IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES

Padmaja Balachandran Kamath

Padmaja Balachandran Kamath

In silico Modeling of Chemical and Biological Interactions at Different Scales

DOCTORAL THESIS

Supervised by:

Prof. Alberto Fernández Sabater

and

Prof. Robert Rallo Moya

Department of Chemical Engineering

UNIVERSITAT ROVIRA I VIRGILI

Tarragona 2016

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES
Padmaja Balachandran Kamath

Departament d´Enginyeria Química
Universitat Rovira i Virgili
Campus Sescelades
Avda. Països Catalans, 26
43007 Tarragona
Tel: 977 55 84 57
Fax: 977 55 96 21

Prof. Alberto Fernández Sabater & Prof. Robert Rallo Moya

CERTIFY

That the present study, entitled "*In silico* modeling of chemical and biological interactions at different scales" presented by Padmaja Balachandran Kamath for the award of the degree of Doctor, has been carried out under our supervision at the Chemical Engineering Department of the University Rovira i Virgili, and that it fulfills the requirements to obtain the Doctor International Mention.

Tarragona, 31st August 2016

Prof. Alberto Fernández Sabater Prof. Robert Rallo Moya

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES
Padmaja Balachandran Kamath

Acknowledgement

The PhD journey is not just one person's, as it often appears to be. My journey as a PhD student has been no different in this sense. I had the privilege of working with many people and establishing professional and personal relationships with them. I would like to thank everyone who has provided me with guidance and assistance at different times in my thesis. Small or big, it always helped me get closer to getting a PhD!

I am deeply grateful to my advisor, Prof. Alberto Fernández for his excellent guidance and support throughout my PhD tenure. Without his patience and dedication to helping me learn I would not have made it to this point. I really appreciate his predisposition to help in scientific and personal matters. All the data manipulation and management skills that I have been appreciated for in my jobs after my PhD term, are a result of meticulous and in-depth trainings that I have received from him.

I would like to thank my co-advisor Prof. Robert Rallo for all the research lines, schemes and ideas for my PhD at large. Initially it was quite intimidating for me to have discussions with him but later on having him as my co-advisor brought new dimensions to my knowledge and thesis. Despite his busy schedule and constant travels, he has been available for meetings and discussions and I cannot thank him enough for that.

I would like to thank Dr. Emilio Benfenati, my research stay advisor for giving me a chance to work in his group and being supportive throughout my stay in Milan and beyond. He has been very kind to accommodate my work and me amidst his busy schedule and big group. I shall always be a fan of his strategic and dynamic research schemes and truly cherish his guidance and leadership. Despite his exceptional professional accomplishments, he is a personification of genuine modesty.

Also thanks to Nuria Juanpere for patiently helping me with all the administrative tasks right from visa documentation to thesis submission. Facebook, email and telephone, I have contacted her everywhere and she has been very kind to respond! Merche Maurin has also been very welcoming and helpful.

Office mates also become an integral part of the PhD experience after having spent long working hours with them. I have been lucky to have amicable office mates all through my PhD (except for the times I was alone in Lab 224). I would like to thank Tatiana, Chandan, Xavier and Roger with whom I have shared two offices in Tarragona. I would like to thank Josep especially for being an integral pillar and helping me with a long list of programming tasks during my PhD. No favour was ever denied or left unattended. I would also like to thank Laura for all her help during my stay in Milan and thesis submission. Muchas gracias also to Lila, for her kind words and heart to heart conversations, which have made me feel better.

Sometimes having a friend as next lab neighbour is a great source of fun and entertainment. I would also like to thank my next office neighbour and friend Viji for all the silly fights, funny moments and dinners. Our experimental v/s computational biology arguments still bring a broad smile on my face.

I would like to extend my gratitude to all the coolest group members and friends I had in Milan! It was a pleasurable experience for me not only to work but also hangout with them and I had the time of my life. I would like to thank all my girls (and boys), Nazanin, Rudy, Marco, Alla, Alessandra, Andrey, Caterina, Sabrina, Azadi, Francesca and Maria, Alice, Fabiola, Yuri, Pier Paolo, Vittorio, Anna and Alberto for all the conversations we have had about everything under the sun and their friendship. I want to thank Nelly for her endless support and patience with data, methods and software and also for the Italian gesture lessons (along with Alice (sometimes in wonderland) and Yuri) to cheer me up whenever I was upset about research. I would also like to thank Claudina, a best friend who made Milan a wonderful memory for me! Despite of not being able to see each other every day, she still has kept encouraging and motivating me.

A big thank you ciccina! Serena, the crazy Serenella, a big thank you for all the late night conversations, skype calls and our little crazy discussions. Lastly, I also want to thank Sabrina Bidoli, for making my stay at Mario Negri for making seamless and for all her efforts to resolve the issues in my apartment.

My stay at Milan will be incomplete without thanking Prof. Hitesh Shrimali. I wish to thank him for all the weekend sightseeings, dinners, art sessions, VLSI and most importantly lessons on PhD life. I now wholeheartedly agree, that PhD is more about patience and perseverance and not about obtaining perfect results and publishing all the time!

Talking about friends, truly I was lucky enough to come across very nice people. I would like to thank my friends, Laxmi, Pinkie, Behrouz, Denisa, Mimmy, Mabel, Ewelina and Ali for this great experience.

I want to acknowledge Priya and Subba who became close friends in the second year of my PhD, for all the nice times we have spent in Tarragona. Special dinners, chit chats, jokes and conversations I just can't forget them. I have immensely missed their company after moving to Leuven.

Some relationships will always be special. I was lucky to have met two very magnanimous people called Shirley and Carlos, who now are like family to me. I want to extend my thanks to them for everything that they have done for me right from the first year of my PhD. Language and cultural differences have never come in their way while showering love and concern on me and I shall always be grateful to them.

As it is said, finishing is more difficult than starting. Writing papers and thesis got difficult as I relocated to Leuven after my PhD term. I would like to thank all my friends, Joris, Barbara, Ravi, Nirosha, Hari, and Akshay for constantly enquiring about my thesis writing. Leuven now feels home because of you guys. I want to thank Marissa, for the motivation and company during writing. A friend in the same domain and similar phase in PhD is the best

comforter. Special thanks also go to Damini for patiently putting up with the thesis-saga! ;)

I would also like to thank my colleagues at the WIV-ISP, Els Van Hoeck, Fabien, Els Collijs, Marie Noëlle, Melissa, Luc, Annemie, Birgit, Tatyana and Philippe for all the knowledge, support, encouragement and skills they gave me on regulatory toxicology. I would like to specially thank the head of our Toxicology unit, Dr. Christiane Vleminckx for her inputs and explanations that made me look at my PhD work in a different light altogether. It is in this time that I began to understand how my work was a part of the big picture.

In addition, I also want to thank the very vibrant and special part of my life, my rocking train buddies (also colleagues of WIV-ISP). I really cherish each day I spend with them. I would like to thank Celine, Patrick, Eveline, Sigrid, Sona, Shreeya, Sara, Sarah, Sylvanus and Tadek for constantly enquiring and motivating me to finish my thesis.

I wish to express my love and gratitude to the most wonderful people in my life, my parents. With my PhD, I feel my mother also graduates with me. Although she studied Economics and programming and has no background in Science, Technology or Engineering, she has always tried to expand her knowledge in the subjects I have studied to understand my work and give me quality inputs. I am really fortunate to have invaluable inputs (and intense hypothetical discussions) from my father, a passionate engineer who has always motivated me think out of the box and instilled the importance of education in me during my formative years. I could not have better mentors than my parents, who did everything they could do selflessly, to lay a strong foundation for my career. Education indeed is the best asset I have. I want to thank my brother Girish, for standing by and protecting me in the most difficult times of my life. If not for him, I am sure working towards a PhD would just have been a dream. I want to thank my sister-in-law Preeti, for her constant banters, pranks and laughter she has filled my life with. I am deeply grateful to Ajith Sir, Saroja Aunty and my dear Ajitha akka, for their endless love and encouragement that has always brought out the best in me. Special

thanks go also to my in-laws, Maa, Papa, Pritee and Vaibhav for all their prayers and wishes they have invested in me.

Last but not the least, a hearty thank you goes to a kind gentleman who has supported me as a friend, partner and husband. Despite of his share of struggles during his PhD, Prashant has been putting up with the roller coaster ride that I have had in mine. Without him, life would not be as beautiful and fun as it is.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES
Padmaja Balachandran Kamath

To my lovely parents

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES
Padmaja Balachandran Kamath

Summary

In the past decades, government, society and industry at large have taken keen interest in the impact at different scales that exposure to chemicals has on humans and environment. Hence, in many countries governments have imposed regulations as per which it has become important to establish the potential effects of these chemical entities with respect to human health and environmental endpoints. Given the time taken by traditional tests, costs and large number of chemicals to be evaluated, there has been a rapid growth in the number of computational models that link the structure of chemical substances to their biological activity.

Nanoparticles are also being used increasingly across different classes of consumers' products. Since, in physiological context, the protein corona constitutes the interface between the nanoparticle and cells, it plays a fundamental role in nanoparticle-cell association. In this line of work, the physicochemical properties of protein corona were used to develop a model to predict cell association.

To extend the basis of knowledge that currently exists in Structure Activity Relationship (SAR) models for chemicals, a similar approach was used to develop a new model and generate sets of metabolic triggers which can be used together with Q(SAR) methods. This work presents SAR rules for prediction of

mutagenicity *in vitro*, along with metabolic triggers for prediction of mutagenicity *in vitro* and *in vivo*. Furthermore, the metabolic triggers approach can also be used to obtain a preliminary idea if a chemical exhibits the same or contrary mutagenic behaviour *in vitro* and *in vivo*.

Lastly, this thesis focuses on the topic of drug resistance in bacteria, which has become a matter of global concern. With bacteria growing resistant to antibiotics at a faster pace than the discovery of new antibiotics, it is important to have information on the response that new bacterial proteins would have to the currently available antibiotics, based on their similarity with the known antibiotic-resistant proteins. In this work an alignment-free method was developed to improve the resistance profile classification of bacterial proteins based on their physicochemical properties.

Resumen

En las últimas décadas, el gobierno, la sociedad y la industria en general han tomado gran interés en el impacto a diferentes escalas que la exposición a los productos químicos tiene sobre los seres humanos y el medio ambiente. Por lo tanto, en muchos países los gobiernos han impuesto regulaciones según las cuales se ha vuelto importante establecer los efectos potenciales de estas sustancias químicas con respecto a la salud humana y a criterios medio ambientales. Teniendo en cuenta el tiempo necesario para las pruebas tradicionales, los costes y el gran número de productos químicos a evaluar, se ha producido un rápido aumento en el número de modelos computacionales que relacionan la estructura de las sustancias químicas con su actividad biológica.

Las nanopartículas se están utilizando cada vez más a través de diferentes clases de productos usados por los consumidores. Dado que, en un contexto fisiológico, la corona de las proteínas constituye la interfaz entre las nanopartículas y las células, ésta desempeña un papel fundamental en la asociación entre nanopartículas y células. En este trabajo, las propiedades físico-químicas de la corona de las proteínas se han utilizado para desarrollar un modelo para predecir la asociación celular.

Con el fin de ampliar la base de conocimientos que existe actualmente en los modelos de relación estructura-actividad (SAR) para productos químicos, se ha utilizado un enfoque similar para desarrollar un nuevo modelo y generar conjuntos de alertas metabólicas que puedan utilizarse junto con los métodos Q(SAR). Este trabajo presenta reglas SAR para la predicción de mutagenicidad *in vitro*, junto con alertas metabólicas para la predicción de mutagenicidad *in vitro* e *in vivo*. Además, el enfoque con alertas metabólicas también se puede utilizar para obtener una idea preliminar acerca de si un producto químico exhibe el mismo comportamiento mutagénico *in vitro* e *in vivo*.

Por último, esta tesis se centra en el tema de la resistencia a los fármacos en las bacterias, que se ha convertido en un asunto de interés global. Con el aumento de la resistencia de las bacterias a los antibióticos a un ritmo más rápido que el descubrimiento de nuevos antibióticos, es importante disponer de información sobre la respuesta que las nuevas proteínas bacterianas tendrían sobre los antibióticos actualmente disponibles, en función de su similitud con las proteínas que sabemos resistentes a los antibióticos. En este trabajo se ha desarrollado un método de alineación libre para mejorar la clasificación en perfiles de resistencia de las proteínas bacterianas en base a sus propiedades físico-químicas.

List of Publications

1. Padmaja Kamath, Alberto Fernandez, Francesc Giralt, Robert Rallo. Predicting cell association of surface-modified nanoparticles using protein corona structure–activity relationships (PCSAR). *Current Topics in Medicinal Chemistry*, 15 (2015), pp. 1930–1937.

2. Padmaja Kamath, Giuseppa Raitano, Alberto Fernandez, Robert Rallo, Emilio Benfenati. *In silico* exploratory study using structure-activity relationship models and metabolic information for prediction of mutagenicity based on the Ames test and rodent micronucleus assay. *SAR and QSAR in Environmental Research*, 26 (2015), pp. 1017–1031.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES
Padmaja Balachandran Kamath

Abbreviations

ANTARES	Alternative Non-Testing methods Assessed for REACH substances
AP	Averaged physicochemical descriptor
ARDB	Antibiotic Resistance Database
BLAST	Basic Local Alignment Search Tool
CAS	Chemical Abstracts Service
CD-HIT	Cluster Database at High Identity with Tolerance
CMR	Carcinogenic, Mutagenic and toxic for Reproduction
CRAFT	Chemical Reactivity and Fate Tool
CWBM	Interference with Cell Wall synthesis and disruption of Bacterial membrane structure
DNA	Deoxyribonucleic acid
ECB	European Chemicals Bureau
ECHA	European Chemicals Agency
EMBOSS	European Molecular Biology Open Software Suite
ENM	Engineered Nanomaterials
FN	False Negative
FP	False Positive
IMP	Inhibition of Metabolic Pathway
INAS	Interference with Nucleic Acid Synthesis
IPS	Inhibition of Protein Synthesis
ITS	Integrated Testing Strategies
LMO	Leave-many-out
LOO	Leave-one-out
MCC	Matthews' correlation coefficient
NCBI	National Center for Biotechnology Information
NP-PC	Nanoparticle-Protein Corona
OECD	Organization for Economic Co-operation and Development

PCSAR	Protein Corona Structure – Activity Relationships
(Q)SAR	(Quantitative) Structure – Activity Relationships
R2	Squared correlation coefficient
REACH	Registration, Evaluation, Authorization and restriction of Chemicals
SA	Structural Alerts
SAR	Structure – Activity Relationships
SARpy	SAR in python
SMILES	Simplified Molecular Input Line Entry System
TN	True Negative
TP	True Positive
VEGA	Virtual models for property Evaluation of chemicals within a Global Architecture
WEKA	Waikato Environment for Knowledge Analysis

Contents

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES
Padmaja Balachandran Kamath

Index of Figures

Index of Tables

Chapter 1 Introduction

1.1 Introduction

Potentially pathogenic organisms and environmental pollutants can harm human health through a series of complex transport and exposure pathways. Humans are subjected to a range of bacterial and chemical exposures from various sources in the environment. Bacteria and chemicals in air, water, soil and food, occupational exposures and lifestyle factors, all contribute to a complex exposure situation in our daily life.

Due to the complex nature of these exposures, which could have different possible impacts on human health and the environment, there is an immense concern and need to know about the potential effects of interactions of different chemical entities with biological systems. In-depth understanding of interactions between chemical and biological entities is a prodigious task, encompassing important aspects like amount and kind of exposure to the biological entity it is coming in contact with.

Advances in computational approaches in Chemistry and Biology at large have been able to help in understanding how these chemical agents interact with biological systems and the potential effects they have on public health and environment, by elucidating the mechanisms and severity of these interactions.

This thesis focuses on the use of computational approaches, specifically linking the structure of the chemical or biological entity with its activity, to predict outcomes of these interactions based on biological properties as well as chemical properties. Models working on this concept are called Structure Activity Relationship (SAR) models. With the growth in the number of chemical entities whose biological responses need to be known, in different biological scenarios and for different endpoints in a short time, the use of SAR models has become a cost effective way and a basic necessity of toxicological investigations [1].

1.2 SAR models for Engineered Nanomaterials (ENM)

Nanoparticles have distinct properties relative to bulk materials of the same chemical composition due to their small size. With the advent of materials science and nanotechnology, engineered nanomaterials (ENMs) with a diversity of sizes, shapes and compositions can now be synthesized, characterized and applied accordingly with their specific properties. ENMs are currently used in biomedicine, food, electronics, and textiles. They have a wide array of applications in biomedicine that include medical imaging, targeted drug delivery, vaccines, prosthetics, biosensing and other therapies [2-4]. Nanomedicine applications are possible mainly because nanosize facilitates the entrance of

ENMs to almost all systems and units of the body, including cells and organelles.

Gold is one of the materials that has been used in the past for medicinal purposes both in its bulk state and in the form of a nanoparticle. Diseases such as smallpox, measles, skin ulcers and rheumatoid arthritis, to name just a few, have traditionally been treated with gold. Its biocompatibility and inertness was valued in traditional Chinese and Ayurvedic medicine. Gold nanoparticles have been investigated due to their capability to integrate into biological systems (i.e., biocompatibility). This is a result of the thermal, physical and chemical stability of gold nanoparticles. They have also been widely used in diagnosis, surgery and medicine at large as anti-inflammatory, anti-cancer and anti-microbial agent [4-6].

A key challenge for the use of gold or other nanoparticles in medicinal applications is to understand the interactions that occur when nanoparticles are introduced in a biological medium and interact with different types of biomolecules such as proteins, lipids and metabolites [7]. The high reactivity and surface energy of ENMs enhance the interactions with the proteins present in biological media. As a result, nanoparticles are covered by a dynamic layer of proteins that form the so-called protein corona, which has a large impact on bioactivity [8-11].

1.2.1 Protein Corona

Proteins are essential biomolecules made up of one or more polypeptides and with a certain conformation based on the amino acids that form them. Proteins carry a net surface charge depending on the pH of their environment. Initially, when a nanoparticle has just entered in a biological environment, the most abundant proteins get adsorbed onto its surface. The structure and composition of this protein corona depends on the properties of the nanoparticle (e.g., size, shape, composition, surface functional groups, surface charges), on the nature of the biological fluid (e.g., blood, interstitial fluid, cell cytoplasm), as well as on the duration of exposure.

The formation of the corona largely depends on the type of proteins that are adsorbed on the nanoparticle surface. The hard corona is formed by proteins that bind directly onto the surface of the nanoparticle, with a high affinity and larger exchange times, hence making this bonding irreversible. In contrast, the soft corona is formed by reversible bindings of proteins onto the hard corona as a result of weak protein-protein interactions occurring over a short exchange time.

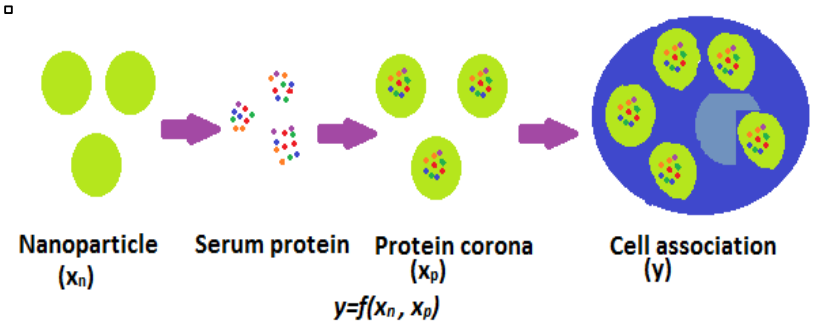

Figure 1: *Prediction of cell association based on the physicochemical properties of nanoparticle and protein corona*

As shown in **Figure 1**, proteins, once adsorbed, provide the nanoparticle with a biological identity that is different from its original synthetic material identity. Biological entities subsequently recognize nanoparticles from their biological identity.

The biological identity of ENMs determines their physiological response including agglomeration, cellular uptake circulation lifetime, signalling, kinetics, transport, accumulation and toxicity [12-14]. It is also known that the formation of the protein corona induces changes in nanoparticle properties such as size, shape and aggregation, and hence configuration that affects the mechanisms and efficiency of their cellular uptake [15,16]. Since the protein corona is the primary representative of the nanoparticle that comes in contact with the cell surface in the physiological environment, its composition plays a crucial role in the biological response. Specifically, cell association is a relevant

endpoint known to be involved in *in vivo* inflammatory responses, biodistribution and toxicity [17-20].

This thesis presents a work on nanoparticles which are likely to interact in real-case application scenarios with mixtures of proteins and biomolecules that will absorb onto their surface forming the so-called protein corona. Information related to the composition of the protein corona and net cell association was collected from literature for a library of surface-modified gold and silver nanoparticles. For each protein in the corona, sequence information was extracted and used to calculate physicochemical properties and statistical descriptors. Data cleaning and pre-processing techniques including statistical analysis and feature selection methods were applied to remove highly correlated, redundant and non-significant features. A weighting technique was applied to construct specific signatures that represent the corona composition for each nanoparticle. Using this basic set of protein descriptors, a new Protein Corona Structure-Activity Relationship (PCSAR) model that relates net cell association with the physicochemical descriptors of the proteins that form the corona was developed and validated.

1.3 Mutagenicity of Chemicals

Mutagenicity is the ability to cause permanent mutations in the DNA sequence. Several years of dedicated scientific research on the introduction of mutations in the DNA caused by interactions

with chemicals have shown a strong connection between mutagenesis and carcinogenesis [21]. Hence, mutagenic substances are part of a class of CMR (carcinogenic, mutagenic and toxic for reproduction) compounds that are emphasized in the European regulation REACH (Registration, Evaluation, Authorization and restriction of Chemicals) [22] due to their irreversible and adverse consequences on human health. REACH encourages scientific innovation and has provisions that facilitate the use of data generated by non-testing methods, specifically of (Q)SAR models (also referred to as *in silico* tools), for ethical and economic reasons [23]. Results obtained by using (Q)SAR tools have the advantage of minimized time, cost, and number of animals needed for testing a substance. Hence, they can be used as an alternative to experimental testing or in a weight-of-evidence approach [24]. Additionally, REACH directives also encourage the use of Integrated Testing Strategies (ITS), which make use of data generated from test batteries to gain a comprehensive information basis for making decisions regarding hazard or risk. As a result, the work described in this part of the thesis is in line with these directives of REACH, using tests of existing data on mutagenicity *in vitro* to predict mutagenicity *in vivo* and establish a link between the *in vitro* and *in vivo* tests. This approach is also bolstered by the fact that the tiered structure for *in vivo* testing incorporates inferences from *in vitro* results [25].

The most commonly used and validated *in vitro* test for mutagenicity is the *Salmonella typhimurium* assay (Ames test) [26, 27], which is primarily used in investigating mutation-

inducing activity and for genotoxicity screening [28]. In the Ames test, frame-shift mutations or base-pair substitutions are detected by exposure of histidine-dependent strains of *Salmonella typhimurium* to the chemical to be tested. When these strains are exposed to a mutagen, reverse mutations restore the capability of the bacteria to synthesize histidine and to grow on a medium deficient in this amino acid [29]. Since the Ames test is a long established and reproducible method offering a broad basis, mutagenicity is one of the most modeled endpoints. In addition, standard protocol has made available abundance of consistent data, which has enabled the development of many *in silico* models based on Ames test.

The reliability of the Ames test has also been acknowledged widely in the literature. Zeiger evaluated the predictivity of the Ames test with respect to data from rodent carcinogenicity from the U.S. National Toxicology Program [30]. According to this evaluation, predictions obtained from Ames test are reliable, and positive chemicals from Ames test can be considered as potential genotoxic carcinogens in rodents. The work also throws light on the capabilities of Ames test, which outperformed a consensus approach with other integrated tests such as chromosome aberration and mammalian cell mutagenicity.

Coming to the *in vivo* tests, as per an assessment carried out by the former European Chemicals Bureau (ECB), there is an immense need of new test-related solutions for studying mutagenicity *in vivo* [31,32]. The micronucleus test in rodents is the most commonly used method to investigate the *in vivo*

mutagenic potential of chemicals following the positive result of *in vitro* Ames test for mutagenicity. Hence, development of (Q)SAR, read-across and grouping of chemicals could be very useful for this endpoint [33,34].

To computationally obtain binary predictions of chemicals as mutagens or non-mutagens, preliminary approaches comprise rules obtained from SAR models. These methods are based on the concept of discovering structural alerts (SA), e.g., molecular fragments that are representative of each toxicity class, based on the dataset of chemicals provided to them. Subsequently, for chemicals for which SAR models are not able to provide binary classifications, biology based methods, which explain the mutagenic or non-mutagenic behavior of chemicals with the help of metabolic triggers (i.e., intermediate and final products into which the chemical is finally broken down) can be used.

In this thesis rules from SAR models and metabolic triggers were made to explain mutagenicity of chemicals *in vitro* and *in vivo*. Hence, a knowledge-based approach combining information from SAR models and metabolic fate of chemicals is presented. Work in this area comprises three parts, all of which are based on SAR rules and metabolic triggers obtained from metabolic fate of chemicals.

In the first part of this work, a model was developed for predict *in vitro* mutagenicity based on Ames test. For this, mutagenicity of chemicals was predicted using newly generated SAR rules. Chemicals predicted as unknowns by SAR rules, were

then predicted using the complementary approach of metabolic triggers. The second part comprises a list of metabolic triggers, obtained from metabolic fate methods, which aid in the prediction of mutagenicity *in vivo*, based on rodent micronucleus assay. Finally, the third part presents two sets of metabolic triggers, with one indicating contrasting mutagenic behaviour of chemicals *in vitro* and *in vivo* and the second indicating non-mutagenic behaviour of chemicals *in vitro* and *in vivo*.

1.4 Resistance of Bacterial Targets to Antibiotics

The discovery of penicillin by Alexander Fleming, in 1928, added a new dimension to the world of medicine. Soon after the discovery of antibiotics, bacterial strains started to develop antibiotic resistance as a result of which the action of antibiotics on life-threatening infections is becoming a cause of major concern [35]. As per estimates, multidrug-resistant S. aureus alone causes more deaths per year than HIV/AIDS [36].

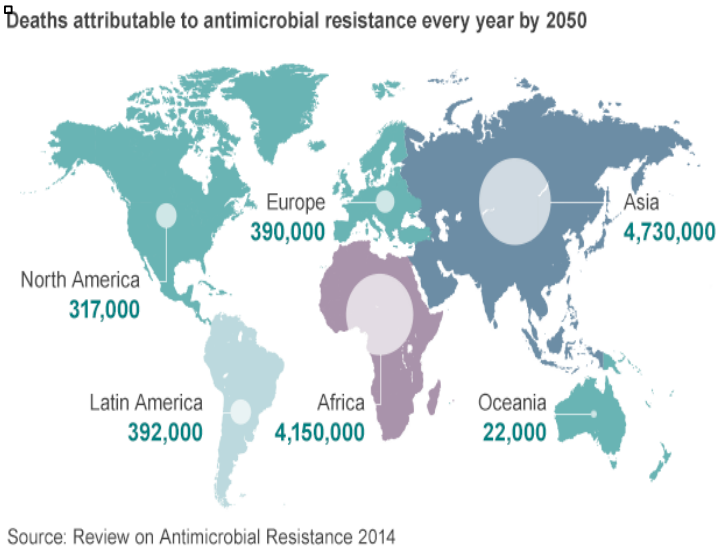

Figure 2: *Number of deaths caused by drug resistance foreseen in 2050 (figure adapted from http://www.bbc.com/news/health-30416844)*

An analysis presented by a famous British economist, Jim O'Neill, attributed that by 2050 death toll caused by antibiotic resistance, which is currently the fourth major cause of death after tetanus, cholera and measles, would be about 10 million (**Figure 2**).

Extensively studied modes of action of antibiotics include interference with cell membrane and cell wall synthesis, inhibition of protein synthesis, interference with nucleic acid synthesis, or inhibition of metabolic pathway [37]. Bacteria develop resistance to these modes of action by:

- Acquiring genes that code for enzymes that destroy the antibiotic prior to its action.

- Attaining efflux pumps that pump out the antibacterial agent prior to its interaction with the target.
- Several genes from a metabolic pathway cause the cell wall to change or mutate porin genes, which restrict the uptake of antibiotics.

Genes responsible for antibiotic resistance can be found in several bacterial genomes, which implies that antibiotic resistance is widespread [38]. In addition to being toxic, antibiotics play an important role in physiology of cell signaling and interaction between species in bacteria [39]. As a consequence, antibiotics render bacterial resistance even without the presence of evolved genes [40]. Hence, it is time to establish a direct relationship between antibiotics and fast evolving bacteria. This is a challenging task, considering a given bacterial genome it is difficult to predict which of the above-mentioned resistance mechanisms will be the first to be manifested by bacteria [41].

Evidences collected from biochemical and genetic studies of bacterial resistance have resulted in the development of *in vitro* and *in silico* methods for the detection of antibiotic resistance. In the current scenario, where development of new antibiotics has almost stagnated [42], and already targeted sites of bacteria are continuously evolving and becoming resistant to antibiotics, the most feasible solution is the search and analysis of new targets in the proteome of these bacteria [43]. Amino acid sequences and physicochemical properties computed from them have been widely used to study different aspects of proteins for varied purposes,

given the fact that they elucidate important information about the protein [44, 45]. In this work, physicochemical properties of known antibiotic-resistant protein sequences were used for understanding resistance in new bacterial proteins.

This thesis presents a consensus model combining the traditionally used Basic Local Alignment Search Tool (BLAST) with a new alignment-free method that has been developed to improve the resistance profile classification of bacterial proteins based on their physicochemical properties. The results show excellent classification into the eight classes of resistance profiles that the bacterial proteins were categorized into. An overall performance similar to the one provided by BLAST was obtained and, in addition, the consensus model has been able to classify all antibiotic-resistant proteins without exceptions.

1.5 Structure of the Thesis

Chapter 2 explains the data sources, the data pruning approaches and the software used for the development of models, along with the metrics used for model evaluation.

Chapter 3 presents the results and discussions for all the models developed in this thesis.

Chapter 4 presents the overall conclusions of the thesis and the models.

1.6 References

1. J.D. McKinney, D. James, A. Richard, C. Waller, C.M. Newman and F. Gerberick (2000) The Practice of Structure Activity Relationships (SAR) in Toxicology. Toxicol. Sci. 56(1), 8-17.

2. P. Bouziotis, D. Psimadas, T. Tsotakos, D. Stamopoulos and C. Tsoukalas (2012) Radiolabeled iron oxide nanoparticles as dualmodality SPECT/MRI and PET/MRI agents. Curr. Top. Med.Chem. 12(23), 2694-2702.

3. N.V. Long, C.M. Thi and M. Nogami (2014) The recent patents and highlights of functionally engineered nanoparticles for potential applications in biology, medicine, and nanomedicine. Current Physical Chemistry 4(2), 173-194.

4. M.M. Joseph and T.T. Sreelekha (2014) Gold nanoparticles - synthesis and applications in cancer management. Recent Patents on Materials Science 7(1), 8-25.

5. T. Jennings and G. Strouse (2007) Past, present and future of gold nanoparticles. Adv. Exp. Med. Biol. 620, 34-47.

6. R. Holiday (2008) Use of gold in medicine and surgery. Biomedical Scientist (The Official Gazette of the Institute of Biomedical science, UK), 2008, 962-963.

7. W.H. Suh, Y.H. Suh and G.D. Stucky (2009) Multifunctional nanosystems at the interface of physical and life sciences. Nano Today 4, 27-36.

8. I. Lynch and K.A. Dawson (2008) Protein-nanoparticle interactions. Nano Today 3(1-2), 40-47.

9. M. Mahmoudi, I. Lynch, M.R. Ejtehadi, M.P. Monopoli, F.S. Bombelli and S. Laurent (2011) Protein-nanoparticle interactions: opportunities and challenges. Chem. Rev. 111(9), 5610-5637.

10. M.P. Monopoli, F.B. Bombelli and K.A. Dawson (2011) Nanoparticle coronas take shape. Nat. Nanotechnol, 6, 11-12.

11. A. Naldoni, L. Mollica and V.D. (2012) Santo Nanoparticle-protein conjugates for nanomedicine applications: design and engineering at the nano-bio interface. Recent Patents on Nanomedicine 2, 17-33.

12. C. Gunawan, M. Lim, C.B. Marquis and R. Amal (2014) Nanoparticle protein corona complexes govern the biological fates and functions of nanoparticles. J. Mat. Chem. B 2, 2060-2083.

13. M. Rahman, S. Laurent, N. Tawil, L. Yahia and M. Mahmoudi (2013) Protein-Nanoparticle Interactions, 1st ed.; Springer-Verlag Berlin Heidelberg.

14. S. Tenzer, D. Docter, S. Rosfa, A. Wlodarski, J. Kuharev, A. Rekik, S.K. Knauer, C. Bantz, T. Nawroth, C. Bier, J. Sirirattanapan, W. Mann, L. Treuel, R. Zellner, M. Maskos, H. Schild, R.H. Stauber (2011) Nanoparticle size is a critical physicochemical determinant of the human blood plasma corona: a comprehensive quantitative proteomic analysis. ACS Nano 5(9), 7155-7167.

15. C.D. Walkey and W.C. Chan (2012) Understanding and controlling the interaction of nanomaterials with proteins in a physiological environment. Chem. Soc. Rev. 41(7), 2780-2799.

16. D. Dutta, S.K. Sundaram, J.G. Teeguarden, B.J. Riley, L.S. Fifield, J.M. Jacobs, S.R. Addleman, G.A. Kaysen, B.M. Moudgil and T.J. Weber (2007) Adsorbed proteins influence the biological activity and molecular targeting of nanomaterials. Toxicol. Sci., 2007, 100(1), 303-315.

17. B. Fadeel (2013) Nanosafety: towards safer design of nanomedicines. J.Intern. Med., 274, 578-580.

18. B.C. Schanen, A.S. Karakoti, S. Seal and D.R. Drake (2009) 3rd; Warren, W.L.; Self, W.T. Exposure to titanium dioxide nanomaterials provokes inflammation of an *in vitro* human immune construct. ACS Nano, 3(9), 2523-2532.

19. M.A. Maurer-Jones, K.C. Bantz, S.A. Love, B.J. Marquis and C.L. Haynes (2009) Toxicity of therapeutic nanoparticles. Nanomedicine, 4(2), 219-241.

20. M.A. Dobrovolskaia (2013) In: Handbook of Immunological Properties of Engineered Nanomaterials; Dobrovolskaia, M.A.; McNeils, S.,Eds.; World Scientific: Singapore, Vol. 1, pp. 547-579.

21. L.D. Claxton, A. Umbuzeiro Gde, and D.M. DeMarini (2010) The Salmonella mutagenicity assay: The stethoscope of genetic toxicology for the 21st century, Environ. Health Perspect. 118, pp. 1515–1522.E.

22. European Union, Regulation (EC) No 1907/2006 of the European and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC,

93/67/EEC, 93/105/EC and 2000/21/EC, Off. J. Eur. Union L136 (2008), pp. 3–280.

23. E. Mombelli and S. Ringeissen (2009) The computational prediction of toxicological effects in regulatory contexts, L'Actualité chimique 335, pp. 52–59.

24. European Chemical Agency, Guidance on information requirements and chemical safety assessment; Chapter R.7a: Endpoint specific guidance. Available at http://echa.europa.eu/web/guest/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment.

25. J. Jaworska and S. Hoffman (2010) Integrated Testing Strategy (ITS) – Opportunities to better use existing data and guide future testing in toxicology, ALTEX. 27, pp. 231–242.

26. B.N. Ames (1979) Identifying environmental chemicals causing mutations and cancer, Science 204, pp. 587–593.

27. B.N. Ames, J. McCann, and E. Yamasaki (1975) Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test, Mutat. Res. 31, pp. 347–364.

28. Organization for Economic Co-operation and Development, OECD guideline for testing of the chemicals, n.471, bacterial reverse mutation test, Commission Regulation (EC) No 440/2008. Available at http://www.oecd.org/chemicalsafety/assessmentofchemicals/1948418.pdf.

29. K. Hansen, S. Mika, T. Schroeter, A. Sutter, A. ter Laak, T. Steger-Hartmann, N. Heinrich, and K.R. Müller (2009) Benchmark data set for *in silico* prediction of Ames mutagenicity, J. Chem. Inf. Model. 49, pp. 2077–2081.

30. E. Zeiger (1998) Identification of rodent carcinogens and noncarcinogens using genetic toxicity tests: Premises, promises, and performance, Regul. Toxicol. Pharmacol. 28, pp. 85–95.

31. F. Pedersen, J. de Brujin, S.J. Munn, and K. Van Leeuwen (2003) Assessment of additional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives, JRC report EUR 20863 EN European Commission JointResearch Centre, Ispra, Italy. Available at http://ec.europa.eu/enterprise/sectors/chemicals/files/reach/testing_needs2003_10_29_en.pdf.

32. K. Van der Jagt, S.J. Munn, J. Torslovand, and J. de Brujin (2004) Alternative approaches can reduce the use of test animals under REACH. Addendum to the Report Assessment of additional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives, JRC

Report EUR 21405 EN. European Commission Joint Research Centre, Ispra, Italy. Available at

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC29111/EUR%202 1405%20EN.pdf.

33. A. Bassan and A.P. Worth (2008) The integrated use of models for the properties and effects of chemicals by means of a structured workflow, QSAR Comb. Sci. 27, pp. 6–20.

34. R. Benigni, C. Bossa, and A. Worth (2010) Structural analysis and predictive value of the rodent *in vivo* micronucleus assay results, Mutagenesis 25, pp. 335–341.

35. S.B. Levy and B. Marshall (2004) Antibacterial resistance worldwide: causes, challenges and responses. Nat. Med., 10, S122–S129.

36. R.M. Klevens, M.A. Morrison, J. Nadle, S. Petit, K. Gershman, S. Ray, L.H. Harrison, R. Lynfield, G. Dumyati, J.M. Townes, A.S. Craig, E.R. Zell, G.E.Fosheim, L.K. McDougal, R.B.Carey and S.K. Fridkin (2007) Invasive Methicillin-Resistant Staphylococcus aureus Infections in the United States. JAMA, 298, 1763–1771.

37. H.C. Neu (1992) The crisis in antibiotic resistance. Science, 257, 1064–1073.

38. V.M. D'Costa, K.M. McGrann, D.W. Hughes and G.D. Wright (2006) Sampling the antibiotic resistome. Science, 311, 374–377.

39. A. Fajardo, and J.L. Martinez (2008) Antibiotics as signals that trigger specific bacterial responses. Curr. Opin. Microbiol., 11, 161–167.

40. J.L. Martinez (2008) Antibiotics and antibiotic resistance genes in natural environments. Science, 321, 365–367.

41. R.Jr.Quintiliani, D. Sahm and P.Courvalin (1998) Mechanisms of resistance to antimicrobial agents. In: Manual of clinical microbiology. 7th ed. American Society for Microbiology, Washington, pp. 1505–1525.

42. J.Carlet, V. Jarlier, S. Harbarth, A. Voss, H. Goossens and D. Pittet (2012) Ready for a world without antibiotics? The Pensières Antibiotic Resistance Call to Action. Antimicrobial Resistance and Infection Control, 1, 11.

43. Q. Li and L. Lai (2007) Prediction of potential drug targets based on simple sequence properties. BMC Bioinformatics, 8, 353.

44. A.N. Sarangi, M. Lohani and R. Aggarwal (2013) Prediction of Essential Proteins in Prokaryotes by Incorporating Various Physico-chemical Features

into the General form of Chou's Pseudo Amino Acid Composition. Protein & Peptide Letters, 2013, 20, 781-795

45. A.A.Toropov, A.P. Toropova, I. Raska Jr., E. Benfenati and G. Gini (2012) QSAR modeling of endpoints for peptides which is based on representation of the molecular structure by a sequence of amino acids. Struct. Chem., 23, 1891–1904.

Chapter 2 Materials and Methods

2.1 Datasets

2.1.1 Nanoparticle dataset used for PCSAR approach

Nanoparticle data was extracted from the literature [1] and a library of 105 gold nanoparticles and 16 silver nanoparticles was built. Nanoparticles were labeled as anionic (57 gold and 7 silver), cationic (27 gold and 5 silver), or neutral (21 gold and 4 silver) depending on the nature of the surface ligands. The dataset also included as endpoint the net cell association values for A549 human lung epithelial carcinoma cells exposed to the above nanoparticles. In addition to charge and net cell association, the dataset also listed 785 proteins, for which the names and accession numbers along with spectral counts were provided.

The same records (i.e., same set of nanoparticles and proteins) used in a previous study [1] were retained to facilitate

model comparison. Briefly, nanoparticle compositions with neutral surface ligands were excluded since serum proteins are not absorbed. Accordingly, models were developed and validated using only those nanoparticles with anionic and cationic surface ligands. Two different models were developed for gold and silver nanoparticles, respectively. Silver nanoparticles were also used to test whether the model built on gold nanoparticles was suitable to predict cell association of nanoparticles with a different core. As for the proteins, only 129 out of 785 proteins in the dataset were quantifiable.

2.1.2 Datasets of chemicals

2.1.2.1 Dataset (based on Ames test) for training

The benchmark dataset developed by Hansen and colleagues was used as a training set which consists of chemicals represented using their canonical simplified molecular input line entry system (SMILES), the outcome of the Ames test (mutagen or non-mutagen) and the corresponding literature references [2, 3]. This dataset comprises data compiled from different sources including, Chemical Carcinogenesis Research Information [4], Helma et al. [5], Kazius et al. [6], Feng et al. [7], VITIC [8], and the GeneTox databases [9] using the Software PipelinePilot [10]. The dataset was preprocessed to remove duplicates (with the same CAS (Chemical Abstracts Service) number, structure and experimental values), salts, mixtures and ambiguous compounds.

After these steps, the number of chemicals retained was 6065 which comprised 3305 (54%) of mutagens and 2760 (46%) of non-mutagens [11]. Furthermore, to ensure the quality of data, the outcome of the Ames test was generated using OECD QSAR Toolbox 3.1.0.21 and Leadscope software to retain only those chemicals that had the same outcome for Ames test in all the three sources, namely Hansen dataset, OECD QSAR Toolbox 3.1.0.21 and Leadscope [12].

Table 1: *Number of chemicals in each dataset used for studying in vitro and in vivo mutagenicity*

Dataset	Mutagen	Non-mutagen
In vitro Ames training set	557	494
In vitro Ames external validation set	42	595
In vivo rodent micronucleus assay dataset (also have responses for *in vitro* Ames test)	202	195

Since the Ames test uses prokaryotic cells that are different from mammalian cells in terms of uptake, metabolism, chromosome structure and DNA repair processes, an exogenous metabolic activation system (i.e., supplemented post-mitochondrial fraction (S9)) is commonly used [2]. For the purpose of analysing the mutagenicity caused by S9 activation, studies with information about the response of the bacterial strains to chemicals before and after S9 activation were selected. To take into account this S9 activation, only chemicals for which studies on the same strain were carried out before and after S9 activation were used. Chemicals that had shown transformation from non-mutagen to mutagen on the same strain for at least one strain

in the same study were categorized as mutagens. On the other hand, chemicals that were found to be non-mutagenic for all strains with and without S9 metabolic activation in the same study were categorized as non-mutagens. Once these steps were completed, 1051 chemicals were retained comprising of 557 (53%) mutagenic and 494 (47%) non-mutagenic chemicals, as shown in **Table 1**.

2.1.2.2 Dataset (based on Ames test) for validation

Data were taken from ECHA CHEM database [13], which comprised 27144 studies on a total of 2975 unique CAS. Data corresponding to the ECHA CHEM database were retrieved from the graphical user interface of the OECD QSAR Toolbox 3.1.0.21 [12]. The database includes chemical substances manufactured or imported in Europe and the information about these is obtained from the registration dossiers, submitted by companies to ECHA in the framework of the European REACH regulation [14]. Pruning criteria were applied to discard information that was not relevant to the study. Firstly, studies which had a Klimisch's code of 1 and 2 (i.e., reliable studies) were only considered [15]. Further, data obtained by applying OECD (Organisation for Economic Co-operation and Development) guideline 471, which corresponds to the bacterial reverse mutation test, were retained [3]. Studies based on multi constituent substances from inorganic origin were discarded to keep only mono constitutent substances the origin of

which was known to be organic. In addition, 221 chemicals, which were already present in the training set, were also excluded.

In the end, as shown in **Table 1**, after applying all the above data pruning criteria only 637 unique chemicals were retained to have a highly imbalanced dataset with 42 (7%) mutagenic chemicals and 595 (93%) non-mutagenic chemicals.

2.1.2.3 Dataset with chemicals evaluated for rodent micronucleus assay and Ames test

This dataset includes chemicals that have been tested following the mammalian erythrocyte micronucleus test in the OECD 474 guideline [16]. The dataset integrates data from the following three data sources, which are present in the OECD QSAR Toolbox 3.1.0.21:

1) Micronucleus ISSMIC from the Istituto Superiore di Sanita, Rome, Italy, and Federal Office of Public Health, Switzerland, which comprises information on 564 chemicals.
2) Micronucleus Oasis from the Laboratory of Mathematical Chemistry, Bourgas, Bulgaria, which comprises information on 557 chemicals.
3) Toxicity Japan MHLW from the Donators Ministry of Health, Labour and Welfare, Japan, which comprises information on 252 chemicals.

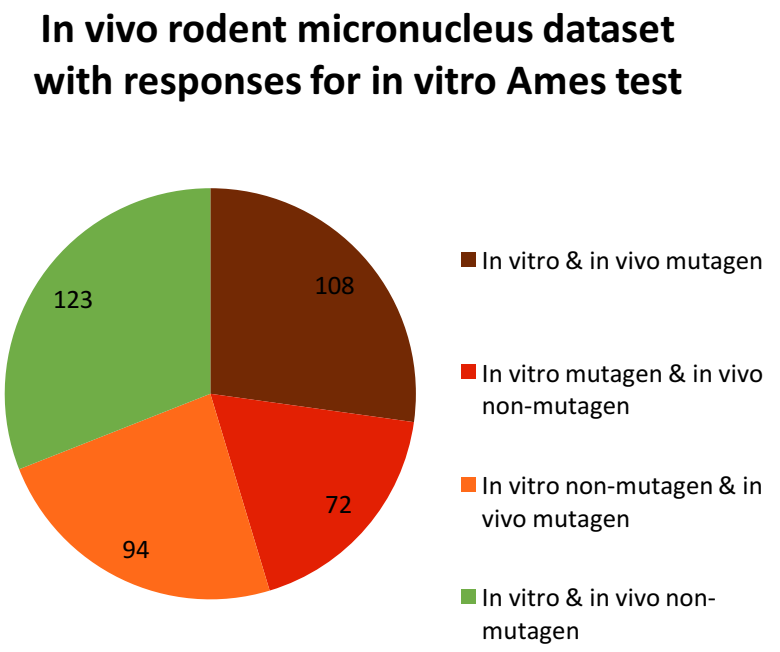

Figure 3: *In vivo rodent micronucleus dataset with responses for in vitro Ames test*

The initial set of chemicals from all the above datasets were put together and their CAS numbers and SMILES were checked to remove chemicals for which studies showed contradictory results. Finally, as shown in **Table 1**, this dataset comprises 397 chemicals, 202 (51%) were mutagenic and 195 (49%) chemicals were non-mutagenic *in vivo*. These chemicals were then compiled to have results for both *in vivo* mutagenicity as per OECD 474 guideline and *in vitro* mutagenicity as per OECD 471 guideline from the ISSCAN v3a database [17].

Figure 3 explains the number of chemicals for which *in vitro* Ames test and *in vivo* test rodent micronucleus assay give the same and opposite mutagenic responses. Of the total number of chemicals, 94 (24%) mutagenic *in vivo* and non-mutagenic *in vitro* chemicals, and 72 (18%) non-mutagenic *in vivo* and mutagenic *in vitro* chemicals were further used to generate metabolic triggers to identify similar and contradictory mutagenic response of chemicals.

2.1.3 Dataset with resistance profiles for antibiotics

Data on known antibiotic-resistant genes were retrieved from the Antibiotic Resistance Database (ARDB) [18], which is a rich source that includes information about antibiotic-resistant genes and their corresponding resistance profiles. The resistance profile of a given gene is defined by the set of antibiotics to which the gene is resistant. ARDB encompasses information on 13293 genes belonging to 3369 different species and showing 377 types of resistance profile for 257 antibiotics. The protein products of these antibiotic-resistant genes are also given along with information on their resistance profile. The relationships between protein sequences and antibiotics in the database are given in the form of resistance types. In the current work, protein names corresponding to each antibiotic-resistant gene together with the corresponding resistance type were collected from ARDB. For each protein sequence, data were retrieved from the National Center for Biotechnology Information (NCBI) Batch Entrez

service (http://www.ncbi.nlm.nih.gov/sites/batchentrez). 1315 similar/homologous sequences were removed using CD-HIT with a 90% sequence identity threshold to avoid undesirable biases [19, 20].

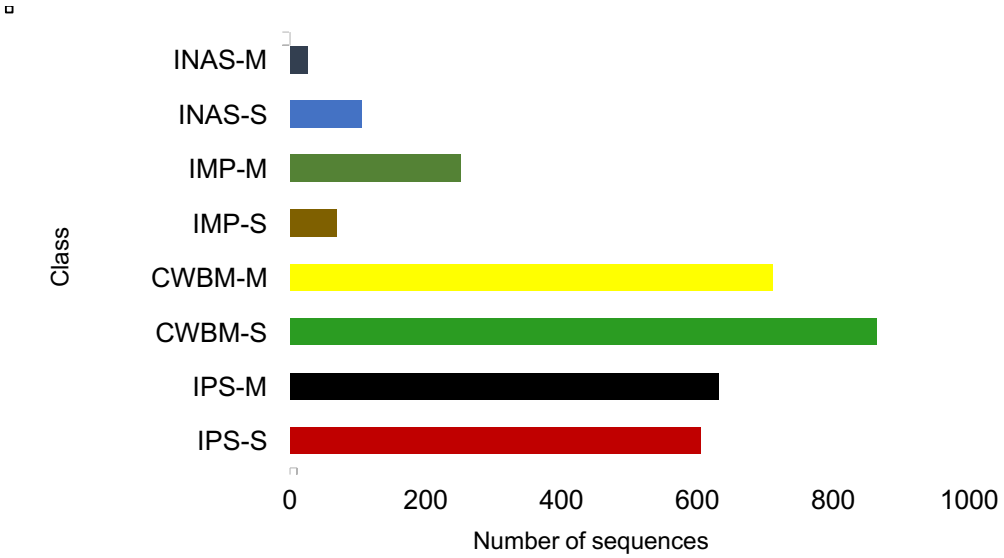

Figure 4: *Distribution of protein sequences across classes in the dataset for studying antibiotic resistance in bacteria*

Each one of the 3263 observations in the dataset was associated with a specific type of resistance. Finally, resistance types were categorized into one of eight classes, depending on the number and type of antibiotics included in the resistance profile.

Depending on the mode of action of the antibiotics that made up the resistance type, each observation was assigned one of the following four labels:

- Inhibition of Protein Synthesis (IPS)
- Interference with Cell Wall synthesis and disruption of Bacterial Membrane structure (CWBM)
- Inhibition of Metabolic Pathway (IMP)
- Interference with Nucleic Acid Synthesis (INAS)

Furthermore, resistance profiles containing a single class of antibiotics were labelled as S (single) and the others were labelled as M (multiple). Combining these two labels with the above modes of action, a total of eight different classes were obtained. The distribution of antibiotic-resistant protein sequences in each class is shown in **Figure 4**. As it can be observed, the dataset is highly imbalanced in terms of the number of records in each class, being CWBM-S the majority class with 863 antibiotic-resistant proteins, while the class INAS-M is the smallest one containing only 26 proteins.

2.2 Software

2.2.1 R computing environment

R is a programming language and environment for statistical computing and graphics. Lot of statistical and graphical techniques are incorporated in R such as linear and nonlinear

modelling, classical statistical tests, time-series analysis, classification, clustering, etc., which have made it widely used software for data analysis and modelling [21].

2.2.2 MultiDendrograms

MultiDendrograms is an easy to use and powerful program which is used for hierarchical clustering of data. With distances computed from similarity matrix as a starting point, a dendrogram is generated using the most common agglomerative hierarchical clustering algorithms, allowing many of the graphical representation parameters to be tuned. In addition, the results may be easily exported to file [22, 23].

2.2.3 SARpy

SARpy (SAR in python) is a tool that identifies the most informative fragments from the chemical substances in a given dataset [24]. The software facilitates the generation of rules from the data without needing a priori knowledge. The algorithm generates a set of substructures of arbitrary complexity, from which the ones that result in a better prediction performance in the training set are automatically chosen as SAs. SARpy is available either through a standalone interface or via the VEGA web-based platform (http://home.deib.polimi.it/gini/SARpy.htm).

The basic information used by SARpy for rule generation is the molecular structure, which is converted into SMILES

disregarding chirality information. The resulting rule-based model tags each compound either as mutagen when one or more SAs appear in the molecular structure, or as a non-mutagen if no SAs are present, because SAs are characteristic of mutagens. Prediction uncertainty is not provided by the model but it can be easily deduced from the likelihood ratio of each SA present in a given chemical [25].

2.2.4 Chemical Reactivity and Fate Tool (CRAFT)

The CRAFT software suite was used to generate reactions for the fate of chemicals. CRAFT is used in areas such as product safety, hazard and risk assessment, and toxicology to interactively evaluate the reactivity, persistence, biodegradability and fate profiles of chemicals in the environment [26]. The software provides the metabolic pathways of chemicals by evaluating their reactivity and fate, along with their conceivable products based on different conditions and organisms. In addition, each step of the metabolic pathway is ranked according to its likelihood. Metabolic reactions were generated using CRAFT Explorer, which includes the UM-BBD likelihood model and the Ester hydrolysis sample model. The UM-BBD likelihood model is an implementation of the biotransformation rules provided in the University of Minnesota Biocatalysis/Biodegradation Database [27]. Most of the biotransformation rules are taken from the above database and are implemented with their exact likelihood. On the other hand, the

Ester hydrolysis sample model comprises only one reaction rule in order to demonstrate (Q)SAR based reactivity model usage [28].

2.2.5 istChemFeat

Functional groups and atom centred fragments were computed using the software istChemFeat, which is a JAVA application based on the VEGA core libraries developed under the EU funded project ANTARES (http://www.antareslife.eu) by Kode s.r.l. (http://kode-solutions.net). Using this software, the chemical space of the datasets as well as the metabolic triggers that were generated using CRAFT were analysed.

2.2.6 Emboss-pepstats

The implementation of pepstats in Emboss, available at (http://emboss.bioinformatics.nl/cgi-bin/emboss/pepstats), was used. pepstats is a program that computes properties of proteins, and it was used to obtain a set of features that describe the protein sequences corresponding to the antibiotic-resistant genes. Typical output of pepstats consists of values for peptide properties such as molecular weight, charge, isoelectric point, probability of expression in inclusion bodies, and A280 molar extinction coefficient. The counts of each of the 20 amino acids are represented as numbers, molar percentage and DayhoffStat. DayhoffStat corresponds to the molar percentage of the respective amino acids divided by their Dayhoff Statistic. In addition, each amino acid is counted, depending on its nature, into tiny, small,

aliphatic, aromatic, non-polar, polar, charged, basic, and acidic, which are represented in terms numbers and molar percent [29].

2.2.7 Rapidminer

RapidMiner is a data mining and machine learning software that facilitates visualization, predictive analytics and statistical modelling, evaluation and deployment [30].

2.2.8 Waikato Environment for Knowledge Analysis(WEKA)

Weka machine learning is a workbench that provides an environment for automatic classification, regression, clustering and feature selection in common data mining problems. It comprises of a graphical interface to a wide number of machine learning algorithms and data pre-processing methods that can be used for data exploration and the experimental comparison of different machine learning techniques on the same problem. Weka can process data given in the form of a single relational table. Its main objectives are to assist users in extracting useful information from data and to enable them to easily identify a suitable algorithm for generating an accurate predictive model from it [31].

2.2.9 BioEdit

BioEdit is a biological sequence alignment editor written only for Windows. It has multiple document interface with

user-friendly features that make alignment and manipulation of sequences easy. With different sequence manipulation and analysis options available along with links to external analysis programs, sequence manipulation is absolutely easy [32].

2.3 Feature generation, selection and model development

2.3.1 For PCSAR model

Information about the 129 quantifiable proteins that form the corona was recorded in the form of spectral counts, which were then converted to relative abundance (RA) using the following equation:

$$RA(n,p) = \frac{SpC(n,p)}{\sum_{k=1}^{129} SpC(n,k)} \qquad (1)$$

where $RA(n, p)$ is the relative abundance of protein p in the nanoparticle formulation n, and $SpC(n, p)$ corresponds to the number of spectral counts recorded for a nanoparticle formulation n and protein p, respectively. The sum of the relative abundances for all proteins over a given nanoparticle formulation is 1. Relative abundances of the proteins were used in a previous study [1] to define a quantitative RA-based fingerprint for each nanoparticle formulation.

The approach in this thesis develops a new type of protein corona fingerprint that is based on physicochemical properties. Protein properties were computed using the EMBOSS Pepstats program from their amino acid sequences. The final descriptors were computed averaging the physicochemical properties of the proteins weighted by the relative abundance of the corresponding protein:

$$AP(n,d) = \sum_{p=1}^{129} RA(n,p) \times Pepstats(p,d) \qquad (2)$$

where $AP(n, d)$ is the averaged value of a physicochemical descriptor d for a nanoparticle formulation n, and $Pepstats(p, d)$ is the value of a physicochemical descriptor d for protein p. Averaged physicochemical descriptor (AP) were normalized by using a z-score transformation (i.e., subtracting the mean value and dividing by the standard deviation of descriptor values). Finally, the normalized physicochemical descriptors were used to form a new AP-based fingerprint (i.e., vector of averaged descriptor values), independent of the specific protein composition of the biological media.

To assess whether the two different fingerprints for nanoparticles carried essentially the same information, hierarchical clustering of nanoparticles was performed using the MultiDendrograms software. The partitions (i.e., clusters) obtained from the two alternative fingerprints were compared. The distance between each pair of nanoparticles, d_{ij}, was computed from Pearson correlation coefficient, r_{ij}, using the metric:

$$d_{ij} = \sqrt{1 - r_{ij}} \qquad (3)$$

PCSAR based on multilinear regression were developed to predict net cell association. The selection of the best set of features (i.e., averaged protein descriptors) for the PCSAR was based on the adjusted correlation coefficient, which measures if the addition of a new descriptor increases the explanatory power of the resulting model.

2.3.2 For chemicals evaluated with Ames test and Rodent micronucleus test

The data processing workflow used to develop the *in vitro* mutagenicity model is based on the use of the rules generated by SARpy. Chemicals were evaluated against the ruleset and tagged as mutagenic, non-mutagenic or unknown. Metabolic pathways were subsequently generated using CRAFT for those chemicals identified as unknown by SARpy rules. Reactions were generated in the same way as in the case of metabolic triggers, i.e. using a likelihood threshold of 0.61 which is value that includes only likely and very likely reactions. All possible reaction steps under aerobic and biotic conditions were computed.

From the resulting reactions, a set of unique metabolic triggers was generated by enumerating the occurrence of each metabolic trigger in the metabolic pathways of mutagenic and non-mutagenic chemicals. Parent compounds, intermediates and products of chemicals which were labeled unknown by SARpy

were then compared to the metabolic triggers. Based on the presence of these metabolic alerts as parent compounds, intermediates or products, the parent compounds were classified as mutagenic or non-mutagenic.

2.3.3 For drug resistance model

pepstats was used to generate protein descriptors which were subsequently the features of the models. In order to partition the data into training and validation sets, the most representative bacterial proteins were used to evaluate the classifier, while the remaining observations were kept for training it. To this end, the original dataset of 3263 proteins was partitioned into validation and training sets using the RapidMiner implementation of Kennard and Stone's algorithm [33]. The construction of the validation set is incremental and starts by incorporating the most dissimilar proteins. Further, for the next candidate proteins, the distance with the nearest protein that has already been selected is calculated, and the protein at the largest distance among the small distances is chosen. The result of this exercise was a validation set that consisted of 654 representative proteins chosen from the original dataset of 3263 proteins. The remaining 2609 proteins were used for training.

In a subsequent preprocessing step, aimed to select the most representative physicochemical properties, the data values in the training set were standardized by using a z-score transformation (i.e., subtracting the mean and dividing by the

standard deviation). Normalized features were then ranked using WEKA, which has been used widely for different classification problems [31, 34-36]. For ranking of our normalized features we used the Infogain method, which ranks features in terms of the amount of information they contribute towards correct classification. For every particular feature, Infogain measures the amount of information this feature gives to the prediction of classes and the reduction in entropy (uncertainty associated with a random feature). This feature selection method has been extensively used for classification problems in biology [37-38].

Once the representative proteins were selected to be part of the validation set, five different data partitions for training were randomly generated to avoid bias in the selection of the features due to the structure of the training set. For all five training sets, different classification models were incrementally developed by adding a new feature at a time, following the decreasing order of significance given by Infogain. This wrapper approach [39] has the purpose of evaluating the performance of a classifier with different subsets of features to identify the most suitable subset.

An alignment-free classifier using the K* algorithm [40] to discriminate between eight different resistance profiles was developed. The algorithm is an instance-based classifier that associates a given bacterial protein with the resistance type of the most similar protein in a training set, and the similarity is based on an entropy distance function.

Finally, the five classifiers were tested on the same validation set and the responses of these classifiers on the validation set were used to compute a consensus classifier, which assigned a class to every observation in the validation set based on the majority response of the five classifiers.

BLAST [41] is the most common approach for identifying sequence-based homologies. However, alignment-based methods require the *a priori* definition of a minimum similarity threshold to recognize homologous sequences. As a result, protein sequences with similarity values below the threshold cannot be classified. In the current work, a local BLAST implementation from BioEdit was used to develop a baseline resistance profile classifier. Homologous sequences were identified with a similarity threshold of 0.001 [42, 43]. Accordingly, bacterial protein sequences were assigned to the resistance profile of the most similar bacterial protein identified by BLAST based on the best bit-score and E-value.

2.4 Performance evaluation of models

2.4.1 For PCSAR model

The PCSAR was subsequently validated using the R statistical programming framework via the following methods:

- Bootstrapping cross-validation with 1000 bootstrap samples, using the 0.632 method [44]. This approach corrects

performance estimates by taking into account the probabilities for each observation to be included in the training and test sets. The correction is based on a weighted average of the re-substitution and bootstrap error estimates.

- Leave-one-out (LOO) cross-validation, where the model is iteratively trained on all observations except one that is used for testing.
- Leave-many-out (LMO) cross-validation, where the dataset was randomly partitioned into two parts, 75% of observations for training and 25% for testing. This partitioning and subsequent training and testing was repeated 100 times.
- 10-fold cross-validation repeated 10 times.

In all the above validation methods, the number of repetitions was optimized to ensure stable performance metrics.

2.4.2 For model based on Ames test

Prediction results were evaluated using traditional Cooper statistics, following the QSAR characterization guidance developed by the Joint Research Centre [45, 46]. Accordingly, accuracy, sensitivity and specificity are defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} \quad (4)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (6)$$

where *TP*, *TN*, *FN* and *FP* represent the number of true positives, true negatives, false negatives and false positives, respectively. In the current analysis, mutagenic compounds form the positive class. Accordingly, true positive instances refer to mutagenic chemicals predicted mutagenic and true negative instances refer to non-mutagenic chemicals predicted non-mutagenic, whereas false negative instances refer to mutagenic chemicals predicted as non-mutagenic and false positive instances refer to non-mutagenic chemicals predicted as mutagenic. It is important to note that, for regulatory purposes, sensitivity is more important than specificity since it is crucial not to consider as safe a toxic chemical.

In order to get a more reliable metric, the above performance measures were complemented with the Matthews' correlation coefficient (MCC). The MCC metric is well suited for skewed (i.e., imbalanced) datasets used for binary classifications [47]. The MCC is computed as follows:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \qquad (7)$$

Possible values of MCC are in the range of -1 to +1, where +1 indicates perfect prediction, 0 indicates a random prediction, and -1 indicates total disagreement between predicted and real values.

2.4.3 For drug resistance model

The model to predict the drug resistance of protein targets was evaluated using the following metrics:

- Given the data imbalance (see **Figure 4**), the κ index [48] was used to assess the performance of the classifiers. The κ index takes into consideration the chance agreement of the observation, rather than weighing classes, based on the number of observations they have [49]. This index varies from 0 (no agreement at all) to 1 (complete agreement).
- Precision (P), this is the fraction of antibiotic-resistant proteins belonging to a given class, which are predicted into the same class, and all antibiotic-resistant proteins belonging to the given class.
- Recall (R), this is the fraction of antibiotic-resistant proteins belonging to a given class, which are predicted into the same class, and all the antibiotic-resistant proteins predicted into the given class and other classes.
- F-measure, which measures the goodness of a classifier in the presence of rare classes:

$$F - \text{measure} = \frac{2PR}{P+R} \qquad (8)$$

For PCSAR and drug resistance models Y-randomization was performed, which consists in the assessment of the performance of the corresponding model, built on randomized

or permuted data. This is used to discard the effect of chance correlations in predictions made by the model.

2.5 References

1. C.D. Walkey, J.B. Olsen, F. Song, R. Liu, H. Guo, D.W.H. Olsen, Y. Cohen, A. Emili and W.C.W. Chan (2014) Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. ACS Nano 8(3), 2439-2455.

2. Organization for Economic Co-operation and Development, OECD guideline for testing of the chemicals, n.471, bacterial reverse mutation test, Commission Regulation (EC) No 440/2008. Available at http://www.oecd.org/chemicalsafety/assessmentofchemicals/1948418.pdf.

3. K. Hansen, S. Mika, T. Schroeter, A. Sutter, A. ter Laak, T. Steger-Hartmann, N. Heinrich, and K.R. Müller (2009) Benchmark data set for *in silico* prediction of Ames mutagenicity, J. Chem. Inf. Model. 49, pp. 2077–2081.

4. Chemical Carcinogenesis Research Information System, NCRI informatics Initiative Homepage 2009. Available at http://www.cancerinformatics.org.uk/matrix/CCRIS.htm.

5. C. Helma, T. Cramer, S. Kramer, and L.D. Raedt (2004) Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds, J. Chem. Inf. Comput. Sci. 44, pp. 1402–1411.

6. J. Kazius, R. McGuire, and R. Bursi (2005) Derivation and validation of toxicophores for mutagenicity prediction, J. Med. Chem. 48, pp. 312–320.

7. J. Feng, L. Lurati, H. Ouyang, T. Robinson, Y. Wang, S. Yuan, and S.S. Young (2003) Predictive toxicology: Benchmarking molecular descriptors and statistical methods, J. Chem. Inf. Comput. Sci. 43, pp. 1463–1470.

8. P.N. Judson, P.A. Cooke, N.G. Doerrer, N. Greene, R.P. Hanzlik, C. Hardy, A. Hartmann, D. Hinchliffe, J. Holder, L. Müller, T. Steger-Hartmann, A. Rothfuss, M. Smith, K. Thomas, J.D. Vessey, and E. Zeiger (2005) Towards the creation of an international toxicology information centre, Toxicology 213, pp. 117–128.

9. Genetic Toxicity, Reproductive and Developmental Toxicity, and Carcinogenicity Database; 2009. Available at http://www.fda.gov/AboutFDA/CentersOffices/CDER/ucm092217.htm.

10. Accelrys Inc., A.S. Scitegic Pipeline Pilot, Version 7.0; 2009; software available at http://accelrys.com/products/pipeline-pilot/.

11. N.G. Bakhtyari, G. Raitano, E. Benfenati, T. Martin, and D. Young (2013) Comparison of *in silico* models for prediction of mutagenicity, J. Environ. Sci. Health C, Environ. Carcinog. Ecotoxicol. Rev. 31, pp. 45–66.

12. The OECD QSAR toolbox for grouping chemicals into categories; software available at http://www.qsartoolbox.org/.

13.European Chemicals Agency. Registered substances, 2014. Available at http://echa.europa.eu/web/guest/information-on-chemicals/registered-substances.

14. E. Mombelli and S. Ringeissen (2009) The computational prediction of toxicological effects in regulatory contexts, L'Actualité chimique 335, pp. 52–59.

15. H.J. Klimisch, M. Andreae, and U. Tillmann (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data, Regul. Toxicol. Pharmacol. 25, pp. 1–5.

16. Organization for Economic Co-operation and Development guideline for testing of the chemicals,n.474, Mammalian Erythrocyte Micronucleus Test, OECD Publishing, Paris Section 4, 1997. Available at http://www.oecd.org/chemicalsafety/risk-assessment/1948442.pdf

17. R. Benigni, C. Bossa, A.M. Richard, and C. Yang (2008) A novel approach: Chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity, Ann. Ist. Super Sanita` 44, pp. 48–56.

18. B.Liu, and M. Pop (2009) ARDB–antibiotic resistance genes database. Nucleic Acids Res., 37, D443–D447.

19. L. Weizhong and A. Godzik (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22, 1658–1659.

20. Y. Huang, B. Niu, Y. Gao, L. Fu and W. Li (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics, 26, 680–682.

21. R Core Team. R: a language and environment for statistical computing. R

Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org (Accessed March 2, 2015).

22. S. Gómez, A. Fernández, C. Granell, and A. Arenas (2013) Structural patterns in complex systems using multidendrograms. Entropy, 15, 5464-5474.

23. A. Fernández and S. Gómez (2008) Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. J. Classif., 25, 43-65.

24. T. Ferrari, G. Gini, N.G. Bakhtyari, and E. Benfenati (2011) Mining Structural Alerts from SMILES: A new way to derive structure activity relationships, CIDM—IEEE Symposium Series on Computational Intelligence, Paris.

25. TOPKAT, version 3.1, User Guide, Accelrys software Inc.: San Diego, CA, USA; software available at http://accelrys.com.

26. CRAFT, developed by Molecular Networks on behalf of the European Commission's Joint Research Centre, Italy, 2008; software available at http://www.molecularnetworks.com/products/craft.

27. L.B.M. Ellis, D. Roe, and L.P. Wackett (2006) The University of Minnesota biocatalysis/biodegradation database: The first decade, Nucleic Acids Res. 34, pp. 517–521.

28. D.M. Bender, J.A. Peterson, J.R. McCarthy, H. Gunaydin, Y. Takano, and K.N. Houk (2008) Cyclopropanecarboxylic acid esters as potential prodrugs with enhanced hydrolytic stability, Org. Lett. 10, pp. 509–511.

29. P. Rice, I. Longden and A. Bleasby (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet., 16(6), 276-277.

30. H. Markus and R. Klinkenberg (2013) Rapidminer: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC.

31. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten (2009) The WEKA data mining software: an update. SIGKDD explorations, Vol. 11.

32. T.A. Hall (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, Vol. 41 (1999), pp. 95-98

33. R.W. Kennard and L.A. Stone (1969) Computer Aided Design of Experiments. Technometrics,Vol. 11, No. 1, pp. 137-148

34. O. Ivanciuc (2008) Weka Machine Learning for Predicting the Phospholipidosis Inducing Potential.Curr Top Med Chem. 2008;8(18):1691-709.

35. K. Wang, X. Hu, Z. Wang and A. Yan (2012) Classification of acetylcholinesterase inhibitors and decoys by a support vector machine.Comb Chem High Throughput Screen. 2012 Jul;15(6):492-502.

36. K. Kadam, P. Prabhakar and V.K. Jayaraman (2012) SVM Prediction of Ligand-binding Sites in Bacterial Lipoproteins Employing Shape and Physio-chemical Descriptors. Protein Pept Lett. 2012 Nov;19(11):1155-62.

37. M. Zheng, X. Luo, Q. Shen, Y. Wang, Y. Du, W. Zhu and H. Jiang (2009) Site of metabolism prediction for six biotransformations mediated by cytochromes P450. 25, 1251–1258.

38. A.K. Mishra and D.K. Lobiyal (2011) miRNA prediction using computational approach software tools and algorithms for biological systems. In: Advances in Experimental Medicine and Biology, Vol. 696, pp 75–82.

39. R. Kohavi and G.H. John (1997) Wrappers for feature subset selection. Volume 97, Issues 1–2, December 1997, Pages 273–324

40. J.G. Cleary and L.E. Trigg (1995) K*: An Instance-based Learner Using an Entropic Distance Measure.In Proceedings of the 12th International Conference on Machine Learning.

41. S.F.Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410.

42. J. Ramana and D. Gupta (2010) Machine Learning Methods for Prediction of CDK-Inhibitors. PLoS ONE 5, e13357

43. R. Kumar, B. Panwar, J.S. Chauhan and G.P.S. Raghava (2011) Analysis and prediction of cancerlectins using evolutionary and domain information.BMC Research Notes 2011, 4:237

44. B. Efron and R. Tibshirani, R (1997) Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc., 92(438), 548-560.

45. J.A. Cooper, R. Saracci, and P. Cole (1979) Describing the validity of carcinogen screening tests, Br. J. Cancer 39, pp. 87–89.

46. A.P. Worth, A. Bassan, A. Gallegos, T.I. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska, and M. Vracko, The characterisation of (Quantitative) Structure-Activity Relationships: Preliminary guidance, EUR 21866 EN. Available at http://ihcp.jrc.ec.europa.eu/ourlabs/predictivetoxicology/information-

sources/qsar-document-area/QSARcharacterisationEUR21866EN.pdf.

47. P. Dao, K. Wang, C. Collins, M. Ester, A. Lapuk, and S.C. Sahinalp (2011) Optimally discriminative subnetwork markers predict response to chemotherapy, Bioinformatics 27, pp. 205–213.

48. J. Cohen (1960) A coefficient of agreement for nominal scale. Educational and Psychological Measurement 20 (1): 37–46.

49. M. Fatourechi, R.K. Ward, S.G. Mason, J. Huggins, A. Schlögl and G.E. Birch (2008) Comparison of evaluation metrics in classification applications with imbalanced datasets. In: Seventh International Conference on Machine Learning and Applications. ICMLA'08, pp. 777–782.

UNIVERSITAT ROVIRA I VIRGILI
IN SILICO MODELING OF CHEMICAL AND BIOLOGICAL INTERACTIONS AT DIFFERENT SCALES
Padmaja Balachandran Kamath

Chapter 3 Results and Discussions

3.1 PCSAR approach

In this section the results of the PCSAR approach are presented, wherein the fingerprints obtained from the literature AP-based Fingerprints and the fingerprints generated using RA-based Fingerprints were compared. To summarize the approach in a nutshell, information related to the composition of the protein corona and net cell association was collected from literature for a library of surface-modified gold and silver nanoparticles. For each protein in the corona, sequence information was extracted and used to calculate physicochemical properties and statistical descriptors. Data cleaning and preprocessing techniques including statistical analysis and feature selection methods were applied to remove highly correlated, redundant and non-significant features. A weighting technique was applied to construct specific signatures that represent the corona composition for each nanoparticle. Using this basic set of protein descriptors, a new PCSAR that relates net

cell association with the physicochemical descriptors of the proteins that form the corona was developed and validated.

3.1.1 Comparison of fingerprint-based partitions

To understand the similarity between the clusters obtained from the two descriptions of nanoparticles, a partition of ten clusters obtained from the RA-based fingerprints was compared with a partition of seven clusters generated from the AP-based fingerprints.

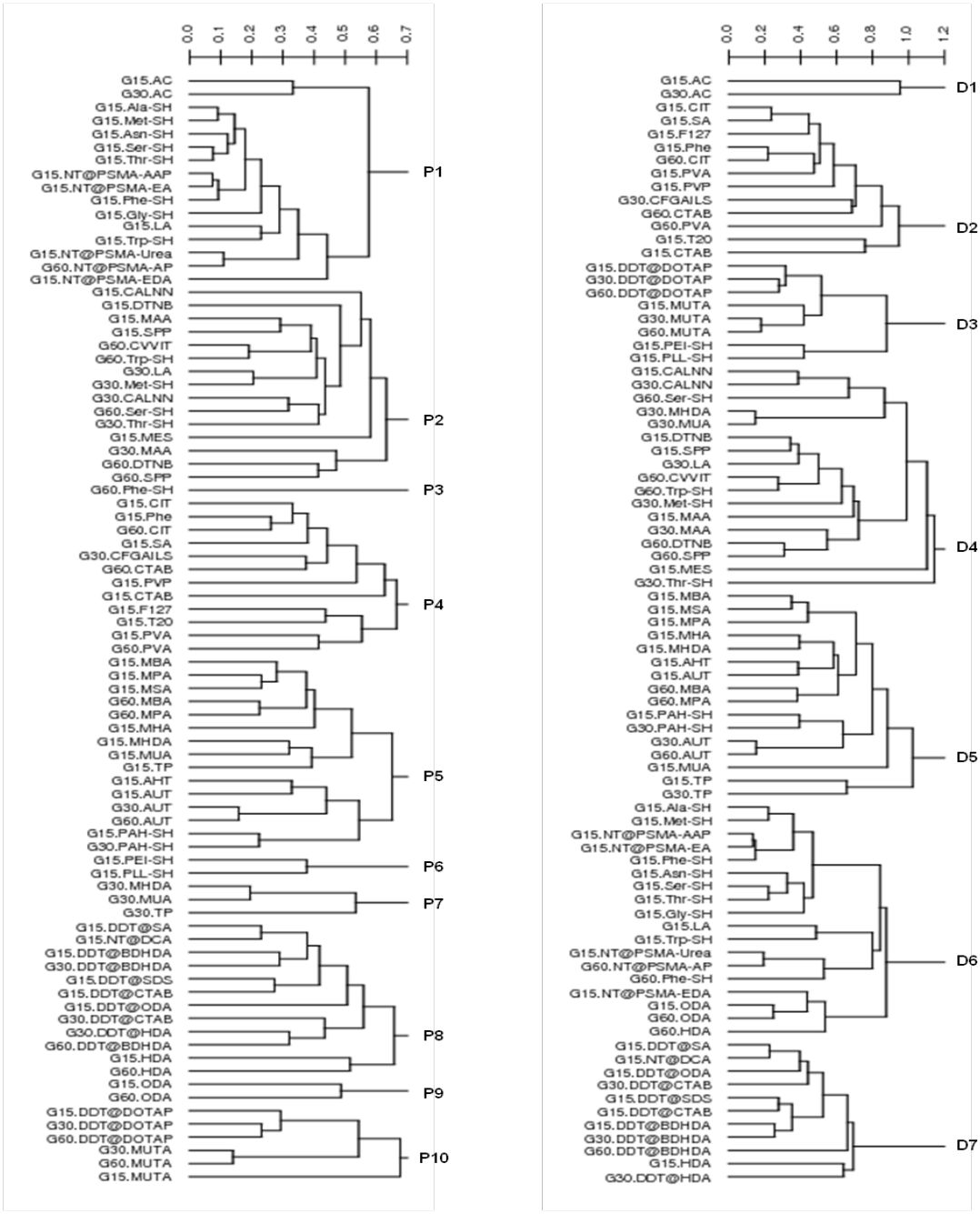

Figure 5: *Hierarchical clusterings obtained from two different descriptions of nanoparticles, i.e., in terms of proteins relative abundance (left); and in terms of physicochemical descriptors of the protein corona (right). The dendrograms were computed using the MultiDendrograms software and they were cut at heights (i.e., similarity level) where the correspondence between the two partitions obtained is maximized.*

Table 2: *Contingency table comparing ten clusters obtained from fingerprints based on proteins relative abundance, and seven clusters obtained from fingerprints based on physicochemical descriptors of the protein corona.*

		Averaged Physicochemical Descriptors (AP-Based Fingerprints)						
		D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
Proteins Relative Abundance (RA-Based Fingerprints)	**P1**	2	0	0	0	0	14	0
	P2	0	0	0	15	0	0	0
	P3	0	0	0	0	0	1	0
	P4	0	12	0	0	0	0	0
	P5	0	0	0	0	15	0	0
	P6	0	0	2	0	0	0	0
	P7	0	0	0	2	1	0	0
	P8	0	0	0	0	0	1	11
	P9	0	0	0	0	0	2	0
	P10	0	0	6	0	0	0	0

Figure 5 shows the two dendrograms obtained using the unweighted average method of hierarchical clustering. The resulting contingency table is given in **Table 2**. In most of the partitions there is a neat overlap between the clusters obtained from the two fingerprints. For example, the nanoparticles in cluster P4 are exactly the same as those in cluster D2 (P4 = D2). Also, cluster D3 matches exactly with the union of clusters P6 and P10 (D3 = P6 ∪ P10). Significant overlaps are found between clusters P1 and D6, and between the following pairs of clusters, where one

cluster strictly contains the other: P2 ⊂ D4, P5 ⊂ D5, and P8 ⊃ D7. The normalized mutual information index for the two partitions was 0.88, confirming that RA-based fingerprints and AP-based fingerprints convey similar information up to a large extent.

The model for gold nanoparticles is a multilinear regression model is based on fingerprints developed from the relative abundance and physiochemical properties of the proteins that form the corona. This model obtained using the adjusted correlation coefficient for feature selection, retained only 7 out of the 35 initial averaged physicochemical properties:

$$\begin{aligned} \log_2(\textit{net cell association}) \;=\; & \\ & -4.56 \\ & +4.92 \times \textit{probability of expression in inclusion bodies} \\ & +1.32 \times \textit{tiny amino acids percentage} \\ & +1.04 \times \textit{basic amino acids percentage} \\ & +0.93 \times \textit{aspartic acid DayhoffStat} \\ & -0.86 \times \textit{molecular weight} \\ & -1.24 \times \textit{polar amino acids percentage} \\ & -3.80 \times \textit{acidic amino acids percentage} \end{aligned} \tag{1}$$

The normalized AP-fingerprints were used to develop the model in Eq. 1.

Table 3: *Comparison of the performance of the two net cell association predictive models. Squared correlation coefficient values are given for the entire dataset, leave-one-out (LOO), and leave-many-out (LMO25%) cross-validations.*

Model	No. of parameters	R^2	R^2_{LOO}	$R^2_{LMO\ 25\%}$
RA-based fingerprint	64	0.93	0.81	0.61 ± 0.18
AD-based fingerprint	7	0.80	0.76	0.72 ± 0.11

Table 3 compares the performance of this model (Eq. 1) with the previously reported model [1] developed from RA-based fingerprints. Although the performance of the model based on RA-based fingerprints is very high for the entire dataset (R^2 =0.93), it can be observed that there is a significant variability in the LOO and LMO cross validations. Specifically, a substantial decay is observed in the LOO cross-validation, where the squared correlation coefficient (R^2 LOO) drops to 0.81. A similar decrease in performance is observed in the LMO cross-validation, where the value of R^2 LMO 25% decreases until 0.61. In contrast, the current model (Eq. 1) developed from AP-based fingerprints shows a more consistent performance with R^2 values of 0.80, 0.76 and 0.72 for the entire dataset, the LOO and the LMO 25% cross-validations, respectively. Literature suggests that models with $R^2 > 0.70$ for LOO cross-validation can be considered to be acceptable [2]. The number of 7 parameters of the present model is ten-fold lower than the number of 64 parameters needed for the model based on

RA-based fingerprints. OECD guidelines for QSAR development and validation suggest that simple relationships are preferred to more complex ones since they are easier to construct, interpret and use.

Additionally, for bootstrap cross-validation tests, the new PCSAR model did not show a significant decay in performance. The results of the 0.632 bootstrapping were well in agreement, with a R^2 value of 0.77 ± 0.07 after 1000 bootstrap samples. A similar R^2 value of 0.77 ± 0.14 was obtained with a 10-fold cross-validation repeated 10 times.

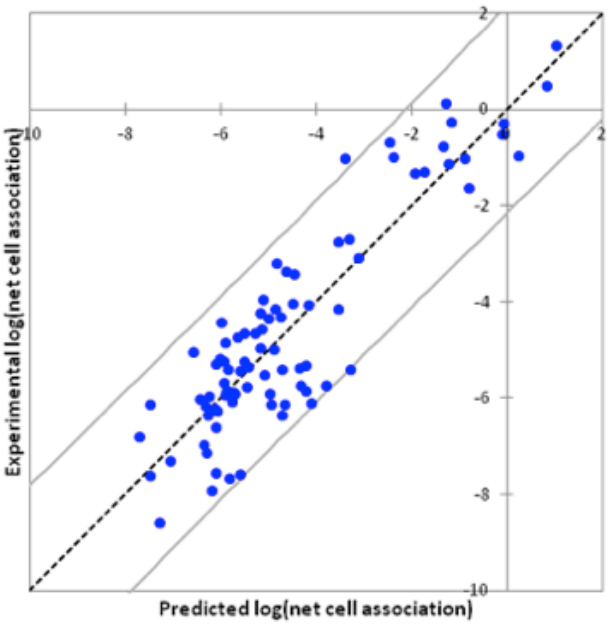

Figure 6: *Predicted versus experimental log_2(net cell association) values for 84 gold nanoparticles. The 95% confidence interval, centred in the dotted line, is shown for visual reference.*

Finally, to discard chance correlations, the values of the net cell association were changed randomly and a new model was constructed using the randomized data. The R^2 for Y randomized PCSAR dropped to 0.16, indicating that predicted values were not obtained by chance. **Figure 6** shows the predicted versus experimental net cell association values for the set of 84 gold nanoparticles. As it can be observed, all predictions except two fall inside the 95% confidence interval.

3.1.2 Validation of the approach with silver nanoparticles

Silver nanoparticles were also considered to check the applicability of the AP-based fingerprint for PCSAR development. The subset of silver ENMs in the original dataset was formed by 12 nanoparticles, with anionic and cationic surface ligand formulations. The primary observations were consistent with those previously reported in the literature [1], where the model developed for gold nanoparticles could not accurately predict nanoparticles with a different core (i.e., silver).

Following the same approach used for gold nanoparticles, a separate multilinear regression model was developed for silver nanoparticles:

$$\begin{aligned} \log_2(\textit{net cell association}) = \\ & - 2.17 \\ & + 21.8 \times \textit{glutamine DayhoffStat} \\ & + 11.6 \times \textit{A280 molar extinction coefficient} \\ & + 11.4 \times \textit{isoleucine DayhoffStat} \\ & - 3.42 \times \textit{tyrosine DayhoffStat} \\ & - 7.40 \times \textit{acidic amino acids percentage} \\ & - 7.71 \times \textit{basic amino acids percentage} \\ & - 26.0 \times \textit{aliphatic amino acids percentage} \end{aligned} \quad (2)$$

The normalized AP-fingerprints were used to develop the model in Eq. 2. The performance of the PCSAR model for silver (Eq. 2) in terms of R^2 for the entire dataset is 0.96. A complete cross-validation analysis was also conducted for the above model. The LOO cross-validation yields a similar value for R^2 of 0.96. The 10-fold cross-validation, which was repeated 10 times, had a R^2 of 0.98 ± 0.05. Similarly, the LMO cross-validation, which was repeated 10 times, had a lower R^2 value of 0.71 ± 0.39 with a significant variability. The 0.632 bootstrap corrections yield a R^2 of 0.78 ± 0.36. These results are explained by the fact that the number of observations in the dataset is very limited (i.e., only 12 nanoparticles) and the number of features included in the model given in Eq. 2 is relatively high (i.e., 7 physicochemical properties).

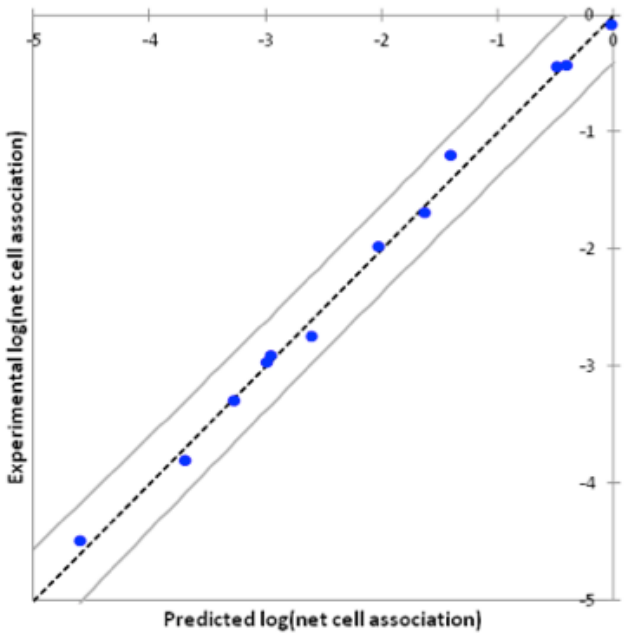

Figure 7: *Predicted versus experimental log_2(net cell association) values for 12 silver nanoparticles. The 95% confidence interval, centered in the dotted line, is shown for visual reference.*

The model perfectly fits, in general terms, the small dataset of silver nanoparticles. However, when it is tested stringently, it tends to have a R^2 value that is in the same order as that of the model for gold nanoparticles, although the variability of the estimate is very high. **Figure 7** depicts predicted versus experimental net cell association values for silver nanoparticles. As it can be observed, all predictions fall inside the 95% confidence interval

3.1.3 Biological interpretation of the predictive model for gold nanoparticles

The role of the protein corona features in the downstream processing of the nanoparticle-protein corona (NP-PC) complex system has been extensively studied. The types of proteins that bound onto the surface of the nanoparticle depend largely on the type of ligand associated with the nanoparticle [3]. The formation of protein corona affects the cellular uptake of the nanoparticle, trafficking and *in vivo* biodistribution of nanoparticles [4].

The biological interpretation of the different model parameters in Eq. 1 can be summarized as follows:

- The probability of expression in inclusion bodies can be interpreted as a measure of the solubility potential [5]. Protein and peptide solubility controls the concentration of proteins in solution, which in turn increases protein adsorption.
- The so-called tiny amino acids include alanine, cysteine, glycine, serine and threonine. This parameter plays an important role in the model since cellular uptake is related to size and factors that depend on the selective permeability of the cell membrane. According to the model, the presence of a high percentage of tiny amino acids increases the net cell association.
- Basic amino acids, which include the positively charged amino acids (namely histidine, lysine and arginine), play an important role in the electrostatic interactions of the nanoparticle with the

protein corona [6]. Positively charged proteins will be attracted to a negatively charged membrane by nonspecific electrostatic interactions.

- Molecular weight is a fundamental parameter for most interactions in biology. In particular, molecular weight is related to the size of the protein. A protein corona formed by large proteins will have a lower net cell association potential.
- Polar amino acids are usually found at the surface of proteins. Some proteins destined for the membrane contain groups of nonpolar amino acid side chains that create a water-shunning (hydrophobic) region on their surface [7]. Accordingly, the percentage of polar amino acids in the corona contributes to decrease net cell association.
- Acidic amino acids (i.e., negatively charged amino acids, and aspartic and glutamic acid) are known to play a vital role in the electrostatic interactions between nanoparticles and their protein corona [8, 9], as well as in the interaction with cell membranes. The presence of acidic amino acids contributes, via electrostatic repulsion, to decrease the cell association to negatively charged membranes.

3.2 Mutagenicity of chemicals

3.2.1 *In vitro* model based on Ames dataset

For the *in vitro* Ames test training dataset, comprising 557 mutagenic chemicals and 494 non-mutagenic chemicals, the chemical space was analysed in terms of the abundance of functional groups of the chemicals in the dataset using istChemFeat. The most abundant functional groups in the dataset were acceptor atoms for H-bonds (N, O, F), aromatic C(sp2), unsubstituted benzene C(sp2), substituted benzene C(sp2), terminal primary C(sp3), donor atoms for H-bonds (N and O), total secondary C(sp3), non-aromatic conjugated C(sp2), hydroxyl groups, and donor atoms for H-bonds (N and O).

Further, metabolites of these 557 mutagenic chemicals and 494 non-mutagenic chemicals were generated by CRAFT, and compiled along with their parent compounds into a list comprising 10380 records. The number of times that each metabolite occurred in the reactions was also recorded. In specific, for each metabolite, its occurrence in the metabolic pathways of both mutagenic and non-mutagenic chemicals was counted.

Further these 10380 metabolites were sorted as per their frequency of occurrence in each type of metabolic pathway (i.e., mutagenic or non-mutagenic). A total of 1769 (17%) metabolites were found to be occurring more than two times in the metabolic pathways of mutagenic chemicals and completely

absent in the metabolic pathways of non-mutagenic chemicals. Similarly, there were 443 (4%) metabolites present more than two times in the metabolic pathways of non-mutagenic chemicals and they were lacking in the metabolic pathways of mutagenic chemicals. These subsets of 1769 (17%) and 443 (4%) metabolites were labelled as metabolic triggers for mutagenic and non-mutagenic chemicals, respectively.

The identified metabolic triggers were used to predict the mutagenicity of chemicals that were predicted as unknowns by SARpy. Using the *in vitro* Ames training dataset, which comprised 1051 chemicals, SARpy only accepted 1004 (96%) chemicals, based on which it generated a set of 35 mutagenic and 33 non-mutagenic rules. For 553 mutagenic chemicals, the classifier based on SARpy rules classified 456 (82.46%) as mutagenic, 75 (13.56%) as non-mutagenic, and 22 (3.98%) as unknowns. For 451 non-mutagenic chemicals, 345 (76.50%) were classified as non-mutagenic, 32 (7.10%) as mutagenic, and 74 (16.41%) as unknowns. The metabolic pathways for the set of 96 (10% of the dataset) chemicals labelled as unknowns by SARpy were generated using CRAFT and the occurrence of the metabolic triggers was checked. Based on the use of the identified metabolites, the metabolic triggers approach was able to correctly predict additional 45 chemicals (5% of the dataset), of which six were mutagenic and 39 were non-mutagenic chemicals, therefore slightly increasing the accuracy and significantly bringing down the percentage of unknowns from 9.56% to 5.08%. For the

remaining 51 unknown chemicals (5% of the dataset), of which 16 were mutagenic and 35 were non-mutagenic, concrete evidence in the form of metabolic triggers was not obtained from CRAFT.

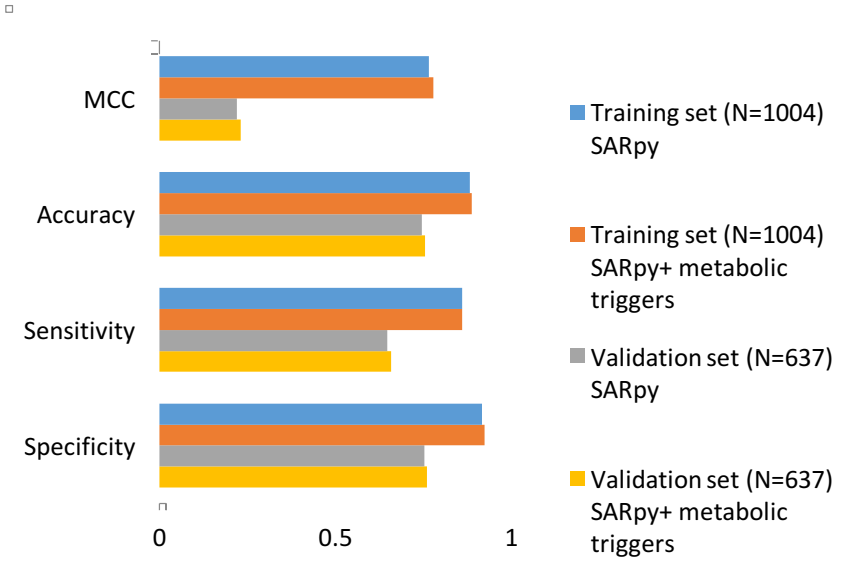

Figure 8: *Results of prediction of in vitro mutagenic and non-mutagenic chemicals, using SARpy and SARpy + metabolic triggers approach.*

Figure 8 provides a comparison of the performances of the two approaches. It should be noted that the SARpy + metabolic triggers approach was able to predict correctly 846 out of 953 total predictions that were made in the training set, obtaining an accuracy of 89%. SARpy provides processed results that are easy to interpret. However, the results from CRAFT are in the form of reactions whose reactants and products must be compared with the list of metabolic triggers. For example, the experimental non-

mutagenic chemical CN(C)C1=CC=CC=C1 has a very likely product, CNC1=CC=CC=C1, which could further have two products, CNC1=CC(O)C(O)C=C1 or NC1=CC=CC=C1, before the reaction comes to an end. In this case, the parent compound itself and all the metabolites are metabolic triggers for *in vitro* non-mutagenic chemicals. In the case of mutagenic chemicals, a chemical was considered to be mutagenic if one of the subsequent metabolites matched with the metabolic triggers. For example, the experimental mutagenic chemical BrCCBr has a likely product of OCCBr. Subsequently, OCCBr will likely have BrCC=O. This chain of reactions has a likely occurrence, and OCCBr is a metabolic trigger for an *in vitro* mutagenic chemical. Hence, the parent compound BrCCBr was predicted as mutagenic, since the likely reaction it will go through will produce a mutagenic metabolite. In this way, taking into account metabolic triggers and likelihoods of the reactions, the mutagenicity of chemicals was predicted based on the parent compound itself or based on the subsequent metabolites.

The same approach was further tested on the *in vitro* Ames test external validation set. Before testing, the chemical space of this dataset was analyzed using istChemFeat, to understand its similarity in terms of functional groups with the *in vitro* Ames test training set. Six out of the ten most abundant functional groups of this dataset are also in the list of the ten most abundant functional groups of the *in vitro* Ames test training set, hence emphasizing the fairly good similarity between the two datasets. These comprise

acceptor atoms for H-bonds (N, O, F), terminal primary C(sp3), total secondary C(sp3), aromatic C(sp2), hydroxyl groups and non-aromatic conjugated C(sp2). Other abundant functional groups, absent in the *in vitro* Ames test training set, were donor atoms for H-bonds (N and O), substituted benzene C(sp2), unsubstituted benzene C(sp2) (multiple occurrences), and total tertiary C(sp3).

Owing to the imbalance in this dataset, wherein the percentage of mutagenic chemicals was very small and a bit different in terms of the abundant functional groups, a new set of rules was generated for the validation dataset using SARpy, and then the metabolic triggers of unknowns from SARpy were compared with the metabolic triggers generated from the original training dataset. The newly generated set of SARpy rules included a total of 38 rules, of which 7 correspond to mutagenic chemicals and the remaining 31 were for non-mutagenic substances. The structure of the new set of rules was in line with the imbalanced composition of the validation dataset (i.e., a small number of mutagenic chemicals and a large number of non-mutagenic chemicals).

Using the above 38 rules for the 42 (7% of the dataset) mutagenic chemicals, the SARpy system classified 22 of them (52.38%) as mutagenic, 12 (28.57%) as non-mutagenic, and 8 (19.05%) as unknowns. For the 595 non-mutagenic chemicals (93% of the dataset), 377 (63.36%) were classified as non-mutagenic, 124 (20.84%) as mutagenic, and 94 (15.80%) as unknowns. Metabolic pathways for the unknown chemicals were

generated using CRAFT and the intermediate reaction products were compared with the metabolic triggers identified during the analysis of the training dataset. Using metabolic triggers, the approach could predict additionally 16 chemicals (2.51%): 1 mutagenic and 15 non-mutagenic. A summary of the results is shown in **Figure 8**. Hence, out of 551 predictions made by the SARpy + metabolic triggers approach, 415 predictions were correct, obtaining an accuracy of 75%. Due to the low numbers of true positives and false positives, a drop in the Matthews correlation coefficient has been observed, which was only 0.23.

3.2.2 Metabolic triggers to predict *in vivo* mutagenicity based on rodent micronucleus assay

As for the *in vitro* datasets, an analysis of the chemical space was also performed for this dataset in terms of abundance of functional groups. The most abundant functional groups in this dataset were acceptor atoms for H-bonds (N, O, F), aromatic C(sp2), unsubstituted benzene C(sp2), terminal primary C(sp3), substituted benzene C(sp2), donor atoms for H-bonds (N and O), total secondary C(sp3), non-aromatic conjugated C(sp2), and hydroxyl groups. In particular, functional groups acceptor atoms for H-bonds (N, O, F) and non-aromatic conjugated C(sp2), both of which have multiple occurrences in the same chemicals, are unique to this dataset and are not present in the two *in vitro* datasets.

This dataset contained chemicals that have been tested *in vivo* in rodents using the micronucleus assay, which is one of the tests most widely used for confirmation of mutagenicity. Using CRAFT, a list of 2144 metabolites was obtained using the approach described previously. Further, only the metabolites which had a count of two or more for the class they belong to and zero counts for the other class were retained and labelled as metabolic trigger for their corresponding class. A final list of 326 (15% of total) metabolic triggers for mutagenic chemicals and 276 (13% of total) metabolic triggers for non-mutagenic chemicals, respectively, was obtained using this approach.

3.2.3 Comparison between *in vitro* and *in vivo*

The dataset used to analyse chemicals for *in vivo* mutagenicity included information about the bioactivity of chemicals in the Ames test in addition to the micronucleus assay in rodents. The comparative analysis of this information is critical to establish a link between the *in vitro* and *in vivo* mutagenic responses. As explained previously in Chapter 2, a preliminary analysis of the mutagenicity data revealed that 72 (15%) chemicals were mutagenic *in vitro* and non-mutagenic *in vivo*. Similarly, 94 (19%) chemicals were mutagenic *in vivo* and non-mutagenic in vitro. The criteria used above to identify a metabolite as a metabolic trigger was applied to the combination of *in vitro/in vivo data*. The analysis resulted in 12 metabolic triggers for the class mutagenic *in vitro* and non-mutagenic *in vivo*, and 13 metabolic

triggers for the class mutagenic *in vivo* and non-mutagenic *in vitro*. Reasons for opposite *in vitro* and *in vivo* responses could be differences in the test (duration, quantity of exposure and so on) and dynamics of the metabolic machinery used by bacteria relative to higher order animals.

In addition to understanding the contradictory behaviour of chemicals *in vitro* and *in vivo*, the identification of chemicals which are non-mutagenic in both assays may facilitate the development of new *in silico* systems that can contribute to reducing the experimental efforts needed for mutagenicity screening. Hence, a list of 26 metabolic triggers that uniquely belong to chemicals which are non-mutagenic after S9 activation in the Ames test as well as in the rodent micronucleus assay were generated from the 123 chemicals that were non-mutagenic both *in vitro* and *in vivo*.

3.2.4 Analysis and validation of metabolic triggers

Further, a deepened analysis of metabolic triggers was performed wherein presence of features (i.e., chemical substructures) that contribute to distinguishing between *in vitro* and *in vivo* mutagenicity were looked for. To this end, relevant chemical features (i.e., functional groups and atom-centred fragments) were first identified using the software istChemFeat. These features have been searched separately in the three sets of metabolites selected: *in vitro* mutagens, *in vitro* non-mutagens and *in vivo* mutagens. The chemical groups that uniquely characterize

each class of metabolites were identified by comparing the different outputs that istChemFeat generated for each set of triggers.

Table 4: *Groups belonging uniquely to the class of metabolites in vitro mutagenic. istChemFeat distinguishes between aromatic and aliphatic form of groups and it shows whether a specific group is present one or more times in the metabolites (single or multiple occurrences).*

Group	**Total matches**
R--CX..X (single occurrence)	31
Nitro groups (aliphatic) (single occurrence)	29
Hydroxylamines (aliphatic) (single occurrence)	21
Isothiazoles (single occurrence)	14
X--CH--X (single occurrence)	14
Isocyanates (aromatic)	11
Al2-NH (multiple occurrences)	11

For example, specific fragments like nitro group or Isocyanates (aromatic) have been recognized only in metabolites generated from chemicals that are mutagenic in vitro. **Table 4** summarizes the chemical groups that were found the most in metabolites. In particular, for istChemFeat X–CH—X, R–CX..X, Al2-NH are atom-centred fragments, where R represents any group linked through carbon, X any electronegative atom (O, N, S, P, Se, halogens) and Al represents an aliphatic group (in this case two aliphatic groups). It should be noted that istChemFeat has also recognized several functional groups that are typically known in the literature for being associated to mutagenicity (i.e., SAs) like ketones (aromatic) (e.g., SA12_Ames in Benigni Bossa rules) or nitro groups (aromatic) (e.g., SA27_Ames of Benigni Bossa rules).

However, these groups have not been included in **Table 4** because they are not specific to *in vitro* mutagenic compounds [10].

Table 5: *Results of istChemFeat with groups having highest counts uniquely in the class of in vitro non-mutagenic.*

Group	**Total matches**
Total quaternary C(sp3) (multiple occurrences)	23
CR4 (multiple occurences)	23
Ring quaternary C(sp3) (multiple occurrences)	20
Oxetanes (single occurrence)	15
N+ (positively charged) (single occurrence)	9
Imides (-thio) (multiple occurrences)	8
Urea (-thio) derivatives (multiple occurrences)	7
Carboxylic acids (aromatic) (multiple occurrences)	5

Similarly, the analysis of the metabolites specific to the *in vitro* non-mutagenic chemicals identified several groups that are not present in the other two-bioactivity classes. **Table 5** summarizes the functional groups that appear more frequently.

Table 6: *Functional groups specific to in vivo mutagenic.*

Group	**Total matches**
Primary amides (aromatic) (single occurrence)	2
Imines (aliphatic) (multiple occurrences)	4
Pyrroles (multiple occurrences)	3

The *in vivo* mutagenic chemicals, as shown in **Table 6,** have only three groups corresponding to the metabolites that are specific for chemicals that belong to this bioactivity class and were not present in the other two classes(*in vitro* mutagenic and *in vitro*

non-mutagenic). The frequency of occurrence of these groups is very low (i.e., have been found only in very few compounds).

Table 7: *Functional groups of mutagenic chemicals both in vitro and in vivo tests.*

Group	**Total matches**	**Total *in vitro* positive**	**% *in vitro* positive**	**Total *in vivo* positive**	**% *in vivo* positive**
Secondary amines (aliphatic) (multiple occurrences)	13	11	85	2	15
Br attached to C1(sp3) (multiple occurrences)	5	4	80	1	20
Imines (aliphatic) (single occurrence)	5	3	60	2	40
R=CRX (single occurrence)	5	3	60	2	40
Guanidine derivatives (single occurrence)	11	4	36	7	64
(C-020)=CX2 (single occurrence)	6	2	33	4	67

Coincidences in the functional group counts and atom-centred fragments that are representative of both *in vitro* and *in vivo* mutagenicity are summarized in **Table 7**. Similarly, groups and atom-centred fragments that are characteristic of both *in vitro* non-mutagenic and *in vivo* mutagenic were considered as false negative alerts. Examples of these false negative alerts are sulfonamides (thio-/dithio-) (single occurrence), pyridines (multiple occurrences), urea (-thio) derivatives (single occurrence), 1-3-5-Triazines (single occurrence) and tertiary amides (aromatic) (single occurrence).

In order to check the validity of metabolic triggers, they were also compared to the SAs available in literature. *In vitro*

mutagenic metabolic triggers were compared to Toxtree SAs [9]. A total of 997 *in vitro* mutagenic triggers had the final prediction as mutagens. Assessments within VEGA comprise of the prediction along with their reliability, which could "experimental value" or "good reliability" or "moderate reliability" or "low reliability", with "experimental value" being the most reliable and "low reliability" being the least. Of these triggers, the assessment for 40 was of experimental value, 268 had good reliability and 689 had low reliability. Overall, the similarity index, calculated based on the molecules' fingerprint and structural aspects (count of atoms, rings and relevant fragments), was between 0.7–1, hence sufficient to bolster the validity of the *in vitro* metabolic triggers.

To check the validity of the *in vivo* mutagenic metabolic triggers, they were compared with the SAs of the *in vivo* micronucleus assay in rodents [9], wherein 317 metabolic triggers had between 1–4 mutagenic SAs for the micronucleus assay (Class I) similar to them.

3.3 *In silico* classification of bacterial proteins into antibiotic-resistance profiles

In this work an alignment free classifier which complemented alignment based BLAST was generated.

Table 8: *Confusion matrix of the combined classifiers and cumulative confusion matrix of the 5 classifiers on the validation set, which had 654 observations. Predicted classes are given in columns and real ones are in rows, following the same order.*

IPS		CWBM		IMP		INAS	
S	**M**	**S**	**M**	**S**	**M**	**S**	**M**
BLAST							
132	0	0	0	0	0	0	0
3	124	0	1	0	0	0	0
0	0	179	1	0	0	0	0
0	0	0	114	0	0	0	0
0	0	0	0	11	0	0	0
0	0	1	1	0	53	0	0
0	0	0	0	0	0	21	0
0	0	0	0	0	0	0	4
Consensus classifier							
120	7	5	4	1	0	0	1
4	117	1	2	2	0	2	0
0	4	167	7	0	0	3	0
1	0	4	110	0	0	1	0
0	0	0	0	11	0	0	0
0	1	1	2	0	51	0	0
2	0	0	0	0	0	19	0
0	0	0	0	0	0	0	4
Combined approach							
135	0	1	2	0	0	0	0
3	124	0	1	0	0	0	0
0	0	180	1	0	0	0	0
0	0	0	116	0	0	0	0
0	0	0	0	11	0	0	0
0	0	1	1	0	53	0	0
0	0	0	0	0	0	21	0
0	0	0	0	0	0	0	4

Table 8 showcases in terms of a confusion matrix, the class wise performance of all three methods, namely, alignment based BLAST, the alignment free consensus classifier and the combined approach.

3.3.1.1 Classification using BLAST

The capability of BLAST to classify antibiotic-resistant proteins into the eight classes was analysed, based on homology search, using the amino acid sequences of these antibiotic-resistant proteins. Each one of the 654 antibiotic-resistant proteins in the validation set was searched for sequence similarity against the entire training set. Using a non-so-stringent E-value threshold of 10^{-10}, BLAST was able to classify 645 out of 654 proteins in the validation set, showing that it cannot classify all the antibiotic-resistant proteins. Although the number of instances correctly classified by BLAST and shown in **Table 8** is impressive, its major drawback is its inability to find some hits. Another pitfall of BLAST was the quality of alignment: even though some predictions were correct, their bit score was found to be less than 100, which is usually considered a threshold for quality alignments. In addition to these problems, there were antibiotic-resistant proteins in the validation set for which their most similar antibiotic-resistant proteins in the training set belonged to more than one class. These drawbacks hence signify that BLAST cannot be used as a sole method for classifying antibiotic-resistant proteins, and there is a strong need for an alignment-free method to complement BLAST in classifying unknown antibiotic-resistant proteins rendered by BLAST.

3.3.1.2 Classification using the alignment-free method

3.3.1.2.1 Feature selection

A total of 33 features from the pepstats program were used as the initial set of features according to the information they provided. The features comprised of simple yet important protein properties, namely: molecular weight, charge, isoelectric point, A280 molar extinction coefficient, probability of expression in inclusion bodies, Dayhoff Stat for each amino acid, and molar percent for each physicochemical class of amino acid (tiny, small, aliphatic, aromatic, non-polar, polar, charged, basic, and acidic).

Table 9: *Order of attributes obtained from the Infogain Ranker method of WEKA. Value of information gain measures the amount of information gained, in terms of class separation, when a new attribute is added.*

Attribute	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Trp DayhoffStat	1.081	1.043	1.141	1.092	1.081
Molecular weight	1.078	1.054	1.045	1.074	1.078
Cys DayhoffStat	0.912	0.902	0.844	0.935	0.912
His DayhoffStat	0.828	0.772	0.863	0.718	0.828
Tyr DayhoffStat	0.807	0.828	0.807	0.846	0.807
A280 Molar Extinction Coefficient	0.748	0.743	0.726	0.867	0.748
Lys DayhoffStat	0.73	0.741	0.746	0.716	0.73
Aliphatic mole	0.716	0.758	0.714	0.704	0.716
Phe DayhoffStat	0.704	0.835	0.658	0.81	0.704

All these features were sorted by the value of information gain obtained from Infogain feature selection method, which assigns values to features based on their contribution to reduction in entropy, as summarized in **Table 9**. With this input, different classifiers were constructed incrementally, with one ranked feature added at a time to the classifier.

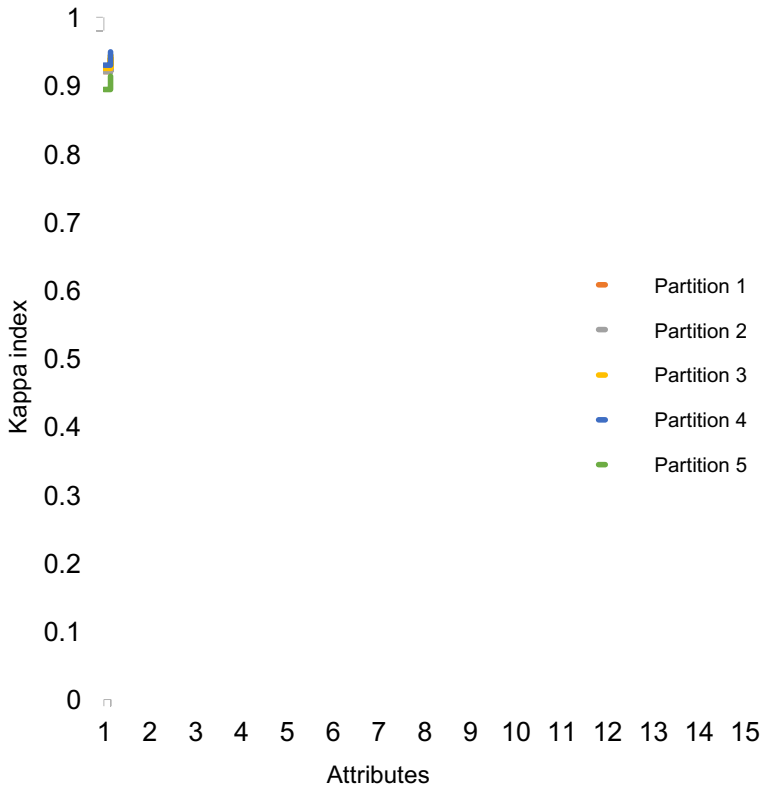

Figure 9: *Performance of all the training sets, partitioned using Kennard & Stone sampling, on classification of instances in increasing order of ranked attributes. The x-axis corresponds to the number of attributes and the y-axis corresponds to the* κ *index.*

As shown in **Figure 9**, it was observed that the κ index kept increasing until the addition of the ninth ranked feature. Further on, when the next five ranked features were added, the index showed stagnation. Therefore, the addition of new features was stopped and the first nine ranked features were used for the final classification models, since the addition of more features did not make any significant improvement in the performance of the classifiers.

The attributes selected in the five datasets, in decreasing order of average value of information gain (**Table 9**), are: Trp DayhoffStat, molecular weight, Cys DayhoffStat, A280 molar extinction coefficient, and His DayhoffStat. From a biological point of view, it can be concluded that our results are in line with results of antibiotic resistance already published in literature:

- Tryptophan (given by the feature "Trp DayhoffStat") is an important amino acid in terms of antibiotic resistance in bacteria, because highly resistant bacteria show increased production of Indole from Tryptophan, with the help of tryptophanase, which is known to have a role to play in imparting resistance to bacteria [10].
- Ranking of molecular weight as the second attribute is an expected result considering that it is a central value for any macromolecule, providing basic information about it.
- Molar extinction coefficient at an absorbance above 275nm is a measure of side chains of tryptophan, tyrosine and cystine, also known as chromophores [11]. Note that cystine, which is

a contributor to A280 molar extinction coefficient, also appears as a highly ranked attribute, as shown in **Table 9**.

Table 10: *Kappa indices of the three approaches for the validation set.*

Approach	**κ index**
BLAST	0.99
Consensus	0.91
Combined	0.98

3.3.1.2.2 Validation

The five alignment-free classifiers were validated using the validation sets, and the majority response of these classifiers for each protein was the result of the consensus classifier. As shown in **Table 10**, a κ index of 0.91 was obtained for the consensus classifier.

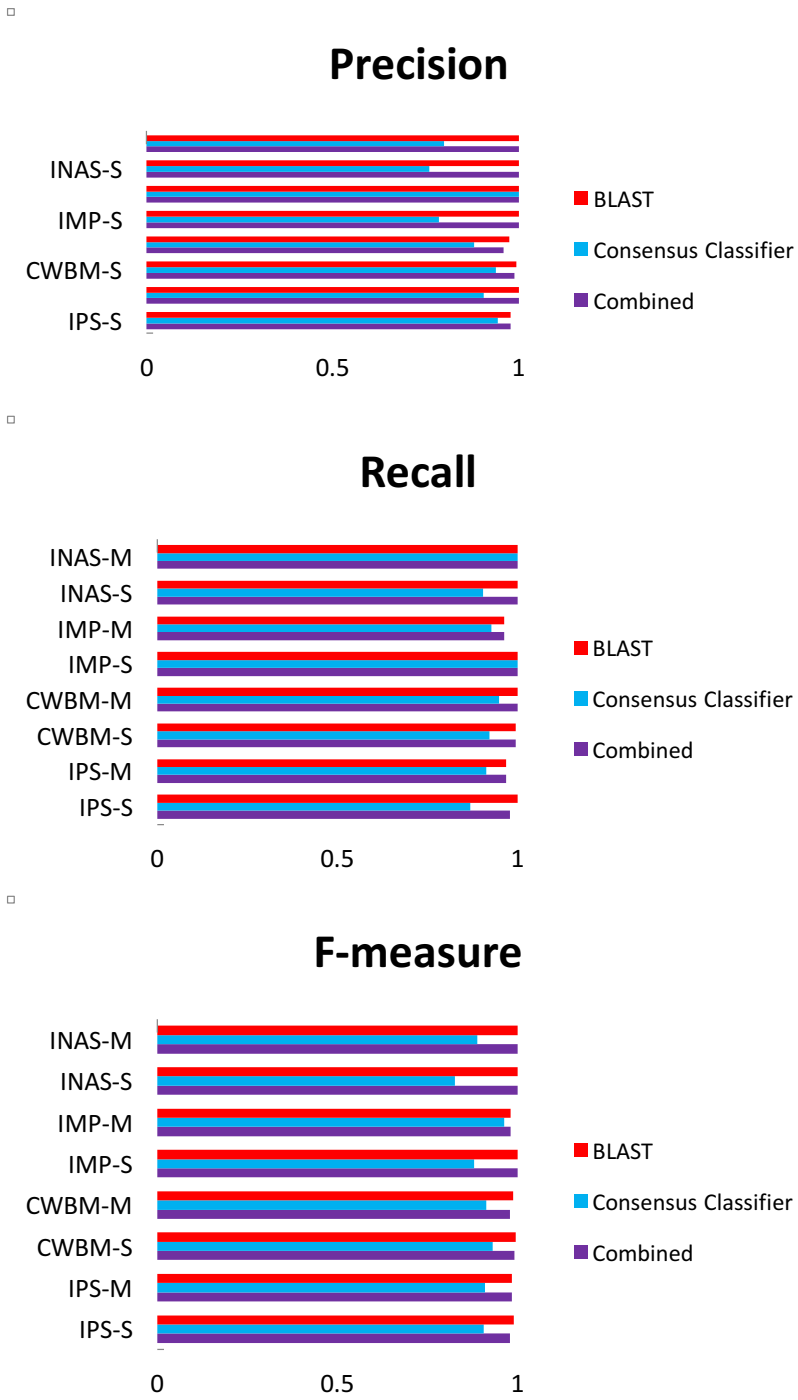

Figure 10: *Evaluation based on precision, recall and F-measure metrics, for the three approaches and on the same validation set.*

Precision, recall and F-measure have been represented in **Figure 10**d for the consensus classifier.

. The precision value for IMP-M, and the recall value for IMP-S and INAS-M was 1, whereas the highest F-measure value was obtained for IMP-M, which was 0.96.

The relationship between the eight resistance profiles and the selected features was concretized when the class labels of antibiotic-resistant proteins were randomly changed with the purpose of performing Y-randomization. The statistical significance of the above mentioned relationship in the alignment-free classifier was demonstrated constructing a new classifier based on these randomly changed class labels of the resistance profiles. The classifier constructed on the permuted class labels showed a significant drop in the κ index which was now equal to 0.31, hence asserting that our classifier was statistically significant and the results obtained for the five incrementally constructed classifiers were not obtained by chance.

3.3.1.3 Combined approach

The alignment-free classifier was constructed as a complementary model to BLAST, making together a combined approach with the aim of classifying antibiotic-resistant proteins that BLAST is unable to classify. Initially, BLAST was applied to obtain the resistance profile of each query protein in the validation set. For the cases where there were no hits, or hits had a bit score

lower than 100, or the query protein was associated with more than one resistance profile, the prediction given by the alignment-free classifier was considered.

From the confusion matrix given in **Table 8**, it can be observed that the combined approach gives a fair tradeoff, with a high efficiency similar to that of BLAST and complete prediction coverage. Hence, with the combined approach predictions for all 654 antibiotic-resistant proteins in the validation set were obtained and the quality of predictions obtained with BLAST was also retained, since the total number of misclassified instances was only 10. As shown in **Table 10**, the combined approach has a κ index of 0.98. This result justifies the quality of the combined approach, especially taking into account that it allows the classification of all antibiotic-resistant proteins without exceptions. In terms of the other metrics, the values obtained by the combined approach are close to the ones obtained by BLAST, as it can be seen in **Figure 10**.

3.4 References

1. Walkey, C.D.; Olsen, J.B.; Song, F.; Liu, R.; Guo, H.; Olsen, D.W.H.; Cohen, Y.; Emili, A.; Chan, W.C.W. Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. ACS Nano, 2014, 8(3), 2439-2455.

2. Kou, P.M; Pallassana, N.; Bowdena, R.; Cunningham, B.; Joy, A.; Kohn, J.; Babensee, J.E. Predicting biomaterial property-dendritic cell phenotype

relationships from the multivariate analysis of responses to polymethacrylates. Biomaterials, 2012, 33(6), 1699-1713.

3. Casals, E.; Pfaller, T.; Duschl, A.; Oostingh, G.J.; Puntes, V. Time evolution of the nanoparticle protein corona. ACS Nano, 2010, 4(7), 3623-3632.

4. Treuel, L.; Brandholt, S.; Maffre, P.; Wiegele, S.; Shang, L.; Nienhaus, G.U. Impact of protein modification on the protein corona on nanoparticles and nanoparticle cell interactions. ACS Nano, 2014, 8(1), 503-513.

5. Harrison, R.G. Expression of soluble heterologous proteins via fusion with NusA protein. inNovations (Novagen), 2000, 11, 4-7.

6. Rocha, A.; Zhou, Y.; Kundu, S.; González, J.M.; Vinson, S. B.; Liang, H. *In vivo* observation of gold nanoparticles in the central nervous system of Blaberus discoidalis. J. Nanobiotechnol., 2011, 9(5).

7. Kucki, M.; Kaiser, J.P.; Clift, M.J.D.; Rothen-Rutishauser, B.; Petri-Fink, A.; Wick, P. The role of the protein corona in fiber structure-activity relationships. Fibers, 2014, 2(3), 187-210.

8. Tiwari, A.; Turner, A.P.F. Biosensors Nanotechnology; Wiley & Sons: New York, 2014.

9. R. Benigni, C. Bossa, N. Jeliazkova, and A. Worth, The Benigni / Bossa rulebase for mutagenicity and carcinogenicity – A module of Toxtree, EUR 23241 EN 2008.

10. Lee, J.; Lee, J. Indole as an intercellular signal in microbial communities. FEMS Microbiol Rev., 2010, 34(4), 426-44.

11. Wetlaufer, D.B. Ultraviolet Spectra of Proteins and amino acids. Adv. Protein Chem., 1962, 17, 303-391.

Chapter 4 Conclusions

This thesis puts forth three non-testing approaches to evaluate emerging chemicals by deducing information available on chemical entities which have been already tested using laboratory tests. Three *in silico* approaches described in this thesis can be used for:

1. assessing cellular interactions of nanoparticles that determine their further course *in vivo*, taking into account physicochemical properties of both nanoparticles and protein corona;
2. analysing toxicity of chemicals in the context of different *in vitro* and *in vivo* tests, based on SAR rules and taking into account the intermediates and products of these chemicals;
3. exploring the know-how of drug resistance profiles in bacteria by taking into account physicochemical properties of drug targets.

4.1 Predicting cell association of surface-modified nanoparticles using Protein Corona Structure-Activity Relationships (PCSAR)

This work presents a comprehensive framework for the prediction of net cell association of the NP-PC complex based on combined information derived from relative abundance and physicochemical properties of the protein corona. The use of protein corona features to predict a biological endpoint provides an alternative and effective approach for developing structure-activity relationships for nanoparticles. In this regard, Protein Corona Structure–Activity Relationships (PCSAR) can be used to link the composition and properties of the corona with nanoparticle's bioactivity profile.

The proposed model uses fingerprints based on physicochemical descriptors of the proteins attached to the nanoparticle. Selected descriptors are simple and can be easily computed with low computational cost for any protein. The advantage of the current model, relative to existing models developed from fingerprints and based on protein abundance, is that is not restricted to datasets or serums that contain exactly the same proteins as those used for training the model. In addition, the combination of using physicochemical descriptors weighted by relative abundance results in more general models with a larger applicability domain. Models can be used to predict cell

association of nanoparticles with protein coronas containing proteins different than those used for training the model. The only information needed by the model is the spectral counts and the primary sequence of the protein. This is a key issue, especially taking into account that the composition of the hard corona, which was once considered to be stable, has been recently found to be evolving with the migration of nanoparticles across different biological fluids [1, 2]. Accordingly, models of cell association must be able to predict cell interactions with a great variety of proteins in order to take into account the dynamic behavior of the corona.

The two fingerprint techniques have been compared by calculating the normalized mutual information index between the partitions obtained after clustering the nanoparticles represented in terms of each fingerprint. Clustering results indicate that the information conveyed by both fingerprints is essentially the same. In addition, the modelling approach proposed here for gold and silver nanoparticles outperforms models based only on relative abundances in terms of applicability, size and stability. Models based on physicochemical descriptors can be applied to a larger set of proteins, as long as the primary sequence and the spectral counts for proteins of the corona are available. Whereas, models based exclusively on relative abundances will only work for proteins specified within the training set. The models developed for gold and silver nanoparticles use very few and easy to obtain

physicochemical descriptors, and their performances are stable under different validation conditions.

The low performance obtained when applying the model developed from gold nanoparticles to predict cell association of silver nanoparticles indicates that the nanoparticle core is a key factor that determines the structure and composition of the protein corona. These results are in line with previous work [3], which concluded that a model to predict cell association for gold nanoparticles is not suitable to predict cell association for silver nanoparticles. Nevertheless, the results obtained with the model developed for silver nanoparticles demonstrate that the current modelling approach can be successfully applied to develop individual models for nanoparticles with different cores.

From the point of view of the protein corona, the models developed here are able to explain the biological relevance of each amino acid in the protein corona with respect to cell association. A lot has been written about the importance of protein corona and its influence on the nanoparticle–cell interactions, but there is limited literature on the contribution of individual amino acids to the interaction between cell entities and the NP-PC complex.

With the advent of varied research on different nanoparticles, it would be interesting to take the PCSAR predictive model introduced here and extend it further to nanoparticles with different cores as well as to other bioactivity endpoints.

4.2 *In silico* exploratory study using structure-activity relationship models and metabolic information

The approach consisting in the determination of mutagenicity by taking into account the presence of specific metabolic triggers contributes to predicting mutagenicity for a higher number of chemicals, in comparison with only SAR approaches. In addition, the major advantage of this method is that it takes into account the possibility of not only the chemical but also the metabolites it generates to be mutagenic.

This approach could also be considered as a component of read-across approaches for chemicals whose mutagenic potential is not known; based on metabolites and reaction by-products, such chemicals would have metabolic triggers or data highly similar to metabolic triggers.

Also, the application of the present methodology results in detailed information of the metabolic pathway of a chemical along with the mutagenic and/or non-mutagenic intermediates that this chemical will produce, which provides insights into the mechanisms of mutagenicity.

The information obtained by the identification and analysis of bioactivity and metabolic triggers has helped to establish links across various experimental tests (e.g., Ames test and rodent micronucleus test). Using this method it is also possible to predict the *in vivo* mutagenicity of a chemical based on its *in vitro* mutagenicity information. The predictions thus obtained would be very reliable given that the identified metabolic triggers are generated from the reactions that are likely or very likely to occur. Future research, with more widely studied chemicals, will result in the identification of more metabolic triggers and will contribute to form more links across various mutagenicity assays in a similar fashion.

The most immediate application of the current approach could be its incorporation within the ToxRead software, which currently provides a reproducible read-across evaluation by identifying similar chemicals via the use of SAs and common relevant features [4]. The addition of information regarding metabolic pathways would give these evaluations a new dimension by providing a method to identify common metabolic triggers.

4.3 *In silico* classification of bacterial proteins into antibiotic resistance profiles

Considering the current scenario where the development of new antibiotics has almost stagnated, the option of exploring new bacterial proteins seems to be the most feasible approach. The

proposed approach which makes use of both alignment based BLAST and alignment free consensus classifier promises to be a useful tool to predict the antibiotic resistant profile of bacterial proteins.

4.4 References

1. Treuel, L.; Brandholt, S.; Maffre, P.; Wiegele, S.; Shang, L.; Nienhaus, G.U. Impact of protein modification on the protein corona on nanoparticles and nanoparticle cell interactions. ACS Nano, 2014, 8(1), 503-513.

2. Lundqvist, M.; Stigler, J.; Cedervall, T.; Berggård, T.; Flanagan, M.B.; Lynch, I.; Elia, G.; Dawson, K. The evolution of the protein corona around nanoparticles: a test study. ACS Nano, 2011, 5(9), 7503-7509.

3. Walkey, C.D.; Olsen, J.B.; Song, F.; Liu, R.; Guo, H.; Olsen, D.W.H.; Cohen, Y.; Emili, A.; Chan, W.C.W. Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. ACS Nano, 2014, 8(3), 2439-2455.

4. G. Gini, A.M. Franchi, A. Manganaro, A. Golbamaki, and E. Benfenati, ToxRead: A tool to assist in read across and its use to assess mutagenicity of chemicals, SAR QSAR Environ. Res., 2014, 25, 999-1011.