, and reproducible extraction of meaningful measurements describing what is observed. Digital image processing and analysis is based on mathematical models of the information contained in image data, and allows for automated extraction of quantitative measurements. Automated methods are reproducible and, if applied consistently and accurately across experiments with positive as well as negative controls, also unbiased. Digital image processing is further motivated by the development of scanning microscopes and digital cameras that can capture image data in multiple spatial-, time-, and spectral-dimensions, making visual assessment cumbersome or even impossible due to the complexity and size of the collected data.

The process of analyzing a digital image is usually divided into several steps, where the objects of interest are first identified, or ‘segmented’, followed by extraction of measurements and statistical analysis. This chapter starts from the basics of describing images as matrices of pixel intensities. Emphasis is thereafter put on image segmentation, which is often the most crucial and complicated step. A number of common mathematical models used in digital image processing of microscopy images from biomedical experiments are presented, followed by a brief description of large-scale image-based biomedical screening.

**Keywords** Image cytometry • Fluorescence microscopy • Cell segmentation • Image analysis

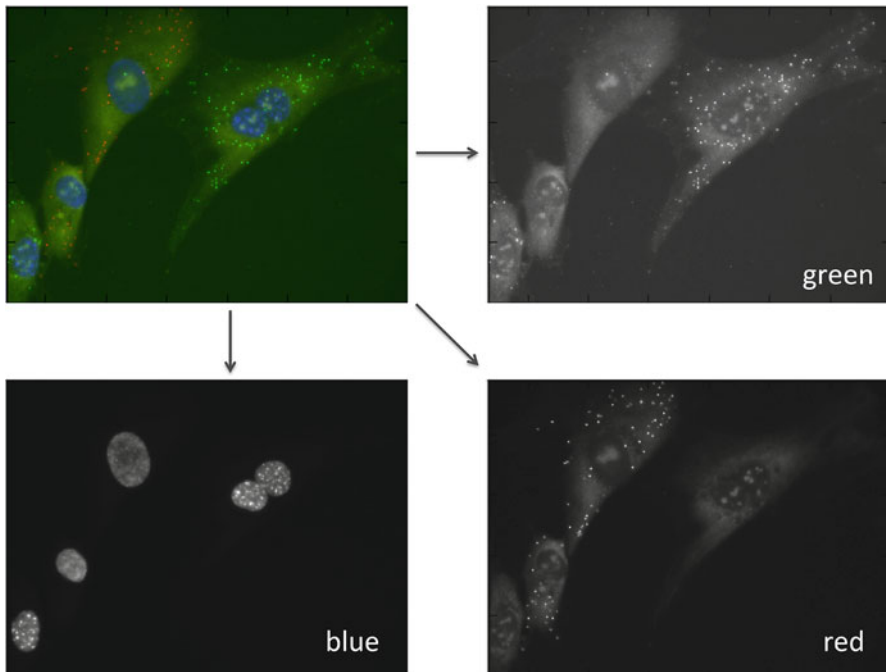
---

C. Wählby (✉)

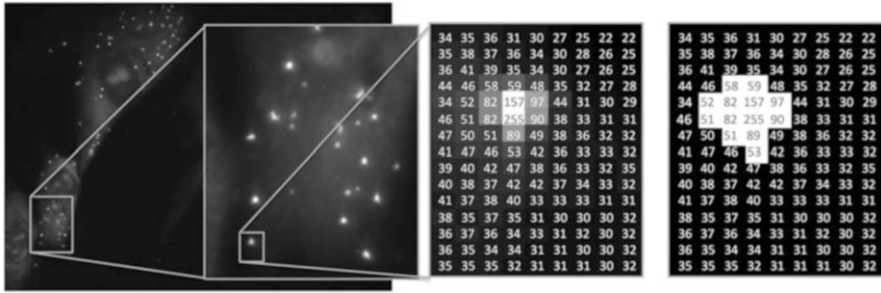
Centre for Image Analysis, Division of Visual Information and Interaction,  
Department of Information Technology, Uppsala University, Uppsala, Sweden  
e-mail: [carolina@cb.uu.se](mailto:carolina@cb.uu.se)

## 1 Pixels and Color Channels

A digital image is not continuous, but consists of discrete picture elements, or pixels. A typical fluorescence microscopy image is built up of multiple fluorescence channels, each representing a separate fluorescence stain, usually bound to DNA or an antibody probing a specific protein or subcellular structure. Figure 1 shows a fluorescence microscopy image where cell nuclei are stained with DAPI binding DNA, and red and green dots representing mRNA molecules (for details see [12]). Imagine that the goal of the analysis is to count the number of red and green dots per cell. The color image in Fig. 1 can be split into its constituent image channels, leading to one image representing the red, green and blue fluorescence respectively. If we take a closer look at the red channel, and zoom in on one of the dots, we can see that the image is built up of square picture elements, or ‘pixels’ for short, see Fig. 2. Each of these pixels is represented as a number in the computer, where a higher number means a brighter pixel, and the whole image can be thought of as a matrix of numbers. In a color image, the three image channels represent the



**Fig. 1** Using three different filter sets, three different fluorescence labels were imaged using fluorescence microscopy. Top left is a composite image of all three image channels; cell nuclei are stained with DAPI binding DNA, and *red* and *green dots* represent mRNA molecules (for details see [12]). Due to autofluorescence and unspecific fluorophore binding, the cells’ cytoplasm can be seen as a weak background staining in the *red* and *green* image channels



**Fig. 2** An image is built up of pixels. If we zoom in on a sub part of the larger image, we see that the image is built up of square pixels, where the graylevel, or intensity of each pixel, can be represented as a number in the computer memory. Image segmentation thresholding (*left*) assigns the maximum value (*white*) to all pixels above a given threshold (in this case intensities  $> 50$ ), while other pixels are assigned the minimum value (*black*)

amount of red, green and blue respectively, and any image analysis operation can work either on a single image channel, or a combination of multiple image channels in two or more spatial dimensions, as well as with time sequences. Here we focus on operations that work on a single channel in two dimensions, but the general idea of images as matrices of pixels, each represented by a number, holds true in any number of dimensions.

## 2 Image Segmentation

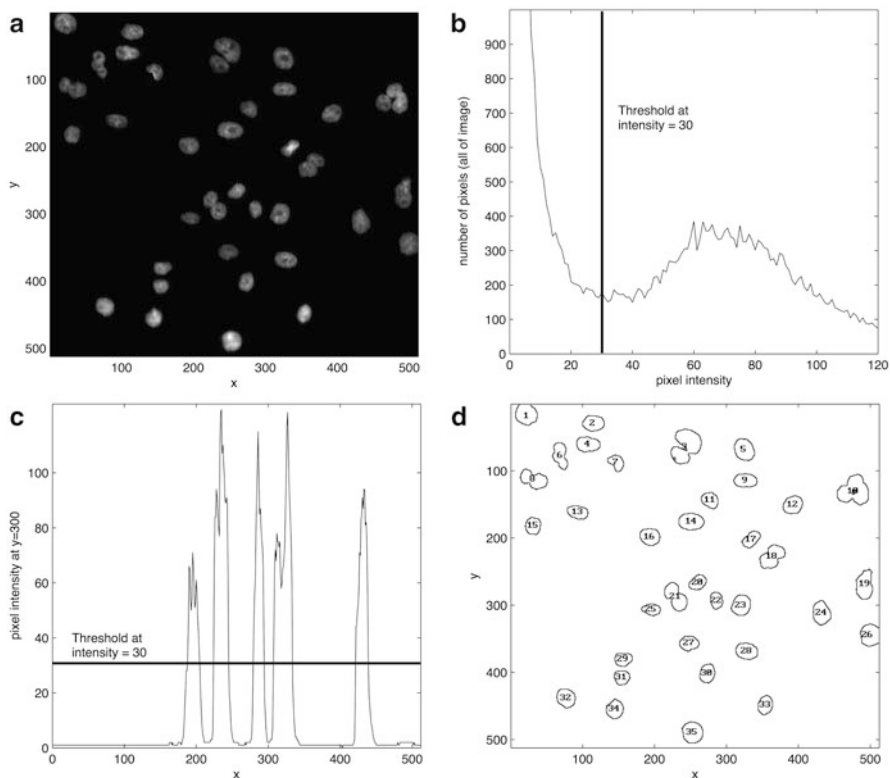
Segmentation is the process in which an image is divided into its constituent objects or parts, and background. It is often the most vital *and* most difficult step in an image analysis task. The segmentation result usually determines eventual success of the analysis. For this reason, many segmentation techniques have been developed, and there exist almost as many segmentation algorithms as there are segmentation problems. The construction of a segmentation algorithm can be thought of as defining a model of the objects that we want to detect in the image. This model is then the basis for the segmentation algorithm.

### 2.1 Thresholding

In the simplest case, we create a model that says that objects are brighter than the image background, and individual objects are well separated from each other. If we can find a suitable intensity threshold that separates the bright objects from the dark background, the segmentation is completed. We simply find all connected pixels brighter than the threshold, and say that they are our objects, as shown



in Fig. 2, left, where all pixels above a given threshold (in this case intensities  $>50$ ) are assigned the maximum value (white) while other pixels are assigned the minimum value (black). The tricky part is to find a suitable threshold. There are many different automated thresholding methods, see [29, 30] for a review. One approach is to look for valleys in the image histogram. Plotting the number of pixels per intensity-level against intensity-level creates an image histogram. If objects are bright and background is dark, the histogram will have one peak for objects and one for background, and a valley will be present between the peaks. Figure 3a shows an image of fluorescence labeled nuclei of cultured cells. The image histogram is shown in Fig. 3b, and a threshold is placed at intensity 30. In Fig. 3c, the intensity variation along a row in (a) is plotted against x-position, and the threshold is shown as a horizontal line. The result of thresholding the image at this level, and labeling the different connected components, is shown in Fig. 3d. Clustered objects will not be separated by simple intensity thresholding.



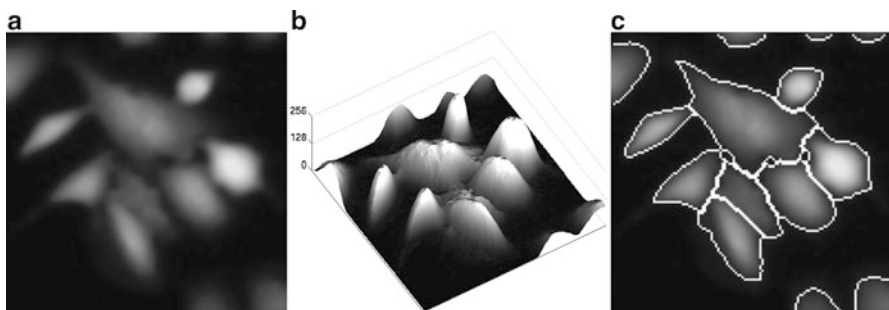
**Fig. 3** Image segmentation by thresholding. (a) Fluorescence stained nuclei of cultured cells. (b) Image histogram of (a). A threshold is placed where the histogram shows a local minimum. The vertical line corresponds to a threshold at intensity 30. (c) An intensity profile along the row  $y = 300$  of (a), with the intensity threshold represented by a horizontal line. (d) The result after thresholding and labeling of connected pixels. Note that not all nuclei are separated by thresholding

## 2.2 Watershed Segmentation

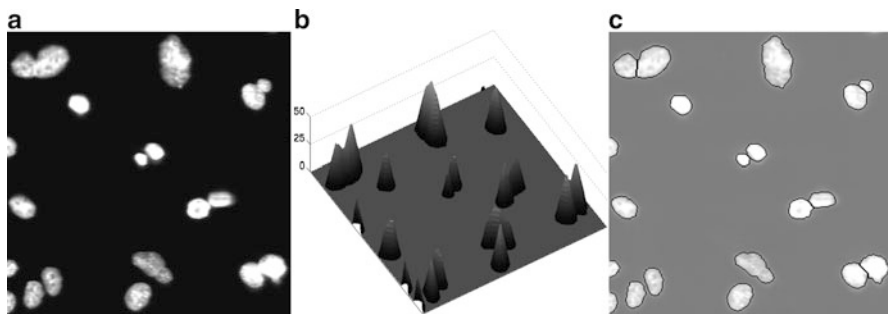
If all the objects are brighter than the image background, but clustered, as in the image of cytoplasm in Fig. 4a, thresholding will only separate the objects from the image background, and not separate the individual objects from each other. There is no single threshold that will separate all cells and at the same time find all cells. We can, however, create a model that says that objects have high intensity, and are less intense at borders towards other objects. If image intensity is thought of as height, the cells can be thought of as mountains separated by valleys in an intensity landscape, see Fig. 4b. The segmentation task is then to find the mountains in the landscape.

A segmentation algorithm that has proven to be very useful for many areas of image segmentation where landscape-like image models can be used is watershed segmentation. The method was originally suggested by Digabel and Lantuéjoul, and extended to a more general framework by Beucher et al. [2]. Watershed segmentation has then been refined and used in many situations; see, e.g., Meyer and Beucher [22] or Vincent [34] for an overview. The watershed algorithm works through intensity layer by intensity layer and splits the image into regions similar to the drainage regions of a landscape. If the intensity of the image is thought of as height of a landscape, watershed segmentation can be described as submerging the image landscape in water, and allowing water to rise from each minimum in the landscape. Each minimum will thus give rise to a catchment basin, and when the water rising from two different catchment basins meet, a watershed, or border, is built in the image landscape. All pixels associated with the same catchment basin are assigned the same label. Watershed segmentation can be implemented with sorted pixel lists [35] so that essentially only one pass through the image is required. This implies that the segmentation can be done very fast.

In the case where we want to find bright mountains separated by less bright valleys, we simply turn the landscape up-side-down, inverting the image, and think



**Fig. 4** Image segmentation by watershed segmentation. (a) Fluorescence stained cytoplasm of cultured cells. (b) The intensity of (a) plotted as a landscape. (c) The result of watershed segmentation of the inverted image



**Fig. 5** Shape-based watershed segmentation. (a) Free and clustered cell nuclei. (b) Distance transformation applied to a thresholded version of (a). The distance from each object pixel to the image background is coded as intensity and displayed as height in a landscape. (c) The result of watershed segmentation of (b) together with (a)

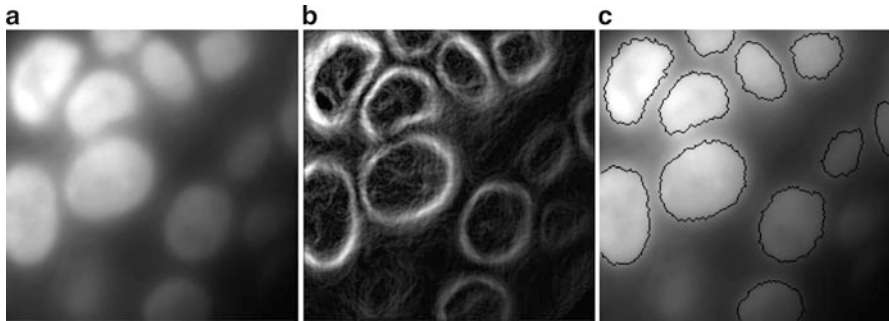
of the mountains as lakes separated by ridges instead of mountains separated by valleys. The result after applying watershed segmentation to the image of the cytoplasm can be seen in Fig. 4c.

### 2.3 *Shape-Based Watershed Segmentation*

If the clustered objects are not separated by less intense borders, they may have some other feature, or combination of features, that can be included in the segmentation model. One example of such a feature is roundness. The cell nuclei in Fig. 5a are all fairly round in shape, but have internal intensity variations that are sometimes greater than those between the individual nuclei. The clustered nuclei can easily be separated from the background using thresholding. The thresholded image can then be transformed into a distance image, where the intensity of each object pixel corresponds to the distance to the nearest background pixel. Calculation of the distance transformation is very fast, calculated by two passes through the image [3, 4]. The result will be an image showing bright cones, each corresponding to a round object, see Fig. 5b. Watershed segmentation can then be applied to the inverted distance image, and the clustered objects are separated based on roundness, see result in Fig. 5c. Shape-based segmentation has proven useful for segmentation of cell nuclei in a number of studies [13, 20, 24, 27].

### 2.4 *Edge-Based Watershed Segmentation*

Intensity variations in the image background often make it difficult to separate the objects from the image background using thresholding. In some cases, it is possible



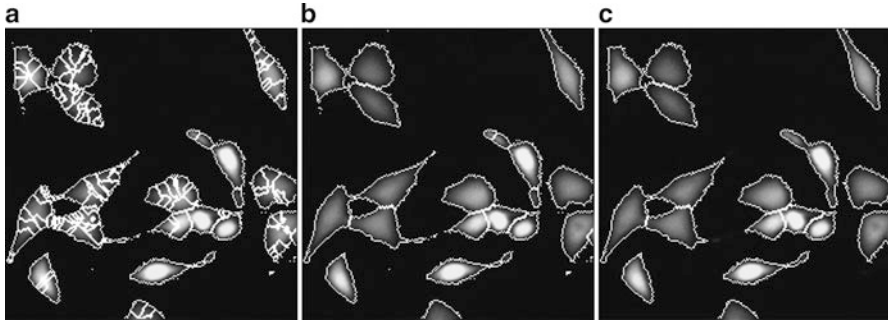
**Fig. 6** Edge-based watershed segmentation. (a) Fluorescence stained cell nuclei in a section from tumor tissue. Due to background variation, separation of nuclei and background by thresholding is not possible. (b) The gradient magnitude of (a). (c) Result after applying watershed segmentation to the gradient magnitude and overlaying the result with the original image

to reduce these background variations by pre-processing steps that computationally reduce these variations [15, 26]. In other cases, a more advanced model saying that the transition between objects and background is marked by a fast change in image intensity may be applied. In Fig. 6a, the background pixels in the upper left corner of the image have the same intensity as the object pixels in the lower right corner of the image. The objects are still visually clearly detectable as their local intensity is different from the local background.

Intensity changes can be described as the magnitude of the image gradient. The magnitude of the gradient expresses the local contrast in the image, i.e., sharp edges have a large gradient magnitude, while more uniform areas in the image have a gradient magnitude close to zero. The local maximum of the gradient amplitude marks the position of the strongest edge between object and background. The commonly used Sobel operators [32] are a set of linear filters for approximating gradients in the  $x$ ,  $y$  (and  $z$ ) directions of an image. Adding the absolute values of the convolutions of the image with the different Sobel operators approximates the gradient magnitude image. Figure 6b shows the gradient magnitude, where large magnitude is shown as high image intensity. If watershed segmentation is applied to the gradient magnitude image, the water will rise and meet at the highest points of the ridges, as shown in Fig. 6c. This corresponds to the location of the fastest change in intensity, just as in our segmentation model.

## 2.5 Merging

When watershed segmentation is applied to an image, water will rise from every minimum in the image, i.e., a unique label will be given to each image minimum. In many cases, not all image minima are relevant. Only the larger intensity variations mark relevant borders of objects. This means that applying watershed segmentation



**Fig. 7** Edge-based merging. (a) Fluorescence labeled cytoplasm with internal intensity variations leading to over-segmentation. (b) Result after merging on minimum height of separating ridge. Some over-segmentation still remains. (c) Result after further merging of all small objects with the neighbor towards which it has its weakest ridge

will lead to over-segmentation, i.e., objects in the image will be divided into several parts, see Fig. 7a. Over-segmentation can be reduced by a pre-processing step reducing the number of local image minima, e.g., by smoothing the image with a mean or median filter. Smoothing may, however, remove important structures, such as edges, in the image. An alternative to pre-processing is post-processing. After applying watershed segmentation, over-segmented objects can be merged.

Merging can be performed according to different rules, based on the segmentation model. One example is merging based on the height of the ridge separating two catchment basins, as compared to the depth of the catchment basins. The model says that a true separating ridge must have a height greater than a given threshold. All pairs of lakes that at some point along their separating ridge have a height lower than the threshold are merged. The result of merging Fig. 7a at height 10 is shown in Fig. 7b.

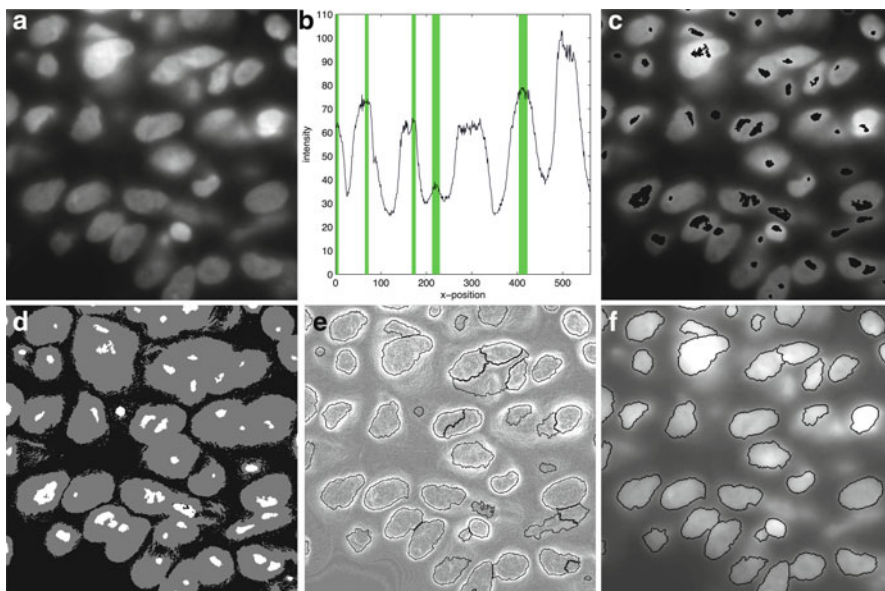
Other merging criteria may also be used. For example, if we know that an object must have a certain size, we can include this in our model and say that every object smaller than this size should be merged with one of its neighbors. If there are several neighbors to choose from, we say that merging should be with the neighbor towards which the small object has, e.g., its weakest ridge [37]. The result of this merging method applied to Fig. 7b is shown in (c). The length of the border between two objects can also be used to decide if neighboring objects should be merged or not [33]. Defining the strength of a border as the weakest point along the border may lead to merging of many correctly segmented objects due to single weak border pixels or weak border parts originating from locally less steep gradients. Another simple measure, which is less sensitive to noise and local variations, is the mean value of all pixels along the object border [38].

## 2.6 Seeded Watershed Segmentation

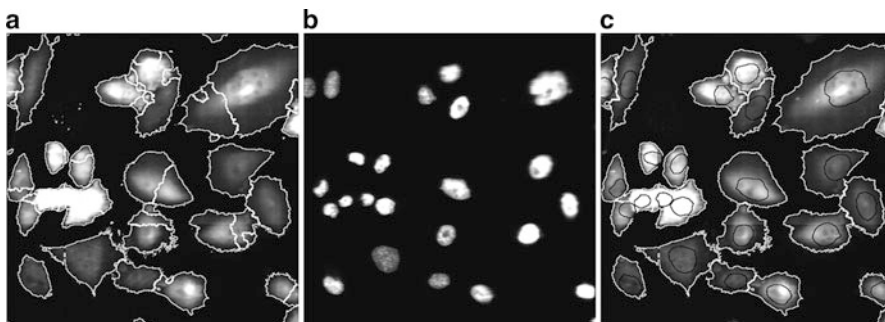
Including a priori information in our model before applying watershed segmentation can also reduce both over- and under-segmentation. Seeded watershed segmentation [1, 22, 34] means that starting regions, called seeds, are given as input to the watershed segmentation. Water is then only allowed to rise from these seeded regions, and the water rising from the seeds floods all other image minima. The water will continue to rise until the water rising from one seeded region meets the water rising from another seeded region, or a pre-defined object/background threshold. This means that we will always end up with exactly as many regions as we had input seeds.

Seeds can be set manually [19], or in an automated way. For example, we may know that, despite variations in both object and background, each object has a certain contrast compared to its local neighborhood. Such regions can be detected using morphological filters. One example is the extended h-maxima transform, which filters out the relevant maxima using a contrast criterion [31]. All maxima are compared to their local neighborhood, and only those maxima greater than a given threshold  $h$  are kept. A low  $h$  will result in many seeds, often more than one seed per object. A high  $h$  will result in fewer seeds, and some objects may not get a seed at all. An example is shown in Fig. 8. The intensity along a pixel row in Fig. 8a is shown in (b), and the h-maxima are marked in gray. Note that maxima that do not contain gray markers do so in a different image row. Despite background variation, h-maxima are found in all cells, as shown in (c). Seeded watershed segmentation is very useful if we perform our segmentation in the gradient magnitude of the image. We can find seeds in object and background regions based on intensity information in the original image, and then let the water rise from these seeds placed in the gradient magnitude image. Object and background seeds are shown in Fig. 8d and the result after watershed segmentation is shown in (e). The result of this segmentation approach, combined with merging based on edge-strength is shown in Fig. 8f.

Seeds may also come from a parallel image. Cells often vary very much in shape and size, and touch each other. Watershed segmentation will not always give a satisfactory result, as seen in Fig. 9a. If we have a single seed per cell, the task of finding the borders of the cell is greatly simplified. The h-maxima transformation results in useless seeds due to great intensity variations within the cells. The nice thing about cells is, however, that each cytoplasm has a natural marker that may be included in the segmentation model: the nucleus. If the nuclei, which are fairly round in shape and usually nicely separated, are stained and imaged in parallel with the cells, they can be used as seeds for watershed segmentation of the cells. The nuclei of the cells in Fig. 9a are shown in (b), and the result of seeded watershed segmentation using the nuclei as seeds is shown in (c).



**Fig. 8** Seeded watershed segmentation. **(a)** Fluorescence stained cell nuclei in tumor tissue. Due to background variation, separation of nuclei and background by thresholding is not possible. **(b)** Intensity profile across one row of pixels of **(a)**, and h-maxima at  $h = 5$  shown as vertical bars in gray. Nuclei without h-maxima have h-maxima in a different row. **(c)** The original image with the h-maxima overlaid. **(d)** Object seeds found by h-minima transformation of the gradient magnitude image of **(c)** followed by removal of small objects (black). **(e)** Result after seeded watershed segmentation of the gradient magnitude image. More than one seed per object leads to over-segmentation. **(f)** Merging on edge strength reduces over-segmentation. Also poorly focused objects may be removed by this step



**Fig. 9** Cell segmentation using nuclei as seeds. **(a)** Clustered fluorescence labeled cells with varying shapes and intensities are difficult to separate from each other. Watershed segmentation will result in both over- and under-segmentation (white lines). **(b)** A parallel image showing the cell nuclei can be used as a seed for watershed segmentation of the cells. **(c)** The result of watershed segmentation (white lines) using the nuclei (black lines) as seeds

## ***2.7 Extension to Volume Images and Time-Lapse Experiments***

Most of the discussed methods can be extended to volume (three dimensional) images. For most methods, the only difference is that instead of working with two-dimensional pixel neighborhoods, we work with three-dimensional voxel neighborhoods. For example, seeded watershed segmentation can be applied to three-dimensional images of fluorescence stained cell nuclei in tumor samples [33, 34]. Here, the 26 side-, edge- and corner- neighbors surrounding each pixel (or voxel) in 3D are considered in each step, starting with finding h-maxima to the final merging of weak edges.

In time-lapse experiments, nuclear stains are often undesirable as they may interfere with the natural behavior of living cells. A nuclear stain may, however, still be used for image segmentation if the cells are fairly stationary. The stain is simply added after the completed experiment, and the same image of the nuclei is thereafter used for segmentation of all time-lapse images [16]. If the cells move, an image of the nuclei can be used as a starting point for backtracking of cell motion.

## ***2.8 Other Segmentation Methods***

Many other models for cell segmentation exist, such as iteratively refined active shape models [11] and snake algorithms [25]. A comprehensive review of cell segmentation approaches can be found in Meijering [21]. However, the approaches described here (and summarized in [38]) are used in a wide range of commercially available software for microscopy image analysis, and they are also available in a range of free and open source software, as reviewed by Eliceiri et al. [8]. Free and open source solutions make it easy to share methods between labs, and are valuable for the reproducibility of research. The methods described here are fast and allow large scale screening studies.

## **3 Feature Extraction and Classification in Large-Scale Image Based Biomedical Screening**

Once the objects of interest have been segmented from each other and from the image background, a large number of descriptive features can be extracted from the individual objects in the image, see [28] for an overview. Features may include object size, shape, distribution of sub-structures such as membranes and signals from specific molecular detection methods, local intensity patterns (texture), number of protrusions, number of nearby neighboring cells etc. All features that we can extract are based on the actual pixel values and their spatial arrangements within the object. Features may also include relationships between objects (such as number

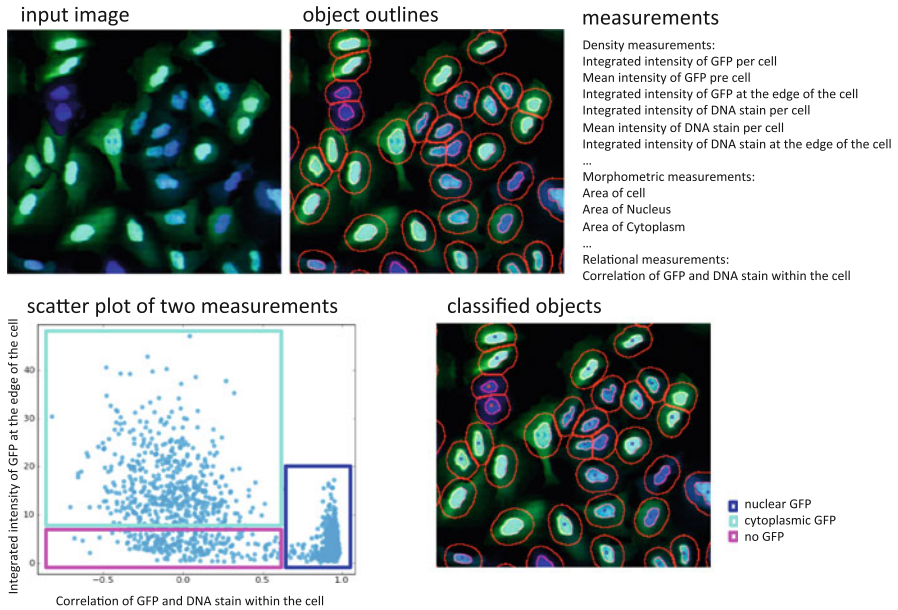


of neighbors within a fixed distance), or relationships between larger objects (e.g. cells) and sub-objects (e.g. organelles).

Morphometric, or shape features are features that are based solely on the spatial arrangements of pixels/voxels and include for example area, which is the number of pixels belonging to the object. Perimeter is another morphometric feature, defined as the sum of steps taken when walking around the edge pixels of a 2D object [6, 14], and compactness index is a measure of how compact the object is, described as the object perimeter squared divided by the area. Densitometric, or intensity features are features that describe the gray-level values (without considering the spatial distribution), and include for example the integrated intensity, which is the sum of the intensity values of all pixels/voxels belonging to the object, and the mean object intensity. Another group of features are the more complex textural or structural features that combine spatial and gray-level information. Examples are the gray-level co-occurrence measurement [23] and mass displacement, which is the distance between the center of mass given by the gray-level image of the object and the center of mass given by a binary mask of the object.

The numerical data produced by feature extraction may not always be the desired end result, and it may be difficult to interpret. A simple example is if the goal of the analysis is to decide the percentage of small, medium-sized and large objects in an image. Numerical data representing the area of each object does not provide the final answer. The numerical data has to be analyzed, and each object has to be classified as small, medium, or large in order to calculate the desired percentages. In many cases, the goal of the analysis is to retrieve more complicated information from the images, and a single feature like area is not sufficient for object description and classification. Figure 10 illustrates an example where cells are segmented, features are extracted, and cells are classified into three different classes based on the two most discriminatory feature measurements. If the phenotypes of interest are complex, a larger number of features may be needed, and the different classes have to be separated by multivariate statistic analysis. It is very important to note that increasing the number of features for object classification will not necessarily improve the classification result. It has been observed that, beyond a certain point, inclusion of additional features leads to worse rather than better performance [7]. One should instead try to pick a limited set of features that can discriminate between the relevant populations as well as possible, or use automated feature selection methods. In fact, a very efficient way of automated selection and reduction of features is by iterative feedback and machine learning [10].

Image processing and analysis is often used in high-content analysis/high-throughput screening (HCA/HTS) experiments, searching large libraries of chemical or genetic perturbants, to find new treatments for a disease or to better understand disease pathways. Automated image segmentation, processing and analysis has also more recently shown to be a powerful tool for grouping chemical compounds based on their mechanism of action [17] and to predict the potential performance of novel chemicals in compound libraries [36]. Large-scale experiments analyzed by automated methods require robust models for object detection, such as the ones described here. Robust and diverse staining approaches also increase the chance of



**Fig. 10** Illustration of object classification based on feature measurement. The image comes from a translocation assay (image set BBBC013v1 provided by Ilya Ravkin, available from the Broad Bioimage Benchmark Collection, [18]) where one can observe a cytoplasm to nucleus translocation of the Forkhead (FKHR-EGFP) fusion protein in stably transfected human osteosarcoma cells, U2OS. In proliferating cells, FKHR is localized in the cytoplasm. Even without stimulation, Forkhead is constantly moving into the nucleus, but is transported out again by export proteins. Upon inhibition of nuclear export, FKHR accumulates in the nucleus. In this assay, export is inhibited by blocking PI3 kinase/PKB signaling by incubating cells for 1 h with Wortmannin. Nuclei are stained with DRAQ, a DNA stain. The goal is to classify cells into three phenotypes; cells with nuclear GFP (i.e. FKHR-EGFP), cytoplasmic GFP, or no GFP expression. Input images are segmented into cell nuclei and surrounding cytoplasm, where the cytoplasm is defined as any pixels within a fixed distance of a nucleus. Next, a large number of different feature measurements are extracted from each cell. In the classification step, cells are classified into the three classes nuclear GFP (*blue box*), cytoplasmic GFP (*green box*), or no GFP expression (*magenta box*) based on two feature measurements, namely integrated intensity of GFP at the edge of the cell and correlation of GFP and DNA stain within the cell. The distribution of feature measurements is shown as a scatter plot, and the final classification result as small colored squares overlaid the segmented cells

detecting subtle changes of cellular states [9]. Different methods for quantification of image quality are also desirable, especially if the number of images is very large, and system failures such as errors in autofocusing, image saturation and debris may introduce errors in the final screening results [5].

In the most common case, cultured cells model biological processes and disease pathways. Studying disease by culturing cells allows for efficient analysis and exploration. However, many diseases and biological pathways can be better studied in whole animals—particularly diseases that involve organ systems and multicellular

interactions, such as metabolism and infection. The worm *Caenorhabditis elegans* (*C. elegans*) is a well-established and effective model organism, used by thousands of researchers worldwide to study complex biological processes. Samples of *C. elegans* can be robotically prepared and imaged by high-throughput microscopy, just as with cells, mathematical models of worm shape and appearance are required for efficient analysis [39, 40].

Automated analysis of cells as well as model organisms are typical examples where biological questions and their real-life constraints together with the possibilities and limitations of mathematical models require expertise from biologists as well as mathematicians. Image based research as such requires different scientific communities to closely engage and communicate, and there is broad potential for new discoveries in this fast growing field of science.

## References

1. Beucher, S.: The watershed transformation applied to image segmentation. *Scanning Microsc.* **6**, 299–314 (1992)
2. Beucher, S., Lantuejoul, C.: Use of watersheds in contour detection. In: *International Workshop on Image Processing: Real-Time and Motion Detection/Estimation*, Rennes (1979)
3. Borgfors, G.: Distance transformations in digital images. *Comput. Vis. Graphics Image Process.* **34**, 344–371 (1986)
4. Borgfors, G.: On digital distance transforms in three dimensions. *Comput. Vis. Image Underst.* **44**(3), 368–376 (1996)
5. Bray, M.A., Fraser, A.N., Hasaka, T.P., Carpenter, A.E.: Workflow and metrics for image quality control in large-scale high-content screens. *J. Biomol. Screen.* **17**(2), 266–274 (2012)
6. Dorst, L., Smeulders, A.W.M.: Length estimators for digitized contours. *Comput. Vis. Graph. Image Process.* **40**, 311–333 (1987)
7. Duda, R.O., Hart, P.E.: *Pattern Classification and Scene Analysis*. Wiley, New York (1973)
8. Eliceiri, K.W., Berthold, M.R., Goldberg, I.G., Ibáñez, L., Manjunath, B.S., Martone, M.E., Murphy, R.F., Peng, H., Plant, A.L., Roysam, B., Stuurman, N., Swedlow, J.R., Tomancak, P., Carpenter, A.E.: Biological imaging software tools. *Nat. Methods* **9**(7), 697–710 (2012)
9. Gustafsdottir, S.M., Ljosa, V., Sokolnicki, K.L., Anthony Wilson, J., Walpita, D., Kemp, M.M., Petri Seiler, K., Carrel, H.A., Golub, T.R., Schreiber, S.L., Clemons, P.A., Carpenter, A.E., Shamji, A.F.: Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* **8**(12), e80999 (2013)
10. Jones, T.R., Carpenter, A.E., Lamprecht, M.R., Moffat, J., Silver, S.J., Grenier, J.K., Castoreno, A.B., Eggert, U.S., Root, D.E., Golland, P., Sabatini, D.M.: Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **106**(6), 1826–1831 (2009)
11. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. *Int. J. Comput. Vis.* **1**, 321–331 (1988)
12. Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., Nilsson, M.: In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013)
13. Krtolica, A., Ortiz de Solorzano, C., Lockett, S., Campisi, J.: Quantification of epithelial cells in coculture with fibroblasts by fluorescence image analysis. *Cytometry* **49**, 73–82 (2002)
14. Kulpa, Z.: Area and perimeter measurement of blobs in discrete binary pictures. *Comput. Graph. Image Process.* **6**:434–454 (1977)

15. Likar, B., Maintz, J.B., Viergever, M.A., Pernus, F.: Retrospective shading correction based on entropy minimization. *J. Microsc.* **197**(3), 285–295 (2000)
16. Lindblad, J., Wählby, C., Bengtsson, E., Zaltsman, A.: Image analysis for automatic segmentation of cells and classification of Rac1 activation. *Cytometry A*. **57**(1), 22–33 (2004)
17. Ljosa, V., Caie, P.D., Ter Horst, R., Sokolnicki, K.L., Jenkins, E.L., Daya, S., Roberts, M.E., Jones, T.R., Singh, S., Genovesio, A., Clemons, P.A., Carragher, N.O., Carpenter, A.E.: Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* **18**(10), 1321–1329 (2013)
18. Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. *Nat. Methods* **9**(7), 637 (2012)
19. Lockett, S.J., Sudar, D., Thompson, C.T., Pinkel, D., Gray, J.W.: Efficient, interactive, and three-dimensional segmentation of cell nuclei in thick tissue sections. *Cytometry* **31**, 275–286 (1998)
20. Malpica, N., Ortiz de Solorzano, C., Vaquero, J.J., Santos, A., Vallcorba, I., Garcia-Sagredo, J.M., del Pozo, F.: Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry* **28**(4), 289–297 (1997)
21. Meijering, E.: Cell segmentation: 50 years down the road. *IEEE Signal Process. Mag.* **29**, 140–145 (2012)
22. Meyer, F., Beucher, S.: Morphological segmentation. *J. Vis. Commun. Image Represent.* **1**(1), 21–46 (1990)
23. Nielsen, B., Albrechtsen, F., Danielsen, H.E.: Statistical nuclear texture analysis in cancer research: a review of methods and applications. *Crit. Rev. Oncog.* **14**, 89–164 (2008)
24. Ortiz de Solorzano, C., Garcia Rodriguez, E., Jones, A., Pinkel, D., Gray, J., Sudar, D., Lockett, S.: Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *J. Microsc.* **193**, 212–226 (1999)
25. Park, J., Keller, J.M.: Snakes on the watershed. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(10):1201–1205 (2001)
26. Piccinini, F., Lucarelli, E., Gherardi, A., Bevilacqua, A.: Multi-image based method to correct vignetting effect in light microscopy images. *J. Microsc.* **248**(1), 6–22 (2012)
27. Ranefall, P., Wester, K., Bengtsson, E.: Automatic quantification of immunohistochemically stained cell nuclei using unsupervised image analysis. *Anal. Cell. Pathol.* **16**, 29–43 (1998)
28. Rodenacker, K., Bengtsson, E.: A feature set for cytometry on digitized microscopic images. *Anal. Cell. Pathol.* **25**, 1–36 (2003)
29. Sahoo, P.K., Soltani, S., Wong, A.K.C., Chen, Y.C.: A survey of thresholding techniques. *Comput. Vis. Graph. Image Process.* **41**, 233–260 (1988)
30. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. *J. Electron. Imaging* **13**(1), 146–165 (2004)
31. Soille, P.: *Morphological Image Analysis: Principles and Applications*. Springer-Verlag, Berlin Heidelberg (1999)
32. Sonka, M., Hlavac, V., Boyle, R.: *Image Processing Analysis and Machine Vision*, 2nd edn. Brooks/Cole Publishing Company, Pacific Grove (1999)
33. Umesh Adiga, P.S., Chaudhuri, B.B.: An efficient method based on watershed and rule-based merging for segmentation of 3-D histo-pathological images. *Pattern Recogn.* **34**, 1449–1458 (2001)
34. Vincent, L.: Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *IEEE Trans. Image Process.* **2**(2), 176–201 (1993)
35. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(6), 583–597 (1991)
36. Wawer, M.J., Li, K., Gustafsdottir, S.M., Ljosa, V., Bodycombe, N.E., Marton, M.A., Sokolnicki, K.L., Bray, M.A., Kemp, M.M., Winchester, E., Taylor, B., Grant, G.B., Hon, C.S., Duvall, J.R., Wilson, J.A., Bittker, J.A., Dančák, V., Narayan, R., Subramanian, A., Winckler, W., Golub, T.R., Carpenter, A.E., Shamji, A.F., Schreiber, S.L., Clemons, P.A.: Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci. U. S. A.* **111**(30), 10911–10916 (2014)

37. Wählby, C., Lindblad, J., Vondrus, M., Bengtsson, E., Björkesten, L.: Algorithms for cytoplasm segmentation of fluorescence labeled cells. *Anal. Cell. Pathol.* **24**(2–3), 101–111 (2002)
38. Wählby, C: Algorithms for applied digital image cytometry. PhD thesis Uppsala University, Sweden (2003)
39. Wählby, C., Sintorn, I.-M., Erlandsson, F., Borgefors, G., Bengtsson, E.: Combining intensity, edge, and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. *J. Microsc.* **215**(1), 67–76 (2004)
40. Wählby, C., Kamensky, L., Liu, Z.H., Riklin-Raviv, T., Conery, A.L., O'Rourke, E.J., Sokolnicki, K.L., Visvikis, O., Ljosa, V., Irazoqui, J.E., Golland, P., Ruvkun, G., Ausubel, F.M., Carpenter, A.E.: An image analysis toolbox for high-throughput *C. elegans* assays. *Nat. Methods* **9**(7), 714–716 (2012)
41. Wählby, C., Conery, A.L., Bray, M.A., Kamensky, L., Larkins-Ford, J., Sokolnicki, K.L., Veneskey, M., Michaels, K., Carpenter, A.E., O'Rourke, E.J.: High- and low-throughput scoring of fat mass and body fat distribution in *C. elegans*. *Methods* **68**(3), 492–499 (2014)

# Image Analysis and Classification for High-Throughput Screening of Embryonic Stem Cells

Laura Casalino, Pasqua D'Ambra, Mario R. Guarracino, Antonio Irpino, Lucia Maddalena, Francesco Maiorano, Gabriella Minchiotti, and Eduardo Jorge Patriarca

**Abstract** Embryonic Stem Cells (ESCs) are of great interest for providing a resource to generate useful cell types for transplantation or novel therapeutic studies. However, molecular events controlling the unique ability of ESCs to self-renew as pluripotent cells or to differentiate producing somatic progeny have not been fully elucidated yet. In this context, the Colony Forming (CF) assay provides a simple, reliable, broadly applicable, and highly specific functional assay for quantifying undifferentiated pluripotent mouse ESCs (mESCs) with self-renewal potential. In this paper, we discuss first results obtained by developing and using automatic software tools, interfacing image processing modules with machine learning algorithms, for morphological analysis and classification of digital images of mESC colonies grown under standardized assay conditions. We believe that the combined use of CF assay and the software tool should enhance future elucidation of the mechanisms that regulate mESCs propagation, metastability, and early differentiation.

**Keywords** Classification • Colony assay • Imaging • Segmentation • Stem cells

---

L. Casalino • G. Minchiotti • E. Jorge Patriarca  
Institute of Genetics and Biophysics “A. Buzzati-Traverso”, CNR, Naples, Italy

P. D'Ambra (✉) • M.R. Guarracino • L. Maddalena • F. Maiorano  
Institute for High-Performance Computing and Networking, CNR, Naples, Italy  
e-mail: [pasqua.dambra@na.icar.cnr.it](mailto:pasqua.dambra@na.icar.cnr.it)

A. Irpino  
Department of Political Science “J. Monnet”, Second University of Naples, Caserta, Italy

## 1 Introduction

Application of image analysis and machine learning algorithms and tools to cell biology is a very active research field aimed to provide fast and objective methods for analyzing the large amount of images produced by modern high-throughput screening platforms available in biological research laboratories [3, 20, 22, 25].

In this work we describe first results related to the development of a multi-component software framework devoted to define automated morphological analysis and classification of Embryonic Stem Cells (ESCs) colonies. The colony-forming (CF) assay is widely used for monitoring the quality of ESC cultures as it currently offers the most sensitive and specific method to quantify the frequency of undifferentiated cells present in a culture. Moreover, it provides a reliable tool also for evaluating quantitative changes in pluripotent cell numbers, following manipulations that may affect the self-renewal and differentiation properties of the treated cells. In a clonogenic assay, under specific supportive conditions, pluripotent mouse embryonic stem cells (mESCs) form tridimensional round-shaped (*domed*) colonies. After the exposure to conditions that affect their metastability promoting an Epiblast-like phenotype [4] or induce differentiation, cells loose the ability to grow tridimensionally and form irregular and flattened (*flat*) colonies. The 4-day CF assay, by detecting the ability of mESCs to form domed or flat colonies, allows the composition of test cell populations to be quantified at the single cell level. However, to achieve acceptable statistical accuracy, a high number of cell colonies is required. In this regard, manual counting and classification are tedious, time-consuming, resource-intensive and subjective (operator-dependent). Therefore, the development of a reliable automated colony counter and classifier for such clonogenic assays would reduce time and resources required, while allowing greater statistical accuracy, standardization and reproducibility, thus offering the possibility for greater throughput over extended periods. To this aim, we proposed an experimental software tool which is able to automatically discriminate and quantify domed colonies, raising from undifferentiated self-renewing mESCs, and flat colonies, derived from undifferentiated Epiblast-like or differentiating cells. It is a multi-component framework interfacing different general-purpose software modules, implementing highly accurate algorithms for image pre-processing (spanning from region of interest identification to background removal), segmentation, and feature-based classification. Trained on untreated reference samples as well as on samples treated with reference compounds, the prototypal version of the software has been tested on a dataset of 40 microscopy images of single wells containing cells grown in different conditions. Comparison of the first results from automated versus manual colony segmentation and classification on randomly chosen images proved that the proposed tool is promising to be used as blind tool to support reliable analysis of molecular screening.

The paper is organized as follows. In Sect. 2 we describe the software components, both in terms of functionality and of main models and algorithms. In Sect. 3 we present first results and provide a quantitative performance analysis referring to usual performance metrics for segmentation and classification problems. Conclusions are drawn in Sect. 4.

## 2 Software System Components

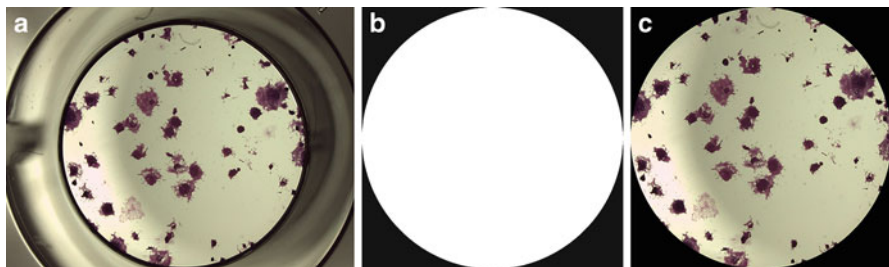
According to the typical image analysis pipeline for High-Content Analysis [22], the proposed steps for automatically discriminating domed and flat colonies in microscopy images can be summarized as follows:

1. **Pre-processing**, to reduce image artifacts caused by imperfections in the image acquisition process.
2. **Segmentation**, to separate the cell colonies in each well image.
3. **Feature computation**, to provide numerical descriptors of each segmented colony.
4. **Classification**, to finally provide the discrimination and quantification of domed and flat colonies, based on the most discriminating features.

### 2.1 Pre-Processing

Input images are initially pre-processed, in order to allow an easier and more accurate segmentation of ESC colonies.

First of all, the well area is extracted by the whole image, in order to focus on the actual region of interest (ROI) in all subsequent steps, as exemplified in Fig. 1. This allows us to reduce not only the computational complexity of the entire procedure, but also the potentially misleading influence of non interesting image details (e.g., shadows and dark areas around the well, as shown in Fig. 1-a).



**Fig. 1** Extraction of the ROI: (a) original image; (b) circular ROI of the well; (c) obtained sub-image, which will be the input for subsequent steps (non interesting pixels shown in *black*)



Relying on the strong intensity discontinuities (edges) provided by the well borders, the circular ROI is accurately obtained through the Hough Transform [10], a powerful tool for the detection of parametric curves in images. It implements a voting process that maps image edge points into manifolds in an appropriately defined parameter space; peaks in this space correspond to the parameters of detected curves. Specifically, the Circle Hough Transform is designed to determine the parameters of a circle when a number of points that fall on its perimeter are known. A circle with radius  $r$  and center  $(x_0, y_0)$  can be described with the parametric equations

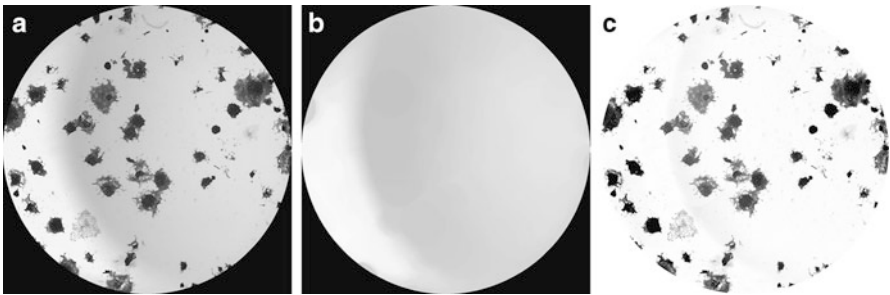
$$\begin{cases} x = x_0 + r \cos \theta \\ y = y_0 + r \sin \theta \end{cases}$$

When the angle  $\theta$  sweeps through the full 360 degree range, the points  $(x, y)$  trace the perimeter of the circle. An image edge point  $(x, y)$  is mapped to the 3D parameter space  $(x_0, y_0, r)$ , voting for all the circles that it could lie on. These votes are accumulated in a 3D array, whose maxima provide the parameters of most prominent circles. In the case of our single-well images, the accumulator maximum provides the well center pixel coordinates and the radius.

The next pre-processing step estimates and then removes the well background, in order to avoid the influence of eventual uneven illumination, as exemplified in Fig. 2. The goal is achieved through mathematical morphology operations [21], that, based on set theory, provide a tool to extract image components useful for the representation and description of region shape, and for pre- and post-processing of images.

Specifically, we relied on closing top-hat filtering, an operation that extracts small elements and details from a given image. Let  $f$  be a grayscale image, and let  $b$  be a grayscale structuring element. The closing top-hat transform of  $f$  (sometimes called the bottom-hat transform or the black top-hat transform) is given by:

$$T_b(f) = f \bullet b - f,$$



**Fig. 2** Background estimation and removal: (a) input image (green color band of the image in Fig. 1-c); (b) estimated background; (c) obtained image, with background removed

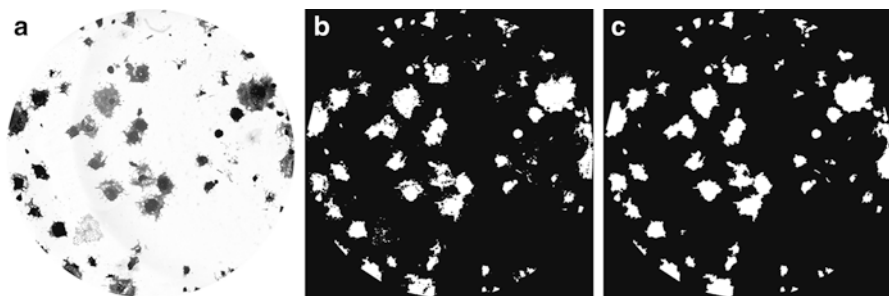
where  $\bullet$  indicates the closing operation. It returns an image containing the objects that are smaller than the structuring element and are darker than their surroundings. In the case of our images, morphological closing  $f \bullet b$  using a structuring element  $b$  having size greater than any expected cell colony removes all the colonies, thus providing a faithful approximation of the well background (see Fig. 2-b). The subsequent subtraction of the original image by the estimated background provides a uniformly illuminated image of the cell colonies, as shown in the normalized result reported in Fig. 2-c.

## 2.2 Segmentation

In order to partition each well image into its constituent objects (ESC colonies and background), we devised two different approaches. In the first approach, the image resulting from pre-processing is binarized through Otsu's method [17] (see Fig. 3-b) and then refined through binary morphological operations. Specifically, refinement involves morphological closing for removing small holes, hole-filling for removing internal holes, and removal of very small objects, i.e., small cell colonies or generic particles of no biological interest (see Fig. 3-c).

A second approach is based on a well-known variational model, for which we recently proposed efficient numerical solvers [9] with the final aim to develop software modules for modern high-performance environments [8]. In the following we briefly describe the model and the main features of the numerical approach and discuss the subsequent refinement of variational-based segmentation for application to the cell identification problem.

The image segmentation problem can be mathematically formulated in terms of a variational model, i.e., in terms of an energy minimization criterion. We look for a piecewise smooth function  $u$  which approximates the original image function  $f$ , with  $u$  discontinuous across a closed set  $K$  that is included in the image domain  $\Omega$  and represented by a suitable function  $z$ . In more details, let  $\Omega \subset \mathfrak{R}^2$  be a bounded open



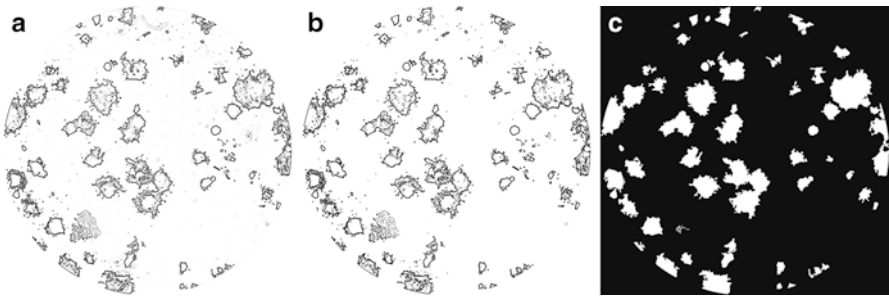
**Fig. 3** First approach to ESC colonies segmentation: (a) input image (the pre-processed image of Fig. 2-c); (b) binarized image; (c) refined binarized image

set and  $f \in L^\infty(\Omega)$  the observed gray-level image. The problem can be described in terms of the minimization of the following functional:

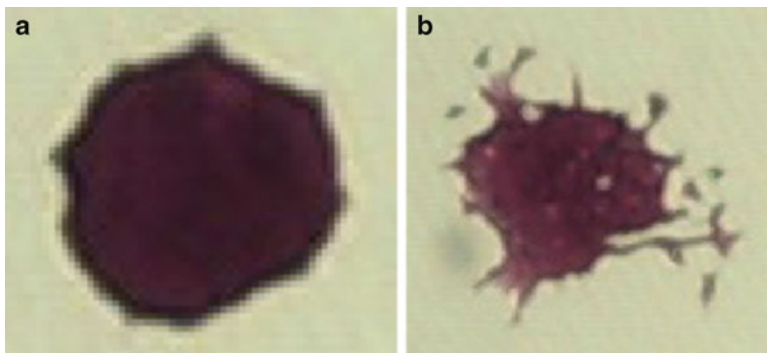
$$E_\epsilon(u, z) = \int_{\Omega} (u - f)^2 dx dy + \beta \int_{\Omega} z^2 |\nabla u|^2 dx dy + \alpha \int_{\Omega} \left( \epsilon |\nabla z|^2 + \frac{(z - 1)^2}{4\epsilon} \right) dx dy,$$

where  $u \in C^1(\Omega \setminus K)$ ,  $0 \leq z \leq 1$  is a function which controls  $|\nabla u|$  and gives an approximate representation of the set  $K$ ,  $\alpha$  and  $\beta$  are positive coefficients and  $\epsilon$  is a positive sufficiently small parameter. The choice of the parameter set, especially of  $\epsilon$ , affects the accuracy in the detection of  $K$  and it is a main drawback in the numerical solution of **1**. Our approach for the solution of **1**, known as the Ambrosio-Tortorelli model or phase-field approximation of the Mumford-Shah model [1], is based on the classical approach of Calculus of Variations which consists in writing the corresponding Euler-Lagrange equations, representing the necessary conditions for the minimizer. Euler-Lagrange equations for **1** are a non-linear system of elliptic partial differential equations coupled with Neumann conditions, which can be solved by finite-difference discretization. We proposed to apply a second-order finite-difference scheme and, starting from a block form of the resulting algebraic equations, we applied a first-order alternating minimization scheme based on the non-linear Gauss-Seidel method, accelerated by inner linear iterations. Our numerical results and comparisons with standard numerical techniques show the efficiency and the robustness of our solution approach for increasing sizes of cell colony images. Details on the numerical scheme and on the above results are discussed in [9]. Here, we only point out that results discussed in the following have been obtained, starting from the preprocessed image obtained by the green band of the original RGB image, when the parameters in **1** were set as follows:  $\alpha = 1$ ,  $\beta = 2$  and  $\epsilon = 10^{-3}$ .

The obtained function  $z$  (see Fig. 4-a), also known as the edge set of the image, represents the set  $K$  where the piecewise smooth function  $u$  (also known as the restored image) is discontinuous and allow us to identify the borders of the cell colonies. Indeed, starting from the edge set, we applied subsequent refinements,



**Fig. 4** Second approach to ESC colonies segmentation: (a) edge set of the preprocessed image of Fig. 2-c; (b) binarized version of the edge set; (c) refined binarized image



**Fig. 5** Examples of ESC colonies extracted from the image in Fig. 1: **(a)** a domed colony (size about 4 % of the original image); **(b)** a flat colony (size about 7 % of the original image)

based on binarization (see Fig. 4-b), hole-filling for removing internal holes, and finally removal of very small objects, analogous to those of the previously described approach, which leads to the cell colony segmentation (see Fig. 4-c).

### 2.3 Features Computation

In order to classify the ESC colonies and be able to differentiate between domed and flat ones, we designed a general approach to feature selection and classification, based only on distinctive feature estimates that can be computed by the available images.

The most distinctive features that allow the biologist to discriminate the two kinds of colonies concern their shape: domed colonies appear compact and rounded, as opposed to flat colonies that spread more or less throughout the well (see Fig. 5). This prior knowledge is exploited by using geometric models locally estimated for each colony in a well. Among the considered shape features, the most significant are those reported in Table 1. The *Area*  $A(X)$  provides the number of pixels of the object (colony)  $X$  and the *Perimeter*  $P(X)$  is computed as the sum of the distances between consecutive pairs of boundary pixels of  $X$ . *Solidity*  $S(X)$  is given by the ratio of the area  $A(X)$  of the object  $X$  and the area  $A(CH(X))$  of its convex hull [5]. Solidity ranges in  $[0,1]$ , producing low values when the shape of the object shows many concavities (as in the flat colonies) and high values when the shape of the object shows few or zero concavities (as in the domed colonies). The irregularity of a contour is expressed through *Compactness*  $C(X)$  [7], given by the ratio between the area  $A(X)$  of the object and  $P^2(X)/4\pi$ , the area of a circle having the same perimeter  $P(X)$  as the object. Indeed, this feature, ranging in  $[0, 1]$ , provides low values for scarcely compact shapes (as in the flat colonies) and high values for compact shapes (as in the domed colonies), reaching its maximum for the most

**Table 1** Some of the adopted shape (S) and texture (T) features and their values for the ESC colonies of Fig. 5

| Feature         | Description   | Value for Fig. 5-a | Value for Fig. 5-b |
|-----------------|---|--------------------|--------------------|
| (S) Area        | $A(X) = \# \text{ pixels of object } X$   | 1512               | 3258               |
| (S) Perimeter   | $P(X) = \text{length of border of object } X$   | 153.09             | 496.54             |
| (S) Solidity    | $S(X) = A(X)/A(CH(X))$  | 0.93               | 0.65               |
| (S) Compactness | $C(X) = 4\pi A(X)/P^2(X)$   | 0.81               | 0.18               |
| (T) Entropy     | $H(X) = -\sum_i \sum_j M(i, j) \log_2 M(i, j)$  | 6.75               | 6.96               |
| (T) Contrast    | $CN(X) = \sum_i \sum_j (i - j)^2 M(i, j)$   | 0.28               | 0.30               |
| (T) Energy      | $E(X) = \sum_i \sum_j M(i, j)^2$  | 0.37               | 0.14               |
| (T) Correlation | $CR(X) = \frac{\sum_i \sum_j (i - \mu_i)(j - \mu_j)M(i, j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}$ | 0.96               | 0.95               |

compact shape: the circle. Other adopted shape features are: *Eccentricity*, that is the eccentricity of the minimum area ellipse including the colony image; *Min* and *Max axis length*, that are the minimum and the maximum axis length of the ellipse containing the colony image, respectively; *Nsegments*, that gives the ratio of the number of contour segments and the area of a colony image.

Besides shape estimates, textures have also been considered. In an image, the intensity variations which define a texture are mostly related to physical variations in the scene (such as pebbles on the ground). It is very difficult to model these variations and no precise definition of texture is present in computer vision literature [6]. For this reason, textures are usually characterized by intensity value variations in the two-dimensional space of an image. The adopted texture features of each object can be described in terms of the *gray-level co-occurrence matrix* (GLCM) [13], that allows us to capture the spatial dependence of gray-level values which contribute to the perception of texture, by showing how often different combinations of pixel brightness values occur in an image. The GLCM matrix is a square matrix  $M$  of dimension  $n$ , where  $n$  is the number of different gray levels in the image. Each element  $M(i, j)$  is generated by counting the number of times a pixel in  $X$  with gray value  $i$  is adjacent to a pixel in  $X$  having gray value  $j$ . Each element  $M(i, j)$  is then normalized so that the sum of all elements of  $M$  is equal to 1, and can thus be considered as the probability of occurrence of adjacent pixel pairs having gray level values  $i$  and  $j$  in the image. Among the considered texture features, the most significant are those reported in Table 1. The *Entropy*  $H(X)$  measures the randomness of the gray-level distribution in the colony image  $X$ , while the *Contrast*  $CN(X)$  is a measure of the intensity contrast between a pixel and its neighbors over  $X$ , assuming null value for a constant image. The *Energy*  $E(X)$  yields the smallest value when all entries in  $M$  are equal; it is 1 for a constant image. The *Correlation*  $CR(X)$  measures the dependency of gray levels on those of neighboring pixels. Further texture features that have been considered are the *Homogeneity*, that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal, and *Centropy* and *Bentropy*, that measure the entropy of color and grayscale colony image, respectively.

## 2.4 Classification

In machine learning, supervised classification methods aim at inferring a classification rule from a class-labeled set of examples described by a set of features. The inferred rule is then used for predicting the class of further unlabeled data. Therefore, a classifier can be considered as a mapping from a feature space to a set of classes. There exist several classifiers in literature, since there are several ways of building up such a mapping. For example, the mapping can be described by a set of induction rules, like in the tree-based classifiers, or the classifiers may be expressed as a linear separator in the original feature space, like for the perceptron or the support vector machines (SVM) classifiers. The choice of a good classifier is not a simple task, as it depends on several choices [14], including the complexity of the classification rule, the size of the training set, and the number of features.

For the classification of ESC colonies into domed and flat ones, we trained a set of binary classifiers from a set of labeled images described by the selected features (see Sect. 2.3). The classifiers have been chosen based on the criterion that the classification rule can be expressed either by a set of induction rules or by an easy interpretable mathematical function, where the importance of each feature in the classification rule is easily interpretable. This suggests which features are more relevant for labeling an image.

Specifically, we adopted three classifiers based on decision trees, that return induction rules for deciding the label of an instance: CART [2], J48 [19] and Adaboost [11] using as weak learners CART classifiers (Adaboost+CART). Among the classifiers searching for a linear separator in the feature space between instances with different labels, we used two versions of the SVM classifier: the SVM linear [24] and the SVM linear with hinge loss [18]. Finally, we adopted the Naïve Bayes classifier [15], a probabilistic algorithm which assigns a probability to each instance for each label, according to the highest label probability. All the above classifiers are available through the open source Weka software [12], that allows testing several classifiers in a very user friendly way.

## 3 Experimental Results

The proposed software framework has been tested on images of a CF assay done by plating mESCs in 96-well plates at the density of 500 cells/cm<sup>2</sup> in medium for propagation of undifferentiated cells (with Lif and serum). After cell adhesion (4 h), cells were treated with reference compounds controlling their metastability or differentiation, L-Proline [4] and Retinoic Acid [23], respectively, added at three different concentrations and allowed to grow for 4 days. Single well bright field images were acquired under a Leica MZ16FA stereo microscope.

Randomly chosen images of treated and untreated wells have been manually annotated, in order to provide the ground-truths for objective evaluation of the software components.

### 3.1 Performance Metrics

Validation of all the system modules has been performed in terms of different metrics frequently adopted in the literature [16], namely *Precision* and *Recall*

$$Recall = \frac{TP}{TP + FN}, \quad Precision = \frac{TP}{TP + FP},$$

where  $TP$ ,  $FN$ , and  $FP$  indicate the total number of true positives, false negatives, and false positives, respectively. *Recall*, also known as *detection rate* or *sensitivity*, gives the percentage of detected true positives as compared to the total number of true positives in the ground truth. *Precision*, also known as *positive prediction*, gives the percentage of detected true positives as compared to the total number of items detected by the method. Using the above mentioned metrics, generally a method is considered *good* if it reaches high *Recall* values, without sacrificing *Precision*. A further metric  $F_1$ , also known as *F-score* or *F-measure*, given by the weighted harmonic mean of *Precision* and *Recall*

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision},$$

provides a single measure that can be used to “rank” different methods.

For validating the classification accuracy, we adopted also the *False Positive Rate* (*FPR*)

$$FPR = \frac{FP}{FP + TN},$$

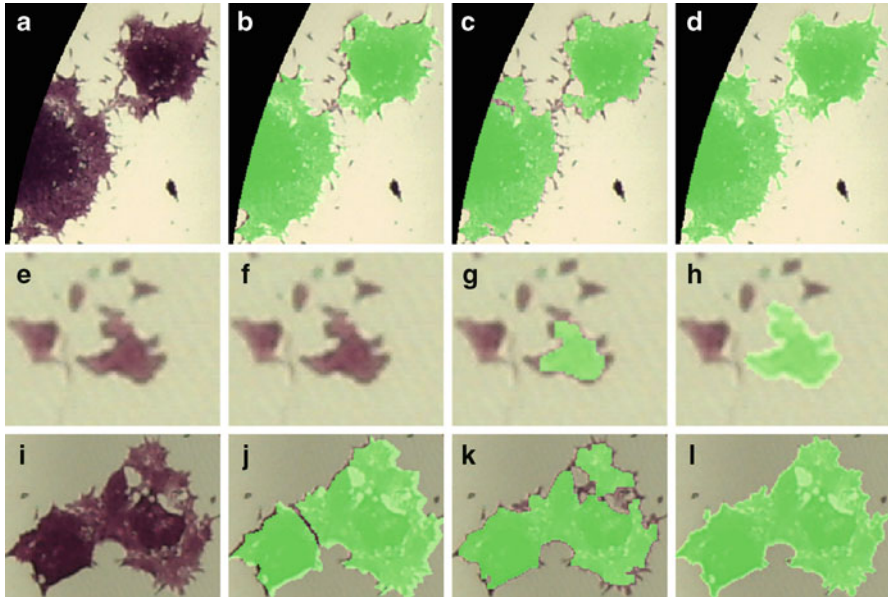
where  $TN$  indicates the total number of true negatives.

### 3.2 Evaluation of the Segmentation Step

For the segmentation step, *Recall*, *Precision*, and  $F_1$  metrics have been adopted both as pixel-based and as object-based measures, depending on whether *true positives* (resp. *false negatives*) are intended as true positive (resp. false negative) pixels or objects (colonies), respectively. While the pixel-based measures provide hints on the fine-grain segmentation accuracy, the object-based measures provide hints on the colony counting ability of the proposed modules. Average segmentation accuracy of

**Table 2** Average segmentation accuracy of the two segmentation approaches

| Metric               | Approach 1  |              | Approach 2  |              |
|----------------------|-------------|--------------|-------------|--------------|
|                      | Pixel-based | Object-based | Pixel-based | Object-based |
| <i>Recall</i>        | 0.7246      | 0.9194       | 0.8793      | 0.8387       |
| <i>Precision</i>     | 0.9663      | 0.9344       | 0.8826      | 0.9630       |
| <i>F<sub>1</sub></i> | 0.8282      | 0.9268       | 0.8809      | 0.8966       |

**Fig. 6** Evaluation of the segmentation step: details of the original image of Fig. 1 (first column), of the ground truth (second column), and of corresponding results obtained by the first and second segmentation approaches (third and fourth column, respectively). Green pixels superimposed on the original image indicate pixels included into the segmentation

the two segmentation approaches in term of the above metrics is reported in Table 2 and some details of the achieved results are reported in Fig. 6. The higher pixel-based accuracy values of the second approach (fourth column of Table 2) indicate a better ability of the variational model-based segmentation to precisely detect the colony contours (e.g., compare Figs. 6-g and 6-h). However, this ability leads also to fuse different colonies that are very close (e.g., see Fig. 6-d), resulting in lower object-based accuracy values (fifth column of Table 2).

Both the approaches report very few false positives and false negatives. Examples of false positives, e.g., segmented colonies that have no interest for the biologist, are reported in Figs. 6-g and 6-h, where the detected small colony has been excluded by the biologist in the ground truth reported in Fig. 6-f. Most of the false negatives are linked to adjacent colonies, barely distinguishable by an untrained human



eye, that are segmented as a single colony. An example is given in the third row of Fig. 6, where the adjacent colonies of the original image (Fig. 6-i) have been manually segmented in the ground truth as two separate colonies (Fig. 6-j) based only on the biologist experience. The two approaches, instead, both provide a single segmented colony (Figs. 6-k and 6-l). This analysis of false positives and false negatives suggests that features other than contrast or edges, able to better describe the discrimination ability of the trained eye, could help avoiding the few segmentation errors reported. Hints could be provided by the subsequent steps of feature computation and feature-based classification.

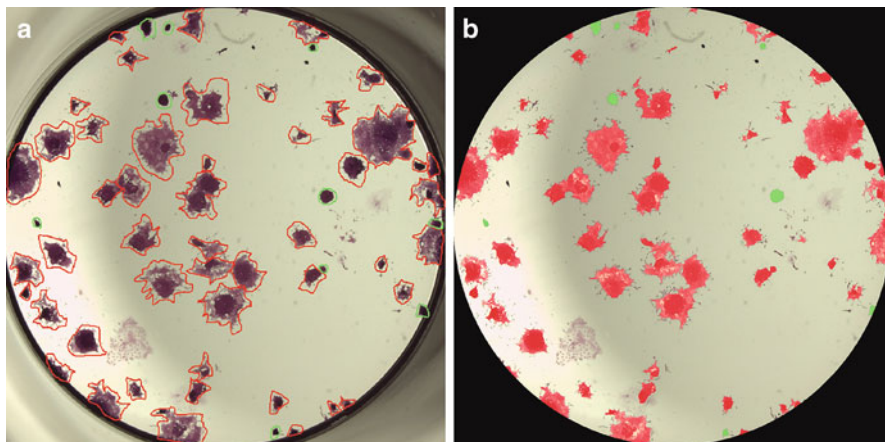
Although the segmentation results are quite accurate for both the proposed approaches, a suitable combination of them, together with a feedback from the subsequent steps of feature computation and classification, could help further improving segmentation accuracy.

### 3.3 First Classification Results

Each of the selected classifiers (see Sect. 2.4) has been trained using the selected features (see Sect. 2.3) of the colonies included into the constructed ground truths. The statistical validation of results has been obtained using a 10-fold cross validation. This procedure consists in splitting the training set in ten parts; at each step, one of those parts is used for testing and the remaining nine are used for training. Therefore, the performance metrics have been evaluated as the average among the one hundred repetitions of the ten-fold cross validation. Table 3 reports the classification performance values for each classifier, suggesting that all the six classifiers are very accurate in predicting the label of an instance. In the case of SVM with hinge loss, the number of false positives is reduced up to 1%. This is exemplified in Fig. 7, where we compare the manual classification provided by the biologist with the automatic classification computed using the SVM with hinge loss for the well of Fig. 1. Table 4 reports the weights (one for each feature) of the normalized equation of the hyperplane that separates domed vs. flat colonies in the feature space. Positive weights are associated to domed colonies. The higher the absolute value of the weights, the more the feature is relevant for discriminating

**Table 3** 10-fold cross validation of classification results

| Classifier                 | <i>FPR</i>   | <i>Recall</i>  | <i>Precision</i> | <i>F-measure</i> |
|----------------------------|--------------|----------------|------------------|------------------|
| CART                       | 0.040        | 0.96585        | 0.96593          | 0.96587          |
| J48                        | 0.032        | 0.96585        | 0.96635          | 0.96589          |
| Adaboost+CART              | 0.024        | 0.97561        | 0.97563          | 0.97560          |
| SVM linear                 | 0.038        | 0.97073        | 0.97097          | 0.97076          |
| SVM linear with hinge loss | <b>0.010</b> | <b>0.98537</b> | <b>0.98540</b>   | <b>0.98536</b>   |
| Naïve Bayes                | 0.024        | 0.97561        | 0.97568          | 0.97562          |



**Fig. 7** Example of classification results: (a) manual classification; (b) automatic classification using SVM with hinge loss

**Table 4** SVM with hinge loss: shape (S) and texture (T) features ordered by the absolute value of (normalised) weights

| Feature         | Weight  | Feature             | Weight  | Feature         | Weight  |
|-----------------|---------|---------------------|---------|-----------------|---------|
| (S) Solidity    | 8.3231  | (S) Eccentricity    | 3.4247  | (T) Contrast    | 1.8952  |
| (S) Compactness | 7.5965  | (T) Color           | -2.7622 | (S) Perimeter   | -1.8120 |
| (T) Correlation | -6.4780 | (S) Min axis length | -2.6790 | (T) Homogeneity | 1.6923  |
| (S) Nsegments   | 5.9170  | (T) Centropy        | -2.4002 | (S) Area        | -0.7793 |
| (T) Energy      | 5.1318  | (S) Max axis length | -2.1554 | (T) Bentropy    | 0.3709  |

the classes. Also the other classifiers<sup>1</sup> confirmed that, as expected, *Solidity* and *Compactness* shape features are the most discriminant features for our binary classification. Among texture features, *Correlation* is the most relevant.

## 4 Concluding Remarks

In this paper we present a novel software framework for the segmentation and classification of microscopic images obtained by CF assays of mESC and suitable for high-throughput applications. Our main aims were to use general-purpose software components for image processing and machine learning, eventually developed in different projects, to support biological experiments within a multidisciplinary context. First results show that our approach is promising to be used as blind tool

<sup>1</sup>Detailed results may be supplied on demand.

to support reliable analysis of molecular high-throughput screening. Future works will include comparisons in terms of efficiency, reliability and flexibility with some existing application-specific software tools and the development of a user-friendly software interface for biologists without deep knowledge in computer programming.

**Acknowledgements** This work was partially supported by public-private laboratory for the development of integrated informatics tools for genomics, proteomics and transcriptomics (LAB GPT), funded by MIUR. We also thank the Integrated Microscopy Facility at the IGB-ABT, CNR.

## References

1. Bar, L., et al.: Mumford and Shah model and its applications to image segmentation and image restoration. *Handbook of Mathematical Methods in Imaging*, vol. I, pp. 1095–1157. Springer, New York (2011)
2. Breiman, L., Friedman, J., Olshen, R., Stone, C.: *Classification and Regression Trees*. Chapman & Hall, New York (1984)
3. Carpenter, A., Jones, T., Lamprecht, M., Clarke, C., Kang, I., Friman, O., Guertin, D., Chang, J., Lindquist, R., Moffat, J., Golland, P., Sabatini, D.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**(R100) (2006)
4. Casalino, L., Comes, S., Lambazzi, G., De Stefano, B., Filosa, S., De Falco, S., De Cesare, D., Minchiotti, G., Patriarca, E.: Control of embryonic stem cell metastability by l-proline catabolism. *J. Mol. Cell Biol.* **3**(2), 108– (2011)
5. Celebi, M., Kingravi, H., Uddin, B., Iyatomi, H., Aslandogan, Y., Stoecker, W., Moss, R.: A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph.* **31**(6), 362–373 (2007)
6. Chen CH Pau LF, W.P.: *The Handbook of Pattern Recognition and Computer Vision* (2nd Edition). World Scientific Publishing Co, Singapore (1998)
7. Cozza, V., Guarracino, M.R., Maddalena, L., Baroni, A.: Dynamic clustering detection through multi-valued descriptors of dermoscopic images. *Stat. Med.* **30**, 2536–2550 (2011)
8. D’Ambra, P., Filippone, S.: A parallel generalized relaxation method for high-performance image segmentation on gpus. *J. of Comput. Appl. Math.* **293**, 34–44 (2016)
9. D’Ambra, P., Tartaglione, G.: Solution of Ambrosio-Tortorelli model for image segmentation by generalized relaxation method. *Commun. Nonlinear Sci. Numer. Simul.* **20**, 819–831 (2015)
10. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM* **15**(1), 11–15 (1972)
11. Freund, Y., Schapire, R.E.: A short introduction to boosting. In: *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 1401–1406. Morgan Kaufmann (1999)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. *SIGKDD Explor. Newsl.* **11**(1), 10–18 (2009)
13. Haralick, R., Shanmugam, K.: Computer classification of reservoir sandstones. *IEEE Trans. Geosci. Electron.* **11**, 171–177 (1973)
14. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2 edn. Springer, New York (2013)
15. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: *Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345. Morgan Kaufmann, San Mateo (1995)
16. Maddalena, L., Petrosino, A.: The 3dSOBS+ algorithm for moving object detection. *Comp. Vision Image Underst.* **122**(0), 65–73 (2014)

17. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man and Cybern.* **9**(1), 62–66 (1979)
18. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B, Burges, C, Smola, A. (eds.) *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge (1998)
19. Quinlan, R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo (1993)
20. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al.: Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**(7), 676–682 (2012)
21. Serra, J.: *Image Analysis and Mathematical Morphology*. Academic Press, Inc, New York (1983)
22. Shariff, A., Kangas, J., Coelho, L., Quinn, S., Murphy, R.: Automated image analysis for high content screening and analysis. *J. Biomol. Screening* **15**, 726–734 (2010)
23. Tighe, A., Gudas, L.: Retinoic acid inhibits leukemia inhibitory factor signaling pathways in mouse embryonic stem cells. *J Cell Physiol.* **198**(2), 223–229 (2004)
24. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, New York (1995)
25. Zhou, X., Wong, S.T.: Informatics challenges of high-throughput microscopy. *IEEE Signal Process. Mag.* **23**, 63–72 (2006)

# Exploiting “Mental” Images in Artificial Neural Network Computation

Massimo De Gregorio and Maurizio Giordano

**Abstract** In Artificial Neural Network (ANN) computing the learned knowledge about a problem domain is “implicitly” used by ANN-based system to carry on Machine Learning, Pattern Recognition and Reasoning in several application domains. In this work, by adopting a Weightless Neural Network (WNN) model of computation called DRASiW, we show how the knowledge of a problem, internally stored in a data representation called “Mental” Image (MI), can be made “explicit” both to perform additional and useful tasks in the same domain, and to better tune and adapt WNN behavior in order to improve its performance in the target domain. In this paper, three case studies of MI processing in the realm of WNN applications are discussed with the aim of proving the viability and the potentialities of exploiting internal knowledge of WNNs to self-adapt and improve their performance.

**Keywords** Weightless systems - Mental images

## 1 Introduction

In traditional ANNs the knowledge about a problem domain is coded in the configuration of synaptic weights between neurons. The goal of the ANN training phase is to find the optimal configuration of weights that allows the network to properly generate the expected outputs in the classification/recognition phase. The configuration of weights can be considered as the internal state of the network. How it is obtained and changed during the network operation is a matter of the particular

---

M. De Gregorio (✉)  
Istituto di Scienze Applicate e Sistemi Intelligenti “Eduardo Caianiello” – CNR, Via Campi  
Flegrei 34, 80078 Pozzuoli, NA, Italy  
e-mail: [massimo.degregorio@cnr.it](mailto:massimo.degregorio@cnr.it)

M. Giordano  
Istituto di Calcolo e Reti ad Alte Prestazioni – CNR, Via Pietro Castellino 111,  
80131 Naples, Italy  
e-mail: [maurizio.giordano@cnr.it](mailto:maurizio.giordano@cnr.it)

ANN model adopted. What is important is that, once the network architecture (i.e., layers, number of neurons per layer, and connection paths) is set, the ANN configuration of weights fully characterizes its behavior.

RAM-based neural networks are alternative models of ANNs in which the learned knowledge about the problem domain is coded inside RAM-neuron contents rather than on their interconnections. As in the classical weight-based ANNs, the particular configuration of RAM cells is obtained in the training phases, either if they are carried out in super-, semi- or unsupervised manner. At any time during a RAM-based neural network operation, the configuration of RAM contents represents the internal state of the network. Once we have set the RAM-based ANN architecture (i.e., layers, number of neurons per layer, RAM bit address, type of data stored, etc.), the “image” (snapshot) of RAM contents fully characterizes the internal state of the ANN and, as a consequence, the image represents the knowledge and the behavior of the ANN functioning.

Generally in ANN models this internal state is *implicit*. Although the internal state of the learning process is coded by the information stored in the ANN data structures (either weights or RAM contents), this information may not be accessible by the neural-based system to be exploited in a computational meta-level. In the RAM-based model of the ANN adopted in this work, the DRASiW weightless model, this is possible thanks to a particular feature: the contents of RAM-neurons not only characterize the network behavior, but they are also an additional information explicitly available to the neural-based system, in such a way that the ANN can process this information in a computing meta-level in order to adapt and to tune its future behavior.

As in [10], our approach tries to make explicit the internal representation of knowledge of an ANN with the aim of facilitating an interpretation (that can be geometrical, physical, symbolic, etc.) of the learning process and of discovering its correlation to the input. While authors do not suggest applications of the ANN inner knowledge processing, in our work we prove with real case studies how to exploit this knowledge to adapt to domain changes as well as to improve ANN performance in the target domain.

Works like [13, 18] propose methods to interpret and to make explicit the ANN internal knowledge by extracting the knowledge in form of rules (either symbolic or fuzzy) with the only aim of using such rules to simulate the ANN behavior. On the contrary, in our approach we exploit learned knowledge of an ANN to improve and/or to adapt the performance of the same ANN, automatically and/or with the user feedback, to a data domain which may change in time or may contain incomplete and/or ambiguous information.

The fact that we start from an already trained ANN and we refine its performance, by extracting and exploiting its internal knowledge, makes our approach also different from others, like [21], in which the knowledge of an ANN trained on a problem domain is used to extract a set of concise and intelligible symbolic rules that can be used to “refine” an already existing rule-based system, which may have an incomplete or even incorrect initial knowledge of the target problem.

Another close topic is how to integrate explicit and implicit knowledge in neuro-symbolic (hybrid) processing [15]. In this perspective the solution presented in our work can be considered hybrid too. Regardless of how we explicitly represent the mental images of ANNs, our intent is to exploit high-order characteristics of the learned knowledge (macro quantities, invariants, etc.) as an additional information to the neural-based system to improve its performance.

This chapter is organised as follows. Sections 2 and 3 are devoted to the introduction of the DRASiW model and its internal knowledge representation (“Mental” Images). Section 4 shows and discusses three different applications in which the use of “Mental” Images in the computational process improves the performance of the DRASiW systems. Finally, Sect. 5 sums up concluding remarks and perspectives.

## 2 DRASiW Model

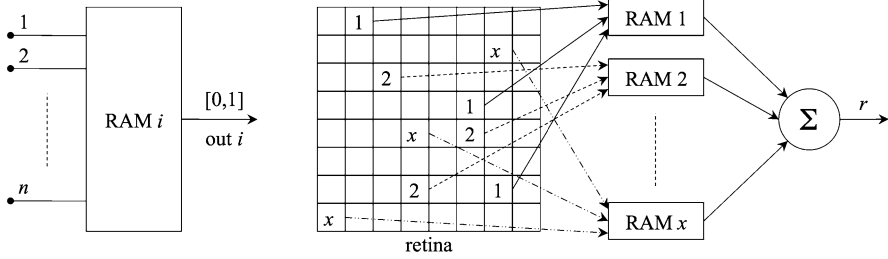
Weightless Neural Networks (WNNs) [1, 12], differently from classical ANNs, adopt a RAM-based model of neuron by which learning information about a data domain is stored into RAM contents instead of computed weights of neuron connections. A RAM-neuron receives an  $n$ -bit input that is interpreted as a unique address (stimulus) of a RAM cell, and used to access it either in writing (learning) or reading (classification) mode. WNNs have proved to provide fast and flexible learning algorithms [2].

WiSARD systems are a particular type of WNN, that can be developed directly on reprogrammable hardware [3]. A WiSARD is composed by a set of classifiers, called *discriminators*, each one assigned to learn binary patterns belonging to a particular class. The WiSARD, also called *multi-discriminator architecture*, has as many discriminators as the number of classes it should be able to distinguish.

Each discriminator consists of a set of RAM-neurons, which store the information of occurrences of binary patterns during the learning stage. Given a binary pattern of size  $s$ , the so-called *retina*, it can be classified by a set of WiSARD discriminators, each one having  $m$  RAMs with  $2^n$  cells such that  $s = m \times n$ . Since each RAM cell is uniquely addressed by an  $n$ -tuple of bits, the input pattern can be partitioned into a set of  $n$ -tuples of bits, each one addressing one cell of a RAM.  $n$ -tuples of bits are pseudo-randomly selected and biunivocally mapped to RAMs (see right part of Fig. 1), in such a way that the input binary pattern is completely covered.

The WiSARD training phase works as follows:

1. *Initialization*: all RAMs cells for each discriminator are set to 0.
2. *Training set selection*: a training set of binary patterns, all with the same size, is selected; each pattern is known to belong to (and to represent) only one class.



**Fig. 1** RAM-neuron (left) and WiSARD discriminator (right)

3. *Training*: for each training pattern the discriminator assigned to the belonging class is selected; the pseudo-random mapping is used to define, from the binary pattern, all  $n$ -tuples; each  $n$ -tuple forms a unique address of a RAM cell of the discriminator, whose content is set to 1.

After training, if a RAM cell is set to 0 then the  $n$ -tuple of bits in the retina, corresponding to physical address (in binary notation) of that memory cell, never occurred across all samples in the training set, otherwise it occurred at least in one sample.

The WiSARD classification phase works as follows:

1. *Test set selection*: a test set of binary patterns, all with the same size, is selected; for each sample of the test set we want to know which category it belongs to.
2. *Classification*: the pseudo-random mapping is used to extract, from each test pattern, the  $n$ -tuples of bits in such a way to identify RAM cells to be accessed across all discriminators; contents of accessed cells are summed by an adder ( $\Sigma$ ) so giving the number  $r$  of RAMs that output 1;  $r$  is called *discriminator response*.

It is easy to see that  $r = m$  if the input pattern belongs to the training set. While  $r = 0$  if no  $n$ -tuple of bits in the input pattern appears in the training set. Intermediate values of  $r$  express a “similarity measure” of the input pattern with respect to training patterns. The adder enables a network of RAM-neurons to exhibit (like ANN models based on synaptic weights) generalization and noise tolerance [2].

DRASiW [8] is an extension of WiSARD: instead of having RAM cells set to 1 once accessed during training, they are incremented by 1 at each access. Thus, after training, RAM contents store the number of occurrences (frequency) of a specific  $n$ -tuple of bits across training patterns. The new domain of memory cells contents (non negative integers) produces the same classification capability of a WiSARD provided that  $\Sigma$  counts the number of addressed non-zero memory cells.

The DRASiW model augments the WiSARD model adding a backward classification capability by which it is possible to generate *prototypes* (i.e., representative samples) of classes learned from training patterns [11, 19]. In DRASiW, RAM-neuron cells act as access counters, whose contents can be reversed to an internal “retina” storing a “Mental” Image (MI). Memory cell contents of DRASiW

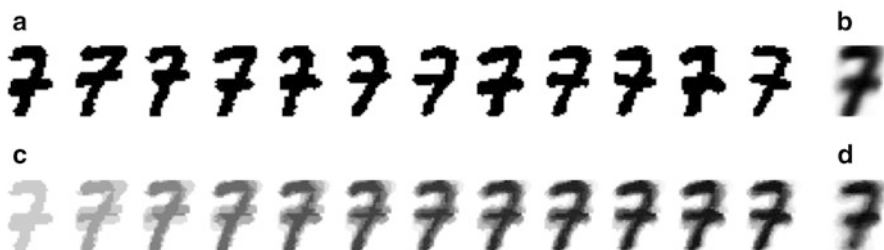


discriminators can hence be interpreted as sub-pattern frequencies in the training set. The MI and the internal “retina” metaphors were originally explored and discussed with respect to their cognitive plausibilities in [6].

### 3 Mental Images

There are two different ways a DRASiW system can produce MIs: statically and dynamically. Static MIs are generated after the training phase and do not change anymore. They represent a pictorial representation of the discriminator internal information. Let consider the 12 instances of black “7”s, reported in Fig. 2a, as the training set for the class “seven”. An example of static MI produced by a DRASiW system trained on this training set, is reported in Fig. 2b. This gray level, non-crisp example of class “seven” is the result of how the sub-patterns appear in the training set. In fact, the gray levels are generated taking into account the sub-pattern frequencies.

Another way of producing MIs is to update them each time the system receives a new training set pattern. This mode, also called *online training*, is by far one of the more interesting operation mode of a DRASiW system. There are many applications in which the system has to adapt to the new and changing appearance of the pattern to classify. The only way to face this problem is to update and store the new information in the MIs. The system updates the MIs each time it receives a new pattern. In Fig. 2c, the reader can notice how the MI changes with respect to the input of patterns. The first MI is produced just with the first “7”. The second one is produced by increasing the gray level of those pixels in common with the previous pattern (more frequent pixels). All the other MIs are the result of applying this procedure each time a new pattern is presented to the system. The MI in Fig. 2d is the result at the end of the process. To sum up, RAM contents corresponding to sub-patterns of the binary input on the retina are increased by one (*reinforcement*), while RAM contents corresponding to those sub-patterns which were not present in the binary input image on the retina are decremented by one (*forgetting*). In other words,



**Fig. 2** Static and dynamic mental images: (a) training patterns; (b) static MI; (c) dynamic MI after each training pattern; (d) dynamic MI at the end of training

the *Reinforcement & Forgetting* strategy (*RF*) allows to store the frequency of sub-patterns occurrences during training time. In this way the MI stored and updated in time represents a sort of dynamic prototype (history) of the corresponding class.

## 4 Improving DRASiW Performance

The first two applications reported in the following subsections deal with the problems of tracking deformable objects and of isolating the background in video sequences. Both of these applications take advantage of dynamic MIs. The third application faces the problem of classification through features. In this case, the information coded in MIs is exploited by the system to identify a set of “metaclasses” used to better refine and improve the classification process.

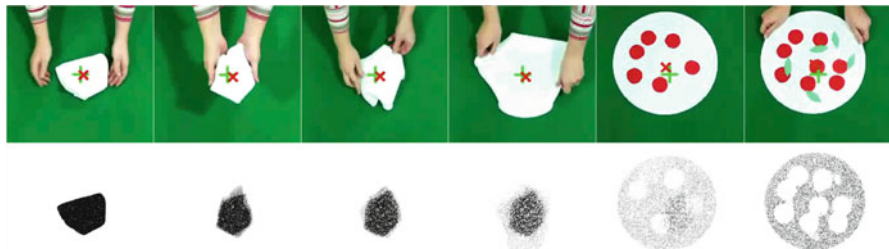
### 4.1 Tracking Deformable Objects

In the realm of object tracking systems [22], many real life scenarios, which span from domestic interaction to industrial manufacturing processes, pose hard challenges. In particular, when the object is non-rigid, deformable, and/or manipulated during the tracking, both its position and deformation have to be followed.

In [20] we present a DRASiW system designed and implemented for tracking deformable objects. It supports online training on texture and shape of the object, with the aim of adapting in real-time to changes and of coping with occlusions. This object tracking system deals with *Pizza Making* problem. Pizza is a non-rigid deformable object that can assume whatever shape we want. Hence, it is not possible to define a model for the tracking. In this context, the system should be able to dynamically identify the pizza dough and robustly track it without prior knowledge.

At the beginning, the tracking system is fed with an image representing the object to follow with its initial shape and position. This image is used to train a set of DRASiW discriminators: one discriminator is placed at the target position, the remaining discriminators are placed all around the target position with increasing displacements in the *XY* directions. The configured set of discriminators forms the so called *prediction window* of the tracking system. When the object starts moving, the DRASiW-based tracking system tries to localize the object through the discriminator responses. The higher is the response the more probable the object is in that part of the prediction window processed by that discriminator. Once the system localizes the object in a new position, it uses this image to train again the set of discriminators in the prediction window which is also displaced jointly to the target. So doing, the MI of the object is updated and, hence, it will represent the more recent object shapes.

Figure 3 shows snapshots of pizza making actions: manipulation, dough stretching, seasoning, and baking. The outputs of the DRASiW system are represented



**Fig. 3** Sketches of the DRASiW tracking system results in a frame sequence (*top row*), and corresponding MIs (*bottom row*)

by colored crosses. The green cross represents the retina center of the discriminator with the higher response; while the red cross is the mass center of the current MI.

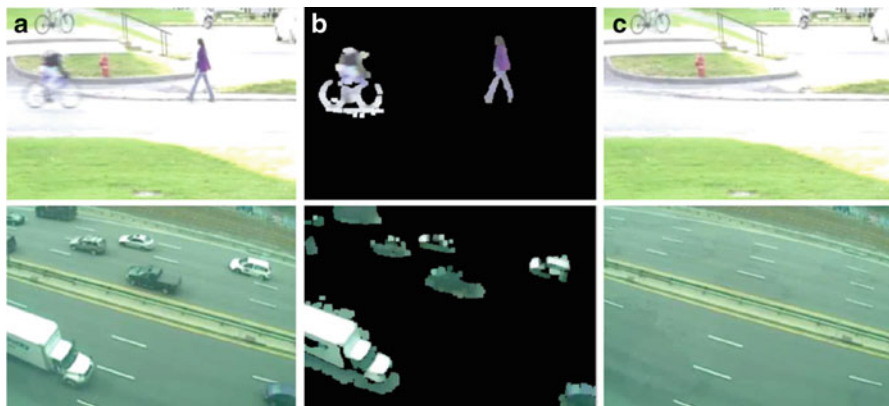
As one can notice, the tracking results improve if the DRASiW system takes into account the information given by the updated MI. We could not reach the same performance if the information contained in the current MI had not been exploited by the system tracking procedure.

## 4.2 Generating Background Models

Change Detection (CD) is the problem of separating foreground objects from background areas in a video sequence. Several techniques and solutions have been proposed to face the CD problem. Evaluation and comparison surveys of existing techniques can be found in [5, 14, 16, 17]. Regardless of the specific applied method,<sup>1</sup> most approaches share the basic idea of insulating moving objects from the background by comparing image areas of new video frames with respect to either a *background model* or a *model of the target moving objects*. Background models can also be classified as pixel-based or region-based depending on whether computation is based on only the pixel color or a neighborhood of pixels.

In [9] a CD method based on DRASiW is proposed. It exploits a pixel-based background model built around the notion of MI. In the approach, pixel processing is carried out by a DRASiW discriminator. The information stored in neurons is related to the evolution of changeable pixel color in the video timeline. The dynamic MI associated to each pixel represents the *dynamic background model* of it, that is the storing in time of more frequent and up-to-date RGB values assumed by that pixel in video frames. The *RF* mechanism (see Sect. 3) allows to dynamically adapt the MI in such a way that, during the video timeline, not up-to-date RGB values gradually disappear from the background model while new and stable colors of recent frames

<sup>1</sup>Just to mentions a fews: physical models, statistical methods with Gaussian mixtures, pixel clustering, image filtering (Kalman, Grabcut, etc.), particle filters and neuron network modeling.



**Fig. 4** Outputs of the DRASiW-based CD method: (a) original frame; (b) moving objects highlighting; and (c) MI background model

will contribute more in the background model. The dynamic MI of pixels allows to better adapt the background model to gradual changes in brightness of lights and shadows as well as to natural background noise. Foreground object detection is carried out by evaluating whether the difference between the current pixel color and the stored MI model of the background overcomes a certain threshold. A queue of more recent foreground samples is used to control the time the pixel stays in the foreground. When the queue is full it means that an object was moved to a position of the scene and it has become part of the background.

In Fig. 4 snapshots showing the outputs of the DRASiW-based CD method<sup>2</sup> are reported. As one can notice, MIs are not only used to fully control the change detection process, but also to filter the input video in order to accomplish two important tasks in video surveillance: 1) moving objects highlighting (see Fig. 4b); 2) subtracting changeable areas from video frames (see Fig. 4c).

### 4.3 Improving Classification

Activity Recognition aims at identifying the actions carried out by a person given a set of observations of itself and the surrounding environment [7]. Recognition can be accomplished, for example, by exploiting the information retrieved from inertial sensors, such as accelerometers. In some smartphones these sensors are embedded by default and one can benefit from this to classify a set of physical activities (standing, sitting, laying, walking, walking upstairs and walking downstairs) by

<sup>2</sup>The proposed method participated in the international competition of CD methods on the video repository [ChangeDetection.net](http://ChangeDetection.net) in 2014, reporting the 3rd best score.

processing inertial body signals through a supervised Machine Learning algorithm for hardware with limited resources [4].

We tried to classify this set of physical activities with DRASiW trained and tested on the HAR<sup>3</sup> (Human Activity Recognition) data set of the UCI Machine Learning Repository. The data set consists of 10,299 instances: 7352 for the training set, and 2947 for the test set. Each instance is formed by 561 features with time and frequency domain variables.

The confusion matrix obtained with the best DRASiW system configuration (16-bit addressing for RAM cells) performing an *F-measure* of 89.7, is shown in Table 1a. The confusion matrix of the Table 1b reports the *F-measure* obtained by the same DRASiW system configuration but exploiting the information content of the static MIs. The DRASiW system automatically analyses the MIs to identify features with a very high discriminating power. The analysis outcome is that the six classes can be grouped in three different “metaclasses”: walking (classes 1, 2, and 3), vertical activity (classes 4 and 5), horizontal activity (class 6). This is automatically discovered by the DRASiW system finding out MI overlappings. For the above metaclasses, the MIs have no intersection (no confusion). At this point, when the DRASiW system has to classify a test sample, it first selects the best-matched metaclass, then it classifies the test sample using only discriminators belonging to that metaclass. The result of this new two-level classification approach is that the confusion matrix is now almost diagonalized (see italic values in Table 1b), and the *F-measure* reaches the value of 94.1, that is, the system improved its classification power by 4.4 %.

## 5 Conclusions

The DRASiW model makes available the learned knowledge in form of an internal data structure called “Mental” Image. This information, which is the synthesis of the learning process, is explicitly available at the programming level and it can be used in several application domains. In this paper we showed how exploitation of MIs, in the context of a DRASiW computational process, allows to pursue different goals/tasks: 1) using global metrics and/or invariants of MIs as additional information (feedback) the system can take advantage of in order to control its functioning (self-healing); 2) verifying the correctness of a training procedure; 3) tuning/adapting classification process by detecting and exploiting more discriminating regions in MIs; 4) facilitating user-system interface and communication.

---

<sup>3</sup><https://archive.ics.uci.edu/ml/machine-learning-databases/00240/>.

**Table 1** Confusion matrices on *HAR* data set: (a) DRASIW; (b) DRASIW with MI processing

| (a)            |      |      |      |      |      |      | (b)         |                |      |      |      |      |      |     |             |
|----------------|------|------|------|------|------|------|-------------|----------------|------|------|------|------|------|-----|-------------|
| Class          | 1    | 2    | 3    | 4    | 5    | 6    | Recall      | Class          | 1    | 2    | 3    | 4    | 5    | 6   | Recall      |
| 1 – Walking    | 477  | 2    | 17   | 0    | 0    | 0    | 96.1        | 1 – Walking    | 482  | 1    | 13   | 0    | 0    | 0   | 97.2        |
| 2 – Upstairs   | 29   | 431  | 9    | 0    | 2    | 0    | 91.5        | 2 – Upstairs   | 12   | 457  | 2    | 0    | 0    | 0   | 97.0        |
| 3 – Downstairs | 47   | 60   | 310  | 0    | 2    | 1    | 73.8        | 3 – Downstairs | 18   | 47   | 355  | 0    | 0    | 0   | 84.5        |
| 4 – Standing   | 0    | 2    | 0    | 400  | 88   | 1    | 81.4        | 4 – Standing   | 0    | 0    | 0    | 416  | 75   | 0   | 84.7        |
| 5 – Sitting    | 0    | 0    | 1    | 33   | 498  | 0    | 93.6        | 5 – Sitting    | 0    | 0    | 0    | 7    | 525  | 0   | 98.7        |
| 6 – Laying     | 0    | 0    | 1    | 0    | 2    | 534  | 99.4        | 6 – Laying     | 0    | 0    | 0    | 0    | 0    | 537 | 100         |
| Precision      | 86.2 | 87.1 | 91.7 | 91.9 | 84.4 | 99.6 | <b>89.7</b> | Precision      | 94.1 | 90.5 | 95.9 | 98.4 | 87.5 | 100 | <b>94.1</b> |

We are aware that the natural unfolding of this work is looking for new ways of using MIs in the context of neurosymbolic systems. Indeed, this is the main investigation direction we will pursue in the next future on this topics. Although it would be nice to have a general formalism and/or (rule-based) high-order language to express the information contained in MIs, we are afraid that any choice would be inevitably effective only in a specific (or class of) problem domains.

## References

1. Aleksander, I., De Gregorio, M., França, F.M.G., Lima, P.M.V., Morton, H.: A brief introduction to weightless neural systems. In: Proceedings of the 17th European Symposium on Artificial Neural Networks, pp. 299–305 (2009)
2. Aleksander, I., Morton, H.: An Introduction to Neural Computing. Chapman & Hall, London (1990)
3. Aleksander, I., Thomas, W.V., Bowden, P.A.: WISARD a radical step forward in image recognition. *Sensor Rev.* **4**, 120–124 (1984)
4. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes–Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In: Proceedings of the 21st European Symposium on Artificial Neural Networks, pp. 437–442 (2013)
5. Bouwmans, T.: Recent advanced statistical background modeling for foreground detection: a systematic survey. *Recent Patents Comput. Sci.* **4**(3), 147–176 (2011)
6. Burattini, E., De Gregorio, M., Tamburrini, G.: Generation and classification of recall images by neurosymbolic computation. In: Proceedings of the 2nd European Conference on Cognitive Modelling, pp. 127–134 (1998)
7. Davies, N., Siewiorek, D.P., Sukthankar, R.: Activity-based computing. *IEEE Pervasive Comput.* **7**(2), 20–21 (2008)
8. De Gregorio, M.: On the reversibility of multi-discriminator systems. Technical Report 125/97, Istituto di Cibernetica, CNR (1997)
9. De Gregorio, M., Giordano, M.: Change Detection with Weightless Neural Networks, *IEEE Change Detection Workshop – CVPR 2014*, pp. 403–407 (2014)
10. Feng, T.J., Houkes, Z., Korsten, M., Spreeuwiers, L.: Internal measuring models in trained neural networks for parameter estimation from images. In: IPA, pp. 230–233 (1992)
11. Grieco, B.P., Lima, P.M.V., De Gregorio, M., França, F.M.G.: Producing pattern examples from “mental” images. *Neurocomputing* **73**(79), 1057–1064 (2010)
12. Ludermir, T.B., Carvalho, A.C., Braga, A.P., Souto, M.C.P.: Weightless neural models: a review of current and past works. *Neural Comput. Surv.* **2**, 41–61 (1999)
13. Mantas, C., Puche, J., Mantas, J.: Extraction of similarity based fuzzy rules from artificial neural networks. *Int. J. Approx. Reason.* **43**(2), 202–221 (2006)
14. Mc Ivor, A.: Background subtraction techniques. In: International Conference on Image and Vision Computing New Zealand, IVCNZ (2000)
15. Neagu, C.D., Palade, V.: Neural explicit and implicit knowledge representation. In: Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies, vol. 1, pp. 213–216 (2000)
16. Panahi, S., Sheikhi, S., Hadadan, S., Gheissari, N.: Evaluation of background subtraction methods. In: *Digital Image Computing: Techniques and Applications*, pp. 357–364 (2008)
17. Piccardi, M.: Background subtraction techniques: a review. In: *IEEE International Conference on Systems, Man and Cybernetics*, pp. 3099–3104 (2004)
18. Sato, M., Tsukimoto, H.: Rule extraction from neural networks via decision tree induction. In: *International Joint Conference on Neural Networks*, vol. 3, pp. 1870–1875 (2001)

19. Soares, C.M., da Silva, C.L.F., De Gregorio, M., França, F.M.G.: Uma implementação em software do classificador WiSARD. In: 5th SBRN, pp. 225–229 (1998)
20. Staffa, M., Rossi, S., Giordano, M., De Gregorio, M., Siciliano, B.: Segmentation performance in tracking deformable objects via WNNs. In: Robotics and Automation (ICRA) (2015)
21. Towell, G., Shavlik, J.: Extracting refined rules from knowledge-based neural networks. *Mach. Learn.* **13**(1), pp. 71–101 (1993)
22. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. *ACM Comput. Surv.* **38**(4), pp. 1–45 (2006)



# Applying Design of Experiments Methodology to PEI Toxicity Assay on Neural Progenitor Cells

**Sara Mancinelli, Valeria Zazzu, Andrea Turcato, Giuseppina Lacerra, Filomena Anna Digilio, Anna Mascia, Marta Di Carlo, Anna Maria Cirafici, Antonella Bongiovanni, Gianni Colotti, Annamaria Kisslinger, Antonella Lanati, and Giovanna L. Liguori**

**Abstract** Design of Experiments (DoE) statistical methodology permits the simultaneous evaluation of the effects of different factors on experimental performance and the analysis of their interactions in order to identify their optimal combinations. Compared to classical approaches based on changing only one factor at a time (OFAT), DoE facilitates the exploration of a broader range of parameters combinations, as well as providing the possibility to select a limited number of combinations covering the whole frame. The advantage of DoE is to maximise the amount of information provided and to save both time and money. DoE has been primarily used in industry to maximise process robustness, but recently it has also been applied in biomedical research to different types of multivariable analyses,

---

The authors “Sara Mancinelli”, “Valeria Zazzu” and “Andrea Turcato” contributed equally to this work.

The authors “Antonella Lanati” and “Giovanna L. Liguori” share senior co-authorship.

S. Mancinelli • V. Zazzu • G. Lacerra • G.L. Liguori (✉)

Institute of Genetics and Biophysics “A. Buzzati Traverso” (IGB), CNR, Naples, Italy  
e-mail: [giovanna.liguori@igb.cnr.it](mailto:giovanna.liguori@igb.cnr.it)

A. Turcato • A. Lanati,  
Valore Qualità, Pavia, Italy

F.A. Digilio  
Institute of Biosciences and Bioresources (IBBR), CNR, Naples, Italy

A. Mascia • A.M. Cirafici • A. Kisslinger  
Institute of Experimental Endocrinology and Oncology “G. Salvatore” (IEOS),  
CNR, Naples, Italy

M. Di Carlo • A. Bongiovanni  
Institute of Biomedicine and Molecular Immunology “A. Monroy” (IBIM),  
CNR, Palermo, Italy

G. Colotti  
Department of Biochemical Sciences, Institute of Molecular Biology and Pathology (IBPM),  
CNR, Sapienza University, Rome, Italy

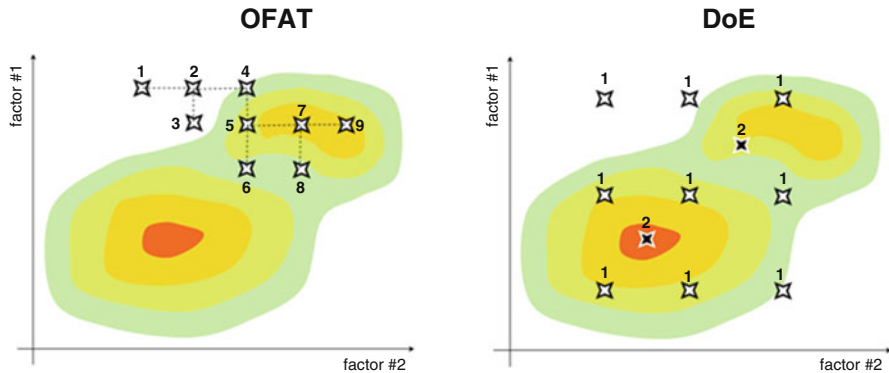
from determination of the best cell media composition to the optimisation of entire multi-step laboratory protocols such as cell transfection.

Our case study is the optimisation of a transfection protocol for neural progenitor cell lines. These cells are very hard to transfect and are refractory to lipidic reagents, so we decided to set-up a protocol based on the non-lipidic Polyethylenimine (PEI) reagent. However, the effect of PEI toxicity on cells has to be correctly evaluated in the experimental design, since it can affect output computation. For this reason, we decided to apply DoE methodology to investigate the effect of PEI, both concentration and type, on cell viability and its interaction with other factors, such as DNA and cell density. The statistics-based DoE approach allowed us to express analytically the neural cell viability dependence on PEI amount/cell and efficiently identify the dose levels of PEI suitable for transfection experiments.

**Keywords** Design of experiments • PEI toxicity • Neural cell transfection  
• Factorial analysis

## 1 Design of Experiments as a Method for Protocol Optimization

Design of experiments (DoE) or experimental design is a methodology whose purpose is planning experiments and analysing their results by optimising the use of resources and time. At the same time, DoE is an effective tool for both maximising the amount of information and minimising the amount of data to be collected. In DoE, a process is seen as an input–output system with a measurable output that depends on the variations of multiple factors. Given a process or system, changes are made to the input variables and the effects on response variables (output) are measured. On this formal basis, factorial experimental designs allow the investigation of the effects of factors by varying them simultaneously instead of changing only one factor at a time (OFAT). OFAT is the most immediate approach to experimentation and is carried out by performing one or more tests for each value (level) of the independent variable (factor), leaving all the other conditions unchanged. The evaluation of the output effects induced by the variation of other factors should thus be obtained by repeating the same type of procedure for every single factor. Moreover, the evaluation of the effect of each factor, in correspondence with a precise combination of all the others, does not consider the interactions between the modelled factors. Therefore, a procedure based uniquely on the OFAT scientific method would omit the study of the effects of contemporary variations of two or more factors. Otherwise, a full OFAT model including all the possible interactions would require an uneven expense of time and resources. For this reason, choosing DoE methodology is the best option to optimise laboratory practices, such as transfection protocols, quickly and efficiently. A figurative representation of the two different methodological approaches is reported in Fig. 1.



**Fig. 1** One factor at a time (OFAT) vs. Design of Experiments (DoE) methodology. Figurative representation of two different approaches to finding the optimal configuration of factors for the same process. System outputs are portrayed with a colour scale (from *white* to *red*) in a continuous bi-dimensional space that is to be explored by experimentation. The four-pointed stars represent all of the attempts made by researchers. Every attempt is a combination of two values, one for each factor, and the ordinal number associated with each attempt refers to the experimental step in which it is made. By varying one factor at a time, a stronger exploitation of attempts, steps and time is needed. Moreover, the results suffer from a higher risk of sub-optimality because of the possibility of arriving at a relative maximum without gaining a general understanding of the system

Candidate factors and their specific levels are selected, depending on the magnitude of their effect on the final result. The possibility of succeeding in arriving at a thorough comprehension of a biological process is strongly connected to the capability of identifying the most influential factors. DoE gives an estimate of the sensitivity of the output as a function of each factor, as well as of the combined effect of two or more factors, in a reduced number of trials (treatments or runs). A treatment corresponds to a determined set of factor levels, and the total number of runs depends on: (1) the experimental design, (2) the number of factors and (3) the replication factor of each experiment.

## 2 Main Aspects of DoE Methodology

DoE is mainly used for:

- Screening many factors and selecting the most relevant ones
- Discovering interactions among factors
- Executing an experiment lowering the risk of biases
- Verifying experimental assumptions and data consistency
- Analysing, interpreting and presenting the results
- Designing statistically robust protocols

- Establishing and maintaining Quality control, possibly by re-iteration of DoE and refinement of the underlying mathematical model

Some principles of DoE are Randomisation, Replication, Blocking, Orthogonality and Factorial experimentation. When applied, these characteristics/principles contribute to improve the robustness of scientific investigation and help researchers in developing experimental settings, so that successive trials can validate previous experimentations in a rigorous way.

Randomisation is accomplished by randomising the testing sequence. In this way, experimental results are protected against biases (e.g. temporal, order, operator-dependent).

Replication is a fundamental operation by which estimation precision is increased and uncontrollable noise is reduced at the same time. Signal-to-noise ratio is augmented by means of a replicate which is a complete repetition of the same experimental treatments and conditions, possibly in a randomised order. The higher costs due to an increase in the number of tests are thus balanced by a more accurate model parameters estimation.

Blocking improves accuracy by removing the effect of known nuisance factors when it is known that identical procedures are applied to each batch. The difference between two procedures is not influenced by the batch-to-batch differences. Blocking is a restriction of complete randomisation that, thanks to the subtraction of batch-to-batch variability from the “experimental error”, increases estimation precision.

Orthogonality is used to generate results whose effects are uncorrelated and therefore can be more easily interpreted. The factors in an orthogonal experiment design are varied independently of each other. This makes it possible to summarise the collected data by taking differences of averages and to show main results graphically by using simple plots of suitably chosen sets of averages.

Finally, factorial experimentation requires that experimental designs include simultaneous, independent and orthogonal variations of all the factors. Since the total number of combinations increases exponentially with the number of factors studied, fractions of the full factorial design can also be constructed. The drawback of a reduction of tests in a fractional factorial design is the possibility of confounding between main effects and factors combinations effects.

Different experimental designs suitable for an experiment are: Plackett–Burman design, Box–Wilson (central composite) designs, Box–Behnken design, Factorial designs, equiradial designs (among them, Dohelert design), mixture designs and combined designs. The full factorial designs [3, 26] allow to estimate primary effects together with the effects of combinations of factors, called interactions, with a limited experimental and statistical complexity. Due to its experimental simplicity coupled with improved statistical efficiency [17], the full factorial design seems to be one of the most eligible approach in biological studies, in which the analysis of interactions between factors (e.g. genetic interactions, protein-protein interactions, gene-protein interactions) is becoming increasingly crucial. In this study we chose a full factorial design, constructing at least a twofold replicate design by keeping at

the same time the number of total runs as reasonably low as possible, without the risk of confounding effects. Non linearities were not studied in this initial phase, so a two-level full factorial approach was used.

In a two-level factorial screening experiment, every chosen factor varies between two levels: qualitative factors have two categorical values (e.g. low/high, A/B, left/right), while quantitative factors vary between two numeric values. Given this design, if  $N$  is the number of candidate factors, the number of requested different runs is  $r = 2^N$ . The number of different runs  $r$  must always be greater than the rank of the design matrix  $X$  that has to be estimated. Finally the number of total runs  $t$  is obtained by multiplying the number of treatments by the replication factor  $k$ , that is the number of times each single treatment is repeated, leading to  $t = k * r = k * 2^N$ . The coefficient  $k$  is 1 if the experiment is not replicated. The treatments are then executed in a randomised order to avoid having uncontrolled variables (i.e. not modelled as factors) contribute to the repeatability variance, affecting the results in a systematic way [3, 17, 26]. This method relies upon the statistical estimation of parameters, which are factors with main effects and factorial interactions. Every parameter is estimated by a mathematical model whose aim is to explain the variability of the output by a combination of factor effects and their interactions, in the form  $y = ax_1 + bx_2 + ..cx_1x_2 \dots$  where  $x_i$  are the modelled factors and  $a$ ,  $b$ ,  $c$  the parameters identified (e.g.  $a$ ,  $b$ : main effects;  $c$ : factors interaction). The accuracy of parameter estimation is calculated by the coefficient of determination  $R^2$ , that is a measure of the percentage of data variability explained by the model and it is used in the multiple linear regression analysis.

An additional measure is adjusted  $R^2$ , that integrates knowledge on the number of modelled variables into a score for the goodness of fit. Its choice is suggested when two models with a different number of factors are compared.

The appropriateness of the estimated mathematical model can be effectively visualised by means of simple plots that show magnitude, whiteness and distribution of residuals. Residual distribution analysis is a way to visualize the mathematical model's fitness for the system under study. Residuals must be low in magnitude and distributed normally. A model's failure to fit must lead to a redefinition of factor levels or to revision of the design itself. An ANOVA analysis can be associated with the DoE analysis, enriching the statistical evaluation of the experiment.

### 3 DoE Applications from Industry to Biological Research

DoE methodology has been widely used in the field of industrial design for the development of processes in order to improve performance. In this field, the primary objectives of these experiments were to: (1) determine the most influential variables on the response, (2) increase product volume, (3) reduce variability. Factorial design has been used for the optimisation of protocols in a variety of industrial fields including manufacturing [10, 11, 13, 25], as well as for pharmaceutical studies

within the Quality-by-design approach to define the design space for standardised production processes (ICH Q8 2009; [28]).

Recently, DoE has been playing an important role also in scientific research areas such as food science [15], chemistry [4] and engineering [22]. The application of DoE has brought very good results in biological fields, among them chromatography [17], metabolomics [29] and especially cellular biology and tissue engineering [2, 4, 7, 12, 18–21, 23, 24]. Optimising the conditions for a specific process in an OFAT manner is a time consuming operation and does not take into account interdependency between factors, which is likely to play a role in most biological processes. Moreover, since the definition of biological protocols has to deal with different environmental conditions, robust estimations of variable parameters and easy visualisation of results are needed to really understand the biological system under observation.

In cellular biology, DoE-based strategies have been applied to develop and optimise serum-free media for culturing mesenchymal stem cell (MSC) spheroids by systematically evaluating media mixtures and a panel of different components for their effects on cell proliferation [2]. Moreover, a factorial Quality-by-design approach has been combined with other approaches such as high-throughput mRNA profiling of a customised chondrogenesis-related gene set as a tool to study in vitro chondrogenesis of human bone marrow derived MSC. The analysis identified the best of the tested differentiation cocktails [21]. Scientists have taken advantage of DoE methodology not just to screen different components of a culture medium, but also to apply it to optimise an entire protocol, such as specific cell line transfection and protein production, obtaining promising results. It has been shown that DoE significantly improves transfection efficiency by a global economy of materials and time [5]. Transfection is the transient or stable introduction of exogenous molecules and genetic material, DNA or RNA, into cultured mammalian cells and is commonly utilised in biological laboratories to study gene function, modulation of gene expression, biochemical mapping, mutational analysis, and protein production. No single delivery method or transfection reagent can be applied to all types of cells, and neural cells are among the most difficult cells to transfect [9, 20]. Very importantly, cellular cytotoxicity and transfection efficiencies vary dramatically depending on the reagent, protocol and cell type being utilised.

## **4 Our Case Study: The Set-Up and Optimisation of a Transfection Protocol for Neural Progenitor Cells**

A groundbreaking project named Quality and Project Management OpenLab (qPMO), inspired by Quality and Project Management principles, has been implemented by a network of Italian National Research Council (CNR) Institutes with the aim of realising and disseminating within the scientific community an innovative way to plan and organise research activity (<http://quality4lab.cnr.it>) [6].

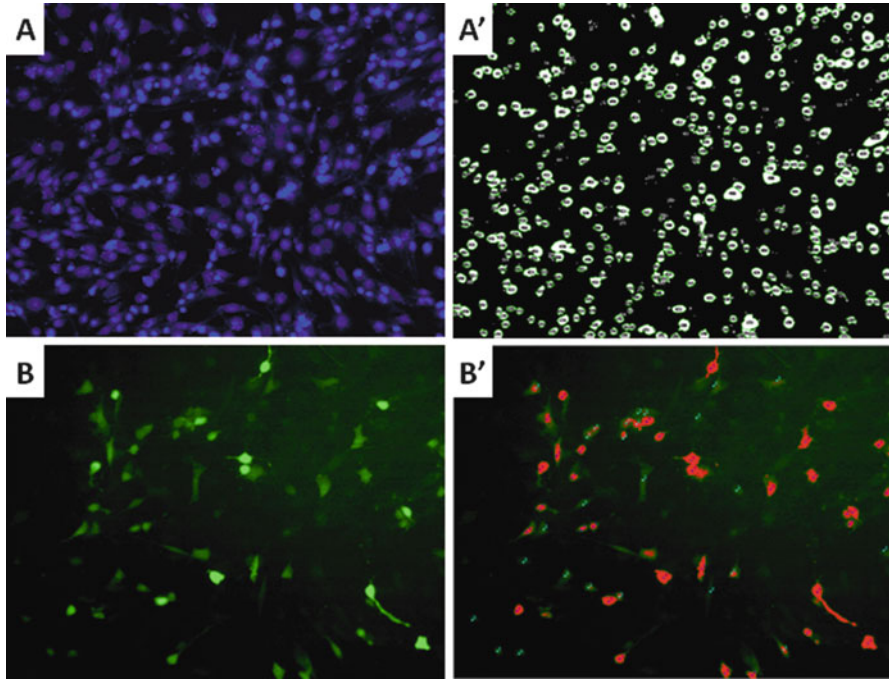
In this context, we selected the DoE model as a very interesting and promising methodology suitable for different kinds of scientific experiments, and we planned

to apply it to both simple assays and medium-high throughput experiments with the final aim of identifying a general guideline for the application of DoE to the set-up and optimisation of scientific protocols.

Our case study is the optimisation of a transfection protocol for neural progenitor for mes-c-myc A1 cell line to obtain a standardised and reproducible laboratory procedure. Mes-c-myc A1 cells are immortalised neural progenitor cells (NPCs) derived from mesencephalon of mouse embryos at 11 days of development which have the characteristics of self-renewal and multipotency [8]. It has been shown that mes-c-myc A1 cells, as is typical of all neural cells, are difficult to transfect and are low-responding to traditional lipidic transfection methods [20]. For this reason, Polyethyleneimine (PEI) was chosen as a transfection reagent because PEI is a cationic non-lipidic transfection reagent normally chosen to achieve higher transfection efficiencies in cell lines that are refractory to liposome-based transfection [19]. A number of PEI molecules have been described in detail with varying molecular size or structure: branched (B) PEI with an average molecular weight of 800 kDa (PEI800) and 25 kDa (PEI25) and a linear (L) form with an average molecular weight of 22 kDa (PEI22) with high transfection activity in vitro and in vivo [28].

Among the factors relevant for transfection efficiency, we selected three quantitative factors: (1) concentration of PEI, (2) DNA amount, (3) cell density and a qualitative one, the type of PEI, L (22 kDa) vs. B (25 kDa). PEI transfectant, as a polyethyleneimine molecule, binds DNA by the presence of nitrogen positive cations (N) in its structure that attract phosphate negative ions (P) of DNA. N/P ratios, depending on PEI amount, have a dramatic impact on transfection efficiency and cytotoxicity [14]. A high concentration of PEI provides a high number of nitrogen cations that can bind DNA efficiently, but which can also be toxic for cells. For this reason, it is important to investigate the correct proportions of PEI and DNA amount, depending of cell line and cell density. Likewise, PEI concentration as well as PEI structure play an important role in both transfection efficiency and cytotoxicity. B conformation of PEI provides a large amount of amine groups that bind DNA more efficiently with respect to the L type. B-PEIs have stronger binding affinity which can condense DNA more efficiently, but they have a less effective release, leading to reduced transfection efficiency. B-PEIs are also more toxic, reducing the viability of cells for transgene expression [19].

To evaluate the percentage of transfected cells, we used ImageJ software to minimise error and variation in downstream analysis [1]. A DNA plasmid containing a green fluorescent protein (GFP) reporter was chosen for transfection to help the output calculation: cells expressing pIRES-EGFP (transfected cells) were visualised directly by fluorescence microscope after PFA 4 % fixation and Hoechst counterstaining for nuclei (all cells). ImageJ java-based image processing software was used for images processing. A plug-in was created to automate and standardize cell counting of total cells (blue labelled nuclei). GFP positive cells (Green labelled,



**Fig. 2** Transfection efficiency calculation through ImageJ software. Transfected Mes-c-myc A1 cells were visualized by fluorescence microscope. (A) Hoechst-labelled mes-c-myc A1 cells (all cells) and (B) cells expressing GFP (transfected cells). (A') a plug-in was created using ImageJ java-based image processing program to automate and standardize cell counting of total cells. GFP positive cells (transfected cells) were counted by summing the number of cells captured by the threshold (*red cells, B'*) and the cells presenting a weak signal undetectable for the threshold calculated by the software (B, B')

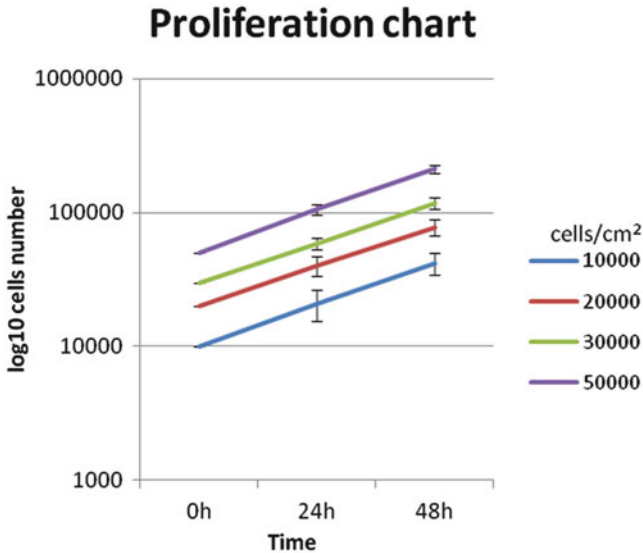
transfected cells) were counted by summing the number of cells captured by the default threshold calculated by the software (red cells) and the cells presenting a weak signal undetectable for the threshold (Fig. 2).

## 5 DoE Methodology Applied to Toxicity Assay

In transfection experiments an important issue to consider is the cell toxicity of the transfectant agent. In our case, for example, a low transfection efficiency could be due to two opposite conditions, an inadequate amount of transfectant and/or DNA or too much transfectant, which can be toxic for the cells.

For this reason, we decided to apply DoE methodology to evaluate cell viability depending on PEI concentration and PEI type in the culture medium, which are the most critical factors, and their interactions with other important factors, such as





**Fig. 3** Proliferation Chart. Proliferation trend of the mes-c-myc-A1 cell line related to the number of cells seeded. The four lines are parallel, indicating that the proliferation rate does not vary with the increase of cell density

DNA concentration and cell density. Before that, we verified that between 10,000 and 50,000 cells/cm<sup>2</sup> the proliferation rate was constant and corresponded to the expected one [8]. In this interval, the cells are in a logarithmic phase of proliferation (Fig. 3), and this behaviour is fundamental for DNA uptake efficiency.

Here we show a flowchart describing every experimental and analytical phase of our DoE approach (Fig. 4). Once the calculation of the output (cell viability) was defined, factors and their levels were chosen, the design of the experiment was generated by means of Minitab<sup>®</sup> Statistical Software and stored in a worksheet. Cells were seeded into 24-well plates (each well is approximately 2 cm<sup>2</sup>) the day before the treatment. The treatment was performed by adding to cells a solution made with an appropriate PEI amount, with and without DNA, according to the design created, in a final volume of 100  $\mu$ l of MEM/F12 culture medium without antibiotic and serum. The incubation lasted overnight at 37 °C in 5 % CO<sub>2</sub>. The next day, culture medium containing the testing amount of PEI and DNA was replaced with fresh complete culture medium. Twenty-four hours after the treatment, viability (output) was calculated as percentage of alive cells respect to the total ones. Alive and dead cells were calculated as follows: culture medium from each condition was collected in different 1.5 ml tubes in order to draw up dead cells in suspension. Next, attached-alive cells were treated with 0.1 % Trypsin for 1 min at 37 °C; thereafter, cells were collected together with their correspondent cells suspension in the same tube. Subsequently, each tube was centrifuged at 900 times gravity for 3 min and the pellet was re-suspended in 200–400  $\mu$ l of culture medium. At the end, 10  $\mu$ l of cell suspension were mixed with 10  $\mu$ l of Trypan Blue (Dye of dead

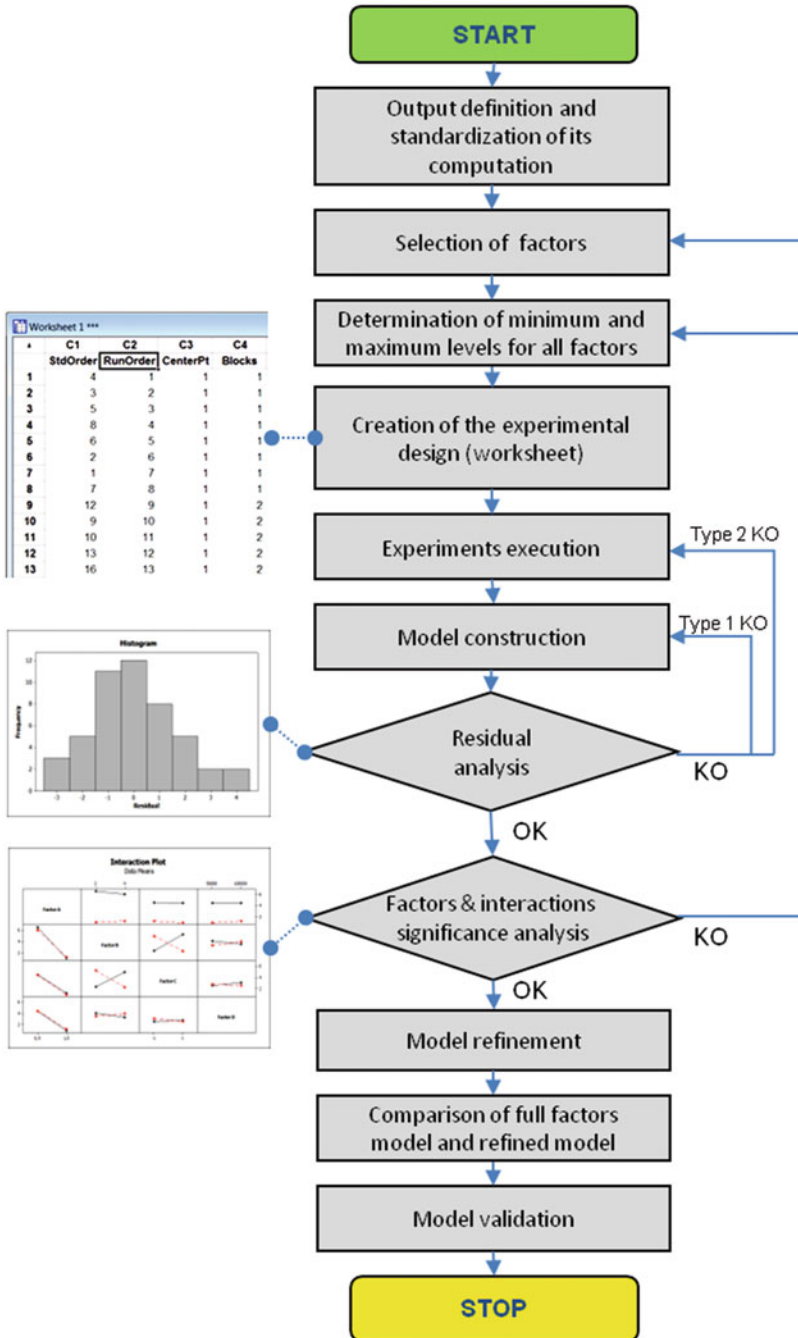


Fig. 4 (continued)

cells) and placed into a haemocytometer. Alive (unstained) and dead cells (stained) were counted under a microscope at 10× magnification. The number of total cells for each well was calculated by summing up alive and dead cells. After collecting data, a linear model was fitted to the data and graphs were generated to evaluate the effects. Residuals plot graphs allowed the evaluation of how accurately the model (model with full factors and interactions) fitted the data. Main Effects and Interaction Effects Plots were analysed to understand the effects on the response of each factor and their interactions. Subsequently the full factor model was compared with a refined model that included only significant factors and interactions.

The model chosen was validated by random experiments using different PEI concentrations and cell densities. After the executions of these experiments, we calculate the amount of PEI/cell used in each condition. Once calculated the output of the random conditions experimented (cell viability) we plotted the results together with the refined model to verify the accuracy of the model constructed by DoE approach.

## 5.1 Factorial Designs and Residual Analysis

Different open source and commercial software are available to create factorial designs and analyse the responses: e.g. Minitab<sup>®</sup> Statistical Software, DOE++, Design-ease, JMP, Develve, MaxStat professional, MacANOVA. Among them we chose Minitab<sup>®</sup> Statistical Software which offers four types of experimental designs: factorial, response surface, mixture and Taguchi (robust). By default, Minitab<sup>®</sup> Statistical Software randomises the run order of all design types. Randomisation helps to ensure that the model meets certain statistical assumptions and allows the reduction of the effects of factors not included in the study. Four factors were put under study: (1) PEI concentration, (2) PEI type, (3) presence or absence of DNA and (4) cell density. We chose a *two level* (each factor varies between two levels) *full* (all combinations are included; no reduction of the design) *factorial design* (all factors are varied at the same time) to analyse the effects of the four factors considered on the output. The execution order of each treatment was performed according to a 'RunOrder' column, which insures the correct randomisation of runs (16 combinations in duplicate). The worksheet and the rough results obtained are shown in Fig. 5.



**Fig. 4** DoE experimental flowchart. Schematic representation of the experimental workflow for DoE statistical analysis. To create the experimental design and analyse the results, Minitab<sup>®</sup> Statistical Software was used. *Rectangles* represent processes; *rhombuses* represent flow checkpoints. Two main checkpoint are considered: Residual Analysis and Factors and interactions significance analysis. KO indicate that checkpoint has not been overcome. Type 1 KO: residuals have not a normal distribution with mean 0. Type 2 KO: residual analysis shows presence of bias or outliers. At the end of the analysis the refined model was validated

**a Two level full factorial design worksheet for Toxicity assay**

| StdOrder | RunOrder | DNA (0,5µg/ml) | PEI type | PEI conc. (mg/L) | Cell Density (cells/cm <sup>2</sup> ) | Viability (%) |
|----------|----------|----------------|----------|------------------|---------------------------------------|---------------|
| 2        | 1        | yes            | B        | 6                | 25000                                 | 85.81         |
| 12       | 2        | yes            | L        | 6                | 50000                                 | 100.00        |
| 1        | 3        | no             | B        | 6                | 25000                                 | 97.73         |
| 18       | 4        | yes            | B        | 6                | 25000                                 | 85.90         |
| 9        | 5        | no             | B        | 6                | 50000                                 | 97.34         |
| 22       | 6        | yes            | B        | 18               | 25000                                 | 5.88          |
| 31       | 7        | no             | L        | 18               | 50000                                 | 61.32         |
| 26       | 8        | yes            | B        | 6                | 50000                                 | 91.36         |
| 32       | 9        | yes            | L        | 18               | 50000                                 | 40.50         |
| 24       | 10       | yes            | L        | 18               | 25000                                 | 15.44         |
| 19       | 11       | no             | L        | 6                | 25000                                 | 100.00        |
| 16       | 12       | yes            | L        | 18               | 50000                                 | 45.42         |
| 10       | 13       | yes            | B        | 6                | 50000                                 | 88.27         |
| 8        | 14       | yes            | L        | 18               | 25000                                 | 25.97         |
| 20       | 15       | yes            | L        | 6                | 25000                                 | 96.94         |
| 3        | 16       | no             | L        | 6                | 25000                                 | 99.48         |
| 30       | 17       | yes            | B        | 18               | 50000                                 | 37.35         |
| 27       | 18       | no             | L        | 6                | 50000                                 | 100.00        |
| 5        | 19       | no             | B        | 18               | 25000                                 | 7.94          |
| 11       | 20       | no             | L        | 6                | 50000                                 | 98.48         |
| 15       | 21       | no             | L        | 18               | 50000                                 | 26.47         |
| 29       | 22       | no             | B        | 18               | 50000                                 | 37.00         |
| 14       | 23       | yes            | B        | 18               | 50000                                 | 42.45         |
| 21       | 24       | no             | B        | 18               | 25000                                 | 0.00          |
| 17       | 25       | no             | B        | 6                | 25000                                 | 100.00        |
| 28       | 26       | yes            | L        | 6                | 50000                                 | 100.00        |
| 7        | 27       | no             | L        | 18               | 25000                                 | 27.60         |
| 23       | 28       | no             | L        | 18               | 25000                                 | 32.52         |
| 4        | 29       | yes            | L        | 6                | 25000                                 | 94.39         |
| 25       | 30       | no             | B        | 6                | 50000                                 | 93.45         |
| 6        | 31       | yes            | B        | 18               | 25000                                 | 4.83          |
| 13       | 32       | no             | B        | 18               | 50000                                 | 26.94         |

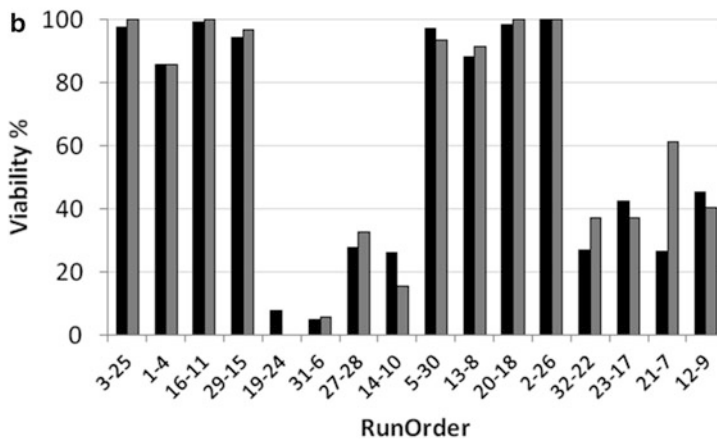


Fig. 5 (continued)

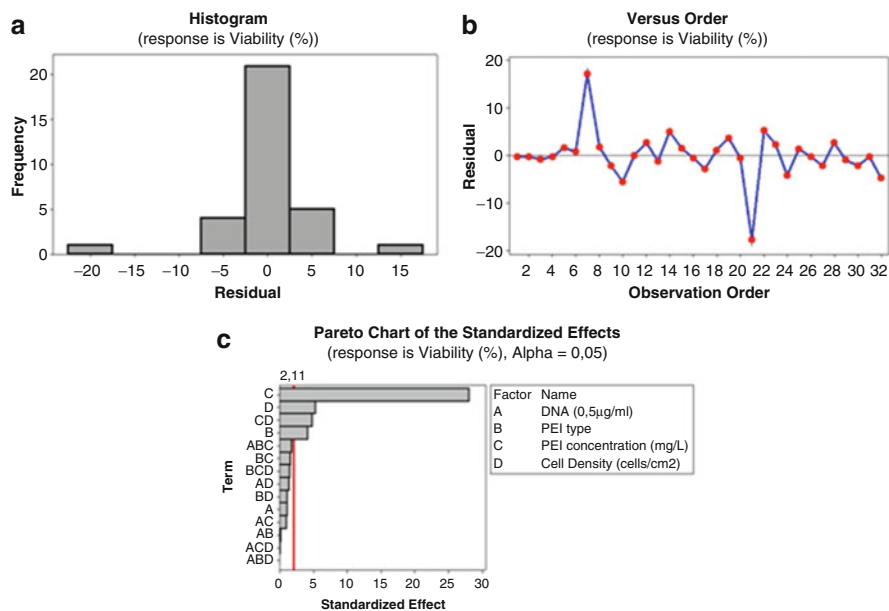
For the statistical analysis of the results, a regressive linear model was fitted to the observations and residuals were analysed (Fig. 6). The distribution appeared to fit the linear model: the Histogram, the graph that correlates the residuals with their frequency, appears to be approximately symmetric and bell-shaped, confirming the normal distribution of residuals (Fig. 6A). The Versus Order plot showed randomly scattered residuals with the absence of significant patterns in the distribution: this demonstrates the time independence of residuals, non-constant variance and missing higher-order terms (Fig. 6B). Once the normality distribution of the residuals was determined, the analysis of factor interactions was performed.

## 5.2 Factors and Interactions Significance Analysis

To verify the significance of the factors and their interactions, a Pareto Chart was generated in which any effect extending beyond the reference red line was significant at the default level of 0.05 (Fig. 6C). As shown by the graphs, three different factors were found to be significant (PEI type, PEI concentration and cell density) and only one two-factor interaction (between PEI concentration and cell density). The presence of DNA did not influence the response. This observation might support the idea that negatively charged DNA molecules do not balance positive charges of PEI transfectant, affecting toxicity in some way.

To interpret the results, Main Effects Plots and an Interactions Plot were analysed (Fig. 7). A main Effects Plot shows the one-factor effect called main effect (Fig. 7A). The horizontal line, corresponding to about 60 (60 % of cell viability), represents the mean of the response of all the runs. The line for DNA confirmed what was shown in the Pareto Chart, that is, in both conditions of absence (no) or presence (yes) of DNA the mean of all conditions was approximately 60 %, resulting in a line with slope close to 0 that indicates the non-influence of the factor. The most important factor was PEI amount. The effect of this factor was the line with the highest slope. In detail, all runs in which 6 mg/L of PEI transfectant were utilised showed a higher cell viability (close to 100 %, total cell viability) with respect to the condition with 18 mg/L. PEI type and cell density have a similarly low slope, indicating that these factors had a comparable small effect on the response. The PEI type plot shows that the mean output of all the runs performed with L-PEI is higher if compared to the one of all the runs performed with B-PEI, confirming the lower

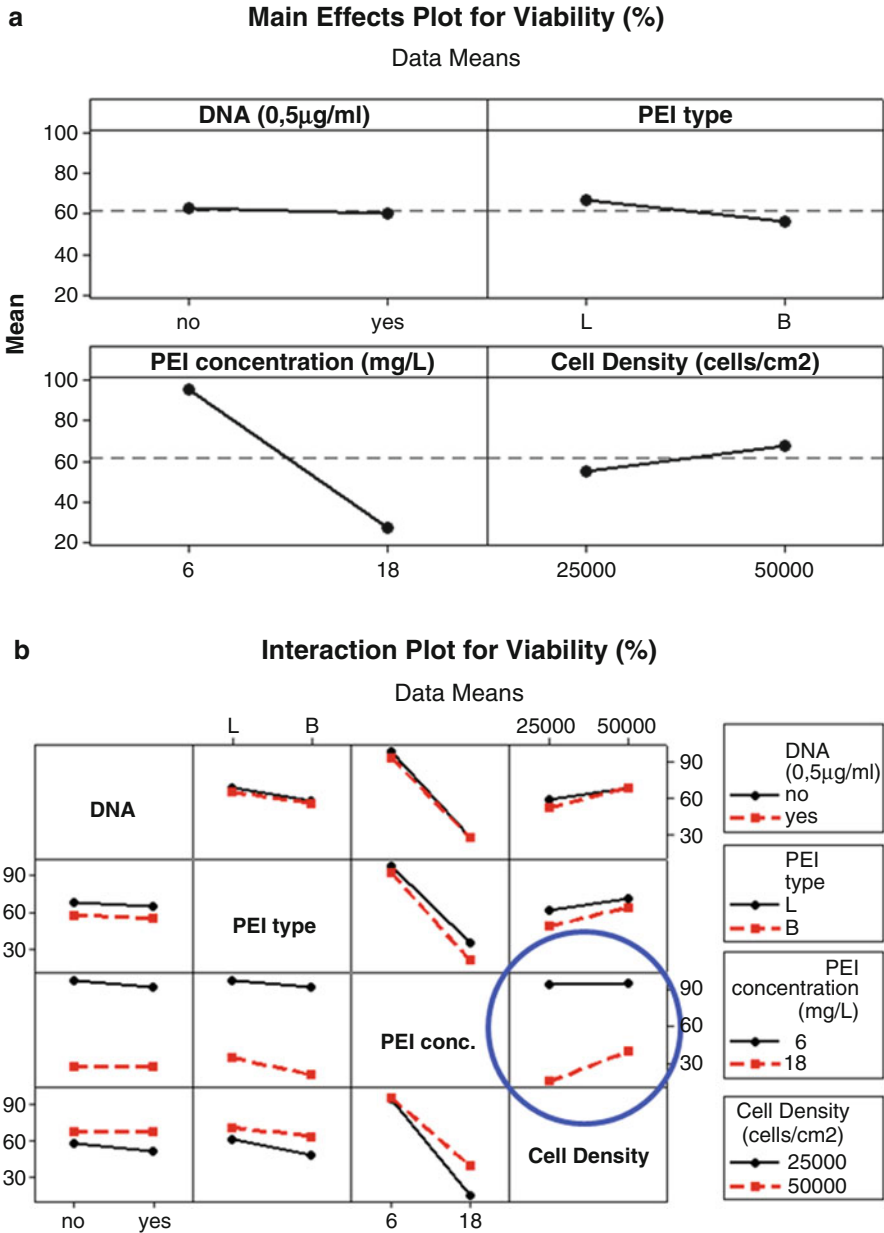
←  
**Fig. 5** Toxicity assay performed with a two-level full factorial experimental design. (A) Worksheet created by Minitab® Statistical Software representing combinations of tested factors and levels. Standard order indicates the order in which combinations are generated according to the design chosen. Run order corresponds to the randomisation of generated combinations. DNA, PEI type, PEI concentration and Cell density are the factors analysed. Viability represents the calculated output. (B) Graphical representation of the results, in which the two series represent the two replicates (*black*, replicate 1; *gray*, replicate 2)



**Fig. 6** Residual and distribution analysis for Toxicity assay. **(A)** Histogram shows that the distribution of the residuals in the experiment is normal. **(B)** Versus order plot illustrates that observation order of the residual is well randomised, due to the absence of a repetitive pattern in the plot. **(C)** Pareto Chart of the Standardized Effects graphically represents the significant factors and interactions. Any significant factor or interaction is characterised by columns that extend beyond the *red line*; the greater the distance from the line, the higher the influence of that factor. Significant factors and interactions were, in order: PEI concentration (factor C), cell density (factor D), interaction between PEI concentration and cell density (CD) and PEI type (factor B)

toxicity of linear conformation of the PEI molecule with respect to the branched one [19]. In the cell density plot, all runs with the factor of 25,000 cells/cm<sup>2</sup> showed a lower cell viability with respect to runs with 50,000 cells/cm<sup>2</sup>, demonstrating cell viability depended on the amount of transfectant per cell. This analysis confirmed PEI concentration as the most critical parameter and identified PEI type and cell density as influencing factors.

To identify significant interactions among factors, the Interactions Plot showing the effect of multiple factors was also analysed (Fig. 7B). Evaluating interactions is extremely important because an interaction can magnify or diminish main effects. All the interactions were not significant because the two lines had the same slope, except for the plot showing the interaction between PEI (mean of the values of both L and B) concentration and cell density. In this case, the two different levels of PEI concentration exhibited two different behaviours: at the lower PEI concentration (black line), cell viability was always approximately 100 % for both cell density conditions, while with the higher PEI concentration (red line), cell viability was higher when cell density was 50,000 cells/cm<sup>2</sup> (around 40–50 %) and



**Fig. 7** Factorial analysis for Toxicity assay. (A) The Main effect plot shows the mean of outputs in correspondence with the different levels of each of the four factors. The dotted line corresponds to a viability of 60 %. The plot shows that PEI concentration is the factor with the strongest effect on cell viability, due to the higher slope of the line connecting the two analysed levels. (B) The Interaction plot is represented by a matrix showing interactions among factors. Similar slope of the lines indicates no significant interaction. The only significant interaction is the one between PEI concentration and cell density, highlighted in the circle

lower when cell density was 25,000 cells/cm<sup>2</sup> (approximately 0 %, corresponding to full mortality). These data show that the minimum amount of PEI used (6 mg/L) never affected cell viability, while the maximum amount (18 mg/L) decreased cell viability between 0 and 50 %, depending on cell density.

### 5.3 *Model Refining and Validation*

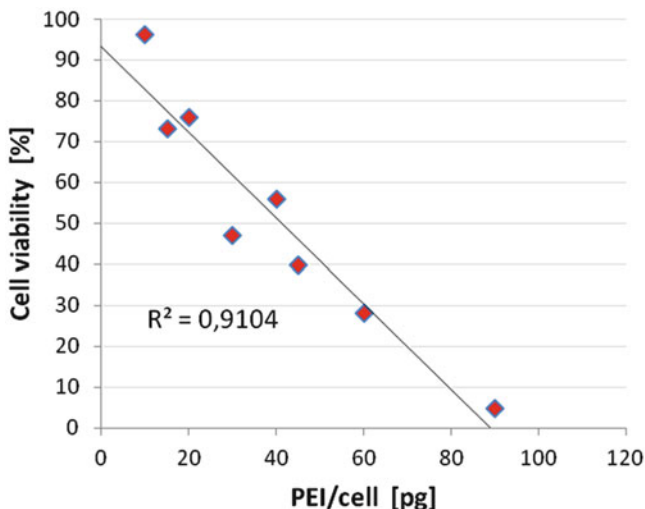
A refined model was constructed discarding all the not significant main effects and interactions: only the significant factors PEI type (B), PEI concentration (C), cells density (D) and the interaction between these last two factors (C\*D) were saved for the model refinement. Subsequently, the full factors model and the refined model were compared. Since full factors model had given as a fitness measure R<sup>2</sup> adjusted of 96.24 % and the refined model gave R<sup>2</sup> adjusted of 95.96 %, the loss of explained variance was considered minimal and the refined model was chosen for further investigations. The analysis of R<sup>2</sup>, including the adjustment of extra explanatory variables, gave us additional evidence to exclude from the analysis not significant factors, such as DNA amount, and irrelevant factors interactions.

## 6 **Conclusions**

The DoE approach applied to our Toxicity assay let us clearly determine which factors most influenced the output of the process under study: PEI concentration and cell density. Moreover, the simultaneous variation of these factors and the subsequent statistical analysis let us identify a significant interaction between PEI concentration and cell density, unmasking the real significant factor influencing cell viability, that is PEI amount per cell. By relating cell viability measured in the Toxicity assay to the amount of PEI, both B and L, per cell tested (PEI pg/cell) we could finally generate a refined model (Fig. 8). To validate this model, we determined cell viability at different levels of PEI pg/cell in the examined interval. All the conditions tested were reasonably close to the line, indicating that the model generated by the factorial analysis had good accuracy and could be used to predict cell viability variation connected to the dose of PEI per cell. The Toxicity assay let us select the upper value of PEI pg/cell to be tested in following transfection experiments, in order to obtain the maximum transfection efficiency avoiding extreme cell death. Thus, we decided to set the maximum level at a concentration of PEI in the culture medium of 12 mg/L, corresponding to 30 PEI pg/cell (cell density 50,000 cells/cm<sup>2</sup>), which would not reduce cell viability more than 50 %.

In summary, taking advantage of statistics-based factorial experimental design we could express analytically mes-c-myc A1 cell viability dependence on PEI amount per cell. Application of DoE allowed to determine the maximum amount





**Fig. 8** Dependence of cell viability on PEI amount/cell. A polynomial model was fitted with Toxicity assay viability results,  $y = -1.05x + 93.19$  where  $x$ : PEI/cell,  $y$ : viability ( $R^2 = 0.9104$ ). *Red rhombuses* correspond to experimental conditions chosen to test the factorial analysis

of PEI usable in mes-c-myc A1 cells transfection experiments coupled to the higher cell viability. The main advantages of using DoE were saving time and resources for the complete experimental plan (leading to efficiency), the evaluation of both main and interaction effects of the selected factors in an easy graphical way, and statistical information that allows data to be robust and reliable (both of which lead to effectiveness). Nowadays, with an increasing scientific competition asking researchers to produce in a short time reliable and reproducible results, our data support the application of DoE to scientific studies for obtaining the best results and optimising the use of resources.

**Acknowledgments** We would like to thank Umberto di Porzio and Giancarlo Bellenchi (Institute of Genetics and Biophysics 'Adriano Buzzati-Traverso') for helpful suggestions and for providing us with the mes-c-myc A1 cells, Salvatore Pulcrano and Valerio Piscopo (Institute of Genetics and Biophysics 'Adriano Buzzati-Traverso') for providing A1 cell culturing protocols, Genesia Manganelli and Emilia Giorgio (Institute of Genetics and Biophysics 'Adriano Buzzati-Traverso') for helping set up the preliminary assays, and Teresa Nutile (Institute of Genetics and Biophysics 'Adriano Buzzati-Traverso') for her statistical support. Portions of information contained in this work are printed with permission of Minitab Inc. All such material remains the exclusive property and copyright of Minitab Inc. All rights reserved. We also thank Richard E. Burket for editing and English revision.

This work was supported by grants from the Ministero dell'Economia (Ministry of Economics and Finance in Italy, CNR FaReBio di Qualità, qPMO Project), the Ministero Istruzione Università Ricerca (Medical Research in Italy RBNE08LN4P\_002) and 'Fondazione con il Sud' (2011-PDR-13) to G.L.L.

## References

1. Abramoff, M.D., Magalhaes, P.J., Ram, S.J.: Image processing with ImageJ. *Biophoton. Int.* **11**, 36–42 (2004)
2. Alimperti, S., Lei, P., Wen, Y., Tian, J., Campbell, A.M.: Serum-free spheroid suspension culture maintains mesenchymal stem cell proliferation and differentiation potential. *Biotechnol. Prog.* **30**, 974–983 (2014)
3. Anderson, M.J., Whitcomb, P.J.: DOE Simplified. Practical tools for effective experimentation. Productivity Press, New York (2007)
4. Bezerra, M.A., Santelli, R.E., Oliveira, E.P., Villar, L.S., Escalera, L.A.: Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta* **76**, 965–977 (2008)
5. Bollin, F., Dechavanne, V., Chevalet, L.: Design of experiment in CHO and HEK transient transfection condition optimization. *Protein Expr. Purif.* **78**, 61–68 (2011)
6. Bongiovanni, A., Colotti, G., Liguori, G.L., Di Carlo, M., Digilio, F.A., Lacerra, G., Mascia, A., Cirafici, A.M., Barra, A., Lanati, A., Kisslinger, A.: Applying quality principles and project management methodologies in biomedical research: a public research network's case study. *Accred. Qual. Assur.* **20**, 203–213 (2015)
7. Chen, Y., Bloemen, V., Impens, S., Moesen, M., Luyten, F.P., Schrooten, J.: Characterization and optimization of cell seeding in scaffolds by factorial design: quality by design approach for skeletal tissue engineering. *Tissue Eng. Part C Methods* **17**, 1211–1221 (2011)
8. Colucci-D'Amato, G.L., Tino, A., Pernas-Alonso, R., French-Mullen, J.M., di Porzio, U.: Neuronal and glial properties coexist in a novel mouse CNS immortalized cell line. *Exp. Cell Res.* **252**, 383–391 (1999)
9. Dalby, B., Cates, S., Harris, A., Ohki, E.C., Tilkins, M.L., Price, P.J., Ciccarone V.C.: Advanced transfection with Lipofectamine 2000 reagent: primary neurons, siRNA, and high-throughput applications. *Methods.* **33**, 95–103 (2004)
10. Dhas, E.R., Dhas, J.H.: A review on optimization of welding process. *Procedia Eng.* **38**, 544–554 (2012)
11. Dirion, B., Cobb, Z., Schillinger, E., Andersson, L.I., Sellergren, B.: Water-compatible molecularly imprinted polymers obtained via high-throughput synthesis and experimental design. *J. Am. Chem. Soc.* **125**, 15101–15109 (2003)
12. Enochson, L., Brittberg, M., Lindahl, A.: Optimization of a chondrogenic medium through the use of factorial design of experiments. *Biores Open Access* **1**, 306–313 (2012)
13. Fei, N.C., Mehat, N.M., Kamaruddin, S.: Practical applications of Taguchi method for optimization of processing parameters for plastic injection moulding: a retrospective review. *ISRN Ind. Eng.* **2013**, 1–11 (2013)
14. Florea, B.I., Meaney, C., Junginger, H.E., Borchard, G.: Transfection efficiency and toxicity of polyethylenimine in differentiated Calu-3 and nondifferentiated COS-1 cell cultures. *AAPS PharmSci.* **4**, E12 (2002)
15. Gacula, M.C.: The design of experiments for shelf life study. *J. Food Sci.* **40**, 399–403 (2006)
16. Green, S., Liu, P.Y., O'Sullivan, J.: Factorial design considerations. *JCO* **16**, 3424–3430 (2002)
17. Hibbert, D.B.: Experimental design in chromatography: a tutorial review. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **910**, 2–13 (2012)
18. Hubert, C., Lebrun, P., Houari, S., Ziemons, E., Rozet, E., Hubert, P.: Improvement of a stability-indicating method by Quality-by-Design versus Quality-by-Testing: a case of a learning process. *J. Pharm. Biomed. Anal.* **88**, 401–409 (2014)
19. Hsu, C.Y., Uludağ, H.: A simple and rapid nonviral approach to efficiently transfect primary tissue-derived cells using polyethylenimine. *Nat. Protoc.* **7**, 935–945 (2012)
20. Karra, D., Dahm, R.: Transfection techniques for neuronal cells. *J. Neurosci.* **30**, 6171–6177 (2010)
21. Knöspel, F., Schindler, R.K., Lübberstedt, M., Petzolt, S., Gerlach, J.C., Zeilinger, K.: Optimization of a serum-free culture medium for mouse embryonic stem cells using design of experiments (DoE) methodology. *Cytotechnology* **62**, 557–571 (2010)

22. Iizarbe, J., Alvarez, M.J., Viles, E., Tanco, M.: Practical applications of design of experiments in the field of engineering: a bibliographical review. *Qual. Reliab. Eng. Int.* **24**, 417–428 (2008)
23. Jakobsen, R.B., Østrup, E., Zhang, X., Mikkelsen, T.S., Brinchmann, J.E.: Analysis of the effects of five factors relevant to in vitro chondrogenesis of human mesenchymal stem cells using factorial design and high throughput mRNA-profiling. *PLoS One* **9**, e96615 (2014)
24. Liu, G., Kawaguchi, H., Ogasawara, T., Asawa, Y., Kishimoto, J., Takahashi, T., Chung, U.I., Yamaoka, H., Asato, H., Nakamura, K., Takato, T., Hoshi, K.: Optimal combination of soluble factors for tissue engineering of permanent cartilage from cultured human chondrocytes. *J. Biol. Chem.* **282**, 20407–20415 (2007)
25. Mandenius, C.F., Brundin, A.: Biocatalists and bioreactor design—bioprocess optimization using design-of-experiments methodology. *Biotechnol. Prog.* **24**, 1191–1203 (2008)
26. Montgomery, D.C.: *Design and Analysis of Experiments*. Wiley, New York (1997)
27. Weissman, S.A., Anderson, N.G.: *Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications*. Organic Process Research & Development. ACS Publications (2014)
28. Wightman, L., Kircheis, R., Rössler, V., Carotta, S., Ruzicka, R., Kursá, M., Wagner, E.: Different behavior of branched and linear polyethylenimine for gene delivery in vitro and in vivo. *J. Gene Med.* **3**(4), 362–372 (2001)
29. Zheng, H., Clausen, M.R., Dalsgaard, T.K., Mortensen, G., Bertram, H.C.: Time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches. *Anal. Chem.* **85**, 7109–7116 (2013)

# A Design of Experiment Approach to Optimize an Image Analysis Protocol for Drug Screening

Antonella Lanati, Cecilia Poli, Massimo Imberti, Andrea Menegon,  
and Fabio Grohovaz

**Abstract** The Design of Experiment, DoE, was applied to support the development of an innovative optical platform for ion channel drug screening. In this work, DoE was exploited to investigate a set of software parameters instead of process variables, an approach that has been only rarely explored. In particular, it was used to define a standard analytical configuration for a MatLab-based image analysis software that has been developed in the laboratory to extract information from images acquired under the drug screening conditions. Since the choice of the type of analysis and filtering, as well as their interactions, was known to affect the final result, the aim was to identify a robust set of conditions in order to obtain reliable concentration-response (sigmoidal) curves in an automated way. We considered five parameters as factors (all qualitative) and two characteristics of the sigmoidal curve as reference outputs. A first DoE screening was performed to reduce the number of needed levels for one factor (an unconventional approach) and a second optimization study to define the best configuration setting. Image stacks from three different experimentation days were used for the analysis and modelled by blocks to investigate inter-day variations. The optimized set of parameters identified in this way was successfully validated on different cell lines exposed to their references drugs. Thanks to this study, we were able to: find the optimized configuration for the analysis, with a reduced number of trials compared to the classical “One Variable at A Time” approach; acquire information about the interactions of different analytical conditions as well as the inter-day influence; and, finally, obtain statistical evidence to make results more robust.

---

A. Lanati (✉) • C. Poli  
Valore Qualità, 27100 Pavia, Italy  
e-mail: [alanati@valorequalita.eu](mailto:alanati@valorequalita.eu)

M. Imberti  
Open Sistemi, 26100 Cremona, Italy

A. Menegon  
San Raffaele Scientific Institute, 20132 Milan, Italy

F. Grohovaz (✉)  
San Raffaele Scientific Institute, 20132 Milan, Italy

Vita-Salute San Raffaele University, 20132 Milan, Italy  
e-mail: [grohovaz.fabio@hsr.it](mailto:grohovaz.fabio@hsr.it)

**Keywords** Design of Experiment • Ion Channels • Drug Screening • Image analysis

## 1 Introduction

### 1.1 Use of Design of Experiment in Research

The design of experiment (DOE), also known as experimental design, was developed by Ronald Fisher in the early 1920s. It is a statistical method aimed to plan and analyze experiments, in order to extract the maximum amount of information with the fewest possible number of runs. It allows also to build regression models and to optimize the output by choosing proper variable settings. The traditional way of conducting experiments is intended to change One Variable at A Time (OVAT), thus accepting the risks to become trapped in a local optimum, missing the global optimum. The DoE allows changing simultaneously all the variables, helping in finding their best combination [6, 12, 14]. Moreover, it provides a regression model that, in the range of variables used to build it, can make predictions for values different from those used in the study (see: [1, 18]).

During the last decades, DOE has been successfully applied to optimize processes in chemistry and engineering [13] as well as in pharmaceutical and biopharmaceutical industry, both in development and production [2, 7, 9–11, 14, 22, 25, 26]. New applications are emerging in biomedical research, specifically in medium-high throughput assays and in the optimization of laboratory protocols [3, 4]. In particular, DoE has recently been used in drug screening, where progress in molecular biology and advanced technologies has given new opportunities to test large chemical libraries against biological targets. However, the introduction of combinatorial chemistry and high-throughput screening has not met the expectations, rather it has been accompanied by a decline in productivity [20]. This can be ascribed to a number of reasons, including the fact that the process of selection leaves behind many potentially interesting molecules [16, 23]. This has drawn the attention to cell-based assays and to more robust screening approaches in order to increase R&D efficiency/efficacy, and thus productivity. In this respect, attention has also grown toward methods for an efficient development and setup of the assays. In the last decades, there are several examples of the application of DoE in drug screening [5, 11, 15, 17, 24], mainly related to the optimization of biological and biochemical process condition. Other examples are oriented to optimization of data processing in metabolomics [8, 27]. The use of experimental design for optimizing software parameters is still poorly explored (e.g. [21]).

In this paper, we report the use of DoE for the fine set up of the analytical processing in a newly developed drug screening approach.

## 1.2 *Dedicated Image Analysis Software for a New Drug Screening Approach*

An innovative optical platform for ion channel drug screening, based on a proprietary approach, has been developed by a multidisciplinary team. Briefly, cells expressing the channel of interest are loaded with a fluorescent voltage-sensitive dye and the effect of a drug is revealed by the fluorescence values recorded before and during exposure to electrical stimulation (see EP2457088 patent for more details). Images acquired under the above conditions are processed by a MatLab-based Image Analysis (MaLIA; a program developed in the laboratory) that offers the possibility to employ different filters, parameters and types of analysis. Data representative of the cellular response to the electric pulses are used to extrapolate changes in resistance/conductance; these values are put in relation with increasing concentrations of the molecule of interest, thereby obtaining a typical sigmoidal Concentration-Response (CR) curve defined by a set of qualitative and quantitative parameters. In the course of the project, the MaLIA has progressively evolved, gaining in flexibility, to explore multiple analytical options. In this development phase, different analysis configurations were experimented: the parameter space was narrowed to a set of five, four of which varying between two values only. The final goal of the project was to define the optimized values of these parameters, in order to perform a standard analysis in full automation, without external, arbitrary interventions. To this end, we employed the DoE to evaluate the effects of different parameters/filters implemented in the MaLIA as well as their interactions.

## 2 The Design of Experiment (DoE) Method

The different values assumed by each *factor* (the experimental variables) are called *levels* (typically only two, codified as  $-1$  or  $+1$ ) and can be either qualitative or quantitative. The DoE allows evaluating both the influence of single variables (*main factors*) and the interplay among factors (*interactions*), i.e. when the effect of a factor depends on the level of one or more other factors. A specific combination of levels is called *treatment* (or *run*). Each treatment is evaluated in terms of *outputs* or *responses*, which are representative of the behaviour of the system. The magnitude of a change in response, when factors are varied, is called *effect*.

In order to achieve statistically relevant conclusions from experiments, it is necessary to adopt different statistical principles: *randomization* (i.e. scrambling the running order of treatments), *replication* (i.e. repeating each treatment twice or more) and *blocking* (i.e. modelling extraneous sources of variation as special variables). These principles minimize experimental bias that may mask the responses of the significant factors (see [18]).

The term *factorial design* identifies the most used set of treatments employed to investigate the effects of factors on a response. The simplest factorial design is called *two-level factorial design*, and is used in two forms: *full factorial* and *fractional factorial design*. Full factorial design requires  $2^K$  runs, where  $K$  is the number of factors, and generates, with the increasing number of factors, a considerable (even unmanageable) amount of runs. In case of many factors (e.g.  $>5$ ), we can reduce the number of requested runs, based on specific assumptions, such as ignoring interaction of more than three factors: in this way we perform a fractional factorial design. Experimental designs reduced to  $2^{K-1}$  runs are called Resolution V designs. In these designs, no main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions. There is also the possibility to consider more than two levels for each factor, and to create a *general factorial design* [1].

After performing experiments according to the planned design, results are analyzed through a graphical interpretation (*Factorial Plots* and *Statistical Plots*) and a set of statistical parameters. The Factorial Plots include the *Main Effect Plot* and the *Interactions Plot*. The former is represented as a straight line: the slope indicates the direction while its magnitude the strength of the effect. On the other hand, the *Interactions Plot* shows how different combinations of factor settings affect the response: non-parallel lines show interaction between couple of factors. The *Normal Probability Plot* (one of the *Statistical Plots*) is a different representation of a distribution, with the cumulative percentage on the logarithmic Y-axis and the ordered values of the observations on the X-axis. In this representation, the Gaussian distribution appears as a straight line. It is used to check normality of the data and to find out the most significant ones: non-significant data are dispersed along a straight line, whereas significant data are apart. In the experimental design, the Normal Probability Plot is used to evaluate significance and normality of both main and interaction effects. The *Pareto Plot* (another *Statistical Plot*) displays the absolute values of main and interaction effects: a reference line shows statistically significant values ( $P < 0.05$ ). The *Normal Plot for Residuals* is conceptually the same as the one used for effects and interactions and estimates the difference (residuals) between actual and predicted values (calculated by the regression model obtained from the DoE analysis), to verify whether the data have a Gaussian distribution.

This analysis can be complemented by a number of statistical parameters, including a regression model describing each response as a function of the selected factors and information coming from the ANOVA analysis (see: [18]).

As a general approach, a *screening analysis* is first performed with less stringent conditions to identify the most significant factors. Subsequently, an *optimization analysis* is applied to a narrower set of factors to find the best condition that optimizes the output(s).

Along with the factorial designs, DoE offers a rich set of other designs, to suit most requirements. Few examples are:

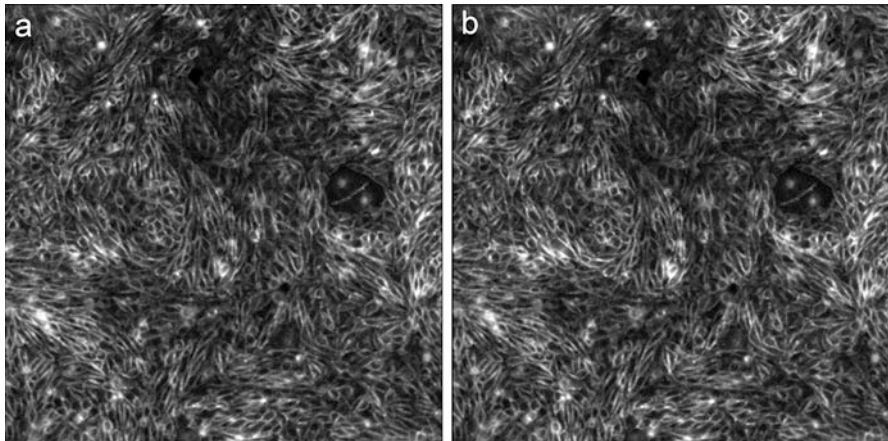
1. *Plackett–Burman* design, which evaluates the effects of main factors only, with a small set of runs. It is mainly used in the screening phase;

2. *Response surface* designs (e.g. Central Composite, Box–Behnken), which are used to identify points of absolute maximum, and to highlight possible nonlinearities (for quantitative factors only). They are mainly used in the optimization phase.
3. *Mixture design*, which is used when factors are components of a blend, and the output depends on their relative proportion.

### 3 Experimental Setup

Experiments were performed on a Chinese hamster ovary (CHO) cell line expressing the human transient receptor potential (TRPV1) channel (kindly provided by Axxam S.p.A) using capsaicin as reference agonist. CHO-TRPV1 cells were stained with a voltage sensitive dye (VSD; di-4-ANEPPS), and exposed to a square electric pulse. Local fluorescence values were measured before and during the pulse (Fig. 1, left and right, respectively) both in the absence and in the presence of capsaicin. As expected from the poor sensitivity of the VSD (8 % fluorescence variation/100 mV), changes are hardly appreciated at first sight and a sophisticated analysis is necessary to automatically isolate and evaluate subcellular responsive areas. Further details are available on the patent EP2457088 and will be reported in a full paper on this new approach (Menegon et al. in preparation).

Among the different types of DoE designs, we decided to use factorial designs for two reasons. On the one hand, we needed to evaluate second order interactions



**Fig. 1** CHO-TRPV1 cells images before (a) and during (b) exposure to an electrical square pulse. The signal (differences in fluorescence intensity in specific subcellular regions) is not easily appreciable without proper data processing



and Plackett–Burman was not suitable. On the other hand, our factors were typically at two levels, making inappropriate other analyses such as response surface designs.

The very same stack of images was processed many times with MaLIA, to cover all the combinations of parameters indicated by the experimental design. Randomization was not required, because no external bias factors could affect the running of the software analysis. For DoE analysis, we selected, among the parameters implemented in MaLIA, the following five factors (variables) that appeared to influence the output data:

- a. Binning: to reduce image noise by combining cluster of pixels into single pixels;
- b. Shape-mask (ShapeM): to select the membrane responsive areas;
- c. Minimum responses filtering (MinRespFilt): to discard signal values lying inside the noise range;
- d. Response calculation (RespCalc): Fold Change (FC) or Normalized Fold Change (NFC);
- e. Output data filtering (OutputDataFilt): pure statistical or functional (to exclude variations not coherent with the expected biological response)

We defined also two outputs to evaluate the influence of these parameters on CR curves:

1. R-squared (Rsq), as a measure of good fitting of the sigmoidal curve;
2. Top minus bottom (T-B), as the difference between highest and lowest values in the sigmoidal curve (a measure of the efficacy of tested drug).

Finally, in order to account for possible inter-day variations (due to biological variability and/or changes in the process), we repeated the same set of treatments on image stacks obtained in three different experimental days, and modelled each of these replications by blocks.

The MaLIA parameters are qualitative and at two levels only, with the sole exception of Binning that has three possible levels: for a full evaluation, a general factorial design with five factors should be employed. According to Anderson and Whitcomb (see Chap. 7, pp. 133–134): (1) a general factorial design is to be avoided when the number of factors increases (typically higher than 3), (2) a reduction of a general factorial design requires ad hoc elaboration. The same authors suggest making preliminary tests to attempt to reduce the analysis to a two-level factorial. In our case, a complete general factorial design (5 factors, one of them at 3 levels) would require  $2^{(5-1)} \times 3 = 48$  runs per replicate that, multiplied by the 3 foreseen replicates, give a total of 144 runs. As expected, the DoE software we use does not allow for reducing general factorial designs. In line with the suggestions of Anderson and Whitcomb, we evaluated the possibility to reduce the number of levels for Binning. Therefore, we first set an unconventional screening analysis, by considering the sole two factors directly involved in the extraction of data from images: Binning (three levels) and Shape Mask (two levels), by using a standard set of the other parameters. Thanks to the reduction of Binning to two levels, in the second analysis we were able to evaluate all factors at two levels with a fractional factorial design. In this way, it was possible to perform the analysis with only 6 + 16

runs for each replicate. Finally, we made a validation to verify: (1) that there is no significant interaction between Binning and the other factors; and (2) that the discarded Binning level was less suitable for optimal results.

We generated Normal Plot and Pareto Plot, to identify statistically significant factors and interactions, as well as the Main Effect Plot and the Interaction Plot to evaluate factors influence on each output. Goodness of fit was judged by the Residuals Plots and other statistical parameters. Blocks provided information about the influence of different experimental days (inter-day variations). All the DoE analysis was performed by Minitab, a statistics package developed at the Pennsylvania State University (Minitab Inc., State College, PA, USA).

## 4 Results of the First Analysis (Screening)

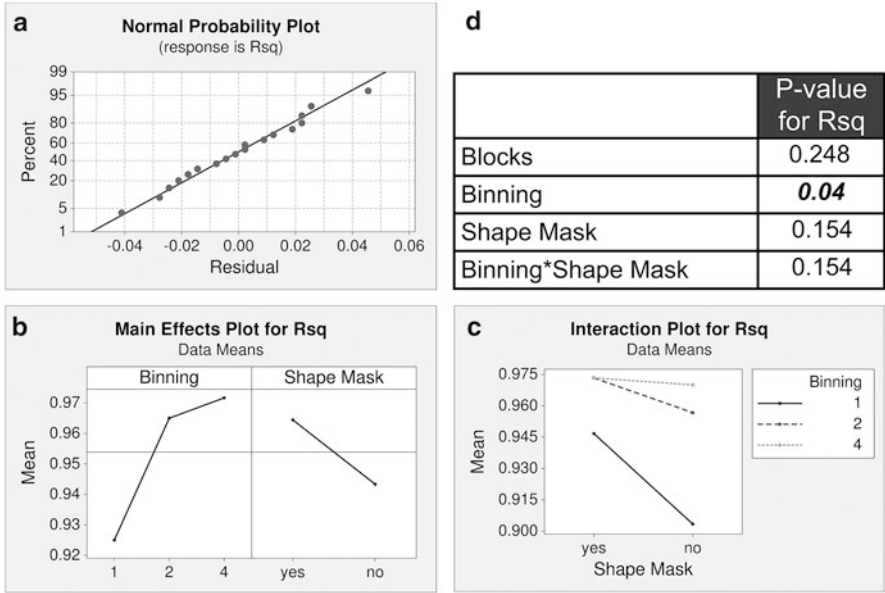
The first analysis was aimed to find the two most significant values out of the three possible levels of image Binning and was performed considering the Shape-Mask as the sole factor able to interact significantly with the Binning. In fact, only these two variables are directly related to the pixels of the image. Factors and their levels were as follows:

1. Binning ( $1 \times 1$ , i.e. no binning;  $2 \times 2$ ;  $4 \times 4$ ; referred to as 1, 2 and 4, respectively);
2. Shape-mask (yes; no).

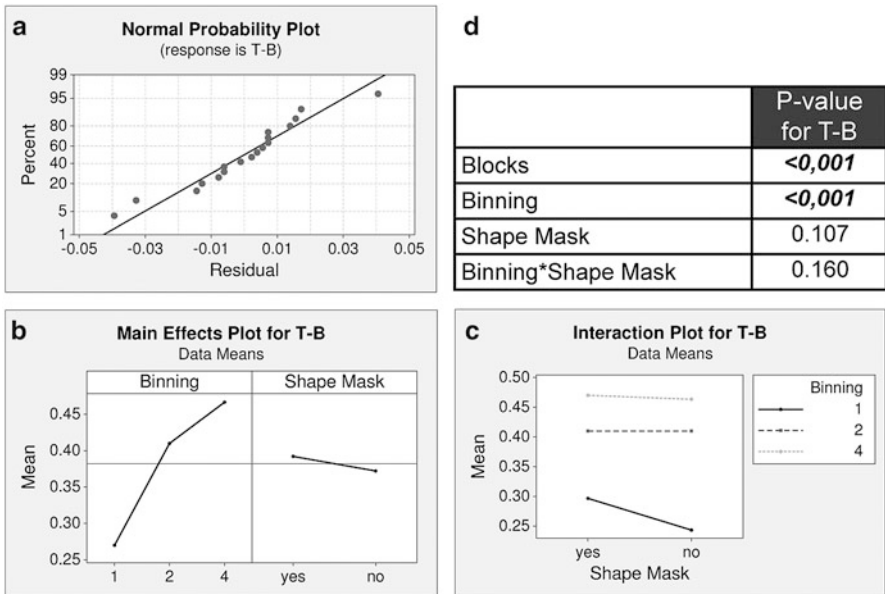
Because of the three-levels Binning factor, a General full factorial design was used (18 runs, 3 replicates). The screening analysis clearly demonstrates an interaction between Binning and Shape Mask on the Rsq output (Fig. 2) but not on the T-B output (Fig. 3).

Figure 2a indicates that residuals for Rsq are normally distributed—i.e. very close to the line representing the normal distribution—a condition necessary to proceed with a standard analysis without doing a variable transformation (see: [1]). The analysis shows a significantly lower Rsq for Binning 1 compared with Binning 2 and even more with Binning 4 ( $P = 0.04$ , Fig. 2d). An improvement in Rsq is observed when Shape Mask is applied (Fig. 2b). The interaction Plot (Fig. 2c) confirms that Binning 1 gives lower Rsq while Binning 2 and 4 show the best results. The influence of Shape Mask is maximal with Binning 1, moderate with 2 and negligible with 4. A  $P$ -value = 0.248 for the variable Blocks shows no influence of inter-day conditions for Rsq.

Figure 3a indicates that residuals are normally distributed also for T-B. The effect of Binning on the T-B output confirms Binning 1 as the worst condition, but also shows a trend, with Binning 4 better than 2 (see Fig. 3b); interestingly, Shape Mask has no influence on the T-B considered alone or even in combination with Binning as shown by the interaction plot (Fig. 3c), where lines are almost parallel. A  $P$ -value  $< 0.001$  for the variable Blocks indicates a significant influence of inter-day conditions on T-B. Overall, Binning was the sole significant factor ( $P < 0.01$ , Fig. 3d).



**Fig. 2** Minitab graphs of the screening analysis for Rsq: Normal probability plot for Residuals (a), Main Effect Plot (b) and Interaction Plot (c). The table in (d) shows P for the chosen factors and their interactions



**Fig. 3** Minitab graphs of the screening analysis for T-B: Normal probability plot for Residuals (a), Main Effect Plot (b) and Interaction Plot (c). The table in (d) shows P values for the chosen factors and their interactions

## 5 Results of the Second Analysis (Optimization)

The aim of the second analysis was to define an optimal parameter configuration, by considering the following factors/levels:

1. Binning (2; 4);
2. Shape-mask (yes; no);
3. Minimum response filtering (yes; no);
4. Output data filtering (Stat; Funct);
5. Response calculation (NFC; FC).

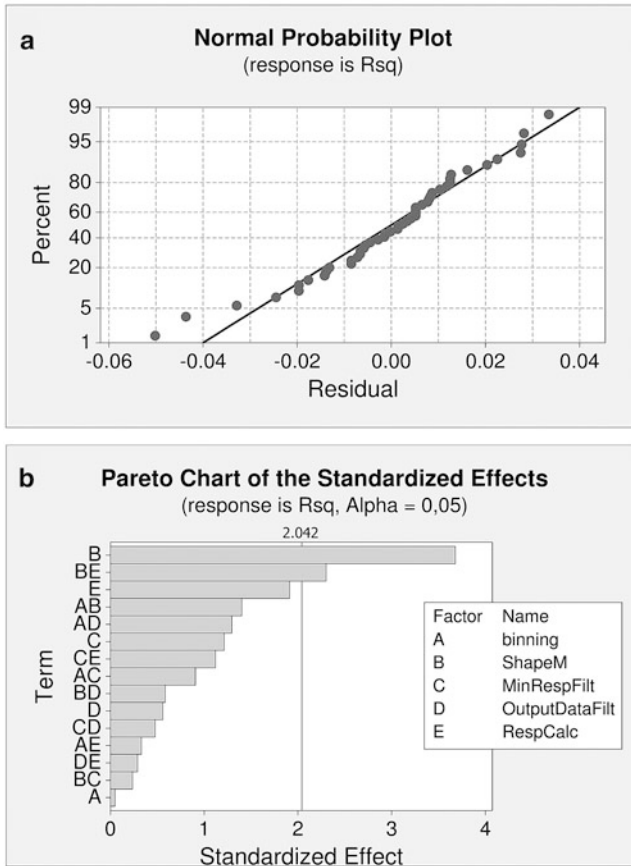
Under these conditions, a Full Factorial design would have required 32 runs per replicate, i.e. the same number of runs needed for an OVAT approach, however, with the advantage of providing information about interactions. Considering that we were interested also on the influence of inter-day variability, a minimum of 3 replicates (performed on image stacks produced in different days) had to be performed. This would have required a total of 96 runs. In order to reduce this number, we made the assumption that the interactions of the second order (i. e. interactions of two factors at a time) were sufficient for a correct approximation in our analysis, also considering that higher interactions (three factors at a time or more) are expected to be negligible in most cases (see [19]). Based on these considerations, we reduced the number of trials by employing the fractional factorial design with resolution V, which required 48 runs for 3 replicates, at the expenses of the assessment of third order interactions. Figure 4 shows the results of DoE Analysis for the Rsq output. Normal probability Plot (Fig. 4a) for Residuals show good fitting. The Pareto Chart of the Standardized Effects (Fig. 4b) indicates that the only statistically significant factor is Shape Mask ( $P = 0.001$ ) while the only significant interaction is Shape Mask with Response Calculation ( $P = 0.029$ ).

Taking into consideration the results shown in Fig. 5a, b, we can assume that, as far as Rsq is concerned, best results are obtained with: Shape-mask, Binning 4, no Minimum response filtering, Statistical Output data filtering and NFC Response calculation.

Similar analysis was performed considering T-B as the Output. Normal probability Plot (Fig. 6a) for residuals show good fitting. The Pareto chart of the standardized effects indicates that all the main factors, but Minimum response filtering, are statistically significant (Fig. 6b): Output data filtering ( $P < 0.001$ ); Binning ( $P < 0.001$ ); Shape-mask ( $P = 0.002$ ); and Response calculation ( $P = 0.004$ ). Minimum response filtering has a significant interaction with Output Data Filtering ( $P = 0.021$ ).

Considering the Main Effects Plot (Fig. 7a) and the Interaction Plot (Fig. 7b) for T-B, we can infer that best results are obtained with Shape-mask, Binning 4, Statistical Output data filtering, no Minimum response filtering and NFC Response calculation.

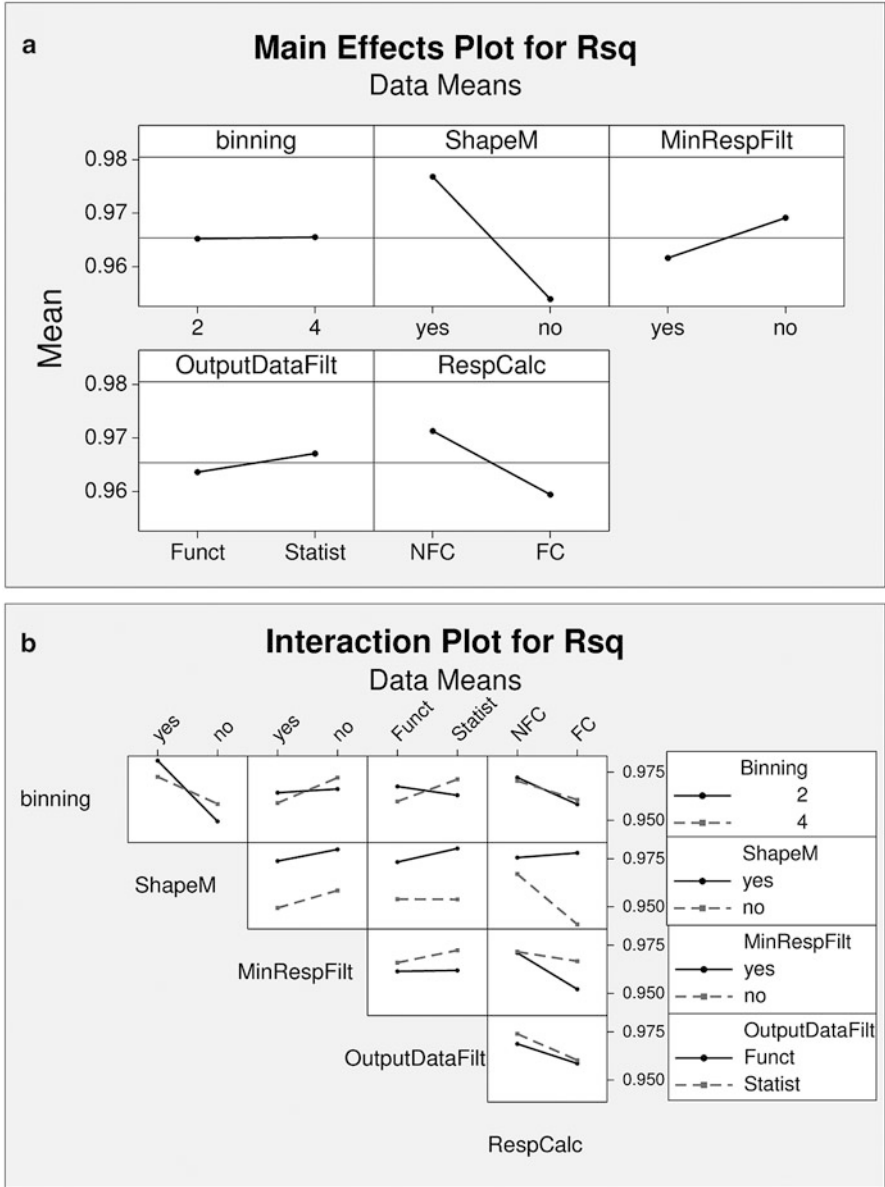
Based on the above results, we were able to define an optimized configuration (Table 1) and a suboptimal one (Table 2).



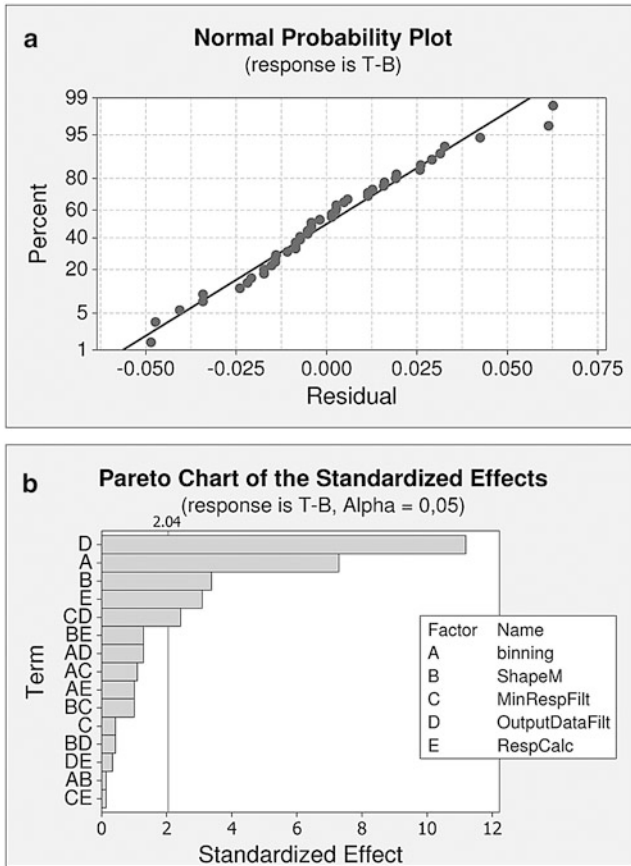
**Fig. 4** Minitab graphs for Rsq optimization analysis: the Normal Probability Plot for Residuals (a) indicates an suitable distribution of residuals; the Pareto of the Standardized Effects (b) indicates that there are only two significant effects (i. e. laying beyond the vertical line that marks the threshold for Alpha = 0.05): Shape Mask and the interaction between Shape Mask and Response Calculation

C-R curves were then calculated with both the optimized and the suboptimal set on the same data used for DoE analysis. Figure 8 illustrates an example in which the C-R curve obtained with the optimized set exhibits an Rsq value improved from 0.91 to 0.99 and a T-B value from 0.28 to 0.49, which represent a percent improvement (defined as  $(PS_{opt} - PS_{subopt})/PS_{subopt}$ , where PS = parameter set) of respectively +8.8 % (Rsq) and +75 % (T-B).

As a final consideration, P-value for BLOCKS showed an influence of inter-day conditions that is significant for T-B ( $P < 0.001$ ) but not for Rsq ( $P = 0.147$ ).



**Fig. 5** Minitab graphs for Rsq optimization analysis (factorial plots): the Main Effects Plot (a) confirms that the Shape Mask effect is the most important among single factors and that best results are obtained when the mask is applied: the Interaction Plot (b), shows the best combination for Shape Mask and Response Calculation (if ShapeM = yes, both values for RespCalc are suitable)

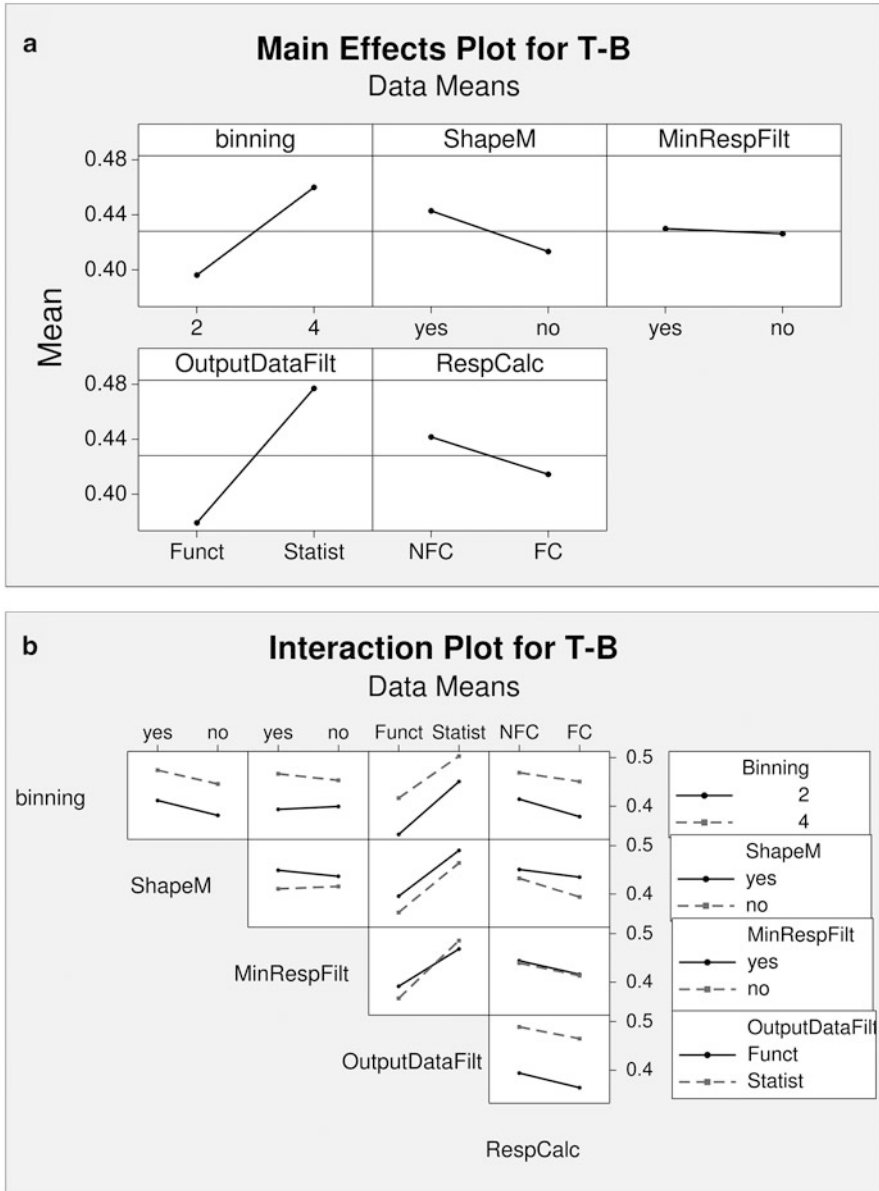


**Fig. 6** Minitab graphs for T-B optimization analysis: the Normal Probability Plot for Residuals (a) indicates a suitable distribution of residuals; the Pareto of the Standardized Effects (b) shows that all single factors, but the Minimum Response Filter (MinRespFilt), are significant, while only one interaction, the one between MinRespFilt and OutputDataFilt, lays beyond the vertical line (threshold for Alpha = 0.05)

## 6 Validation of Obtained Optimized Configuration

The optimized parameter configuration we obtained with the previous analysis was then validated.

As a first step, we verified the initial hypothesis that Binning had no significant interactions with factors other than ShapeM. Indeed, Figs. 4a and 6a show that the interactions between Binning and the other factors do not reach statistical significance. Of note, in the same Fig. 4a we can appreciate that ShapeM has a significant interaction with RespCalc, clearly indicating that it is not possible to separate the pixel-related factors from the others.



**Fig. 7** Minitab graphs for T-B optimization analysis (factorial plots): the Main Effects Plot (a) indicates the best values for the significant factors: Binning = 4, ShapeM = yes, OutputDataFilt = Statist and RespCalc = NFC. In the Interaction Plot (b), the value MinRespFilt = no together with OutputDataFilt = Statist are the significant interacting factors values that optimize the output T-B



**Table 1** Optimized parameter set for Rsq and T-B

| Optimized configuration    |                         |                         |                  |
|----------------------------|-------------------------|-------------------------|------------------|
| Factor                     | Optimized level for Rsq | Optimized level for T-B | Optimized choice |
| Binning                    | n.i.                    | 4                       | 4                |
| ShapeMask                  | YES                     | YES                     | YES              |
| Minimum response filtering | n.i.                    | NO                      | NO               |
| Output data filtering      | n.i.                    | Stat                    | Stat             |
| Response calculation       | NFC                     | NFC                     | NFC              |

*n.i.* not influent

**Table 2** Suboptimal parameter set for Rsq and T-B

| Suboptimal configuration   |                          |                          |                   |
|----------------------------|--------------------------|--------------------------|-------------------|
| Factor                     | Suboptimal level for Rsq | Suboptimal level for T-B | Suboptimal choice |
| Binning                    | n.i.                     | 2                        | 2                 |
| ShapeMask                  | NO                       | NO                       | NO                |
| Minimum response filtering | n.i.                     | YES                      | YES               |
| Output data filtering      | n.i.                     | Funct                    | Funct             |
| Response calculation       | FC                       | FC                       | FC                |

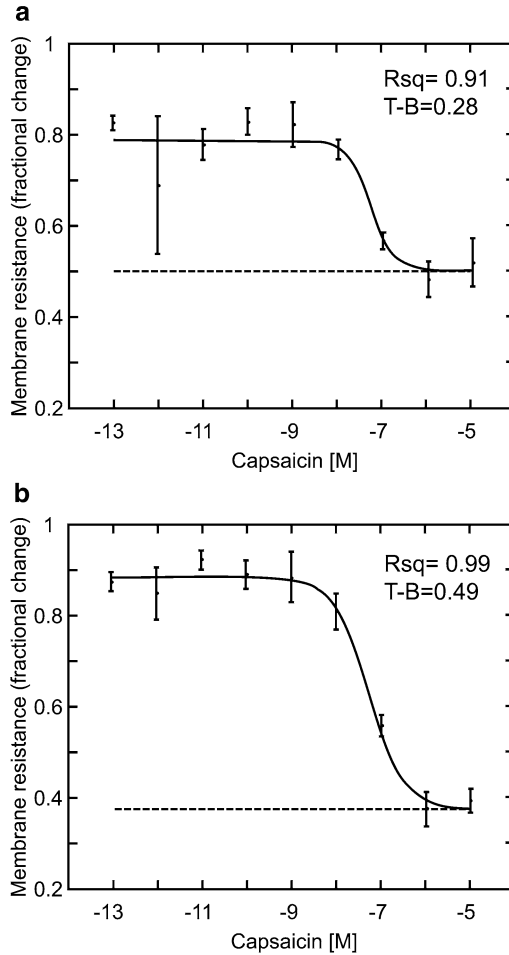
*n.i.* not influent

Afterwards, to validate the rejection of Binning = 1 in the first analysis, we ran MaLIA on 8 different image stacks with the same parameters employed in the optimized (Table 1) and suboptimal (Table 2) configurations, with the exception of Binning value set to 1. The substitution of Binning = 1 worsened the value of Rsq and T-B in both the optimized (−6 % and −30 %, respectively) and the suboptimal configuration (−17 % and −54 %, respectively). We can conclude that, as suggested by the experience during the development of the MaLIA program and assumed during the design of the first analysis, Binning = 1 minimized the overall performance. This ex post validation also confirms the validity of the assumptions we made in the first analysis of this unconventional DoE design.

Afterward, we produced C-R curves with both the optimized and the suboptimal sets on data from different experiments in order to validate the results in a wide range of cell and drug types (see Table 3).

Experimental data were randomly selected within a time interval of 2 years, representing five cellular lines exposed to their reference drugs. Two experiments for each cell line were considered. Such a wide time interval was used to take into account also changes due to the evolution of both biological protocols and screening processes.

The Rsq and the T-B values of the C-R curves obtained with the optimized and suboptimal parameters sets are compared in Fig. 9a, b and shown as percent variation  $((PS_{opt} - PS_{subopt})/PS_{subopt})$  in Fig. 9c. The charts clearly indicate that the optimized set consistently produces better CR curve: Rsq benefits of a slight



**Fig. 8** CR curves obtained with optimized and suboptimal parameter sets: the CR curve shows the fractional changes of the membrane resistance at different drug concentrations (log). The CR curves obtained with the suboptimal (a) and with the optimized (b) parameter sets, on the same images stack, are compared to put in evidence the marked improvement:  $Rsq$  from 0.91 to 0.99,  $T-B$  from 0.28 to 0.49

improvement (up to 4.3 %), while  $T-B$  takes much more advantage (up to 90.4 %). The only exception is represented by an experiment (HEK-293 GABA-A exp. 1), in which  $Rsq$  is lower (-1.2 %) with the optimized set, even though the  $T-B$  response maintains a positive performance of +9.3 %.

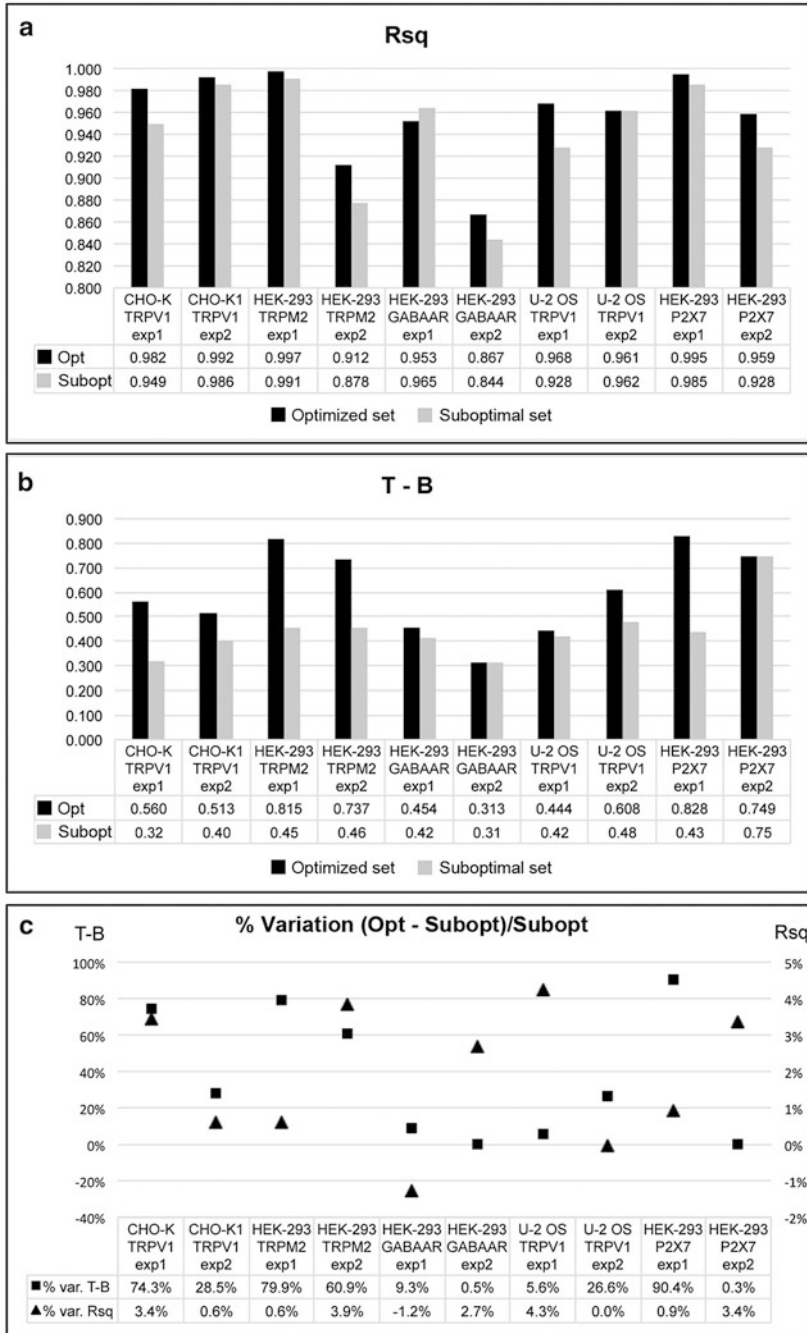


Fig. 9 (continued)


**Table 3** Pharmacological targets used to validate the optimized parameter set

| Pharmacological target |  | Reference drug   |
|------------------------|--|--|
| CHO-K1/TRPV1           | Chinese hamster ovary K1 cells expressing the transient receptor potential cation channel subfamily V member 1     | Capsaicin  |
| HEK-293/TRPM2          | Human Embryonic Kidney 293 cells expressing the transient receptor potential cation channel, subfamily M, member 2 | (agonist)  |
| HEK-293/GABAAR         | Human Embryonic Kidney 293 cells expressing the $\gamma$ -aminobutyric acid type A receptor                        | $\gamma$ -aminobutyric acid  |
| U-2 OS/TRPV1           | human osteosarcoma U2OS cell line expressing the transient receptor potential cation channel subfamily V member 1  | Capsaicin  |
| HEK-293/P2X7R          | Human Embryonic Kidney 293 cells expressing the purinergic receptor subclass P2X7                                  | BzATP, (2'/(3')-O-(4-Benzoylbenzoyl)adenosine-5'-triphosphate tri(triethylammonium)) |

## 7 Discussion and Conclusions

DoE was performed to optimize the set of analytical parameters used in a new drug screening procedure. This is a simple application of the method that provided useful results with good efficiency (time and resources vs results).

We adopted an unconventional DoE approach: the screening design, instead of being employed to reduce the number of factors, was used to reduce the number of levels of one of the factors, taking advantage of the specific knowledge of the image analysis process. This simplification made possible, in a second analysis, a direct and simpler comparison among all main parameters, thereby avoiding the more complex General Factorial design. This second design, which we called optimization in this study, was performed with a Resolution V Fractional Factorial design. A standard Response Surface design could not be employed since all factors are qualitative and intermediated values could not be envisaged (see: [18]). The choice of using a Fractional Factorial design, which ignores third order interactions, is largely supported by the evidence that results are little influenced by pairs of factors (interaction of the second order), validating the initial assumption of a negligible contribution of higher order interactions.

 **Fig. 9** Validation of the optimized set on different cells/drugs (see Table 2): Rsq (a) and T-B (b) values are always improved with the optimized set of parameters (*black columns*) rather than with the suboptimal one (*gray columns*), with a single exception (HEK-293 GABAAR exp.1,  $Rsq_{opt} < Rsq_{subopt}$ ). The percent variations are shown in (c), where Rsq variations refer to the *right y-axis*, while the T-B variations to the *left y-axis*

The unconventional choice of a general full factorial, followed by a fractional factorial design, allowed us to downsize the number of runs. A traditional general factorial, with 4 factors at 2 levels and 1 at 3 levels, would have required 48 ( $2^4 \times 3$ ) runs for each replicate. With our approach, we made 6 ( $3 \times 2$ ) runs for the first phase and 16 ( $2^{(5-1)}$ ) for the second analysis, with a total of 22 and a saving of 26 runs with respect to a single general factorial design. Considering 3 replicates, we saved 78 runs. Each MaLIA run (inputting data, setting parameters, waiting for analysis elaboration and collecting results) takes at list 7–10 min to a skilled operator. Accordingly, we saved up to 13 h on a total forecasted effort of 24 h, i.e. 54 % saving. Overall, by this DoE approach, we saved time, gaining more information.

Finally, the use of the Blocking reveals that the impact of the experimental day could not be neglected in this study, which embraces 2 years' work of development of the drug screening platform. Interestingly, only the T-B, but not the Rsq, was subjected to inter-day variation. This result can lead to two conclusions: first, this influence may deserve further analysis after final validation of the screening platform. Secondly, experimental data could have been transformed to correct for blocks, thus obtaining a result independent of day-to-day variability (see [1]; Chap. 2). However, the excellent validation of the optimized set of parameter demonstrates that the result is robust enough to make a more sophisticated analysis not necessary. Further investigation might involve a refinement of the quantitative thresholds used for some of the parameters (e.g. threshold for Minimum response filtering). If we consider the results in terms of the final application, the observed inter-day variation appears to reflect the process of optimization of the biological and biochemical conditions during the progressive development of the drug screening platform. On the other hand, they provide direct evidence that also in sub-optimal experimental conditions, the set of choice guarantees the best possible result. This evidence receives further confirmation by the fact that a consistent improvement was observed independently of the cell lines and drug type. Overall, this is an important prerequisite to consider this new approach for the study of different pharmacological targets, in an unbiased way and in an industrial context.

In conclusion, our work demonstrates that the application of DoE on the selection of software parameters, although still poorly exploited, can provide very useful results by reducing the number of trials compared to a complete OVAT approach. In this respect, it is worth noticing how a conscious introduction of constraints to reduce the degrees of interactions, along with a two-stage design, can greatly simplify the modelling and thus the obtaining of the result.

**Acknowledgments** The authors wish to thank Axxam S.p.A. for kindly providing the cell lines employed in this work. We greatly acknowledge the scientific contribution of Dr. Riccardo Fesce during the development of MaLIA. The project was developed in the R&D laboratory of the Advanced Light and Electron Microscopy BioImaging Center (Experimental Imaging Center-San Raffaele Scientific Institute) within the framework of the project “Optical Method for Ion Channel Screening”, “Progetto Metadistretti Tecnologici”, Regione Lombardia.

## References

1. Anderson, M.J., Whitcomb, P.J.: DOE Simplified. Practical Tools for Effective Experimentation. Productivity Press, New York (2007)
2. Beyer, H.G., Sendhoff, B.: Robust optimization—a comprehensive survey. *Comput. Methods Appl. Mech. Eng.* **196**, 3190–3218 (2007)
3. Bollin, F., Dechavanne, V., Chevalet, L.: Design of experiment in CHO and HEK transient transfection condition optimization. *Protein Expr. Purif.* **78**(1), 61–68 (2011)
4. Bongiovanni, A., Colotti, G., Liguori, G.L., Cirafici, A.M., Di Carlo, M., Digilio, F.A., Lacerra, G., Mascia, A., Lanati, A., Kisslinger, A.: Applying quality principles and project management methodologies in biomedical research: a public research network's case study. *Accred. Qual. Assur.* **20**(3), 203–213 (2015)
5. Chou, T.C.: Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacol. Rev.* **58**, 621–681 (2006)
6. Dejaegher, B., Vander Heyden, Y.: Experimental designs and their recent advances in set-up, data interpretation, and analytical applications. *J. Pharm. Biomed. Anal.* **56**, 141–158 (2011)
7. Ebrahimi-Najafabadi, H., Leardi, R., Jalali-Heravi, M.: Experimental design in analytical chemistry—part I: theory. *J. AOAC Int.* **97**(1), 3–11 (2014)
8. Eliasson, M., Rännar, S., Madsen, R., Donten, M.A., Marsden-Edwards, E., Moritz, T., Shockcor, J.P., Johansson, E., Trygg, J.: Strategy for optimizing LC-MS data processing in metabolomics: a design of experiments approach. *Anal. Chem.* **84**, 6869–6876 (2012)
9. Elliott, P., Billingham, S., Bi, J., Zhang, H.: Quality by design for biopharmaceuticals: a historical review and guide for implementation. *Pharm. Bioprocess.* **1**(1), 105–122 (2013)
10. Glasnov, T.N., Tye, H., Kappe, O.: Integration of high speed microwave chemistry and a statistical design of experiment' approach for the synthesis of the mitotic kinesin Eg5 inhibitor monastrol. *Tetrahedron* **64**, 2035–2041 (2008)
11. Gooding, O.W.: Process optimization using combinatorial design principles: parallel synthesis and design of experiment methods. *Curr. Opin. Chem. Biol.* **8**, 297–304 (2004)
12. Hibbert, D.B.: Experimental design in chromatography: a tutorial review. *J. Chromatogr. B* **910**, 2–13 (2012)
13. Ilzarbe, J., Alvarez, M.J., Viles, E., Tanco, M.: Practical applications of design of experiments in the field of engineering: a bibliographical review. *Qual. Reliab. Eng. Int.* **24**, 417–428 (2008)
14. Leardi, R.: Experimental design in chemistry: a tutorial. *Anal. Chim. Acta* **652**, 161–172 (2009)
15. Lutz, M.W., Menius, J.A., Choi, T.D., Gooding Laskody, R., Domanico, P.L., Goetz, A.S., Saussy, D.L.: Experimental design for high-throughput screening. *Drug Discov. Today* **1**(7), 277–286 (1996)
16. Macarron, R., Banks, M.N., Bojanic, D., Burns, D.J., Cirovic, D.A., Garyantes, T., Green, D.V., Hertzberg, R.P., Janzen, W.P., Paslay, J.W., Schopfer, U., Sittampalam, G.S.: Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10**(3), 188–195 (2011)
17. Malo, N., Hanley, J.A., Carlile, G., Liu, J., Pelletier, J., Thomas, D., Nadon, R.: Experimental design and statistical methods for improved hit detection in high-throughput screening. *J. Biomol. Screen.* **15**(8) (2010)
18. Montgomery, D.C.: Design and Analysis of Experiments. Wiley, New York (1997)
19. Montgomery, D.C., Runger, G.C.: Applied Statistics and Probability for Engineers. Wiley, New York (2003)
20. Posner, B.A.: High-throughput screening-driven lead discovery: meeting the challenges of finding new therapeutics. *Curr. Opin. Drug Discov. Devel.* **8**(4), 487–494 (2005)
21. Pota, M., Pedone, A., Malavasi, G., Durante, C., Cocchi, M., Menziani, M.C.: Molecular dynamics simulations of sodium silicate glasses: optimization and limits of the computational procedure. *Comput. Mater. Sci.* **47**(3), 739–751 (2009)
22. Ranga, S., Jaimini, M., Sharma, S.K., Chauhan, B.S., Kumar, A.: A review on Design of Experiments (DOE). *Int. J. Pharm. Chem. Sci.* **3**(1), 216–224 (2014)

23. Scannell, J.W., Blanckley, A., Boldon, H., Warrington, B.: Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200 (2012)
24. Tye, H.: Application of statistical ‘design of experiments’ methods in drug discovery. *Drug Discov. Today* **9**(11), 485–491 (2004)
25. Weissman, S.A., Anderson, N.G.: Design of Experiments (DoE) and process optimization. A review of recent publications. *Org. Process. Res. Dev.* ACS Publications (2014)
26. Zang, C., Friswell, M.I., Mottershead, J.E.: A review of robust optimal design and its application in dynamics. *Comput. Struct.* **83**, 315–326 (2005)
27. Zheng, H., Clausen, M.R., Kastrup Dalsgaard, T., Mortensen, G., Bertram, H.C.: Time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches. *Anal. Chem.* **85**, 7109–7116 (2013)

# Computational Modeling of miRNA Biogenesis

Brian Caffrey and Annalisa Marsico

**Abstract** Over the past few years it has been observed, thanks in no small part to high-throughput methods, that a large proportion of the human genome is transcribed in a tissue- and time-specific manner. Most of the detected transcripts are non-coding RNAs and their functional consequences are not yet fully understood. Among the different classes of non-coding transcripts, microRNAs (miRNAs) are small RNAs that post-transcriptionally regulate gene expression. Despite great progress in understanding the biological role of miRNAs, our understanding of how miRNAs are regulated and processed is still developing. High-throughput sequencing data have provided a robust platform for transcriptome-level, as well as gene-promoter analyses. *In silico* predictive models help shed light on the transcriptional and post-transcriptional regulation of miRNAs, including their role in gene regulatory networks. Here we discuss the advances in computational methods that model different aspects of miRNA biogenesis, from transcriptional regulation to post-transcriptional processing. In particular, we show how the predicted miRNA promoters from PROMiRNA, a miRNA promoter prediction tool, can be used to identify the most probable regulatory factors for a miRNA in a specific tissue. As differential miRNA post-transcriptional processing also affects gene-regulatory networks, especially in diseases like cancer, we also describe a statistical model proposed in the literature to predict efficient miRNA processing from sequence features.

**Keywords** Mirna regulation • Promoter prediction • Mirna processing • Gene regulatory networks

## 1 The Role of miRNAs in Gene-Regulatory Networks

In biological research, diverse high-throughput techniques enable the investigation of whole systems at the molecular level. One of the main challenges for computational biologists is the integrated analysis of gene expression, interactions between

---

B. Caffrey • A. Marsico (✉)

Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

e-mail: [caffrey@molgen.mpg.de](mailto:caffrey@molgen.mpg.de); [marsico@molgen.mpg.de](mailto:marsico@molgen.mpg.de)



genes and the associated regulatory mechanisms. The two most important types of regulators, Transcription Factors (TFs) and microRNA (miRNAs) often cooperate in complex networks at the transcriptional level and at the post-transcriptional level, thus enabling a combinatorial and highly complex regulation of cellular processes [1].

While TFs regulate genes at the transcriptional level by binding to proximal or distal regulatory elements within gene promoters [1], microRNAs (miRNAs) act at the post-transcriptional level on roughly half of the human genes. These short non-coding RNAs of 18–24 nucleotides in length which can bind to the 3'-untranslated regions (3' UTRs) or coding regions of target genes, leading to the degradation of target mRNAs or translational repression [2].

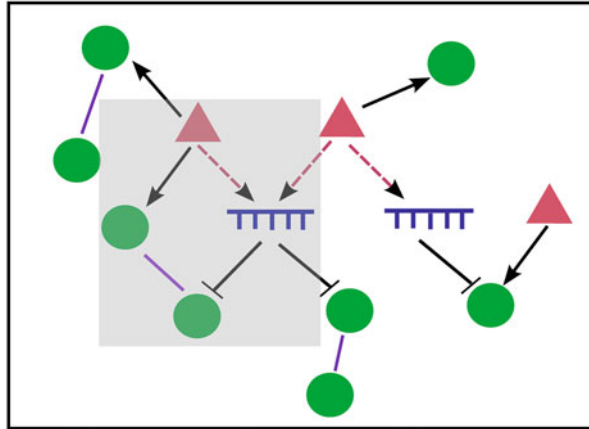
MiRNAs are associated with an array of biological processes, such as embryonic development and stem cell functions in mammals [3], and a crucial role of miRNAs in gene regulatory networks has been recognized in the last decade in the context of cancer and other diseases [4, 5]. Altered miRNA expression profiles have often been associated with cancer development, progression and prognosis [6]. MiRNAs which negatively regulate tumor suppressor genes can be amplified in association with cancer development. On the other hand, deletions or mutations in miRNAs targeting oncogenes can lead to the over-expression of their targets [5, 6].

MiRNAs also affect several aspects of the immune system response [7]. For example, cells of the hematopoietic system can be distinguished from other tissues by their miRNA expression profiles, including, among the others, the highly expressed miRNA hsa-miR-155 [7]. Other immune system-related miRNAs are activated in response to viral or bacterial infections (e.g. hsa-miR-146a) and they affect the expression of several cytokines downstream [8].

Given the growing prevalence of miRNA functions in contributing to the control of gene expression, gene regulatory networks have been expanded to become rather complex incorporating the involvement of miRNAs. The general framework for inferring gene regulatory networks involving Transcription Factors (TFs) and miRNAs is usually built using the following steps:

- **1:** When expression data are available under a certain condition, the first step is to identify those genes which are mostly expressed in that particular condition or de-regulated compared to a control experiment.
- **2:** miRNAs responsible for the observed co-expression or de-regulation of a set of genes are identified by identifying enriched miRNA binding sites in the 3'-UTRs of such genes. This is usually done by mining publicly available databases for miRNA-target interactions [9, 10].
- **3:** MiRNA-target interactions are filtered based on the miRNA expression level (when available) or by using a cutoff score indicating the reliability of the predicted interaction. In addition, it is expected that when a miRNA regulates a gene, the miRNA and the gene show typical correlated expression patterns across multiple samples. This can be used as a criterium to further filter miRNA-gene interactions which do not show any such correlation [9].

**Fig. 1** Cooperative action of miRNAs and TFs in gene regulatory networks. miRNAs are colored in *blue*, TFs in *red* and their regulated target genes, as well as genes involved in potential protein–protein interactions are colored in *green*. Dotted red arrows indicate potential regulators of miRNAs and purple lines indicate protein–protein interactions extracted from databases. A typical feedback loop is highlighted in *grey*



- **4:** TFs regulating this set of genes can be inferred by means of prediction algorithms which scan for known TF binding sites in the proximal gene promoter regions using Position Weight Matrices (PWMs) [11].
- **5:** Protein–protein interaction databases, such as STRING, BioGrid and KEGG can be inspected to find possible interactors of such genes and the cellular pathways that they affect.

These steps give rise to a network as depicted in Fig. 1. In this schematic representation nodes represent the significant set of genes, miRNAs and transcription factors in the process under study and the links between them represent predicted regulatory interactions.

It is well known that miRNAs are involved in negative regulation and/or positive feedback loops which can also involve the transcription factors that regulate their activity [12]. The knowledge of the transcription factors which regulated the miRNAs in question often provide the missing links in the aforementioned regulatory network (Fig. 1, red dotted arrows). The identification of TF–miRNA interactions remains a difficult task without which a full understanding of the underlying processes is hampered. In recent years there has been an increase in the development of computational methods to predict miRNA promoters and their regulating TFs in order to unravel the TF–miRNA interactions missing in such typical regulatory networks.

## 2 MiRNA Transcriptional Regulation

### 2.1 Challenges of *in silico* miRNA Promoter Identification

MiRNA promoter recognition is a crucial step towards the understanding of miRNA regulation. Knowing the location of the miRNA transcription start site (TSS) enables the location of the core promoter, the region upstream of the TSS which contains

the TFs binding sites necessary to initiate and regulate transcription. Predictions of binding sites in the core promoter elements can enable the identification of regulatory factors for a specific miRNA (or a class of miRNAs), greatly improving our understanding of miRNA function, and their role in tissue-specific feedback loops.

Genome-wide identification of miRNA promoters has been hindered for many years by two main factors. The first reason is the deficit in mammalian promoter sequence characterization, which makes promoter prediction a challenging task in general [13]. Although promoter regions contain short conserved sequence features that distinguish them from other genomic regions, they lack overall sequence similarity, preventing detection by sequence-similarity-based search methods such as BLAST. Promoter recognition methods in the early 90s exploited the fact that promoters contain few specific sequence features or TF binding sites that distinguish them from other genomic features [13]. This observation could be used to build a consensus model, such as Position Weight Matrices (PWMs) or Logos to search for new promoters in the genome. It soon became clear that such methods could not be generalized to all existing promoters and more advanced strategies for pattern recognition utilized machine learning models trained on sequence k-mers.

The second reason for the lack of knowledge in miRNA transcriptional regulation is due to the complexity of the miRNA biogenesis pathway: miRNAs, whether they are located in intergenic regions or within introns of protein-coding genes, often referred to as host genes, are generally generated from long primary transcripts which are rapidly cleaved in the nucleus by the enzyme Droscha [2]. This presents a technical barrier for large-scale identification of miRNA TSSs as they can be located in regions far away from the mature miRNA and cannot be inferred simply from the annotation of the processed mature miRNA, as done for stable protein coding gene transcripts [14]. In addition, the situation is further complicated by the fact that recent studies indicate that several alternative miRNA biogenesis pathways exist, especially for intragenic miRNAs. Indeed, if co-transcription with the host gene were the only mechanism to generate intragenic miRNAs, then the mirna and hostgene expression should be highly correlated among different tissues or conditions. Many recent studies, however, show many instances of poor correlation between mirna and host gene, pointing to an independent regulation of the mirna, utilizing an alternative intronic promoter [15]. There is evidence that intragenic miRNAs may act as negative feedback regulatory elements of their hosts interactomes [16] but the contribution of host gene promoter versus intronic miRNA promoters, and the mechanisms that control intronic promoter usage are still interesting open questions in the biology of miRNA biogenesis.

Although overall similarity in promoters is not a general phenomena, it does exist in the form of phylogenetic footprinting. Based on this observation, one of the first methods for miRNA promoter detection identifies about 60 miRNA transcriptional start regions by scanning for highly conserved genomic blocks within 100 kb of each mature miRNA and searching for a core promoter element in the consensus

sequence regions extracted from these blocks [17]. Although this method proved to be valid in the identification of evolutionary conserved promoters, the sensitivity of such an evolutionary approach is very low, given the high number of non-conserved miRNAs annotated in MiRBase [18].

## ***2.2 Next Generation Sequencing (NGS) Technology Leads to Significant Advances in miRNA Promoter Prediction***

Recently, thanks to the advent of next-generation sequencing technologies combined with Chromatin Immunoprecipitation (CHIP-Seq technology [19]) and nascent transcript capturing methods, such as Cap Analysis of Gene Expression coupled to NGS sequencing (deepCAGE) [20] or Global run on sequencing (GRO-seq) [21], several computational methods for miRNA promoter prediction genome-wide have been developed, providing valuable understanding in the detailed mechanisms of miRNA transcriptional regulation. For example, the epigenetic mark H3K4me3 has been identified as a hallmark of active promoters, and computational methods for promoter recognition have begun exploiting this information systematically.

The deepCAGE technique enables the mapping of the location of TSSs genome-wide. In the FANTOM4 Consortium this technique was applied across various different tissues and conditions in order to profile transcriptional activities and promoter usage among different libraries.

GRO-seq is a technique to capture nascent RNAs genome-wide by quantifying the signal of transcriptionally engaged PolIII at gene promoters. Both deepCAGE and GRO-seq read density is sharply peaked around transcripts TSS and it can be successfully used to locate the TSSs of miRNA primary transcripts [14, 22]. Finally, recent RNA-seq studies with increased sequencing depth can also be used to identify the transient and lowly expressed pri-miRNA transcripts [22].

## ***2.3 Classification and Comparison of miRNA Promoter Prediction Methods***

A limited number of miRNA promoter recognition methods have been developed in the past few years and can be classified either according to the methodology used, supervised versus unsupervised learning approaches, or based on the nature of their predictions, tissue specific versus general promoter predictions and intergenic versus all predicted promoters, including intronic promoters. The main features of existing miRNA promoter prediction methods can be summarized in Table 1.

According to the model used to describe the data one can distinguish two categories of miRNA promoter recognition methods:

**Table 1** Comparison of different methods for miRNA promoter prediction

|                      |  |   |  |   |   |  |  |   |
|----------------------|--|---|--|---|---|--|--|---|
| Cell line            | Fujita [17]  | Ozsolak [25]  | Marson [23]  | Barski [24]   | S-Peaker: Megraw [27]   | miRStart: Chien [26]                                     | PROMiRNA: Marsico [14]   | microTSS: Georgakilas [22]  |
|                      | -  | UACC62, MALME, MCF cells  | mESC, hESC cells   | CD4+ T cells  | -   | 36 different tissues                                     | 33 different tissues   | mESC, hESC, IMR90 cells   |
| Data used            | Blastz genomic alignments from UCSC  | Nucleosome positions from ChIP-chip data  | H3K4me3 ChIP-seq data  | H3K4me3, H3K9ac, H2AZ, and PolII ChIP-seq data  | CAGE data (FANTOM4)   | 36 deepCAGE libraries (FANTOM4) and 14 TSS-seq libraries | 33 deepCAGE libraries (FANTOM4)  | RNA-seq data, HeK4me2, PolII ChIP-seq and DNase-seq   |
| Methodology          | Unsupervised approach: identification of conserved blocks upstream of miRNAs | Unsupervised approach: empirical score of nucleosome-free regions based on sequence features (TFBs) | Unsupervised approach: empirical score based on HeK4me3 conservation and proximity to the mature miRNA | Unsupervised approach: score accounting for evidence of peaks from four ChIP-seq signals, plus EST evidence | Supervised approach: L1-logistic regression model trained on gene promoter TFBS | Supervised model: SVM trained on protein coding genes    | Semi-supervised mixture model built on CAGE data and sequence features | Supervised model: SVM trained on chromatin features at gene promoters and then used to score miRNA RNA-seq enriched regions |
| Intergenic promoters | yes  | yes   | yes  | yes   | yes   | yes  | yes  | yes   |
| Intronic promoters   | Not reported   | yes   | no   | no  | no  | not reported   | yes  | no  |

- *De novo approaches*, which identify and score miRNA TSS in an unsupervised manner. These include models based on experimentally determined histone mark profiles [23, 24] or nucleosome positioning patterns [25]. For example Marson [23] and Barski [24] consider regions enriched in H3K4me3 signal as putative promoters and assign them a score. Oszolak [25] combine nucleosome positioning patterns with ChIP-chip screens to score putative transcription initiation regions upstream of active miRNAs.
- *Supervised methods*, based on the evidence that miRNA promoters present the typical characteristics of Polymerase II transcription and therefore trained on protein coding gene promoter features and subsequently used to predict miRNA promoters. Such methods include mirStart [26] and microTSS [22] (Table 1). MirStart trains a SVM model on protein coding gene features (CAGE tags, TSS-Seq and HeK4me3 ChIP-Seq data), and uses the trained model to identify putative miRNA promoters [26]. microTSS also uses a combination of three SVM models trained on HEK4me3 and PolII occupancy at protein-coding gene promoters to score putative initial miRNA TSSs candidates derived from deeply sequenced RNA-Seq data [22].

One of the latest miRNA promoter prediction tools, PROMiRNA, is a method in between these two categories [14]. PROMiRNA is based on a semi-supervised classification model which does not make any assumption about the nature of miRNA promoters and their similarities to protein-coding genes. On the contrary, PROMiRNA tries to learn the separation between putative miRNA promoters and transcriptional noise based on few features, such as CAGE tag clusters upstream of annotated miRNAs and sequence features.

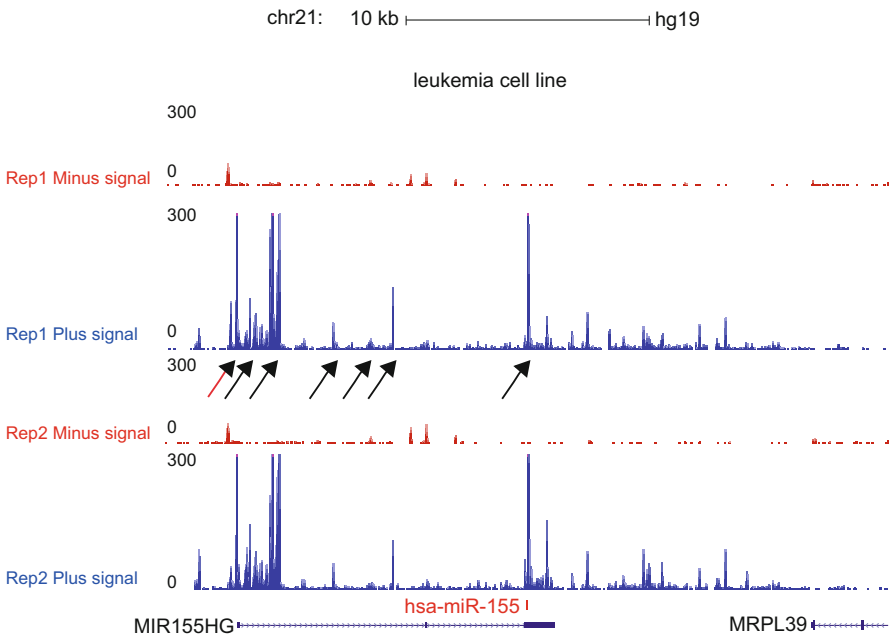
Each of the described methods has advantages and disadvantages. Methods for miRNA promoter recognition based solely on sequence features, such as the evolutionary framework proposed by Fujita [17] or S-Peaker [27], based on TF binding sites and proposed by Megraw et al., are very accurate in identifying putative promoter regions. MiRNAs are, however, known to mediate gene regulation in a highly tissue-specific manner, therefore it is expected that their regulation also happens in a tissue-specific way. Such methods cannot distinguish between promoters potentially active in different tissues, given that sequence features are invariant features, but merely suggest possible locations for miRNA promoters. On the other hand, methods based on chromatin features have been designed for specific cell lines, therefore providing a snapshot of the active promoters. Histone mark-based methods provide a broad view of promoter regions, rather than high-resolution predictions, hampering the detection of multiple TSSs close to each other in the genome. In addition, most chromatin-based methods can predict the promoters of independently transcribed intergenic miRNAs, but lack sensitivity in discovering alternative or intronic promoters.

MicroTSS overcomes the problem of the non-informative broad predictions by making use of deep-coverage RNA-seq data and pre-selecting RNA-seq islands of transcription upstream of intergenic pre-miRNAs at single-nucleotide resolution. Such initial miRNA promoter candidates are then given as input to the SVM model

which returns the predictions. Due to the nature of the RNA-seq used to pre-select candidate TSSs, microTSS works well for intergenic miRNAs but is not suitable for identifying intronic promoters due to the difficulty in discriminating transcription initiation events from the read coverage signal corresponding to the host transcript.

The method from Ozsolak [25] and PROMiRNA [14] are the only two methodologies which report predictions of intronic promoters. In particular, in PROMiRNA miRNA promoter predictions are derived from multiple high-coverage deepCAGE libraries, and correspond to highly expressed, as well as lowly expressed tissue-specific intronic promoters.

Figure 2 shows seven predicted TSSs for hsa-miR-155, six of which are intronic promoters in a leukemia cell line, indicating that alternative promoters are able to drive the expression of this miRNA in this cell line. However, due to the low expression of alternative intronic promoters, compared to intergenic promoters, and to the difficulty of validating such promoter predictions (a gold standard data-set for miRNA promoters is missing), predictions of intronic promoters may suffer from higher false discovery rates compared to intergenic promoters.



**Fig. 2** PROMiRNA predicted promoters for hsa-miR-155, a human miRNA located in the non-coding BIC host transcript (also called MIR155HG). The *red arrow* indicates the TSS of the host gene and the *other arrows* point to the predicted alternative intronic promoters located in the genomic range between 677 bp and 12 kb upstream of the mature miRNA. The promoter predictions were consistent in two different CAGE replicates

### 2.3.1 miRNA-Mediated Regulatory Network Reconstruction

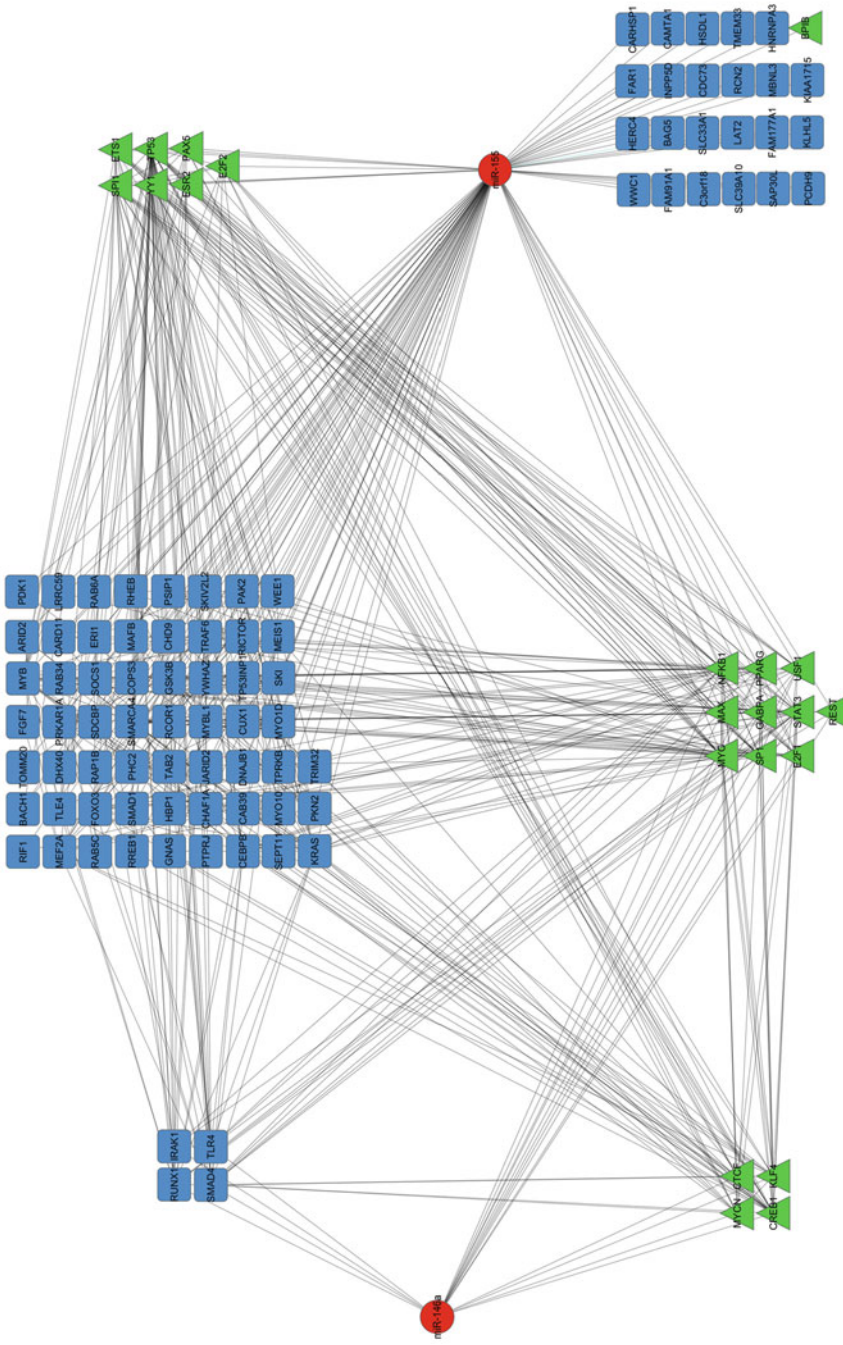
Here we show an application of PROMiRNA to the derivation of the tissue-specific miRNA-mediated regulatory network in three immune cell line libraries from the FANTOM4 CAGE data. For simplicity we include in this network only two of the main human miRNAs which are highly expressed and known to play a role in the Immune System: hsa-miR-146a and hsa-miR-155. MiR-146a is an intergenic miRNA known to be involved in regulation of inflammation and other processes related to innate immune response [8]. Mir-155 resides in the non-coding host gene MIR155HG and is known to play a role in cancer, as well as viral and bacterial infection processes [28]

PROMiRNA predicts two alternative promoters in the leukemia cell line for hsa-miR-146a, one located 17 kb upstream of the mature miRNA and the other 16.6 kb, and six alternative intronic promoters (in addition to the host gene promoter) for hsa-miR-155 (as already shown in Fig. 2). Starting from these predictions, we scanned the 1000 bp regions around each predicted miRNA TSSs for putative transcription factor binding sites with the TRAP tool [29]. Given a database of TFs motif models, TRAP computes the affinity of each factor for a certain genomic sequence. For each predicted promoter we ranked the TFs based on their computed binding affinities. The top ten factors regulating each miRNA are selected and included in the network if they are expressed in Immune System cell lines, according to the Human Protein Atlas database [30]. Potential regulatory factors are connected by means of edges to the corresponding miRNA (Fig. 3). Also potential miRNA targets extracted from TargetScan and other miRNA target databases [9], as well as interactions between gene–gene and gene-TF are extracted from the STRING database [31] and, if expressed in the Immune system, added to the network (Fig. 3). This partial reconstruction of the regulatory network involving hsa-mir-146a and hsa-mir-155 in the Immune System shows that a portion of the top target predictions is shared between the two miRNAs, while other targets are specific to one or the other miRNA. Also, hsa-miR-146a and hsa-miR-155 seem to be targeted by a set of common transcription factors, among which we find the NFKB1, a well known Immune System factor.

## 3 Predictive Models of miRNA Processing

Global mature miRNA expression is not only regulated at transcriptional level, but several post-transcriptional steps influence the final miRNA expression level and contribute to define a particular phenotype. In detail, miRNA initially generated in the nucleus as long primary transcripts are processed by the Microprocessor complex (Drosha/DGCR8) to produce stem-loop structured precursors which are then further processed in the cytoplasm by Dicer [32]. While signatures of miRNA expression may be used as biomarkers for cancer diagnosis and stratification in several cancers, it has become clear in recent years that specifically aberrant processing,

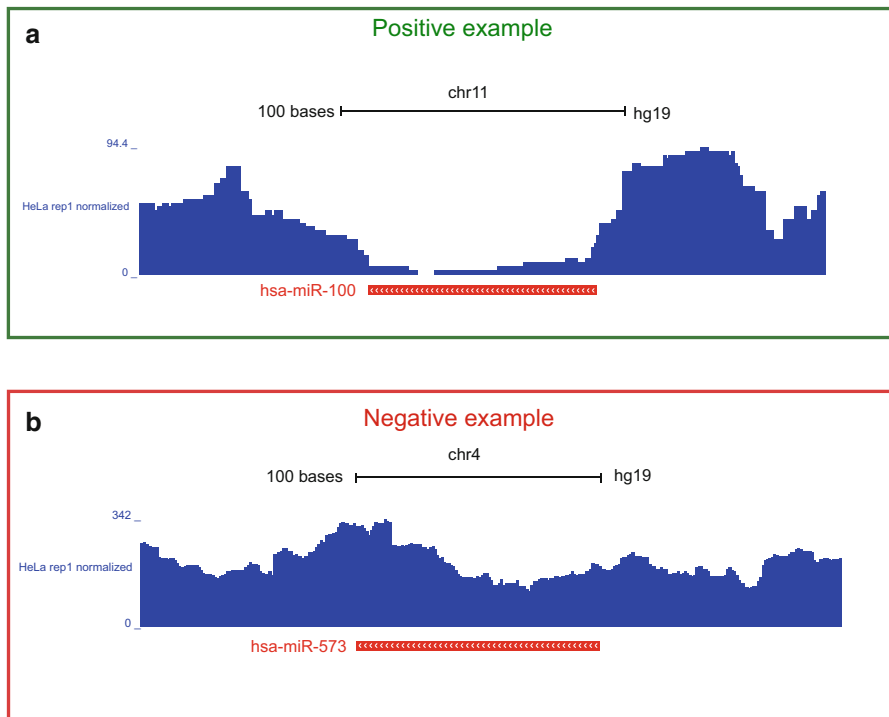




**Fig. 3** Regulatory network mediated by miRNAs hsa-miR-155 and hsa-miR-146a, including experimentally validated (mirTarbase) and predicted (overlap of miranda & TargetScan) miRNA-target interactions, transcription factor-miRNA interactions detected by means of the TRAP tool in the PROmiRNA predicted miRNA regions, and gene-gene interactions extracted from the STRING database

rather than altered transcription, correlates with cell invasion or progression of inflammation. The method by which the Microprocessor is able to distinguish miRNA hairpins from random hairpin structures along the genome and efficiently process them is still a subject of investigation. Recent studies have shown that sequence motifs flanking precursor miRNAs play a significant role in primary transcript cleavage [33].

In a recent study [34] we have quantified the effect of different sequence motifs on the Microprocessor activity in an endogenous setting. We have performed high-throughput RNA sequencing experiments of nascent transcripts associated to the chromatin fraction in different cell lines. Since processing of primary miRNA transcripts occurs co-transcriptionally, while the transcript is still associated to chromatin, the read coverage pattern at miRNA loci shows the typical Microprocessor signature, where Droscha cleavage is reflected in a significant drop in the read coverage in the precursor region. We have defined a quantitative measure of processing efficiency called Microprocessing Index (MPI), as the logarithm of the ratio between the read density adjacent to the pre-miRNA and the read density in the precursor region. On the basis of MPI values, miRNAs could be divided into *efficiently processed* (Fig. 4a  $MPI \leq -1.0$ , also called positive examples) and *non-efficiently processed* (Fig. 4b  $MPI \geq -0.4$ , also called negative examples).



**Fig. 4** Genomic regions around miRNAs hsa-miR-100 (a) and hsa-miR-573 (b), respectively, and normalized read coverage at the miRNA loci. The significant drop in read coverage at the miR-100 precursor indicates that this miRNA is efficiently processed in HeLa cells (a), while miR-573 is not

A classification model based on sequence features was built in order to discriminate between these two classes. We used L1-regularized logistic regression for training and classification of the miRNA in positives and negatives. In detail, given a binary variable  $Y$ , where  $y_i = 0$  or  $y_i = 1$  for each data point  $i$ , the probability of the outcome of  $Y$ , given the data  $\mathbf{X}$ , is given by the following sigmoid function:

$$P(y = 1|x, \theta) = \frac{1}{1 + \exp(-\theta^T x)} \quad (1)$$

where  $\theta$  is the parameter vector of the logistic regression model. The optimization problem (Maximum Likelihood Estimate of  $\theta$ ) in the case of L1-regularization is formulated as the following:

$$\min_{\theta} \left( \sum_{i=1}^M -\log P(y_i|x_i, \theta) + \beta \|\theta\|_1 \right) \quad (2)$$

In our case the features used in the model were either dinucleotide counts (dinucleotide-based model) or counts of short motifs (motif-based model) in the regions upstream and downstream of miRNA precursors. L1-regularized logistic regression performs automatic feature selection penalizing dinucleotides or motifs which are not significant in distinguishing efficiently processed miRNAs from non-efficiently processed. We found that the most important features associated with enhanced processing are: a GNNU motif (N indicates any nucleotide) directly upstream of the 5' of the miRNA, a CNNC motif between 17 and 21 positions downstream of 3' of the miRNAs and dinucleotides GC and CU enriched at the base of the miRNA stem loop.

## 4 Conclusions

*In silico* methods for studying miRNA biogenesis, ranging from statistical models of promoter recognition and transcription factor binding site prediction to predictive models of miRNA processing, enable a better understanding of miRNA-mediated regulation in tissue-specific networks. Recent progress in the field of NGS resulted in a plethora of high-throughput and high-quality datasets in the last few years. This enabled the development of data-driven computational approaches which make use of such data and combine them with traditional sequence signals, in order to get more accurate prediction of miRNA promoters. Although the basics of the miRNA biogenesis pathway are known, there are still many unsolved questions. For example, several regulatory factors might be involved in miRNA regulation at different levels. Although some regulators of miRNA transcription and processing

have been predicted and experimentally validated, more sophisticated *in silico* methods are needed to discover more of these factors and predict how they affect miRNA biogenesis.

RNA binding proteins interact with both pri-miRNAs in addition to intermediate miRNA products at different stages of their regulation. High-throughput sequencing of RNA sites bound by a particular protein will reveal more aspects about miRNA regulation, as well as enable more reliable identification of targets which are physiologically relevant.

Although observations from different sources need to be unified in a coherent framework, it is clear that targeted computational approaches can help linking different evidence from several genomic datasets and give a significant contribution to discover additional details about miRNA-mediated regulation.

## References

1. Guo, Z.: Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci. Rep.* **4**, 5150 (2014)
2. Davis, B.N.: Regulation of MicroRNA biogenesis: a miRiad of mechanisms. *Cell Commun. Signal* **10**, 7–18 (2014)
3. Bartel, D.P.: MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009)
4. Plaisier, C.L.: A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. *Genome Res.* **22**, 2302–2314 (2012)
5. Esquela-Kerscher, A.: Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer* **6**, 259–269 (2006)
6. Takahashi, R.U.: The role of microRNAs in the regulation of cancer stem cells. *Front Genet* **4**, 295 (2014)
7. Davidson-Moncada, J.: MicroRNAs of the immune system: roles in inflammation and cancer. *Ann. N. Y. Acad. Sci.* **1183**, 183–194 (2010)
8. Ma, X.: MicrorNAs in NF-kappaB signaling. *J. Mol. Cell Biol.* **3**, 159–166 (2011)
9. Lewis, B.P.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005)
10. Betel, D.: Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11**, R90 (2010)
11. Sandelin, A.: JASPAR: an open-access database for eukaryotic transcription factor binding profiles *Nucl. Acids Res.* **32** D91–D94 (2004)
12. Krol, J.: The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* **11**, 597–610 (2010)
13. Fickett, J.: Eukaryotic promoter recognition. *Genome Res.* **7**, 861–878 (1997)
14. Marsico, A.: PROMiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol.* **14**, R84 (2013)
15. Monteys, A.M.: Structure and activity of putative intronic miRNA promoters. *RNA* **16**, 495–505 (2010)
16. Hinske, L.C.: A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics* **11**, 533 (2010)
17. Fujita, S.: Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates. *Bioinformatics* **24**, 303–308 (2008)

18. Kozomara, A.: miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152–D157 (2011)
19. Kozomara, A.: ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009)
20. de Hoon, M.: Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques* **44**, 627–628 (2008)
21. Core, L.J.: Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008)
22. Georgakilas, G.: microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat. Commun.* **10**, 5700 (2014)
23. Barski, A.: Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521–533 (2008)
24. Barski, A.: Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res.* **19**, 1742–1751 (2009)
25. Ozsolak, F.: Chromatin structure analyses identify miRNA promoters. *Gene Dev.* **22**, 3172–3183 (2008)
26. Chien, C.H.: Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res.* **39**, 9345–9356 (2011)
27. Megraw, M.: A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res.* **19**, 644–656 (2009)
28. Eis, P.: Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc. Natl. Acad. Sci.* **102**, 3627–3632 (2003)
29. Thomas-Chollier, M.: Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat. Protoc.* **102**, 3627–3632 (2003)
30. Uhlen, M.: Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015)
31. Szklarczyk, D.: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acid Res.* **39**, D561–568 (2011)
32. Ha, M.: Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–524 (2014)
33. Auyeung, V.C.: Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* **152**, 844–858 (2013)
34. Conrad, T.: Microprocessor activity controls differential miRNA biogenesis in vivo. *Cell Rep.* **9**, 542–554 (2014)

# Tunicate Neurogenesis: The Case of the *SoxB2* Missing CNE

Evgeniya Anishchenko and Salvatore D'Aniello

**Abstract** The discovery of the *SoxB2/Sox21* regulatory element, conserved from basal metazoa to human, opened novel perspectives to study the conservation among distant related genomes. This discovery represents exceptional maintenance of an almost identical enhancer structure controlling a gene that is fundamental for nervous system development. The activity of metazoan *SoxB2* enhancers was previously demonstrated in zebrafish embryos by cross-species experiments.

Here we tested the activity of human and amphioxus orthologue *cis*-regulatory sequences in embryos of the tunicate *Ciona intestinalis* through a transgenic approach, and found out that *SoxB2* enhancers retained their activity in neuronal differentiation even in a non-vertebrate chordate.

This result was unexpected since the conserved *SoxB2* enhancer was not found in *Ciona* in previous studies. Nevertheless, we adopted a different comparative approach and performed a phylogenetic footprinting analysis using two congeneric tunicate species, *C. intestinalis* and *Ciona savignyi*, that, in fact, evidenced a conserved *SoxB2* 3' element. The discovered element could potentially be the missing orthologous *SoxB2* enhancer previously identified in human, zebrafish, and amphioxus.

A detailed search for possible transcription factors revealed the massive presence of Sox, Pou and Fox binding sites as found in other deuterostomes. Nevertheless, whether the conserved *SoxB2* element of *Ciona* possesses a functional ability as gene transcriptional enhancer remains to be demonstrated experimentally.

**Keywords** Evolution • Transgenesis • Nervous system • *Ciona* • *Cis*-regulatory enhancers

---

E. Anishchenko • S. D'Aniello (✉)

Department of Biology and Evolution of Marine Organisms (BEOM), Stazione Zoologica Anton Dorhn, Villa Comunale, 80121 Napoli, Italy

e-mail: [salvatore.daniello@szn.it](mailto:salvatore.daniello@szn.it)

## 1 Introduction

One of the most intriguing mechanisms of nervous system (NS) development is related to the neuronal lineage specification, which is of great interest to science and subject of numerous and intensive studies. Nevertheless, we are still far from a complete understanding of these processes. The discovery of an evolutionary conserved non-coding element (CNE) in the animal kingdom by Royo et al. [16] started a discussion about this key aspect of neural state regulation in animal development. It was found to be an example of gene regulatory element conservation in all metazoans, from the cnidarian *Nematostella* to human. Only the exonic regions of genes were known to have such a degree of conservation among animals so different in body shape and complexity, diverging from a common ancestor around 600 My ago [14].

Royo and colleagues discovered a highly conserved CNE that regulates the *SoxB2* gene (SRY-box B), recognizable at the sequence level within metazoans, and explored its functional significance in transphyletic *cis*-regulatory DNA experiments. The invertebrate *SoxB2*, and the orthologous gene in vertebrates *Sox21*, are involved in neuronal development, differentiation and regeneration, indicating that these genes are responsible for the pluripotent features of presumptive neuronal tissues in animal [7, 9, 10, 15, 17–19, 21]. The sequence comparison of *SoxB2/Sox21* CNE among human (*Homo sapiens*), zebrafish (*Danio rerio*), amphioxus (*Branchistoma floridae*), acorn worm (*Saccoglossus kowalevskii*), sea urchin (*Strongylocentrotus purpuratus*) and cnidarian (*Nematostella vectensis*) showed an evolutionary conserved region of 200 bp, located at the 3' of the gene in all analysed loci [16]. Royo and collaborators demonstrated through transgenic experiments on zebrafish embryos that CNEs from diverse animal genomes were functional regulative elements for different stages of neurogenesis, including patterning and development of the vertebrate forebrain. Similarly, the reporter gene expression driven by human *SOX21* CNE and sea urchin *SoxB2* CNE was functional in developing the nervous system of *Drosophila*, despite absence of clear sequence orthology. This was the first study pointing to the fact that the regulatory state recognized by a conserved DNA sequence may have been redeployed at different levels of the developmental regulatory program during evolution of the complex central nervous system (CNS).

A detailed study focused on the regulation of *Sox21b* (fish ortholog of *SoxB2*) expression highlighted 19 regulatory DNA elements conserved between vertebrates (human, chicken, mouse, frog, zebrafish and fugu) [13]. Transgenic experiments using conserved fragments from the fugu genome in zebrafish showed that the majority of these CNEs were able to generate tissue-specific expression patterns in the CNS and sensory organs, in agreement with *Sox21b* expression domains. As expected, one of the enhancers analysed in this study corresponded to the evolutionary conserved element discovered by Royo and colleagues, the CNE17 in the *Sox21b* locus [13]. Nevertheless, CNE17 and CNE6 were the only enhancer elements able to drive the expression of the reporter gene in the lens, which

represents an innovation in vertebrates. A possible explanation for this could be that CNE17, orthologous to the highly conserved metazoan CNE, was co-opted in the fish lineage for the lens expression, as a consequence of the sub-functionalization of the two fish paralogs, *Sox21a* and *Sox21b* [10].

An evolutionary puzzling case remained to be solved. As mentioned above, one of *SoxB2/Sox21* enhancer was found to be conserved in highly distant related animals and transcriptionally active during CNS development, indicating a key role in nervous system evolution. Nevertheless, in previous studies it was not possible to detect any trace of the *SoxB2* enhancer conservation in the lineage of tunicates, the sister group of vertebrates [5] which are considered important model systems for the study of evolution and development in chordates. Tunicates, differently from cephalochordates, are highly diverged from the common chordate ancestor, both morphologically and genetically, and this represents an additional difficulty for evolutionary biologists that take advantage of homologies between body structures and sequence conservation as main principles. Here we tried, therefore, to reveal the potentiality of tunicates in our understanding of deuterostome NS evolution.

The ascidian *C. intestinalis* represents a very useful animal model to perform *in vivo* transgenic assays because it possesses most of the molecular pathways and gene repertoire as the rest of chordates. Nevertheless, the *Ciona* genome shows divergent characteristics that sometimes can represent a limitation to experimental approaches and on the other hand species-specific genomic events, such as gene loss, can be considered an experimental advantage in evolutionary devoted studies.

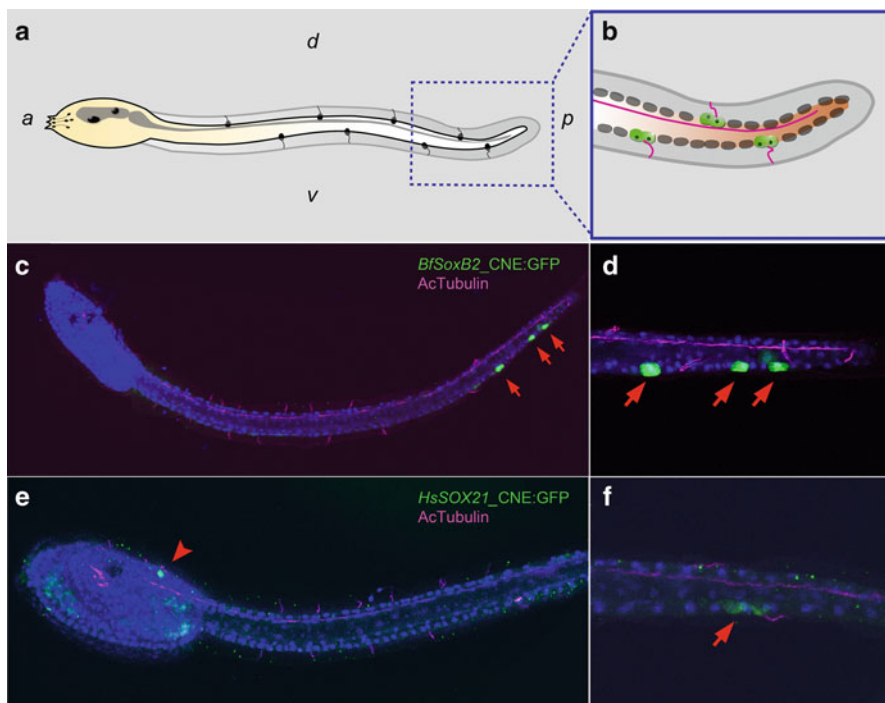
## 2 Results

Two main evolutionary questions prompted us to choose *C. intestinalis* as the model organism for this study, taking into account the advantage of the easy application of transgenesis approaches that are very well established for *Ciona*.

First, is the ascidian embryonic transcription factors (TF) machinery able to recognize the transcriptional information contained in cross-species enhancers, considering the loss of the evolutionary conserved *SoxB2* CNE? Second, could the presence of the conserved *SoxB2* CNE be masked at sequence level by the highly divergent genome of ascidians?

To answer these questions, that are interesting per se from an evolutionary point of view, we performed a series of computational and transgenic experiments. To understand whether the regulation of the *SoxB2* enhancer is maintained in *Ciona*, despite the loss of the orthologous region, we carried out transgenic experiments in *C. intestinalis*, introducing exogenous DNA regulative fragments in developing embryos. More in detail, we used the technique of transgenesis by electroporation of a purified plasmid containing the putative enhancer with a GFP reporter gene into fertilized eggs. In the first series of *in vivo* experiments we used CNE fragments, amplified by PCR on genomic DNA, corresponding to *SoxB2/Sox21* CNEs from



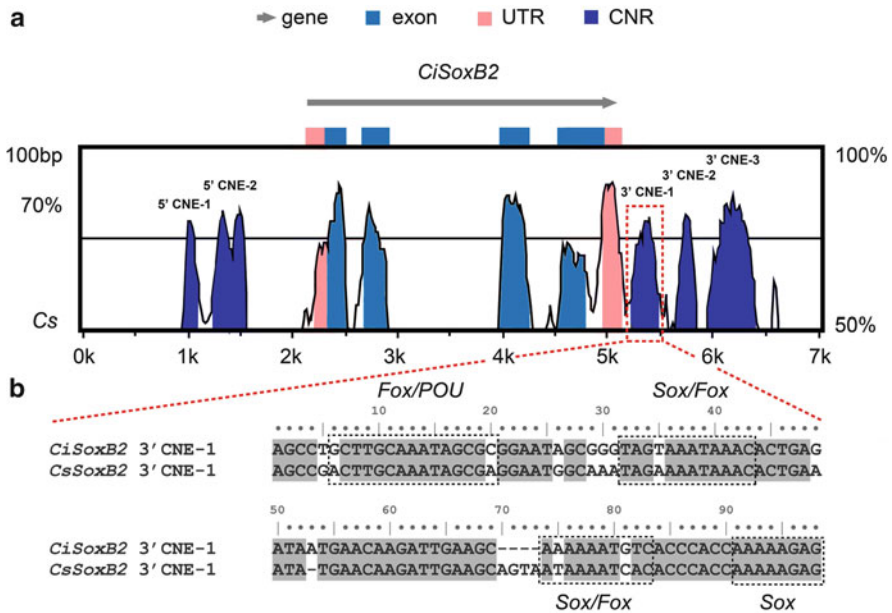


**Fig. 1** Transgenic larvae using human and amphioxus cnes driving GFP expression in neural territories. **(a)** *Ciona intestinalis* 20 hpf larvae body plan in a schematic representation: *a* anterior, *p* posterior, *d* dorsal, *v* ventral. The CNS is indicated in grey. **(b)** Magnification of larval tail. Pairs of caudal epidermal neurons are indicated in green. **(c, d)** amphioxus *SoxB2* CNE drives GFP expression in *Ciona* ectodermal neurons in the tail (red arrows). **(e, f)** Human *Sox21* CNE resulted active in *Ciona* ectodermal neurons in the tail (red arrows) and in the head (red arrowhead)

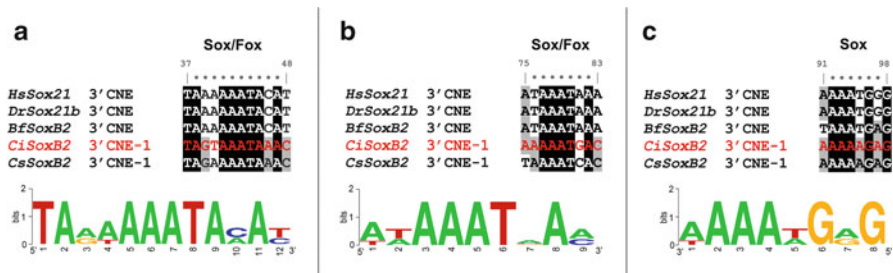
different animal models: acorn worm *S. kowalewskii* (hemichordate), sea urchin *S. purpuratus* (echinoderm), *B. floridae* (cephalochordate) and human *H. sapiens* (vertebrate). These DNA fragments correspond to the enhancers previously used in transgenic experiments on zebrafish by Royo et al. [16].

Transgenic experiments using *B. floridae* and human *H. sapiens* CNEs gave positive results as shown in Fig. 1. The *SoxB2* enhancer from *B. floridae* was able to drive the expression of the GFP reporter gene in paired tail epidermal neurons in about 70 % of the larvae (red arrows in Fig. 1c, d). This result was confirmed by the human *SOXB2* enhancer activity in about 60 % of the larvae (red arrow in Fig. 1f), albeit that we recorded fainter signals. Interestingly, the human DNA construct was also able to drive the GFP expression in another neuronal compartment in the head region (arrowhead in Fig. 1e). In Fig. 1a we present a diagram of the *Ciona* larval body plan with the central nervous system in grey. Figure 1b is a magnification of the tail region showing in green the paired epidermal neurons, GFP positive with both the human and cephalochordate exogenous DNA constructs.

To answer the second evolutionary question we performed phylogenetic footprinting analyses that showed an extremely low degree of conservation between *C. intestinalis* and other deuterostomal orthologous regions containing the conserved *SoxB2* CNE. Nevertheless, to go deeper into this comparative analysis and trying to find the orthologous CNE in tunicates, we compared two congeneric species that diverged 180 My ago [1], with the aim to reveal conserved non-coding regions that were unrecognizable when searched between tunicates and distant related species. We therefore performed a Vista analysis that allowed to discover three main regions highly conserved between *C. intestinalis* and *C. savignyi* in the 3' of *SoxB2* (Fig. 2a), that could correspond to the orthologous region previously found conserved in other deuterostomes. A local alignment was performed showing a very high degree of sequence conservation between the two *Ciona* species, at least 70 % identical in non-coding regions (Fig. 2b). Hence, a detailed in silico analysis was performed using Jaspas software in order to reveal potential transcription factor binding sites (TFBS), and compare them with those predicted in sea urchin, amphioxus and human *SoxB2* enhancers [16]. This allowed the identification of a cluster of several Sox and Fox binding sites in the 3' CNE-1 peak, which was not detected in other 5' and 3' CNEs (Figs. 2b and 3a-c).



**Fig. 2** Phylogenetic footprinting and *Ciona*'s CNE alignment. **(a)** Vista analysis between *SoxB2* loci of *Ciona intestinalis* and *Ciona savignyi*. Dark blue peaks represent conserved non-coding elements between the two species, pink indicates the *SoxB2* 5' and 3' UTRs and blue the exons. **(b)** Alignment of *Ciona*'s *SoxB2* 3' CNE-1. Potential binding sites for Pou, Sox and Fox are highlighted by frames



**Fig. 3** Diagrams of TF binding sequences conserved in deuterostomes. *Ciona*'s *SoxB2* 3' CNE-1 contains Sox and Fox (a, b) and Sox (c) binding sites that are found in orthologous sequences from *Ciona savignyi*, amphioxus, zebrafish and human. Black background indicates a 100 % match of identity between all species considered

Three short sequence fragments were found to be highly conserved in 3' CNE-1, which could be the potential binding targets for Sox and Fox (Fig. 3a–c). The level of conservation of the multiple sequence alignment using other chordates (amphioxus, zebrafish and human) was 67 % (Fig. 3a), 56 % (Fig. 3b) and 63 % (Fig. 3c).

### 3 Discussion

The regulatory landscape of genes involved in developmental processes is constrained by enhancers that remained conserved during evolution. Therefore, *cis*-regulatory elements conserved between orthologous genes in vertebrates have been readily recognized in comparative genomic studies as soon as multiple genomes sequencing projects become available. The exceptional case of the discovery of an ancient enhancer retained in metazoans has opened new perspectives in the research field of *cis*-regulatory elements.

The direct comparison between distantly related animals can be inconclusive when the degree of nucleotide conservation is low, while on the contrary the choice of congeneric species is often fruitless because the high homology becomes uninformative in the search for non-coding active elements. In this perspective, the availability of genomes from numerous metazoan species help greatly in reconstruction of metazoan evolution. We applied a transgenic approach using human *SOX21* and amphioxus *SoxB2* enhancers exogenously in *C. intestinalis* embryos and more important we demonstrated that they were functional in proneural tissues, as previously demonstrated in a related study on zebrafish and *Drosophila* embryos. Here we found the putative ancestral enhancer of the *SoxB2* gene by comparing two tunicates, which was thought to be lost in such fast evolving genomes. A detailed bioinformatics search in the conserved non-coding regions on

*Ciona*'s *SoxB2* loci revealed a cluster of four TFBS of the Sox and Fox class in the 3' CNE-1 (Figs. 2b and 3a–c), who correspond to the *SoxB2* CNEs reported by Royo et al. [16].

However recent studies demonstrated that the ancestral regulatory function of *SoxB2* CNE is still conserved, despite the lack of sequence similarity among different phyla [6, 11]. Furthermore, similar to our results of *Ciona* transgenic experiments, the human *SOX21* and sea urchin *SoxB2* CNEs were demonstrated to be functional in the neuroblasts of the presumptive brain and ventral nerve cord of *D. melanogaster* embryos [16]. These transgenic approaches highlighted the deep functional conservation of metazoan *SoxB2* CNEs in neurogenesis, not only in species possessing high sequence similarity but also in animals showing significantly divergent *SoxB2* regulatory elements. Recently a finding was reported of so called FCNEs (Functional Conserved Non-coding Elements) concerning those *cis*-regulatory elements that, despite a low sequence similarity across distant related species, still keep the ancestral function during developmental processes [20].

The potential transcriptional activity of the *SoxB2* CNEs identified in the present study in two *Ciona* species, despite missing a high degree of sequence similarity with other deuterostomes, remains to be experimentally confirmed in future studies.

## 4 Materials and Methods

### 4.1 Animals and Embryos

Adult individuals of *C. intestinalis* used in this study were collected from the Gulf of Naples (Italy) and kept in tanks at 18 °C until further use. To prevent spontaneous spawning in captivity, ripe animals were exposed to continuous light. Gametes were collected from the gonoducts of several animals and used for in vitro fertilization.

### 4.2 Comparative Genomics

To obtain DNA sequences for *SoxB2* loci, a series of databases was used: ANISEED database ([www.aniseed.cnrs.fr/](http://www.aniseed.cnrs.fr/)) for *C. intestinalis* and *C. savignyi* sequences; SpBase ([www.spbase.org](http://www.spbase.org/)) for sea urchin *S. purpuratus*; JGI (<http://genome.jgi-psf.org/Brafl1/Brafl1.home.html>) for amphioxus *B. floridae*, and NCBI (<http://www.ncbi.nlm.nih.gov/>) for human and acorn worm *S. kowalevskii* sequences.

### 4.3 *Phylogenetic Footprinting and in Silico Analyses*

Genomic sequences from the two congeneric *Ciona* species, including *SoxB2* locus plus 5 kb upstream and 5 kb downstream, were aligned using the AVID software [2]. Sequences were compared using mVISTA ([8]; <http://genome.lbl.gov/vista/mvista/submit.shtml>), with the following parameters: 100 bp of fragment length with 70 % of sequence identity.

In order to reveal TFBSs in *SoxB2* CNEs, dna sequences were analysed using Jaspar (<http://jaspar.genereg.net/>), a TF binding profile database [12]. Diagrams of POU, Sox and Fox binding sequences conserved between two *Cionas*, human and amphioxus were generated using WebLogo software [3].

### 4.4 *Transgenic Experiments*

Four CNEs from *S. kovalenskii*, *S. purpuratus*, *B. floridae* and *H. sapiens*, were amplified by PCR and cloned in the pSP72:CNE:2XGFP:SV40 vector, containing the GFP reporter gene and SV40 polyadenylation sequence. *C. intestinalis* transgenic embryos were obtained via electroporation experiments, as previously described [4], and observed with confocal microscopy after immunohistochemical detection. Each experiment was performed in triplicate, comparing at least 100 embryos for each single construct. Briefly, eggs were dechorionated, before fertilization to be ready to incorporate the exogenous DNA using a solution containing: 1 % sodium thioglycolate, 0.05 % proteinase E and 1N sodium hydroxide (NaOH), and afterwards washed in filtered sea water (FSW). The 200  $\mu$ l of dechorionated and fertilized eggs were transferred into Bio-Rad Gene Pulser 0.4 cm cuvettes containing a 0.77 M mannitol solution and 100  $\mu$ g of the exogenous DNA plasmids, and subsequently electroporated using a Bio-Rad Gene Pulser II<sup>TM</sup> with the following settings: constant 50 V and 800  $\mu$ F. Electroporated eggs were transferred into petri dishes with 1 % agarose bottom with FSW and let develop at 18 °C until the desired developmental stage.

### 4.5 *Whole Mount Immunohistochemistry*

Embryos were fixed in 4 % formaldehyde during 30 min at room temperature and washed with PBT (PBS 1x, 0.1 % Tween20). Embryos were dehydrated gradually in 70 % ethanol, followed by rehydration in PBS 1x four times. To permeabilize the embryos, they were incubated in PBS containing 0.01 % Triton-100 for 30 min. Embryos were incubated in blocking buffer (PBS 1x, 0.01 % Triton-100, 30 % goat serum) over night. Next, embryos were kept in blocking buffer with 1:300 polyclonal anti-GFP Ab from rabbit (TP401; Torrey Pines Bionabs) and 1:300

monoclonal anti-Acetylated Tubulin (AcTubulin) Ab from mice (T7451; Sigma) for 2 days at 4 °C and subsequently washed with PBT changing the solution every 15 min for 4 h. Then, embryos were incubated with the secondary anti-mouse Alexa 488 Ab or anti-rabbit Alexa 633 Ab in PBT (1:500), over night at 4 °C, then washed in PBT and incubated with DAPI (D9542; Sigma) 1:10<sup>4</sup> in PBT for 10 min. Laser scanning confocal images were obtained with a Zeiss LSM 510 META confocal microscope.

**Acknowledgments** The authors are grateful to Maria Ina Arnone, Margherita Branno, Annamaria Locascio, Filomena Ristoratore and Antonietta Spagnuolo for their suggestions about in vivo experiments and for sharing with us DNA plasmids for transgenesis. We thank Mara Francone for technical assistance with DNA maxi-preparation, the Marine Resources for Research Unit of Stazione Zoologica Anton Dohrn for animal fishing and maintenance. Heartfelt gratitude to Rita Marino for her excellent suggestions concerning immunohistochemical experiments. Evgeniya Anishchenko has been supported by a SZN PhD fellowship (2011–2014). This work was supported by a Marie Curie Career Integration Grant (FP7-PEOPLE-2011-CIG, PCIG09-GA-2011-293871) to Salvatore D’Aniello.

## References

1. Bernà, L., Alvarez-Valin, F., D’Onofrio, G.: How fast is the sessile *Ciona*? *Comp. Funct. Genomics* (2009). 875901
2. Bray, N., Dubchak, I., Pachter, L.: AVID: a global alignment program. *Genome Res.* **13**, 97–102 (2003)
3. Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E.: WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004)
4. D’Aniello, S., D’Aniello, E., Locascio, A., Memoli, A., Corrado, M., Russo, M.T., Aniello, F., Fucci, L., Brown, E.R., Branno, M.: The ascidian homologue of the vertebrate homeobox gene *Rx* is essential for ocellus development and function. *Differentiation* **74**, 222–234 (2006)
5. Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H.: Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968 (2006)
6. Doglio, L., Goode, D.K., Pelleri, M.C., Pauls, S., Frabetti, F., Shimeld, S.M., Vavouri, T., Elgar, G.: Parallel evolution of chordate cis-regulatory code for development. *PLoS Genet.* **9**, e1003904 (2013)
7. Ferrero, E., Fischer, B., Russell, S.: *SoxNeuro* orchestrates central nervous system specification and differentiation in *Drosophila* and is only partially redundant with *Dichaete*. *Genome Biol.* **15**, R74 (2014)
8. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I.: VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279 (2004)
9. Kamachi, Y., Kondoh, H.: Sox proteins: regulators of cell fate specification and differentiation. *Development* **140**, 4129–4144 (2013)
10. Lan, X., Wen, L., Li, K., Liu, X., Luo, B., Chen, F., Xie, D., Kung, H.F.: Comparative analysis of duplicated *Sox21* genes in zebrafish. *Dev. Growth Differ.* **53**, 347–356 (2011)
11. Maeso, I., Irimia, M., Tena, J.J., Casares, F., Gómez-Skarmeta, J.L.: Deep conservation of cis-regulatory elements in metazoans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130020 (2013)

12. Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., Wasserman, W.W.: JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2013)
13. Pauls, S., Smith, S.F., Elgar, G.: Lens development depends on a pair of highly conserved *Sox21* regulatory elements. *Dev. Biol.* **3665**, 310–318 (2012)
14. Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., Jurka, J., Genikhovich, G., Grigoriev, I.V., Lucas, S.M., Steele, R.E., Finnerty, J.R., Technau, U., Martindale, M.Q., Rokhsar, D.S.: Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94 (2007)
15. Reiprich, S., Wegner, M.: From CNS stem cells to neurons and glia: *Sox* for everyone. *Cell Tissue Res.* **359**, 111–124 (2014)
16. Royo, J.L., Maeso, I., Irimia, M., Gao, F., Peter, I.S., Lopes, C.S., D'Aniello, S., Casares, F., Davidson, E.H., Garcia-Fernández, J., Gómez-Skarmeta, J.L.: Transphyletic conservation of developmental regulatory state in animal evolution. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 14186–14191 (2011)
17. Sandberg, M., Kallstrom, M., Muhr, J.: *Sox21* promotes the progression of vertebrate neurogenesis. *Nat. Neurosci.* **8**, 955–1001 (2005)
18. Taguchi, S., Tagawa, K., Humphreys, T., Satoh, N.: Group B *Sox* genes that contribute to specification of the vertebrate brain are expressed in the apical organ and ciliary bands of hemichordate larvae. *Zoolog. Sci.* **19**, 57–66 (2002)
19. Uchikawa, M., Yoshida, M., Iwafuchi-Doi, M., Matsuda, K., Ishida, Y., Takemoto, T., Kondoh, H.: B1 and B2 *Sox* gene expression during neural plate development in chicken and mouse embryos: universal versus species-dependent features. *Dev. Growth. Differ.* **53**, 761–771 (2011)
20. Vassalli, Q.A., Anishchenko, E., Caputi, L., Sordino, P., D'Aniello, S., Locascio, A.: Regulatory elements retained during chordate evolution: coming across tunicates. *Genesis* **53**, 66–81 (2015)
21. Whittington, N., Cunningham, D., Le, T.K., De Maria, D., Silva, E.M.: *Sox21* regulates the progression of neuronal differentiation in a dose-dependent manner. *Dev. Biol.* **397**, 237–247 (2015)

# MECP2: A Multifunctional Protein Supporting Brain Complexity

Marcella Vacca, Floriana Della Ragione, Kumar Parijat Tripathi,  
Francesco Scalabrì, and Maurizio D'Esposito

**Abstract** After more than 20 years from its discovery, MECP2 roles are far from the fully understanding. MeCP2 binds the genome globally, with the need of a single, methylated CG and is enriched in heterochromatic foci. Early hypothesis proposed it as a generalized repressor and modulator of genome architecture that keeps down the transcriptional noise. Its modulation of L1 retrotransposition and the regulation of pericentric heterochromatin condensation might be conceivably associated with this function. Interestingly, MECP2 is mutated in the paradigmatic chromatin disease Rett syndrome, an X linked neurodevelopmental disease affecting females. This highlighted a different function of MECP2, as repressor of downstream genes and the identification of few downstream genes corroborated this hypothesis. Rather recently, however, with the help of high throughput technologies and a number of appropriate mouse models finely dissecting MECP2 functional domains, new and somehow unexpected roles for MECP2 have been highlighted. Expression profiling studies of specific brain areas support a role of MeCP2 not only as a transcriptional silencer but also as activator of gene expression. Beyond its binding to DNA, MeCP2 is also able to influence alternative splicing, promoting inclusion of hypermethylated exons in alternatively spliced transcripts. MeCP2 has been also found to bind non CG methylated residues in brain. Overall, MECP2 appears to be a multifunctional protein, exquisitely adapted to support the functional complexity of the brain.

---

M. Vacca

Institute of Genetics and Biophysics, "A. Buzzati Traverso", CNR, Naples, Italy

e-mail: [marcella.vacca@igb.cnr.it](mailto:marcella.vacca@igb.cnr.it)

F. Della Ragione • M. D'Esposito (✉)

Institute of Genetics and Biophysics, "A. Buzzati Traverso", CNR, Naples, Italy

IRCCS Neuromed, Pozzilli (Is), Italy

e-mail: [floriana.dellaragione@igb.cnr.it](mailto:floriana.dellaragione@igb.cnr.it); [maurizio.desposito@igb.cnr.it](mailto:maurizio.desposito@igb.cnr.it)

K.P. Tripathi

ICAR-CNR, Naples, Italy

e-mail: [kumpar@na.icar.cnr.it](mailto:kumpar@na.icar.cnr.it)

F. Scalabrì

IRCCS Neuromed, Pozzilli (Is), Italy

e-mail: [francesco.scalabrì@neuromed.it](mailto:francesco.scalabrì@neuromed.it)



**Keywords** Neuro-developmental disease • Rett syndrome • MeCP2 • Splicing • DNA methylation

## 1 Mouse Models and Their Contributions to Disentangle MECP2 Functions

MeCP2 (methyl-CG binding protein 2) is an ubiquitous transcription factor encoding two different splicing isoforms, MeCP2A and MeCP2B (Fig. 1a), both containing two main domains, Methyl-Binding Domain (MBD) and Transcriptional Repression Domain (TRD) and predominantly expressed in brain [8]. It is mutated in Rett syndrome (RTT, OMIM 321750), a progressive neurodevelopmental disorder affecting almost exclusively females [36]. The first murine model carrying constitutive ablation of *Mecp2* gene showed embryonic lethality [41], but, just after the cloning of RTT causative gene [3] several new mouse models have been generated helping to depict MECP2 biological role. By crossing floxed animals with Cre deleter mice, ubiquitously expressing Cre transgene, mice lacking MeCP2 in all tissues were provided. Floxed mice were crossed also with Nestin-Cre mice, generating a progeny lacking MeCP2 selectively in the brain. Both *Mecp2*-null constitutive and Nestin-Cre conditional mutants recapitulate symptomatic manifestations of RTT: they are apparently healthy and fertile for the first few weeks of age but develop neurological phenotype at 5–6 weeks and die at 10–12 weeks of age. These findings strongly suggest a primary role of MeCP2 in the brain [13, 18]. Female mice, representing the true RTT model, develop symptoms at 12 weeks of age and survive beyond 12 months. The overall brain structure is conserved in the absence of MeCP2 but brains are smaller than age-matched controls due to reduced neuronal size [13] and neural dendritic arborization [34] as observed in humans. MECP2 absence impairs both excitatory and inhibitory transmission in neurons. Like many other X-linked intellectual disability genes [5], MECP2 impacts also on dendritic spine density and synaptic plasticity.

A helpful model to study the effect of truncating mutation similar to those found in RTT patients is the *Mecp2*-308/Y mouse, carrying a mutation that introduces a premature stop codon. This model shows progressive RTT-like neurological phenotypes as well, but symptoms onset and age of death are significantly delayed. Moreover, these mice show hyperacetylation of histone H3, suggesting again that MeCP2 dysfunction has an effect on chromatin architecture [38]. Noteworthy, in the last years, murine models have been generated carrying *Mecp2* mutations for loss and/or gain of function in specific brain regions or sub-neuronal populations. This detailed analysis allowed to hypothesize that specific phenotypes observed in RTT models may be ascribed to different brain compartments [10].

Lately, also knock-in (KI) mice carrying disease causing point mutations or MeCP2 derivatives (i.e. protein forms no more phosphorylable in specific amino

acids) have been generated. For example, mice expressing MeCP2 with the common RTT causing mutation R306C or those carrying phosphorylation-defective T308A derivative have been useful to demonstrate the importance of the interaction between MeCP2 and the NCoR complex and how this is regulated by the activity dependent T308 phosphorylation. MeCP2 R306C mutant cannot neither binds the NCoR complex nor be phosphorylated in T308, whereas the MeCP2 T308A phosphorylation-defective derivative constitutively binds the NCoR complex, independently from neuronal activity [15]. It has been noticed that MECP2 R306C KI mice, compatible with a model of MECP2 loss of function, are more dramatically affected than T308A KI mice, compatible with a model of MECP2 gain of function, even if both models exhibit Rett like features [26].

## 2 MECP2 Deficiency and Changes in Transcriptional Profiling

Despite the hypothesized role of MeCP2 as a transcriptional repressor, transcriptome profiling of total brains from *Mecp2*-null mice revealed only slight changes in gene expression [11]. On the contrary, dysregulation of thousands of genes came out by profiling specific brain regions relevant to RTT symptoms, such as the hypothalamus and cerebellum of *Mecp2*-null and -overexpressing animals (Fig. 1b, left panel, i). Interestingly, these studies support a role of MeCP2 not only as a transcriptional silencer but also as an activator of gene expression, through its association with the transcriptional activator CREB1 [6, 12] (Fig. 1b, left panel, ii). The brain-derived neurotrophic factor (*Bdnf*) gene, encoding a signalling molecule with crucial roles in brain development and neuronal plasticity has been found consistently deregulated in the absence of MeCP2 and is thus considered a bona fide MeCP2 target gene. It has been shown that in resting neurons MeCP2 is bound to methylated promoter of BDNF, while in depolarized neurons, which cause BDNF activation, MeCP2 become phosphorylated in Ser421 and dissociates from BDNF promoter [14, 27]. More recently it has becoming clear that a dual operation model could explain the MeCP2 dependent-BDNF expression control [24].

Not surprisingly, among the MeCP2 targets identified in two different transcriptional profiling studies of *Mecp2*-null brains there are several non-coding RNAs, including hundreds of miRNAs [19] and long non-coding RNAs (lncRNAs) [35]. Even if a major challenge is to understand the molecular consequences of deregulated lncRNAs, the GABA receptor subunit Rho 2 gene was proposed as an interesting target. Altogether, these findings suggest a global role of MeCP2 in the transcriptional regulation of many classes of genes, underlying the importance of its integrity and the devastating effect of its mutations in RTT patients.

### 3 MECP2 Global Role in Genome Architecture

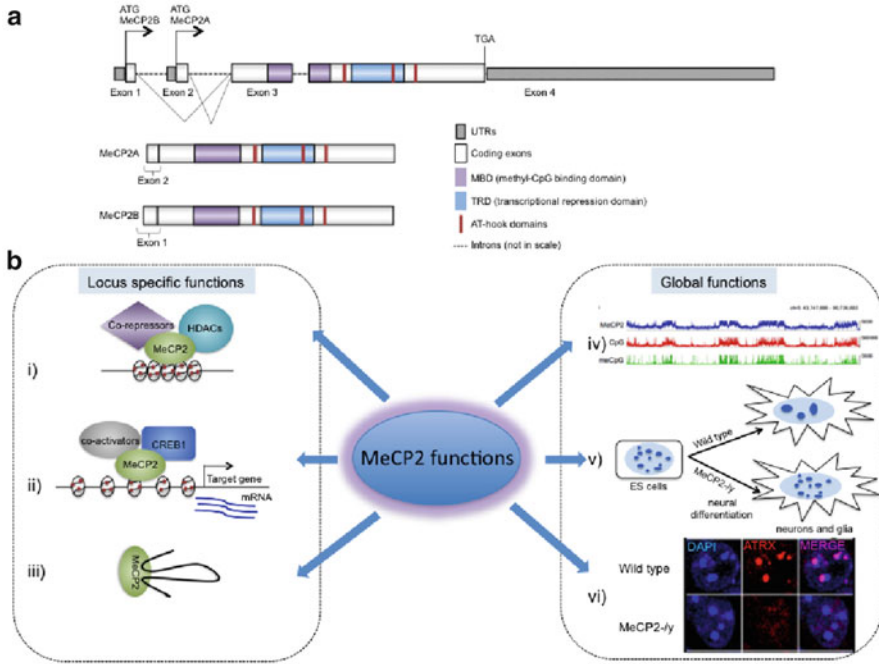
A clue to understand the crucial role of MeCP2 for brain functions has been provided by large-scale analysis of MeCP2 distribution. Skene and co-workers [39] reported that in neuronal nuclei, as opposed to other cell-types, MeCP2 is very abundant: its levels approach those of the histone octamer allowing the protein to be genome-wide bound in these cells, tracking methylated-CG moieties (Fig. 1b, right panel, iv). In neurons, MeCP2 may therefore act as a global organizer of chromatin structure, which is also supported by the fact that brains of *MeCP2*-null mice are characterized by increased histone acetylation and a doubling of histone H1 levels. Moreover, an intriguing finding is that the lack of MeCP2 in neurons from mature brain is responsible for the de-repression of spurious transcription of repetitive elements, such as L1 retrotransposon [32, 39].

The association of MeCP2 with chromatin seems also to be involved in generating higher order chromatin structures. In part, the silencing of an imprinted gene cluster on chromosome 6, including *Dlx5* and *Dlx6*, was proposed to depend on the formation of a MeCP2-dependent chromatin loop enriched in methylated H3 lysine 9 (H3K9), a mark of silent chromatin (Fig. 1b, left panel, iii) [8]. Moreover, MeCP2 accumulates at pericentromeric heterochromatin containing densely methylated major satellite DNA, forming specific chromatin structures called chromocenters [23]. It was revealed a crucial role of MeCP2 and the necessity of the MBD for the condensation of these chromatin structures during myogenic differentiation [9], and later this has been demonstrated also during neural differentiation (Fig. 1b, right panel, v) [7]. Furthermore, mutated MeCP2 forms (carrying different mutations frequently found in RTT) fail to correctly localize in heterochromatin, and many of them are unable to induce a correct chromocenter clustering [2].

More recently, MeCP2 has been proposed to be the major 5-hydroxymethylcytosine (5hmC)-binding protein in brain [31]; the high abundance of 5hmC in neurons and in particular in the gene body of transcribed genes probably ensures a cell specific epigenetic control of MeCP2 on chromatin structure and gene expression (see also Chap. 5). These findings parallel new discoveries suggesting that post-translational modifications of MeCP2 (i.e. phosphorylation) in response to multiple stimuli may provide novel keys to understand how MeCP2 can specifically modulate neuronal chromatin remodelling in response to neuronal activity [15].

### 4 MECP2 and Regulation of Alternative Splicing

An exciting developing area of research is highlighted by the complex relationships between epigenetics and splicing regulation. 5-hydroxymethylcytosine is highly enriched at the exon-intron junction in the brain, while 5-methyl cytosine (5-mC) is enriched at the exon-intron junction in non-neuronal cells [44]. Interestingly, in non-neuronal context MeCP2 has been found enriched in highly methylated



**Fig. 1** (a) Schematic representation of MECP2 gene structure (*upper panel*). The main protein domains are indicated with *different colors*. Alternative splicing producing the two isoforms, MeCP2A or MeCP2-beta (486 amino acids) and MeCP2B or MeCP2-alpha (498 amino acids), is indicated by *solid lines*. (b) Main MeCP2 functions are schematized. The *left panel* reassumes locus-specific roles: MeCP2 is able to repress the transcription of specific target genes by recruiting co-repressors and histone deacetylases (i) [14, 22, 33], to activate the transcription of target genes in the hypothalamus by binding co-activators and CREB1 (ii) [12] and to silence Dlx5/6 imprinted locus by promoting the formation of a higher order chromatin loop (iii) [20]. The *right panel* summarizes the main global roles: MeCP2 is globally distributed in the mouse neurons tracking the methyl-CG density (iv) [39], it plays a crucial role for the chromocenter clustering during neural differentiation of mouse ES cells (v) [7] and is important for the correct sub-nuclear localization of ATRX protein in the brain (vi) [4]

included alternatively spliced exons [28]. Also histone modifications can influence alternative splicing: it was proposed that loss of HDAC1 activity increased histone H4 acetylation surrounding alternative exons [44]. In Rett patients, expression of MECP2 mutated alleles is specifically associated with an increased monoacetylation level of H4 [42]. In turn, this phenomenon may result in over-expression of MeCP2 target genes providing functional implications in RTT pathogenesis.

Interestingly, tri-methylation of H3 lysine 9 (H3K9-me3) is a functional histone mark to recruit the heterochromatin protein HP1 and foster the inclusion of alternative exons [44]. Remarkably, MeCP2 physically interacts with HP1 proteins [1] and H3K9-me3 is mainly enriched at pericentric heterochromatin, an already known landscape for MeCP2 binding.

Recent findings highlighted that chromatin remodeling is mediated also by lncRNAs [37]. LncRNAs are involved, indeed, in the recruitment of epigenetic factors to specific genomic loci [43]. In brain tissues MeCP2 binds a number of lncRNAs as RNCR3 and MALAT1, this latter interacts with splicing factors too [29]. These data provide evidence that MeCP2 could be a bridge between epigenetic modification and alternative splicing regulation, taking also into account that MeCP2 binds several spliceosome components [29].

## 5 MECP2 Functions and the Brain DNA Methylation Landscape

Emergence of new experimental approaches analyzing the genome-wide single base resolution profiling of DNA methylation and hydroxymethylation [25] has made feasible to reconsider reading-mechanisms of DNA methylation signature. Not surprisingly, spotlight has been focusing on MECP2 to better define its role both in physiological and pathological conditions, such as Rett syndrome.

Lister and co-workers report an extensive DNA methylation re-assessment during postnatal mouse development. New roles are emerging for non CG methylation, such as CH methylation (mCH, in which H = A, C or T) and hydroxymethylation. In the latter case, genes losing CG methylation thus acquiring hmC signature, become active. Conversely, CH-methylation in neurons is depleted in expressed genes, representing an additional marker of gene repression.

In human and mouse CNS neurons mCH level significantly rises during brain postnatal development, reaching levels as abundant as methylated CGs [21, 25]. DNA hydroxymethylation is also enriched in neurons, 10 times more than in embryonic stem cells, with a postnatal increase. 5hmC profiling revealed its enrichment in gene bodies of expressed genes concomitantly to depletion around transcriptional start sites. If 5hmC represents a stable epigenetic mark or an intermediate molecule, tagging active sites of DNA demethylation, remains to be clarified [21, 25].

Quantitative modulation of CG and non-CG methylation in brain mirrors those of specific epigenetic factors, primarily methyl binding proteins. For instance, MeCP2 level increases synchronously with mCH and 5hmC rate [21, 39]. Additionally, MeCP2 is capable to bind mCH and repress transcription [17], in contrast to earlier experiments reporting a strong preference for mCG [30]. This discrepancy has been recently clarified by testing the binding capacity of MeCP2 towards all known forms of methylated DNA. Gabel and colleagues demonstrated that MeCP2 binds efficiently mCG and poorly 5hmCG in contrast to what reported previously [31]; actually, it shows higher affinity binding to mCA and hmCA [16]. Furthermore, MeCP2 binding to mCA is biased towards long genes expressed in brain; these genes become up-regulated upon MECP2 mutations, possibly causing neurological symptoms of Rett syndrome [16]. Similarly, neuronal overexpression of long genes has been already noticed in a MeCP2-loss of function mouse model [40]. Thus, transcriptional up-regulation of long genes is becoming a specific feature of Rett brain, over other neurological pathologies [16].

**Acknowledgements** The authors gratefully acknowledge Epigenomics Flagship Project EPIGEN MIUR-CNR and Associazione Italiana Sindrome di Rett. FS is recipient of a Neuromed fellowship.

## References

1. Agarwal, N., Hardt, T., Brero, A., Nowak, D., Rothbauer, U., Becker, A., Leonhardt, H., Cardoso, M.C.: Mecp2 interacts with Hp1 and modulates its heterochromatin association during myogenic differentiation. *Nucleic Acids Res.* **35**, 5402–5408 (2007)
2. Agarwal, N., Becker, A., Jost, K.L., Haase, S., Thakur, B.K., Brero, A., Hardt, T., Kudo, S., Leonhardt, H., Cardoso, M.C.: MeCP2 Rett mutations affect large scale chromatin organization. *Hum. Mol. Genet.* **20**, 4187–95 (2011)
3. Amir, R.E., Van Den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., Zoghbi, H. Y.: Rett syndrome is caused by mutations in X-linked Mecp2, encoding methyl-Cpg-binding protein 2. *Nat. Genet.* **23**, 185–188 (1999)
4. Baker, S.A., Chen, L., Wilkins, A.D., Yu, P., Lichtarge, O., Zoghbi, H.Y.: An at-hook domain in Mecp2 determines the clinical course of Rett syndrome and related disorders. *Cell* **152**, 984–996 (2013)
5. Bassani, S., Zapata, J., Gerosa, L., Moretto, E., Murru, L., Passafaro, M.: The neurobiology of X-linked intellectual disability. *Neuroscientist* **19**, 541–552 (2013)
6. Ben-Shachar, S., Chahrouh, M., Thaller, C., Shaw, C.A., Zoghbi, H.Y.: Mouse models of Mecp2 disorders share gene expression changes in the cerebellum and hypothalamus. *Hum Mol Genet.* **18**, 2431–2442 (2009)
7. Bertulat, B., De Bonis, M.L., Della Ragione, F., Lehmkuhl, A., Mildner, M., Storm, C., Jost, K.L., Scala, S., Hendrich, B., D’esposito, M., Cardoso, M.C.: Mecp2 dependent heterochromatin reorganization during neural differentiation of a novel Mecp2-deficient embryonic stem cell reporter line. *Plos One* **7**, E47848 (2012)
8. Bienvenu, T., Chelly, J.: Molecular genetics of Rett syndrome: when DNA methylation goes unrecognized. *Nat. Rev. Genet.* **7**, 415–426 (2006)
9. Brero, A., Easwaran, H.P., Nowak, D., Grunewald, I., Cremer, T., Leonhardt, H., Cardoso, M.: Methyl Cpg-binding proteins induce large-scale chromatin reorganization during terminal differentiation. *J. Cell Biol.* **169**, 733–743 (2005)
10. Calfa, G., Percy, A.K., Pozzo-Miller, L.: Experimental models of Rett syndrome based on Mecp2 dysfunction. *Exp. Biol. Med. (Maywood)* **236**, 3–19 (2011)
11. Chadwick, L.H., Wade, P.A.: Mecp2 in Rett syndrome: transcriptional repressor or chromatin architectural protein? *Curr. Opin. Genet. Dev.* **17**, 121–125 (2007)
12. Chahrouh, M., Jung, S.Y., Shaw, C., Zhou, X., Wong, S.T., Qin, J., Zoghbi, H.Y.: Mecp2, a key contributor to neurological disease, activates and represses transcription. *Science* **320**, 1224–1229 (2008)
13. Chen, R.Z., Akbarian, S., Tudor, M., Jaenisch, R.: Deficiency of methyl-Cpg binding protein-2 in Cns neurons results in a Rett-like phenotype in mice. *Nat. Genet.* **27**, 327–331 (2001)
14. Chen, W.G., Chang, Q., Lin, Y., Meissner, A., West, A.E., Griffith, E.C., Jaenisch, R., Greenberg, M.E.: Derepression of Bdnf transcription involves calcium-dependent phosphorylation of Mecp2. *Science* **302**, 885–889 (2003)
15. Ebert, D.H., Gabel, H.W., Robinson, N.D., Kastan, N.R., Hu, L.S., Cohen, S., Navarro, A.J., Lyst, M.J., Ekiert, R., Bird, A.P., Greenberg, M.E.: Activity-dependent phosphorylation of Mecp2 threonine 308 regulates interaction with Ncor. *Nature* **499**, 341–345 (2013)
16. Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., Greenberg, M.E.: Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522**, 89–93 (2015)

17. Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., Zhu, H., Chang, Q., Gao, Y., Ming, G.L., Song, H.: Distribution, recognition and regulation of non-Cpg methylation in the adult mammalian brain. *Nat. Neurosci.* **17**, 215–222 (2014)
18. Guy, J., Hendrich, B., Holmes, M., Martin, J.E., Bird, A.: A mouse *Mecp2*-null mutation causes neurological symptoms that mimic Rett syndrome. *Nat. Genet.* **27**, 322–326 (2001)
19. Guy, J., Cheval, H., Selfridge, J., Bird, A.: The role of *Mecp2* in the brain. *Ann. Rev. Cell Dev. Biol.* **27**, 631–652 (2011)
20. Horike, S., Cai, S., Miyano, M., Cheng, J.F., Kohwi-Shigematsu, T.: Loss of silent-chromatin looping and impaired imprinting of *Dlx5* in Rett syndrome. *Nat. Genet.* **37**, 31–40 (2005)
21. Kinde, B., Gabel, H.W., Gilbert, C.S., Griffith, E.C., Greenberg, M.E.: Reading the unique DNA methylation landscape of the brain: non-Cpg methylation, hydroxymethylation, and *Mecp2*. *Proc. Natl. Acad. Sci. USA* **112**, 6800–6806 (2015)
22. Klose, R., Bird, A.: Molecular biology. *Mecp2* repression goes nonglobal. *Science* **302**, 793–795 (2003)
23. Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., Bird, A.: Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**, 905–914 (1992)
24. Li, W., Pozzo-Miller, L.: *Bdnf* deregulation in Rett syndrome. *Neuropharmacology* **76**, 737–746 (2013)
25. Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J.C., Rao, A., Esteller, M., He, C., Haghghi, F.G., Sejnowski, T.J., Behrens, M.M., Ecker, J.R.: Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013)
26. Lyst, M.J., Ekiert, R., Ebert, D.H., Merusi, C., Nowak, J., Selfridge, J., Guy, J., Kastan, N.R., Robinson, N.D., De Lima Alves, F., Rappsilber, J., Greenberg, M.E., Bird, A.: Rett syndrome mutations abolish the interaction of *Mecp2* with the *Ncor/Smrt* co-repressor. *Nat. Neurosci.* **16**, 898–902 (2013)
27. Martinowich, K., Hattori, D., Wu, H., Fouse, S., He, F., Hu, Y., Fan, G., Sun, Y.E.: DNA methylation-related chromatin remodeling in activity-dependent *Bdnf* gene regulation. *Science* **302**, 890–893 (2003)
28. Maunakea, A.K., Chepelev, I., Cui, K., Zhao, K.: Intragenic DNA methylation modulates alternative splicing by recruiting *Mecp2* to promote exon recognition. *Cell Res.* **23**, 1256–1269 (2013)
29. Maxwell, S.S., Pelka, G.J., Tam, P.P., El-Osta, A.: Chromatin context and ncRNA highlight targets of *Mecp2* in brain. *RNA Biol.* **10**, 1741–1757 (2013)
30. Meehan, R.R., Lewis, J.D., Bird, A.P.: Characterization of *Mecp2*, a vertebrate DNA binding protein with affinity for methylated DNA. *Nucleic Acids Res.* **20**, 5085–5092 (1992)
31. Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S., Heintz, N.: *Mecp2* binds to 5hmc enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**, 1417–1430 (2012)
32. Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., Gage, F.H.: L1 Retrotransposition in neurons is modulated by *Mecp2*. *Nature* **468**, 443–446 (2010)
33. Nan, X., Cross, S., Bird, A.: Gene Silencing by methyl-Cpg-binding proteins. *Novartis Found Symp.* **214**, 6–16 (1998); Discussion 16–21, 46–50
34. Nguyen, M.V., Du, F., Felice, C.A., Shan, X., Nigam, A., Mandel, G., Robinson, J. K., Ballas, N.: *Mecp2* is critical for maintaining mature neuronal networks and global brain anatomy during late stages of postnatal brain development and in the mature adult brain. *J. Neurosci.* **32**, 10021–34 (2012)
35. Petazzi, P., Sandoval, J., Szczesna, K., Jorge, O.C., Roa, L., Sayols, S., Gomez, A., Huertas, D., Esteller, M. Dysregulation of the long non-coding RNA transcriptome in a Rett syndrome mouse model. *RNA Biol.* **10**, 1197–1203 (2013)
36. Rett, A.: On A unusual brain atrophy syndrome in hyperammonemia in childhood. *Wien Med Wochenschr* **116**, 723–726 (1966)

37. Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Bruggmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., Chang, H.Y.: Functional demarcation of active and silent chromatin domains in human Hox loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007)
38. Shahbazian, M., Young, J., Yuva-Paylor, L., Spencer, C., Antalffy, B., Noebels, J., Armstrong, D., Paylor, R., Zoghbi, H.: Mice with truncated Mecp2 recapitulate many Rett syndrome features and display hyperacetylation of histone H3. *Neuron* **35**, 243–254 (2002)
39. Skene, P.J., Illingworth, R.S., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Andrews, R., Bird, A.P.: Neuronal Mecp2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol. Cell* **37**, 457–468 (2010)
40. Sugino, K., Hempel, C.M., Okaty, B.W., Arnson, H.A., Kato, S., Dani, V.S., Nelson, S.B.: Cell-type-specific repression by methyl-Cpg-binding protein 2 is biased toward long genes. *J. Neurosci.* **34**, 12877–12883 (2014)
41. Tate, P., Skarnes, W., Bird, A.: The methyl-Cpg binding protein Mecp2 is essential for embryonic development in the mouse. *Nat. Genet.* **12**, 205–208 (1996)
42. Wan, M., Zhao, K., Lee, S.S., Francke, U.: Mecp2 truncating mutations cause histone H4 hyperacetylation in Rett syndrome. *Hum. Mol. Genet.* **10**, 1085–1092 (2001)
43. Wang, K.C., Chang, H.Y.: Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **43**, 904–914 (2011)
44. Zhou, H.L., Luo, G., Wise, J.A., Lou, H.: Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res.* **42**, 701–713 (2014)



# DNA Barcode Classification Using General Regression Neural Network with Different Distance Models

Massimo La Rosa, Antonino Fiannaca, Riccardo Rizzo, and Alfonso Urso

**Abstract** The “cythosome c oxidase subunits 1” (COI) gene is used for identification of species, and it is one of the so-called *DNA barcode* genes. Identification of species, even using DNA barcoding can be difficult if the biological examples are degraded. Spectral representation of sequences and the General Regression Neural Network (GRNN) can give some interesting results in these difficult cases. The GRNN is based on the distance between the memorized examples of sequence and the input unknown sequence, both represented using a vector space spectral representation. In this paper we will analyse the effectiveness of different distance models in the GRNN implementation and will compare the obtained results in the classification of full length sequences and degraded samples.

**Keywords** Barcode classification • Alignment-free • GRNN

## 1 Introduction

The so-called *DNA barcode sequence* is a small segment (~650 bp) of DNA, usually from “cythosome c oxidase subunits 1” mitochondrial gene (COI) [8, 13]. The sequence is a good marker for DNA and is widely used for identification and taxonomic rank assignment of many species [5].

DNA barcoding is difficult if the biological samples under analysis are degraded: in this case only fragments of the barcode sequence is available. A suitable solution for this problem is studied in [14]: in this work the barcode sequence is analysed in order to find small subsequences that are still useful for identification of the sample specie.

We started from a different point of view: we addressed the identification and rank assignment of degraded barcode sequences, usually sequence fragments of about 200 bp, building a robust classifier based on the spectral representation and a modified version of the General Regression Neural Network (GRNN).

---

M. La Rosa (✉) • A. Fiannaca • R. Rizzo • A. Urso  
ICAR-CNR, viale delle Scienze Ed. 11, 90128 Palermo, Italy  
e-mail: [larosa@pa.icar.cnr.it](mailto:larosa@pa.icar.cnr.it); [fiannaca@pa.icar.cnr.it](mailto:fiannaca@pa.icar.cnr.it); [ricrizzo@pa.icar.cnr.it](mailto:ricrizzo@pa.icar.cnr.it); [urso@pa.icar.cnr.it](mailto:urso@pa.icar.cnr.it)

Using spectral representation the DNA barcode sequence is represented using the frequency of very short strings of length  $k = 3, 4, \dots$ , called  $k$ -mers. This sequence representation is often addressed as  $k$ -mers decomposition or, more generally, as alignment-free sequence decomposition. In this representation the order of  $k$ -mers in the sequence is discarded and only their count is considered; if a sequence fragment has a  $k$ -mers frequency distribution similar to the one of the whole barcode sequence then the two will have a similar representation.

The set of the frequencies of the  $k$ -mers in a sequence constitutes the representing vector for the sequence in a vector space. The dimension of the representation space is  $4^k$  and the distance among these representing vector can be calculated using Euclidean norm in  $\mathfrak{R}^{4^k}$ .

The GRNN is a neural network originally developed for regression and adapted to classification of DNA sequences in [17]. This modification made the network a prototype-based classification tool that classifies a new input looking at the distance from the memorized training samples. It is clear that different distance models, like *Euclidean*, *manhattan* and so on, can change the performances of the network, as we found in [17].

In this paper we want to go further in this study and analyse and compare the performances of other distance models on the GRNN, considering classification results of both full length sequences and degraded samples.

With regards to barcode classification, very interesting results have been obtained in the works presented in [10, 15, 20]. In particular both the algorithms described in [15, 20] propose alignment-based methods in order to classify barcode specimen. In [20], after the training sequences are aligned, a set of logic rules are extracted in the form “if pos35 is G and pos300 is A then the sequence is classified as ...”, where  $\text{pos}X$  represents a sequence locus. In [15], first a phylogenetic tree of input sequences is computed; then at each branching node, a set of “characteristic attributes” (CA) is identified for the corresponding leaf nodes. Considering a branch node, CAs are single nucleotide position or multiple nucleotide positions that are shared only by one of the branch descending from that node. Another alignment-free approach more similar to our proposed method is the one presented in [10]. There authors introduce the spectral representation for the barcode sequences and they use two machine learning algorithms,  $k$ -Nearest Neighbour (kNN) [2] and Support Vector Machine (SVM) [18], to train different classifiers. In this paper we are going to compare our GRNN approach with the classifiers proposed in [10] because they represent alignment-free approaches, differently from [15] and [20], that also implement the spectral representation. The comparison between our GRNN method and the SVM classifier has been already done in [17], where we demonstrated our method outperforms SVM when dealing with sequence fragments. Therefore in this paper we compare our GRNN method against the  $k$ -NN classifier.

## 2 Methods

Prototype-based classification tools are based on sequence distance; there are many algorithms to evaluate sequence distance besides the evolutionary distance, for example the compression distance used in [11, 12]. The vector space representation is obtained by considering the frequency of all possible 5-letters substring in the DNA barcode sequence (k-mers), these k-mers are obtained by using a sliding window on the sequence. A deeper discussion on this representation can be found in [3, 10]. In the following sections the GRNN modified algorithm is explained and the different distance measures applied are described; moreover the barcode dataset used is introduced.

### 2.1 The General Regression Neural Network

Artificial neural networks (ANN) are a set of algorithms used to approximate functions or cluster large sets of input values. A neural network usually have a very large set of parameters (the network weights) adapted using a set of training examples and a specific learning algorithm (the training algorithm). The training phase is aimed at reducing the error of the network on a specific task, classification or regression, by changing the weight values.

Among the neural networks the GRNN [19] is a network created for regression i.e. the approximation of the values of a dependent continuous variable  $y$  given a set of samples  $(\mathbf{x}_i, y_i) \ i = 1, 2, \dots N$ .

In the following we will discuss the one dimensional output case, the extension to an output vector  $\mathbf{y}$  being straightforward (see [19] for details).

The GRNN do not have a training phase, it is based on the memorization of all the training examples in the hidden layer: one neural unit for each training samples (see Fig. 1). When a new pattern  $\mathbf{x}'$  is presented to the network input the output  $y$  is calculated using the following equation:

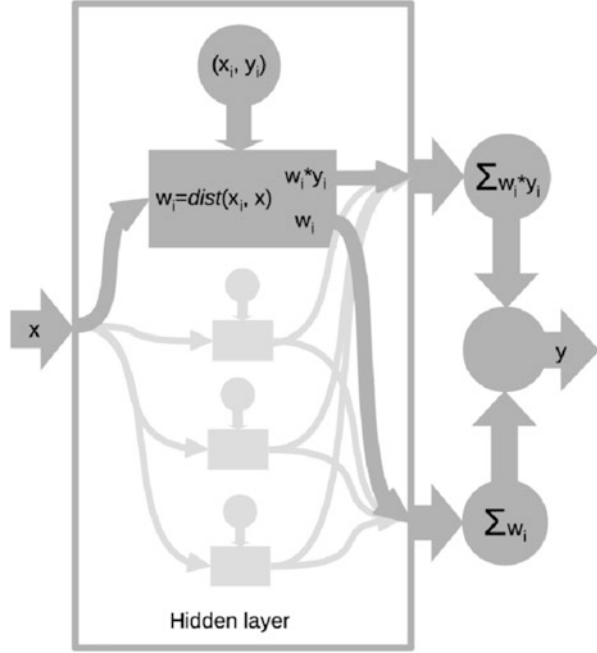
$$y' = \frac{\sum w_i * y_i}{\sum w_i} \quad (1)$$

where the weight  $w_i$  are obtained from each hidden unit as

$$w_i = \exp \left\{ -\frac{d(\mathbf{x}', \mathbf{x}_i)}{2\sigma^2} \right\} \quad (2)$$

The  $\sigma$  value, called spread factor, is the only parameter of the GRNN network. The weight  $w_i$  is considered by some literature the excitation level of the neural unit  $i$  corresponding to the input  $\mathbf{x}'$ .

**Fig. 1** The representation of the GRNN neural network. The hidden layer contains all the training patterns and calculates the  $w_i$  considering the distance from the input pattern  $x$ . These weights are used to calculate the output. On the *right* there is the output layer composed by three units: the *upper* one collects all the terms  $w_i * y_i$  and the *lower* one collects the terms  $w_i$ : these terms are combined in the third unit that generates the output



There are some studies on the optimal value of  $\sigma$  that can be a single value for the whole network or a specific value for each hidden unit. In [7] it is suggested a formula that depends on the maximum distance and number of patterns in the training set.

The GRNN can be used in classification problems: considering a set of classification examples  $(\mathbf{x}_i, c_h)$  where  $\mathbf{x}_i$  (with  $i = 1, 2, \dots, N$ ) is the input pattern and  $c_h$  (with  $h = 1, 2, \dots, H$ ;  $H$  is the number of available classes) is the class assigned to the pattern  $\mathbf{x}_i$  it is possible to build a set of training examples for a GRNN network as  $(\mathbf{x}_i, \mathbf{y}_i)$  where  $\mathbf{y}_i = [y_{i,1}, y_{i,2}, \dots, y_{i,H}]$  is given by

$$y_{i,j} = \begin{cases} 0 & \text{if } j \neq h \\ 1 & \text{if } j = h \end{cases} \quad (3)$$

where  $c_h$  is the class of the pattern  $x_i$ .

The set of couples  $(\mathbf{x}_i, \mathbf{y}_i)$  can be used as a training set for the GRNN and the class of the new input  $\mathbf{x}'$  can be calculated as

$$c_h(\mathbf{x}') = \arg \max_j \{y'_j | j = 1, 2, \dots, H\} \quad (4)$$

In order to implement our classification tool for DNA sequences, we obtained the vector representation of the DNA sequences using a  $k$ -mer decomposition, as shown in [10], in which sequences are coded as fixed size vectors whose components are

the number of occurrences of short DNA snippets of  $k$  fixed-length, called  $k$ -mers. Considering  $k = 5$ , as proposed in [10], we have vectors of dimension  $4^5 = 1024$  to represent genomic sequences.

The GRNN is used with different distance models, in particular some of the  $L_p$  norms, the correlation norm and the cosine norm.

## 2.2 The Distance Models

In this section the  $L_p$  norms used are introduced, together with the cosine and correlation distances.

### 2.2.1 $L_p$ Norms

The norm is a function that assigns a strict positive number to a vector in a vector space  $f : (\mathbf{x}) \rightarrow \Re$  that satisfies the following properties:

$$f(\alpha \mathbf{x}) = |\alpha| f(\mathbf{x}) \quad (5)$$

$$f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y}) \quad (6)$$

$$\text{if } f(\mathbf{x}) = 0 \text{ then } \mathbf{x} \text{ is the vector zero} \quad (7)$$

the  $L_p$  family norms, or  $p$ -norms, defined as:

$$\|\mathbf{x}\|_p = \left( \sum_i |x_i|^p \right)^{\frac{1}{p}}. \quad (8)$$

The most common norm is the Euclidean norm with  $p = 2$ , but are also used the  $p = 1$  norm namely City-block or Manhattan, and the Chebyshev norm, or  $L_\infty$ . Although should be  $p \geq 1$  there are also fractional norms with  $p < 1$ , that are interesting in the case of high dimensional spaces.

In case of high-dimensionality data, such as the 1024 sized vectors representing DNA sequences, the Euclidean norm used to define the distance tend to *concentrate* [4]. That means all pairwise distances between high-dimensional objects appear to be very similar. Authors in [4] also state that the concentration phenomenon is intrinsic to the norm. In order to overcome this phenomenon, fractional norms can be used in place of Euclidean norm [1, 9]; whereas with  $0 < p < 1$   $L_p$  norms are called fractional norms, which induce fractional distances. Moreover, fractional norms are able to deal with non-Gaussian noise [4]. In this work we adopted fractional norms, considering different values of  $p$ , in order to compute Eq. (2) and to limit the effects of the curse of dimensionality.

If  $p = 1$  in Eq. (8) the norm is called the *Manhattan norm*, or *taxicab norm*, and is defined as

$$L_1 = \sum_i |x'_i - x_i|. \quad (9)$$

both the names are related to the distance a taxi as to drive in a city with a rectangular grid.

The Chebyshev distance is obtained from the formula:

$$d(\mathbf{x}', \mathbf{x}_i) = \max_i (|x'_i - x_i|). \quad (10)$$

this is usually considered as  $L_\infty$  norm.

### 2.2.2 Cosine and Correlation Distance

Cosine and correlation distance are both based on scalar product  $\mathbf{x}' \cdot \mathbf{x}_i$ , instead of the difference  $\mathbf{x}' - \mathbf{x}_i$ . The cosine distance is defined by the following equation:

$$d(\mathbf{x}', \mathbf{x}_i) = 1 - \frac{\mathbf{x}' \cdot \mathbf{x}_i}{\|\mathbf{x}'\| \|\mathbf{x}_i\|} \quad (11)$$

where the  $\|\cdot\|$  is the Euclidean norm. The correlation distance is defined by:

$$d(\mathbf{x}', \mathbf{x}_i) = 1 - \frac{(\mathbf{x}' - \bar{\mathbf{x}}') \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_i)}{\|\mathbf{x}' - \bar{\mathbf{x}}'\| \|\mathbf{x}_i - \bar{\mathbf{x}}_i\|} \quad (12)$$

where  $\bar{\mathbf{x}}'$  is the mean of the input vectors  $\mathbf{x}'$  and  $\bar{\mathbf{x}}_i$  is the mean of the training samples.

## 2.3 Barcode Dataset

We downloaded barcode sequences from the Barcode of Life Database (BOLD) [16]. In our study, we considered 10 barcode datasets belonging to different BOLD projects and living organisms. These datasets have been selected according to some criteria: we chose only *barcode compliant* dataset, i.e certified by BOLD as true barcode sequences, with sequence length not shorter than 500 bp and not longer than 800 bp. These datasets differ each other on the basis of the number of species and specimen, the sequence length and the sequence quality (in terms of undefined nucleotides). Following these criteria, we collected 2210 sequences. The dataset composition, in terms of number of different taxa and number of specimen for each taxa, is summarized in Table 1, where it is possible to note how the dataset is unbalanced.

**Table 1** Barcode dataset composition at each taxonomic level

| Sequence distribution for each taxa |           |         |           |           |         |           |           |         |
|-------------------------------------|-----------|---------|-----------|-----------|---------|-----------|-----------|---------|
| Phylum                              |           |         | Class     |           |         | Order     |           |         |
| # Classes                           | # Seqs    | % Seqs  | # Classes | # Seqs    | % Seqs  | # Classes | # Seqs    | % Seqs  |
| 1                                   | 1361      | 61.9 %  | 1         | 1361      | 61.9 %  | 1         | 1049      | 47.46 % |
| 2                                   | [219,386] | 27.4 %  | 2         | [219,286] | 22.85 % | 3         | [209,286] | 32.30 % |
| 2                                   | [111,133] | 11.0 %  | 3         | [100,133] | 15.56 % | 4         | [100,133] | 20.22 % |
| Family                              |           |         | Genus     |           |         | Species   |           |         |
| # Classes                           | # Seqs    | % Seqs  | # Classes | # Seqs    | % Seqs  | # Classes | # Seqs    | % Seqs  |
| 1                                   | 885       | 40.04 % | 1         | 386       | 17.46 % | 1         | 279       | 12.64 % |
| 3                                   | [209,274] | 31.76 % | 3         | [209,290] | 32.48 % | 4         | [105,140] | 22.30 % |
| 4                                   | [103,164] | 23.12 % | 6         | [103,164] | 35.15 % | 30        | [14,92]   | 49.50 % |
| 7                                   | [4,46]    | 5.06 %  | 15        | [4,71]    | 14.91 % | 35        | [1,11]    | 15.56 % |

Numbers between square brackets represent range of values

### 3 Results and Discussion

In this section, we describe the parameter setup for the GRNN algorithm and the adopted training/testing procedure. Then we report classification results in terms of accuracy, precision and recall scores, and finally we discuss those results.

#### 3.1 Experimental Setup

The only parameter of the GRNN algorithm is the spread factor  $\sigma$  (Eq. 2). In our experiments, we tuned the  $\sigma$  value by means of a ten fold cross validation procedure, considering as training set the dataset composed of the full length sequences. This procedure has been carried out implementing each distance model (see Sect. 2.2), and for values of  $\sigma$  ranging from 0.5 to 0.8, with a step of 0.1. For each value of  $\sigma$  we noticed that the behaviour of the GRNN was substantially the same regardless the distance model, and the best results, in terms of error rate, were obtained with  $\sigma = 0.6$ . As for the fractional distances, Eq. (8) with  $p < 1$ , we considered three values for  $p$ : 0.3, 0.5, 0.7. All the experiments have been done using Python scripts on a Windows 7 machine equipped with i7 Intel CPU at 2.8 GHz with 8 GB of RAM. Computational times of the GRNN algorithm are about 1 min for a single experiment.

The classification performances of the GRNN algorithm have been tested considering full length barcode sequences and sequence fragments of 200 consecutive bp randomly extracted from the original sequences. We want to assess the GRNN predictive power and its robustness with regards to the sequence sizes. In fact, in the study of environmental species, for example, usually only small portions of the barcode sequences are available.

For each distance model, the training and testing procedures have been done in two ways. In the first case, we adopted a ten fold cross validation method: in each fold, we trained the GRNN with the 90 % of the full-length sequences and we used as test set the remaining 10 % of both the full-length sequences and their corresponding sequence fragments of 200 bp. In the second case, we trained the GRNN with the whole dataset of the full-length sequences and then we tested it with all the sequence fragments. In the first scenario, we want to assess the classification performances of the GRNN considering full-length sequences and its generalization degree when used to classify sequence fragments whose corresponding original sequence does not belong to the training set. In the second scenario, we supposed the GRNN is used to recognize small random fragments, by “knowing” all the original full-length sequences. Comparison with the k-NN classifier has been carried out following the same training and testing procedure. We used the  $k$ -NN implementation provided by the Weka Experimenter Platform [6], considering  $k = 1$  and  $k = 3$ , as done similarly in [10].

### 3.2 Classification Results

Classification scores have been evaluated by means of the accuracy, precision and recall performance measures.

These scores are summarized in Tables 2, 3, and 4, respectively. Each table is composed of three parts, according to the adopted training/testing procedure. “Full-length” means the classification results are obtained through a ten fold cross validation scheme considering full length sequences both for training and testing; the scores are averaged over the ten folds. “Full vs. 200-bp” means the classification results are obtained through a ten fold cross validation scheme considering full-length sequences for training and 200 bp fragments for testing; once again the scores are averaged over the ten folds. “200-bp” means the classification results are obtained training with the whole dataset of full-length sequences and tested with all the sequence fragments. In each table, in the first column there is the distance model used to train the GRNN, and in the second row there is the taxonomic level, from Phylum to Species. The last two rows of each table part show the results obtained from the k-NN classifiers.

### 3.3 Discussion

From the classification results shown in Tables 2, 3, and 4, it is evident that the GRNN and the k-NN algorithms are able to correctly classify full-length barcode sequences, with scores around 100 % at each taxonomic level. The GRNN reaches those scores with all the distance models except for the correlation and the cosine distances.



**Table 2** Accuracy scores at each taxonomic level the GRNN algorithm, considering each distance model, and the k-NN classifier

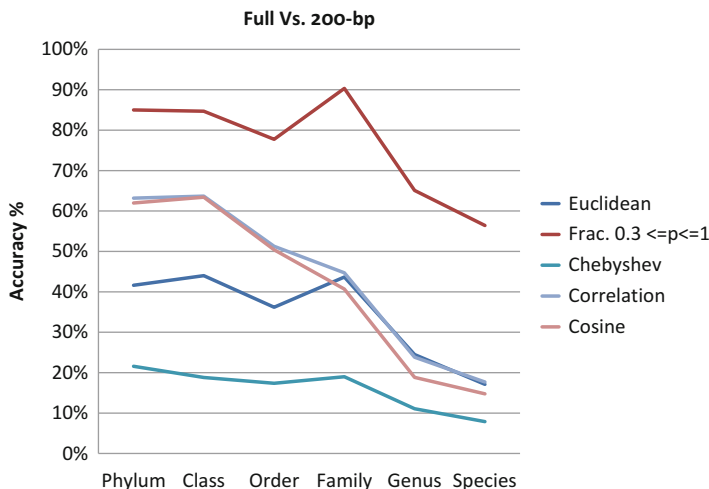
| PRECISION              |        |        |        |        |        |         |
|------------------------|--------|--------|--------|--------|--------|---------|
| Algorithm              | Phylum | Class  | Order  | Family | Genus  | Species |
| <b>FULL-LENGTH</b>     |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 100.0% | 100.0% | 100.0% | 100.0% | 99.9%  | 96.1%   |
| Frac. p=0.3            | 100.0% | 100.0% | 100.0% | 100.0% | 99.6%  | 94.7%   |
| Frac. p=0.5            | 100.0% | 100.0% | 100.0% | 100.0% | 99.8%  | 96.6%   |
| Frac. p=0.7            | 100.0% | 100.0% | 100.0% | 100.0% | 99.1%  | 96.1%   |
| Chebyshev              | 100.0% | 100.0% | 100.0% | 100.0% | 98.6%  | 91.9%   |
| City Block             | 100.0% | 100.0% | 100.0% | 100.0% | 99.8%  | 97.0%   |
| Correlation            | 12.5%  | 10.6%  | 18.7%  | 12.9%  | 6.6%   | 2.3%    |
| Cosine                 | 12.4%  | 10.6%  | 6.3%   | 3.3%   | 0.8%   | 0.3%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 99.3%   |
| K=3                    | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 98.9%   |
| <b>FULL Vs. 200-bp</b> |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 62.7%  | 65.2%  | 54.1%  | 53.9%  | 24.3%  | 13.8%   |
| Frac. p=0.3            | 77.8%  | 80.2%  | 76.6%  | 87.0%  | 55.2%  | 46.7%   |
| Frac. p=0.5            | 77.2%  | 76.6%  | 77.8%  | 83.9%  | 60.6%  | 46.1%   |
| Frac. p=0.7            | 77.3%  | 77.2%  | 76.4%  | 80.8%  | 58.6%  | 49.9%   |
| Chebyshev              | 25.9%  | 26.4%  | 13.2%  | 10.9%  | 5.4%   | 2.0%    |
| City Block             | 79.0%  | 76.3%  | 75.8%  | 85.4%  | 57.0%  | 41.0%   |
| Correlation            | 22.2%  | 10.6%  | 17.7%  | 12.7%  | 9.2%   | 1.9%    |
| Cosine                 | 12.4%  | 10.6%  | 6.3%   | 4.9%   | 7.4%   | 1.8%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 83.9%  | 83.9%  | 81.2%  | 80.6%  | 74.7%  | 61.7%   |
| K=3                    | 83.8%  | 83.8%  | 82.0%  | 80.3%  | 75.3%  | 62.8%   |
| <b>200-bp</b>          |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 83.4%  | 86.0%  | 68.1%  | 67.2%  | 44.7%  | 29.8%   |
| Frac. p=0.3            | 98.3%  | 98.6%  | 97.9%  | 91.1%  | 80.6%  | 78.4%   |
| Frac. p=0.5            | 99.3%  | 99.4%  | 93.7%  | 90.8%  | 81.6%  | 78.8%   |
| Frac. p=0.7            | 100.0% | 100.0% | 98.5%  | 89.9%  | 81.9%  | 79.8%   |
| Chebyshev              | 45.6%  | 44.0%  | 23.4%  | 14.1%  | 8.8%   | 3.8%    |
| City Block             | 100.0% | 100.0% | 98.2%  | 89.1%  | 75.8%  | 70.6%   |
| Correlation            | 32.6%  | 10.5%  | 18.7%  | 9.5%   | 10.0%  | 1.5%    |
| Cosine                 | 12.5%  | 10.5%  | 6.0%   | 9.4%   | 8.2%   | 1.6%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 87.7%  | 87.7%  | 85.1%  | 85.0%  | 78.6%  | 71.0%   |
| K=3                    | 86.5%  | 86.5%  | 84.3%  | 83.4%  | 78.6%  | 73.9%   |

**Table 3** Precision scores at each taxonomic level the GRNN algorithm, considering each distance model, and the k-NN classifier

| PRECISION              |        |        |        |        |        |         |
|------------------------|--------|--------|--------|--------|--------|---------|
| Algorithm              | Phylum | Class  | Order  | Family | Genus  | Species |
| <b>FULL-LENGTH</b>     |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 100.0% | 100.0% | 100.0% | 100.0% | 99.9%  | 96.1%   |
| Frac. p=0.3            | 100.0% | 100.0% | 100.0% | 100.0% | 99.6%  | 94.7%   |
| Frac. p=0.5            | 100.0% | 100.0% | 100.0% | 100.0% | 99.8%  | 96.6%   |
| Frac. p=0.7            | 100.0% | 100.0% | 100.0% | 100.0% | 99.1%  | 96.1%   |
| Chebyshev              | 100.0% | 100.0% | 100.0% | 100.0% | 98.6%  | 91.9%   |
| City Block             | 100.0% | 100.0% | 100.0% | 100.0% | 99.8%  | 97.0%   |
| Correlation            | 12.5%  | 10.6%  | 18.7%  | 12.9%  | 6.6%   | 2.3%    |
| Cosine                 | 12.4%  | 10.6%  | 6.3%   | 3.3%   | 0.8%   | 0.3%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 99.3%   |
| K=3                    | 83.9%  | 83.9%  | 81.2%  | 80.6%  | 74.7%  | 61.7%   |
| <b>FULL Vs. 200-bp</b> |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 62.7%  | 65.2%  | 54.1%  | 53.9%  | 24.3%  | 13.8%   |
| Frac. p=0.3            | 77.8%  | 80.2%  | 76.6%  | 87.0%  | 55.2%  | 46.7%   |
| Frac. p=0.5            | 77.2%  | 76.6%  | 77.8%  | 83.9%  | 60.6%  | 46.1%   |
| Frac. p=0.7            | 77.3%  | 77.2%  | 76.4%  | 80.8%  | 58.6%  | 49.9%   |
| Chebyshev              | 25.9%  | 26.4%  | 13.2%  | 10.9%  | 5.4%   | 2.0%    |
| City Block             | 79.0%  | 76.3%  | 75.8%  | 85.4%  | 57.0%  | 41.0%   |
| Correlation            | 22.2%  | 10.6%  | 17.7%  | 12.7%  | 9.2%   | 1.9%    |
| Cosine                 | 12.4%  | 10.6%  | 6.3%   | 4.9%   | 7.4%   | 1.8%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 83.9%  | 83.9%  | 81.2%  | 80.6%  | 74.7%  | 61.7%   |
| K=3                    | 83.8%  | 83.8%  | 82.0%  | 80.3%  | 75.3%  | 62.8%   |
| <b>200-bp</b>          |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 83.4%  | 86.0%  | 68.1%  | 67.2%  | 44.7%  | 29.8%   |
| Frac. p=0.3            | 98.3%  | 98.6%  | 97.9%  | 91.1%  | 80.6%  | 78.4%   |
| Frac. p=0.5            | 99.3%  | 99.4%  | 93.7%  | 90.8%  | 81.6%  | 78.8%   |
| Frac. p=0.7            | 100.0% | 100.0% | 98.5%  | 89.9%  | 81.9%  | 79.8%   |
| Chebyshev              | 45.6%  | 44.0%  | 23.4%  | 14.1%  | 8.8%   | 3.8%    |
| City Block             | 100.0% | 100.0% | 98.2%  | 89.1%  | 75.8%  | 70.6%   |
| Correlation            | 32.6%  | 10.5%  | 18.7%  | 9.5%   | 10.0%  | 1.5%    |
| Cosine                 | 12.5%  | 10.5%  | 6.0%   | 9.4%   | 8.2%   | 1.6%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 87.7%  | 87.7%  | 85.1%  | 85.0%  | 78.6%  | 71.0%   |
| K=3                    | 86.5%  | 86.5%  | 84.3%  | 83.4%  | 78.6%  | 73.9%   |

**Table 4** Recall scores at each taxonomic level the GRNN algorithm, considering each distance model, and the k-NN classifier

| RECALL                 |        |        |        |        |        |         |
|------------------------|--------|--------|--------|--------|--------|---------|
| Algorithm              | Phylum | Class  | Order  | Family | Genus  | Species |
| <b>FULL-LENGTH</b>     |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 96.5%   |
| Frac. p=0.3            | 100.0% | 100.0% | 100.0% | 100.0% | 99.7%  | 94.6%   |
| Frac. p=0.5            | 100.0% | 100.0% | 100.0% | 100.0% | 99.4%  | 96.8%   |
| Frac. p=0.7            | 100.0% | 100.0% | 100.0% | 100.0% | 99.3%  | 96.5%   |
| Chebyshev              | 100.0% | 100.0% | 100.0% | 100.0% | 99.0%  | 92.8%   |
| City Block             | 100.0% | 100.0% | 100.0% | 100.0% | 99.7%  | 97.1%   |
| Correlation            | 20.0%  | 16.7%  | 13.8%  | 16.5%  | 8.5%   | 4.3%    |
| Cosine                 | 20.0%  | 16.7%  | 12.5%  | 8.0%   | 4.8%   | 2.1%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 99.3%   |
| K=3                    | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 99.0%   |
| <b>FULL Vs. 200-bp</b> |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 52.6%  | 59.2%  | 49.9%  | 50.4%  | 26.7%  | 15.4%   |
| Frac. p=0.3            | 67.8%  | 67.1%  | 65.3%  | 77.8%  | 48.4%  | 44.7%   |
| Frac. p=0.5            | 72.7%  | 72.1%  | 69.8%  | 78.3%  | 56.7%  | 45.1%   |
| Frac. p=0.7            | 75.6%  | 80.9%  | 72.6%  | 78.5%  | 56.1%  | 50.3%   |
| Chebyshev              | 26.8%  | 27.5%  | 21.8%  | 15.6%  | 10.3%  | 4.8%    |
| City Block             | 79.7%  | 74.3%  | 69.1%  | 81.4%  | 56.2%  | 42.5%   |
| Correlation            | 20.9%  | 16.7%  | 14.7%  | 12.5%  | 9.0%   | 3.5%    |
| Cosine                 | 20.0%  | 16.7%  | 12.5%  | 8.1%   | 6.2%   | 2.6%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 84.5%  | 84.5%  | 78.0%  | 77.3%  | 67.5%  | 57.3%   |
| K=3                    | 82.4%  | 82.4%  | 76.9%  | 75.6%  | 65.8%  | 55.4%   |
| <b>200-bp</b>          |        |        |        |        |        |         |
| <b>GRNN</b>            |        |        |        |        |        |         |
| Euclidean              | 69.0%  | 73.1%  | 60.2%  | 57.2%  | 37.3%  | 23.6%   |
| Frac. p=0.3            | 80.5%  | 83.7%  | 81.5%  | 81.0%  | 69.1%  | 62.9%   |
| Frac. p=0.5            | 92.4%  | 93.7%  | 88.3%  | 82.1%  | 72.2%  | 67.2%   |
| Frac. p=0.7            | 99.5%  | 99.6%  | 90.9%  | 84.0%  | 74.6%  | 70.0%   |
| Chebyshev              | 30.7%  | 32.2%  | 24.4%  | 14.2%  | 11.1%  | 4.3%    |
| City Block             | 100.0% | 100.0% | 89.1%  | 82.5%  | 74.1%  | 63.6%   |
| Correlation            | 20.7%  | 16.7%  | 14.6%  | 9.5%   | 8.8%   | 2.5%    |
| Cosine                 | 20.0%  | 16.7%  | 12.5%  | 6.7%   | 5.3%   | 1.8%    |
| <b>K-NN</b>            |        |        |        |        |        |         |
| k=1                    | 84.7%  | 84.7%  | 78.2%  | 77.6%  | 67.8%  | 57.7%   |
| K=3                    | 82.7%  | 82.7%  | 76.6%  | 75.5%  | 65.8%  | 55.7%   |



**Fig. 2** Accuracy scores at each taxonomic level for the “Full vs. 200-bp” training/testing scheme of the GRNN classifier with different distance models

Using those distances, the performances of the GRNN drop significantly, reaching about 62 % in terms of accuracy at phylum level, and only about 20 and 12 % in terms of recall and precision respectively. That means distances based on scalar product of the patterns are not suitable with the GRNN algorithm. The most interesting results are therefore the ones obtained during the classification evaluation of the sequence fragments. First of all, the performances decrease with respect to taxonomic level, as it is also evident in the chart of Fig. 2. As the taxonomic rank goes down, indeed, the number of categories to classify increases (see Table 1) and, as a consequence, it is more difficult to correctly classify the patterns. Considering the “Full vs. 200-bp” part, the only meaningful scores are provided by the GRNN implementing fractional and city block distances. In particular while the correlation and the cosine distances keep on giving low scores as in the case of full-length sequences, the Chebyshev and the Euclidean distance have a strong drop of performances, with scores about 40 % for Euclidean distance at Phylum level and about 20 % for Chebyshev distance at Phylum level. The same drop of performances also affects the k-NN classifiers, with very similar scores regardless the value of k. On the other hand, considering fractional and city block distances, the GRNN is still able to provide acceptable classification results for sequence fragments, with scores ranging from about 85 % at phylum level to about 57 % at Species level. These results further confirm that fractional norms contrast the effects of distance concentration. It is important to remember that in the case of “Full vs. 200-bp” the GRNN network classify the sequence fragments without “knowing” the corresponding full length sequences during the training phase. It is interesting to note (see Fig. 2) that at the family level there are the best scores: that because the distribution of specimen at family level is very unbalanced, with one family collecting about the 40 % of

available samples, as reported in Table 1. Finally, considering the “200-bp” part of Tables 2, 3, and 4, once again only the GRNN implementing the fractional and the city block distances are able to provide a proper classification for sequence fragments. In this last case, the performance scores are higher than the “Full vs. 200-bp” scenario, because in this situation we carried out a complete training procedure of the GRNN considering all full-length sequences. Of course, because the spectral representation of full-length and sequence fragments are different from each other, no sequence fragment used in the test set belong to the training set.

## 4 Conclusion

In this work, a modified version of the GRNN algorithm implementing different distance models for barcode sequence classification is presented. The GRNN classification performances have been assessed with regards to sequence sizes. Experimental trials have been carried out considering full-length sequences and sequence fragments that simulate a very common scenario in which only environmental samples are available. In the case of full-length sequences, 6 out of 8 distance models provided near perfect results, in terms of accuracy, precision and recall, with scores ranging between 100 % at Phylum level and 90 % at Species level. The same scores are reached using the k-NN classifier. Only correlation and cosine distance did not provide acceptable results. In the case of sequence fragments, fractional and city block distances only gave meaningful results: in the “Full vs. 200-bp” scenario, accuracy ranged from 85 % at Phylum level to 57 % at Species level; in the “200-bp” scenario, accuracy ranged from 95 to 100 % at Phylum level to 70–79 % at Species level. In both scenarios our GRNN approach outperformed the k-NN classifier. That means GRNN implementing fractional and city block distances was able to correctly predict the similarity between original full-length sequences and their corresponding sequence fragments. All the other distance model were affected by a strong classification performance drop.

## References

1. Aggarwal, C., Hinnenburg, A., Keim, D.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) *Database Theory ICDT 2001. Lecture Notes in Computer Science*, vol. 1973, pp. 420–434. Springer, Berlin/Heidelberg (2001)
2. Cover, T., Hart, P.: Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**(1), 21–27 (1967)
3. Fiannaca, A., La Rosa, M., Rizzo, R., Urso, A.: Analysis of DNA barcode sequences using neural gas and spectral representation. In: Iliadis, L., Papadopoulos, H., Jayne, C. (eds.) *Engineering Applications of Neural Networks. Communications in Computer and Information Science*, vol. 384, pp. 212–221. Springer, Berlin/Heidelberg (2013)
4. Francois, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. *IEEE Trans. Knowl. Data Eng.* **19**(7), 873–886 (2007)

5. Hajibabaei, M., Singer, G.A.C., Hebert, P.D.N., Hickey, D.A.: DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* **23**(4), 167–172 (2007)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. *ACM SIGKDD Explorations Newsletter* **11**(1), 10–18 (2009)
7. Haykin, S.: *Neural Networks: A Comprehensive Foundation*, 2nd edn. Prentice Hall, Upper Saddle River (1998)
8. Hebert, P.D.N., Ratnasingham, S., DeWaard, J.R.: Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Ser. B, Biol. Sci.* **270** **Suppl.**, S96–S99 (2003)
9. Hinnenburg, A., Aggarwal, C., Keim, D.: What is the nearest neighbor in high dimensional spaces? In: *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB '00*, pp. 506–515. Morgan Kaufmann, San Francisco (2000)
10. Kuksa, P., Pavlovic, V.: Efficient alignment-free DNA barcode analytics. *BMC Bioinf.* **10**(Suppl.14), S9 (2009)
11. La Rosa, M., Fiannaca, A., Rizzo, R., Urso, A.: A study of compression-based methods for the analysis of barcode sequences. In: Peterson, L.E., Masulli, F., Russo, G. (eds.) *Computational Intelligence Methods for Bioinformatics and Biostatistics. Lecture Notes in Computer Science*, vol. 7845, pp. 105–116. Springer, Berlin/Heidelberg (2013)
12. La Rosa, M., Fiannaca, A., Rizzo, R., Urso, A.: Alignment-free analysis of barcode sequences by means of compression-based methods. *BMC Bioinf.* **14**, S4 (2013)
13. Marshall, E.: Taxonomy. Will DNA bar codes breathe life into classification? *Science (New York, N.Y.)* **307**(5712), 1037 (2005)
14. Meusnier, I., Singer, G.A.C., Landry, J.F., Hickey, D.A., Hebert, P.D.N., Hajibabaei, M.: A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* **9**, 214 (2008)
15. Rach, J., Desalle, R., Sarkar, I.N., Schierwater, B., Hadrys, H.: Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc. Biol. Sci. R. Soc.* **275**(1632), 237–247 (2008)
16. Ratnasingham, S., Hebert, P.D.N.: Bold: the barcode of life data system (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* **7**(3), 355–364 (2007)
17. Rizzo, R., Fiannaca, A., La Rosa, M., Urso, A.: The general regression neural network to classify barcode and mini-barcode DNA. In: *Computational Intelligence Methods for Bioinformatics and Biostatistics. Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg (2015)
18. Scholkopf, B., Smola, A.: *Learning with Kernels*. MIT, Cambridge (2002)
19. Specht, D.F.: A general regression neural network. *IEEE Trans. Neural Netw.* **2**(6), 568–576 (1991)
20. Weitschek, E., Van Velzen, R., Felici, G., Bertolazzi, P.: BLOG 2.0: a software system for character-based species classification with DNA barcode sequences. What it does, how to use it. *Mol. Ecol. Resour.* **13**(6), 1043–1046 (2013)

# First Application of a Distance-Based Outlier Approach to Detect Highly Differentiated Genomic Regions Across Human Populations

Stefano Lodi, Fabrizio Angiulli, Stefano Basta, Donata Luiselli,  
Luca Pagani, and Claudio Sartori

**Abstract** Genomic scans for positive selection or population differentiation are often used in evolutionary genetics to shortlist genetic loci with potentially adaptive biological functions. However, the vast majority of such tests relies on empirical ranking methods, which suffer from high false positive rates. In this work we computed a modified genetic distance on a 10,000 bp sliding window between sets of three samples each from CHB, CEU and YRI samples from the 1000 Genomes Project. We applied SOLVINGSET, a distance-based outlier detection method capable of mining hundreds of thousands of multivariate entries in a computationally efficient manner, to the average pairwise distances obtained from each window for each CHB-CEU, CHB-YRI and CEU-YRI to compute the top- $n$  genic windows exhibiting the highest scores for the three distances. The outliers detected by this approach were screened for their biological significance, showing good overlap with previously known targets of differentiation and positive selection in human populations.

**Keywords** Distance-based outlier • Whole-Genome scan

---

S. Lodi (✉) • C. Sartori

Department of Computer Science and Engineering, University of Bologna, 40136 Bologna, Italy  
e-mail: [stefano.lodi@unibo.it](mailto:stefano.lodi@unibo.it)

F. Angiulli

Department of Computer Engineering, Modelling, Electronics, and Systems,  
University of Calabria, 87036 Rende, Italy

S. Basta

Institute of High Performance Computing and Networking, Italian National Research Council,  
87036 Rende, Italy

D. Luiselli

Department of Biological, Geological and Environment Sciences, University of Bologna,  
40126 Bologna, Italy

L. Pagani

Department of Archaeology and Anthropology, University of Cambridge, Cambridge, UK

## 1 Introduction

*Homo sapiens* is one of the most widespread species in the world, inhabiting environments that span from the arid Savannah to the icy Siberia [25]. During their journey to the colonization of most of the emerged lands, humans developed cultural and genetic adaptation to cope with the diverse challenges posed by the many environments they encountered.

While the “smoking-gun” non-synonymous mutations are relatively easy to detect in genes for which a specific genotype-phenotype relationship is already established, genome wide scans for signatures of positive selection pose a higher degree of complexity. Firstly, the managing of whole genome sequences needs the development of specific tools capable of handling millions of positions in a computationally effective manner. Secondly, the random processes responsible for the neutral accumulation of mutations and their frequency in a population (genetic drift) constitute the major confounder in this kind of quest. Some of the most popular selection tests in the field of human evolutionary genetics ( $F_{ST}$  [32], iHS [31], XP-EHH [28], PBS [33]), designed to detect signature of positive selection in the genome of a given population, indeed rely on the empirical ranking of the statistic they calculate. While genomic loci that underwent positive selection would surely fall in the top ranking regions for any carefully designed statistic, the same distribution subset would also be populated by merely drifted regions. Furthermore should a population show no signs of positive selection, it would still be possible to define a set of top ranking regions, which would obviously be simply representing genetic drift in that population. Therefore more accurate tools capable of efficiently managing whole genome data and of focussing on highly differentiated regions without relying on empirical ranking, are crucial to the development of the field. Additionally, as sequencing efforts are gradually shifting from a “many samples, low coverage” to a “few samples, high coverage” strategy, high accuracy when dealing with as little as 5 samples per population would be desirable feature of such tools. The deployment of units of samples per each population would also reduce the accuracy of popular frequentist methods, such as  $F_{ST}$ , for which such a number of individuals would not be sufficient to generate reliable frequency estimates.

Outliers, or anomalies, are observations which deviate significantly from the remaining data [18]. The occurrence of outliers signals the presence of a different data generating mechanism, which in turn may be correlated to harmful conditions, economic losses, malfunctioning devices, but also interesting novelties from which new knowledge can be extracted. Outlier detection is an important data mining problem, which is computationally difficult to solve for large data sets meeting the temporal requirements of its typical applications, which include fraud detection, intrusion detection, data cleaning, medical diagnosis, mechanical failure forecasting, and network analysis.

In this paper we will be concerned with the *unsupervised* outlier detection problem, in which exceptional and normal data must be separated without the help of training examples. A prominent approach to the unsupervised problem is



*distance-based* outlier detection, which bases the distinction on distances to a subset of all data set objects [4, 6, 7, 10, 17, 21, 26, 30]. Most of these approaches define a *weight* or *score* for every object, which summarizes its dissimilarity to its  $k$  nearest neighbors by means of a function of their distances.

Distance-based outlier detection is non-parametric, in that no assumption is required on the distribution of data. It is thus more widely applicable than model-based approaches, which base the identification of an object as outlier on the probability of suitably defined tail regions of the assumed distribution. Distance-based outlier detection is also computationally more difficult than model-based outlier detection. For this reason, numerous parallel and distributed distance-based outlier detection methods have been proposed obtaining large speed-up over sequential ones. Some of these contributions present algorithms for Graphic Processing Units, which contain thousands of computing cores that can execute general purpose programs and cost a fraction of a computer cluster. Therefore, we believe that distance-based outlier detection is mature for application to the genomic domain, and, in particular, to the analysis of the human genome. In this study we set out to apply SOLVINGSET, a multidimensional outlier detection method, to search for genomic regions highly differentiated among modern human populations.

## 2 Related Work

To the best of our knowledge, no widely adopted test to detect signatures of positive selection relies on multi-dimensional outlier detection. The tests available so far can be divided into three major classes. SNP based tests (such as  $F_{ST}$  [32] or PBS [33] tests) focus on genetic signals stemming from a single genetic position which can be subsequently used to calculate the average or maximum value over a genetic window of a given size. Haplotype based test, which include iHS [28, 31] and XP-EHH [28], evaluate the length and frequency of haplotypes in a given populations, flagging out genomic regions showing outstanding haplotype patterns. The third class of tests (Tajima's  $D$  [29], Fay and Wu's  $H$  [16]) relies on the site frequency spectrum (calculated per each genomic window of a given length) and still flags outstanding regions based on their overall genomic ranking.

Many methodologies for outlier detection have been proposed in the literature of statistics, machine learning and data mining; [19] and [12] are comprehensive reviews of work in the field. In the sequel we recall the most relevant contributions.

Barnett and Lewis [9] present a large collection of univariate, distribution-based outlier tests. More recently, many works in the data mining field have addressed the issue of efficiency of outlier detection in very large data sets motivated by its usefulness in information technology, finance, medicine and mechanics.

Knorr and Ng proposed the  $NL$  algorithm [21], in which an outlier is defined as an object  $o$  such that the fraction of all objects belonging to a closed ball of radius  $D$  and center  $o$  is smaller than a fixed threshold  $s$ . The authors proved elsewhere that their definition can be unified to a distribution-based one. In fact, for popular distributions

and appropriate choices of  $D$  and  $p$ , the outlier property is equivalent to membership in parametrically defined outlier tail regions. Such outliers are distance-based; they are also related to local density, in that the outlier property of an object depends the number of objects in its neighbourhood. This approach was improved in later works. Ramaswamy et al. [27] estimate local density as the  $k$ th nearest neighbour distance and base a ranking of outliers on such distance. Breunig et al. [11] also define a degree for the property of being an outlier, the Local Outlier Factor (LOF), but in contrast to previous proposals, the degree is relative to the density of neighbouring objects. Therefore, in their approach two objects may have similar degree even if their distance to the nearest cluster is very different, because such clusters have very different densities. Bay and Schwabacher [10] first introduced an outlier detection framework in which a running threshold on an object's score allows to exclude objects that cannot be outliers from the computation.

In many applications, the running time of an outlier detection implementation must fall into a feasible range, due to the size of data sets, and, in many cases, because it is instrumental in prompt user intervention. For this reason, there has been an increasing interest in parallel and distributed methods for outlier detection. Hung and Cheung [20] proposed the *PENL* algorithm, which is a parallelization of *NL* [21]. *PENL* transfers the entire dataset among nodes; therefore it has limited applicability in distributed mining. Lozano and Acuna [23] proposed a parallel version of Bay's algorithm [10] which showed restricted scalability in some experiments. Support-based methods for distributed high-dimensional data sets, have been proposed by Otey et al. [24] and Koufakou and Georgiopoulos [22]. Finally, Dutta et al. [14] proposed a top- $k$  outlier detection method which discovers objects that are exceptions to the overall correlation structure of the data, as presented by its principal components.

### 3 Methods

Here we adopt the average pairwise difference (*APD*) as the most basic yet stable statistic, considering over a given number of base pairs the total number of differences between two individuals and dividing by the total number of explored sequence. *APD* was calculated per each 10,000 bp region in all possible  $(p_{1i}, p_{2j})$  pairs between populations, where  $i$  and  $j$  are the  $i$ -th and  $j$ -th individuals of population 1 ( $p_1$ ) and population 2 ( $p_2$ ), taking five individuals for each CEU (European), CHB (Asian), YRI (African) populations, of the ones available from the 1000 Genomes Project. To the *APD* between populations was then subtracted the *APD* within populations and the total divided by the *APD* between:

$$APD = \frac{APD \text{ between populations} - APD \text{ within populations}}{APD \text{ between populations}} \quad (1)$$

Let  $D$  be data set of objects, which is a finite subset of a given metric space. For any object  $p \in D$ , its *weight*  $w_k(p, D)$  in  $D$  is the sum of the distances from  $p$  to its  $k$  nearest neighbors in  $D$ .

Let then  $T$  be a subset of  $D$  of size  $n$ . If there are no objects  $x \in T$  and  $y$  in  $(D \setminus T)$  such that  $w_k(y, D) > w_k(x, D)$ , then we say that  $T$  is the *set of the top  $n$  outliers in  $D$* , that  $w^* = \min_{x \in T} w_k(x, D)$  is the *weight of the top  $n$ -th outlier*, and finally that the objects in  $T$  are the *top  $n$  outliers in  $D$* . If ties on the weight values occur, for some object  $y$  in  $(D \setminus T)$ ,  $w_k(y, D) = w^*$  might hold. In such a case, the objects  $x$  in  $T$  such that  $w_k(x, D) = w^*$  are nondeterministically selected among the ones having the same value of weight.

**Fig. 1** Definition of the top- $n$  outliers of a data set

The same window approach was also taken to calculate  $F_{ST}$  [32] between pairs of the same populations, taking 80 samples for each group. The genomic distribution of  $APD$  calculated using only 5 genomes per population was then compared with the  $F_{ST}$  distribution calculated on 80 samples and taken as the benchmark. We also simulated 50 genomic regions of 200 kbp each in three populations using the MSMS algorithm [15] (command line: `-ms 120 1 -t 136 -r 80 -N 10000 -SF 0 1 1 -Sc 0 1 10000 100 0 -I 3 40 40 40 10 Smark`) applying the specified selection strength only at the beginning of the first two regions on population 1 (labelled as YRI). The remaining 48 regions were run without the selection flags and allowed to differentiate only through genetic drift.

Each 10,000 bp window was then processed for its  $APD$  searching for distance-based outliers. To this end, we adopted the outlier definition given in [4], which we briefly recall; a formal definition is given in Fig. 1.

Every data set object is associated to a *weight*, that is, the sum of the distances from the object to its  $k$  nearest neighbors. Object weight measures the degree of anomaly of an object. Note that the weight of an object can be large even if the object has one or more close neighbours, if the number of such neighbours is much smaller than  $k$ . Object weight induces an ordering of all objects according to their degree of anomaly. Rather than fixing a threshold on weight to select outliers, only the top- $n$  objects having largest weights are selected as outliers, where  $n$  is an additional integer parameter of the definition.

The computation of the top- $n$  outliers of a data set is straightforward by the following basic algorithm. Compute a  $n \times k$  matrix  $A$  in which  $A_{jr}$  is a pair  $(q_{jr}, d_{jr})$  where  $q_{jr}$  is the  $r$ -th neighbour of object  $p_j$  and  $d_{jr}$  is its distance to  $p_j$ . Compute a weight vector  $w$  by  $w_j = \sum_{r=1}^k d_{jr}$ , sort it in descending order, finally select elements from  $w_1$  to  $w_n$ . Such an algorithm is however expensive for large data sets both in terms of computation time and occupied memory, due to the size of matrix  $A$  and the complexity of computing nearest neighbours.

The SOLVINGSET family of algorithms [2, 3, 5, 6] allows for an efficient computation of top- $n$  outliers in a variety of sequential, parallel (GPU), and distributed computing environments, with or without GPU parallel co-processing.

**Algorithm 1** SOLVINGSET**Input:** Multivariate data set  $D$ , distance function  $d$ , positive integers  $k, n, m$ .**Output:** The top- $n$  outliers of  $D$ .

---

```

1: Select  $m$  objects randomly from  $D$  and insert them into an initially empty candidate set  $C$ 
2: while  $C \neq \emptyset$  do
3:   Move all candidates in  $C$  from the data set  $D$  to the solving set  $S$ 
4:   for all  $p \in D$  do
5:     for all  $c \in C$  do
6:       Compute the distance  $d$  between  $c$  and  $p$ 
7:       Update the max-heap  $H_c$  of  $c$  with  $(p, d)$ 
8:       Update the max-heap  $H_p$  of  $c$  with  $(c, d)$ 
9:     end for
10:    Compute the weight  $w$  of  $p$  from the distances in  $H_p$ 
11:    if  $w$  is smaller than the smallest weight in  $T$  then
12:      Remove  $p$  from  $D$ 
13:    end if
14:    Update the min heap  $NC$  of candidates for the next iteration with  $(p, w)$ 
15:  end for
16:  for all  $c \in C$  do
17:    Update the min-heap  $T$  of top- $n$  outliers with  $(c, w)$ 
18:  end for
19:  Copy the  $m$  objects with largest weights from  $NC$  to  $C$ 
20: end while

```

---

Large data sets, with a number of objects of order  $10^6$ , can be processed by the distributed version several times in a few minutes, allowing for a broader exploration of the  $(k, n)$  parameter space.

The sequential SOLVINGSET algorithm [6] for computing top- $n$  outliers is described in Algorithm 1. In contrast to the basic algorithm, SOLVINGSET computes neighbours for all  $|D|$  objects in batches: At each iteration, the algorithm computes the  $k$  nearest neighbours in  $D$  of only a fresh set  $C$  of *candidate* objects of size  $m$ , where  $m$  is a parameter (lines 4–7). The algorithm ensures at the end of each iteration that the top  $n$  objects by weight among all processed candidates are stored in  $T$  (line 16). At termination,  $T$  clearly contains the top- $n$  outliers.

Computing the top- $n$  outliers in batches allows to discard many objects in the course of the computation. Line 8 ensures that the  $k$  nearest neighbours in the set of all current and past candidates of any object  $p \in D$  are stored in  $H_p$  at the end of each iteration. The distance sum in  $H_p$  is thus monotonically non increasing during the computation; it is also an upper bound to the actual weight of  $p$ . Storing both  $T$  and  $H_p$  allows for a large reduction in the number of objects that can be elected as candidates at each iteration. In fact, the weight of the  $n$ -th object in  $T$  is a lower bound to the weight of the  $n$ -th lightest object among the actual top- $n$  outliers of  $D$ . All objects having a smaller weight than the weight of the  $n$ -th object in  $T$  cannot be top- $n$  outliers; thus, they can be removed from the data set (line 10–13).

Storing provisional  $k$ -neighbourhoods in  $H_p$  also allows to make an informed choice of the candidates for the next iteration. Lines 14 and 19 ensure that  $C$  contains the  $m$  objects having largest weight, basing weight on the distances in  $H_p$

at the previous iteration. Such a selection of  $C$  accelerates the increase speed of the smallest weight in  $T$ , thereby allowing for a more effective removal of objects that cannot be outliers at lines 10–13.

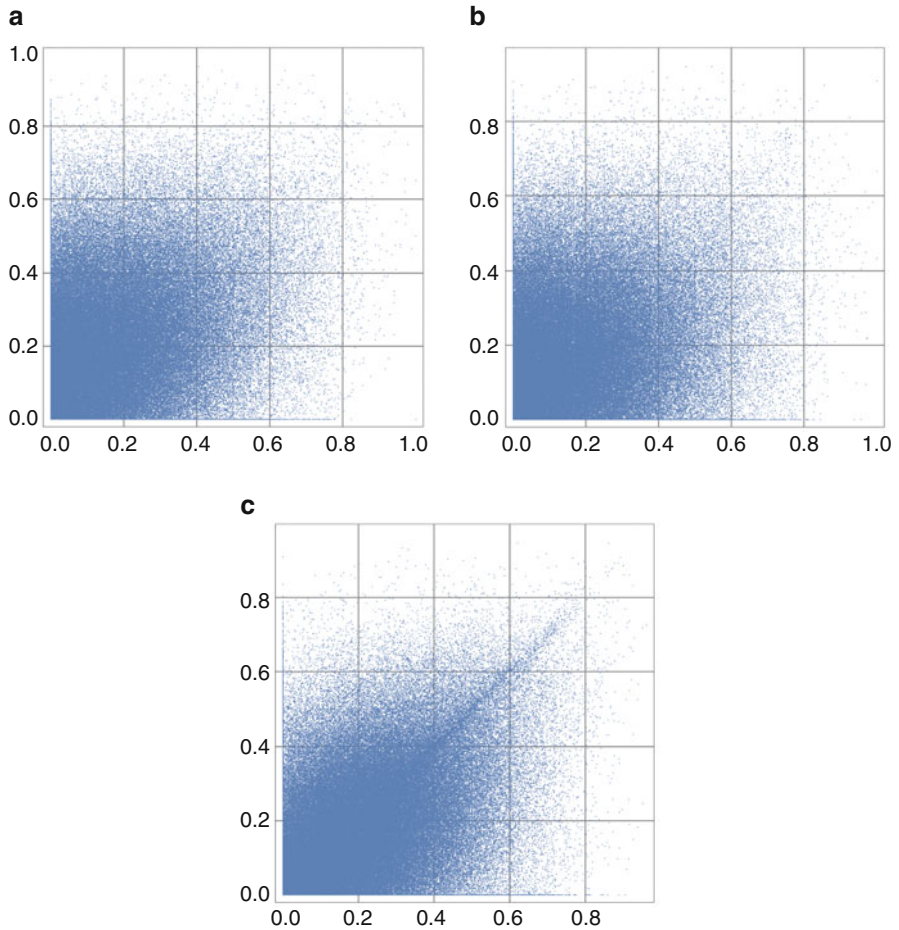
## 4 Experimental Results

A distributed version of the SOLVINGSET algorithm was applied to a six-column data table with schema  $(chr, start, end, CHB-CEU, CHB-YRI, CEU-YRI)$ , where  $chr$ ,  $start$ , and  $end$  are the chromosome, the start base and the end base position of a region, respectively, and  $CHB-CEU$ ,  $CHB-YRI$ ,  $CEU-YRI$  are the average pairwise differences between the Asian and European, Asian and African, and European and African populations.

The preliminary run on simulated data showed the ability of SOLVINGSET to capture 100 % of true positive when taking the 50 windows above the elbow of the ranked distribution of weights. Furthermore we noticed that the  $\sim 50$  % false positives present in the above set could be further reduced, since the true positives all fell in the top of the weight distribution.

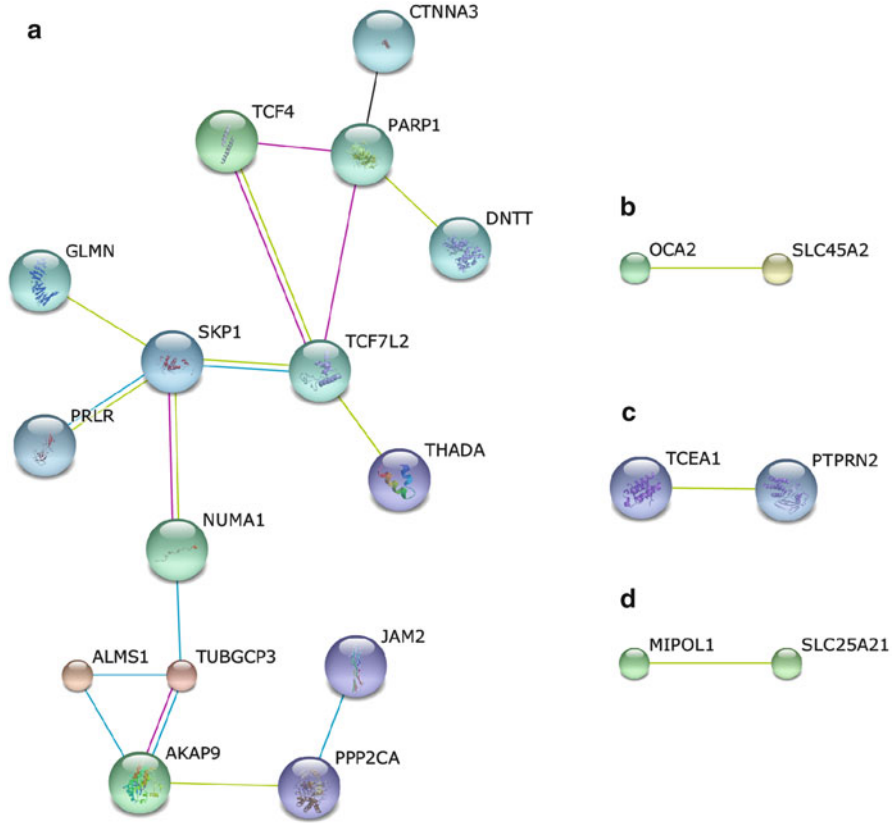
Figure 2 shows instead a plot of the genetic distances calculated on empirical data. We chose to focus only on autosomic regions, to avoid spurious results due to the different demographic dynamics of sex-chromosomes and mtDNA. The distance between regions was computed as the Euclidean distance in the three-dimensional space having coordinates  $CHB-CEU$ ,  $CHB-YRI$ ,  $CEU-YRI$ . The number  $n$  of regions to retrieve as distance-based outliers was set to 260, which is about 0.1 % of the total number of regions (264,908). The number of  $k$  of nearest neighbours was set to 50. Therefore, groups of outliers which are close to each other but separated from the rest of the data will have a weight that decreases as their size increases and will eventually become flatter at sizes  $\geq 50$ .

144 out of the 261 (55 %) outlier regions, each spanning 10,000 base pairs (bp), contained at least 1 gene. After inspecting all 10,000 bp autosomal windows against the human GTF file at Ensemble we observed that 16 % of windows contained at least 1 gene. Therefore the 3.44 folds enrichment in the number of gene-containing and, accordingly, functionally meaningful windows accounts for the biological relevance of the obtained outliers as a whole. The 98 unique genes covered by the 144 windows mentioned above were searched for protein-protein interactions using String 9.1 ([www.string-db.org](http://www.string-db.org)). Four protein-protein networks were identified (Fig. 3) of which one (Fig. 3a) contained the TCF7L2 and THADA genes, known to be associated with Type II Diabetes [8] and showed an increased presence (among others) of genes linked with the positive regulation of macromolecules metabolism. Another network (Fig. 3b) linked OCA2 and SLC45A2, known to regulate the skin pigmentation phenotype [25]. When compared with the top 1 % of the  $F_{ST}$  results obtained for the CEU-YRI, CEU-CHB and YRI-CHB pairs, 31, 28, and 23 % of the 98 genes were found respectively. Furthermore, 16 % of these



**Fig. 2** Plot of average pairwise differences between the Asian and European, Asian and African, and European and African populations. (a) CHB-CEU and CHB-YRI; (b) CHB-CEU and CEU-YRI; (c) CHB-YRI and CEU-YRI

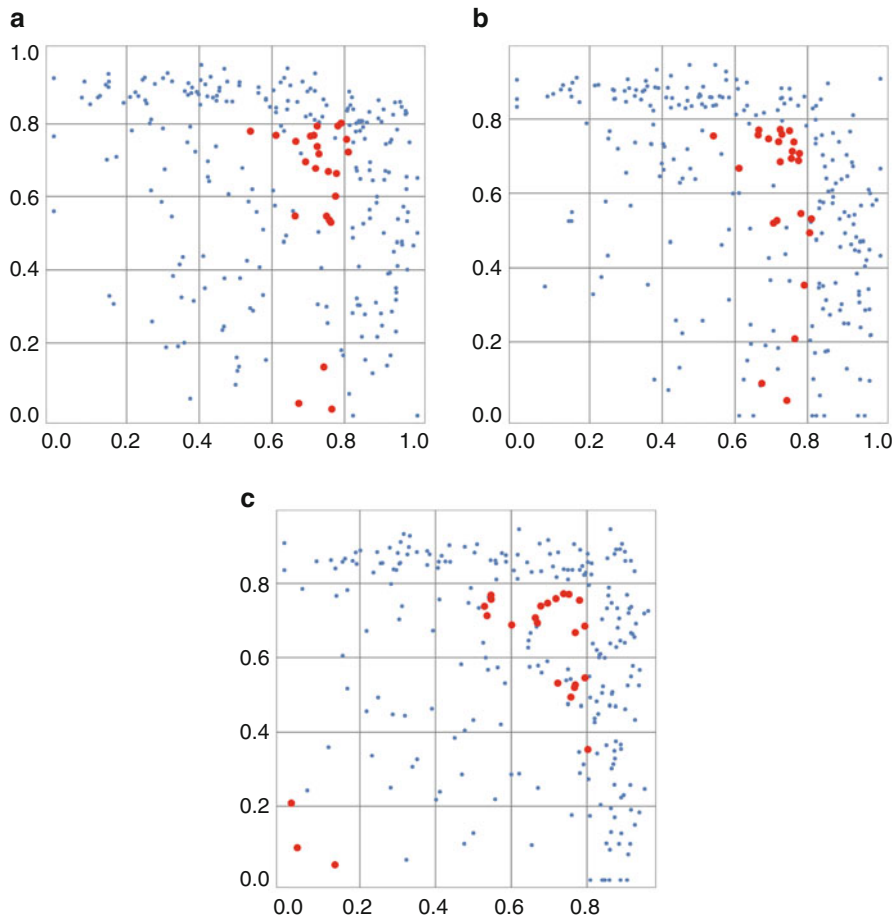
genes were already reported in the literature as known target of positive selection [1, 13, 25]. Combining the  $F_{ST}$  and literature evidences, 55% of the 98 genes characterized by SOLVINGSET (our method) were validated as putative candidate of positive selection in worldwide human populations. Finally, note that simple methods which exclude all windows having a value in the difference columns CHB-CEU, CHB-YRI, CEU-YRI below a given threshold are not equivalent to ours, when the number of retrieved windows is the same. In fact, in our experiment the set union of all top 0.1% windows in the CHB-CEU, CHB-YRI, CEU-YRI does not cover entirely the set of outlier windows which have been found by our method, which is shown in Fig. 4; red points represent the subset of such windows



**Fig. 3** Protein-protein interaction network (from *string-db.org*) of the genes covered by the outlier regions. The 98 unique genes covered by the outlier regions were inputted into the *string-db.org* database to search for protein-protein interaction networks. Twenty-one of these genes formed 4 distinct networks. The first (a) encompassed 15 genes including TCF7L2 and THADA, associated with Type-II Diabetes, and showed significant enrichment of genes involved in the metabolism of macromolecules. The second (b) network includes two of the best characterized genes involved in skin pigmentation, while the other 2 genes (c and d) did not offer a biologically univocal interpretation

which are outside the top 0.1 % windows in all three difference columns CHB-CEU, CHB-YRI, CEU-YRI. Such windows missed by threshold methods include in particular network c of Fig. 3.

**Discussion** The outlier windows detected with our method, as a whole, must be seen as genomic regions which either underwent extreme differentiation in one of the three assessed populations or experienced different genetic pressures in those populations. As a result, the gene list identified with this approach is not necessarily linked with putative selective pressure acting on a specific population. On the contrary, such list should be intended as an overview on the genomic regions which



**Fig. 4** Plot of outlier windows found by our method; red points are windows which are not included in the top 0.1% of any of the three axes CHB-CEU, CHB-YRI, CEU-YRI. **(a)** CHB-CEU and CHB-YRI; **(b)** CHB-CEU and CEU-YRI; **(c)** CHB-YRI and CEU-YRI

are mostly differentiated across continents. In this light it is remarkable to note how these regions show a 3.44 folds enrichment in gene contents. Furthermore they feature protein-protein interaction networks such as the ones involving type-II diabetes or skin pigmentation genes, both known to have played a major role in the genetic adaptation to the various environments encountered during the human worldwide expansion.



## 5 Conclusions

We have presented a first application of SOLVINGSET, a distance-based outlier detection algorithm to the problem of extracting outliers from a three-dimensional data set in which each dimension records differences in base pairs between homologous windows over the genomes of individuals from two populations among CHB, CEU, and YRI from the 1000 Genomes Project, to the purpose of identifying candidate genes for positive selection. The outlier windows found by our method have been compared with the results of  $F_{ST}$  taken as benchmark and have been examined for their biological significance; the comparison yielded a validation of 55 % of the genes contained in the found outlier windows as putative candidates of positive selection. Furthermore, its computational efficiency in tackling a multidimensional problem makes SOLVINGSET a tool capable of improving the current knowledge on highly differentiated genomic regions across multiple populations.

**Acknowledgements** This work has been partially supported by the Italian Ministry of Education, Universities and Research under PRIN Data-Centric Genomic Computing (GenData 2020) and by CINECA ISCR project HIOXICGP. Luca Pagani would like to thank Guy Jacobs for his help with simulations. The authors have no conflict of interests to declare.

## References

1. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.: An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65 (2012)
2. Angiulli, F., Basta, S., Lodi, S., Sartori, C.: Distributed strategies for mining outliers in large data sets. *IEEE Trans. Knowl. Data Eng.* **25**(7), 1520–1532 (2013)
3. Angiulli, F., Basta, S., Lodi, S., Sartori, C.: Fast outlier detection using a gpu. In: International Conference on High Performance Computing and Simulation (HPCS), pp. 143–150 (2013)
4. Angiulli, F., Pizzuti, C.: Outlier mining in large high-dimensional data sets. *Trans. Knowl. Data Eng.* **2**(17), 203–215 (2005)
5. Angiulli, F., Basta, S., Lodi, S., Sartori, C.: Accelerating outlier detection with intra- and inter-node parallelism. In: International Conference on High Performance Computing and Simulation (HPCS), pp. 476–483. IEEE, Bologna, Italy, 21–25 July (2014)
6. Angiulli, F., Basta, S., Pizzuti, C.: Distance-based detection and prediction of outliers. *Trans. Knowl. Data Eng.* **18**(2), 145–160 (2006)
7. Angiulli, F., Fassetti, F.: Dolphin: an efficient algorithm for mining distance-based outliers in very large datasets. *ACM Trans. Knowl. Disc. Data* **3**(1), 4:1–4:57 (2009)
8. Ayub, Q., Moutsianas, L., Chen, Y., Panoutsopoulou, K., Colonna, V., Pagani, L., Prokopenko, I., Ritchie, G.R.S., Smith, T.C., McCarthy, M.I., et al.: Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *Am. J Hum. Genet.* **94**(2), 176–185 (2014)
9. Barnett, V., Lewis, T.: *Outliers in Statistical Data*, 3rd edn. Wiley, Chichester (1994)
10. Bay, S.D., Schwabacher, M.: Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: *Knowledge Discovery and Data Mining* (2003)

11. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104. ACM, New York, USA (2000)
12. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. *ACM Comput. Surv.* **41**(3), 15:1–15:58 (2009)
13. Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., Garrison, E., Xue, Y., Tyler-Smith, C., et al.: Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol.* **15**(6), R88 (2014)
14. Dutta, H., Giannella, C., Borne, K.D., Kargupta, H.: Distributed top-k outlier detection from astronomy catalogs using the DEMAC system. In: *SDM* (2007)
15. Ewing, G., Hermisson, J.: Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**(26), 2064–2065 (2010)
16. Fay, J.C., Wu, C.I.: The neutral theory in the genomic era. *Curr. Opin. Genet. Dev.* **11**(6), 642–646 (2001)
17. Ghoting, A., Parthasarathy, S., Otey, M.E.: Fast mining of distance-based outliers in high-dimensional datasets. *Data Min. Knowl. Disc.* **16**(3), 349–364 (2008)
18. Han, J., Kamber, M.: *Data Mining, Concepts and Technique*. Morgan Kaufmann, San Francisco (2001)
19. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. *Artif. Intell. Rev.* **22**, 85–126 (2004)
20. Hung, E., Cheung, D.W.: Parallel mining of outliers in large database. *Distrib. Parallel Dat.* **12**(1), 5–26 (2002)
21. Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: *VLDB*. pp. 392–403 (1998)
22. Koufakou, A., Georgiopoulos, M.: A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. *Data Min. Knowl. Disc.* (2009, Published online)
23. Lozano, E., Acuña, E.: Parallel algorithms for distance-based and density-based outliers. In: *ICDM*. pp. 729–732 (2005)
24. Otey, M.E., Ghoting, A., Parthasarathy, S.: Fast distributed outlier detection in mixed-attribute data sets. *Data Min. Knowl. Disc.* **12**(2–3), 203–228 (2006)
25. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., et al.: Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**(5), 826–837 (2009)
26. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: *SIGMOD*, pp. 427–438 (2000)
27. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 427–438. ACM, New York, USA (2000)
28. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.: Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**(7164), 913–918 (2007)
29. Tajima, F.: Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* **123**(3), 585–595 (1989)
30. Tao, Y., Xiao, X., Zhou, S.: Mining distance-based outliers from large databases in any metric space. In: *KDD*, pp. 394–403 (2006)
31. Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K.: A map of recent positive selection in the human genome. *PLoS Biol.* **4**(3), e72 (2006)
32. Wright, S.: Isolation by distance under diverse systems of mating. *Genetics* **31**(1), 39 (1946)
33. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al.: Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**(5987), 75–78 (2010)

# Predicting the Metagenomics Content with Multiple CART Trees

Dante Travisany, Diego Galarce, Alejandro Maass, and Rodrigo Assar

**Abstract** Metagenomics is a technique for the characterization and identification of microbial genomes using direct isolation of genomic DNA from the environment without cultivation. One of the key step in this process is the taxonomic classification and clustering of the DNA fragments, process also known as *binning*. To date, the most common practice is classifying through alignments to public databases. When a representing specie is present in this database the process is simple and successful, if not, an underestimation of taxonomic abundances is produced. In this work we propose a alignment-free method capable of assign taxa to each read in the sample by analyzing the statistical properties of the reads. Given an environment, we collect genomes from public available databases and generate genomic fragments libraries. Then, statistics of k-mer frequencies, GC ratio and GC skew are computed for each read and stored in an environment-associated dataset used to build a robust machine learning procedure based on multiple CART trees. Finally, for each read the CART trees are asked about their taxa and the most voted ones are selected. The method was tested using simulated and public human gut microbiome data sets. The database was constructed using 98 genera present in Gastrointestinal Tract available at Human Microbiome Project. A multiple CART tree with 558-trees predictor was generated, capable to estimate the genus and abundance in the sample with 47 % of accuracy in read assignments. Performance rates are comparable with those from semi-supervised methods and also the computation times were reduced due to alignment-free methodology. Restricted to 17 early considered genera, our method increases its accuracy to 77 %.

**Keywords** Metagenomics content prediction • Human gut microbiome • CART trees • K-mer frequencies

---

D. Travisany • D. Galarce • A. Maass

Departamento de Ingeniería Matemática, Center for Mathematical Modeling,

Universidad de Chile, Santiago, Chile

e-mail: [dtravisany@dim.uchile.cl](mailto:dtravisany@dim.uchile.cl); [d.galarce.castro@gmail.com](mailto:d.galarce.castro@gmail.com); [amaass@dim.uchile.cl](mailto:amaass@dim.uchile.cl)

R. Assar (✉)

Instituto de Ciencias Biomédicas, Escuela de Medicina, Universidad de Chile, Santiago, Chile

e-mail: [rodrigo.assar@gmail.com](mailto:rodrigo.assar@gmail.com)

## 1 Background

During the last years, metagenomics has consolidated as the technique to characterize and identify microbial genomes using a direct isolation of genomic DNA sequences from the environment [12, 24]. A common metagenomics project begins with the selection of an environment where genomic DNA from the microorganisms is obtained and transformed into digital information by high-throughput sequencing technologies like 454 pyrosequencing [18] and Illumina [22]. Using several algorithms and bioinformatics tools, metagenomics attempt to answer three main questions:

Who is out there? know what microorganisms live in that specific environment and estimate the relative abundance of each taxa.

What are they doing?, to identify genes and metabolic profiles of the samples.

Who is doing what? to associate gene functions to the different microorganisms present in the sample.

Consequently, a bioinformatics pipeline could be divided into the following steps:

Pre-processing and normalization, the results obtained from sequencing are filtered from low quality reads and sequencing artifacts, in order to conserve only the high quality data.

Binning and assembly, consist in the taxonomic classification of the reads (Who is out there?) and assembly to extend the reads into contigs or scaffolds.

Annotation process to find coding regions and assign a function to the reads (contigs, scaffolds) using public databases (What are they doing?) and finally the storage to retrieve the results and associate microorganisms and genes (Who is doing what?).

An emblematic example is the study of the Sargasso sea [27] in which researchers discover new bacteria species and adaptation mechanisms. More recently, the study of the human microbiome and its relation with infections and chronic diseases [19] has emerged as a new approach for diagnosis and treatments in medicine [1]. Other studies demonstrate how the unbalances in microbiota are associated with modulation of complex diseases such as some types of cancer (e.g. [26]). The list of examples, projects and metagenomics data is extensive and it is growing very fast. Nevertheless, the computational methods for metagenome analysis still present many problems. In contrast, genomic sequences of single species can usually be assembled and analyzed by many available methods. In metagenomics, the coexistence of different microorganisms and the short length of the DNA reads make difficult their assignments and assembly [7, 23]. Thus, the main difficulties arise while binning, to face this challenge, a variety of tools and methods have been developed, the most common consists in alignment matching against protein/nucleotide databases. Unsupervised methods for clustering (CompostBin [5], AbundanceBin [29]) coupled with the alignment matching and semi-supervised methods (like phymmBL for 454 technology [3]) are able to process reads of short lengths and also, phymmBL can improve results using higher taxonomy levels for

reads of unknown genus. Although that tool is accurate at genus-level, is still based on computationally intensive alignments.

An alignment-free method is proposed. This method assigns a taxa to each read by analyzing its statistical properties. Given a studied environment, the method receives as input a list of possible taxa that should cohabitate within the environment and a sample of DNA reads. With the list of possible taxa synthetic libraries of DNA reads from known genome sequences are simulated using 454-Sim [16]. For each synthetic read, k-mer frequencies (k from 1 to 4), GC ratio and GC skew are computed, generating the environment-associated training and validation dataset. Both datasets are used to build a machine learning procedure based on multiple CART trees [4, 11]. For each tree the training data set is used to iteratively ask about specific pattern frequencies, until obtain a tree with high taxonomic homogeneity leaves. Afterward, the tree is pruned to improve homogeneity of leaves using the validation dataset. Subsequently, for each read within the sample, the same pattern frequencies are computed and then, the multiple CART trees are interrogated about the taxa. The most voted taxon is chosen for the read, subsequently the abundance consists in the sum of reads classified in a taxon. In addition, the importance of each DNA pattern for the built predictor is obtained.

The approach was applied to the human gut microbiome. Given a set of new samples of DNA reads from a human gut, the method is able to estimate the genus of each read and consequently the abundance of each genera. Among the most important pattern predictors outstand *GC ratio*, *GC*, 3-mers *GCG* and its reverse complementary *CGC*. The method was constructed using simulated data from 454-sim, tested using BEAR [13] simulated dataset and one public real datasets of guts from geographically distributed children [30]. The accuracy of the method reach 47 % in total read assignments, confusing mostly by closed related genera.

## 1.1 Organization of the Paper

The following section, *Methods*, provides a brief explanation about the data, techniques and algorithms used in this work. Next in Sect. 3, *Results*, we show and analyze the classifications on real and simulated data. Finally in Sect. 4, *Conclusions*, we discuss the scopes of our results and future developments.

## 2 Methods

This alignment-free classification method based on CART trees [4] was generated using a three step methodology.

First, a compilation of the genomic sequences available from putative candidate bacteria present within the studied environment is obtained, then, we produce simulated datasets from these taxa. The second step consist in using a priori

chosen DNA pattern statistics (explanatory variables) and computing these over the simulated and real data sets. The predictor is built using the statistics of these simulated data set. Then is interrogated for each data set in order to retrieve the taxa and finally obtain the abundance.

The definition of the classification problem, the process of data acquisition and preprocessing, the construction of CART trees, the implementation, the estimation of DNA-pattern importance and the validation are described in Sects. 2.1–2.6.

## 2.1 *The Classification Problem: Two Cases*

The classification problem consists in assign a taxon to a read given the frequency of DNA-patterns. The class to predict is the taxon, the explanatory variables are 342 consisting in k-mer frequencies (with  $k \in \{1, \dots, 4\}$ ), GC ratio, and GC skew. The frequencies are computed for each variable and standardized to avoid the effect of the read length.

In this manuscript we consider two training data set to choose the taxa:

- The set of taxa consists of 17 genera that were early studied at the gastrointestinal tract: *Acidaminococcus*, *Akkermansia*, *Alistipes*, *Bacteroides*, *Bifidobacterium*, *Coprococcus*, *Eggerthella*, *Escherichia*, *Eubacterium*, *Faecalibacterium*, *Megasphaera*, *Parabacteroides*, *Prevotella*, *Roseburia*, *Ruminococcus*, *Shigella*, *Streptococcus*.
- The 98 genera available to date in the gastrointestinal tract from the Human Microbiome Project. More details in Sect. 2.2. This list is available at <http://metagenomics.cmm.uchile.cl/CART>.

## 2.2 *Metagenomics Data Acquisition and Processing*

In order to produce taxon-associated training and pruning datasets. Four hundred and fifty four genomes comprising 98 genera available as projects for Gastrointestinal Tract in the Human Microbiome Project (HMPGIT) [25] were downloaded, every genome was fragmented using fragsim ( -l 1000 -c 10000 ). All fragments were filter to remove ambiguous nucleotides. Next, using GATB (Genome Assembly & Analysis Tool Box) [8] and in-house PERL scripts, statistics of k-mer (1,2,3,4-mer) frequencies, GC ratio and GC skew were computed for each fragment.

To construct the synthetic test dataset BEAR (Better Emulation for Artificial Reads) was used. Briefly, BEAR can emulate reads from various sequencing platforms, using a unique method for deriving run-specific error rates extracting useful statistics from the metagenomic data itself, such as quality-error models [13]. Consequently, to generate the error model the input for BEAR were reads from a real 454-GS Junior experiment, using high complexity abundance flag and the HMPGIT

data set as genome database, the maximum and minimum lengths for synthetic test dataset were set to 50 and 1000 basepair respectively. After the creation of the metagenomics synthetic sample, artificial duplicates were removed using CD-HIT-454 [9] (default parameters). The low quality reads were also removed using FASTX toolkit [10] (phred quality  $\geq 20$ , minimum read length 300). The next step consist in compute the statistics for each read and finally the classification with Multiple CART trees.

### 2.3 *Building Multiple CART Trees*

The classification method CART (Classification And Regression Trees [4, 11]) was used as prediction tool. This machine learning method is based on simple questions about explanatory variables which generate a decision tree. Starting at the root with training data, reads are iteratively asked about specific pattern frequencies until obtaining nodes with low class impurity. The class impurity of a given node is computed by the Gini index, and the class is chosen minimizing the node classification error (Fig. 1).

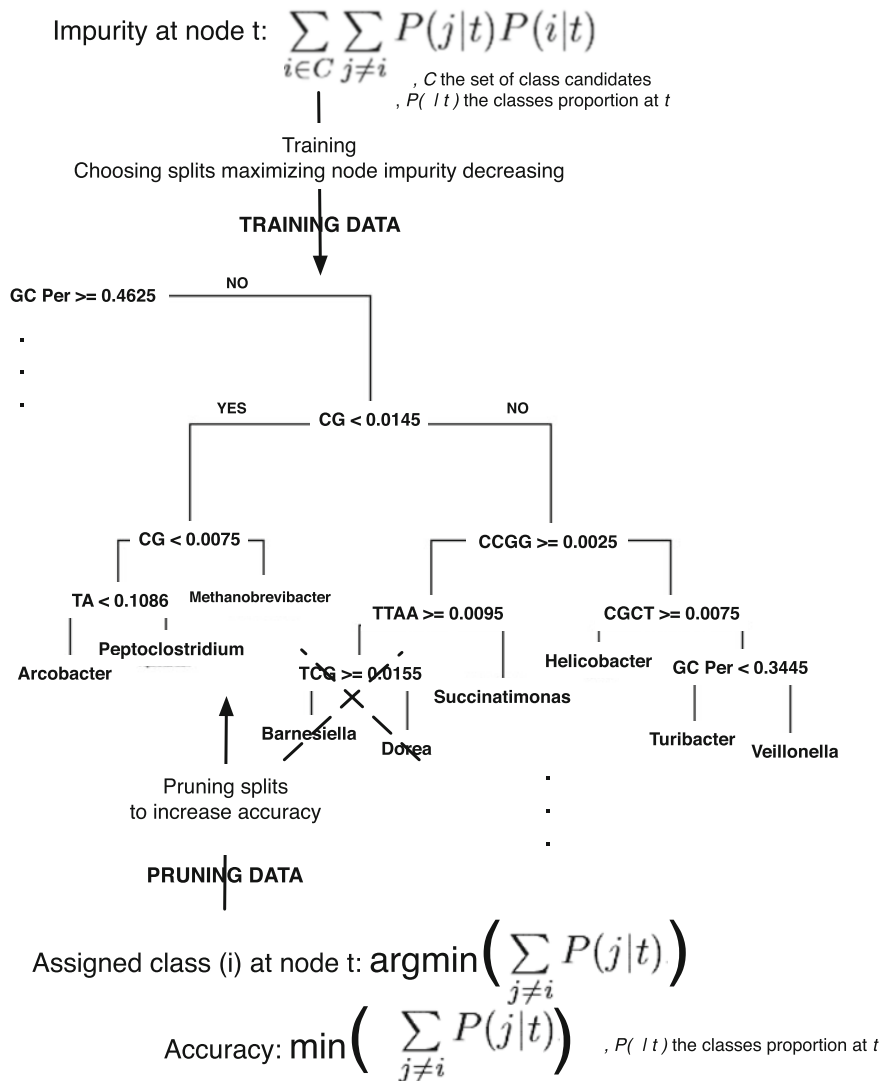
To decrease overtraining, the trees were pruned to improve the results on pruning data. During the pruning process we began with a deep tree and sequentially pruned it until minimize the misclassification of the tree. In addition, multiple trees were generated by randomly choosing of training and pruning datasets. Thus, the same classification problem (with the whole set of taxa) was solved independently by each tree. The set of trees, forest, is the final predictor.

Consequently (see Sect. 2.1) two multiple CART predictors were built:

- CART17 assigns the taxon among the 17 genera early described in literature.
- CART98 assigns the taxon among the whole set of genera at the gastrointestinal tract database.

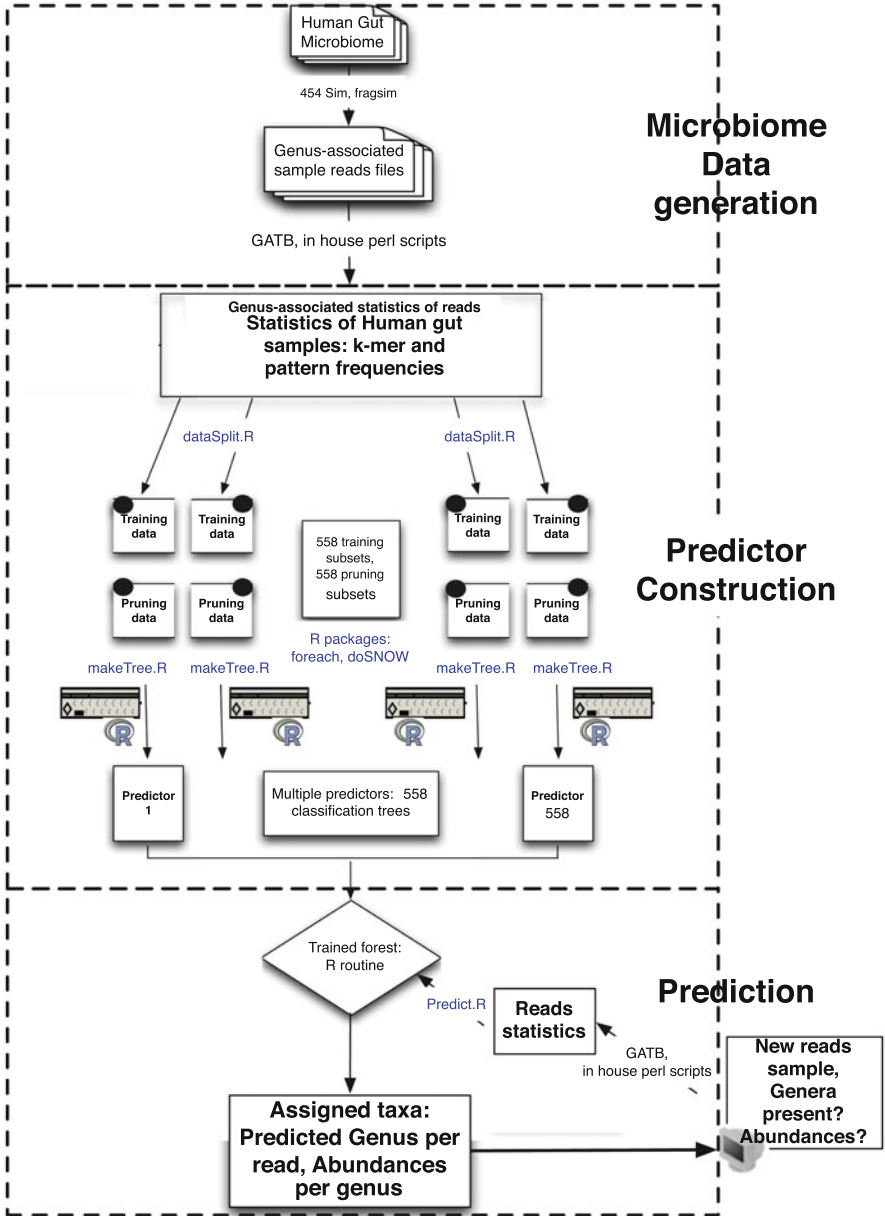
### 2.4 *Implementation*

Trees are built using the R-package rpart. In order to control each tree and run them in parallel, RandomForest was not used. Three implementation levels were considered: Data generation, Construction, Prediction (Fig. 2). Multiple trees were built separately and packed into a R-list object. The structure of the implementation of each tree is summarized in the Fig. 3. Processing times were reduced computing each tree in parallel. In the prediction steps, for each read the pattern frequencies are computed and the genus predicted over the a-priori built forest. Every tree vote for a taxon and the read is assigned to the most voted taxa. This three-level separation allows to store the forest within a host server and calling genus predictions from local machines.

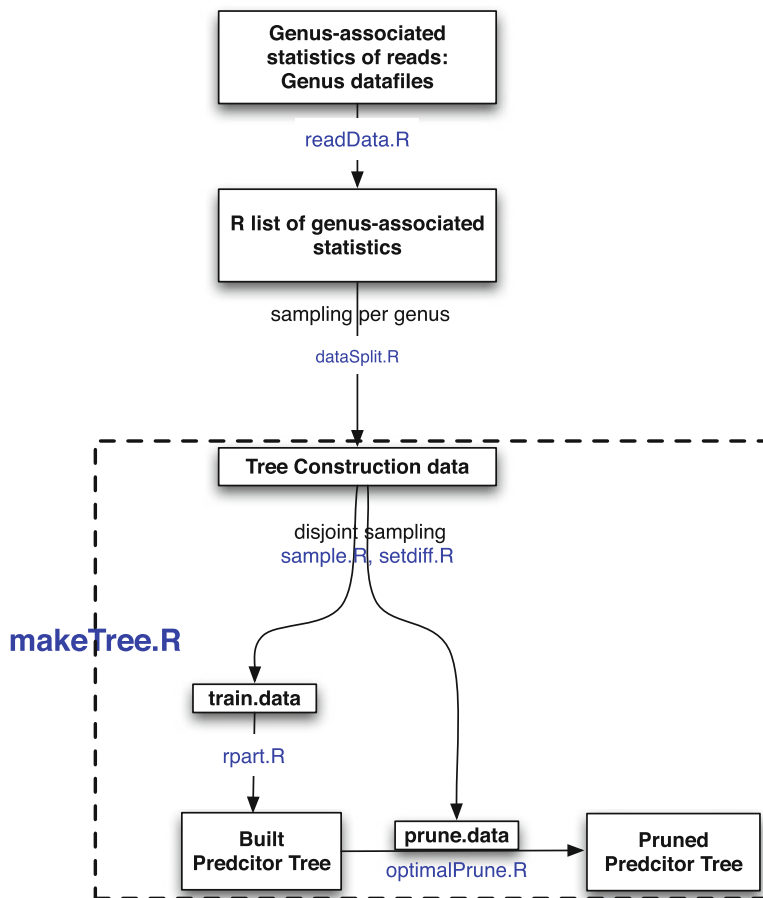


**Fig. 1** Classification tree. Using the training data, each node of the tree is iteratively subdivided into two finer nodes depending on values of variables until obtaining lowest impurities. After that, trees are pruned to improve accuracy on pruning data. Majority class is assigned at each terminal node





**Fig. 2** Multiple CART trees method. Three implementation levels: Data generation, Construction, and Prediction. Specification for the microbiome application



**Fig. 3** Implementation of tree generation with R. Function `readData.R` builds a R list of genus-associated statistics, which is used by `dataSplit.R` to create the construction data.frame with random selection of reads assuring equal genus representation. In `makeTree.R` the construction data.frame is randomly separated into two data.frames for building the tree and pruning it

## 2.5 DNA-Patterns Importance

The importance of each DNA-pattern to classify reads by genera was computed as the number of times a pattern is used to split a node and the classification-improvements were measured by decreasing impurity due to these splits (Fig. 1). Classification improvements were averaged across the CART trees produced by the method.

## 2.6 Validation and Comparison with Other Binning Methods

Validation datasets were generated independently from training and pruning data. They were composed by:

- Metasim generated training and pruning data sets for the 17 taxa [20],
- 454-Sim generated training and pruning data sets for the 98 taxa [16],
- BEAR-generated test data sets [13],
- Geographically distributed (*GD*) data sets, three random samples were taken from children of Venezuela (Amazonia), Malawi and United States of America [30].

In silico generated datasets (first, second and third) have known genera. In case of the fourth dataset [30], estimated abundances predicted by MG-RAST [17] were considered..

Thus, the classification methods that we considered to compare are:

- MG-RAST [17].
- PhymmBL [3].

## 3 Results

As described in Sect. 2.1, two predictors were built: one with 17 genera early studied at the gastrointestinal tract, and the current 98 genera of the gastrointestinal tract currently available at The Human Microbiome Project. The last case was deeply cover. R codes for building our predictor, data sets for building it and validating it are available in <http://metagenomics.cmm.uchile.cl/CART>.

### 3.1 Results for CART17

A set of 150,000 reads were generated for training and pruning using the MetaSim software [20] emulating 454-reads. Distribution of reads per genus was chosen according to available sequences: 245 Acidaminococcus, 256 Akkermansia, 2406 Alistipes, 40,981 Bacteroides, 31,404 Bifidobacterium, 1277 Coprococcus, 331 Eggerthella, 29,118 Escherichia, 2378 Eubacterium, 4468 Faecalibacterium, 220 Megasphaera, 427 Parabacteroides, 1693 Prevotella, 4401 Roseburia, 2731 Ruminococcus, 4340 Shigella, 23,324 Streptococcus.

For this case, 1500 trees predictor *CART17* with accuracy of 77 % was generated. Read assignments for each genus are shown in Table 1. The highest misclassification rate is due to *Shigella-Escherichia* confusion, which are often declared within the same genus [14, 31].

**Table 1** Read assignments matrix for *CART17*

| Real genus | (1) | (2) | (3) | (4)  | (5)  | (6)  | (7) | (8)  | (9)  | (10) | (11) | (12) | (13) | (14) | (15) | (16) | (17) |
|------------|-----|-----|-----|------|------|------|-----|------|------|------|------|------|------|------|------|------|------|
| (1)        | 100 | 0   | 0   | 0    | 0    | 0    | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| (2)        | 0   | 100 | 0   | 0    | 0    | 0    | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| (3)        | 0   | 0   | 100 | 0    | 0    | 0    | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| (4)        | 0.3 | 0.1 | 0.7 | 69.2 | 1.6  | 2    | 0   | 4.9  | 3.1  | 1.6  | 0.1  | 0.7  | 1.2  | 6.3  | 4.1  | 0.7  | 3.2  |
| (5)        | 0   | 0.3 | 4.6 | 1.8  | 83.2 | 0    | 0.7 | 1.6  | 0    | 5.8  | 0.2  | 0    | 0.7  | 0.4  | 0.4  | 0.1  | 0.1  |
| (6)        | 0   | 0   | 3.6 | 16.1 | 1.8  | 21.4 | 0   | 0    | 7.1  | 23.2 | 0    | 0    | 0    | 16.1 | 10.7 | 0    | 0    |
| (7)        | 0   | 0   | 0   | 0    | 0    | 0    | 100 | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| (8)        | 0.2 | 0.3 | 1.1 | 3.4  | 3.2  | 0.1  | 0.2 | 73.4 | 0.9  | 3    | 0.2  | 0    | 0.6  | 1.5  | 0.5  | 11.2 | 0.5  |
| (9)        | 0   | 0   | 0   | 0    | 0    | 0    | 0   | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| (10)       | 0.4 | 1.5 | 0   | 0.2  | 0    | 0    | 0   | 1    | 0.4  | 90.1 | 0.2  | 0    | 0    | 1.3  | 4.8  | 0    | 0    |
| (11)       | 0   | 0   | 0   | 0    | 0    | 0    | 0   | 0    | 0    | 0    | 100  | 0    | 0    | 0    | 0    | 0    | 0    |
| (12)       | 0   | 0   | 0   | 0    | 0    | 0    | 0   | 0    | 0    | 0    | 0    | 100  | 0    | 0    | 0    | 0    | 0    |
| (13)       | 0   | 0   | 0   | 0    | 0    | 0    | 0   | 0    | 0    | 0    | 0    | 0    | 100  | 0    | 0    | 0    | 0    |
| (14)       | 0   | 0   | 0   | 0    | 0    | 6.9  | 0   | 0    | 1.3  | 0    | 0    | 0    | 0    | 81.5 | 10.3 | 0    | 0    |
| (15)       | 0   | 0   | 8.6 | 43.2 | 1.4  | 12.9 | 0   | 0    | 12.9 | 7.2  | 0    | 0    | 0    | 10.1 | 3.6  | 0    | 0    |
| (16)       | 0   | 0   | 0   | 0    | 0    | 0    | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 100  | 0    |
| (17)       | 0   | 0   | 0.1 | 9.3  | 0.2  | 0.4  | 0   | 2.6  | 1.5  | 0.4  | 0    | 0.1  | 2.3  | 0.7  | 0.7  | 0.3  | 81.4 |

Genera listed by: (1) Acidaminococcus, (2) Akkermansia, (3) Alistipes, (4) Bacteroides, (5) Bifidobacterium, (6) Coprococcus, (7) Eggerthella, (8) Escherichia, (9) Eubacterium, (10) Faecalibacterium, (11) Megaspheara, (12) Parabacteroides, (13) Prevotella, (14) Roseburia, (15) Ruminococcus, (16) Shigella, (17) Streptococcus. At position  $(i, j)$  is shown the mean percentage (%) of reads actually belonging to the  $i$ -th genus that are assigned to the  $j$ -th genus. Estimations are done using data simulated independently of construction datasets

### 3.2 Results for CART98

The training dataset considered for *CART98* consist in fragment libraries composed by 4,540,000 reads equally distributed in 450 species forming 98 genera (considered as classes). The predictor was formed by 558 CART trees. The classification-importance of DNA patterns is shown in Fig. 4. Only 17 from 342 patterns show to be important. The GC ratio (*GC Per*) appears as the most important variable to discriminate between genera, being a key pattern to decreasing Gini impurity in first node separations. Other patterns that out-stand are *GC* and *TA* duples and codons such as *CGA* and *CGC* (and its reverse complements). Only the tetranucleotide *GATC* appears among the 10 most important patterns.

Our method reaches accuracy 47% in read-genus assignments for fragsim 454-simulated data sets. For most important microorganisms (Fig. 5), our method mostly recovers abundance of BEAR-generated data sets, coinciding with PhymmBL [3].

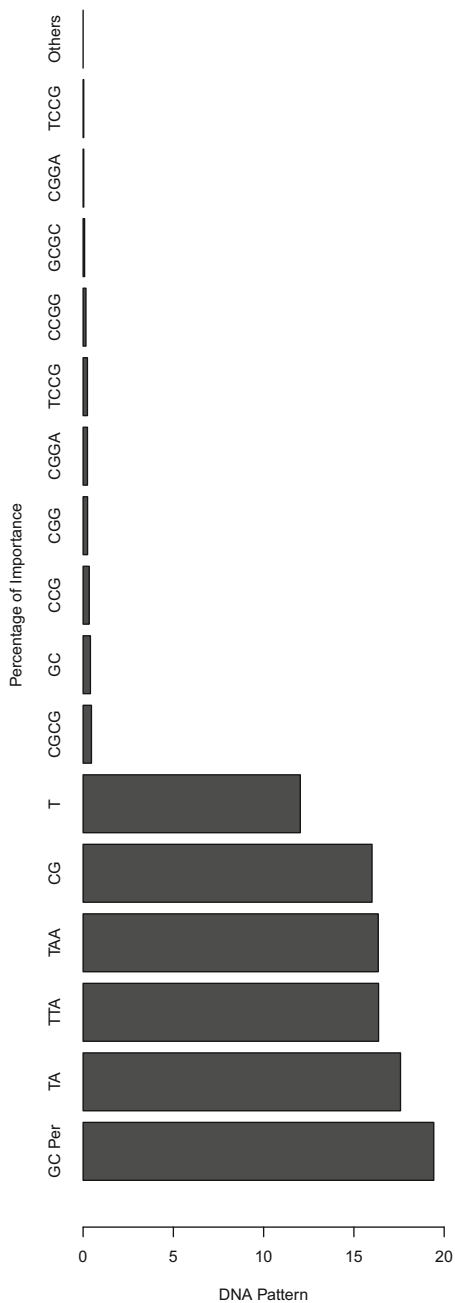
For *GD* reads from human gut microbiomes, we observed that, excepting *Bacteroides*, most significant abundance order differences reported by Yatsunenko et al. [30] (using MG-RAST) are also obtained with our method. In particular, results show that *Bifidobacterium* are more abundant in the sample from Malawi, and *Prevotella* are more abundant in the sample from Venezuela. We obtained the same behavior for *Paraprevotella*, which was not reported by Meyer et al. [17] and using PhymmBL [3] (Table 2).

## 4 Conclusions and Discussion

Multiple CART trees approach stands out as an accurate new way to predict the taxa present in a environment and their abundance. Restricted to 17 representative genera, our method has accuracy 77% in read assignments. Read assignment results strengthen with the hypothesis of *Shigella* belonging to *Escherichia* genus [14, 31].

GC ratio is recognized as the most important DNA pattern in genus assignments, also the importance of this pattern in taxa discrimination is widely known. Due to GC par are connected by three hydrogen bonds, GC ratio is higher in coding sequences than in non-coding zones [2]. Thus, bacteria showing more zones with high GC ratio have higher proportions of coding sequences. In fact, GC content is used to characterize some bacteria from phylum *Actinobacteria*, those present high GC ratio (70%) and *Plasmodium falciparum* with very low GC ratio (20%). In addition, two of our most important classification patterns (*GATC* and *CG*) are regions of DNA methylation in bacteria. The *GATC* tetranucleotide has some important physical properties [21]. Currently, approaches to identify *GATC* regions in the microbiome are being performed [15] and how different human associated bacteria interacts with their host in these regions [6] should provide diversity of bacterial functions between healthy and disease states [15].

**Fig. 4** Classification-importance (relative percentage) of DNA patterns for *CART98* predictor. Only 17 from 342 patterns are relevant. *GC Per* appears as the most important



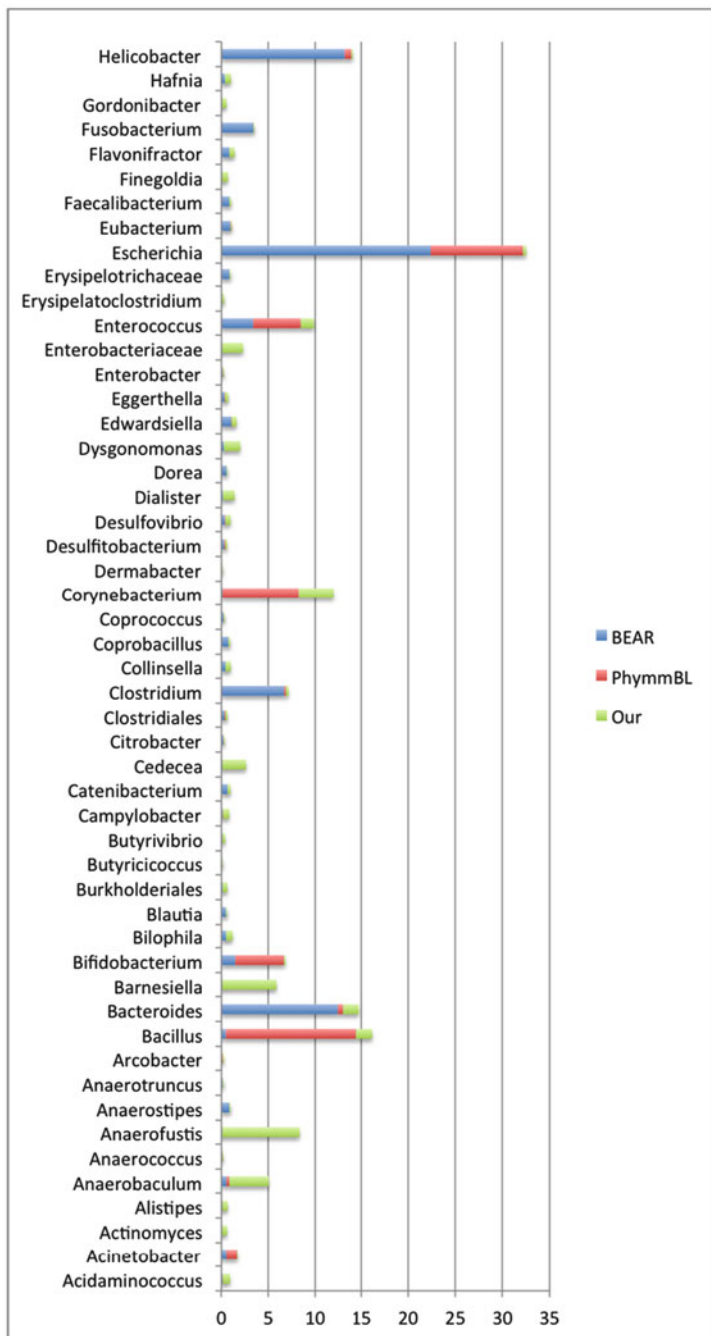


Fig. 5 Genus abundances (%) for first 50 genera with CART98 predictor. Reference values from BEAR-generated data sets, results using PhymmBL, and results using CART98 method

**Table 2** Comparison of abundance estimations

| Microorganism   | Venezuela |        |       | Malawi |        |        | USA    |        |       |
|-----------------|-----------|--------|-------|--------|--------|--------|--------|--------|-------|
|                 | [17]      | [3]    | Our   | [17]   | [3]    | CART98 | [17]   | [3]    | Our   |
| Bacteroides     | 19.8 %    | 11.8 % | 2.7 % | 26 %   | 24.5 % | 3.8 %  | 35.2 % | 15.7 % | 2.7 % |
| Bifidobacterium | 1.2 %     | 3.9 %  | 0.9 % | 35.2 % | 32.8 % | 6 %    | 0.3 %  | 2.5 %  | 0.5 % |
| Escherichia     | 0.6 %     | 4.6 %  | 0.3 % | 1.8 %  | 4.2 %  | 0.5 %  | 0 %    | 4.2 %  | 0.1 % |
| Prevotella      | 12 %      | 0 %    | 4.2 % | 5.5 %  | 0 %    | 2.4 %  | 1.7 %  | 0 %    | 1.7 % |
| Paraprevotella  | 0 %       | 0 %    | 5.2 % | 0 %    | 0 %    | 4 %    | 0 %    | 0 %    | 5.1 % |

Methods: MG-RAST [17], PhymmBL [3], and Multiple CART trees (CART98). Testing data set: geographically distributed children

For 98 genera, our predictor shows accuracy of 47 % on 454-Sim-generated DNA reads. For real samples, we recovered phenotype differences across gut microbiomes from Venezuela, Malawi and USA. Results are comparable with those obtained by MG-RAST and PhymmBL methods. For BEAR-generated data sets, in which genera are known, we obtained abundance predictions so accurate as those obtained using PhymmBL.

The alignment-free feature of the classifier allows to use it on non-assembled or unknown species. Actually we are working in a way to improve genus predictions using a variant of our method using hierarchical classification, thus reaching deeper taxonomies on target microorganisms. The next challenges we will face are incorporating other alignment-free distances [28], and focusing on target microorganisms-diseases to discriminate sick from healthy patients.

**Acknowledgements** This work was financed by ICBM U. Chile, Project CIRIC-INRIA Chile, Fondap Grant 15090007, Fondecyt 3130762, Basal Grant to the Center for Mathematical Modeling (grant no: PFB 03). We are also grateful to excellence PhD fellowships of U. Adolfo Ibañez. We acknowledge the support of the National Laboratory for High Performance Computing at the Center for Mathematical Modeling (PIA ECM-02- CONICYT).

## References

1. Abubucker, S., Segata, N, Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S.T., Meth, B., Schloss, P.D., Gevers, D., Mitreva, M., Huttenhower, C.: Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**(6), e1002358 (2012)
2. Bentley, S.D., Parkhill, J.: Comparative genomic structure of prokaryotes. *Ann. Rev. Genet.* **38**(1), 771–791 (2004)
3. Brady, A., Salzberg, S.L.: Phymm and phymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* **6**(9), 673–676 (2009)
4. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: *Classification and Regression Trees*, 1st edn. Chapman and Hall/CRC, New York (1984)
5. Chatterji, S., Yamazaki, I., Bai, Z., Eisen, J.A.: Compostbin: a dna composition-based algorithm for binning environmental shotgun reads. In: *Research in Computational Molecular Biology*, pp. 17–28. Springer, Heidelberg (2008)



6. Chernov, A.V., Reyes, L., Xu, Z., Gonzalez, B., Golovko, G., Peterson, S., Perucho, M., Fofanov, Y., Strongin, A.Y.: Mycoplasma CG- and GATC-specific DNA methyltransferases selectively and efficiently methylate the host genome and alter the epigenetic landscape in human cells. *Epigenetics* **10**(4), 303–318 (2015)
7. Dong, H., Chen, Y., Shen, Y., Wang, S., Zhao, G., Jin, W.: Artificial duplicate reads in sequencing data of 454 genome sequencer flx system. *Acta Biochim. Biophys. Sin.* **43**(6), 496–500 (2011)
8. Drezen, E., Rizk, G., Chikhi, R., Deltel, C., Lemaitre, C., Peterlongo, P., Lavenier, D.: Gatk: genome assembly & analysis tool box. *Bioinformatics* **30**(20), 2959–2961 (2014)
9. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(23), 3150–3152 (2012)
10. Gordon, A., Hannon, G.J.: Fastx-toolkit. FASTQ/A short-reads pre-processing tools. [http://hannonlab.cshl.edu/fastx\\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit) (2010, unpublished)
11. Hdar, C., Assar, R., Colombres, M., Aravena, A., Pavez, L., Gonzlez, M., Martnez, S., Inestrosa, N.C., Maass, A.: Genome-wide identification of new Wnt/-catenin target genes in the human genome using CART method. *BMC Genomics* **11**(1), 348 (2010)
12. Hugenholtz, P., Tyson, G.W.: Microbiology: metagenomics. *Nature* **455**(7212), 481–483 (2008)
13. Johnson, S., Trost, B., Long, J.R., Pittet, V., Kusalik, A.: A better sequence-read simulator program for metagenomics. *BMC Bioinf.* **15**(Suppl 9), S14 (2014)
14. Lan, R., Reeves, P.R.: Escherichia coli in disguise: molecular origins of shigella. *Microbes Infect.* **4**(11), 1125–1132 (2002)
15. Leonard, M.T., Davis-Richardson, A.G., Ardisson, A.N., Kempainen, K.M., Drew, J.C., Ilonen, J., Knip, M., Simell, O., Toppari, J., Veijola, R. et al.: The methylome of the gut microbiome: disparate dam methylation patterns in intestinal bacteroides dorei. *Front. Microbiol.* **5**, 361 (2014)
16. Lysholm, F., Andersson, B., Persson, B.: An efficient simulator of 454 data using configurable statistical models. *BMC Res. Notes* **4**(1), 449 (2011)
17. Meyer, F., Paarmann, D., D’Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A. et al.: The metagenomics rast server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinf.* **9**(1), 386 (2008)
18. Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R.D.E., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., Rampp, M., Miller, W., Schuster, S.C.: Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**(5759), 392–394 (2006)
19. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Paslier, D.L., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dor, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D., Wang, J.: A human gut microbial gene catalog established by metagenomic sequencing. *Nature* **464**(7285), 59–65 (2010)
20. Richter, D.C., Ott, F., Auch, A.F., Schmid, R., Huson, D.H.: MetaSimA sequencing simulator for genomics and metagenomics. *PLoS ONE* **3**(10), e3373 (2008)
21. Riva, A., Delorme, M.-O., Chevalier, T., Guilhot, N., Hénaut, C., Hénaut, A.: The difficult interpretation of transcriptome data: the case of the gatc regulatory network. *Comput. Biol. Chem.* **28**(2), 109–118 (2004)
22. Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J., Chisholm, S.W.: Unlocking short read sequencing for metagenomics. *PLoS ONE* **5**(7), e11840 (2010)

23. Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S., Huttenhower, C.: Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**(1), 666 (2013)
24. Tringe, S.G., Rubin, E.M.: Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6**(11), 805–814 (2005)
25. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I.: The human microbiome project. *Nature* **449**(7164), 804–810 (2007)
26. Valenzuela, M., Bravo, D., Canales, J., Sanhueza, C., Daz, N., Almarza, O., Toledo, H., Quest, A.F.G.: Helicobacter pylori-induced loss of survivin and gastric cell viability is attributable to secreted bacterial Gamma-glutamyl transpeptidase activity. *J. Infect. Dis.* **208**(7), jit286 (2013)
27. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., Smith, H.O.: Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**(5667), 66–74 (2004)
28. Weitschek, E., Santoni, D., Fiscon, G., De Cola, M.C., Bertolazzi, P., Felici, G.: Next generation sequencing reads comparison with an alignment-free distance. *BMC Research Notes* **7**, 869 (2014)
29. Wu, Y.-W., Ye, Y.: A novel abundance-based algorithm for binning metagenomic sequences using 1-tuples. *J. Comput. Biol.* **18**(3), 523–534 (2011)
30. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., Heath, A.C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J.G., Lozupone, C.A., Lauber, C., Clemente, J.C., Knights, D., Knight, R., Gordon, J.I.: Human gut microbiome viewed across age and geography. *Nature* **486**(7402), 222–227 (2012)
31. Zuo, G., Xu, Z., Hao, B.: Shigella strains are not clones of Escherichia coli but sister species in the genus Escherichia. *Genomics Proteomics Bioinformatics* **11**(1), 61–65 (2013)

# A Statistical Approach to Infer 3D Chromatin Structure

Claudia Caudai, Emanuele Salerno, Monica Zoppè, and Anna Tonazzini

**Abstract** We propose a new algorithm to estimate the 3D configuration of a chromatin chain from the contact frequency data provided by HI-C experiments. Since the data originate from a population of cells, we rather aim at obtaining a set of structures that are compatible with both the data and our prior knowledge. Our method overcomes some drawbacks presented by other state-of-the-art methods, including the problems related to the translation of contact frequencies into Euclidean distances. Indeed, such a translation always produces a geometrically inconsistent distance set. Our multiscale chromatin model and our probabilistic solution approach allow us to partition the problem, thus speeding up the solution, to include suitable constraints, and to get multiple feasible structures. Moreover, the density function we use to sample the solution space does not require any translation from contact frequencies into distances.

**Keywords** 3D chromatin structure • Chromosome Conformation Capture • Quaternions

## 1 Introduction

The nuclear DNA is arranged in a 30 nm fiber called chromatin, and in human cells has a length of about 2 m in total, folded in 46 chromosomes. Its spatial organization ensures the continuous accessibility of DNA to translation, replication, regulation and repair machinery. Understanding how DNA is organized will help to discover its functional features and the epigenetic mechanisms involved. A first important step in describing the organization of DNA within the nucleus was done with the experiments of fluorescence in situ hybridization (FISH) [1], a technique used to

---

C. Caudai (✉) • E. Salerno • A. Tonazzini  
National Research Council of Italy, Institute of Information Science and Technologies,  
Via Moruzzi 1, 56124 Pisa, Italy  
e-mail: [claudia.caudai@isti.cnr.it](mailto:claudia.caudai@isti.cnr.it)

M. Zoppè  
National Research Council of Italy, Institute of Clinical Physiology,  
Via Moruzzi 1, 56124 Pisa, Italy

detect and localize specific DNA sequences. Recently, high resolution techniques have been developed, called Chromosome Conformation Capture (3C) [4], which provide contact frequencies between pairs of DNA fragments in the whole genome. The latest such technique, called HI-C [17], has a very high genomic resolution, reaching a few kbp, depending on the enzyme used in the procedure.

From HI-C information, it is possible to formulate hypotheses about the three-dimensional chromatin configurations. Many approaches have been proposed to address this problem. They can be divided into three main categories, each offering specific advantages and criticalities: constrained optimization, Bayesian inference, and polymer models. The new reconstruction method we propose in this chapter was conceived to exploit the benefits of the state-of-the-art methods while avoiding some of their drawbacks.

All the constrained optimization strategies proposed to introduce a model for the solution, a set of constraints, and a cost function to be optimized against the available data. As mentioned, the 3C data available are contact frequencies evaluated over the whole population of cells in the experiment, typically many millions. The first attempts to translate these data into geometrical information assume that the chromatin configurations are not very different throughout the population, and that pairs of fragments often found in contact are closer than pairs with low contact frequencies. On this basis, most of the existing methods propose some formula to translate the contact frequencies into Euclidean distances, to be fitted by the reconstructed structures. Duan et al. [7] propose a three-dimensional model of yeast genome, in which chromatin is modeled as a bead chain, with partially impenetrable beads, forced to stay in a spherical nucleus of  $1 \mu\text{m}$ . The objective function to be minimized exploits an inverse proportionality relationship between contacts and distances. The same deterministic law is also adopted by Fraser et al. [8] and Dekker et al. [4]. In Sect. 2, we show how this translation leads to severe geometric inconsistencies. Baù and Marti-Renom [2] translate the contact frequencies into harmonic forces, calibrating the distances between beads. The constrained optimization approach has the advantage of introducing geometric and biophysical constraints into the model, but has two big disadvantages: the high dimensionality of the systems and the absence of confidence intervals to evaluate the uncertainty of the solutions obtained.

The data are affected by errors and biases and, as mentioned, derive from experiments on millions of cells. This makes necessary the adoption of a probabilistic approach to sample the space of the feasible solutions. The first probabilistic approach has been published by Rousseau et al. [16], who use a Markov Chain Monte Carlo sampling on a Gaussian likelihood, built through an inverse-quadratic law between contacts and distances (MCMC5C). Hu et al. [9] use the same relationship, proposing an algorithm called BACH (Bayesian 3D Constructor for HI-C data), to build consensus 3D structures. The novelty of the cited Bayesian approaches is the possibility to introduce biases into the data model (as in BACH). Another important advantage is the possibility of sampling the solution space: this aspect is essential, since it is more meaningful to search for sets of possible solutions rather than a single consensus. The major drawbacks of BACH are its computational

complexity, due to the large number of parameters to be estimated, and the absence of suitable topological constraints.

Another interesting approach is the integration of polymer physics into the 3D chromatin structure model. This has the advantage of not requiring the translation from frequencies into distances, and permits the adoption of iterative adaptive methods. Meluzzi and Arya [14] propose a coarse-grained bead-chain polymer model approximating the physical behavior of a 30 nm chromatin fiber; the system evolves adjusting iteratively the model parameters, until a match with contact frequency data is reached. This approach is highly reliable but very expensive computationally. For this reason, it cannot yet be applied to experimental data: a validation has only been performed against reference data sets obtained from simulations of systems with up to 45 beads.

An analysis of the different solutions mentioned above reveals a number of drawbacks that must be overcome to obtain more reliable results. Our main point is the questionable adequacy of the translation of contact frequencies into Euclidean distances. In Sect. 2, we show that this strategy produces a set of distances often severely incompatible with the Euclidean geometry. Then, in Sect. 3, we briefly describe our solution model, our cost function, which does not include an explicit contact-to-distance relationship, and the stochastic algorithm we used to sample the solution space. Section 4 concludes the chapter, with some reference to our first experimental results.

## 2 Geometrical Consistency of the Frequency-Distance Translation

The problem of the geometrical inconsistencies derived from translating contact frequencies into Euclidean distances has been overlooked by almost all groups that have worked with contact frequency data. An exception is the work of Duggal et al. [6], who propose a filtering technique to select subsets of interactions obeying to metric constraints. This method is very interesting, but has a high computational cost.

It is important to exert some caution with the extraction of topological information (measurements, distances) from interaction data, because contacts are discrete events (sums of dichotomous events) with causal and random components, whereas spatial distances are continuous quantities forced to undergo precise geometric laws. It is necessary to check whether the distances meet the basic geometrical consistency conditions, *e.g.* the triangular inequality. The non-violation of these conditions is a necessary but not sufficient condition for geometric consistency. If geometric consistency conditions are severely violated, the set of distances cannot be used as a target to achieve sensible geometric conformations of chromatin. However, the fact that these inequalities are not violated, or are violated slightly, does not ensure the geometrical consistency of the system. For example, if we have a set of equal

distances (e.g. all equal to 1), the triangular inequalities would never be violated, but no structure in the 3D Euclidean space can show such a distance set, unless it is made of no more than 4 points.

Let us consider a chromatin chain made of  $N$  elements, and any subsequence  $S$  of it, with  $M$  elements, identified by the index set  $I = \{1, 2, \dots, M\}$ . Let us now consider a partition  $P$  of  $S$ , that is, any set of  $L \leq M$  consecutive segments that sum up to  $S$ , identified by the set of index pairs  $K = \{(1, k_2), (k_2, k_3), \dots, (k_L, M)\}$ , with  $1 < k_2 < k_3 < \dots < k_L < M$ . A necessary condition for the Euclidean distances between all the possible pairs in  $S$  to be consistent with the 3D Euclidean geometry is that, for any possible  $K$ :

$$d_{1,M} \leq \sum_{(i,j) \in K} d_{i,j} \quad (1)$$

where  $d_{i,j}$  is the distance between the  $i$ -th and the  $j$ -th elements of  $S$ .

In our preliminary study, we considered two sets of experimental data made available in the literature, from the entire human genome in GM06690 [13] and GSE18199 cells [18], both with genomic resolution of 1 Mbp. Then, for both data sets, for any possible subsequence of all the chromosomes, and for 13 different frequency-to-distance relationships, we evaluated the number and the extent of the violations to condition (1). The results of this analysis are summarized in Table 1, whereas the contributions of each individual chromosome are plotted in Figs. 1 and 2. The number of violations and their weights rapidly decrease by applying the laws  $1/\sqrt[n]{x}$ , with  $n \in \{1, 2, \dots, 5\}$ . This does not mean that these laws are suitable to build a good target function, since they actually tend to produce a set of nearly equal distances, which normally lead to impossible structures.

Also considered from another viewpoint, the inversion process from contact frequencies into distances presents a heuristic gap, because the measured contact frequencies do not depend exclusively on geometric properties, but also on other factors, such as the presence of topological barriers, energy conditions, and random events. In summary, we think that assuming that pairs with many contacts are likely to be close to each other can be justified, whereas pairs with a few contacts are not warranted to be distant from each other. Our analysis demonstrates that experimental frequency data very often lead to distances that are more or less severely incompatible with real configurations in the 3D Euclidean space. For this reason, such distances cannot be used as rigid targets for structure estimation. Indeed, any consistent distance system identifies a precise structure in the Euclidean space (for a general introduction to Distance Geometry, see for example [12]), but 3C data are produced by many distinct cells, so it is very unlikely that a single relationship is able to generate geometrically consistent distances from the contact matrix.

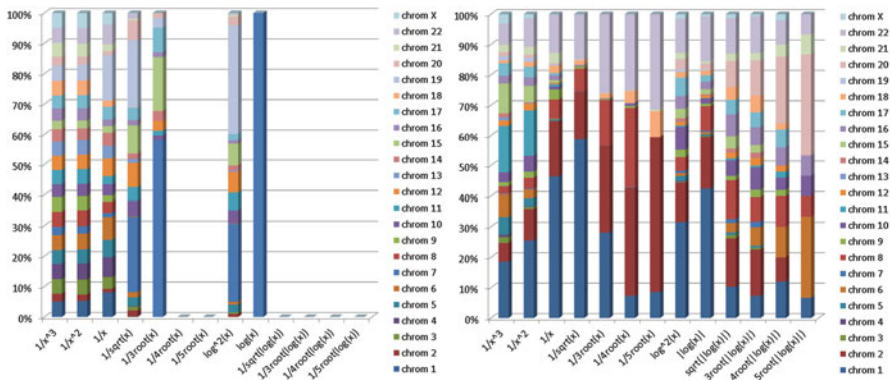
**Table 1** Frequency-distance conversion laws for dataset available in [13, 18]

| Lieberman-Aiden et al. [13]                     |                                   |                              | Yaffe and Tanay [18]                      |                                   |                              |
|---|-----------------------------------|------------------------------|---|-----------------------------------|------------------------------|
| Transformation laws                             | Number of violations <sup>a</sup> | Average percentage violation | Transformation laws <sup>b</sup>          | Number of violations <sup>a</sup> | Average percentage violation |
| $x \rightarrow d = \frac{1}{x^3}$               | 28003.8                           | 3458.5                       | $x \rightarrow d = \frac{1}{x^3}$         | 2464.6                            | $4 \cdot 10^8$               |
| $x \rightarrow d = \frac{1}{x^2}$ [9, 16]       | 26502.8                           | 424                          | $x \rightarrow d = \frac{1}{x^2}$         | 1439.6                            | $6 \cdot 10^5$               |
| $x \rightarrow d = \frac{1}{x}$ [4, 7, 8]       | 8954,1                            | 42.4                         | $x \rightarrow d = \frac{1}{x}$           | 766.7                             | 1501.8                       |
| $x \rightarrow d = \frac{1}{\sqrt{x}}$          | 72,3                              | 8.6                          | $x \rightarrow d = \frac{1}{\sqrt{x}}$    | 604.9                             | 99.4                         |
| $x \rightarrow d = \frac{1}{\sqrt[3]{x}}$       | 2.7                               | 1.5                          | $x \rightarrow d = \frac{1}{\sqrt[3]{x}}$ | 287.9                             | 32                           |
| $x \rightarrow d = \frac{1}{\sqrt[4]{x}}$       | 0                                 | 0                            | $x \rightarrow d = \frac{1}{\sqrt[4]{x}}$ | 55.9                              | 16.3                         |
| $x \rightarrow d = \frac{1}{\sqrt[5]{x}}$       | 0                                 | 0                            | $x \rightarrow d = \frac{1}{\sqrt[5]{x}}$ | 9.1                               | 4.6                          |
| $x \rightarrow d = \frac{1}{\log^2(x)}$         | 65.4                              | 8.7                          | $x \rightarrow d = \log^2(x)$             | 1143                              | 1095.5                       |
| $x \rightarrow d = \frac{1}{\log(x)}$           | 0.3                               | 0.3                          | $x \rightarrow d =  \log(x) $             | 566.7                             | 72.8                         |
| $x \rightarrow d = \frac{1}{\sqrt{\log(x)}}$    | 0                                 | 0                            | $x \rightarrow d = \sqrt{ \log(x) }$      | 34                                | 28                           |
| $x \rightarrow d = \frac{1}{\sqrt[3]{\log(x)}}$ | 0                                 | 0                            | $x \rightarrow d = \sqrt[3]{ \log(x) }$   | 7.1                               | 18.5                         |
| $x \rightarrow d = \frac{1}{\sqrt[4]{\log(x)}}$ | 0                                 | 0                            | $x \rightarrow d = \sqrt[4]{ \log(x) }$   | 2.2                               | 8.5                          |
| $x \rightarrow d = \frac{1}{\sqrt[5]{\log(x)}}$ | 0                                 | 0                            | $x \rightarrow d = \sqrt[5]{ \log(x) }$   | 0.7                               | 5.2                          |

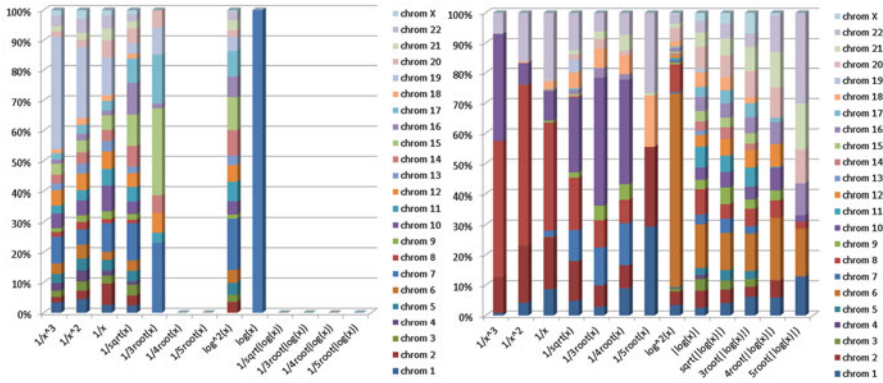
In the formulas  $x$  represents the contact frequency and  $d$  the Euclidean distance

<sup>a</sup> Averaged on chromosomes

<sup>b</sup> Contact frequency values normalized to 1



**Fig. 1** Percent contributions of the different chromosomes to the total number of geometric violations, for the 13 transformation laws considered. *Left*: data from [13]. *Right*: data from [18]. For each column, the contributions of the chromosomes have always the same order: from chromosome 1 at the bottom to chromosomes 22 and X at the top of the column



**Fig. 2** Percent contributions of the different chromosomes to the average extent of the geometric violations, for the 13 transformation laws considered. *Left*: data from [13]. *Right*: data from [18]. For *each column*, the contributions of the chromosomes have always the same order: from chromosome 1 at the *bottom* to chromosomes 22 and X at the *top* of the *column*

### 3 Our Approach

Each of the studies that proposed methods for 3D chromatin reconstruction from contact data presents problems and advantages, summarized in Table 2. As a contribution to the field, we propose a new algorithm that includes a list of desirable features:

1. Possibility to enforce geometrical constraints on the solutions.
2. Computational efficiency, including partitioning and parallel processing capabilities.
3. No deterministic translation from contact frequencies to distances.
4. Possibility to get multiple configurations compatible with the data.

To obtain features (1) and (2), we rely on our chromatin model. If we model the chromatin fiber as a bead chain, we can first impose that it must remain connected, that is, that the beads must maintain their genomic locations, and then introduce constraints on the distances between adjacent beads and on the angles formed by any two consecutive bead pairs. This amounts to constrain the length of any subchain and its maximum curvature. Of course, the appropriate values for these constraints must be decided on the basis of the relevant biological knowledge. Partitioning the problem can enable us to speed up the estimation process. We reach this goal by taking into account the existence of chromatin segments, called *topological domains* [5], that have no important interactions with other genomic regions, and exploiting the multiscale capabilities of our chromatin model. The structure of each topological domain can be estimated from the data coming exclusively from the fragments belonging to it. The resulting structure is then considered as a bead in a lower resolution chain, whose contact frequencies are



**Table 2** Chart of problems and advantages in the previous state of the art

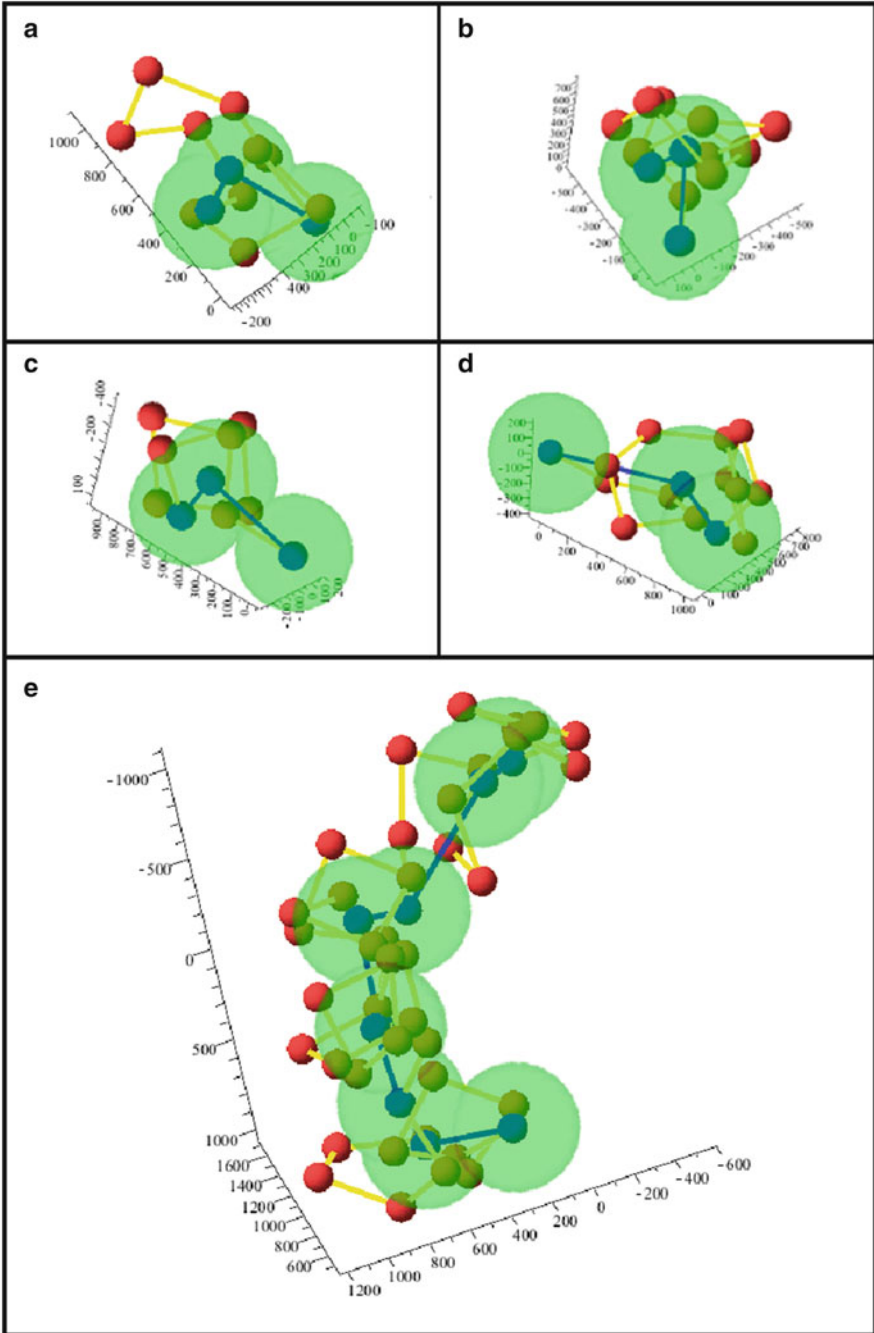
|   | Problems  | Advantages   |
|---|---|--|
| <b>Constrained Optimization</b><br>Dekker et al. [4]<br>Fraser et al. [8]<br>Duan et al. [7]<br>Baù and Marti-Renom [2] | Very high dimensionality<br>No confidence intervals can be computed to measure the uncertainty of the structure obtained  | First attempt of conversion of a set of noisy contact frequencies measurements into more interpretable data<br>Introduction of constraints based on the structure of the chromatin fiber |
| <b>Bayesian Inference</b><br>Russeau et al. [16] (MCMC5C)<br>Hu et al. [9] (BACH)                                       | Any evaluation of structural variations of chromatin at different resolution scales<br>No geometrical constraints<br>Geometrical inconsistencies given by translation of contact frequencies into distances | Bayesian approach to sample the whole space of solutions<br>Introduction of systematic biases into the data model (BACH)   |
| <b>Polymer Models</b><br>Nagano et al. [15]<br>Meluzzi and Arya [14]  | Complexity of the system  | Conversion from frequencies into distances not required<br>Integration of polymer physics into the 3D chromatin structure model  |

evaluated along with possible higher-level isolated domains. The structures of these new topological domains are reconstructed by the same strategy described above. This process can continue recursively, until a data set with a single domain is found. The full-resolution structure is then reconstructed by substituting, recursively, the lower-resolution beads with the subchains reconstructed at finer resolutions. Except for the finest resolution available, our beads are not spheres, but are equipped with the macroscopic properties of the subchains they represent, each being a non-deformable triplet identified by the centroid of the related subchain and its endpoints. Figure 3 depicts an example of this model for two consecutive scales.

Requirement (3) is reached through our cost function. We first observe that, as mentioned in Sect. 2, fragment pairs characterized by high contact frequencies can reliably be considered in close proximity, but the converse does not need to be true: pairs with low contact frequencies do not need to be far apart. We thus avoid to consider the lowest frequencies in our cost function, which, anyway, can sufficiently determine the problem by exploiting the geometrical constraints. The resulting expression is:

$$\Phi(\mathcal{C}) = \sum_{i,j \in \mathcal{L}} n_{i,j} \cdot d_{i,j} \quad (2)$$

where  $\mathcal{C}$  is the configuration of the subchain being estimated,  $\mathcal{L}$  is the set of bead pairs that are likely to be close to each other, and  $n_{i,j}$  is the contact frequency characterizing the  $(i,j)$ -th pair. Thus, no target distance is included in the formula: the contact frequency data are directly used to weight the contributions of the



**Fig. 3** (a)–(d) Consecutive fragments of the chromatin fiber, represented as bead sequences (*red* balls linked by *yellow* segments), and as centroid-endpoints triples (*blue* balls linked by *blue* segments). The larger spheres represent the assumed sizes for the beads at the lower resolution. (e) Lower-resolution chain composed by the fragments in (a)–(d)

individual pairs in the summation. It is apparent that an unconstrained optimization of this cost function would find global minima in each configuration with  $d_{ij} = 0$  for all  $(i, j) \in \mathcal{L}$ . The constraints, however, make these solutions unfeasible.

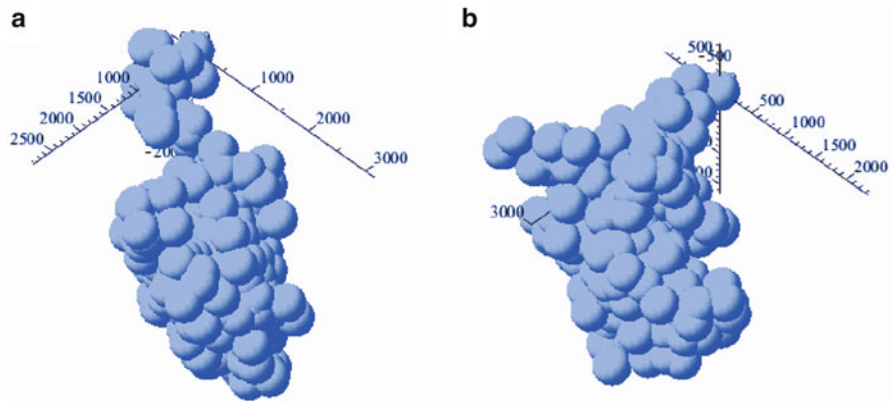
Finally, requirement (4) is satisfied by our estimation algorithm. Although the configurations that are not compatible with the constraints are not feasible solutions, it is expected that the cost function reaches minimum values for many different feasible configurations. To be able to sample the solution space, we treat the objective function as a negative log-density, and use a Monte Carlo approach to find high-probability configurations. In practice, we use a classical simulated annealing procedure [11], where the model updates are proposed through quaternion operators [10]. This choice allows us to maintain automatically the coherence of the reconstructed chain at each update, thus avoiding to check the fit to most of the constraints before continuing with the iteration. Indeed, the compatibility of the current solution with the constraints must only be checked against possible spatial interferences between pairs of beads. Since so many configurations fit well the data and the constraints, different runs of this stochastic procedure will produce different highly reliable results, whose structures should reproduce the variety of the configurations assumed by the chromatin chain in the experimental cell population. Our multiscale approach can also be exploited to generate different configurations of the subchains at any resolution, and then combine them to produce, recursively, different configurations of the overall chain.

## 4 Conclusions

In this chapter, we propose a new approach for the estimation of chromatin configurations starting from HI-C contact frequency data. The main characteristics of our approach are:

- The data-fit function does not require the translation of frequencies into Euclidean distances.
- The multiscale bead-chain model can be equipped with biophysical constraints; any prior information available must be translated into geometrical constraints.
- The probabilistic procedure samples the solution space so that multiple configurations compatible with both the data and the constraints can be found.
- The model evolution during the iterations is obtained through quaternion operators.

Thanks to these features, our procedure avoids some of the drawbacks in the algorithms proposed so far in the literature. Also, our algorithm is conceptually simple, and amenable to be speeded up by exploiting several levels of parallelism. As a proof of principle, we have performed some tests on real HI-C data from human cells [3]. In these tests, we obtained a number of different structures characterized by similar values of the cost function but showing a few distinct spatial behaviors



**Fig. 4** Two typical configurations resulting from our experiments (measurements in nm): (a) more expanded, (b) more compact

(two examples are shown in Fig. 4, from data related to the long arm of the human chromosome 1 [13]). The macroscopic appearance of these structures is compatible with the expected shape of a portion of chromosome.

In conclusion, we have generated an algorithm that can substantially contribute to the elucidation of chromosomal structure, by producing families of structures compatible with biological information. Our procedure is also innovative in the use of quaternions to evolve the model during the estimation process.

**Acknowledgements** This work has been funded by the Italian Ministry of Education, University and Research, and by the National Research Council of Italy, Flagship Project InterOmics, PB.P05. The authors are indebted to Luigi Bedini and Aurora Savino for helpful discussions.

## References

1. Amann, R., Fuchs, B.M.: Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nat. Rev. Microbiol.* **6**, 339–348 (2008)
2. Baù, D., Marti-Renom, M.A.: Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosom. Res.* **19**, 25–35 (2011)
3. Caudai, C., et al.: Reconstructing 3D chromatin structure from chromosome conformation capture data, InterOmics Flagship Project, Report cnr.isti/2015-PR-001, National Research Council of Italy - ISTI, Pisa (2014)
4. Dekker, J., et al.: Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002)
5. Dixon, J.R., et al.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012)
6. Duggal, G., et al.: Resolving spatial inconsistencies in chromosome conformation measurements. *Algorithms Mol. Biol.* **8**, 8 (2013)
7. Duan, Z., et al.: A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010)
8. Fraser, J., et al.: Chromatin conformation signatures of cellular differentiation. *Genome Biol.* **10**, R37 (2009)

9. Hu, M., et al.: Bayesian inference of Spatial organizations of chromosomes. *PLOS Comput. Biol.* **9**, 1002–893 (2013)
10. Karney, C.F.: Quaternions in molecular modeling. *J. Mol. Graph. Model.* **25**, 595–604 (2007)
11. Kirkpatrick, S., et al.: Optimization by simulated annealing. *Science* **229**, 671–680 (1983)
12. Liberti, L., et al.: Euclidean distance geometry and applications. *SIAM Rev.* **56**, 3–69 (2014)
13. Lieberman-Aiden, E., et al.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009)
14. Meluzzi, D., Arya, G.: Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acid Res.* **41**, 63–75 (2013)
15. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., Fraser, P.: Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013)
16. Rousseau, M., et al.: Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinf.* **12**, 414–429 (2011)
17. van Berkum, N.L., et al.: Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39**, 1869–1875 (2010)
18. Yaffe, E., Tanay, A.: Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1067 (2011)

# Basic Exploratory Proteins Analysis with Statistical Methods Applied on Structural Features

Eugenio Del Prete, Serena Dotolo, Anna Marabotti, and Angelo Facchiano

**Abstract** Exploratory Data Analysis (EDA) is an approach for summarizing and visualizing the important characteristics of a data set, in order to make a prearranged data screening and display multivariate data in a graphical way, to render them more comprehensible. Moreover, it reveals hidden aspects within the simple evaluations. In particular, EDA is suitable for datasets with comparable variables, as structural-geometrical protein features. In this work, we analyzed some proteins belonging to ten different architectural families. After retrieval, feature selection and normalization stages, the dataset has been processed by means of simple correlation, partial correlation and principal component analysis (PCA), highlighting family-independent or family-specific relationships, and possible outliers for the dataset itself. The results can be useful to connect these features to functional protein properties.

**Keywords** Correlation • Exploratory data analysis • Global features • Principal component analysis • Protein structure

## 1 Background

Exploratory Data Analysis (EDA) is the process of looking through data to get a basic idea of their structures and attributes, often with visualizations. EDA is a graphical-statistical approach, almost a philosophy of research, applied to data in order to make some aspects clearer and answer some questions about them. It is like a magnifying glass that helps in:

---

E. Del Prete (✉) • S. Dotolo • A. Facchiano  
Institute of Food Science, National Research Council, Via Roma 64, 83100 Avellino, Italy  
e-mail: [eugenio.delprete@isa.cnr.it](mailto:eugenio.delprete@isa.cnr.it); [serenadotolo@hotmail.it](mailto:serenadotolo@hotmail.it); [angelo.facchiano@isa.cnr.it](mailto:angelo.facchiano@isa.cnr.it)

A. Marabotti  
Institute of Food Science, National Research Council, Via Roma 64, 83100 Avellino, Italy  
Department of Chemistry and Biology, University of Salerno, Via Giovanni Paolo II 132,  
84084 Fisciano, SA, Italy  
e-mail: [amarabotti@unisa.it](mailto:amarabotti@unisa.it)

- leading towards a right interpretation of data;
- showing and summarizing data in a clear way;
- finding underlying relationships among observations and, main thing, among variables.

It can be univariate or multivariate and can use graphical or not graphical methods. A historical explanation can be found in [1]. Main point is that EDA is essential in understanding data, because it can reinforce or undermine *a priori* knowledge about observations and prepare data for the following inference step.

In the analysis of large data sets, an inevitable phase that must anticipate the statistical analysis concerns getting and cleaning data. Data can be obtained from a variety of sources: downloaded from online repositories, streamed on-demand from online sources, automatically generated by physical apparatus interfaced to a computer, generated by a computer software, manually entered in a spreadsheet or text file. Data origin, management and storage are other issues related to the getting part of the data analysis. Raw data retrieved are probably not in a convenient format, because of semantic errors, missing entries, inconsistent formatting. Thus, it is recommended to make a control on all variables and, if necessary, integrate new ones from different sources that are coherent with the previous ones, in order to create a tidy final dataset [2].

Investigations on protein structure and function represent a field of research in which experimental techniques as well as computational methods are widely applied [3–6]. Nevertheless, many aspects are still unsolved, in particular concerning the relationships between structure and function of proteins. While successful methods have been developed to “predict” the complex three-dimensional structure of a protein from a simple structural information as the amino acid sequence, and are largely applied in literature and by our research group [7–9], it is less investigated the deep nature of the structural features and their relationships with protein function. In other words, evolution may modify the amino acid sequence of an ancestral protein at a large extent among living species, thus affecting the lower level of structural organization of a protein family member. This has low impact on the three-dimensional structure, i.e. the higher level of structural organization, so that the protein family maintains its specific biochemical function over the species. On the other hand, a single amino acid substitution within a protein can strongly affect structure and function, as in human pathologies due to genetic diseases [10–12]. However, it is still unclear in detail how the modification of amino acid sequence is softened or emphasized when it is reflected at the functional level. In this context, we are interested to exploit graphical and mathematical methods, poorly used in protein science up to now, in order to explore protein structure and function relationships from a new point of view.

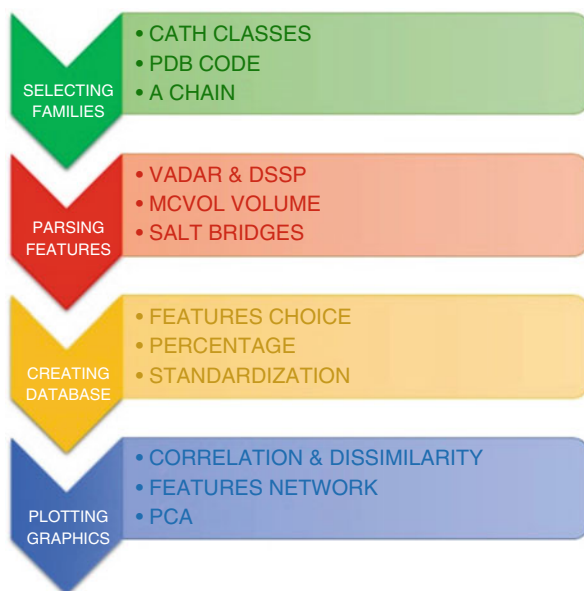
In this study, multivariate graphical methods have been used, because of tabular (observations—variables) data type. Data are composed of protein families chosen for their functional similarity; it is interesting to examine protein structure and analyze conformational features within a family and among the families, in order to find relationships that could be related to functional properties. Ten protein families have been chosen, depending on CATH different architectural classification [13].

## 2 Materials and Methods

### 2.1 Analysis Workflow

The workflow (Fig. 1) consists of four steps, of which the first three concern getting and cleaning data, whereas the fourth step is the real EDA. More in details:

- **Step 1.** 153 crystallographic structures (Table 1) have been retrieved from RCSB PDB [14]. The structures have been selected to represent ten structural protein families, and different architectural classes in CATH. Rules applied to select structures to be analyzed are: families for which a similar number of structures is available (i.e., in the range 13–19); within each family, only one chain per protein (in homo-multimeric proteins, A chain), and structures which differ for less than 50 residues in length;
- **Step 2.** Different online and local tools have been used to extract protein structural properties from PDB files: *Vadar* [15], for secondary structure (also confirmed with DSSP [16]), hydrogens bonds, accessible surface areas, torsion angles, packing defects, charged residues numbers, free energy of folding; *McVol* [17] for volumes, with a mean difference of 4 % from that extracted from Vadar, but using a more robust algorithm; an in-house-developed *R-script* for automatic search of salt bridges conditions [18]. Parsing of Vadar results has been performed by means of regular expression commands in *R*;



**Fig. 1** Analysis graphical workflow. There are four simply identifiable steps: steps 1–3 composed the getting and cleaning data part, step 4 is the Exploratory Data Analysis



**Table 1** Protein families and PDB structures

| CATH code  | PDB files  |
|--|--|
| Beta-Lactamase (BLA)<br>3.30.450, $\alpha/\beta$ 2-layer sandwich                            | ID1J, 1EW0, 1F2K, 1N9L, 1P0Z, 2V9A, 2VK3, 2VV6, 2ZOH, 3BW6, 3BY8, 3CI6, 3CWF, 3EEH                               |
| Cathepsin B (CTS)<br>3.90.70, $\alpha/\beta$ complex   | 1AEC, 1B5F, 1S4V, 2B1M, 2BDZ, 2DC6, 2P7U, 2WBF, 3A18, 2BCN, 3CH2, 3LXS, 3P5U                                     |
| Ferritin (FTL)<br>1.20.1260, $\alpha$ up-down bundle   | 1J14, 1QGH, 1R03, 1S2Z, 1TJO, 2FKZ, 2XJM, 2YW6, 3AK8, 3E1J, 3KA3, 3MPS, 3R2H, 3RAV                               |
| Glycosyltransferase (GTF)<br>1.50.10, $\alpha$ - $\alpha$ barrel                             | 1GAH, 1HVX, 1KRF, 1KS8, 1NXC, 1R76, 1X9D, 2NVP, 2P0V, 2XFG, 2ZZR, 3P2C, 3QRY                                     |
| Hemoglobin (HGB)<br>1.10.490, $\alpha$ orthogonal bundle                                     | 1CG5, 1FLP, 1GCW, 1HLM, 1RQA, 1UVX, 2C0K, 2QSP, 2VYW, 3BJ1, 3NG6, 3QQR, 3WCT, 4IRO, 4NK1                         |
| Lipocalin 2 (LCN)<br>2.40.128, $\beta$ - $\beta$ barrel                                      | 1AQB, 1BEB, 1CBI, 1CBS, 1GGL, 1GM6, 1IIU, 1JYD, 1KQW, 1KT6, 1LPI, 1OPB, 1QWD, 2CBR, 2NND, 2RCQ, 2XST, 3S26, 4TLJ |
| Lysozyme (LYS)<br>1.10.530, $\alpha$ orthogonal bundle                                       | 1BB6, 1FKV, 1GD6, 1GHL, 1HHL, 1IIZ, 1JUG, 1QQY, 1REX, 1TEW, 2EQL, 2GV0, 2IHL, 2ZF, 3QY4                          |
| Proliferating Cell Nuclear Antigen (PCNA)<br>3.70.10, $\alpha/\beta$ box                     | 1AXC, 1B77, 1CZD, 1DML, 1T6L, 1UD9, 1UL1, 1HII, 1IUX, 2OD8, 3HI8, 3LX1, 3P83, 3P91, 4CS5                         |
| Purine Nucleoside Phosphorylase (PNP)<br>3.40.50, $\alpha/\beta$ 3-layer sandwich            | 1A90, 1JP7, 1M73, 1ODK, 1PK9, 1QE5, 1TCU, 1V4N, 1VMK, 1XE3, 1Z33, 2P4S, 3KHS, 3OZE, 3SCZ, 3TL6, 3UAV, 4D98       |
| Superoxide Dismutase (SOD)<br>1.10.287, $\alpha$ orthogonal bundle 2.60.40, $\beta$ sandwich | 1BSM, 1HDS, 1ICV, 1MA1, 1MMM, 1MY6, 1Q0E, 1WB8, 2ADP, 2JLP, 2W7W, 3BFR, 3ECU, 3EVK, 3LIO, 3QVN, 3SDP             |

Left column reports the name of the protein, short name, CATH code, and architecture; right column reports the PDB codes

- **Step 3.** Among all the variables extracted by means of Vadar, percent features have been preferred for their intrinsic homogeneity. More in details, the features related to residues have been transformed in a percent form by means of protein sequence length; on the other hand, the ones related to surfaces have been transformed by means of total accessible surface area. Furthermore, they have been normalized in a standard score form for a better stability relative to the EDA. That is, mean value has been subtracted from the data and the result has been divided by the standard deviation: details are described in the Sect. 2.3. Redundant features, as expected values, have been ignored;
- **Step 4.** Variables have been transformed into correlation and dissimilarity matrices, through the procedures explained under Sect. 2.3.1. Then, they have been used as features for an overall PCA, in order to verify the existence of common information. A comparison with a features network has been showed.

All the work has been executed with *R* [19] inside *R Studio IDE*, using some specific *R packages* to perform getting and cleaning phase and EDA. In particular: *stringr*, to rearrange file names [20]; *RCurl*, to manage connection for downloading [21]; *bio3d*, to compute DSSP inside R and read PDB files [22]; *corrplot*, to plot graphical correlation matrix [23]; *Hmisc*, to calculate correlation matrix with p-value [24]; *ppcor*, to calculate partial and semi-partial correlations with p-value [25]; *dendroextras*, to readjust and color dendrogram [26]; *ggplot2*, to plot PCA clustering [27]; *GeneNet*: to plot features network [28].

## 2.2 Statistical Methods

As part of EDA, two proven statistical procedures have been chosen for our work: correlation and principal component analysis [29], with different developments and additional interpretations.

Correlation has been performed as Pearson's correlation coefficients between pairwise features. Its practice must be carefully implemented, because of a batch of well-known traps (causality, multi-collinearity, outliers and so on). Statistical validation, performed here, procures only a quantitative robustness: an incisive analysis, together with a knowledge of data, allows to reach non-misleading conclusions. Partial correlation can help with collinearity problem, taking away the effects of another variable, or several other variables, on a relationship. Moreover, it can be used to detect possible redundant features.

Principal component analysis (PCA) is a very common multivariate statistical method, simple and powerful: it is an unsupervised approach and it is considered an EDA method. It allows summarizing initial variables in new ones, so-called components, which represent data in a more compact way and their tendency. Furthermore, given the intrinsic orthogonality of the components, PCA can be applied to obtain a kind of clustering [30], depending on inner information derived

from explained variance. This grouping helps in seeking possible outliers when executed on a dataset (it is a good habit searching for outliers, because they could polarize inferred results).

## 2.3 Mathematical Overview

### 2.3.1 Correlation, Partial Correlation and Dissimilarity

Given two variables with continuous values  $X = (x_1, \dots, x_r)$  and  $Y = (y_1, \dots, y_r)$ , where  $r$  is rows-observations number and  $c$  column-variables number, the density  $f(x_i, y_j)$  is represented by a single element in the normalized data table, and it is just a sort of bivariate distribution in a numerical form. A measure of strength and direction of association between the variables is provided by the covariance:

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x\mu_y \quad (1)$$

where

$$E[XY] = \sum_{i=1}^r \sum_{j=1}^c x_i y_j f(x_i, y_j) \quad (2)$$

where  $\mu_x, \mu_y$  are the expected values for a single variable. An index of covariation between  $X$  and  $Y$  is provided by the correlation coefficient:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (3)$$

where  $\sigma_x, \sigma_y$  are the standard deviations for a single variable. Given a third variable  $Z$ , the partial correlation coefficient between  $X$  and  $Y$  after removing the effect of  $Z$  is:

$$\rho_{yx-z} = \frac{\rho_{yz} - \rho_{yx}\rho_{zx}}{\sqrt{1 - \rho_{yx}^2} \sqrt{1 - \rho_{zx}^2}} \quad (4)$$

and it is possible to extend the formula in case of removing the effect of all the variables but one [31, 32]. Furthermore, a transformation from correlation to dissimilarity, by means of the formula:

$$d_{xy} = 1 - |\rho_{xy}| \quad (5)$$

allows to obtain a distance matrix, consistent with a cluster dendrogram on the variables themselves.  $d_{xy}$  is also known as Pearson's distance [33, 34]. Finally, every correlation coefficient has been validated with a  $t$ -test for significance, with the statistic:

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}} \quad (6)$$

where  $\rho$  is a generic correlation and  $n-2$  are the degrees of freedom [32].

### 2.3.2 Principal Component Analysis

Give a data table in a matrix form, it is possible to create new variables as linear combination of the old ones:

$$\begin{cases} PC_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1c}X_c \\ PC_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2c}X_c \\ \dots \\ PC_l = a_{l1}X_1 + a_{l2}X_2 + \dots + a_{lc}X_c \end{cases} \quad (7)$$

that have the largest variance. For a single principal component loading vector  $a_m = (a_{11}, \dots, a_{1c})^T$ , with  $m = 1, \dots, l$ , it is required to resolve an optimization problem:

$$\max_{a_m} \left\{ \frac{1}{r} \sum_{i=1}^r \left( \sum_{j=1}^c a_{1c} x_{ij} \right)^2 \right\} \quad (8)$$

subject to  $\sum_{j=1}^c a_{1j}^2 = 1$ . This is an eigenvalues-eigenvectors problem, numerical and computationally resolvable with Single Value Decomposition factorization, with  $a_m$  determined by:

$$(\Sigma - \lambda_m I) a_m = 0 \quad (9)$$

where  $\Sigma$  is the covariance/correlation matrix of the original data,  $\lambda_m$  are eigenvalues in descending order associated with  $a_m$  eigenvectors and  $I$  is the identity matrix.

After calculating the contribution of every eigenvalue  $\lambda_m / \sum_{k=1}^l \lambda_k$ , it is possible to choose the first several  $\lambda_m$  that cover a preset quantity of explained variability. In other words, the new data table composed by scores  $PC_k$ , always in matrix form,

represents the old one with a reduced dimensionality. Scores and loading vectors are plotted in a single biplot display [35, 36]. The challenge with this method is the new variables interpretation in the reality: that is, they are not so intuitive and their understanding is often delegated to investigator's experience.

### 2.3.3 Standardized Variables

Also known as z-score or standard score, a standardized variable has a mean equal to zero and a variance (standard deviation) equal to one, and it is possible to obtain it by means of the linear transformation:

$$z_x = \frac{x - \mu_x}{\sigma_x} \quad (10)$$

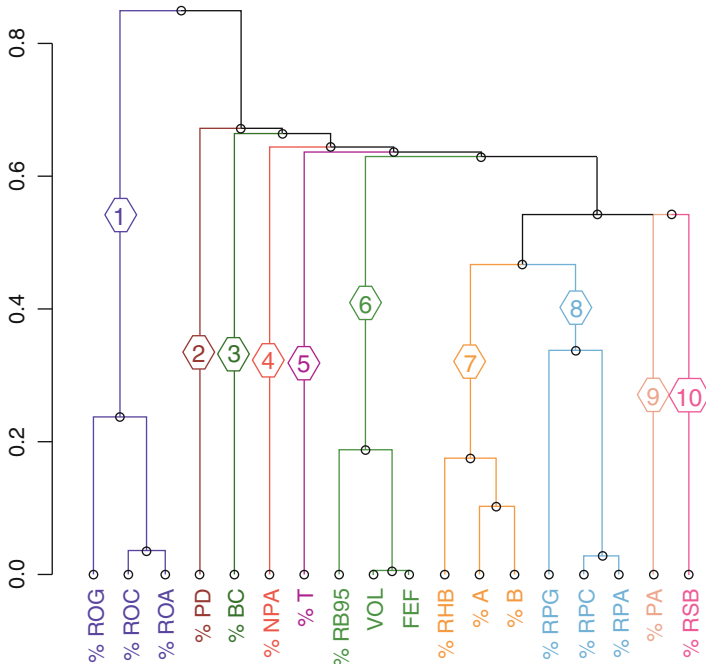
useful for comparing same variables from different distributions or variables with different units of measurement. This kind of normalization is recommended when correlations have to be used [32].

## 3 Results

Dendrogram in Fig. 2, obtained following formula (5), highlights relationships between the features chosen for the entire proteins dataset: it is the landmark about structural and geometrical features, but only in reference to the proteins chosen for assembling the dataset. There are four evident clusters: from the left, the first and the third concern torsion angles, the second concerns volume, free energy of folding and residues buried for the most part, and the fourth concerns secondary structures and residues convolved in hydrogen bonds.

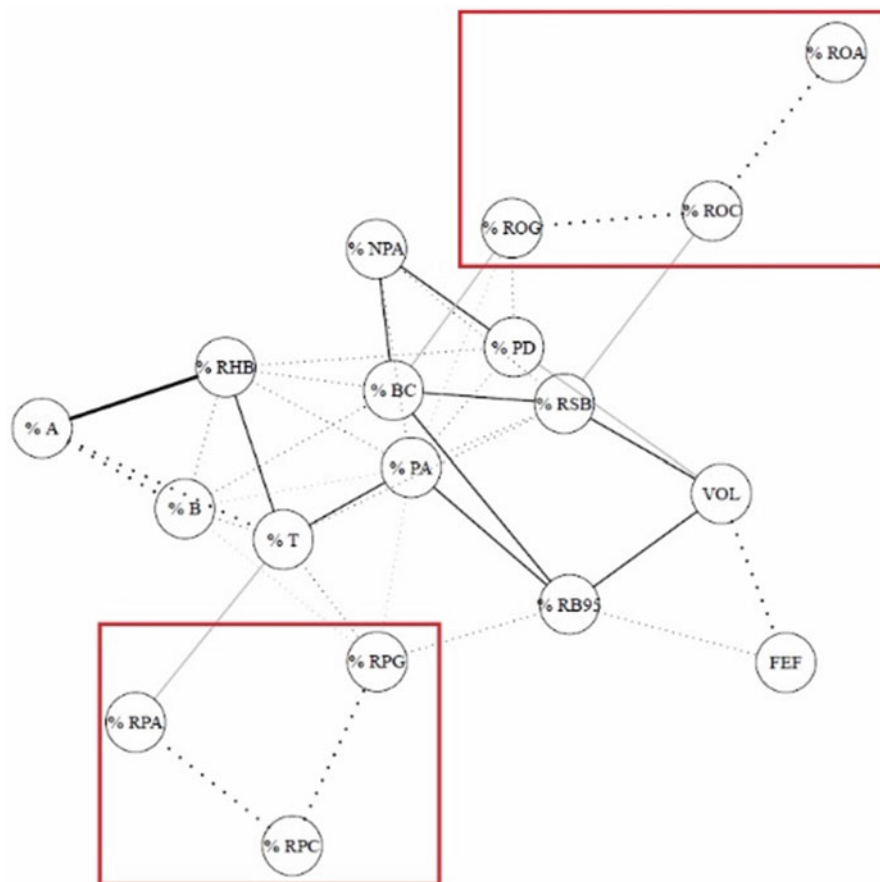
Features network in Fig. 3 has been plotted by means of partial correlations and graphical Gaussian model (GGM) [37]: it helps in seeking spurious correlations and pruning excessive features. For this dataset, torsion angles information results peripheral in the network, therefore they can be considered as unnecessary for the purpose of the work.

PCA, performed on the whole dataset, allows to extract the real important features in term of variability, producing a sort of clustering. In Fig. 4, the first principal component is composed by structural features (%A, %RHB) and second principal component by energy-geometrical ones (VOL, FEF, %RB95). This statistical technique is useful for outliers detection: for example, in the same plot, an isolated protein results so distant that it must be consider an outlier not only for its family (SOD), but also for the entire dataset. PCA performed only on SOD family has confirmed the result.



**Fig. 2** Dissimilarity dendrogram for proteins dataset. Every number (and color) indicates a cluster for the features. Cut-off has been put at 0.4, as deduced from the grafico. *Legend:* *ROx* omega angle core/allowed/generous, *PD* packing defect, *BC* buried charge, *NPA* non polar accessible surface area, *T* turn, *RB95* buried 95 %, *VOL* volume, *FEF* free energy folding, *RHB* hydrogen bond, *A* alpha helix, *B* beta sheet, *RPx* phi-psi angles core/allowed/generous, *PA* polar accessible surface area, *RSB* salt bridge

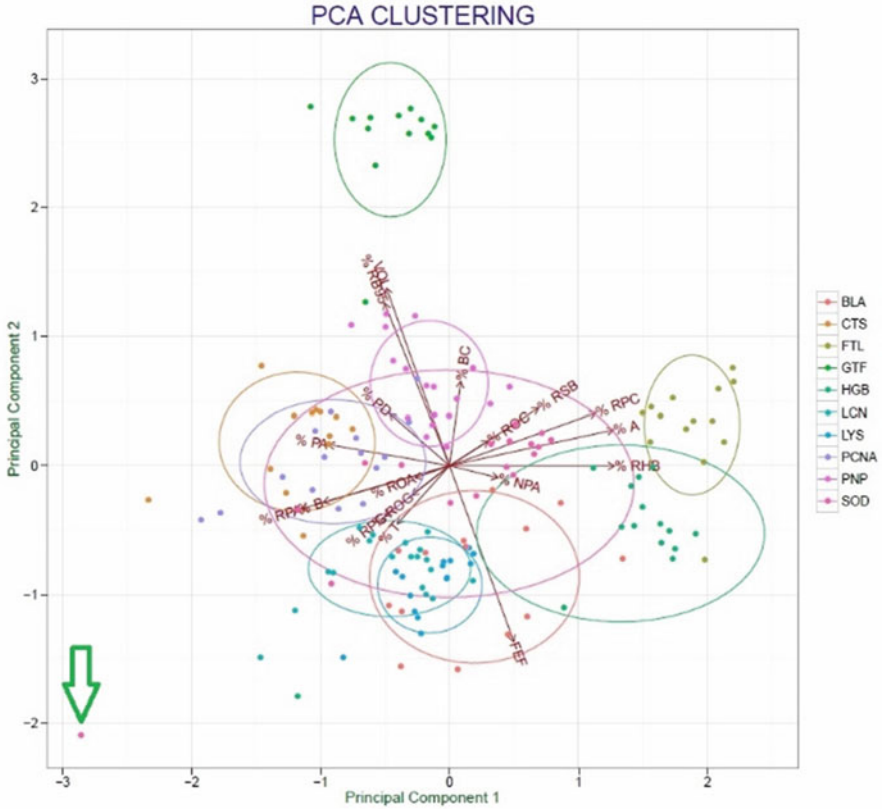
Moreover, previous plot questions if some relationships between the features are family-independent. A graphical correlation matrix for a single family protein may answer to this query. For example, choosing SOD family in Fig. 5 as test, it is possible to notice a strong family-specific “four-relationship” in the bottom left corner, between buried charged residues, secondary structure and free energy of folding. In this work, strong correlation threshold is 0.65, deduced from data. By contrast, some relationships are family-independent: for example, because of intrinsic physical-conformational connection (secondary structure and residues involved in hydrogen bonds) or prediction formula (volume and free energy of folding [15]).



**Fig. 3** Features network for proteins dataset. *Continuous line* represent partial correlation, *dotted line* represent partial anticorrelation (with the support of GGM). Peripheral subnetworks have been showed in the *squares*, which contain phi, psi and omega angle features. Meaning of acronyms as in Fig. 2

## 4 Conclusions and Perspectives

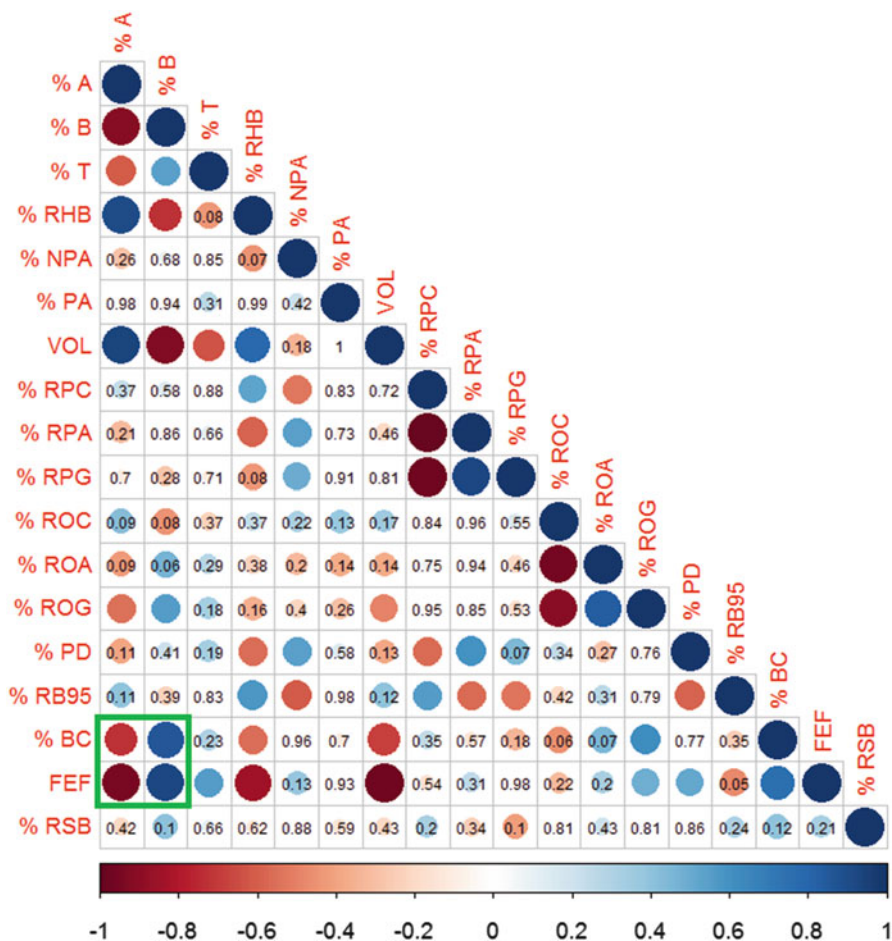
All the procedures that are part of EDA are well-suited for this kind of multivariate data: (a) distance dendrogram shows an overview about features interactions; (b) partial correlation indicates some possible redundant feature, if integrated in a network algorithm; (c) simple correlation helps in seeking family-specific features relationships; (d) principal component analysis is useful in finding family-specific connections to features and possible outliers. Therefore, these graphical multivariate procedures may be good tools in order to create a sort of fingerprint for the protein families themselves.



**Fig. 4** PCA for protein dataset. Centralizing ellipses enclosed each protein family. GTF family is polarized near positive PC2, FTL family near positive PC1 and SOD family is wide open. *Bottom left arrow* points to an outlier: *Pseudomonas putida* SOD A chain (PDB code: 3SDP). Protein families short names refer to legend in Table 1

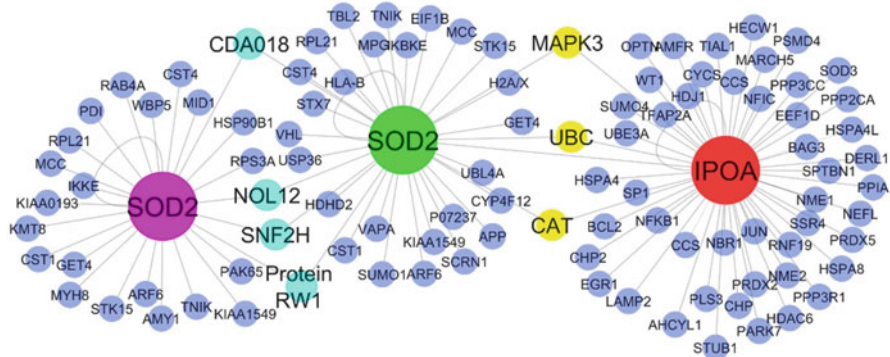
As future perspective, there are two directions of work enhancement: using advanced regression analysis to make a more robust features selection—partial correlation aids to do this, on the other hand PCA is not a real feature selection technique: it is rather a sort of “compression features” method—and integrating functional information (for example, by the analysis of protein interaction networks, as shown in Fig. 6) to highlight connections with the structural-geometrical ones.





**Fig. 5** Circular (lower) triangular correlation matrix for SODs. *Circle dimension* represents correlation strength, while *circle color* represent correlation direction (*blue*: correlation, *red*: anticorrelation). In the *green square*, there is a closed family-specific “four-relationship”. Numbers in the matrix show correlation statistically non-significant ( $p\text{-value} > 0.05$ )

**Acknowledgments** This work is partially supported by the *Flagship InterOmics Project* (PB.P05, funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations).



**Fig. 6** SODs partial interaction network. IPOA is an alternative name of SOD1 soluble; every SOD2 mitochondrial has its gene connections taken from different online database. Network has been drawn with *Cytoscape* [38]

## References

1. Tukey, J.W.: *Exploratory Data Analysis*. Behavioral Science: Quantitative Methods. Addison-Wesley, Reading (1977)
2. De Jong, E., Van der Loo, M.: *An Introduction to Data Cleaning with R*. Statistics Netherlands, The Hague (2013)
3. Branden, C., Tooze, J.: *Introduction to Protein Structure*, 2nd edn. Garland Publishing Inc, New York (1999)
4. Facchiano, A.M., Colonna, G., Ragone, R.: Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study. *Protein Eng.* **11**(9), 753–760 (1998)
5. Marabotti, A., Spyrikis, F., Facchiano, A., Cozzini, P., Alberti, S., Kellogg, G.E., Mozzarelli, A.: Energy-based prediction of amino acid-nucleotide base recognition. *J. Comput. Chem.* **29**, 1955–1969 (2008)
6. Russo, K., Ragone, R., Facchiano, A.M., Capogrossi, M.C., Facchiano, A.: Platelet-derived growth factor-BB and basic fibroblast growth factor directly interact in vitro with high affinity. *J. Biol. Chem.* **277**, 1284–1291 (2002)
7. Buonocore, F., Randelli, E., Bird, S., Secombes, C.J., Facchiano, A., Costantini, S., Scapigliati, G.: Interleukin-10 expression by real-time PCR and homology modelling analysis in the European sea bass (*Dicentrarchus Labrax L.*). *Aquaculture* **270**, 512–522 (2007)
8. Casani, D., Randelli, E., Costantini, S., Facchiano, A.M., Zou, J., Martin, S., Secombes, C.J., Scapigliati, G., Buonocore, F.: Molecular characterisation and structural analysis of an interferon homologue in sea bass (*Dicentrarchus labrax L.*). *Mol. Immunol.* **46**, 943–952 (2009)
9. Marabotti, A., D’Auria, S., Rossi, M., Facchiano, A.M.: Theoretical model of the three-dimensional structure of a sugar binding protein from *Pyrococcus horikoshii*: structural analysis and sugar binding simulations. *Biochem. J.* **380**, 677–684 (2004)
10. Marabotti, A., Facchiano, A.M.: Homology modelling studies on human galactose-1-phosphate uridylyltransferase and on its galactosemia-related mutant Q188R provide an explanation of molecular effects of the mutation on homo- and heterodimers. *J. Med. Chem.* **48**, 773–779 (2005)
11. Facchiano, A., Marabotti, A.: Analysis of galactosemia-linked mutations of GALT enzyme using a computational biology approach. *Proteins Eng. Des. Sel.* **23**, 103–113 (2010)

12. d'Acerno, A., Facchiano, A., Marabotti, A.: GALT protein database: querying structural and functional features of GALT enzyme. *Hum. Mutat.* **35**, 1060–1067 (2014)
13. Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., Yeats, C., Thornton, J.M., Orengo, C.A.: New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* **41** (Database issue):D490–8 (2013). URL <http://www.cathdb.info/>
14. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000)
15. Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R.F., Sykes, B.D., Wishart, D.S.: VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* **31**(13), 3316–3319 (2003)
16. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**(12), 2577–2637 (1983)
17. Till, M.S., Ullmann, G.M.: McVol - a program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. *J. Mol. Model.* **16**, 419–429 (2010)
18. Costantini, S., Colonna, G., Facchiano, A.M.: ESBRI: a web server for evaluating salt bridges in proteins. *Bioinformatics* **3**(3), 137–138 (2008)
19. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2014). <http://www.R-project.org/>
20. Wickham, H.: stringr: Make it easier to work with strings. R package version 0.6.2 (2012). URL <http://CRAN.R-project.org/package=stringr>
21. Temple Lang, D.: RCurl: General network (HTTP/FTP/...) client interface for R. R package version 1.95-4.3 (2014). URL <http://CRAN.R-project.org/package=RCurl>
22. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A., Caves, L.S.: Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**(21), 2695–2696 (2006)
23. Wei, T.: corrplot: Visualization of a correlation matrix. R package version 0.73 (2014) URL <http://CRAN.R-project.org/package=corrplot>
24. Harrell Jr, F.E., Dupontm, C. and al.: Hmisc: Harrell Miscellaneous. R package version 3.14- 5 (2014) URL <http://CRAN.R-project.org/package=Hmisc>
25. Kim, S.: ppcor: Partial and Semi-partial (Part) correlation. R package version 1.0 (2012). URL <http://CRAN.R-project.org/package=ppcor>
26. Jefferis, G.: dendroextras: Extra functions to cut, label and colour dendrogram clusters. R package version 0.2.1 (2014). URL <http://CRAN.R-project.org/package=dendroextras>
27. Wickham, H.: A layered grammar of graphics. *J. Comput. Graph. Stat.* **19**(1), 3–28 (2010)
28. Schaefer, J., Opgen-Rhein, R., Strimmer, K.: GeneNet: Modeling and Inferring Gene Networks. R package version 1.2.10 (2014). URL <http://CRAN.R-project.org/package=GeneNet>
29. Ding, Y., Cai, Y., Han, Y., Zhao, B., Zhu, L.: Application of principal component analysis to determine the key structural features contributing to iron superoxide dismutase thermostability. *Biopolymers* **97**(11), 864–872 (2012)
30. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the 21st International Conference on Machine Learning, Banff, 2004
31. Jobson, J.D.: Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design. Springer Texts in Statistics, 4th edn. Springer, New York (1999)
32. Edwards, A.L.: Multiple Regression and the Analysis of Variance and Covariance, 2nd edn. W.H. Freeman and Company, New York (1985)
33. Quinn, G.P., Keough, M.J.: Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge (2002)
34. Fulekar, M.H.: Bioinformatics: Applications in Life and Environmental Sciences. Springer, Heidelberg (2009)
35. Jolliffe, I.T.: Principal Component Analysis. Springer Series in Statistics, 2nd edn. Springer, New York (2002)

36. James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning: With Application in R*. Springer Texts in Statistics. Springer Science + Business Media, New York (2013)
37. Schaefer, J., Strimmer, K.: An empirical Bayes approach to inferring large scale gene association networks. *Bioinformatics* **6**(21), 754–764 (2005)
38. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Ideker, T.: Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011)

# Modelling of Protein Surface Using Parallel Heterogeneous Architectures

Daniele D'Agostino, Andrea Clematis, Emanuele Danovaro, and Ivan Merelli

**Abstract** A proper representation of protein surfaces is an important task in bioinformatics and biophysics. In a previous work we described a parallel workflow, based on the isosurface extraction and the CUDA architecture, able to produce high-resolution molecular surfaces based on the Van der Waals, Solvent Accessible, Richards-Connolly and Blobby definitions. In particular it is able to create surfaces composed by hundred millions triangles in less than 30 s using a Nvidia GTX 580, with speedup values up to 88. However in most application such number of triangles can be difficult to manage. In this paper we present an extension able to reduce the size of the surfaces by performing a simplification step, keeping however an high quality of the results. In particular the focus of the paper is on the efficient use of heterogeneous compute capabilities available on present workstations: the large surface produced using the CUDA device is progressively transferred and simplified on the host using the multicore CPU.

**Keywords** Protein surface • Heterogeneous architectures

## 1 Introduction

Superficial complementarities play a significant role to determine the possible binds between pairs of molecules [1–3] because mechanisms such as enzyme catalysis and recognition of signals by specific binding sites rely on macromolecular external morphological characteristics [4].

From the physical point of view, in fact, protein–protein interactions occur in two stages [5]. There is a first stage of molecular recognition, where the two macromolecules diffuse near each other until their interfaces come sufficiently close

---

D. D'Agostino (✉) • A. Clematis • E. Danovaro  
Institute of Applied Mathematics and Information Technologies, National Research Council of Italy, 16149 Genoa, Italy  
e-mail: [dagostino@ge.imati.cnr.it](mailto:dagostino@ge.imati.cnr.it); [clematis@ge.imati.cnr.it](mailto:clematis@ge.imati.cnr.it); [danovaro@ge.imati.cnr.it](mailto:danovaro@ge.imati.cnr.it)

I. Merelli  
Institute for Biomedical Technologies, National Research Council of Italy, 20090 Segrate, Italy  
e-mail: [ivan.merelli@itb.cnr.it](mailto:ivan.merelli@itb.cnr.it)

to begin the binding stage, when high affinity interactions are formed by modification of the side-chain and backbone conformations. It means that macromolecular interactions are driven at first by the conformation of the protein surfaces and just in a second phase the local physicochemical properties of the macromolecules are involved in minimizing the free energy of the system.

These interactions are usually analyzed using energetic approaches, but they are computational intensive, therefore a fast pre-processing screening based on the surface characteristics is very useful to reduce the set of the binding possibilities and to speedup the analysis.

Therefore, modeling macromolecular surfaces is an important task in many fields of bioinformatics and biophysics. Also visualization is very important, since a user is interested in the overall rendering quality of the molecular model: classical paradigms triangulate the surface and then visualize the mesh [6] or they are based on analytical models [7].

A molecular structure is represented through its 3D atomic coordinates. This is the format adopted by the Protein Data Bank (PDB) [8], the most important repositories for crystallography and nuclear magnetic resonance (NMR) structural analyses, that is commonly accepted as a standard. Such format well suits structural descriptions and it is the starting point for many other molecular surface definitions presented in the literature, each of them designed to some specific goals.

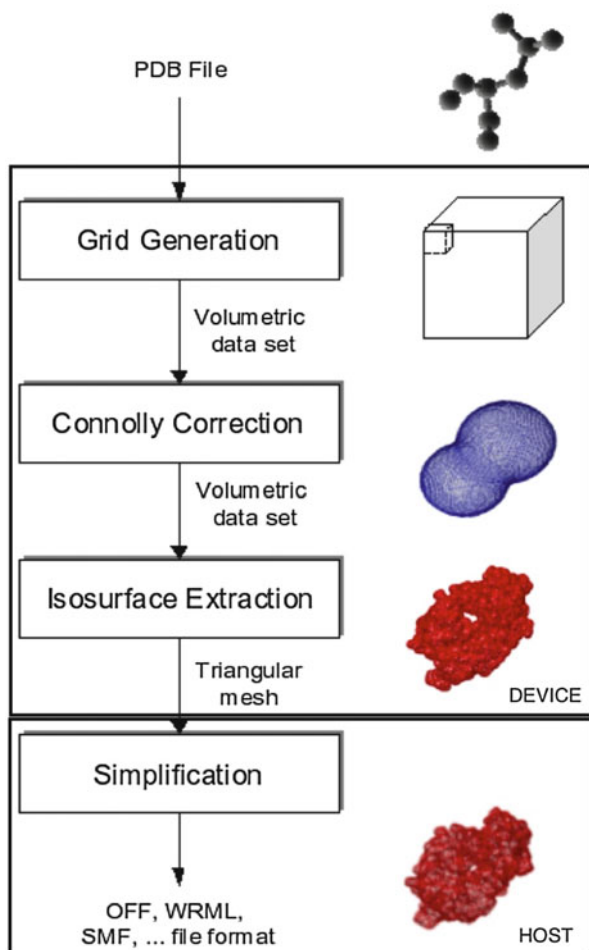
In this paper, we consider four among them, the most used ones, which are: the Van der Waals, the Solvent Accessible, the Richards-Connolly and the Blobby surfaces. We present an improved parallel algorithm based on the isosurface extraction and simplification operations, able to produce high-resolution surfaces for very large molecules using parallel heterogeneous architectures. Current workstations in fact can offer really amazing raw computational power, in the order of TFlops, on a single machine equipped with multiple CPU and GPU devices. The drawback is that the available computational cores, memories, and communication bandwidth can be extremely heterogeneous. The true challenge in using a similar system is the programming of parallel applications that are able to exploit in a efficient and effective way the different levels and capabilities [9].

These aspects represent the main contribution of the paper, that is an extension of [10, 11]. In [10] we presented the parallel workflow, originally designed for traditional homogeneous HPC cluster and therefore based on the message-passing paradigm. In [11] we improved the surface generation operation using the CUDA architecture, obtaining surfaces composed by hundred millions triangles with very high performance figures. However in most applications such number of triangles can be difficult to manage, therefore in this paper we exploit both accelerators and multicore processors to speedup the complete workflow including also the simplification operation.

The paper is organized as follows. Section 2 presents the four operations of the parallel workflow. Experimental results are discussed in Sect. 3, followed by Conclusions and possible future developments.

## 2 The Parallel Workflow for Molecular Surface Generation

The architecture of the workflow is represented in Fig. 1. The main input of the system is a PDB file containing the atomic coordinates of the atoms that form a molecule, while the output is represented by the isosurface corresponding to one of the four supported surface definitions. The workflow is composed by four operations: Grid Generation, Connolly Correction, Isosurface Extraction and Simplification.



**Fig. 1** The four stages of the parallel workflow for the reconstruction of molecular surfaces. The isosurface extraction and the simplification steps are executed concurrently, respectively on the CUDA device and the multicore CPU

## 2.1 Grid Generation

The first step for obtaining one of the four considered kinds of molecular surfaces through the isosurface extraction operation is the generation of a volumetric data set representing it. A volumetric data set can be viewed as a set of  $Z$  elements represented by  $XY$  slices. The values of  $X$ ,  $Y$  and  $Z$  are calculated as the bounding box of the molecule considering the minimum and the maximum atomic coordinates for each axis and taking into account the occupation volume of the bound atoms. The size of this grid is deeply influenced by the space sampling step, that typically varies between 0.7 and 0.1 Å. Smaller step values correspond to dense grids and high resolution surfaces, and vice versa.

The Van der Waals, the Solvent Accessible and the Richard-Connolly surfaces are obtained by modelling atoms as spheres that assumes negative inner values and increases gradually outwards, changing the sign just in correspondence of the Van der Waals surface. For example, in the volume that contains an oxygen atom, sign inversion occurs at 1.52 Å from the atom centre. These objects are positioned in a uniform space grid coherently with the PDB coordinates and are added in a point-wise fashion.

The blobby surface  $S$  is instead defined as

$$S := \{\mathbf{p} \in R^3 : G(\mathbf{p}) = 1\}, \quad G(\mathbf{p}) = \sum_{i=1}^{n_a} e^{B\left(\frac{\|\mathbf{p}-\mathbf{c}_i\|^2}{r_i^2} - 1\right)} \quad (1)$$

In particular,  $B$  is a negative parameter (the blobbyness) that plays the role of the probe radius when compared with the previous surfaces.

Following the CUDA name convention, the GPU card is called device and the CPU is called host. From the computational point of view, the value of each grid point is the result of the influence of all the atoms on it. For large molecules (e.g.  $10^5$  atoms), this means to consider several million points: therefore the device is the most suitable choice to perform this step. Due to performance consideration and hardware limitations it is not possible to generate a thread for each atom-point pair, therefore the parallelization has to be performed by subdividing the points or the atoms among the CUDA threads. We experimented that even if the partitioning on the number of points allows a greater scalability and parallelism degree, the achieved performance is much lower than with the alternative strategy. This is due to the large number of non-local memory accesses and by the fact that each atom influences in a significant way only the points within a limited bounding box surrounding it, so the subdivision of the atoms results in a lower number of operations.



## 2.2 Connolly Correction

This step is performed only when the Connolly surface is required. This surface consists of the border of the molecule that could be accessed by a probe sphere representing the solvent. When this sphere rolls around a pair of atoms closer than its radius, it traces out a saddle-shaped toroidal patch of reentrant surface. If the probe sphere is tangent to three atoms, the result is a concave patch of reentrant surface. The main difference with respect to the Van der Waals and the Solvent Accessible surfaces is that these patches close the superficial small cavities.

The Connolly Correction operation consists in changing the values of the points of the volume that become internal (and so with a negative value) considering these new patches. It is performed in two steps, the identification of the pairs of close atoms and the modification of the values of the points in the neighbourhood of these pairs.

Both of them are executed on the device. Even a middle-sized molecule such as 1GKI, described in Table 1, requires to analyze more than 381 million pairs for identifying which are sufficiently close, while with the largest one this number becomes greater than 8 billion. The resulting number or real pairs is obviously smaller, but much larger than that of the atoms, and this allows a better exploitation of the device compute capabilities with higher speedups. Also in this case the grid update is performed by subdividing the atoms' pair among the Cuda cores.

## 2.3 Isosurface Extraction

The third operation is the extraction of the isosurface representing the molecular shape. The Marching Cubes algorithm [12] is the most popular method used to extract triangulated isosurfaces from volumetric datasets. In the Marching Cubes algorithm, the triangular mesh representing the isosurface is defined piecewise over the cells in which the grid is partitioned. A cell is intersected by the isosurface represented by the isovalue if the isovalue is between the minimum and the maximum of the values assumed by the eight points of the grid that define each cell.

**Table 1** This table summarises the main characteristics of the three considered molecules and the two grid resolutions

| Molecule | Atoms  | Pairs     | Grid               | Bloppy triangles | Connolly triangles |
|----------|--------|-----------|--------------------|------------------|--------------------|
| 1GKI-0.5 | 19,536 | 413,983   | 226 × 235 × 237    | 1,410,816        | 2,082,492          |
| 1GKI-0.1 | 19,536 | 413,983   | 1132 × 1177 × 1185 | 36,583,252       | 61,272,696         |
| 1AON-0.5 | 58,674 | 1,179,430 | 312 × 477 × 469    | 4,256,936        | 6,412,276          |
| 1AON-0.1 | 58,674 | 1,179,430 | 1563 × 2385 × 2346 | 110,392,108      | 181,218,988        |
| 3G71-0.5 | 90,898 | 1,981,810 | 379 × 474 × 491    | 6,485,168        | 8,543,516          |
| 3G71-0.1 | 90,898 | 1,981,810 | 1894 × 2367 × 2458 | 167,548,496      | 258,570,660        |

This kind of cells is called active cells. An active cell contributes to approximate the isosurface for a patch made of triangles, and the union of all the patches represents the isosurface.

The Connolly surface is extracted considering the isovalue 0 if the Connolly Correction is performed. Otherwise the Van der Waals and the Solvent Accessible surfaces are extracted considering, respectively, the isovalues 0 and 1.4 [13]. Both the surfaces share the same geometric features with the difference that the Solvent Accessible surface increase the atomic radii with the probe radius, typically of a water molecule, that allows to minimize the probability of internal interstices and, somehow, takes into account the size and the presence of the solvent molecules.

As regards the Blobby surface, the definition given in Eq. (1) means that this surface can be obtained with the isosurface extraction operation considering the isovalue 1.0.

The parallelization of the original algorithm for the CUDA architecture is quite a straightforward task, because it is achieved by assigning one cell for each thread. But it presents some issues, as the duplication of the triangles vertexes and the need to transfer up to several Gigabytes of results. In [11] we discussed in details an efficient implementation able to produce high-resolution molecular surfaces by overlapping computation and data transfer. A further advantage of this implementation is that it can act as a component in a workflow, because it produces an output suitable for both the direct visualization and the reuse of the generated molecular surface for following analyses.

This implementation is included in the pseudocode for the protein surface generation for heterogeneous architectures presented in Listing 1.

A key feature of the Isosurface extraction operation, as well as of the above two operations, is that they are executed on the device and in an iterative way. As we said before, the size of the grid is determined on the basis of the coordinates of the atomic centres and the required sampling step. Smaller step values correspond to dense grids and high resolution surfaces, and vice versa. As shown in Table 1, the size of volumetric data sets can be of several GBs, while the size of most of the device is more limited. Considering that the isosurface extraction operation requires to process a pair of slices at a time, we implemented this part of the workflow in an iterative way for increasing values of the Z coordinate, as shown in the pseudocode. This means that one slice is created at each iteration (except for the first one, where the slices for  $Z = 0$  and 1 are created) in order to replace the slice having the lowest Z value. In this way, we are able to process very large data sets if the size of a pair of slices does not exceed the device memory. Moreover, we overlapped computation and the transfer of the result from the device to the host, in order to hide this overhead as much as possible. Considering that the size of the largest isosurface is of about 5 GB, the data transfer time would otherwise represent a major issue for the performance.

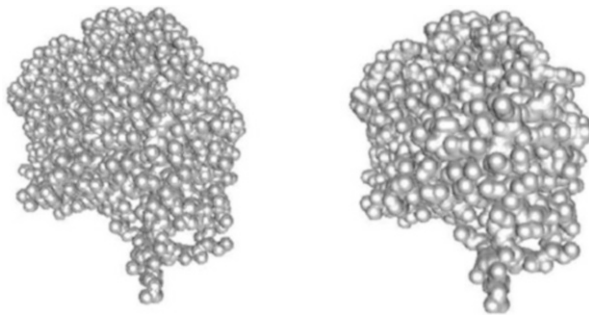
**Listing 1** Pseudocode of the workflow for the parallel protein surface generation for heterogeneous architecture.

```

#pragma omp parallel sections {
  /** PRODUCER **/
  #pragma omp section {
    if (CONNOLLY) Connollypairs(atoms , pairs);
    for (z = 0; z < ZDIM; z++) {
      if (z == 0) { SliceCreation <<<dimG,dimB>>>(sliceInf ,
        z , atoms)
        if (CONNOLLY) Correction <<<dimG,dimB>>>(sliceInf ,
        z , pairs) }
      else { sliceInf = sliceSup;
        SliceCreation <<<dimG,dimB>>>(sliceSup , z+1, atoms)
        if (CONNOLLY) Correction <<<dimG,dimB>>>(sliceSup ,
        z+1, pairs) }
      /** Start Overlap 1 **/
      VerticesCalc <<<dimG,dimB, stream1 >>>(Sliceinf , Slicesup ,
        xlabel , ... , z)
      if (z != 0) copy_async (restri[numtrifound ],TOHOST, stream2)
      /** Stop Overlap 1 **/
      if (numtrifound > THRESHOLD) /* activate the consumers */
      xscan = Scan(xindex) ... zscan = Scan(zindex)
      /* The number of resulting vertices is numtrifound */
      numvertpre = numvert , numvert += numvertfound
      VerticesCompact<<<dimG,dimB>>>(xscan , ... , resvert ,
        numvertpre)
      /** Start Overlap 2 **/
      TrianglesCalc <<<dimG,dimB, stream3 >>>(Sliceinf , Slicesup ,
        trivect , trinumvect)
      copy_async (resvert[numvertfound ],TOHOST, stream4)
      /** Stop Overlap 2 **/
      restrinumvect = Scan(trinumvect)
      /** The number of resulting triangles is numtrifound **/
      TrianglesCompact<<<dimG,dimB, stream5 >>>(trivect ,
        restrinumvect , restri)
      /* The transfer of triangles of the last pair */
      if (z == ZDIM-1) copy (restri[numtrifound ],TOHOST)
      /*Stops the consumer at the end of the simplification
        of the last block of triangles */
    } }

  /** CONSUMERS **/
  #pragma omp section {
    while (!stop) {
      /*Wait for a block of triangles */
      ReadModel(resvert , restri , model) //Done in parallel
      Simpl(model, percsimpl) //Done in parallel
      WriteModel(model)
    } } }

```



**Fig. 2** A comparison of two different representations of the FAB complex [PDB:1A5F] composed by 1.3 million triangles. The *left image* is obtained considering a grid step of 0.3 Å, while the left one is obtained with a grid step of 0.1 Å combined with the simplification of 90 % of the produced triangles. The latter requires more computing time but the surface quality is higher

## 2.4 Simplification

The Marching Cubes algorithm has the characteristics of producing a large number of small planar triangles, and this results in very large triangular meshes without any advantage in terms of available information. With the simplification operation it is possible to tackle this issue because this operation allows to obtain a smaller but equivalent surface by exactly merging small, planar triangles.

If the objective is only the reduction of triangles, the selection of a bigger step for Grid Generation can represent an alternative. However in this case the resulting mesh will have an uniform coarse grain level of detail, while a simplified mesh is characterized by the non-uniform level of detail, where the most irregular zones are represented using more triangles than the regular ones. In Fig. 2 two Connolly surfaces of the same molecule made up by about 1,300,000 triangles are presented. If the left one, obtained using a finer grid and the simplification operation with a simplification percentage of 90 %, is compared with the right one, obtained using a coarser grid, it appears that the best result is obtained using the simplification operation.

This operation is performed on the CPU, because irregular data structures do not suit well on devices, and it is based on the MPI-based parallel simplification algorithm described in [14, 15]. In the present system we re-implemented it using OpenMP, because we exploit the shared memory paradigm where a concurrent set of threads cooperate to update a single, shared mesh representation when an edge collapse operation occurs. The main achievement of this implementation is represented by the fact that the isosurface extraction and this operation work in a producer-consumer way (see Listing 1): when a given number of triangles is produced, the simplification operation is activated on them on the CPU while the device continues its iterative process. Such strategy is based on [16], that is a sliding-window approach that has the advantage to allow the simplification of also the

border regions between the execution of the operation on different sets of triangles. This strategy has the further major advantage to allow overlapping the computation performed on both the heterogeneous computational resources of a workstation, with an important advantage in terms of execution times.

### 3 Experimental Results

Experimental results were collected using a workstation equipped with dual six-cores Intel Xeon E5645 CPUs and one NVIDIA GTX580 device.

Three molecules of the PDB repository, chosen on the basis of their size, and two grid spacings, 0.5 and 0.1 Å corresponding to a medium-detailed and a high-detailed resolution, were considered for the scalar field generation operation. Their characteristics and the resulting volumetric datasets are shown in Table 1. We presented also the size of the resulting Blobby and the Connolly surfaces, because the size of the Van der Waals and Solvent Accessible surfaces are very close to the Blobby one in all cases.

For sake of brevity Table 2 shows the performance of the sequential and parallel implementations for the Blobby surface definition. We can see that about 80% of the sequential time is spent in the simplification operation. The performance for its parallel version are limited by the fact that the adopted data structure for representing the model contains several explicit representations of elements relations (i.e vertexes, triangles and edges), therefore a large number of “critical” regions is required. Nevertheless, a good overlap occurs with the operations performed on the CUDA device: in the largest case we produced a mesh composed by 80 million triangles in about 20 min instead of 1 h.

### 4 Conclusions and Future Developments

In this work we presented a parallel workflow for the modeling of protein surfaces based on the Van der Waals, Solvent Accessible, Richards-Connolly and Blobby definitions. The main characteristic is represented by the fact that the workflow is able to exploit at a time the heterogeneous compute capabilities of present workstation to overlap computations both on multicore CPUs and devices.

In particular the Marching Cubes algorithm used in the isosurface extraction operation has the drawback to produce large triangular meshes composed by small, planar triangles. The execution of the simplification operation allows to get smaller surfaces with higher quality, and the implementation presented here has the advantage to overlap isosurface extraction and simplification operations, in the sense that the large isosurface produced using the device is progressively transferred and simplified on the host using the multicore CPU available on present workstations.

**Table 2** This table presents the times, in seconds, for executing the implementation of the Molecular Surface Generation for heterogeneous architectures considering the Blobby surface definition and a simplification of 50 %

|          | Scalar field generation |             | Isosurface extraction |              | Simplification |              | TOTAL  |              |
|----------|-------------------------|-------------|-----------------------|--------------|----------------|--------------|--------|--------------|
|          | Seq.                    | CUDA        | Seq.                  | CUDA         | Seq.           | OMP          | Seq.   | CUDA+OMP     |
| IGKI-0.5 | 0.78                    | 0.02 (39.0) | 2.56                  | 0.05 (47.4)  | 22.9           | 12.4 (1.8)   | 26.2   | 12.5 (2.1)   |
| IGKI-0.1 | 52.07                   | 1.22 (42.5) | 82.33                 | 1.87 (43.9)  | 587.2          | 279.7 (2.1)  | 721.6  | 281.2 (2.6)  |
| IAON-0.5 | 3.67                    | 0.07 (50.3) | 8.26                  | 0.14 (59.8)  | 68.3           | 39.0 (1.7)   | 80.2   | 39.1 (2.0)   |
| IAON-0.1 | 184.92                  | 3.55 (52.1) | 449.70                | 9.81 (45.8)  | 1771.8         | 738.2 ( 2.4) | 2406.4 | 744.9 (3.2)  |
| 3G71-0.5 | 4.34                    | 0.08 (53.6) | 10.28                 | 0.17 (59.4)  | 104.0          | 57.8 (1.8)   | 118.6  | 58.0 (2.0)   |
| 3G71-0.1 | 229.38                  | 5.05 (45.4) | 502.19                | 12.88 (39.0) | 2689.1         | 1222.3 (2.2) | 3420.7 | 1229.7 (2.8) |

In brackets the achieved speedups. It is worth noting that, in the total time for the CUDA version, we did not consider the initialisation time, that is of about 3 s in all cases

In the present implementation the simplification is the most computational intensive part, requiring about 80 % of the sequential execution time. In a future work we will explore possible alternative data structures to speed up the algorithm, in order to have a better overlap and thus improving the overall performance.

The possibility of performing a fast screening of possible macromolecular interactions using this surface matching algorithm is very important, since it can be the core of a full docking procedure. The perspective is to analyse possible interactions between biological macromolecules, starting from the knowledge of their structures, but working mainly on surface descriptions, which seem to bring effective information about their functional capabilities.

## References

1. Kinoshita, K., Nakamura, H.: Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci.* **14**(3), 711–718 (2005)
2. Binkowski, T.A., Adamian, L., Liang, J.: Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.* **332**(2), 505–526 (2003)
3. Merelli, I., Cozzi, P., D'Agostino, D., Clematis, A., Milanese, Image-based surface matching algorithm oriented to structural biology. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**(4), 1004–1016 (2011)
4. Via, A., Ferre, F., Brannetti, B., Helmer-Citterich, M.: Protein surface similarities: a survey of methods to describe and compare protein surfaces. *Cell. Mol. Life Sci.* **57**, 1970–1977 (2000)
5. Camacho, C.J., Vajda, S.: Protein-protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.* **12**, 36–40 (2002)
6. Yu, Z., Hols, M.J., Cheng, Y., McCammon, J.A.: Feature-preserving adaptive mesh generation for molecular shape modelling and simulation. *J. Mol. Graph. Model.* **26**(8), 1370–1380 (2008)
7. Sanner, M., Olson, A.J., Spehner, J.C.: Reduced Surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**(3), 305–320 (1996)
8. Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H., Westbrook, J.: The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.* **7**(11), 957–959 (2000)
9. Danovaro, E., Clematis, A., Galizia, A., Ripeti, G., Quarati, A., D'Agostino, D.: Heterogeneous architectures for computational intensive applications: A cost-effectiveness analysis. *J. Comput. Appl. Math.* **270**, 63–77 (2014)
10. Merelli, I., Orro, A., D'Agostino, D., Clematis, A., Milanese, L.: A parallel protein surface reconstruction system. *Int. J. Bioinf. Res. Appl.* **4**(3), 221–239 (2008)
11. D'Agostino, D., Clematis, A., Decherchi, S., Rocchia, W., Milanese, L., Merelli, I.: CUDA accelerated molecular surface generation. *Concurr. Comput. Pract. E* **26**(10), 1819–1831 (2014)
12. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3-D surface construction algorithm. *Comput. Graph.* **21**(3), 163–169 (1987)
13. D'Agostino, D., Merelli, I., Clematis, A., Milanese, L., Orro A.: A parallel workflow for the reconstruction of molecular surfaces. *Adv. Parallel Comput.* **15**, 147–154 (2008)
14. Garland, M., Heckbert, P.S.: Surface simplification using quadric error metrics. In: *ACM SIGGRAPH 1997 Proceedings*, pp 209–216 (1997)
15. Clematis, A., D'Agostino, D., Mancini, M., Gianuzzi, V.: Parallel decimation of 3D meshes for efficient web-based isosurface extraction. *Adv. Parallel Comput. Ser.* **13**, 159–166 (2004)
16. Hoppe, H.: Efficient, lossless, continuous-resolution representation of surface triangulations. In: *ACM SIGGRAPH 1996 Proceedings*, pp 99–108 (1996)