UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biologiche e Molecolari

XXVII Ciclo

Investigating and predicting the determinants of protein-protein interactions through computational-structural biology approaches: implications for structural vaccinology

Claudio Peri

PhD Thesis

Scientific tutors: Prof. Martino Bolognesi

Dr. Giorgio Colombo

Academic year: 2013-2014

SSD:BIO/10; BIO/11

Thesis performed at:
Istituto di Chimica del Riconoscimento Molecolare, CNR
Via Mario Bianco 9, 20131 Milano.

Deep into that darkness peering, long I stood there,
wondering, fearing, doubting

Edgar Allan Poe

CONTENTS

PART I: INTRODUCTION

1. SUMMARY 1

2. STATE OF THE ART 2

2.1 Melioidosis: a deadly disease without a vaccine 3
2.2 Reverse Vaccinology: making vaccines in the post-genomic era 6
2.3 The GtA consortium: the RV approach against melioidosis 10
2.4 Structural Vaccinology: from antigen to epitope 13
2.5 The digital revolution of protein science 16
2.6 Molecular Simulations: approaches, functional forms, potential and limitations. 18
2.7 Epitopes and simulations 23

3. AIM OF THE PROJECT 27

4. MAIN RESULTS 29

4.1 Defining the Structural Vaccinology pipeline 29
4.2 Application of the SV pipeline toward the identification of active epitopes 35
4.3 Rational design of active epitopes. 42
4.4 Expansion of the original SV pipeline towards protein targeting and diagnostics 46
4.5 Expansion of MLCE toward MHC-II epitopes and development of a web tool. 49
4.6 The role of surface energetics in the formation of protein-protein complexes 56

5. CONCLUSIONS AND FUTURE PROSPECTS 63

6. REFERENCES 66

7. ACKNOWLEDGEMENTS 73

8. CREDITS 74

PART II: SCIENTIFIC PUBLICATIONS

9. PUBLISHED MANUSCRIPTS 75

9.1 Lassaux P, Peri C, *et al*. Structure 21, 2013 75
9.2 Peri C, Gagni P, *et al*. ACS Chem Biol 8, 2013 75
9.3 Gourlay LJ, Peri C, *et al*. Chem Biol 20, 2013 75
9.4 Gori A, Longhi R, *et al*. Amino Acids 45, 2013 75

10. ACCEPTED AND SUBMITTED MANUSCRIPTS 76

10.1 Peri C, Corrada D, Conchillo-Solè O, *et al*. Method Mol Biol Gen, 2014 76
10.2 Gaudesi D, Peri C, *et al*. ACS Chem Biol, 2014 88

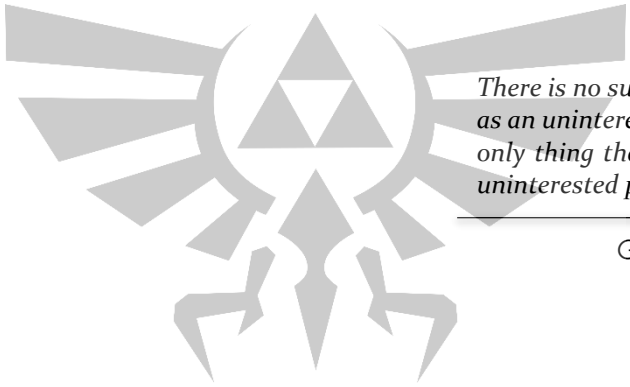
PART III: SUPPLEMENTARY MATERIALS

11.1 Supplementary figures 110
11.2 Supplementary tables 118
11.3 Supplementary methods 122

1. SUMMARY

Clarifying the physico-chemical principles of protein-protein interactions is critically important to understand the relationships between biological structures and functions in all biochemical mechanisms. In this project we aim to develop, validate and apply new computational-theoretical methods to study and predict the binding regions of proteins starting from 3D structural information and from the analysis of the conformational and physico-chemical properties of the constituting amino acids. In particular, this project entails the integrated analysis of the energetic properties of different datasets of proteins solved at high resolution. In this context, we have focused on four main subjects with different, yet highly intertwined, objectives. The first subject will address the application of an energy-based computational predictor for the identification of possible antibody-binding surfaces (epitopes) of protein antigens from the pathogen *Burkholderia pseudomallei,* responsible for human melioidosis. The second will focus on the expansion of the same rationale, adapting the method towards different applications, and including as a novel functionality the prediction of MHC-II coupled epitopes to elicit the intervention of T helper cells. The third objective concerns the design and characterization of peptides and peptidomimetics to optimize the properties of the identified epitopes as better vaccine candidates. The fourth one will pursue the investigation of the energetic determinants of interacting proteins in a more general context (not limited to immunogenic epitopes), aiming at the identification of an energy-based property describing the interaction event at the atomistic level of resolution. This part of the project is aimed at the development of a computational tool based on such property to help improve the understanding of the determinants of protein interactions and help predict their binding interfaces and orientation. All four subjects have been investigated in the broad spectrum of activities of an academic consortium, devoted to the identification of antigens from *B. pseudomallei* showing sufficient immunogenic potential to be considered as components for a vaccine against the pathogen. The computational methods developed and tested within this framework have theoretical as well as practical implications, from the physico-chemical study and characterization of protein-protein interactions, to the design of biologically active molecules.

2 - STATE OF THE ART



There is no such thing on earth as an uninteresting subject; the only thing that can exist is an uninterested person

G. K. Chesterton

This chapter will define the general framework of the thesis project, describing the overarching goal of all the endeavors, methods and applications developed during these three years of Ph.D., namely the development of better candidate immunodiagnostics and vaccines against melioidosis. Therefore, I will first focus on the disease, describing the essential facts concerning this endemic infection, its etiological agent and the state of art of past and current research initiatives. I will explain the reason why a vaccine is extremely desirable to eradicate this life-threatening disease, and why a quick and effective diagnosis may save many human lives. I will recapitulate the approaches and advances of vaccinology, describing how current technology may finally pave the way to an effective vaccine against melioidosis, mentioning the studies that provided us with the necessary information to investigate specific protein antigens with the highest protective potential. Finally, I will be introducing our Structural Vaccinology approach and our work rationale, in which computational biology plays a key role, both at the current level and in future advancements and implementations.

2.1 MELIOIDOSIS: A DEADLY DISEASE WITHOUT A VACCINE.

Melioidosis is a serious infectious disease caused by the Gram-negative environmental saprophyte *Burkholderia pseudomallei*, first described in Burma in 1912 as a "glanders-like" disease[1]. Stanton and Fletcher[2] named Melioidosis from the Greek words *melis*, which means “distemper of asses”, and *eidoes*, which means “resemblance”. To date, acute cases are most frequently reported from northeast Thailand, where it is the third most common cause of death due to infectious diseases, after HIV/AIDS and tuberculosis[3], and from Darwin in northern Australia where it has been the commonest cause of fatal community-acquired bacteremic pneumonia[4]. Melioidosis is also being increasingly reported from many countries across South and East Asia as well as parts of South America, Papua New Guinea and the Caribbean. It is apparently rare in Africa[5], although infection may pass unrecognized because diagnostic confirmation relies on microbiological culture, which is often unavailable in resource-restricted regions of the world. A map representing the endemicity of melioidosis is reported in Figure 1.

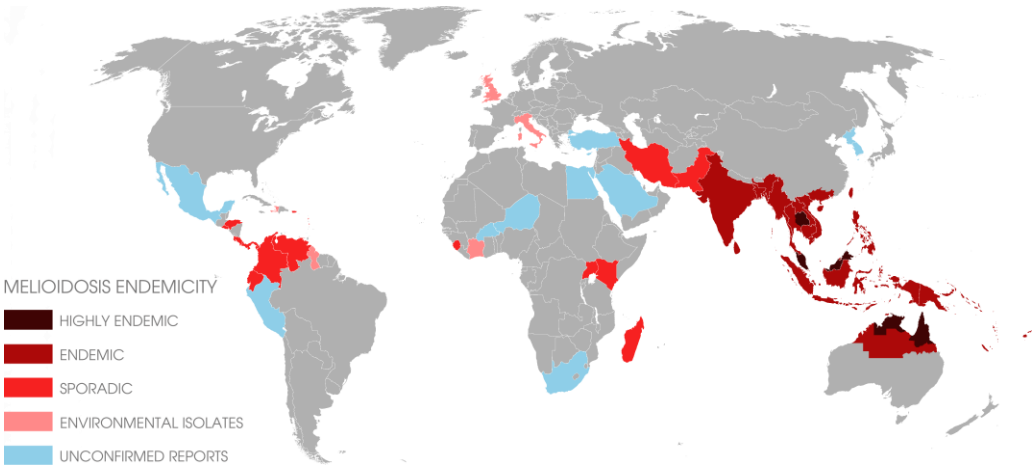


Figure 1: **Endemicity of melioidosis across the globe** (2013)

B. pseudomallei is also considered to have potential in biological warfare, and is regarded as a potential bioterrorist weapon. It appears on the category B list of the critical agents published by the US Centers for Disease Control and Prevention[6].

The preferred host invasion routes for *B. pseudomallei* are mainly inhalation of contaminated dust or droplets, direct contact with contaminated soil or water through penetrating wounds and existing skin abrasions, or ingestion[7]. In few cases it is reported to be nosocomial, sexually transmitted, or laboratory acquired. A number of epidemiological and animal studies have indicated that melioidosis is not contagious[5]. Clinical manifestations are extremely diverse, and vary from acute sepsis to chronic localized pathology, to latent infection, which can reactivate decades later from a yet unknown tissue reservoir[8,9]. The Vietnam War experience drew attention toward more chronic forms of the infection or reactivations long after exposure. Such forms were described as the "Vietnam time-bomb", which manifested later in life as a tuberculosis-like disease[6].

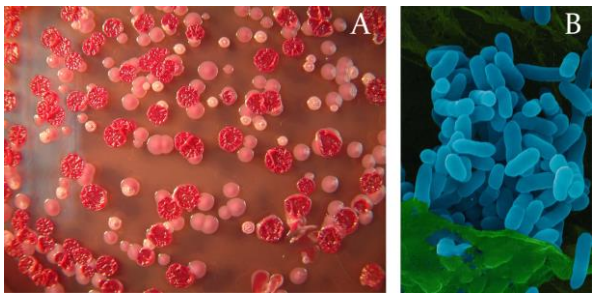


Figure 2: ***Burkholderia pseudomallei***. **A)** Typical colony morphologies. Copyright © 2006 Nature Publishing Group, Nature Reviews | Microbiology. **B)** Scanning Electron Microscopy image of *B. pseudomallei*. Copyright © 2001 Dennis Kunkel Microscopy, inc. /Dennis Kunkel

The most common presentation is community-acquired pneumonia. However, some of the symptomatic presentations such as prolonged fever, weight loss and chest X-ray findings, may be misdiagnosed for tuberculosis leading to wrong therapies. Other than that, there are patients with septic arthritis, multiple abscesses in liver, spleen, prostatic abscess (more common in Australia), suppurative parotitis (more common in Thailand), osteomyelitis, pyomyositis, cellulitis, fasciitis, skin abscesses or ulcers, and bacteremia with or without focus. The protean manifestations of melioidosis often lead to clinical under-diagnosis of this fatal disease, and confirmation of infections relies heavily on time consuming culture methods for bacterial isolation, all of which heavily affects the mortality rate. Healthy individuals can develop melioidosis, but the majority of community acquired cases have some underlying immunosuppressive condition, particularly diabetes and to a lesser extent chronic renal disease, thalassemia or alcoholism[9]. The mortality rate in acute cases can exceed 50% and prolonged treatment with antibiotics may result only in the temporary control of the infection, with 10-15% relapse when antibiotic therapy is withdrawn.

The treatment of melioidosis is often problematic because the bacteria are inherently resistant to many of the commercially available antibiotics and successful therapy often requires extended treatment regimens. Typically, melioidosis patients are treated with parenterally delivered ceftazidime or a carbapenem class drug for 10-14 days, and then with oral trimethoprim-sulfamethoxazole for 12-20 weeks[8,10].

At the present state, there is no vaccine available for use in humans and no candidates are close to licensing[9]. The lack of information on mechanisms of virulence and host resistance has limited work to devise vaccines against melioidosis. However, there are several reports which indicate that it is feasible to protect against the disease, at least in animal models. Atkins *et al*. constructed a mutant of *B. pseudomallei* (2D2) which was auxotrophic in the branched chain amino acid biosynthetic pathway[11]. The median lethal dose of this mutant was greater than 107 cfu in BALB/c mice by the i.p. route whereas the median lethal dose of the parental wild type strain was 80 cfu. Mice which had been dosed with 106 cfu of *B. pseudomallei* 2D2 were protected against a subsequent i.p. challenge with 104 MLD doses of the wild type *B. pseudomallei*[11]. This protection was abolished by mAb depletion of CD4+ but not CD8+ cells from the vaccinated animals prior to challenge, indicating a strong helper cell component to the protection[12]. More recently others have extended this work and demonstrate the ability of other live attenuated mutants to protect against experimental melioidosis in mice[13,14]. While this work with live attenuated mutants provides a clear demonstration of the feasibility of inducing protective immunity, it seems unlikely that a live attenuated mutant derived from a bacterium known to persist in the body would be acceptable as a licensable vaccine, arguing strongly for the development of a non-living vaccine.

There are several reports which indicate that it is feasible to partially protect against melioidosis in animal models of disease, with non-living vaccines. Much of this work has been carried out with polysaccharides derived from *B. pseudomallei*, and the immunization of mice with capsular polysaccharide or lipopolysaccharide is able to provide some protection against the disease[15].

However, polysaccharides alone are generally poor vaccines that do not produce an anamnestic response because of the lack of T cell involvement in the generation of

immunity[16]. In order to convert to a more favorable T-dependent response polysaccharides are often conjugated to proteins. This is particularly important when polysaccharide vaccines are to be used in children, since this group generally responds poorly to this type of vaccine in the absence of a protein carrier[16].

The nature of the protective immune response, at least in murine models of disease, has also been reported in several previous studies. Both antibodies[17,18] and CD4+ T-cells[12,19] have been shown to play key roles in protection. In contrast CD8+ T-cells have previously been shown to play a minor role in the induction of a protective immune response after the immunization of mice with a live attenuated vaccine[19]. As highlighted above, several studies indicate the importance in protection of antibodies against polysaccharide and lipopolysaccharide[15,17,20], but there is also good evidence that protein antigens evoke protective antibody responses[17,20]. Several proteins have been identified as able to provide low levels of protection against melioidosis[21-23]. Starting from this rationale, the development of a non-living vaccine based on specific peptidic antigens appears to be a prime necessity in the effective elicitation of the immune response against meliodisis.

2.2 REVERSE VACCINOLOGY: MAKING VACCINES IN THE POST-GENOMIC ERA

The first practice to be commonly addressed as "vaccination" possibly originated in Asia as early as 1000 CE, based on the inoculation of materials from smallpox lesions to somebody's arm[24]. The practice diffused to Africa and Turkey as well, before approaching Europe, but was formally studied and introduced as a medical practice in 1796 by Edward Jenner, a country doctor living in Berkley, England, who "vaccinated" an eight-year-old boy by inoculating in his arm pus collected from a cowpox lesion on a milkmaid's hand. Six weeks later Jenner repeated the process with material coming from a lesion caused by smallpox, causing no adverse consequences for the boy[25]. After the episode, Jenner conduced one of the first clinical trials in history of medicine, and gathered further evidence from twelve subsequent experiments and sixteen additional case histories. In his

book "*An inquiry into the Causes and Effects of the Variolae Vaccine*"[26] he described and named the first medical immunization, a practice to become in the future one of the most successful scientific advances and medical accomplishments in human history. For an entire century, vaccines referred only to cowpox inoculation for smallpox, and the practice of variolation (the controlled transfer of pus from one person's active smallpox lesion to another person's arm, usually subcutaneously with a lancet) was not short of wariness, hostility or mockery (Figure 3).

A century later, when it was discovered that infections are caused by microbes, French chemist Louis Pasteur started the rational development of vaccines and established the basic rules of vaccinology[27]. He developed what he called a rabies vaccine in 1885, although what he really produced was an antitoxin functioning as a post-infection antidote. Pasteur proposed that in order to make a vaccine, one should "isolate, inactivate

Figure 3: **Caricature exaggerating the widespread debate over the concerns about vaccination**. Cowpox vaccine is being administered to frightened people, while cows emerge from different parts of their bodies. "The Cow-Pock-or-the Wonderful Effects of the New Inoculation!-*vide.* the Publications of y^e^ Anti-Vaccine Society" by satirist artist James Gillray, 1802. (Library of Congress, Prints & Photographs Division, LC-USZC4-3147).

and inject the microorganism"[28] that causes the disease. Pasteur's rules were followed for a century by vaccine developers, introducing vaccines for a number of serious infectious diseases. Jonas Salk used a formaldehyde treatment to kill a poliovirus producing the first vaccine effective against poliomyelitis. The second, oral vaccine was developed by Albert Sabin[29] by attenuating the virus with repeated passages through nonhuman cells at subphysiological temperatures. Maurice Hilleman developed vaccines against measles, mumps, and rubella by viruses attenuation[30]. Others, such as Ramon and Glenny, isolated essential components from bacterial or viral cultures, inactivated them, and paved the way for the development of vaccines against diphtheria and tetanus[31,32], *Neisseria*

meningitidis, Streptococcus pneumoniae, Haemophilus influenzae, and so on. In the case of hepatitis B, it was found that the causative agent could not be cultured in vitro. As a result, the vaccine was initially developed by inactivating viral antigen present in the plasma of chronically infected people[33]. The vaccines developed using Pasteur's rules became powerful tools in the history of medicine and, in less than a century, led to the eradication of some of the most devastating infectious diseases globally.

By the end of the 20th century, live vaccines had been widely used for vaccination against all sort of infectious diseases. Live vaccines have the ability to replicate and are recognized as foreign by human body without causing any pathological or lethal effects. In this period a remarkable progress was made by the introduction of recombinant DNA technologies and chemical conjugation of proteins to polysaccharides, as well as advances in the use of novel adjuvants. Such advancements enabled researchers to specifically inactivate essential gene(s) involved in pathogenesis, ultimately leading to the development of well-defined attenuated live vaccines[34]. A different strategy involves the use of killed whole cell pathogens, non-living vaccines which remain immunogenic for host immune system. This type of vaccine is known to induce a narrow range of immune responses due to the inability to replicate in the host, making them safe for vaccinations on immunocompromised individuals, but usually require multiple doses for long lasting protection. Such strategy includes the vaccines against anthrax, Q fever and whooping cough[35].

However, in spite of all recent technologies, the original empirical approach for vaccinology introduced almost a century before was starting to wear out. The limitations of the traditional formula were evident for pathogens incapable of growing in vitro (e.g. papilloma virus), for pathogens with an intracellular cycle (tuberculosis, malaria), in which the infection is predominantly controlled by T-cell intervention rather than humoral immune response, and for microbes presenting antigenic hypervariability, such as serogroup B meningococcus, or HIV. Most of the vaccines that could be discovered by traditional means were a reality, and the remaining pathogens required new, game-changing technology to be tackled.

A powerful and game-changing tool came from the ability to access the genomes of microorganisms, a new technology that became available in 1995 when Craig Venter published the genome of the first free living organism[36]. This technological revolution allowed for the first time the capacity to move beyond the rules of Pasteur, using the computer resources to rationally design vaccines starting with information present in the genome, without the need to grow the specific microorganisms. This new approach was denominated Reverse Vaccinology (RV)[37].

The first pathogen addressed by the RV approach was Meningococcus B (MenB), a pathogen that causes 50% of the meningococcal meningitis worldwide, it can kill within 24 hours and cause serious life-long disabilities[38]. This bacterium had been refractory to vaccine development because its capsular polysaccharide is identical to a human self-antigen, whereas the bacterial surface proteins are extremely variable[39]. With the full sequencing of MenB genome, researchers were able to identify *a priori* with bioinformatic analyses the sequences of all potential surface located proteins. 600 potential antigens were produced and tested for antigenicity, revealing 29 previously unknown surface exposed proteins capable of inducing antibodies able to kill bacteria in vitro in the presence of complement. Up until that time, only 12 surface antigens of MenB were known, and of these, less than half showed bactericidal activity[40,41], revealing the power of the new approach. In subsequent years, the antigens discovered by this approach inducing the best and broadest bactericidal activity were selected and inserted into prototype vaccines that were able to induce protective immunity against most of the MenB strains in mice[42]. After successful preclinical and clinical studies, the vaccine has been approved by the authorities of Europe, Canada, Australia and USA (2014) with the commercial name Bexsero.

During the last decade, reverse vaccinology has been applied to many other bacterial pathogens, expanding the original scheme to take advantage of the latest technologies in sequencing and bioinformatics. It is the case of B Streptococcus, whose vaccine has been identified with a pan-genomic approach to account for sequence variability across the different strains, developing a combination of antigens which is able to protect against all serotypes[43]. The RV approach was also fundamental in the discovery of an effective vaccine for A Streptococcus.

The intrinsic difficulty here was represented by the cross-reactivity of the antibodies induced against the pathogen, targeting human antigens. RV was essential to ensure that the selected antigens did not have homology to proteins encoded by the human genome[44]. A similar approach was being used also in the development of protein-based vaccines for other antibiotic-resistant pathogens, such as *Staphylococcus aureus* and *Streptococcus pneumoniae*, and vaccine against *Chlamydia* is in progress of development[45].
The Reverse Vaccinology strategy, with its early successes and fast improvements to the original formula, has therefore revived the expectations to conquer pathogens that were previously considered difficult or impossible to address following the traditional rules of vaccinology.
Traditionally, live attenuated vaccines are likely to induce a long-term protection (at least a decade) against the disease[46], and this may constitute one of their major advantages. However, when thinking of a live attenuated vaccine against melioidosis, it may prove difficult to license such a mutant for human immunization. *B. pseudomallei* has the potential to reach inside the cells of its host and persist for long periods of time (even years or decades), establishing latent infections and potentially revert into a life-threatening bacterium. For this reason, among the approaches that are currently pursued to discover and produce an effective vaccine against *B. pseudomallei*, RV plays a crucial role.

2.3 THE GtA CONSORTIUM: THE RV APPROACH AGAINST MELIOIDOSIS.

Following a proteome wide scale study to identify proteins recognized in individuals who have been exposed to *B. pseudomallei*[47], the project "From Genome to Antigen: a Multidisciplinary Approach towards the Development of an Effective Vaccine against *Burkholderia pseudomallei*, the Etiological Agent of Melioidosis" (GtA) was established. This academic consortium is devoted to the identification of a handful of *B.pseudomallei* antigens (i.e. 2 to 5) capable of eliciting a broad, sterilizing immunity and that could subsequently lead to the production of a vaccine against this life-threatening pathogen.

In the original work performed by Felgner *et al.*[47], the authors fabricated a protein microarray containing 1,205 proteins from *B. pseudomallei*, probed it with 88 melioidosis patient sera, and identified 170 reactive antigens. After a second testing on a smaller array probed with a collection of 747 individual sera including material from melioidosis patients from different endemic and non-endemic areas of Thailand and Singapore, they were able to identify 49 antigens, that are significantly more reactive in melioidosis patients than in healthy people and in patients with other types of bacterial infections. One of the tasks addressed by the GtA project is the investigation of the immunogenic potential of each individual protein highlighted by this study, and in parallel use means of RV vaccinology to refine the original list and restrict all the efforts toward the most promising targets. In this regard, the GtA consortium has embarked into a deep study in the genomic and transcriptomic characterization of the bacterium.

Thanks to a large multi-chromosomal genome, one of the largest and most complex among any species of bacteria, *B. pseudomallei* is able to persist and survive in a multitude of environments. The first strain to be fully sequenced, *B. pseudomallei* K96243, contained approximately 6,332 predicted coding sequences within 7.25 Mb of DNA spread across two circular chromosomes[48,49]. This large genome encodes an unparalleled arsenal of virulence factors, including three type III secretion systems (T3SS), six type VI secretion systems (T6SS), multiple antibiotic resistance factors, and at least four polysaccharide gene clusters, including a capsular polysaccharide[48,50,51]. In addition, the genome shows a high degree of plasticity, enabling the bacterium to acquire genomic islands by horizontal transfer. In an effort to research and discover effective vaccine component against this pathogen, a broad vision and understanding of the survival and pathogenesis mechanisms are necessary, to focus on the most effective candidate proteins.

In this regard, comparison between *B. thailandensis,* a less aggressive environmental pathogen, seldom reported to cause melioidosis-like opportunistic infections, and *B. pseudomallei* genomes, may highlight the features responsible for the strong difference in virulence between the two species. Genomic studies have shown that, in spite of a strong difference in pathogenicity, *B. thailandensis* shares a large set of putative virulence factors with *B. pseudomallei*[52,53] and may be considered an adequate model to study the

specificities at the expression level of the two species. Among the activities programmed for the GtA project, Peano *et al*.[54] studied the *B. thailandensis* strain CDC272l121, a known clinical isolate[55], and investigated its response to oxygen availability and different growth temperatures (28°C versus 37°C), mimicking host infection conditions and environmental conditions of tropical regions. The global gene expression and cell surface-associated protein productions were determined along with the assessment of motility and capsular polysaccharide production.

Another transcriptional study was performed directly on *B. pseudomallei*[56] via whole-genome transcriptome profiling covering a broad spectrum of conditions and exposures, compiling a so-called "condition compendium". The analysis confirmed many previously-annotated genes and operons, and at the same time identified novel transcripts including anti-sense and non-coding ones. With a systematic approach of expression analysis, held with the comparison between different organisms and growing conditions, the search for putative antigenic proteins is going to be refined concurrently with integration of new data. In addition, the presence of specific conditions can help ascribe putative functions to previously uncharacterized genes and identify novel regulatory elements.

Another important piece of the puzzle has been laid by Moule *et al*. with a comprehensive list of putative essential genes compiled by the construction of a mutant library of transposons, consisting of over 10^6 *B. pseudomallei* K96243 mutants and the sequence analysis of this library[57]. The list includes known housekeeping genes involved in primary metabolic pathways, as well as core lipopolysaccharide biosynthesis genes and many genes encoding hypothetical proteins that have not previously been established as essential.

All these multi-disciplinary efforts contributed to the definition (and refinement) of a list of putative targets to be analysed individually by a second branch of the GtA consortium, that defines the framework of this thesis project.

2.4 STRUCTURAL VACCINOLOGY: FROM ANTIGEN TO EPITOPE

Within the broad framework of the GtA project, which aims to identify a small number of peptide/protein antigens with protective potential against melioidosis, the RV approach for the identification of the best protein candidates plays a crucial role. Most importantly, GtA aims to demonstrate the feasibility of complementing RV with structural information on the selected antigens: this constitutes the basis for the rational design and/or redesign of novel biomolecules, such as peptides, modified antigens or their domains/fragments as actual vaccine components. This approach constitutes a new view of vaccinology termed Structural Vaccinology (SV).

The SV approach combines structural biology, immunology and computational sciences to investigate all aspects of immune recognition from full-length antigens to epitopes. From the advent of DNA recombination, the focus of vaccinology has in fact gradually shifted from the full pathogen to a more rational approach concerning individual antigens, based on the principle that the immune system does not react against a pathogen per se, but recognizes specific antigens functioning as triggers for the elicitation of the host's defences. The adaptive, long term immunity is no exception, and can be stimulated by the combination of potent immunogens in a unique vaccine.

Although a significant protection can be achieved by the use of strong antigens such as bacterial lipopolysaccharide or glyco-conjugates[15,58], these components cannot be processed and presented to the T-helper cells (CD4+), lacking the principal inductor of memory B cells. For this reason they may act as potent adjuvants but must be delivered together with a peptidic component, that is processed in the proteasome (for MHC type I fragments) or in the phagolysosome (for MHC type II fragments), and loaded on MHC molecules to be presented for recognition by CD8+ and CD4+ T cells respectively[59,60]. The short peptidic fragments used by the immune system machinery as beacons to activate the T cells are immunogenic epitopes. The antigen-antibody interaction, which is the mechanism at the core of the humoral immunity, is also based on the recognition of small, and very specific linear and/or discontinuous epitopes on the surface of the antigen. This recognition is at the basis of the recruitment and activation of the complement cascade, marking the pathogen for ingestion by phagocytosis (opsonisation), and forming a

membrane attack complex (MAC), transmembrane channels capable of causing the bacterial cell lysis[60,61]. In addition, the antigen-antibody recognition is essential in the mechanism of agglutination, where antibodies coating the bacteria form aggregates, thus recruiting other effector cells like phagocytes, mast cells, neutrophils and natural killers, actually clearing the infection by phagocytosis and an arsenal of cytotoxic molecules[59,60].

In this context, structure-based antigen design[62], a fundamental constituent of SV techniques, has generated high expectations toward a 'new era' of fast and safe vaccine development. The approach relies on the concept that by knowing which parts of an antigen are effectively responsible for antigenicity, it is possible to engineer the native antigen to optimize its properties as vaccine candidate. In particular, identifying those components within the antigen structure that elicit protective immunity can potentially permit to carry out antigen manipulation driven by structural considerations (e.g., domain stabilization, conformational constraints). In this way, efforts toward vaccine optimization can be focused on those antigen regions that play a significant role in immunity[62]. As a further development of the SV approach, it may then be viable to select only the antigen regions able to elicit an immune response and translate them in the form of peptides, small proteins or conjugates as potential vaccine candidates.

During the last decade, scaling down a vaccine from the full organism to the isolation of some key elements helped overcome the limitation of previous approaches and find a solution for those pathogens against which no effective vaccines existed (see previous chapter). Now the advent of computer-assisted and computer based methodologies for epitope prediction and modification allows further efforts toward structure-based antigen and epitope optimization, to improve over the use of the native antigen structure in different aspects.

For instance, a landmark integration of atomic-resolution information with computational techniques was reported by Schief and collaborators working on a hybrid method for the grafting of functional motifs onto unrelated protein scaffolds to accurately replicate the antigenic surface recognized by target antibodies[63-65]. Epitope scaffolding strategy, by mimicking complex antigenic targets, might be particularly useful to reproduce discontinuous epitopes and enhance the extent of the immune response against antigens for which elicitation of antibodies has been demonstrated to be

particularly challenging. Indeed, the scaffold-supported antigen is 'presented' in a context that lacks pathogen defensive mechanisms that have evolved to elude the immune response[66]. In one example, the authors' work focused on transplanting a two-segment discontinuous HIV gp120 epitope on a scaffold suitable to accommodate the additional motif without altering its original functional conformation[63]. Scaffold selection and design for optimal motif transplantation were computationally assisted, as well as the generation of a small set of mutagenesis libraries to undergo functional screening. The authors were able to generate a scaffold-bound motif displaying specificity and affinity for antibody recognition similar to the original gp120. This strategy may be then potentially suitable for using grafted epitopes as immunogens to elicit neutralizing antibodies. The plasticity of the instruments and methods of the latest SV strategies offer the chance to challenge a whole new level of biomolecular mechanisms, such as the maturation of a germline of broadly neutralizing antibodies (bNAbs). A similar strategy has been employed for VRC01-class bNAbs, a recently discovered class of antibodies targeting the CD4 receptor binding site on the envelope protein of HIV[67-70]. These antibodies are detected in patients after years of maturation post infection, thus very unlikely to be elicited by a common vaccine unless through a complex and lengthy regimen of immunizations. In addition, the precursor antibodies for VRC01-class lack affinity for the wild-type envelope protein[68,69,71], complicating the efforts to produce a vaccine capable of eliciting the maturation of these bNAbs. In a recent study from Jardine *et al.*[72] computational interface design was used in combination with yeast display screening of computation-guided mutagenesis libraries in a process of maturation of the antigen, increasing the affinity for the antibodies and optimizing immunogenicity with nanoparticle-based antigen presentation.

Another application of the antigen and epitope identification paradigm followed by rational design, is the adaptation of SV strategies to the search of vaccines against antigenically variable pathogens, such as the efforts made on antigen fHbp (factor H binding protein) of *Neisseria meningitidis* serogroup B (MenB). In this study, multiple epitopes from variants of fHbp were grafted on a single fHbp structure, in order to elicit a broad immunity across different strains[73]. Another study addressed the structure of type 2a pilus (BP-2a) from Streptococcus Group B (GBS), identifying an epitope-carrying domain responsible for the elicitation of protective antibodies in six different protein

variants. From the 3D structure of one isoform, the authors synthesized a construct composed of all six epitope carrying domains from all variants, eliciting a broad protection in mice[74].

Aside from challenging currently untreatable or unpreventable infectious diseases, similar strategies may be pursued to explore non-conventional solutions to non-infectious diseases, such as neoplasias, and the increased safety demands from regulation authorities are encouraging new attempts to develop protein and peptide-based vaccines for a diverse set of diseases.

In the framework of the GtA project on melioidosis, we tested and validated a novel SV approach for the identification and design of immunogenic epitopes from the structural data on antigens isolated from the bacterium and prioritized using the RV methodologies described in the previous chapter.

2.5 THE DIGITAL REVOLUTION OF PROTEIN SCIENCE.

Computational chemistry and computational biology were born by the combination of physical chemistry and structural biology with modern informatics and physics, and applied in a range of settings to assist in solving chemical or biological problems. The foundations of the theories and methods behind these disciplines date back to the discoveries in the history of quantum mechanics, and the first theoretical calculations on valence bonds by Heitler and London in 1927[75]. Later books influencing the development of computational quantum theories applied on chemistry include works from Linus Pauling and E. Bright Wilson[76], as well as Heitler[77] and Charles Coulson[78] .

With the advent of efficient computer technology in the 1940s and 1950s, it first became possible to think of the solution of complex atomic systems wave equations as something realizable. The first *ab initio* calculations on diatomic molecules were performed in 1956 at MIT using the Hartree-Fock approach with a minimal basis set. In parallel, a novel computationally-effective method to perform simulations of chemical particles and calculate statistical mechanical properties based on classical approximations was in

invented by Stanislav Ulam working on nuclear weapons projects at the Los Alamos National Laboratory, USA. The name Monte Carlo method, was given by Nicholas Metropolis after the Monte Carlo Casino[79]. This new method encompasses algorithms that rely on repeated random sampling in order to obtain the distribution of an unknown probabilistic entity, and were subsequently employed for the simulation of complex chemical and biological systems. Following the success of the Monte Carlo method, during late 50s and 60s, a novel method for the simulation of chemical and biological systems was theorized by Alder and Wainwright, despite the lack of computational resources for that time in order to fulfill the original vision[80]. The Molecular Dynamics method, as it was named, was (and is) not intended for the simulation of atomic systems by the approximation of the actual wave function; rather, it starts from the assumption that an atomistic system may be approximated to the rules of classic mechanics and treated according to their chemical properties, maintaining their functional characteristics intact. In this scenario, the position and force interactions of the simulated particles evolve through time according to a set of parameters set in the *Force Field*.

Nowadays, the advancing pace of computer science providing better computational capabilities at a continuous rate, along with the modern software implementations, are helping the solution of a whole new level of chemical and biological challenges. The modern methods of computational chemistry allow the generation and the manipulation of novel or existing molecular structures, with a range of applications that include the theoretical investigation of physico-chemical properties of biological macromolecules, to the practical design and simulation of new pharmaceutical compounds.

With regards to the investigation of complex biological systems, like ligand-receptor interactions or cell membrane simulations, the principal methods employed are still based on classic potentials, such as Molecular Mechanics, Molecular Dynamics and Docking algorithms, but the scale of the system is rapidly changing, allowing to perform computationally demanding calculations in a reasonable amount of time. Outstanding examples include the use of parallel GPU computing to obtain the first complete atomic structure of the HIV capside, composed of more than 64 million atoms[81], and the introduction of new supercomputers like Anton, allowing to crack the realm of the millisecond simulations for conventional protein systems in 2009[82].

Such technical accomplishments are not merely technological showcases, but are followed by a parallel non-linear growth in the number of deposited and freely accessible protein structures data, including an increasing number of protein and protein-nucleic acid complexes, and full biological assemblies. As of September 2014, the Protein Data Bank features more than 100,000 protein structures, with a growing pace of almost 10,000 structures deposited in the last year and 2,200 items referred to human elements[83]. The increasing amount of data coupled to the latest computational resources are a precious tool in service of molecular modeling, simulation and design approaches, crucial to the field of Structural Vaccinology, as reported in the previous chapter.

2.6 MOLECULAR SIMULATIONS: APPROACHES, FUNCTIONAL FORMS, POTENTIAL AND LIMITATIONS.

As it is not possible to observe and interact with individual atoms or molecules directly, computational chemistry can help describe and/or predict the properties of a system by modeling. Every molecular system presented in a computational model will be strongly dependent upon the sophistication of the model itself. Since a higher fidelity comes at the price of computational resources and simulation time, it is common practice to choose the simplest representation that will illustrate the property of interest satisfactorily. Since all macromolecules (DNA, proteins, lipids, etc.) are dynamic entities, most experimental properties can be investigated by reconstructing the motions or dynamics of a molecule numerically. This can be done by computing a *trajectory*, i.e. a series of molecular configurations as a function of time by the simultaneous integration of the Newton's equations of motion. MD simulations are set to explore the time dependent behavior of atomic and molecular systems, providing a detailed description of the way in which that system changes from one conformation or configuration to another. Importantly, simulations can generate ensembles of representative configurations in such a way that accurate values of structural and thermodynamic properties can be obtained with a reasonable and feasible amount of computation.

In Molecular Dynamics simulations, the motions of the particles obey the laws of classical mechanics, which is a suitable approximation for a wide range of applications where electronic motions and reorganizations are not involved (i.e. chemical reactions cannot be studied). This includes complex biomolecular systems or material science to the study of polymers. Each iterative configuration of the system is generated by the application of Newton's laws of motion ($F = ma$) for every atom in the system. The end result is a trajectory that specifies the variation of the positions and velocities of the particles in time, and the properties of interest are calculated as time or ensemble averages.
Newton's second law is actually a differential equation that can be re-written as:

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i} \qquad (\mathbf{1})$$

which describes the motion of particle i of mass m_i along coordinate x_i, and subject to the force F_{xi} originating from the presence and interaction of all other particles in the system with atom i. The force can be expressed as the negative derivative of a potential function $V(r_1, r_2, r_3, .. r_N)$ describing the fundamental types of interactions in the system.

$$F_i = -\frac{\partial V}{\partial r_i} \qquad (\mathbf{2})$$

In this case, V can be considered as the potential energy of the system as a function of atomic positions. The equations are solved simultaneously in small *time steps, dt*. The system is followed for a settled time, keeping the number of particles constant, and temperature and pressure at the required values (NPT conditions). The coordinates are written to an output file at regular intervals generating the trajectory.
A typical potential function for all-atom protein simulations is expressed in the form:

$$V(r_1, r_2, ...r_N) = \sum_{bonds} \frac{1}{2} K_b [b - b_0]^2 + \sum_{angles} \frac{1}{2} K_\theta [\theta - \theta_0]^2 + \sum_{\substack{improp \\ dihedrals}} \frac{1}{2} K_\zeta [\zeta - \zeta_0]^2 +$$

$$+ \sum_{dihedrals} K_\varphi [1 + \cos(n\varphi - \delta)] + \sum_{pairs(i,j)} \left[C_{12}(i,j) / r_{ij}^{12} - C_6(i,j) / r_{ij}^6 + q_i q_j / (4\pi\varepsilon_0 \varepsilon_r r_{ij}) \right] \quad (3)$$

The potential *V* reported through equation **3** is called *Force Field* (FF)[84,85]. Throughout the projects reported in this thesis, we use predominantly the GROMOS 53a6 force field as implemented in the Gromacs Package[86]. Each term describes a pairwise relation, taking into account the physico-chemical interactions present in the system. In particular:

$$\sum_{bonds} \frac{1}{2} K_b [b - b_0]^2$$

This term describes the vibrations of covalent bonds around their equilibrium positions b_0. In a classical framework, the vibration is simply expressed in the form of harmonic potential with a spring constant K_b, since the strength of the bond only allows slight fluctuations around the equilibrium values.

$$\sum_{angles} \frac{1}{2} K_\theta [\theta - \theta_0]^2$$

An analogous formulation for the description of the angular vibration around the equilibrium position.

$$\sum_{\substack{improp \\ dihedrals}} \frac{1}{2} K_\zeta [\zeta - \zeta_0]^2 \text{ and } \sum_{dihedrals} K_\varphi [1 + \cos(n\varphi - \delta)]$$

This term is used to describe dihedral dependencies within the molecule. The "improper" dihedral term, also expressed with harmonic potential, functions as a correction factor for out-of-plane deviations (e.g. to keep benzene rings planar). The dihedral potential for 1-4 torsions is described by a cosine expansion and may take any value within 360° depending on the height of the barrier between the low energy conformations, which makes the precision of the dihedral potential barrier crucial for many polymer properties.

$$\sum_{pairs(i,j)} \left[C_{12}(i,j)/r_{ij}^{12} - C_6(i,j)/r_{ij}^{6} + q_i q_j /(4\pi\varepsilon_0\varepsilon_r r_{ij}) \right]$$

The final term in the equation describes all non-bonded interactions. The first term is a Lennard Jones potential, taking into account Van der Waals interactions, while the second one is a formulation of the classical Coulombian electrostatic potential between two charges.

The Force Field is dependent upon various parameters, defining equilibrium distances or force and dielectric constants. The parameters are determined by either experimental data, or by fitting to high level *ab initio* calculations. Due to their parametric nature, the Force Fields are continuously updated and refined to approximate effectively one or more properties of the system they intend to represent. Adding complexity to a FF to increase a model's fidelity does not automatically produce better simulations, while it automatically increase the computational cost (and time). Thus, the choice of the Force Field must be carefully tailored on the specificity of the system.

The initial state of an MD simulation is the assignment of positions and velocities to all atoms/particles in the system. In the case of all protein simulations performed for this thesis project, the initial positions coincide with the structural coordinates present in the PDB file, surrounded by a pre-equilibrated solvent bath of water molecules enriched with Na+ or Cl- counterions to neutralize the protein net charge. Each atom is assigned a velocity according to a uniform Maxwellian distribution of velocities consistent with the temperature at which the simulation will be run.

Ideally, a simulation should be able to reproduce the behavior of an infinite system, so that macroscopic quantities would be calculated straightforwardly. This is totally out of reach even for the most powerful computers, and one has to study finite-size systems characterized by some boundaries. In order to compensate for this limitation, we used Periodic Boundary Conditions (PBC) to represent our systems.

PBCs enable a simulation to be run using a relatively small number of particles, in such a way that the particles experience interactions and forces as if they were in a bulk fluid. A simple representation of a PBC box of particles is shown in Figure 4.

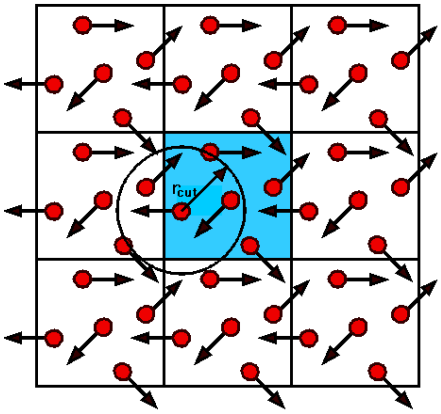


Figure 4**: Example of a bi-dimensional PBC box containing moving particles.** In order to render the system as a continuous one, the original box (blue background) is replicated by translation as virtual images located adjacent the edges. The velocity and interaction forces calculated on every particle (e.g. the particle at the centre of the black circle) are dependent upon the position of the nearest particles inside the box and the nearest image particles present in the image boxes.

The simulation box is replicated by translation at the edges in a series of image boxes, identical in all aspects to the original one. If a particle leaves the box during a simulation, it will be replaced by its image coming in from the opposite side of the box. The number of particles in the simulation box remains fixed, but the solvent behaves as a bulk with no border effects.

PBC conditions are a convenient solution to a major limitation of atomistic simulations, but are not free from drawbacks. For instance, the range of the interactions in the system may be a problem when considering long-range potentials such as the electrostatics. A common solution relies on the use of truncation or cutoff approximations to the treatment of long-range forces, to avoid generating artefacts such as having a particle interacting with its own periodic image (and so with itself).

When the simulation starts, at each step the algorithm calculates the forces acting on every atom then integrate Newton's equations of motion to calculate the position of the particles at the given time step (equation 1), building the trajectory.

2.7 EPITOPES AND SIMULATIONS

Antibody-binding epitopes are conventionally classified as continuous, i.e. sequential and relatively short peptides from the protein sequence able to bind anti-antigen antibodies, or discontinuous, i.e. patches of atoms/fragments from not contiguous protein regions which are brought to close proximity by protein folding and whose antigenicity depends upon the protein conformation. It is worth underlining that epitopes do not exist as discrete and structured entities, but rather have fuzzy boundaries and are defined by their functional ability to bind antibodies[87]. While continuous epitopes can in theory be predicted out of the protein sequence by consensus of available datasets of immunogenic peptides[88], the prediction of discontinuous epitopes relies upon the knowledge of the protein 3D structure. Over the last few decades, several approaches have been tested to run epitope predictions from protein structure. Some groups attempted to correlate antigenicity with protein region properties such as solvent accessibility or flexibility[89,90]. Electrostatic desolvation profiles (EDP) method hypothesizes that surface protein regions with a small free energy penalty for water removal may correspond to preferred interaction sites and may, accordingly, in the case of antigens, constitute binding sites for antibodies[91].
Others based their approach on identifying those regions that protrude out of the protein globular surface and relating them to antigenicity[92,93]. In a reverse approach, Molina and colleagues targeted the identification of epitopes within the protein structure through a bioinformatic analysis of sets of mimitope sequences, i.e. randomly generated peptides mimicking epitopes functional antibody recognition properties[94]. By integrating mimitopes' alignments and consensus, the authors were able to identify the original epitope regions targeted by antibodies within the native antigen in a series of case models where antibody–antigen crystal structures were available.

In this direction, our group has recently developed a new way to approach *in silico* epitope prediction based on the integrated analysis of the energetic properties of an antigen, namely Matrix of Local Coupling Energies[95] (MLCE). MLCE integrates the analysis of the

energetic properties of proteins to identify interaction-networks on the surface of the isolated antigens. Such networks correspond to conformational patches that can aptly be recognized by a binding partner (i.e. the antibody molecule). MLCE was designed to select the protein substructures presenting a low energy contribution to the general protein stability, formally representing those sites in which interactions networks are not energetically optimized. According to the low intensity constraints to the rest of the protein, these substructures are characterized by dynamic properties allowing them to visit multiple conformations, a subset of which can be recognized by the antibody[96,97].

From the computational standpoint, MLCE is based on the calculation of the matrix M_{ij} of inter-residue non-bonded interaction energies between residue pairs (thus including only van der Waals forces and electrostatics) using a MM-GBSA (Molecular Mechanics Generalized Born Surface Area) approximation for implicit solvent treatment. The analysis is performed on individual structures visited during an MD trajectory.
According to the Energy Decomposition method[98], all diagonal elements of self-interactions are neglected, and the matrix M_{ij} is diagonalized and re-expressed in terms of eigenvalues and eigenvectors, in the form:

$$M_{ij} = \sum_{k=1}^{N} \lambda_k w_i^k w_j^k \qquad (4)$$

N is the number of protein aminoacids, λ_k is the eigenvalue associated with the *k-th* eigenvector, and w_i^k is the *i-th* component of the associated normalized eigenvector. Eigenvalues are labelled following an increasing order, so that λ_1 is the most negative. Throughout this thesis, we refer to first eigenvector as the eigenvector corresponding to the eigenvalue λ_1. The total non-bonded energy E_{nb} is defined as:

$$E_{nb} = \sum_{i,j=1}^{N} M_{ij} = \sum_{i,j=1}^{N} \sum_{k=1}^{N} \lambda_k w_i^k w^k{}_j \qquad (5)$$

And since M_{ij} can be effectively approximated[98,99] by:

$$M_{ij} \approx \tilde{M}_{ij} = \lambda_1 w_i^1 w_j^1 \qquad (6)$$

The total non bonded energy can be expressed as:

$$E_{nb} \approx E^{app}{}_{nb} = \sum_{i,j=1}^{N} \tilde{M}_{ij} = \sum_{i,j=1}^{N} \lambda_1 w_i^1 w^1{}_j \qquad (7)$$

From this we can recover an approximation to the global stabilization energy, $E_{nb}{}^{app}$, that was shown to correlate with the relative different stabilities of mutants of several test proteins[99]. This method provides information on the mean coupling energy between two residues in the native state, revealing the network of most interacting and non-interacting residues throughout the structure, whose mutation would have a profound impact on the stability of the protein.

Taking from this analysis, the MLCE methods intersects the energy matrix $E_{nb}{}^{app}$ with the contact map of the representative structure from MD. The contact map (also referred as neighbouring list) is a binary representation of proximity between any two residue pair inside the conformation. If the distance between C_β atoms of any two amino acids is below a cut-off value, the corresponding matrix entry is set to 1, otherwise it is set to 0. The distance cut-off is set to 6.5 Å. For the sake of homogeneity with the energy matrix, also contacts between nearest neighbours *i, i+1* are included. Therefore:

$$C_{ij} = \begin{cases} 1 & r_{ij} \leq 6.5 \\ 0 & ri_j > 6.5 \end{cases} \qquad (8)$$

The matrix resulting from the intersection of the simplified energy matrix and contact map is an energy matrix presenting only the local contributions of non-bonded interaction, highlighting pairs within the contact cut-off that are energetically coupled or uncoupled. This matrix is referred as the Matrix of Local Coupling Energies[95] (MLCE).

The coupling interactions inside MLCE are ranked in increasing order according to their respective intensities (from weaker to stronger). Starting from the minimum value (weakest local coupling interactions), the set of putative interaction sites was defined by including increasing residue-residue coupling values until the number of couplings corresponding to a given threshold. According to the benchmark[95], the optimal threshold varies from 10% to 20% of all low-energy pairs, and is set to 15% as standard value. This corresponds in our approximation to the set of local interactions with minimal intensities, and accordingly they could represent those localized regions that are less energetically coupled with the rest of the protein, and that consequentially fit basic criteria for representing epitope candidates. The method in its full implementation for the prediction of epitopes from the 3D structure of an antigen is named BEPPE (Binding Epitope Prediction from Protein Energetics), and it has been chosen along with the EDP profiles as a key resource for the identification of antibody-binding sites inside the GtA project. Within a broad Structural Vaccinology initiative, the use of modern day computational methods for simulation and prediction represent an attractive and challenging opportunity for both the identification of epitopes and the structure-based rational design of immunogenic peptides.

3 - AIM OF THE PROJECT

From now on we live in a world where man has walked on the Moon. It's not a miracle; we just decided to go.

James A. Lovell. JR.
commander of Apollo 13

The aim of this project, in the broader framework of GtA, may be subdivided into four, distinct main objectives.

First, my research work aims at providing the GtA project with the computational resources necessary to identify putative binding epitopes with immunogenic potential. The relevance of this task becomes clear when considering the GtA project in a global view, since the first element required to its Structural Vaccinology approach is to connect the 3D structure of one antigen to information on predicted binding epitopes. Once the epitopes are defined, a second step takes place, focused on the rational design of effective vaccine-like components starting from the prior knowledge of the antigen structure and the epitopes location. This endeavor aims to develop an interdisciplinary research pipeline, comprehensive of structural biology, molecular biology and computational chemistry, to fill the gap between structure and epitopes. Therefore, one of the main objectives of my project is to integrate different computational methods in the definition of such pipeline. The methods to be employed revolve around the Molecular Dynamics

simulation approaches, Homology modeling and the energy-based computational method for analysis and prediction of antibody-binding sites defined by Scarabelli *et al.*[95]

The second main objective of this project is the expansion of such defined pipeline beyond its original concept, integrating novel functionalities and exploring different application scenarios. Specifically, the energy-based predictor originally designed and validated for antibody-binding sites will be extended and adapted to the prediction of MHC-II loaded epitopes. A similar functionality is extremely desirable, not only in the perspective of a unique, comprehensive tool for epitope prediction, but also in the specificity of melioidosis infections. As described in chapter 2.1, antibodies and T-helper response are the primary routes for recovery and sterilizing immunity. Another significant expansion of the original concept will be explored in the application of the predictive pipeline in the field of immunodiagnostics, to identify epitopes with the potential to be employed as biomarkers in a diagnostic tool.

The next objective is right downstream of the predictive pipeline. Once an immunogenic epitope is identified, it may be conveniently designed into a mimic of the original protein, in order to maintain the epitope in a stable and soluble form, retaining its original folding conformation and displaying maximal exposure to the immune system. Such a design will be the third objective of this project.

The fourth and last aim of this project is more general, and is not strictly related to the development of a vaccine against melioidosis, but is equally relevant. Starting from the experience gained with energy-based predictor of binding epitopes, the reach of my studies has been expanded to include a dataset of interacting proteins and explore in detail their interaction energetics, both at the inter-molecular and intra-molecular level. The purpose of such a comparison is the identification of an energetic property, indicative of the productive positioning of possible PPI patch. Such property would therefore be employed as a descriptor of the interaction process, leading to the development of a computational tool for the analysis and the prediction of this interaction property. The method, once validated, may be employed in a range of applications, from the ab-initio characterization of PPIs, to protein-protein docking algorithms, to the evolutionary and functional investigation of interactions.

4 - MAIN RESULTS

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

Isaac Asimov

In this chapter, I will summarize the results and perspectives of the works presented in PART II in the form of published or submitted articles, putting them in their proper context within the global framework of the project and according to the four main objectives stated in section 3. I will also present the results of one work still unpublished and in phase of completion.

4.1 DEFINING THE STRUCTURAL VACCINOLOGY PIPELINE

As described in Section 3, the initial focus of the project was on the definition of a working pipeline going from the structure of the antigen to the identification of immunogenic epitopes. Based on the results of a seminal study based on protein microarrays to investigate the interaction of the host immune system with *B. pseudomallei* over 1,000 proteins[47], our shortlist of antigenic candidates was composed of 49 proteins.

Among the first antigens to be further studied for their immunogenic properties, we considered the oligopeptide-binding protein A ($OppA_{Bp}$)[100].

$OppA_{Bp}$ is part of an ATP-binding cassette (ABC) transport system, a group of proteins regarded as potential targets for the development of therapeutic interventions against bacterial infections, given their key role in bacterial survival, virulence, and pathogenicity. Components of several ABC transporter systems, from both Gram-negative and Gram-positive bacteria, have been proposed as candidates for vaccine development because they were shown to react with convalescent patient sera[101,102]. OppA is part of the oligopeptide transport system OppABCDF involved in nutrient uptake and recycling of cell-wall peptides[103]. The transport system consists of five subunits: two integral membrane proteins forming the pore, two proteins responsible for ATP hydrolysis, and a substrate-binding protein (SBP)[103]. OppA is a receptor, or substrate-binding protein, and determines the recognition properties of the system, delivering the substrates to its cognate-binding partners. Previous studies on OppA from different pathogens revealed that OppA from *Listeria monocytogenes* is a virulence factor important for intracellular survival[104] and that OppA from *Yersinia pestis* is a protective antigen[102]. With regard to $OppA_{Bp}$, it is recognized by T cells primed by *B. pseudomallei*, triggers IFN-γ production, and stimulates both humoral and cell-mediated responses. However, sera raised against $OppA_{Bp}$ failed to offer protection in a mouse infection model[22]. Therefore, $OppA_{Bp}$ is seen as a suitable target for structure-based antigen analysis and improvement of its antigenic properties, and was chosen as a reference protein for the definition of our Structural Vaccinology pipeline. In this framework, we studied the crystal structure of $OppA_{Bp}$ (gene access BPSL2141) determined for residues 39-554 at 2.1 Å resolution (Rgen and Rfree values of 15.8% and 20.9%, respectively), devoid of its predicted signal peptide (residues 1–38). The 3D structure of this protein, as inferred from its density map, is shown in Figure 5, and is composed of two lobes, comprising respectively domains AB and C.

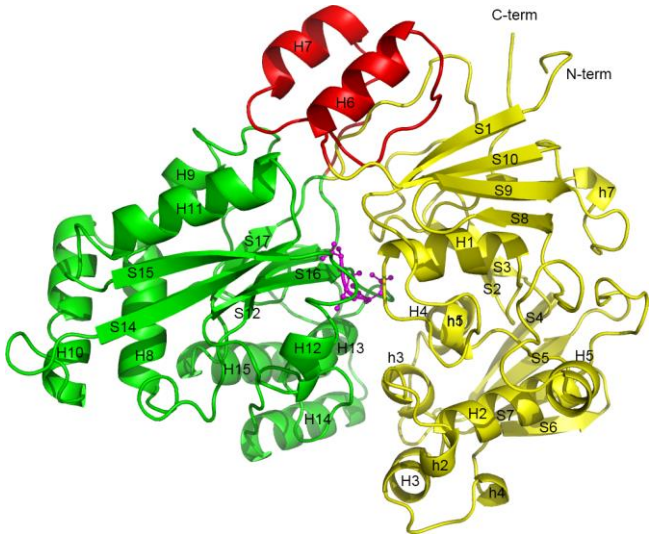


Figure 5: **Tertiary structure of $OppA_{Bp}$**. Secondary structure ribbon representation of $OppA_{Bp}$ bound to its tripeptide ligand (purple balls-and-sticks). Domains A, B and C are illustrated in yellow-red and green, respectively. β Strands, α helices, and 3_{10} helices are labeled S1–S17, H1–H15, and h1–h7, respectively. The N and C termini (N term and C term, respectively) of $OppA_{Bp}$ are labeled. This figure was generated using PyMOL.

Concerning the epitope prediction and detection, we used three different methodologies, two of them consisting in computational methods, and relying on the solved 3D structure and subsequent MD simulation, while the last one consisted of an experimental epitope mapping. The first computational method, named BEPPE, is an energy-based predictor for antigenic epitopes based on the MLCE approach, as described in section 2.6. The second computational method, named Energy Desolvation Profiles[91] (EDP), is more general, having been developed to identify protein-protein interaction interfaces based on the desolvation potential on the surface of proteins, and is being validated for antigen-antibody interactions.

The third, experimental method to be used for the mapping of epitopes used recombinant $OppA_{Bp}$ and cognate polyclonal sera. The approach involves proteolytic digestion (using diverse proteases) of the target antigen prior to immunocapturing and subsequent analysis of antibody-bound peptides by mass spectrometry. The sera collected from three immunized mice were independently analyzed and produced identical results. From the computationally predicted epitopes we drew a consensus prediction leading to the identification of three epitopes (COMP1, COMP2 and COMP3). Similarly, three epitopes were identified via experimental mapping (EXP4, EXP5, EXP6). All six epitopes were produced in the form of BSA-conjugated peptides, and tested for immunoreactivity, against sera collected from 19 healthy donors (subdivided in seronegative and seropositive

categories) and 20 cases of melioidosis recovery. The results of this test are displayed in Figure 6, showing for epitopes COMP1-3 a generally higher reactivity against the recovery sera. COMP1 and COMP3, in particular, were significantly recognized by plasma from seropositive subjects, and interestingly, the reactivity in COMP3 was sufficiently diverse in the three groups to indicate some discriminating potential if used as a biomarker tool. The EXP4–EXP6 peptides showed a distinct reactivity pattern toward plasma of all three groups. In fact, the EXP peptides reacted strongly with the plasma from asymptomatic healthy patients relative to the seronegative individuals. Particularly, EXP5 and EXP6 were not significantly recognized by the plasma from recovered subjects, hinting at the potential application of both peptides for discriminating between asymptomatic versus clinical melioidosis in endemic areas.

From the results of the epitope prediction and mapping, we learnt that both computational and experimental means are effective in the identification of immunogenic antibody binding sites.

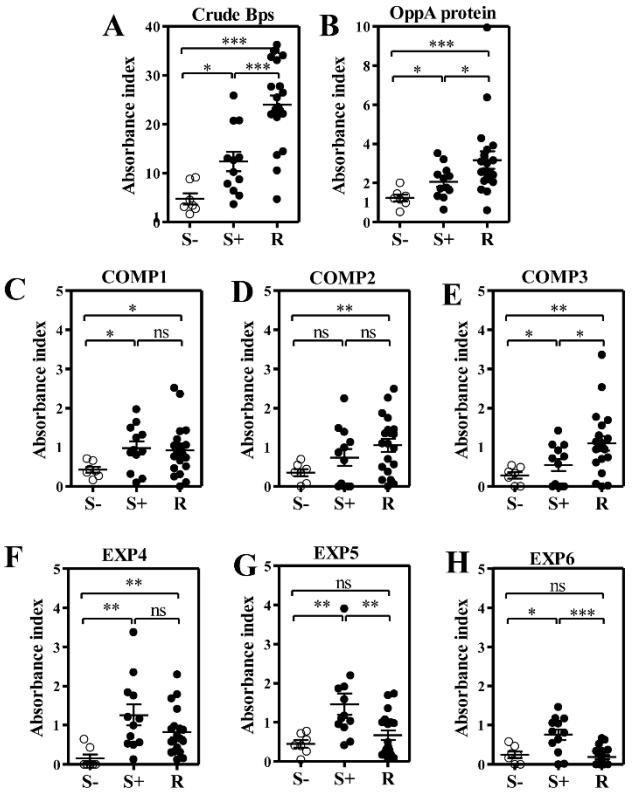


Figure 6: **Antibody response to *B. pseudomallei* $OppA_{Bp}$ epitopes in plasma of healthy and recovered melioidosis subjects**: Crude *B. pseudomallei* antigen (A), recombinant OppABp (B), and synthetic peptides COMP1 (C), COMP2 (D), COMP3 (E), EXP4 (F), EXP5 (G), and EXP6 (H) were coated onto ELISA plates and probed with diluted plasma samples of healthy seropositive individuals (S+; n = 12) tested by means IHA (titer >40), healthy seronegative individuals (S-; n = 7), and recovery melioidosis individuals (R; n = 20), and quantified by indirect ELISA. Data represent the absorbance index (AI) of individual samples = (OD tested - OD uncoated) / OD uncoated.
Experiments were performed in duplicate, and results represent mean (AI) ± SE, Mann-Whitney U test; *p < 0.05, **p < 0.01, and ***p < 0.001 values. ns, not significant.

Immunocaptured fragments are only restricted to linear stretches, while computational methods are able to identify conformational epitopes, that are composed of amino acids in proximity within the 3D structure, even if not close on the primary protein sequence. Conversely, the experimental technique is able to identify potential epitopes located inside the core of the protein (EXP4, EXP5), while MLCE and EDP are restricted to the analysis of the protein solvent accessible area. For this reason perhaps, in addition to the large size of the protein, we were not able to draw a consensus between the experimental and computational predictions, having only three amino acids in common between COMP1 and EXP6. In order to test the validity of the computational predictions in replicating the results from the experimental protein digestion and immunocapturing, we modeled the protein digestion in silico using an energy-based domain decomposition approach that allows in silico dissection of a folded protein into smaller fragments[105]. The underlying hypothesis is that such fragments expose sequence stretches that may be targeted by antibodies under conditions of partial unfolding or degradation of the antigen protein. Application of the domain decomposition algorithm allowed us to identify six cleavage boundaries and cluster the results into three different protein fragments which partially overlap with the $OppA_{Bp}$ structural domains. EXP4 and EXP6 peptides are entirely contained in fragments A' and C', respectively, whereas EXP5 extends across fragments B' and C'. MLCE and EDP predictions were then applied to the isolated fragments, identifying potential epitopes that very satisfactorily overlap with the peptides identified by immunocapture (Figure 7).

Working with $OppA_{Bp}$ we were able to define our prediction and testing pipeline for immunogenic epitopes: the first step is the determination of the antigen's 3D structure, followed by a computational characterization via Molecular Dynamics simulations. The representative conformations extracted from the clustering procedure of the simulations lay the basis for a computational prediction via MLCE and EDP methods. In parallel, an experimental epitope mapping is performed on the full antigen.

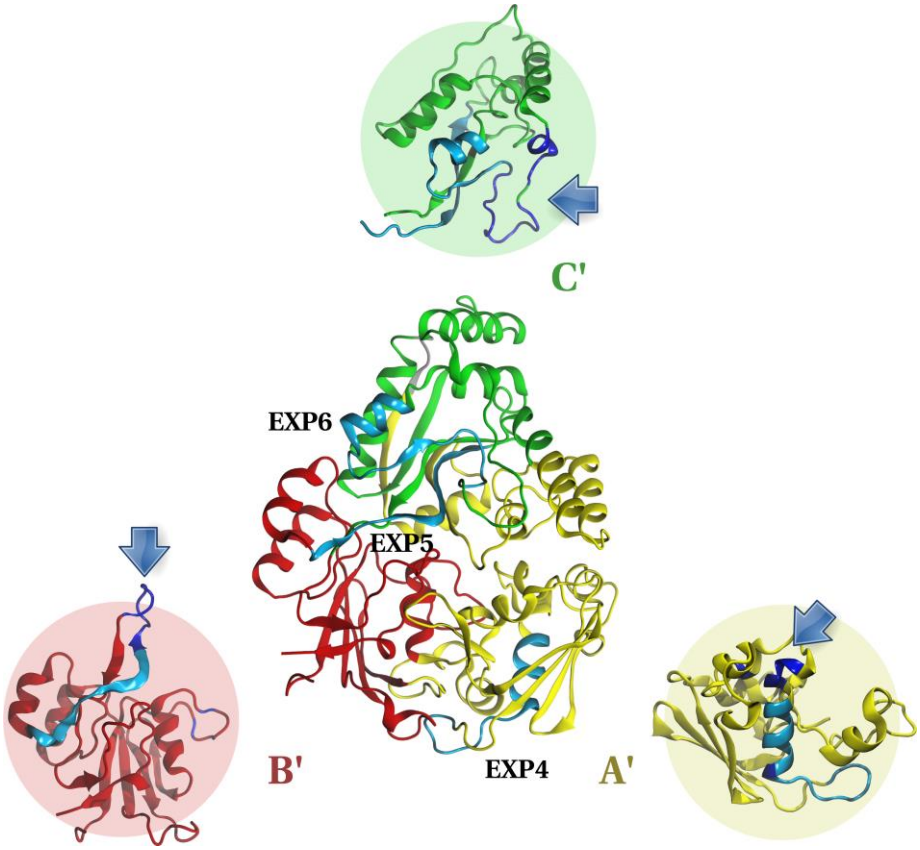


Figure 7**: Energy-Based Domain Decomposition of OppABp and Epitope Prediction**. The 3D structure of OppABp is shown with the defined fragments A', B' , and C' highlighted in yellow, red, and green, respectively. The fragments predicted through the decomposition algorithm are also shown individually in the shaded circles. The epitope regions mapped by immunocapture (EXP4, EXP5, and EXP6) are highlighted in light blue in both the full structure and in the individual fragments. Regions of the EXP epitopes matching those predicted by MLCE plus EDP on the isolated fragments are shown in dark blue and marked by arrows.

All three prediction and mapping methods have weaknesses and differences, so a consensus must be drawn between all three techniques in order to minimize the possible impact of false positives. In case of large, multidomain proteins, if a consensus is indeed not reached, another computational approach based on domain decomposition may help define a consensus not solely based on solvent exposed epitopes. While this is clearly a secondary concern when thinking of antibody binding sites, it may hint at the potential of predicting T-cell epitopes (see section 4.5).

From the analysis of the immunoreactivity of the predicted and synthesized epitopes, it appears very likely that this pipeline may be employed in the successful determination of immunogenic epitopes, to be tested in vitro for opsonic killing activity. In addition, the differential reactivity of COMP3 against seronegative, seropositive and recovered melioidosis patients, suggested the possibility to employ the same pipeline in the serodiagnostic field (see section 4.4).

4.2 APPLICATION OF THE SV PIPELINE TOWARD THE IDENTIFICATION OF ACTIVE EPITOPES.

The work carried out on $OppA_{Bp}$ antigen allowed the validation of our pipeline from the solved 3D structure to the serological tests for immunoreactivity of epitopes. Unfortunately, the reactivities of the epitopes identified within this protein were not strong enough to justify further tests on this candidate, so we moved on according to the priority list. During the three years of this project, our collaborators in Univeristat Autònoma de Barcelona, University of Exeter and CNR-ITB, in charge of refining the shortlist of antigens through RV applications, modified and restricted the original list. Initially comprising 49 antigens, the list was updated at multiple steps, reaching in 2013 the number of 21 targets, subdivided in acute phase antigens (15) and chronic infection antigens (6), and sorted in priority according to expression and antigenicity (Table 1).
We concentrated our efforts on high priority proteins specifically expressed in the acute phase of infection. The first protein in the list is a hemolisin-related protein weighing 325kDa (3,229 amino acids), extremely challenging to study at the structural level. Eventually, during the last two years of this project we analyzed seven antigens, for many of which the work is still in progress at different levels of completion: BPSL2765, BPSL1445, BPSL2520, BPSL3319, BPSL0919, BPSL1050 and BPSL2063 (being respectively number #2, #3, #4, #6, #7, #9 and #12 in priority).

	Name	annotation	Protein length	Priority
ACUTE TARGETS	BPSL1661	putative hemolysin related protein	3229	1
	BPSL2765	putative OmpA lipoprotein	170	2
	BPSL1445	putative lipoprotein	195	3
	BPSL2520	putative exported protein	198	4
	BPSS2053	putative hemolysin	3103	5
	BPSL3319	flagellin FliC	388	5
	BPSL0919	LytB protein	326	6
	BPSL0999	putative OmpA family	215	7
	BPSL1050	hypothetical protein	126	8
	BPSL2277	LolC protein	417	9
	BPSS2013	hypothetical protein	418	10
	BPSL2063	putative membrane protein	1090	11
	BPSS1727	putative haemagglutinin	908	12
	BPSS1384	putative membrane protein	328	13
	BPSS0443	conserved hypothetical protein	449	14
CHRONIC TARGETS	BPSL1897	hypothetical protein	155	1
	BPSL3369	acoD acetaldehyde dehydrogenase	506	2
	BPSL0296	hypothetical protein	74	3
	BPSL3247	put lipoprotein	467	4
	BPSL2287	HesB family protein	107	5
	BPSL1899	fimbriae like assembly protein	56	6

Table 1**: Shortlist of highest-priority antigens of GtA.** The proteins are subdivided according to their disease stage. Acute phase antigens are displayed on top (red), chronic stage antigens are shown at the bottom (yellow). For each protein the gene name is also reported, along with a brief annotation and the protein length (number of amino acids).

The first protein we analyzed is priority #2, BPSL2765, also referred as $OmpA_{Bp}$ or Pal_{Bp}, since it exhibits high similarity to members of the peptidoglycan-associated lipoprotein (Pal) family, involved in maintaining outer membrane integrity and the import of selected organic nutrients. In this work[106] we applied the prediction and testing pipeline validated in the OppA study, using the same computational methods (MLCE and EDP), along with the immunocapturing epitope mapping. In the process we identified three putative epitopes, one of which based on a full consensus of all three techniques. The resulting epitopes were produced in free and BSA-conjugated form, and tested for immunogenicity. The results of the test can be seen in Figure 8, panel A. The epitopes were challenged with antisera collected from healthy donors, once again subdivided in seropositive and seronegative, and patients who recovered from melioidosis. The reactivity against the full protein was used as a reference, and according to the image, it is worth noting that the reactivity of EPITOPE 3, the consensus one, is proportional to the reactivity of the entire protein, both in intensity and in the ability to discriminate between the three categories of individuals. For this reason, this epitope moved to the next phase, the test for opsonic

killing activity (OPK). Antibodies against EPITOPE 3 were produced in rabbit, and added to cultures of neutrophils incubated with *B. pseudomallei*, evaluating the phagocytosis rate, oxidative burst response and total bacterial killing at increasing concentrations of anti-EPITOPE3 antibody (Figure 8, panel B). In addition to the OPK, an agglutination test was also performed with the same antibodies (Figure 8, panel C). According to results, the antibodies raised against the epitope are active in stimulating the immune response in multiple ways, helping the neutrophils clear the culture from the bacteria even in absence of the complement system, and triggering an agglutination response even stronger than the activity registered with antibodies raised against the full protein. For all these reasons, EPITOPE 3 from BPSL2765 appears like an ideal candidate for structure-based design for improvements, and for immunization tests in animal models.

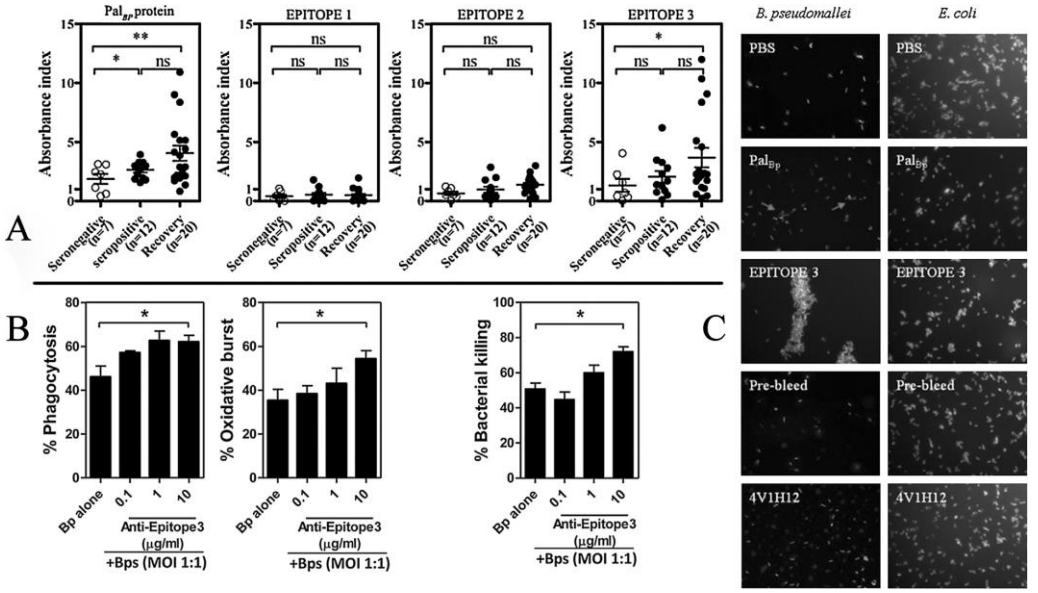


Figure 8**: Antibody reactivity of epitopes 1-3 from Pal$_{Bp}$ and OPK/agglutination tests of anti-Epitope 3 antibodies. A)** antibody response to PalBp protein and peptides in plasma of healthy and recovered melioidosis subjects. **B)** Anti-EPITOPE 3 Increased phagocytosis, oxidative Burst, and *B. pseudomallei* killing by Human Neutrophils. Overall percentages of phagocytosis and oxidative burst by flow cytometry and total bactericidal activity by standard colony counting from three healthy subjects, assayed in the absence or presence of anti-Epitope 3. Data represent means ± SE; *p < 0.05. **C)** Agglutination of *B. pseudomallei* after Exposure to Pal$_{Bp}$ or anti-Epitope 3. RFP-expressing *B. pseudomallei* or *E. coli* were incubated with 1 mg of antibodies for 30 min at 37 C. Bacterial agglutination was observed under epifluorescence microscopy, and the images are representative of three independent experiments. Arrows indicate agglutination of *B. pseudomallei* after incubation with Pal$_{Bp}$.

Another high-ranked candidate to be extensively analyzed was BPSL1050 (priority #9) (see section 10.2 of PART II), an uncharacterized protein of unknown function and novel structure with no annotated analogues in the Protein Data Bank or on the UniProt repository. The 3D structure of this protein was determined via NMR at San Raffaele Scientific Institute, producing a bundle of 20 structures, accounting for some conformational flexibility, especially in the long loops contained in the protein fold.

The regular approach for epitope prediction, validated with a high-resolution structure solved via diffractometric techniques, involves the simulation of the NMR-structure in 3 all-atom MD replicas, covering a simulation time of 50ns each at 300K and explicit water. In this case, the protocol was adapted using as starting coordinates for the replicas the 3D conformations of the first 3 structures in the bundle. The analysis of this target was also extended to include the characterization of the internal dynamics properties of the protein in terms of the principal motions (Essential Dynamics[107]) and the mechanical coordination among different protein parts (Communication Propensity[108]), suggesting an interesting convergence between NMR and computational characterization of the protein dynamic activity. In this case, the epitope prediction and mapping highlighted two epitopes, namely L1050#1 and L1050#2. The epitopes were synthesized in free form and BSA-conjugated for the serological test, as well as KLH-conjugated for rabbit immunization and production of antibodies. The peptide reactivity against human sera, carried out in the same conditions as previously tested with $OppA_{Bp}$ and Pal_{Bp} denoted a poor reactivity of the full protein and the two epitopes in patients' and donors' sera (Figure 9 panel A).

It is however worth of notice that L1050#2 could show an appreciable increase in intensity, compared to the signal of the full protein, indicating one of the advantages of scaling from antigen to epitope: the immune response elicited by the single epitope may vary compared to the response induced with the native antigen. Antibodies were raised against the epitopes and subsequently tested for specificity (they must be able to recognize the full antigen to be protective) (Figure 9 panel B), as well as in agglutination experiments (Figure 9 panel C). While the antibodies appear perfectly able to recognize the protein, they are also marking cross-reactive signals with the peptides, indicating a poor specificity (especially for L1050#1).

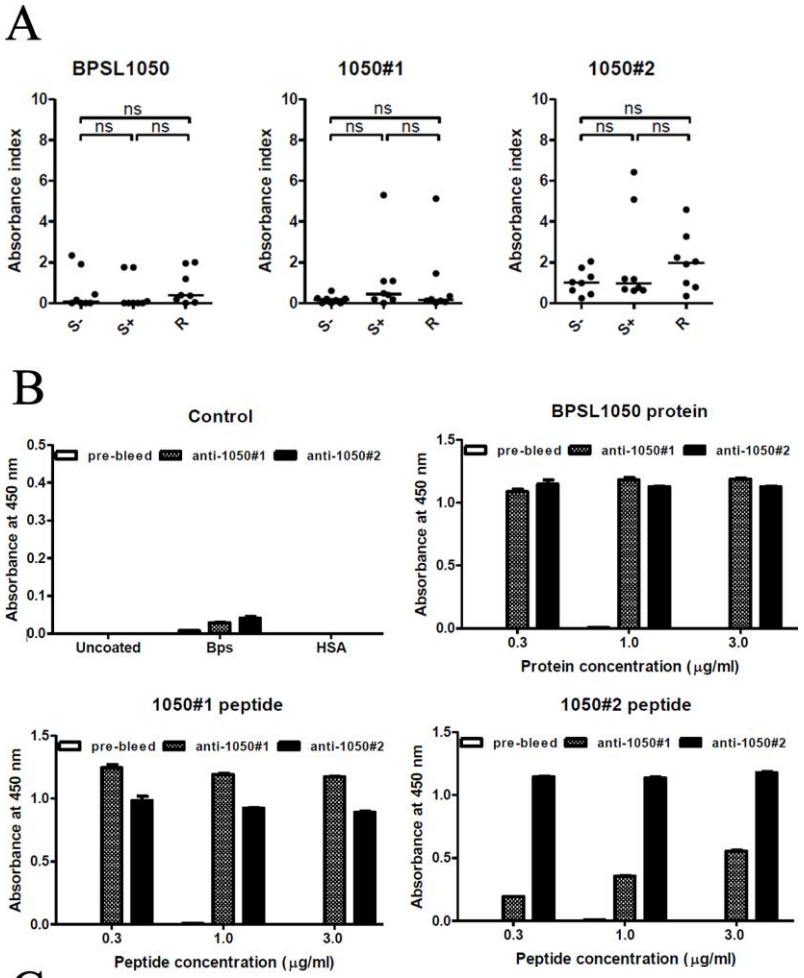


Figure 9**: in vitro characterization of peptides 1050#1, 1050#2 and antibodies anti-1050#1, anti-1050#2. A)** Antibody response to BPSL1050 protein and peptides in plasma of healthy and recovered melioidosis subjects.

B) Recognition of rabbit anti-L1050#1 and anti-L1050#2 sera to BPSL1050 protein/peptides detection by ELISA. Crude Bps antigen, BPSL1050 protein, peptides L1050#1 and L1050#2 were individually coated into 96-well polystyrene plates. 0.5 ug/ml of either un-immunised rabbit sera or sera from rabbits immunised with L1050#1 (hatched bars) or L1050#2 (solid bars) were tested in duplicate for binding to the plate-immobilised antigens. Rabbit antibodies were detected using a suitable HRP-anti-hIgG conjugate.

C) Agglutination of *B. pseudomallei* after exposure to antibodies raised against BPSL1050 or *L1050#1* and L1050#2. RFP-expressing *B. pseudomallei* was incubated with 1 μg of antibodies for 30 min at 37°C.

The agglutination test, on the other hand, is delivering nice agglomerates for both antibodies against both epitopes. The preliminary results on this protein are not sufficiently promising to consider this protein one of our main candidates as a possible vaccine component (unlike EPITOPE 3 from Pal_{Bp}), although LI050#2 and the two antibodies raised against the designed peptides may function as an adjuvant or could be employed as serodiagnostic tools, similarly to $OppA_{Bp}$ Epitope 3.
Concerning the other cadidates, namely BSL1445 (#3), BPSL3319 (#6) and BPSL0919 (#7), we followed a similar approach for epitope discovery, for which we are currently producing experimental data to evaluate their actual response.

BPSL3319, also known as $FliC_{Bp}$, assembles to form the flagellar filament that permits bacterial motility. FliC can induce IFN-γ responses from human T cells and is recognized by antibodies from seropositive individuals living in endemic areas[109,110]. In addition, FliC is recognized as a pathogen associated molecular pattern (PAMP) by Toll-like receptor 5 (TLR5) and nucleotide-binding oligomerization domain (NOD)-like receptor C4 (NLRC4), activating both innate and adaptive immunity[111]. After structural characterization, we employed an even broader approach for epitope discovery, since this target is most likely able to stimulate the response of the T cells, in addition of being targeted by antibodies. In fact, we complemented the MLCE/EDP and immunocapturing epitope mapping approach with sequence-based T cell epitope prediction, (using IEDB, selecting common HLA-DRB1 type in Thailand) and other sequence-based predictors for antibody binding sites (BepiPred[112], BCPred and AAP[113]). From the consensus prediction (see part III, Figure SI) we synthesized four individual epitopes, three were predicted and tested as putative B and T cell epitopes, while the remaining was predicted and tested as a T cell one. Our preliminary, still unpublished results indicate that two out of three B-peptides are strongly recognized in human sera (Figure S2), and the polyclonal antibodies raised against them are also active in OPK tests, enhancing both phagocytosis rate and oxidative burst with similar signal intensity as the full antigen (Figure S3, Figure S4).

When tested for the induction of IFNγ and IL-10 in peripheral blood mononuclear cells (PBMC), all four putative T-cell epitopes were active (Figure S5), confirming this protein as one of our priority ones. Additional tests will be necessary to get an insight of the most reactive epitopes to be advanced to the next phase.

BPSL1445, whose structure was solved via NMR, is another uncharacterized protein for which we identified three epitopes, based on MD simulations started from structures in the NMR bundle (sequences reported in Table S1), which are currently under evaluation for immunoreactivity. BPSL0919 corresponds to the gene locus $LytB_{Bp}$, coding for a 4-hydroxy-3-methylbut-2-enyl diphosphate reductase 1, active in the MEP pathway of isoprenoid biosynthesis. Unlike the other candidates analysed so far, this gene codes for a cytoplasmic enzyme (and not a membrane or extracellular protein), though our evidence suggests that this protein is highly immunogenic and specific of acute infections. According to the structural analysis based on homology modelling (58% identity and 82% similarity with *E.coli* IspH), $LytB_{Bp}$ is composed of three well-defined domains forming a planar shape of three lobes surrounding the iron-sulphur binding site. Because of the difficulties encountered in the cloning and expression of the full protein, epitope predictions were based on the homology model, and produced a single domain of $LytB_{Bp}$ for experimental epitope mapping, corresponding to the fragment on which EDP and MLCE indications were focusing. During our preliminar investigations, antibodies against the full protein were raised and challenged in OPK and agglutination experiments. The antibodies were able to agglutinate *B. pseudomallei*, but were not able to elicit phagocytosis, placing this candidate as another target of secondary priority.

4.3 RATIONAL DESIGN OF ACTIVE EPITOPES.

The reach of the SV pipeline described in the previous sections is the identification of a restricted number of active epitopes. For active we mean capable of inducing an immunoreaction *in vitro*, in other words limited to controlled conditions in which epitopes, as well as the antibodies raised against those epitopes, are used in conjugated and free forms. When considering the next step in the line, meaning the tests *in vivo*, some further issues need to be addressed. The immunogens (the epitopes) need to be delivered as stable and soluble elements. Free peptides are subject to quick degradation inside an organism, and conjugated peptides are a solution better suited for the laboratory than a vaccine component. At the same time, with the structural design of the epitopes we have a chance to increase the immunogenic potential of the original sequences.

From these considerations we started the process of rational design of BPSL2765 (Pal_{Bp}) epitopes. Our first design is a stabilized version of Pal_{Bp} EPITOPE 3, from now on referred as PAL3. This epitope is composed of 20 amino acids, most of them originally part of an alpha helix in the full antigen. According to Circular Dicroism (CD) measures, the synthesized PAL3 in free form is not capable of maintaining the native conformation, and possesses a β-structure propensity instead (Figure S6), a behavior confirmed by MD simulations (1 microsecond at 300, 320 and 340K in explicit solvent). In order to stabilize the helical conformation of the native peptide, we envisioned the use of an extra-sequence rigid oligopeptidic scaffold as folding nucleator to assist the formation and the stabilization of the alpha helix . We then exploited a side-chain cyclized pentapeptide inserted at the C-terminus of the epitope sequence, that we obtained starting from a linear precursor composed of a Gln-Ala-Glu tripeptide encompassed by ω-azido- and ω-yl-α-amino acid residues. Cyclization occurred via intramolecular copper-catalyzed click reaction[114] to afford the resulting cyclo-pentapeptide (side-chain) bridged through a 1,4-disubstituted 1,2,3-triazolyl moiety. The planar formula of the click reaction can be observed in Figure 10, panel A.

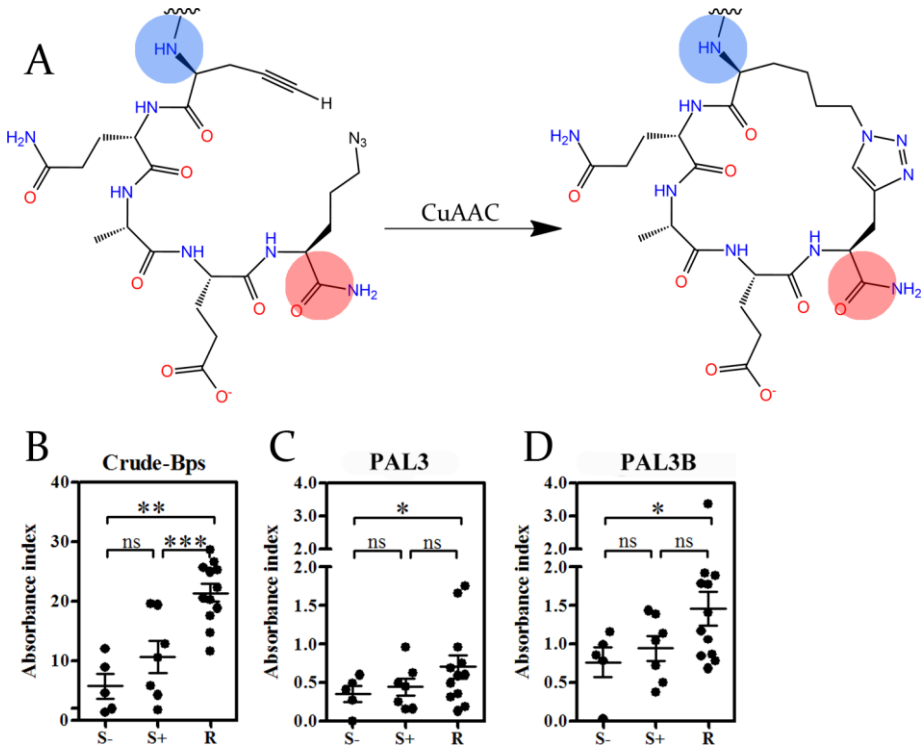


Figure 10**: 2D structure of the helix stabilizer insertion and antibody reactivity against PAL3 and PAL3B in plasma of healthy and recovered melioidosis subjects. A)** Planar configuration of the helix stabilizer. The blue and red circles identify the N-ter and C-ter respectively. **B**) Crude-*B. pseudomallei* (Bps), synthetic peptide PAL3 and PAL3B, were coated onto ELISA plates and probed with diluted plasma samples of healthy controls, seronegative (n=5), seropositive individuals (n = 7) and recovered melioidosis patients (n = 12) and quantified by indirect ELISA.

The addition of this molecule is sufficient to induce the formation of a stable helix, according to CD and simulations (Figure S6). The reactivity of this new epitope (PAL3B) has been assessed by serological tests along with the regular PAL3 as positive control. The results can be seen in Figure 10, panel B, and show an improved absorbance index for PAL3B in comparison to the previous one, especially in recovered melioidosis group. In parallel, the new epitope marks a higher reactivity in the group of seronegative individuals. This may indicate that the insertion is partially immunogenic, or that the helical form is prompting a partially aspecific recognition. Either way, the increased baseline could indicate a self-adjuvating effect, which is not necessarily negative for vaccination purposes. To evaluate the potential efficacy of PAL3B in the elicitation of the

immune system in vivo, we performed experiments of passive immunization in mice. Antibodies against PAL3 and PAL3B were raised in rabbit and used to protect mice from infection. C57BL6 mice were immunized with PBS, a negative control antibody, PAL3 antibodies or PAL3B antibodies intraperitoneal route. 6 hours later, the mice were challenged with Bps576, nasal route. Unfortunately, PAL3 and PAL3B antibodies were unable to elicit any kind of protection (Figure S7), hinting at the fact that an active immunization test may be necessary to display some protective effect, and that heterologous antibodies alone are not sufficient to trigger an appreciable immune reaction in rodents. For this reason, PAL3/PAL3B need to be redesigned into a more stable and active construct, ready to be employed *in vivo*.

One interesting fact about Pal_{Bp} is that PAL3, an antibody-binding site, is predicted to be located just before the location of a T-cell epitope (PALT) on the protein sequence. This proximity can be seen in Figure 11 A, showing the full structure of the protein highlighting PAL3 (teal) and PALT (red). From this observation, we started a process of development of a protein mimic, i.e. a synthetic construct based on a protein fragment, maintaining the original protein fold and presenting the epitopes PAL3 and PALT with maximal solvent exposure.

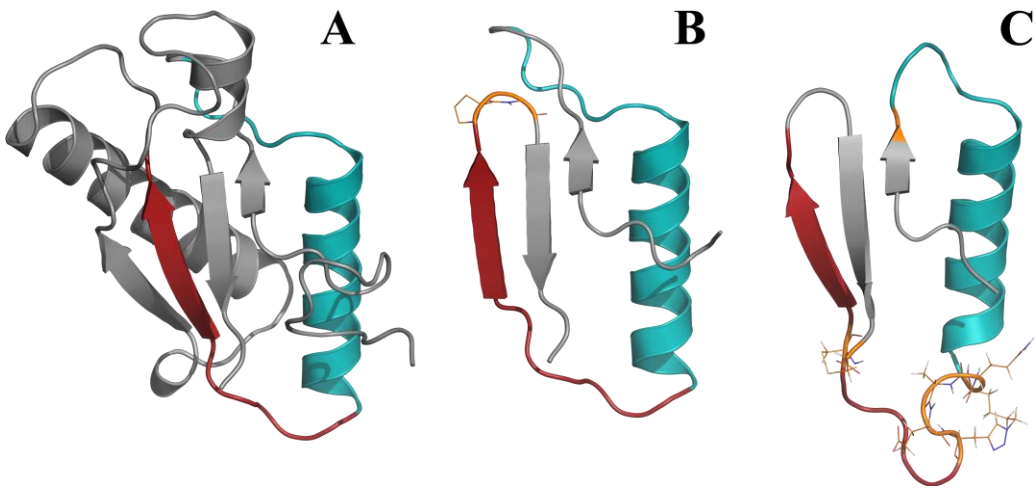


Figure 11**: Structure-based design of Pal_{Bp} into a protein mimic exposing epitopes PAL3 and PALT. A)** Native 3D structure of Pal_{Bp}. The B epitope PAL3 and the T epitope PALT are highlighted in teal and red color, respectively. **B**) First step design: part of the original folding unit comprising the epitopes are excised from the 3D structure and connected by a short turn (DPRO,GLY orange and lines). **C)** Second step design: the mimic is further enhanced with the addition of a helix-inducing insertion between PAL3 and PALT, the reduction of the upper loop and the addition of a disulfide bridge closing the β-hairpin (orange and lines).

All modifications in the design of this mimic and all the testing process have been performed *ab initio* with computational resources for molecular modeling and assessing the protein stability via MD simulations. This project is currently work in progress, so a final fold to be synthesized and employed for preliminary *in vitro* tests is yet to be prepared. Nevertheless, we are drawing close to a final design. The main advances are summarized in panels B and C of Figure 11. In the first step (panel B) we simply operated a virtual excision of the epitopes PAL3 and PALT from the original 3D structure, including part of the surrounding folding unit and joined the gap between the two antiparallel sheets with a bridge composed of a D-Alanine and a Glycine (orange). The sequence of this mimic is reported in Figure S8. Taking this structure as a starting point for simulations, we assessed the ability of an already folded construct to remain stable in solution. In all replicas, part of the helix was lost and the overall stability compromised within the first 50 ns of simulation. By these preliminary indications we identified some crucial spots for subsequent interventions. With the second batch of modifications (Figure 11 C), we reduced the loop located prior to PAL3, we locked the beta sheets by adding a disulfide bridge involving positions 34 and 49, and placed the same helix-inducing insertion we used for epitope PAL3B, positioned between epitopes PAL3 and PALT. The insertion was converted in atomic coordinates and force field parameters, and connected to the mimic in two possible orientations with respect of the original backbone, namely cis and trans (sequence reported in Figure S8). Our latest analysis indicate a substantially improved stability compared to the original fragment depicted in panel B. The insertion is successful in promoting the formation and stabilization of the helix throughout the whole simulation time in both orientations, but introduces an undesired effect. The addition of the circular oligopeptide is forcing the backbone to assume a helical fold before and after its position, meaning that the epitope PALT is subject to an excessive strain resulting in a disrupting effect on the sheets (Figure S9). The solution to this problem might be quite simple. In the next modification iteration, we plan to elongate the loop connecting PAL3 with PALT, allowing the backbone to dissipate the torsion without impairing the following sheets.

Other interventions will include the reduction of the N-ter segment, accompanied by point mutations to induce the formation of a third anti-parallel sheet and attract the nearby helix. At present state, our work of rational design on Pal_{Bp} antigen is in advanced state, and we plan to produce a synthetic construct in the upcoming months, ready to start the testing process.

In conclusion I want to remark that the design of protein mimics is just one of the possible strategies to produce powerful vaccine components. Another approach would be the presentation of multiple active epitopes from different antigens on a scaffolding structure, e.g. silica or gold nanoparticles, allowing the presentation of a considerable quantity and variety of immunogens to the immune system while retaining the simplicity and affordability of peptide synthesis.

4.4 EXPANSION OF THE ORIGINAL S.V. PIPELINE TOWARDS PROTEIN TARGETING AND DIAGNOSTICS

In Section 4.1 of this part, when introducing the results on $OppA_{Bp}$ antigen, I hinted to the fact that epitopes lacking the reactivity necessary to be considered for vaccination purposes could be useful for other applications, specifically for diagnostic purposes. The high mortality of melioidosis is partially to be ascribed to the difficulty of a quick and correct diagnosis of the symptoms. In the cases in which an acute infection is misinterpreted (and consequently mistreated), the chances to survive drop dramatically. In this framework, epitopes showing a differential reactivity profile against human sera collected from healthy and infected individuals may be employed as biomarkers to reveal a disease state. To explore the possibility of expanding the reach of SV, we set out to integrate our epitope prediction methods with high-throughput microarray analysis with the aim to translate molecular understanding of protein-antibody recognition into the design of efficient and selective protein-based analytical and diagnostic tools. We selected two human antigens with known and available tridimensional structure for epitope prediction, namely proteins FABP3 (Fatty acid binding protein 3) and S100B, a calcium-

binding protein involved in the regulation of many cellular pathways, including protein phosphorylation and calcium homeostasis, cell growth and differentiation[115]. Both have been proposed as markers for neurological disorders and damage[116,117]. After simulation and epitope prediction, we raised antibodies against the corresponding peptides and collected the antisera. Then we employed the protein microarray technology to detect the two protein biomarkers by antibody capturing. The overall flowchart of the protocol is summarized in Figure 12. The aim of the test is to verify the ability of the individual epitopes to elicit an immunoreaction with specificity and potential for quantitative detection against the full-length antigen in the sera of patients.

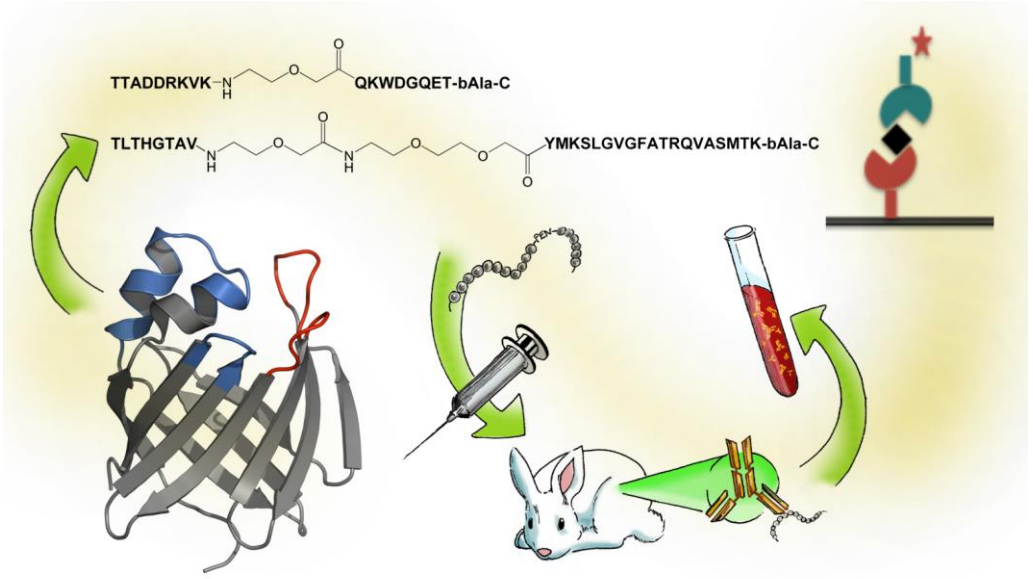


Figure 12**: Graphical representation of the epitope-based protein targeting workflow.** The figure reassumes the key steps of a modified SV pipeline to fit a protein targeting/diagnostic application. Starting from the biomarker's 3D structure (in this case FABP3), the putative antibody binding sites are predicted using the MLCE method (blue and red secondary structure elements). The epitopes are synthesized in form of peptides using PEG moieties to account for the native spatial separation of different elements in discontinuous epitopes. The peptides are inoculated in rabbits for the production of antisera, and the purified antibodies are immobilized on microarray chip surfaces for qualitative and quantitative detection assays.

From the epitope prediction phase we identified two antibody-binding sites on FABP and two on S100. The specificity of interaction was tested with FABP, incubating the full-length protein spotted on the microarray chip with Cy3-labelled antibodies produced against two different mimics of the epitopes.

Nonrelated negative controls include Human Serum Albumin, Ovoalbumin, Ovomucoid, Rabbit, Mouse, and Goat Immunoglobulins. Each of the two anti-FABP antibodies produced a clear fluorescent signal in correspondence of FABP spots (see part II, section 9.2), while control proteins were not recognized, indicating that the predicted epitopes were inducing the production of specific antibodies, capable of recognizing the full-length parent protein. A slight nonspecific interaction is visible on immunoglobulins spots, accounting for antibody aggregation, a well-known drawback in antibody microarrays. In a competition assay the antibodies were challenged with equimolar solutions of the corresponding peptides, producing a strong quenching effect when incubated on the microarray chip. A scrambled version of the peptide did not affect the fluorescence, indicating a direct competition for the recognition site on the antibody.

The epitope mapping on S100, on the other hand, highlighted two epitopes located on opposite sides of the three-dimensional protein structure, separated by 25 Å. This led us to consider the chance of quantitative detection via sandwich immunoassay. In this case, the first set of antibodies (capture antibody) targeted against Epitope 2 were spotted on the microarray chip and incubated with increasing concentration of S100 protein. The second set of antibodies, targeted against Epitope 1 was labelled with Cy3 and employed as detection antibody. The micro-immunoassay generated a dose-dependent fluorescence leading to a limit of detection of 16.06 ng/mL (see part II, section 9.2).

Overall, these results represent a further important validation of the MLCE concept and method that we introduced in the context of epitope prediction. The computational approach could be used to screen libraries of known antigens and efficiently design a diverse, yet focused, collection of epitopes. The combination of this strategy with microarrays opens up the opportunity to generate new targeted and high-throughput diagnostic platforms, and the analysis of the structural and recognition properties of epitopes may further enhance the paradigm allowing for ab-initio design and modification of peptidic antigens to elicit specific responses. Currently, we are applying the same rationale to active epitopes from proteins $OppA_{Bp}$, Pal_{Bp}, $FliC_{Bp}$ and BPSL1050

4.5 EXPANSION OF MLCE TOWARD MHC-II EPITOPES AND DEVELOPMENT OF A WEB TOOL

Another key element of this project is the expansion of the computational methods beyond their original intent, revealing new insights into the physico-chemical properties of epitopes, and providing the Structural Vaccinology approach with additional functionalities. The computational tool we explored and expanded is once more MLCE.
MLCE was designed to select the protein substructures carrying a low energy contribution to the general protein stability, and characterized by a minimal intramolecular energetic couplings to the rest of the protein. These substructures, with their ability to visit multiple conformations, represent putative binding sites for antibodies. The predictive power of MLCE has been validated in the identification and design of immunogenic antibody-binding epitopes of the $OppA_{Bp}$ and Pal_{Bp}[100,106] (see part I, sections 4.1, 4.2). Further applications of this method have been discussed in the design and synthesis of peptidic variants of predicted epitopes, in order to produce antibodies capable of specific interaction with both the peptide and the native protein for diagnostic purposes[118] (part I, section 4.4).
Favorable energetics and conformational accessibility, however, could also underlie the recognition and binding of proteins dedicated to the processing and proteolytic cleavage of antigens, which results in the generation of peptide sequences (T-cell epitopes) for presentation to the major histocompatibility complex (MHC) type II system, the other fundamental component of immune response. A major intrinsic trait of T-cell epitopes is represented by their sequence specificity for different MHC allelic variants, which is *per se* a valid criterion for affinity prediction[119,120]. At the physico-chemical level, however, other intrinsic properties may favor the presentation of a given antigenic peptide. Especially considering MHC-II epitopes, during the antigen processing in the phagolysososme, the selection of the epitopes to be loaded on the MHC-II is dependent upon various criteria, such as the sequence specificity of the proteases (i.e. cathepsins B,D,L,S) and the state of unfolding of the antigen[121,122]. Concurrently, the open groove of the MHC-II may bind to large segments of the antigen during the protein unfolding

guiding the proteolytic cleavage (guided antigen processing[123,124]). At the physico-chemical level, the enthalpic constraints underlying the stability of a protein fold may play a key role in the process, helping define the epitope selection. Strongly coupled residues allegedly oppose strong resistance to the cleavage process, being the principal responsible for the conservation of the fold. Conversely, weakly coupled peptidic segments represent easy targets for protein unfolding and digestion, as well as preferential sites for early MHC-II binding in the guided antigen processing. By this concept the proteolysis of the antigenic protein would not follow a random process, rather a stepwise degradation beginning with the subdivision of the protein in domains, followed by a further fragmentation of the resulting units by the same criteria. In this scenario, the peptides featuring a low energy coupling would be the most obvious antigenic candidates as more exposed to proteolytic cleavage and MHC binding, and thus more represented among the fragments selected for MHC presentation.

Based on the latter considerations, we set out to investigate the possibility of expanding the reach of the MLCE approach to include the prediction of T-cell epitopes. To this end, we selected antigens for which the crystal structures of the full-length proteins and related T-cell epitopes are available. In particular we focused our attention on those T-cell epitopes presented by the MHC type II complex. The data set proposed in this work takes into account 13 antigenic proteins. We gathered 57 non-redundant epitope sequences annotated in the Epitope Database (IEDB)[119], whose 3D structure in complex with the MHC molecule has been experimentally solved (and deposited at the RCSB Protein Data Bank) for 13 of them (Table S2).

The overall prediction workflow following the collected data is shown in Figure 13. The 3D structure of each antigen was relaxed for 5ns by Molecular Dynamics, in NPT conditions and explicit water. The trajectories were clustered to collect the representative frame of each main cluster, in order to get a more relaxed and unconstrained 3D structure for the following analysis. Potential epitopes were then predicted on the first cluster structure using our energy-based approach. In contrast with B epitopes, the T epitopes are oligopeptides that are usually not located on surface protein regions, so MLCE prediction was coupled to a second energy-based computational method, BLOCKS, able to identify

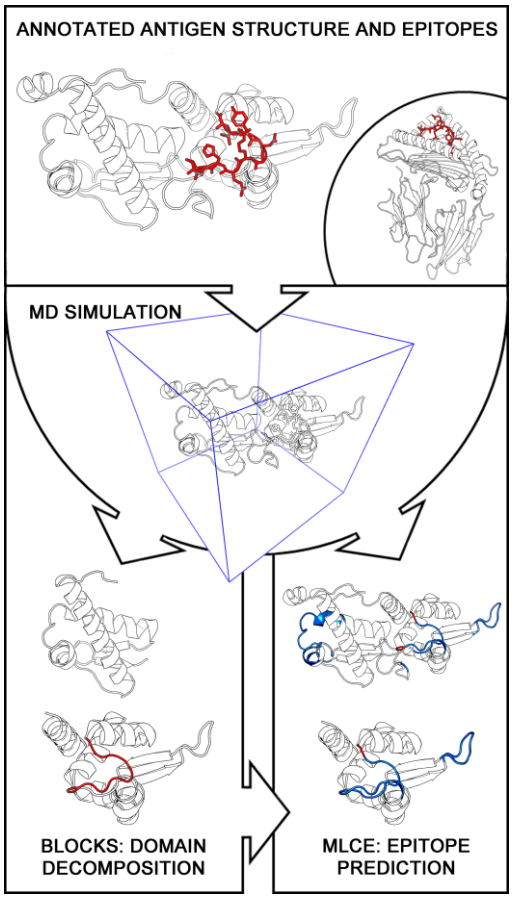


Figure 13: **Schematic representation of the predicton workflow**. We started from the PDB structure (in figure 3BZH) of antigens for which evidence of T epitopes is known (red), and at least one epitope is deposited in crystallographic complex with its MHC molecule (in the upper circle). Then we simulated the full antigen structure and dissected it in domains according to the energy-based domain decomposition. The evaluation of prediction performance was carried out after MLCE prediction over the whole structure as well as the epitope-carrying fragments. Predicted regions for protein 3BZH are shown in blue.

and dissect a protein structure or conformation into individual smaller subdomains, mimicking the initial cleavage of the antigens. This is achieved by grouping different protein fragments into energy clusters that are found to correspond to stable subunits[105]. Epitope prediction is then carried out separately on each resulting individual domain (block): the resulting energetically uncoupled residues may correspond to internal sequence stretches in the initial native, full-length protein structure.

The results of epitope prediction were benchmarked against the available crystal structures of the complexes with MHC-II molecules and with the annotated IEDB epitopes. Statistical analysis of the predictive performance of MLCE was carried out, and evaluated in the presence or absence of the domain decomposition approach.

After building the antigen data set and the T-cell epitope list, we applied the Energy Decomposition method in its original implementation (without BLOCKS decomposition, see also Scarabelli *et al.* 2010[95]) on representative structures of the proteins sampled from short MD simulations. The method returns a binary prediction (a residue belongs either to an epitope or not-an-epitope) in the form of residues gathered in surface patches.

The overall performance of the method is reported in Table 2. A summary of the statistical performance parameters employed is reported in Table S3.

MHC-II EPITOPES

PREDICTION PERFORMANCE ANALYSIS

PDB ID	Cutoff	SENS.	SPEC.	PREC.	ACC.	MCC	p-value
1CB0.A	25%	0.55	0.63	0.06	0.63	0.07	1.42E-059
1D3B.A	15%	0.20	0.94	0.33	0.84	0.17	5.05E-003
1EA3.A	10%	0.14	1.00	1.00	0.36	0.20	4.00E-009
1HA0.A	25%	0.00	0.88	0.00	0.85	-0.06	7.16E-065
1I7Z.A	10%	0.44	0.90	0.16	0.88	0.22	7.32E-021
1OVA.D	25%	0.27	0.59	0.19	0.51	-0.12	3.41E-022
2GIB.B	25%	0.38	1.00	1.00	0.59	0.41	9.80E-008
2JK2.A	15%	0.87	0.86	0.28	0.86	0.44	6.36E-028
2VB1.A	25%	0.34	0.83	0.62	0.60	0.19	1.60E-001
2WA0.A	20%	0.70	0.82	0.16	0.82	0.27	4.27E-031
3BZH.A	25%	0.89	0.74	0.18	0.75	0.33	1.35E-026
3FEY.C	25%	0.50	0.74	0.08	0.73	0.11	1.28E-078
3HLA.A	15%	0.00	0.78	0.00	0.59	-0.25	1.09E-020
AVERAGE	**20%**	**0.41**	**0.82**	**0.31**	**0.69**	**0.15**	

Table 2: **Overall performance analysis of BEPPE in the task of MHC II epitopes prediction**. For every protein (PDB code) is reported the optimal cutoff for prediction and the classification performance parameters sensitivity, specificity, precision and accuracy[97]. The Matthews Correlation Coefficient (MCC)[125] test and statistical significance of the analysis (p-value) are also reported.

The results on the full-length forms of the 13 antigens indicate a statistically significant predictive power of MLCE for T epitopes. However, the use of the single MLCE criterion fails to yield good results in four cases: 1CB0, 1HA0, 1OVA and 3HLA. With the exception of 1CB0, the other proteins present a large multidomain structure, and the prediction may be facilitated by domain decomposition. The concept is not exclusive for large multi-domain proteins. 1EA3 and 2VB1 are antigens presenting at least one epitope sequence which is not solvent exposed, although small in size and composed of a single domain. The combination of domain decomposition and MLCE prediction revealed such epitopes improving the predictor.

The predictions after domain decomposition are depicted in Table 3, showing the results for each protein fragment including mapped epitopes. For such cases we have calculated separate statistics for the two fragments. Incorporating the energy-based decomposition approach to divide the antigenic protein into fragments prior to MLCE results in a general improvement in the performance statistics, but yelding mixed results when predicting over multiple protein fragments. This may be due to the inability to decompose very compact proteins into domains (1OVA), or when producing protein fragments too small (2GIB), or proposing cleavage sites inside annotated epitopes such as in the case of 2VB1 and 3BZH.

AFTER DOMAIN DECOMPOSITION

PDB ID	Cutoff	SENS.	SPEC.	PREC.	ACCU.	MCC.	p-value	DOMAIN	CHANGE
1CB0.A	15%	0.00	0.83	0.00	0.76	-0.13	4.05E-15		
1D3B.A	15%	0.22	0.84	0.29	0.70	0.07	1.61E-02		
1EA3.A	25%	0.25	1.00	1.00	0.33	0.18	5.67E-12	1	
1EA3.A	25%	0.43	0.84	0.80	0.59	0.28	6.61E-03	2	
1HA0.A	15%	0.54	0.78	0.13	0.76	0.18	4.41E-32		
1I7Z.A	10%	0.44	0.94	0.40	0.89	0.36	5.17E-05		
1OVA.D	Failed decomposition								
2GIB.B	10%	0.30	1.00	1.00	0.66	0.43	1.73E-02	1	
2GIB.B	Epitope cleavage							2	
2JK2.A	25%	0.60	0.64	0.18	0.63	0.16	5.29E-19		
2VB1.A	25%	0.00	0.56	0.00	0.36	-0.47	4.13E-07	1	
2VB1.A	Epitope cleavage							2	
2WA0.A	15%	0.90	0.85	0.41	0.86	0.55	1.01E-09		
3BZH.A	Epitope cleavage								
3FEY.C	10%	1.00	0.93	0.53	0.94	0.70	1.50E-08	1	
3FEY.C	10%	0.00	0.88	0.00	0.84	-0.08	5.27E-23	2	
3HLA.A	25%	0.34	0.63	0.34	0.53	-0.03	8.19E-05		

Table 3: **Performance analysis of MLCE on individual protein fragments.** The table includes all the previous statistical parameters, accompanied by a DOMAIN column, indicating the proteins featuring two fragments containing epitopes. The last column on the right indicate whether the predictive performance is enhanced (green), degraded (red) or unchanged (blu) compared to the full protein prediction. The performance calculation was not possible (or not significant) in proteins 1OVA, 2GIB-2, 2VB1-2 and 3BZH.

With all these limitations and caveats in mind, we can consider the domain decomposition method for its ability to improve epitope prediction in all cases in which splitting a protein in subparts is a minor concern. These include moderately large proteins, or proteins composed of multiple domains. Other examples include known protein fragments and domains of interest, for which BLOCKS could suggest the ideal cleavage sites for MLCE predictions.

Overall, the results showed a significantly accurate T-cell epitope identification, with greater performance when epitope prediction is carried out after structural decomposition via the BLOCKS approach, suggesting the energetic uncoupling may be also a key property of MHC epitopes, confined to specific subdomains of the initial protein. From a performance standpoint, it must be noticed that MLCE is significantly more capable in the prediction of B-cell epitopes, confirming that in the context of MHC epitope selection other key factors contribute to the final presentation.

MLCE is a method specifically designed to highlight energetic motifs from a protein 3D structure, revealing in many cases the areas containing potential epitopes. The method can be successfully employed in a dual purpose (B/T epitopes), giving consistent indication of immunogenic sequence location. Building on this consideration, since a modern vaccine should stimulate a protective response originated from both the B-cell and T-cell mediated recognition using a minimum set of immunogenic components and adjuvants, we consider the predictive pipeline of MLCE a flexible tool ready to be employed successfully for vaccine development, diagnostics design and protein targeting purposes.

In conclusion, in this study we have expanded the reach of MLCE to become a more general B and T-cell epitope predictor. These mechanisms constitute two major branches of the adaptive immune response, and are both fundamental in the eradication of the infection and the acquisition of a long-term, active immunological memory.

Based on these new findings, we adopted the new revisions to the MLCE algorithm to set up a free, web-accessible version of the BEPPE server suitable for predictions of B and T cell epitopes, as described in part II, section 10.1. The new server accepts single structures

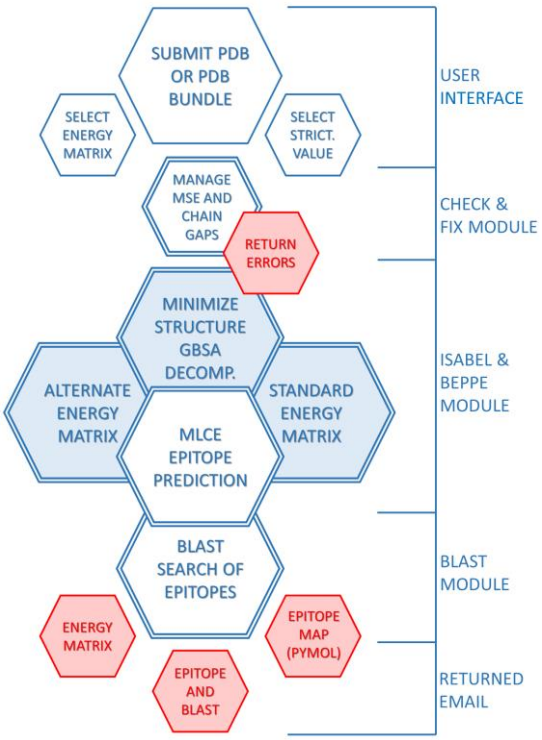


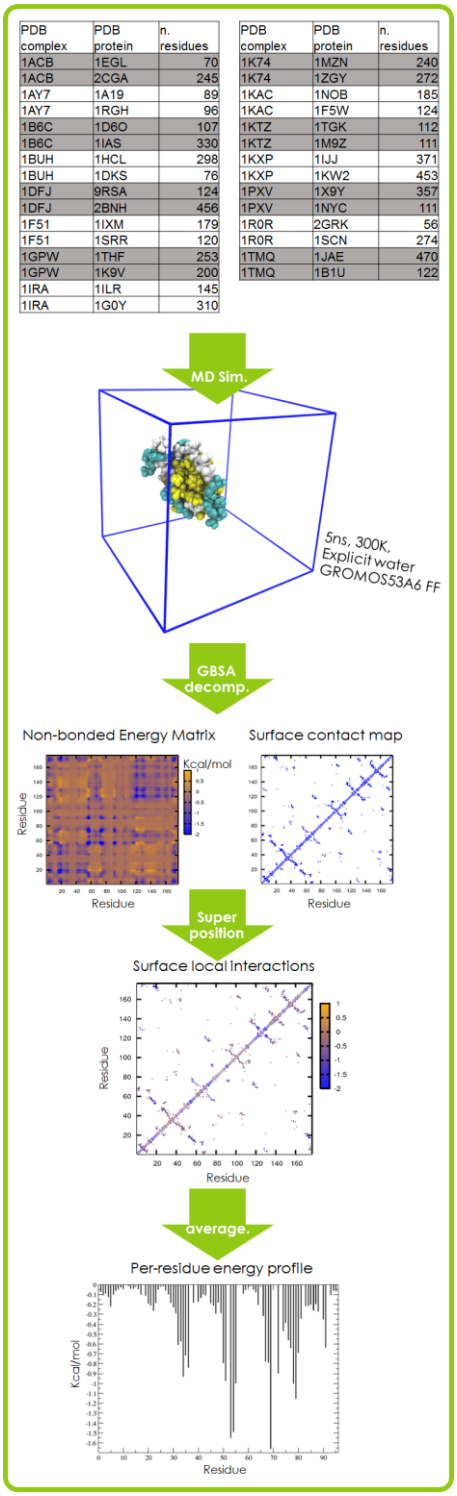
Figure 14**: Layout of the web tool BEPPE**: each individual PDB file is passed by the User Interface to the prediction pipeline (double line exagons). The main modules processing the structure, existent even in stand-alone format, consist of a preliminary input check script, the package ISABEL for the energy decomposition and analysis, the program BEPPE for the prediction of epitopes, and the BLAST script. All results (and eventually all errors) are gathered in a unique text file accompanied by supplementary data, and sent to the user via email (red, shaded background, single line exagons).

of proteins or collections of proteins in PDB format, and returns a list of predicted epitopes in form of patches, along with their BLAST search against the human proteome, and a pictorial representation of the energy matrices produced after the energy decomposition. In this version, the user can access two advanced options to select the energy matrix to be calculated (one is to be preferred in case of multi-domain or multi-chain proteins), and the strictness of prediction, that is the percentage of low-energy contributions to be used for prediction. A schematic representation of the server automation is shown in Figure 14.

4.6 THE ROLE OF SURFACE ENERGETICS IN THE FORMATION OF PROTEIN-PROTEIN COMPLEXES.

The last part of this project, as stated in section 3, is not directly related to the development of a vaccine. Rather it aims at garnering new insights into the biophysical aspects at the basis of the energetic analyses used for epitope prediction. The use of energy-based analysis tools to predict the location of binding epitopes implies the presence of a property defining binding in the general energetics of the protein. Once the property is verified, it can be described with a mathematical formulation. Such binding property could also be predicted with appropriate algorithmic automation. Thus, in the case of MLCE, the observation of a discernible pattern (low coupling) inside the non-bonded intramolecular energy and solvation terms of antibody binding sites, led ultimately to the epitope predictor BEPPE.

We then set out to investigate protein energetics, computed and approximated with molecular mechanics techniques, in search of other possible properties. The fact that the antibody-binding sites are very peculiar interface of protein-protein recognition, led us to formulate the hypothesis that the energetic data could reveal a more general property linked to protein-protein interactions (PPIs).
To investigate these aspects in depth, we gathered a dataset of 30 proteins, subdivided as 15 binary complexes undergoing transient interactions. The dataset was selected as an heterogeneous sample of Benchmark 3.0[126], originally compiled to evaluate the performance of protein docking algorithms, and is composed of proteins for which high resolution structural data is present, completed with detailed annotation of their binding interfaces. The methodological outline of this study, starting from the atomic coordinates of the proteins and getting to a per-residue representation of the energy of the system, is shown in Figure 15. In this work, we collected the structural information of all proteins, both in individual form and in dimeric complex (45 PDB codes), and run simulations for each one (5ns all-atom MD, at 300K and explicit solvent using GROMACS 4.5[127] and GROMOS53A6 force field[86]. The energetic determinants were computed via MM-GBSA



calculation approach (Molecular Mechanics Generalized Born Surface Area) treated in implicit solvent as included in the AMBER 11 software package[128]. The total interaction energy of the system is decomposed into intramolecular and solvation terms on a residue pairwise basis. Energy decomposition is used in addition to reconstruct a simplified energy matrix. The procedure is similar to the methods employed by MLCE/BEPPE to generate the energy matrix, with a difference: the matrix is not reconstructed based on the principal eigenvector of the full covariance matrix, instead it is generated by the combination of a set of principal eigenvectors, selected according to their level of representation of the energetic covariance[105]. The resulting energy matrix (Figure 15) is filtered by a neighbouring list, maintaining solely the energy pairs of surface residues, excluding all distal contributions (the maximum distance allowed is 7 Å).

Figure 15: **Schematic representation of the workflow followed for the energetic description analysis.** From top to bottom, the 3D structure of 15 complexes and their constituting 30 proteins (in table) have been simulated for 5ns for conformational relaxation (MD box). The nonbonded intramolecular and solvation energy terms of each protein have been calculated and retrieved via GBSA calculation, and expressed in form of a pairwise energy matrix via energy decomposition. The energy matrix, filtered by a similar, pairwise neighbouring list of surface residues, is collapsed to a monodimensional profile, averaging all energy contributions for each amino acid. From this energetic profile, the weakest elements (energy close to 0) are selected to be mapped on the protein surface.

The final matrix is simplified as an energetic profile, averaging the energetic interactions of each residue to a unique value (Figure 15).

All preliminary investigations in search of a direct correlation between the intensity of this energy and the presence of a binding interface were unable to highlight any property with significant ability to predict the interaction interface. Instead, all tests highlighted interesting motifs of different energy coupling intensity characterizing protein surfaces.

According to these observations, we started approaching the investigation with a broader perspective. The vast majority of computational methods devoted to the analysis and the prediction of PPIs focus on the identification of the binding interface. The reason for this is the necessity to rely on a direct, non-ambiguous source of information that can be annotated and retrieved for practical applications. It is what we call the Instruction Manual Strategy: to know exactly where two proteins assemble is the most efficient way of reconstructing the complex (Figure 16). This strategy offers a common and intuitive perspective for studying protein-protein interactions, but it is not the only one. Another way to reconstruct the complex or to study the mechanisms of interaction may be represented by what we call the Watercolor Strategy. The concept may be better introduced using a familiar analogy: one can imagine somebody (e.g. a kid) leaving a mark of watercolor on an assembly of toy bricks.

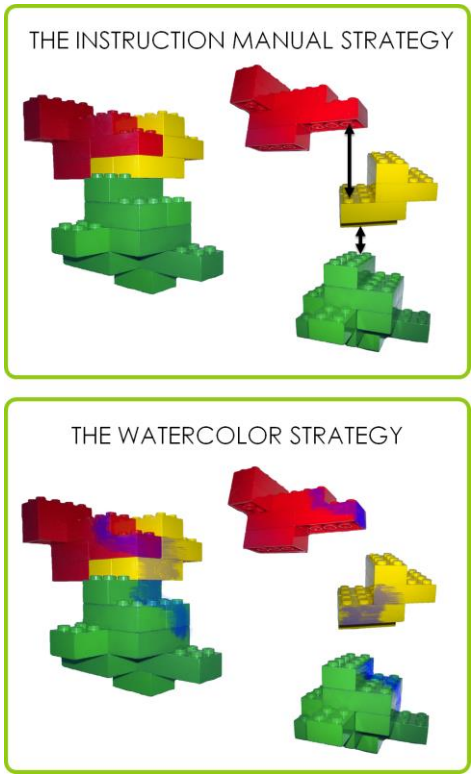


Figure 16: **Conceptual representation of our theoretical strategy, expressed by toy bricks analogy.** On top, The traditional approach to protein-protein interaction pursues the location of the binding sites and use them as guidelines for assembly. In our strategy, a mark on the complex persistent on each individual part would be able to reveal additional information, such as the binding orientation

If the mark is persistent, it will remain visible on its constituent parts, even after disassembling. Based on this mark, one may be able to reconstruct the original assembly following the continuity and the orientation of the color, even in absence of instructions (Figure 16). A similar perspective may be applied to the PPI problem: according to our hypothesis, protein adaptation to their binding partners may have left such a trace in the energy of stabilization during the course of evolution, pushing the constituent parts toward a sort of energetic complementarity across the binding interface. If such motifs are persistent, so that they can be isolated in the energetic footprint of the complexes and the isolated constituent proteins in monomeric form, they would represent a property of protein-protein interactions. A formal description of such property could be employed in the development of a dedicated analysis tool, with potential to reveal the location of a binding site or, most likely, the binding orientation. More generally this would contribute to the investigation and clarification of another physico-chemical aspect underpinning protein-protein interactions.

From the energetic analysis of the isolated form of each protein in our dataset, we observed that surface regions characterized by weak energetic uncoupling define specific motifs. Examples of these areas are shown in Figure 17, upper panel. Interestingly, when calculated and mapped on their respective crystallographic complexes, the motifs are conserved and tend to identify a unique and continuous motif connected across the two proteins in 14 out of 15 protein complexes (Table S4), hinting to the possibility to reveal an energetic complementarity in the complex state (Figure 17 middle panel). It is worth noting the fact that protein 1G0Y undergoes a radical conformational change from the closed conformation of the isolated state and the open conformation of the heterodimeric form, in complex with protein 1ILR. Despite the difference, the motifs identified in the closed conformation produce a sharp continuous mark across the two proteins when mapped on the complex structure. Attempts to predict the binding interface on the isolated form of this protein using known algorithm would clearly fail exactly due to the conformational change. Proteins 1SRR and 1IXM, constituent parts of the complex 1F51, represent another interesting case since the two proteins assemble in a tetrameric biological form. The continuous motif actually emerges in the tetramer (Figure 17).

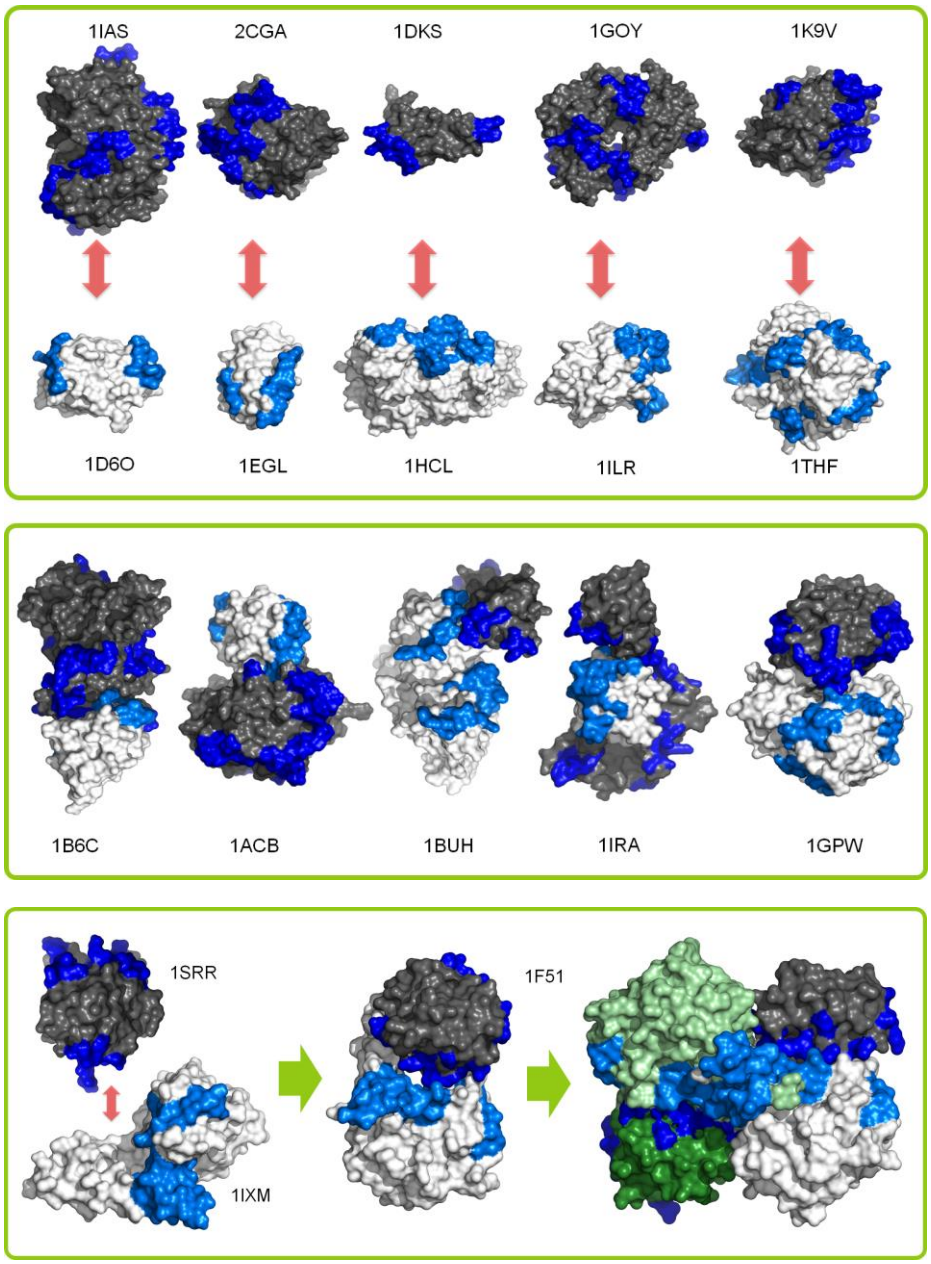


Figure 17: **Motifs of uncoupled energy mapped on the 3D structure of isolate protein and complexes**. On top, The analysis of the local energetic coordination, performed on the monomeric variants of the protein dataset shows that surface regions characterized by weak energetic uncoupling outline specific motifs (in blue). In the middle panel, the same motifs are mapped on their respective complexes, outlining unique and continuous *blue stripes*. In the bottom panel, the motifs calculated on monomeric proteins 1SRR and 1IXM outline a *blue stripe* in both dimeric and tetrameric assemblies. (pale green and green proteins correspond to further 1IXM and 1SRR proteins, respectively)

According to the peculiar representation and colors chosen to display the results of analysis during this study, these motifs have been given, for simplicity, the name of *blue stripes*.

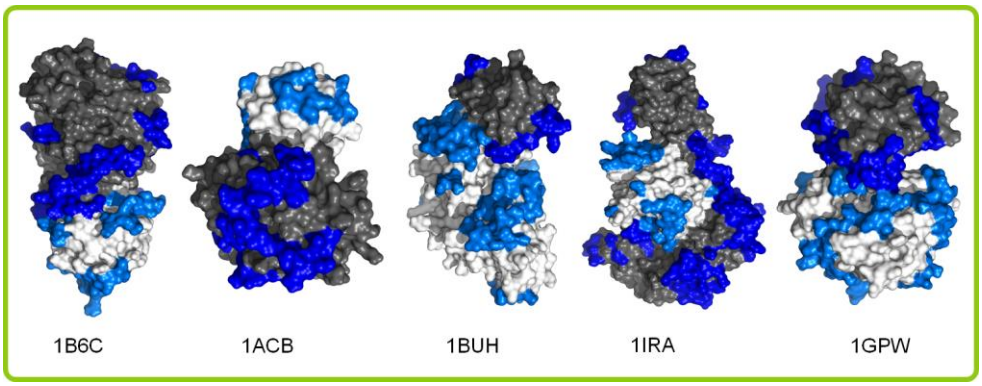


Figure 18: **Motifs of uncoupled energy from the energetic analysis of entire complexes**. This panel depicts an example of the same analysis of energetic uncoupling calculated on the full complex of proteins shown in Figure 17. The comparison between monomers and dimers returned similar results in 1B6C, 1BUH and 1GPW, only affected by minor differences. 1ACB features a very similar blue stripe, but spatially translated upon binding (the structure has been rotated 90 degrees compared to the previous image). 1IRA is one of the few examples of poorly persistent blue stripe, although a motif continuity can be still appreciated. The differences may be partly due to the vast conformational rearrangement of protein 1GOY between free and bound form.

According to our hypothesis, the *stripes* identified on the monomers would be still visible when calculated on the full complex structures, as they are constituents of a continuous mark. The energetic analysis run over all the complexes partially confirmed this expectation. A continuous motif of weak coupling energy is visible for 12 out of 15 complexes (Table S4), and in the majority of cases (9 out of 15) the *stripes* are conserved with respect to the previous results. Of the remaining 6 complexes, 3 still present a similar (if not equivalent) motif, but showing a spatial translation on the protein surface (1ACB, Figure 18). The major difference producing this effect (and producing the majority of other, more subtle changes in other complexes) is due to areas characterized by very low uncoupling energy that upon binding acquire stability, and new, previously "neutral" surface patches that are affected by destabilization, thus being featured in the *blue stripes*. A quantitative indication of identity between results of monomer analysis vs complex analysis is shown in Table S4.

The percentage of identity is very high in the majority of cases, indicating that the analysis is mainly oriented towards the same surface areas, although with differences.
The presence of differences between *stripes* calculated from the isolated protein and the ones generated from the complexes are compatible with the common idea of proteins as dynamic systems, and differences within the energy of stabilization between different levels of protein organizations were not entirely unexpected. The striking detail, in our opinion, is the fact that similar *stripes* can be identified over monomers and complexes, so that the analysis of interacting proteins in isolated form is able to approximate a property of the final complex.

The presence of a discernible pattern inside the local energetic stability of monomeric proteins capable of outlining binding motifs on the complex, suggests a form of energetic complementarity between the constituent parts. This complementarity is visible at the chemico-physical level of residue interaction and is not dependent on the chemical composition of the amino acids involved (Figure S10). This property can be exploited to gain information at the protein-protein interaction level, like the presence of a binding site or the binding orientation. This led to the development of an automated analysis tool, named BLUEPRINT (*Beppe*-Like Uncoupling Energy for PRotein INTeraction). At the current development state, BLUEPRINT is able to process autonomously the energetic data from GBSA calculations of an entire dataset of protein monomers and/or complexes. The analysis produces data on the location and signal intensity of the *blue stripes* in textual and pictorial format, mapping the results on the structure of the complex when available.

This method may be employed in a range of applications, from the ab-initio characterization of PPIs, to protein-protein docking algorithms. One further area of interest would be the study of a possible correlation between evolution and energetic signature, in order to better understand the mechanisms producing this complementarity, and put it in relation with protein function.

5 - CONCLUSIONS AND FUTURE PROSPECTS

We live in a society exquisitely dependent on science and technology, in which hardly anyone knows anything about science and technology

Carl Sagan

Structural Vaccinology is a developing field that aims to devise viable strategies for vaccine design, based on the identification of immunogenic determinants (epitopes), through computational and 3D structure analyses. In the context of this thesis project, I have worked within an international consortium following a Reverse and Structural Vaccinology initiative devoted to the identification of active antigens and epitopes from the pathogen *Burkholderia pseudomallei*, etiological agent of the life-threatening disease melioidosis.

The computational methods I developed and used within this framework, mostly involving molecular modeling and simulation as well as epitope prediction based on the MLCE approach, were successful, in combination with other *in silico* and *in vitro* mapping systems, in the identification of antibody-binding sites on numerous antigenic X-ray and NMR structures.

The integrated pipeline that we first described with $OppA_{Bp}$ antigen was able to deliver a handful of immunogenic peptides identified from different high-priority antigens. These peptides were recognized by antibodies present in the plasma of melioidosis-infected patients, and were able to elicit agglutination and neutrophils opsonization when tested in vitro. Interestingly, the reactivity of the single epitopes, as well as their ability to stimulate the immune response in vitro were slightly different from the results we registered with their corresponding full-length antigens. Although in many cases this led to the identification of less-active epitopes, in other cases such as epitope COMP3 from $OppA_{Bp}$ and Pal_{Bp} PAL3, they were able to deliver peptides with full potential to discriminate between healthy and infected patients. In one case also, Epitope PAL3 from Pal_{Bp}, the single epitope showed enhanced activity with respect of the native antigen. These peptides are ideal targets for the rational engineering of stable and potent vaccine components. Preliminary results on a conformationally constrained version of PAL3 are showing increased immunoreactivity, strongly encouraging future efforts in structure-based design of the identified epitopes.

The results shown here demonstrate the feasibility of a structure-based and epitope-based vaccine initiative, and expand the original concept toward other fields of application. Our method has recently proven that it is possible to use synthetic peptides to raise antibodies that retain full specificity for the epitope region and fully recognize the full-length protein, opening perspectives for the development of antibody-based screening diagnostic platforms.

From the practical point of view, designed biomolecules can have importance in the development of diverse applications, ranging from analytics and diagnostics, to drug-discovery and biotechnology. Hence, by increasing our understanding of the molecular-level origins of protein interactions, we will be able to rationally engineer novel molecules (peptides, peptidomimetics, and de novo designed proteins) with characteristics suitable for a particular application.

With this objective in mind, during this project I have applied novel strategies to extend the possibilities of MLCE toward the prediction of MHC-II binding epitopes, and worked at the code level to implement a full prediction software based on MLCE, BEPPE, on a web-based freely accessible platform. This new version of BEPPE should increment the

versatility of the predictive instrument while considerably reducing the time necessary for analysis, producing a free, convenient, effective, and user friendly tool for the GtA project as well as for other initiatives, whether they are interested in SV or structure-based protein targeting.

The investigation of protein structures and interactions, on the other hand, has other essential implications. From the fundamental point of view, the development of rational approaches to predict and design sequences with specific properties can help in understanding the physical basis of molecular recognition as well as furthering our understanding of the relationships between protein sequence, structure, dynamics, and function. In this framework, I have investigated the energetic footprint of 15 interacting proteins, both in monomeric and dimeric form, in search of a specific property of interaction. This lead to the observation of clear motifs in the local energetic coordination of the monomers, showing that surface regions characterized by weak energetic uncoupling outline specific motifs on the protein isolates. When mapped on their respective crystallographic complexes, the motifs tend to identify a unique and continuous feature, a "*blue stripe*", connected at the interface border. An effect that was also visible across those proteins of the dataset forming trimeric or tetrameric assemblies. Analyzing the energy of the entire complexes, we obtained similar results, suggesting that these marks are persistent. The presence of this pattern in the local energetic coordination of isolate proteins suggests a form of energetic coordination retaining information on the bound forms, like the presence of a binding site or the binding orientation. This approach has been developed into an independent analysis tool, named BLUEPRINT. Future applications of BLUEPRINT to real case scenarios may either validate or reject this model. In case it proves reliable, it may be employed in a range of fundamental and practical applications, from the ab-initio characterization of PPIs, to protein-protein docking algorithms, to the evolutionary and functional investigation of interactions.

6 - REFERENCES

1. Whitmore A, Krishnaswami CS (1912) An account of the discovery of a hitherto undescribed infective disease occuring among the population of Rangoon. *Indian Med Gazette* 47: 262–267.
2. Stanton AT, Fletcher W. Melioidosis, a new disease of the tropics. *Trans Fourth Congr Far East Assoc Trop Med* 1921(2):196-8.
3. Limmathurotsakul D, Wongratanacheewin S, *et al*. (2010) Increasing incidence of human melioidosis in Northeast Thailand. *Am J Trop Med Hyg* 82:1113–1117.
4. Currie BJ, Fisher DA, Howard DM, Burrow JN, Selvanayagam S, *et al*. (2000) The epidemiology of melioidosis in Australia and Papua New Guinea. *Acta Trop* 74:121–127.
5. Currie BJ, Dance DA, Cheng AC (2008) The global distribution of *Burkholderia pseudomallei* and melioidosis: an update. *Trans R Soc Trop Med Hyg* 102(1): 1–4.
6. White NJ. (2003) Melioidosis. *Lancet* 361:1715-22.
7. Cheng AC, Currie BJ. (2005) Melioidosis: Epidemiology, Pathophysiology, and Management. *Clin Microbiol Rev* 18:383-416.
8. Peacock SJ. (2006). Melioidosis. *Curr. Opin. Infect. Dis.* 19(5): 421-428.
9. Wiersinga WJ, van der Poll T, *et al*. (2006). Melioidosis: insights into the pathogenicity of *Burkholderia pseudomallei*. *Nat. Rev. Microbiol*. 4(4): 272-282.
10. Wuthiekanun V and Peacock SJ (2006). Management of melioidosis. *Expert. Rev. Anti. Infect. Ther*. 4:445-455.
11. Atkins T, Prior RG, *et al*. (2002). A mutant of *Burkholderia pseudomallei*, auxotrophic in the branched-chain amino acid biosynthetic pathway, is attenuated and protective in a murine model of melioidosis. *Infect. Immun*. 70:5290-5294.
12. Haque A, Easton A, *et al*. (2006). Role of T cells in innate and adaptive immunity against murine *Burkholderia pseudomallei* infection. *Journal of Infectious Diseases* 193(3): 370-379.
13. Cuccui J, Easton A, *et al*. (2007). Development of signature tagged mutagenesis in *Burkholderia pseudomallei* to identify mutants important in survival, attenuation and pathogenesis. *Infect. Immun.* 75: 1186-1195.
14. Breitbach K, Kohler J, *et al*. (2008). Induction of protective immunity against *Burkholderia pseudomallei* using attenuated mutants with defects in the intracellular life cycle. *Trans R Soc Trop Med Hyg* 102 Suppl 1(1): S89-94.
15. Nelson M, Prior JL, *et al*. (2004). Evaluation of lipopolysaccharide and capsular polysaccharide as subunit vaccines against experimental melioidosis. *J. Med. Microbiol.* 53(12): 1177-1182.
16. Pollard AJ, Perrett KP, *et al*. (2009). Maintaining protection against invasive bacteria with protein-polysaccharide conjugate vaccines. *Nat. Rev. Immunol.* 9(3): 213-20.

17. Jones SM, Ellis JF, *et al*. (2002). Passive protection against *Burkholderia pseudomallei* infection in mice by monoclonal antibodies against capsular polysaccharide, lipopolysaccharide or proteins. *J. Med. Microbiol*. 51: 1055-1062.
18. Healey GD, Elvin SJ, *et al*. (2005). Humoral and cell-mediated adaptive immune responses are required for protection against *Burkholderia pseudomallei* challenge and bacterial clearance post-infection. *Infect. Immun.* 73(9): 5945-5951.
19. Haque A, Chu K, *et al*. (2006). A live experimental vaccine against *Burkholderia pseudomallei* elicits CD4(+) T cell-mediated immunity, priming T cells specific for 2 type III secretion system proteins. *Journal of Infectious Diseases* 194(9): 1241-1248.
20. Sarkar-Tyson M, Smither SJ, *et al*. (2009). Protective efficacy of heat-inactivated *B. thailandensis*, *B.mallei* or *B. pseudomallei* against experimental melioidosis and glanders. *Vaccine* 27(33): 4447-51.
21. Brett PJ and Woods DE. (1996). Structural and immunological characterization of *Burkholderia pseudomallei* O-polysaccharide-flagellin protein conjugates. *Infect. Immun*. 64(7):2824-2828.
22. Harland DN, Chu K, *et al*. (2007). Identification of a LolC homologue in *Burkholderia pseudomallei*, a novel protective antigen for melioidosis. *Infect. Immun*. 75(8):4173-80.
23. Hara Y, Mohamed R, *et al*. (2009). Immunogenic *Burkholderia pseudomallei* outer membrane proteins as potential candidate vaccine targets. *PLoS One* 4(8): e6496.
24. Fenner F, Henderson DA, Arita L, Jezek Z, Ladnyi ID. (1988) Smallpox and its eradication. *Geneva: World Health Organization*.
25. Jenner E. (1801) The Origin of the Vaccines Inoculation. London: Shury.
26. Jenner E. (1798) Inquiry into the Causes and Effects of the *Variolae* Vaccine. London: Sampson Low, 45.
27. Pasteur L. (1880) De l'attenuation du virus du Choléra des poules. *CR Acad. Sci.* 91:673–680.
28. Rappuoli R. (2004) From Pasteur to genomics: progress and challenges in infectious diseases. *Nat. Med*. 10:1177–1185.
29. Levine MM, Lagos R, Esparza J. (2009) Vaccines and Vaccination in Historical Perspective. *New Generation Vaccines. Fourth Edition*. New York: Informa Healthcare USA, Inc.:pp. 1–11.
30. Offit PA. (2007) Vaccinated: One Man's Quest to Defeat the World's Deadliest Diseases. New York: HarperCollins.
31. Glenny AT, Hopkins BE. (1923) Diphtheria toxoid as an immunizing agent. *Br J Exp Pathol*. 4:283–288.
32. Ramon G. (1924) Sur la toxine et sur l'anatoxine diphtériques. Pouvoir floculant et propriétés immunisantes. *Ann Inst Pasteur* 38:1–106.
33. Buynak EB, Roehm RR, *et al*. (1976) Vaccine against human hepatitis B. *J. Amer. Med. Assoc*. 235(26):2832-4.
34. Clare S, Dougan G. (2004). Live recombinant bacterial vaccines in Novel Vaccination Strategies, ed Kaufmann S. H. E., editor. (Weinheim: Wiley-VCH;), 319–341.
35. Ada G. (2001) Vaccines and *vaccination. N. Engl. J. Med*. 345(14):1042-53.
36. Fleischmann RD, Adams MD, *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 269(5223):496-512.
37. Rappuoli R. (2000) Reverse vaccinology. *Curr. Opin. Microbiol*. 3(5):445-50.

38. World Health Organization. Meningococcal meningitis fact sheet. Available at: http://www.who.int/mediacentre/factsheets/fs141/en/
39. Sette A, Rappuoli R. (2010) Reverse Vaccinology: Developing Vaccines in the Era of Genomics. *Immunity* 33(4):530-41.
40. Tettelin H, Saunders NJ, *et al*. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*. 287(5459):1809-15.
41. Pizza M, Scarlato V, *et al*. (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*. 287(5459):1816-20.
42. Giuliani MM, Adu-Bobie J, *et al*. (2006) A universal vaccine for serogroup B meningococcus. *Proc. Natl. Acad. Sci. USA*. 103(29):10834-9.
43. Maione D, Margarit I, *et al*. (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science*. 309(5731):148-50.
44. Rodriguez-Ortega MJ, Norais N, Bensi G, *et al*. (2006) Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. *Nat. Biotechnol.* 24: 191–197.
45. Thorpe C, Edwards L, Snelgrove R, *et al*. (2007) Discovery of a vaccine antigen that protects mice from *Chlamydia pneumoniae* infection. *Vaccine*. 25(12):2252-60.
46. Vidor E. (2010) Evaluation of the persistence of vaccine-induced protection with human vaccines. *J Comp Pathol*. 142(1):S96–101.
47. Felgner PL, Kayala MA, *et al*. (2009). A *Burkholderia pseudomallei* protein microarray reveals serodiagnostic and cross-reactive antigens. *Proc Natl Acad Sci USA* 106(32): 13499-504.
48. Holden MT, Titball RW, Peacock SJ, *et al*. (2004). Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc. Natl. Acad. Sci. USA*. 101:14240–14245.
49. Nandi T, Ong C, Singh AP, *et al*. (2010). A genomic survey of positive selection in *Burkholderia pseudomallei* provides insights into the evolution of accidental virulence. *PLoS Pathog*. 6(4):e1000845.
50. Sarkar-Tyson M, Thwaite JE, Harding SV, *et al*. (2007) Polysaccharides and virulence of *Burkholderia pseudomallei*. *J. Med. Microbiol*. 56:1005–1010.
51. Shalom G, Shaw JG, Thomas MS. (2007) In vivo expression technology identifies a type VI secretion system locus in *Burkholderia pseudomallei* that is induced upon invasion of macrophages. *Microbiology* 153:2689–2699.
52. Kim HS, Schell MA, Yu Y, Ulrich RL, Sarria SH, *et al*. (2005) Bacterial genome adaptation to niches: divergence of the potential virulence genes in three *Burkholderia* species of different survival strategies. *BMC Genomics* 6: 174.
53. Galyov EE, Brett PJ, DeShazer D (2010) Molecular insights into *Burkholderia pseudomallei* and *Burkholderia mallei* pathogenesis. *Annu Rev Microbiol* 64: 495–51.
54. Peano C, Chiaramonte F, Motta S. *et al* (2014) Gene and Protein Expression in Response to Different Growth Temperatures and Oxygen Availability in *Burkholderia thailandensis*. *PLoS One*. 2014; 9(3):e93009.
55. Glass MB, Gee JE, Steigerwalt AG, Cavuoti D, Barton T, *et al*. (2006) Pneumonia and septicemia caused by *Burkholderia thailandensis* in the United States. *J. Clin. Microbiol.* 44: 4601–460.

56. Ooi WF, Ong C, Nandi T, *et al*. (2013) The condition-dependent transcriptional landscape of *Burkholderia pseudomallei*. *PLoS Genet*. 9(9):e1003795.
57. Moule MG, Hemsley CM, Seet Q, *et al*. (2014) Genome-wide saturation mutagenesis of *Burkholderia pseudomallei* K96243 predicts essential genes and novel targets for antimicrobial development. *MBio*.5(1):e00926-13.
58. Conaty S, Watson L, Dinnes J and Waugh N. (2004) The effectiveness of pneumococcal polysaccharide vaccines in adults: a systematic review of observational studies and comparison with results from randomised controlled trials. *Vaccine* 22(23-24): 3214-24.
59. Janeway C. (2001) Immunobiology. (5th ed.). Garland Publishing.
60. Pier GB, Lyczak JB, Wetzler LM. (2004) *Immunology, Infection, and Immunity*. ASM Press.
61. Rus H, Cudrici C, Niculescu F. (2005) The role of the complement system in innate immunity. *Immunol Res* 33 (2): 103–112.
62. Dormitzer PR, Ulmer JB, Rappuoli R. (2008)**.** Structure-based antigen design: a strategy for next generation vaccines. *Trends in Biotechnology* 26: 659-6667.
63. Azoitei ML, Correia BE, Ban Y-EA *et al*. (2011) Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science* 334:373–376.
64. Correia BE, Ban Y-EA, Holmes Ma *et al*. (2010) Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. *Structure* 18:1116–1126.
65. Ofek G, Guenaga FJ, Schief WR *et al*. (2010) Elicitation of structure-specific antibodies by epitope scaffolds. *Proc Natl Acad Sci USA* 107:17880–17887.
66. Burton DR. (2010) Scaffolding to build a rational vaccine design strategy. *Proc Natl Acad Sci USA* 107:17859–17860.
67. Wu X, Yang Z-Y, Li Y, *et al*. (2010) Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* 329:856–861.
68. Zhou T, Georgiev I, Wu X, *et al*. (2010) Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science*. 329:811–817.
69. Scheid JF, Mouquet H, Ueberheide B, *et al*. (2011) Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science*. 333:1633–1637.
70. West AP, Diskin R, Nussenzweig MC, Bjorkman PJ. (2012) Structural basis for germ-line gene usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120. *PNAS*. 109:E2083-E2090.
71. McGuire AT, Hoot S, *et al*. (2013) Engineering HIV envelope protein to activate germline B cell receptors of broadly neutralizing anti-CD4 binding site antibodies. *J Exp Med*. 210(4):655-663.
72. Jardine J, Julien J-P, Menis S, *et al*. (2013) Rational HIV immunogen design to target specific germline B cell receptors. *Science* 340(6):711-716.
73. Scarselli M, Arico B, Brunelli B *et al.* (2011) Rational design of a meningococcal antigen inducing broad protective immunity. *Sci Transl Med* 2011; 3: 91ra62.
74. Nuccitelli A, Cozzi R, Gourlay LJ *et al.* (2011). Structure-based approach to rationally design a chimeric protein for an effective vaccine against Group B Streptococcus infections. *Proc Natl Acad Sci USA* 108: 10278–10283.

75. Heitler W. and London F. (1927) Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik. *Zeitschrift für Physik*, 44, 455—472.
76. Pauling L, Wilson EB. (1935) Introduction to Quantum Mechanics: With Applications to Chemistry. McGraw-Hill Book Company, inc. New York and London.
77. Heitler W. (1945) Elementary Wave Mechanics – With Applications to Quantum Chemistry. Oxford / Clarendon Press, Oxford.
78. Coulson CA. (1952) Valence. Oxford University Press, Oxford.
79. Metropolis N, Ulam S. (1949). The Monte Carlo Method. *J. Am. Stat. Assoc*. 44 (247): 335–341.
80. Alder BJ, Wainwright TE. (1959). Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys*. 31 (2): 459.
81. Zhao G, Perilla JR, Yufenyuy EL, Meng X, *et al*. (2013) Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 497: 643-646.
82. Shaw DE, Dror RO, Salmon JK, *et al*. (2009). Millisecond-Scale Molecular Dynamics Simulations on Anton (Portland, Oregon). *Proceedings of the ACM/IEEE Conference on Supercomputing* (SC09) (New York, NY, USA: ACM): 1–11.
83. http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html
84. van Gunsteren WF, Billeter SR, Eising AA, et al. (1996) Biomolecular simulation: The GROMOS96 manual and user guide. Zürich, Switzerland: Vdf Hochschulverlag.
85. van Gunsteren WF, Daura X, Mark AE. (1998) GROMOS force field. *Encyclopaedia of Comput Chem* 2: 1211–1216.
86. Oostenbrink C, Villa A, Mark AE, van Gunsteren W. (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem*. 25: 1656–1676.
87. Van Regenmortel MHV. (2009) What is a B-cell epitope? *Meth. Mol. Biol.* 524: 3–20.
88. Greenbaum JA, Andersen PH, Blythe M, *et al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit. JMR* 20: 75–82.
89. Novotny´ J, Handschumacher M, Haber E, *et al.* (1986) Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl. Acad. Sci. USA* 83: 226–230.
90. Westhof E, Altschuh D, Moras D, *et al*. (1984) Correlation between segmental mobility and the location of antigenic determinants in proteins. *Nature* 311: 123–126.
91. Fiorucci S, Zacharias M. (2010) Prediction of protein–protein interaction sites using covering algorithms. *Biophys. J.* 98: 1921–1930.
92. Ponomarenko J, Bui H–H, Li W, *et al*. (2008) ElliPro: a new structurebased tool for the prediction of antibody epitopes. *BMC Bioinform* 9: 514.
93. Thornton JM, Edwards MS, Taylor WR, Barlow DJ (1986) Location of "continuous" antigenic determinants in the protruding regions of proteins. *Eur. Mol. Biol. Organiz. J.* 5: 409–413.
94. Moreau V, Granier C, Villard S, *et al*. (2006) Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics* 22: 1088–1095.

95. Scarabelli G, Morra G, Colombo G. (2010) Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. *Biophys. J.* 98:1966–1975.
96. Ma B, Wolfson H, Nussinov R. (2001) Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Curr. Opin. Struct. Biol.* 11:364-369.
97. Ponomarenko JV, Bourne PE. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct. Biol.* 7:64.
98. Tiana G, Simona F, De Mori GM, Broglia RA, Colombo G. (2004) Understanding the determinants of stability and folding of small globular proteins from their energetics. *Protein Sci.* 13(1): 113-24.
99. Morra G, Colombo G. (2008) Relationship between energy distribution and fold stability: insights from molecular dynamics simulations of native and mutant proteins. Proteins 72(2): 660-72
100. Lassaux P, Peri C, Ferrer-Navarro M, *et al.* (2013) A Structure-Based Strategy for Epitope Discovery in *Burkholderia pseudomallei* OppA Antigen. Structure 21(1): 167-175.
101. Garmory HS, Titball RW. (2004) ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies. *Infect. Immun.* 72: 6757–6763.
102. Tanabe M, Atkins HS, Harland DN, Elvin SJ, et al. (2006) The ABC transporter protein OppA provides protection against experimental *Yersinia pestis* infection. *Infect. Immun.* 74: 3687–3691
103. Monnet V. Bacterial oligopeptide-binding proteins. (2003) *Cell. Mol. Life Sci.* 60: 2100–2114..
104. Borezee E, Pellegrini E, Berche P. (2000) OppA of *Listeria monocytogenes*, an oligopeptide-binding protein required for bacterial growth at low temperature and involved in intracellular survival. *Infect. Immun.* 68: 7069–7077.
105. Genoni A, Morra G, Colombo G. (2012) Identification of domains in protein structures from the analysis of intramolecular interactions. *J. Phys. Chem. B* 116: 3331–3343.
106. Gourlay LJ, Peri C, Ferrer-Navarro M, *et al.* (2013) Exploiting the *Burkholderia pseudomallei* Acute Phase Antigen BPSL2765 for Structure-Based Epitope Discovery/Design in Structural Vaccinology. *Chem. Biol.* 20(9): 1147-1156
107. Amadei A, *et al.* (1993) Essential dynamics of proteins. *Proteins* 17(4): 412-425.
108. Morra G, Verkhivker G, Colombo G. (2009) Modeling signal propagation mechanisms and ligand-based conformational dynamics of the Hsp90 molecular chaperone full length dimer. *PLOS Comp. Biol.* 2009, 5:e1000323.
109. Suwannasaen D, Mahawantung J, Chaowagul W, *et al.* (2011) Human immune responses to *Burkholderia pseudomallei* characterized by protein microarray analysis. *J. Infect. Dis.* 203:1002-1011.
110. Chen YS, Hsiao YS, Lin HH, *et al.* (2006) CpG-modified plasmid DNA encoding flagellin improves immunogenicity and provides protection against *Burkholderia pseudomallei* infection in BALB/c mice. *Infect. Immun.* 74:1699-1705.
111. Hayashi F, Smith KD, Ozinsky A, *et al.* (2001) The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* 410:1099-1103.
112. Larsen JE, Lund O, Nielsen M. (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res* 2:2.

113. Saha S, Raghava GP. (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 65: 40-48.

114. Scrima M, Le Chevalier-Isaad A, Rovero P, *et al*. (2010) CuI-Catalyzed Azide–Alkyne Intramolecular *i*-to-(*i*+4) Side-Chain-to-Side-Chain Cyclization Promotes the Formation of Helix-Like Secondary Structures. *Eur. J. Org. Chem*. v2010(3): 446-457.

115. Marenholz I, Heizmann CW, Fritz G (2004). S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochem. Biophys. Res. Commun.* 322(4): 1111–22.

116. Chiasserini, D., Parnetti, L., Andreasson, *et al*. (2010) CSF levels of heart fatty acid binding protein are altered during early phases of Alzheimer's disease. *J. Alzheimer's Dis*. 22: 1281–1288.

117. Steiner J, Bogerts B, Schroeter ML, Bernstein HG. (2011) S100B protein in neurodegenerative disorders. *Clin. Chem. Lab. Med*. 49:409–424.

118. Peri C, Gagni P, Combi F, *et al*. (2013) Rational Epitope Design for Protein Targeting. *ACS Chem. Biol.* 8(2): 397-404.

119. Vita R, Zarebski L, Greenbaum JA, *et al*. (2010) The immune epitope database 2.0. Nucleic Acids Res. 2010 Jan;38(Database issue):D854-62. doi: 10.1093/nar/gkp1004.

120. Singh H, Raghava GP. (2001) ProPred: Prediction of HLA-DR binding sites. *Bioinformatics* 17 (12): 1236-7.

121. Blum JS, Wearsch PA, Cresswell P. (2013) Pathways of antigenic processing. *Annu. Rev. Immunol*. 31: 443-73.

122. Mimura Y, Mimura-Kimura Y, Dorees K, *et al*. (2007) Folding of an MHC class II-restricted tumor antigen controls its antigenicity via MHC-guided processing. *Proc. Natl. Acad. Sci. USA* 104(14): 5983-8.

123. Sercarz EE, Maverakis E. (2003) Mhc-guided processing: binding of large antigen fragments. *Nat. Rev. Immunol*. 3(8):621-9.

124. Moss CX, Tree TI, Watts C. (2007) Reconstruction of a pathway of antigen processing and class II MHC peptide capture. *EMBO J*. 26(8): 2137-47.

125. Matthews BW. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA-protein struct. M* 405(2): 442–451.

126. Hwang H, *et al*. (2008) Protein-Protein Docking Benchmark Version 3.0. *Proteins* 73(3): 705-709.

127. Hess B, Kutzner C, van der Spoel D, Lindahl E. (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput*. 4(3): 435-447.

128. http://ambermd.org

7 – ACKNOLEDGEMENTS

The projects described in this thesis were supported by Fondazione CARIPLO (Progetto Vaccini, contract number 2009-3577) and Regione Lombardia - Fondazione CARIPLO, (Progetto PROVA: "discovery/development of diagnostic PRObes and VAccine candidates targeting *Burkholderia* infections").

I want to thank and acknowledge all the people that worked with me on these projects:

CNR-ICRM, Milano

Alessandro Gori, Dario Corrada, Giulia Morra, Renato Longhi Giorgio Colombo.

UNIMI, dept. Biosciences, Milano

Louise Gourlay, Patricia Lassaux, Lucia Perletti, Martino Bolognesi.

UEX-CLES, Exeter

Rachael J. Thomas, Olivia L. Champion, Stephen L. Michell, Richard W. Titball.

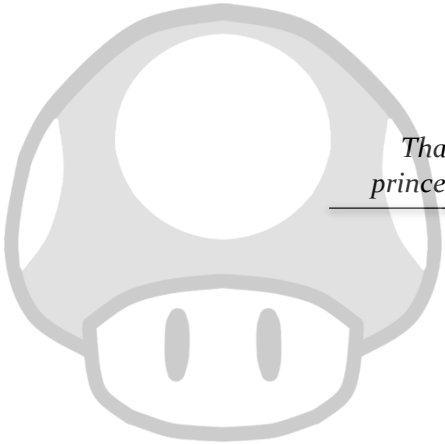
UAB-IBB, Barcelona

Mario Ferrer-Navarro, Oscar Conchillo-Solé, Xavier Daura.

CMDL, Khon Kaen

DarawanRinchai, Chidchamai Kewcharoenwong, Arnone Nithichanon, Ganjana Lertmemongkolchai.

8 – CREDITS



Thank you Mario, but our princess is in another castle!

Toad

C'è stato un istante, durante la stesura della tesi, in cui avuto una strana sensazione di disagio. In quel momento stavo sostituendo il campo in copertina "Name Initials, Last name" con il mio nome. Ho sentito come se non fosse del tutto vero, come se quel nome non bastasse. Certamente, la tesi di dottorato appartiene al dottorando. Lui svolge il lavoro, redige il testo, accompagna un progetto dalla carta da blocco alla carta stampata mettendoci il suo impegno, un po' imparando quello che non sa e un po' facendo a pezzi quello che crede di sapere...

Eppure penso che una tesi non appartenga solo al dottorando. Molto è dovuto ai suoi colleghi, che hanno creato un luogo in cui è piacevole sia lavorare che perdere tempo (perché le idee migliori vengono in pausa caffè, è risaputo). Hanno riportato sulla terra le ansie e le aspettative, e gli hanno fatto capire presto che anche le persone esperte chiedono quando non si sanno rispondere da sole. Di sicuro appartiene al suo supervisore, che prima gli ha spiegato le regole e poi gli ha mostrato come si gioca da professionisti. Gli ha dato entusiasmo, fiducia e carta bianca, permettendogli di tracciare i contorni dei propri limiti in cerca di una via attorno. Ne reclamano la proprietà tutte le persone che hanno contribuito allo stesso progetto con le loro differenti competenze, concertando gli sforzi e proiettando i risultati verso un traguardo più ampio. Un pezzo appartiene infine alla donna che in questi anni di dottorato gli ha custodito il cuore, i pensieri, i sogni.

I miei ringraziamenti sono per voi.

9 – PUBLISHED MANUSCRIPTS

In this section, I am listing the URLs and details of the scientific publications produced within this thesis project. For licensing reasons, the actual journal pages are embedded in the text only in the printed version of this document.
All articles are publicly accessible at the following addresses:

9.1 http://www.sciencedirect.com/science/article/pii/S0969212612003796

Lassaux P, Peri C, Ferrer-Navarro M, *et al.* (2013)

A Structure-Based Strategy for Epitope Discovery in *Burkholderia pseudomallei* OppA Antigen.

Structure 21(1): 167-175. DOI: 10.1016/j.str.2012.10.005

9.2 http://pubs.acs.org/doi/abs/10.1021/cb300487u

Peri C, Gagni P, Combi F, *et al.* (2013)

Rational Epitope Design for Protein Targeting.

ACS Chem. Biol. 8(2): 397-404. DOI: 10.1021/cb300487u

9.3 http://www.sciencedirect.com/science/article/pii/S1074552113002779

Gourlay LJ, Peri C, Ferrer-Navarro M, *et al.* (2013)

Exploiting the *Burkholderia pseudomallei* Acute Phase Antigen BPSL2765 for Structure-Based Epitope Discovery/Design in Structural Vaccinology.

Chem. Biol. 20(9): 1147-1156. DOI: 10.1016/j.chembiol.2013.07.010

9.4 http://link.springer.com/article/10.1007/s00726-013-1526-9

Gori A, Longhi R, Peri C, Colombo G. (2013)

Peptides for immunological purposes: design, strategies and applications.

Amino Acids 45(2): 257-268. DOI: 10.1007/s00726-013-1526-9

10 – ACCEPTED AND SUBMITTED MANUSCRIPTS

Submitted to:

Methods in Molecular Biology, Accepted Oct. 2014

Prediction of antigenic B and T cell epitopes via Energy Decomposition analysis. Description of the web-based prediction tool BEPPE.

Claudio Peri[1], Oscar C. Solé[1], Dario Corrada[1], Alessandro Gori, Xavier Daura and Giorgio Colombo*

[1]All authors contributed equally to this new version of the web tool BEPPE.

*corresponding author

Summary:

Unraveling the molecular basis of immune recognition still represents a challenging task for current biological sciences, both in terms of theoretical knowledge and practical implications. Here, we describe the physical-chemistry methods and computational protocols for the prediction of antibody-binding epitopes and MHC-II loaded epitopes, starting from the atomic coordinates of antigenic proteins (PDB file). These concepts are the base of the web tool BEPPE (Binding Epitope Prediction from Protein Energetics), a free service that returns a list of putative epitope sequences and related blast searches against the Uniprot human complete proteome. BEPPE can be employed for the study of the biophysical processes at the basis of the immune recognition, as well as for immunological purposes such as the rational design of biomarkers and targets for diagnostics, therapeutics and vaccine discovery.

1.Introduction

The increasing availability of experimentally solved protein structures and the advances in theoretical knowledge and computational resources permitted to obtain a deeper understanding of proteins and their physico-chemical properties. This in turn paved the

way for the development of dedicated computational tools for the analysis and the functional prediction of these properties. In this context, BEPPE (Binding Epitope Prediction from Protein Energetics) is a web-based tool for the prediction of antibody binding sites and MHC-II loaded epitopes, based on the analysis of protein internal energetics. BEPPE is rooted in a computational biophysics approach, defined MLCE, for the prediction of antigenic substructures in isolated proteins *(1)*.

The method was initially applied to the prediction of B-cell epitopes, i.e. epitopes that can be recognized and bound by antibodies. *(2, 3, 4)*. The predictive performance *(5)* of the method, originally benchmarked on 19 antibody-antigen interactions, yielded 0.46 percentage in sensitivity and 0.84 specificity, with a PPV (positive predicted value, also referred as precision) of 0.32 and AUC values (Area Under the Curve) *(6)* of 0.71, at the variation of the prediction cutoff. Now we have expanded the possibilities of BEPPE to include the prediction of MHC-II epitopes. Our latest benchmark composed of 13 antigens and 57 non-redundant MHC-II epitopes, indicate a prediction performance scoring sensitivity, specificity and PPV of 0.41, 0.83, and 0.31 respectively, which is very close to the predictivity observed for antibody-binding sites [**Table 1**]. One obvious use of the results is the design of synthetic epitope sequences. Moreover, our methods can be exploited for the design of optimized antigens. This can be achieved, e.g., through the identification of antigen conformations that optimally present immunogenic regions to the recognition by the immune system molecules, and the identification of mutations that can block the protein in such conformation. BEPPE can be accessed as a free web service at the following URL: http://bioinf.uab.es/BEPPE

2. Materials

BEPPE is largely composed of code written in-house, yet relies on some public free and commercial packages to perform the Energy Decomposition Analysis (as described by Tiana *et al. (7)*) and BLAST search *(8)*. Here we describe in detail the conditions and parameters integrated in the automated analysis procedure.

2.1. MM-GBSA calculation parameters

The input structure is treated in implicit solvent by means of the MM-GBSA calculation approach (*Molecular Mechanics Generalized Born Surface Area*) included in the AMBER (Assisted Model Building with Energy Refinements) software package *(9)*.

Molecular Mechanics (MM) parameters for interactions are described by the forcefield *ff03 (10, 11)*. The MM calculations are carried out with the *sander* module.

The polar solvation term is approximated with the Generalized Born (GB) model *(12, 13)* and OBC re-scaling *(14)* for solvation energy. The dielectric constant is set to 80 for the bulk (water as solvent) and 1 for the protein and internal cavities. The GB approach used by AMBER takes into account the shielding effect; we have adopted a physiological salt concentration (0.1 M).

The non-polar solvation term is calculated through the evaluation of the solvent accessible surface area (SA) using the icosahedra approximation. The non polar contribution to the free energy of solvation is calculated as Enp = surften*SA (no offset correction, SURFOFF = 0) and the surface tension (surften) used is 0.0072 kcal/mol/A^2.

2.2. Energy minimization and residue pairwise decomposition

Input structures undergo a preliminary energy minimization procedure before any subsequent analysis, consisting of 200 minimization steps with the steepest descent optimization algorithm. Short range and long-range non-bonded interactions are considered within a 12 Å cutoff.

The interaction terms are used to build a matrix, describing the interaction energy between every residue pair in the sequence. This matrix is simplified through Principal Component Analysis decomposition. The resulting main eigenvectors and eigenvalues are used to generate a simplified matrix recapitulating the most relevant stabilizing interactions within the protein structure.

2.3 Matrix of Local Coupling Energy (MLCE) and epitope prediction

The simplified energy matrix reconstructed after PCA is used as the input for epitope prediction. BEPPE intersects the energy matrix with information on the protein topology, expressed as a contact matrix (6 Å cutoff from beta carbons): the result is the Matrix of

Local Coupling Energies (MLCE) *(1)*. From this, the algorithm selects the list of contacting amino acids with a minimal coupling to the rest of the protein residues. This list forms the final epitope prediction, expressed in the form of patches. The selection of the residues carrying the weakest coupling energy depends upon a cutoff, defined as the percentage of most-uncoupled pairs over all possible ones in the structure. The cutoff is set by default to 15% of all contributions.

2.4 BLAST search of epitopes

As reported in the bibliography, immunoreactive bacterial epitopes should not match sequences present in the human proteome *(15)*. To check for this, predicted epitope sequences are blasted against the human proteome, downloaded from Uniprot ftp *(16)* (this file is updated every four weeks). NCBI blast+ version 2.2.23 is executed with parameters adjusted for short input sequences using a strategy file downloaded from a 16-residue query NCBI blast search result. This file contains parameters optimized for short sequences as described in the blast help *(17)*. Blast results are provided with the prediction output file.

3. Methods

This section explains in details the procedures and calculation steps to run a prediction job. In **Fig. 1** we illustrate the individual modules of the server and the job pipeline.

1. Access the web page for BEPPE at URL http://bioinf.uab.es/BEPPE
2. Upload your antigen pdb file or pdb file bundle in zip format. The program is compatible with single and multiple chain files (*see* **Note 1**). Upon submitting the job, each structure is passed to the core module for analysis and prediction (*see* **Note 2**).
3. The "*check and fix*" module is composed of a Python script, analyzing the PDB for common format anomalies, such as missing residues, presence of selenomethionines (MSE), duplicate residues (alternative sidechain position and mutations), and alternative atom positions. The script will automatically convert any MSE to MET, and remove any atom/residue duplicate, keeping only the first one indicated in the PDB file. In case of chain holes the calculation of the pairwise

energetic contributions may be dramatically affected. In this case the program will be terminated and an error returned via email (*see* **Notes 3 and 4**).

To avoid any unexpected inconvenience, we strongly encourage the user to manually check the coordinates file before submitting.

4. Once the PDB file is accepted, it is passed to a comprehensive automated pipeline, converting the PDB file format to AMBER standard, launching the energy minimization and the energy decomposition, as described in the Materials section. This independent code layer is named "*ISABEL"* (ISabel is Another BEppe Layer). After decomposition, the interaction energy matrix is diagonalized into its eigenvalues and related eigenvectors. By default only the first eigenvector will be taken into account *(7)*. Alternatively, it is possible to select a subset of the most representative eigenvectors according to the method proposed by Genoni *et al. (18)*. On the basis of the eigenvector(s) selected, a simplified interaction energy matrix will be generated (**Fig. 2**); such a procedure is intended to emphasize the more relevant (from an energetic point of view) non-bonded interactions from the initial "raw" interaction energy matrix, as described in details in *(1, 19, 20)*. The user can select the alternative energy matrix by checking the box "*alternative energy matrix*" before submission. The use of the alternative matrix is limited to large, multidomain proteins, or in presence of multi-chain complexes. The default energy matrix should be preferable in the vast majority of cases (*see* **Note 5**)
5. The newly built energy matrix is passed to the module BEPPE, which builds the MLCE matrix as the intersection of energy and contact matrices. The energetic cutoff for the selection of the epitope patches is set by default to 15% of all contributions, which generally scored best for antibody-antigen interfaces *(1)*, but other cutoffs may be chosen depending on the user requirements. A stricter cutoff will limit the results to the most uncoupled residues, usually located in very small patches. Conversely, loosening the cutoff will produce larger patches. The optimal range for prediction goes from 5% to 25% of all low-energy contacts (*see* **Note 6**). The cutoff can be manually adjusted from the drop-down menu "*Strictness of prediction*". Note that the actual percentages have been substituted in the web page with a much simpler notation, being 1 for a strict prediction (5%), and 5 for

a soft one (25%). Clearly, in this simple values notation, the standard cutoff is set to 3.

6. Finally, the patches are blasted against the human proteome for similar motifs. Given the fact that most of the applications of BEPPE, including the discovery of biomarkers for diagnostics and targets for vaccines and therapeutics, would require immunogenic epitopes with no resemblance of pre-existent, endogenous proteins, we provide the end-user with a quick outlook on the sequence alignment of the results.
7. The output provided includes the prediction at the residue level in text format subdivided in patches, along with the blast search results for each one. In attachments, the system provides a pictorial representation of the energy matrix as regenerated after decomposition and diagonalization (such as shown in **Fig. 2**), and a Pymol script *(21)* for an automatic, easy visualization of the results mapped on the protein structure.

4. Notes

1. The program is compatible with PDB files including multiple chains. If the user needs to run a prediction on individual subunits of a protein complex, or different conformations of the same protein, it is required that each structure is submitted as an individual PDB file. In a multi-chain file, the total energy of the system will be estimated, guiding the predictor towards the most uncoupled patches of the whole assembly. Such an analysis may be indicated in the case of complexes undergoing obligate interactions.
2. BEPPE is a web tool returning one prediction for each input PDB file. One can use the reference structure obtained by diffraction techniques or NMR solution, but we would like to stress the possibility of using multiple conformations of the same protein to draw a consensus prediction out of different snapshots (e.g. 3 files), to help reduce the noise and focus on the significant patches. These conformations may be different fits of an NMR bundle, or representative frames from a Molecular Dynamics simulation.

3. The script ignores any missing residue as commonly annotated in the PDB remarks, since occasionally chain gaps are present within the protein structure even if not indicated. For similar reasons, it neither fully relies on the annotated numbering of the amino acids. The program progressively evaluates the distance of alpha carbons of consecutive residues. If large distances are found (>2.8 Å) the chain gap is evident, even if not indicated, and the job for that file is aborted. If smaller distances are found, the annotated chain gap may be just nominal and the chain integrity is likely to be maintained. In this case, the PDB file will be renumbered, since an uneven numbering may produce artifacts. The renumbered PDB file is accepted and a warning is returned (take into account that any prediction will follow the numbering of this new file, not the original one!).
4. The program automatically rejects any input structure containing chain gaps, since the extent of the energy perturbations and their effect on the predictivity have never been addressed in our benchmarks. Nevertheless, if you wish to run a prediction on a protein structure containing chain holes, try submitting it as a multiple chain file.
5. Usually, the first eigenvector is sufficient to reconstruct the energetic signature of a protein *(7)*. However, when attempting to reconstruct the energy matrix of proteins composed of independent domains (or subunits, considering complexes), the first eigenvector may be insufficient for a complete protein coverage. The user can be aware of this issue by looking at the graphic representation of the energy matrix (provided as an email attachment. An example is shown in **Fig. 2**). When the energetic information (the colored clusters) is not spread along the whole energy matrix (i.e. in presence of large, white bands in correspondence of one domain), this may be due to the inability of the first eigenvector to represent the full matrix. The alternative method reconstructs the matrix with the specific task of maximizing coverage while considering only the most cohesive and representative energy modules from the top-ranked principal eigenvectors. This method is to be intended as a recovery system for those areas that would be left undescribed, since its efficacy has never been validated for the purpose of epitope prediction.

6. BEPPE is particularly sensitive to extremely low-coupling energies. Meaning that unstructured regions extended towards the solvent, thus bearing minimal contacts with the rest of the protein (such as long, flexible loops), will inevitably mark the local energy matrix (MLCE) with a strong uncoupled signal. Albeit these regions may actually include an epitope, the presence of such strong signals may mask the presence of other regions worth of attention. Such an event can be easily acknowledged by looking at your reference structural data, and comparing it with the energy matrix provided as an email attachment. If the energy matrix shows an overwhelming signal in correspondence of such motifs, we strongly recommend to use a very low cutoff (values 1-2 from the drop-down menu) for your primary prediction (returning the core of the uncoupled motif/region), and run a subsequent prediction with a loose cutoff (value 5 from the drop-down menu). The second run can extend the prediction to include other, distal significant patches. In this case, one can ignore the "bleaching" noise effect produced at high cutoffs around the primary patch (due to the extremely strong signal) and focus on the core patch identified with the first prediction and the "secondary" patches highlighted at higher cutoff.

Acknowledgments

This project was supported by EU's FP6 ("BacAbs", ref. LSHB-CT-2006-037325) and the Cariplo Foundation ("GtA", ref. 2009-3577).

References

1. Scarabelli, G.; Morra, G.; Colombo, G. (2010) Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. *Biophys J* 98, 1966-1975.
2. Lassaux, P.; Peri, C.; *et al*. (2013) A Structure-Based Strategy for Epitope Discovery in *Burkholderia pseudomallei* OppA Antigen. *Structure* 21, 167-175.
3. Peri, C.; Gagni, P.; *et al*. (2013) Rational Epitope Design for Protein Targeting. *ACS Chem Biol* 8, 397-404.
4. Gourlay, L.J.; Peri, C.; *et al*. (2013) Exploiting the *Burkholderia pseudomallei* Acute Phase Antigen BPSL2765 for Structure-Based Epitope Discovery/Design in Structural Vaccinology. *Chem Biol* 20, 1147-1156.
5. Ponomarenko, J.V.; Bourne P.E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct Biol* 7, 64.
6. Fawcett, T. (2006) *An introduction to ROC analysis. Pattern Recognitt Lett* 27, 861-974.
7. Tiana, G.; Simona, F.; De Mori G.M.; *et al*. (2004) Understanding the determinants of stability and folding of small globular proteins from their energetics. *Protein Sci* 13,113-24.

8. Altschul, S.F.; Gish, W.; Miller, W.; *et al.* (1990) Basic local alignment search tool. *J Mol Biol* 5,403-10.
9. http://ambermd.org/
10. Duan, Y.; Wu, C.; Chowdhury, S.; *et al.* (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24, 1999–2012.
11. Lee, M.C.; Duan, Y. (2004) Distinguish protein decoys by using a scoring function based on a new Amber force field, short molecular dynamics simulations, and the generalized Born solvent model. *Proteins* 55, 620–634.
12. Hawkins, G.D.; Cramer, C.J.; Truhlar, D.G. (1995) Pairwise solute descreening of solute charges from a dielectric medium. *Chem Phys Lett* 246, 122-129.
13. Hawkins, G.D.; Cramer, C.J.; Truhlar, D.G. (1996) Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J Phys Chem* 100,19824-39.
14. Onufriev, A.; Bashford, D.; Case, D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins* 55, 383-394.
15. Amela, I.; Cedano, J.; Querol, E. (2007) Pathogen proteins eliciting antibodies do not share epitopes with host proteins: a bioinformatics approach. *PLoS One*;2(6):e512.
16. The UniProt Consortium. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. 42, 191-198.
17. http://www.ncbi.nlm.nih.gov/BLAST/Why.shtml
18. Genoni, A.; Morra, G.; Colombo, G. (2012) Identification of Domains in Protein Structures from the Analysis of Intramolecular Interactions. *J Phys Chem B* 116, 3331-43
19. Corrada, D.; Morra, G.; Colombo, G. (2013) Investigating allostery in molecular recognition: insights from a computational study of multiple antibody-antigen complexes. *J Phys Chem B* 117(2), 535-52.
20. Corrada, D.; Colombo, G. (2013) Energetic and dynamic aspects of the affinity maturation process: characterizing improved variants from the bevacizumab antibody with molecular simulations. *J Chem Inf Model* 53(11), 2937-50.
21. http://www.pymol.org
22. Mattews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage Lysozyme. *Biochim Biophys Acta* 405, 442-51

Figures

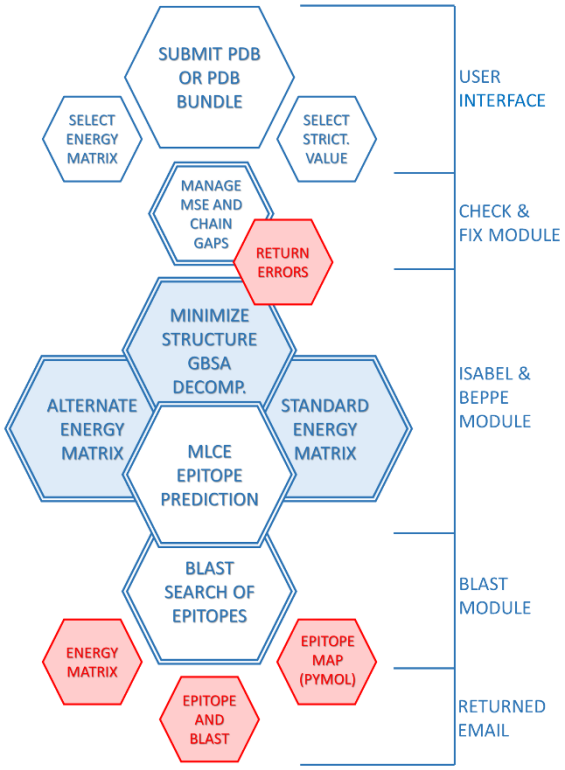


Fig. 1. Layout of the web tool BEPPE: each individual PDB file is passed by the User Interface to the prediction pipeline (double line hexagons). The main modules processing the structure, existent even in stand-alone format, consist of a preliminary input check script, the package ISABEL for the energy decomposition and analysis, the program BEPPE for the prediction of epitopes, and the BLAST script. All results (and eventually all errors) are gathered in a unique text file accompanied by supplementary data, and sent to the user via email (red, shaded background, single line hexagons).

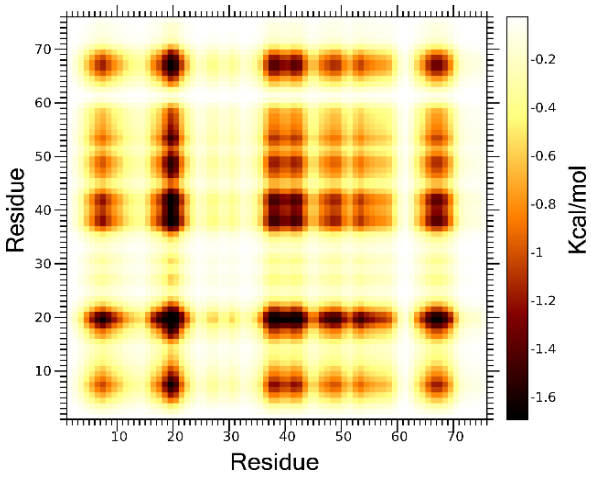


Fig. 2. Example of simplified energy matrix of the human Cyclin-dependent kinases regulatory subunit 1 (CKS1B), reconstructed from the first eigenvector of the non-bonded energy matrix, after diagonalization. The information is expressed on a residue pair basis, so each value represent the coupling energy of two amino acids on the protein sequence (x and y axis). The weaker the signal, the more intense is the color, highlighting the most energetically uncoupled residues. Conversely, those areas that carry the strongest energy of stabilization are depicted here as light bands.

Tables

Table 1. Overall performance analysis of BEPPE in the task of MHC II epitopes prediction. For every protein (PDB code) is reported the optimal cutoff for prediction and the classification performance parameters sensitivity, specificity, precision and accuracy *(5)*. The Matthews Correlation Coefficient (MCC) *(22)* test and statistical significance of the analysis (p-value) are also reported.

MHC-II EPITOPES
PREDICTION PERFORMANCE ANALYSIS

PDB ID	Cutoff	SENS.	SPEC.	PREC.	ACC.	MCC	p-value
1CB0.A	25%	0.55	0.63	0.06	0.63	0.07	1.42E-059
1D3B.A	15%	0.20	0.94	0.33	0.84	0.17	5.05E-003
1EA3.A	10%	0.14	1.00	1.00	0.36	0.20	4.00E-009
1HA0.A	25%	0.00	0.88	0.00	0.85	-0.06	7.16E-065
1I7Z.A	10%	0.44	0.90	0.16	0.88	0.22	7.32E-021
1OVA.D	25%	0.27	0.59	0.19	0.51	-0.12	3.41E-022
2GIB.B	25%	0.38	1.00	1.00	0.59	0.41	9.80E-008
2JK2.A	15%	0.87	0.86	0.28	0.86	0.44	6.36E-028
2VB1.A	25%	0.34	0.83	0.62	0.60	0.19	1.60E-001
2WA0.A	20%	0.70	0.82	0.16	0.82	0.27	4.27E-031
3BZH.A	25%	0.89	0.74	0.18	0.75	0.33	1.35E-026
3FEY.C	25%	0.50	0.74	0.08	0.73	0.11	1.28E-078
3HLA.A	15%	0.00	0.78	0.00	0.59	-0.25	1.09E-020
AVERAGE	**20%**	**0.41**	**0.82**	**0.31**	**0.69**	**0.15**	

Submitted to:

ACS Chemical Biology, Oct. 2014

Towards the Development of Diagnostic and Therapeutic Approaches against *B. pseudomallei*: Structure Based B Cell Epitope Design of BPSL1050 Antigen

Davide Gaudesi*[1], Claudio Peri[2*], Giacomo Quilici[1], Alessandro Gori[2], Mario Ferrer-Navarro[3], Oscar Conchillo-Solé[3], Rachael Thomas[4], Richard Titball[4], Xavier Daura[3,5], Giorgio Colombo[2§] and Giovanna Musco[1§]
[1]Biomolecular NMR Laboratory, Division of Genetics and Cell Biology, S. Raffaele Scientific Institute, Milan, Italy
[2]Department of computational biology, Institute for Molecular Recognition Chemistry, Italian National Research Council, Milan, Italy
[3]Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Spain
[4]College of Life and Environmental Sciences, University of Exeter, Exeter, UK
[5]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[*]Joint First Authors
[§]Joint Corresponding Authors: musco.giovanna@hsr.it; giorgio.colombo@icrm.cnr.it

Abstract

Bulkholderia pseudomallei is the etiological agent of melioidosis, a severe endemic disease in South-East Asia, causing septicemia and organ failure with high mortality rates. Current treatments and diagnostic approaches are largely ineffective. The development of new diagnostic tools and vaccines towards effective therapeutic opportunities against *B. pseudomallei* is therefore an urgent priority. In the framework of a multidisciplinary project tackling melioidosis through reverse and structural vaccinology, BPSL1050 was identified as a candidate for immunodiagnostic and vaccine development based on its reactivity against the sera of melioidosis patients. We determined its NMR solution structure and dynamics, and by novel computational methods we predicted immunogenic epitopes that once synthesized were able to elicit the production of antibodies inducing the agglutination of the bacterium and recognizing both BPSL1050 and *B. pseudomallei* crude extracts. Overall, these results hold promise for novel chemical biology approaches in the discovery of new diagnostic and prophylactic tools against melioidosis.

Introduction

Bulkholderia pseudomallei is the etiological agent of melioidosis, a severe endemic disease in South-East Asia, Australia, the Indian subcontinent, southern China, and with sporadic cases in South America [1]. The global incidence of the disease is unknown, but in rural areas of north Thailand it accounts for 40% of all deaths from community-acquired septicemia [2]. Because of its airborne transmission, high mortality rate, and multidrug resistance the Gram negative *B. pseudomallei* is also considered a potential bioterrorism agent and has been defined as category B priority pathogen by the US Centers for Disease Control and Prevention [3, 4]. *B. pseudomallei* is inherently resistant to many antibiotics, including first, second and third generation cephalosporins, aminoglycosides, penicillins and polymyxin [5], making its treatment extremely difficult. Moreover, it is a facultative intracellular bacterium with the ability to infect different cell types [6, 7]. Conceivably, intracellular replication and survival can assist the bacteria in circumventing the humoral immune system, explaining the long periods of latency, recrudescent or relapsing infections following many years after the initial infection. For all these reasons, development of effective *B. pseudomallei* vaccines and diagnostics is an urgent priority. However, at present there are neither effective methods of diagnosis nor approved vaccines for melioidosis in humans or animals [8, 9]. Of note, several melioidosis symptoms, e.g. pneumonia, chronic disease, septicemia, inner organ failure and localized abscesses, are shared by other pathologies, thus hampering a timely and correct diagnosis. The high mortality rate (between 30% to 70%) is mainly ascribed to the short time (24-48 h) available for antibiotic therapy intervention before the appearance of the acute phase symptoms [10]. In this respect, serological diagnosis assays based on the indirect hemagglutination test (IHA) [11] are preferred to the time consuming bacterial cultures [12]. However, IHA has several limitations: *i.* high background levels of antibody titers found in healthy persons living in endemic areas confuse the diagnostic results[13], *ii.* patient sera aspecific immunorecognition of whole bacterial lysate increases the probability of false-positive results, *iii.* limited test standardization leads to poor characterization of the different *B. pseudomallei* strains [14]. Reliable, specific and rapid diagnostic tools, such as serologic assays, using antigens that can differentiate asymptomatic from clinical infection, are essential for diagnosis prior to antibiotic treatment [10].

The identification, characterization and optimization of such antigens may constitute an important starting point towards the development of efficient management of *B. pseudomallei* infections.

In this context reverse vaccinology (RV), i.e. the process of genome-based antigen discovery, could strongly contribute to the efficient identification of surface-located proteins both as candidate vaccines against *B. pseudomallei* and as diagnostic bioprobes. This approach, complemented by recombinant antigen production and immunological tests to assess their antigenic/protective properties [15, 16] has proven to be essential both for the development of selective diagnostic tools and of therapeutic vaccines for other pathogens [17]. Structural vaccinology (SV) supports RV predicting epitopes recognized by antibodies (Abs) [18, 19, 20]. It provides a structural rationale for the synthesis of epitope mimics as bioprobes to selectively capture disease specific antibodies in patient sera, thus facilitating infection diagnosis [21, 22]. Furthermore, SV can be used as the starting point to rationally design peptides and to optimize the properties of antigenic proteins (or domains) that elicit the production of bacteria-neutralizing antibodies or that trigger a T-cell response.

Recently, a protein array of 1,205 *B. pseudomallei* proteins was used to map the antibody response in 747 serum samples from well-defined melioidosis positive and negative patients. This led to the identification of 49 protein antigens that are significantly more reactive to sera from *B. pseudomallei* infected patients, opening new opportunities for the generation of better diagnostics and vaccines [23]. Herein we present a successful example of a structural vaccinology approach on BPSL1050, a candidate for immunodiagnostic and vaccine development. Its solution structure and dynamics, combined to Matrix of Local Coupling Energy (MLCE) and Electrostatic Desolvation Profiles (EDP) methods [24] allowed for prediction of immunogenic epitopes. Structural and physicochemical principles, complemented and corroborated by experimental epitope mapping methods, guided the synthesis of reactive epitopes that elicited the production of Abs, which in turn effectively induced the bacterium agglutination. Notably, these Abs were even more effective in amplifying this immune response than Abs raised against the whole protein, thus supporting the development of strategies focussing on single antigenic protein elements.

Results and discussion

In this work we adopted a SV pipeline that entails: *i.* NMR structural characterization of an antigenic protein from *B. pseudomallei*; *ii.* computational prediction of epitope sequences; *iii.* their chemical synthesis as free or conjugated peptides; *iv.* elicitation of specific antibodies; *v*. Abs induced bacterial agglutination. We focused on BPSL1050, a highly immunoreactive hit emerging from the analysis of a protein array displaying RV-selected *B. pseudomallei* proteins [23]. The predicted pathogen surface exposure of BPSL1050, a highly conserved protein in the *Burkholderia* family (Figure S1), along with its relative small dimensions made it particularly suitable for SV investigations by means of multidimensional heteronuclear NMR spectroscopy (Table 1). Importantly, the epitope mimics generated on the basis of our SV strategy was able to recapitulate the main immunogenic determinants of the whole protein, eliciting antibodies that recognized recombinant full-length BPLS1050 and crude extract of *B. pseudomallei*. Remarkably, these Abs turned out to be more effective than Abs raised against the whole protein in triggering bacterial agglutination.

3D solution structure of BPSL1050

Recombinant BPSL1050 (residues Met5-Ala130) behaves as a monomer in solution, as assessed by gel-filtration elution volumes [25] and by the rotational correlation time (τc~9 ns, determined from ^{15}N relaxation data) (Figure S2), which is in good agreement with the expected value for a folded 14KDa protein. The structure consists of three helical regions comprising residues Asp8-Leu29 (α1), Thr74-Met81 (α2) and Gly85-Arg107 (α3), that pack onto an antiparallel β sheet, formed by 4 strands spanning residues His34-Val38 (β1), Thr47-Ala53 (β2), Leu66-Thr71 (β3), and Leu120-Gl24 (β6) (Figure 1A, B). Notably, helix α3 is characterized by a pronounced bend around Pro97 and is connected by a short loop (Glu82-Ala83) to helix α2, which is on turn almost perpendicular to the N-terminus of helix α1. Finally, residues Val109-Asp110 (β4), Thr115-Gln116 (β5) form a small β-strand which folds back like a lid towards the C-termini of helices α1 and α3. The overall protein topology is α1, β1, β2, β3, α2, α3, β4, β5, β6. The four stranded antiparallel β-sheet is solvent exposed on one side and packs tightly against the three helices on the other side,

generating a network of hydrophobic and aromatic interactions that contribute to tight packing of the protein (Figure 1B,C).

In agreement with its structural compactness, BPSL1050 is extremely stable, as assessed by the high melting temperature (T_m=67 C°) measured by Circular Dichroism thermal denaturation (Figure 1D). Interestingly, the C-terminal tail (spanning residues Gly125-Glu129) is well ordered and fills a groove formed by helix α2 and α3, as supported by a dense network of NOE contacts between Leu126, Phe128 and distal residues (Figure 1C). Most of the fold is well defined in the bundle of structures, with a root mean square deviation (rmsd) of 0.92 ±0.15 Å on the backbone atoms from residue Thr8-Gly129. Only the N-terminal residues (Met1-Pro7, whereby the first four residues belong to the residual N-terminal tag) and the protruding loops comprising Gly41-Gly44 (L1) and Pro54-Pro65 (L2) are highly disordered, as assessed by the paucity of NOEs. Of note, the amide of Gly42 and Gly44 are not visible in HSQC spectra, due to exchange with the solvent, and residues located in L2 display a clear decrease of their heteronuclear NOE values, indicative of motions in the ps-ns timescale (Figure 1E). Collectively, the high exposure and high internal mobility of L1 and L2 anticipate for these loops a potential role as antibody recognition sites.

Structure similarity searches with the DALI [26] server did not produce a significant match with any other protein except for CYAY from *Burkholderia cenocepacia* (with an rmsd of 4.2 Å over 82 residues and a Z score of 4.2). CYAY is a bacterial homolog of human frataxin, a mitochondrial protein important in iron homeoastasis [27]. Similarly to CYAY and human frataxin, BPSL1050 displays a patch of negatively charged residues on one edge of the protein (Figure 1E), reminiscent of a possible iron binding site. However NMR titrations with Fe^{2+} and Fe^{3+} and other metal ions (Mg^{2+}, Ca^{2+}, Zn^{2+}) did not show any interaction (data not shown), excluding a possible functional relationship between BPSL1050 and CYAY/Frataxin. BPSL1050 is highly conserved along the *Burkholderia* family, suggesting a fundamental yet unidentified function in the life of the pathogen. In this respect, the availability of an atomically detailed characterization of its structure and dynamics will be important in the screening/development of antimicrobial molecules targeting *B. pseudomallei*, as a complementary strategy to vaccine development

Characterization of protein dynamics via MD simulations and comparison with NMR-derived parameters.

Three 100ns MD simulation runs were performed, starting from BPSL1050 NMR solution structure. The Root Mean Square Fluctuation profile of the Cα atoms from the initial structure is well conserved among the three runs, and shows that the region encompassing residues Ala53 to Asn67, including loop L2, displays the largest degree of mobility (Figure 2A), in agreement with heteronuclear NOE data (Figure 1E). This loop is mostly unstructured and undergoes rearrangements paralleled by conformational changes of the lid region Val109-Ala119.

To investigate the major, non-random large-scale displacements of the protein substructures, the pair wise covariance matrix of atomic displacements was calculated for each trajectory. This matrix accounts for correlations in atomic motions and is used to highlight protein regions that move coherently. Principal components analysis (PCA) or Essential Dynamics (ED) analysis [28] can then be used to reduce the dimensionality of the matrix by diagonalization. This emphasizes the amplitude and direction of dominant global protein motions, which correspond to the ones occurring along the matrix eigenvectors associated to the main eigenvalues. PCA shows that the protein core, composed of the four main anti-parallel ß-sheets and part of the two principal α-helices, is stable and rigid throughout the simulations. Conversely, loop L1 and the hinge between helices α2 and α3, together with loop L2 and the Val109-Ala119 lid account for the principal conformational changes (Figures 2B, C). As non-random, cooperative displacements involve two opposite sides of the protein, the atomic details governing their mechanical connection were investigated by means of Coordination Propensity (CP) analysis whereby mechanically coordinated pairs of aminoacids are defined on the basis of their distance fluctuations: the lower the distance fluctuation, the higher the mechanical coordination (Figure 3A). Applying this analysis to all possible pairs within the protein and selecting the ones with minimal distance fluctuations we defined the network that connects the two regions undergoing non-random displacements, that encompasses helix α1, Ala53 and Pro54, and finally loop L2 (Figure 3B, C). In contrast, the region between Leu96 and the C-terminus, including part of helix α3 and β4 and β5 strands, appears mechanically uncoupled from the rest of the protein.

According to the CP profile (Figure 3B) helix *α3* is characterized by a strong internal coordination and behaves as an independent rigid body.

***In silico* epitope prediction**

Next, we investigated the antigenic determinants of the novel structure of BPSL1050. The 3D representative structures obtained from the clustering analysis of the MD simulations were used to predict the location of antibody-binding epitopes. To this end, a consensus prediction was generated using the MLCE and EDP methods. This pipeline for epitope prediction has already been successfully validated in other applications based on the concept that the 3D structure of an epitope, its related dynamics and exposure properties determine its antibody-binding, and consequently its antigenic potential [21, 22]. In this context, MLCE integrates the analysis of the dynamical and energetic properties of proteins to identify non-optimized energetic interaction-networks on the surface of the isolated antigens, which correspond to substructures that can aptly be recognized by a binding partner (the antibody). EDP calculates the free energy penalty for desolvation by placing a neutral probe at various protein surface locations. Surface regions with a small free energy penalty for water removal may correspond to preferred interaction sites. The sequences of the individual predictions resulting from the two methods and their location on the protein structure are reported in Table 2 and Figure 4A, respectively. Based on the resulting consensus, composed of different sequences located on two main areas ($_{39}$**GYGGHGHP**$_{46\text{-}53}$**APHAEHVRGYAP**$_{65}$ and $_{94}$**AALPRK**$_{99\text{-}101}$**AA**$_{102\text{-}104}$**ENARGVD**$_{110\text{-}115}$**T**-$_{117}$**ADA**$_{119}$), we predicted the presence of three main immunogenic sequences: the small loop L1, the main loop L2, and part of helix *α3 extending until Leu120*. These substructures are energetically uncoupled with respect to the rest of the protein, and their contribution to the overall fold stability is minimal. This finding is consistent with the dynamic behavior observed by CP analysis, whereby the main loop shows the highest mobility, and helix *α3* is not mechanically coordinated with respect to the protein core. Combining this information with the previous analysis, we hypothesized that these regions may represent viable candidate epitope sequences.

Experimental epitope mapping

The sequences identified by computational epitope predictions were next cross-validated with experimental epitope mapping experiments that were carried out using recombinant BPSL1050, cognate polyclonal sera and polyclonal rabbit sera using a protocol developed in-house [22, 29, 30]. Rabbit sera were collected before and after immunization. The pre-immune serum was used as control, to discard non-specific binding. Three different peptides were captured by the polyclonal IgGs from BPSL1050-immunised rabbit, with masses 2213.1 Da, 3268.430 Da and 6533.364 Da. In order to unequivocally determine the sequences of these peptides, MS/MS spectra were obtained (Figure S3). The corresponding sequences are $_{61}$**GYAHPLNLALTWNTDEIER**$_{79}$ for the 2213.1 Da ion and $_{93}$**LAAWENARGVDFGSRTQADALVLLGGLDFEA**$_{130}$ (EXP-2) for the 3268.430 Da ion. No MS/MS spectra could be obtained for the peak with a mass-to-charge ratio of 6533.364 Da due to its large mass and scanty ionization. Nevertheless, peptide mass fingerprinting assigns this peak to $_{21}$**IARAIADLLNHRAHTDVVGYGGHGHPTQVRIVAPHAEHVRGYAHPLNLALTWNTDEIER**$_{79}$ (EXP-1) that contains the sequence of 2213.1 Da identified by MS/MS. Interestingly the two peptides are located in the same region of the protein (Figure 4A).

Synthesis and conjugation of predicted epitopes.

To obtain a minimal number of peptides eliciting the production of Abs, we synthesized two sequences recapitulating the consensus between the computational and experimental epitope mapping methods. To this aim we used microwave-assisted Fmoc-chemistry, including triethyleneglycolate (O_2Oc) spacer units, and coupling to keyhole limpet hemocyanin (KLH) to serve as carrier proteins. The two consensus peptides were L1050#1 (connecting L1 and L2 Cys--$(O_2Oc)_{2-38}$V**GYGGHGHP**TQVRIV**APHAEHVRGYAP$_{65}$**) and L1050#2 *(e*xtending from helix α3 $_{84}$DGAARFERYL**AALPRK**L**AA**W**ENARGVD**FGSRTQ**ADA**L$_{120}$-$(O_2Oc)_2$-Cys), with the computationally predicted mapped epitope residues highlighted in bold (Table 2, Figure 4B). In both cases we combined short predicted sequences in two unique constructs to allow for the simultaneous presentation of different potential epitopes.

Evaluation of anti-L1050#1 and anti-L1050#2 rabbit Abs.

The two synthetic epitopes were used to elicit Abs production in rabbits. After purification Abs were evaluated for specificity against full length BPSL1050 and single peptides using indirect ELISA. Notably, Abs raised against the individual peptides recognized full-length BPSL1050 (Figure 5B) as effectively as the immunizing peptide (Figure 5C, D). They also recognized a crude extract of *B. pseudomallei* (Figure 5A). These results confirm that the two epitope mimics recapitulate the immunogenic determinants of the protein. Remarkably, considering the lack of reliable diagnostic tools for melioidosis, the Abs ability to recognize cognate antigen could aptly be exploited to generate biomarkers for the efficient detection of *B. pseudomallei* infections. In this perspective, designed epitopes and Abs could be used as bioprobes in combination for microarray display or for multiplexed ELISA tests, fostering the development of rapid pathogen detection tests, and ultimately allowing a timely pharmacological intervention.

Agglutination ability of anti-BPSL1050, anti-L1050#1 and anti-L1050#2 Abs

We next tested the generated Abs for their agglutination ability (Figure 6A-D). Agglutination is a host defense mechanism in which antibodies recognize and bind to surface antigens on multiple bacteria eventually causing their aggregation and consequent microbial death [31]. It also favors bacterial clearance by phagocytes and impedes pathogens migration through tissues, limiting the infection [32]. Abs raised against BPSL1050, L1050#1 and L1050#2 were all able to agglutinate *B. pseudomallei* cells and not *E. coli* control cells (data not shown), indicating that the effect promoted by our Abs is specific for *B. pseudomallei*. The full length protein was sufficient to induce a faint effect after 30 minutes (as shown by the red arrow in Figure 6B). Of note, Abs against the peptides were more effective than Abs raised against the whole protein in triggering pathogens cells agglutination (Figure 6C, D), suggesting that it is possible to amplify the protective immune response using a single antigenic element of the antigen. Collectively, the agglutination and clearance effects elicited by the Abs against the designed peptides demonstrate the efficacy of strategies focussing on the specific immunogenic determinants of antigens to design biomolecules able to activate protective immune responses.

These Abs could for instance be exploited in passive immunization, as well as adjuvants in combination with other components in new therapeutic vaccines

Conclusions and Perspectives

In conclusion, the results presented on this previously unexplored *B. pseudomallei* protein provide support for the efficacy of strategies combining structural investigations, computational design approaches and experimental immunological analyses to understand the antigenic determinants of protein antigens. Our results might have applications in the design of specific bioprobes (e.g. peptides, and antibodies) that, combined to other antigenic derivatives, could be effective in the development of rapid diagnostic and therapeutic approaches against melioidosis.

The increasing knowledge on the epitopes deriving from different *B. pseudomallei* proteins and their immunoreactivities opens new perspectives in the chemical biology strategies against the pathogen. These entail on the one hand the production of Abs for passive immunization, and on the other hand the design of chemical multipresentation systems [33, 34] as potential immunogens. In particular, nanoparticles and dendrimers, simultaneously displaying multiple antigenic sequences that recapitulate the immunogenic determinants of various targets, could have therapeutic potential for the development of novel vaccination strategies.

Methods

Protein expression and purification

Isotopically enriched BPSL1050 ($^{15}N/^{13}C$, or ^{15}N labelled) was expressed and purified as described previously [25]. NMR samples were prepared in a buffer containing 150 mM NaCl, 20 mM NaH_2PO_4, 20 mM Na_2HPO_4, 2 mM DTT buffer at pH 7 with 0.16 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) and 10% D_2O.

NMR measurements

NMR experiments were performed at 25 °C on a 600-MHz spectrometer (Bruker Avance 600 Ultra Shield TM Plus, Bruker BioSpin) equipped with a triple-resonance TCI

cryoprobe with a z-shielded pulsed-field gradient coil. Data were processed with Topspin (v. 2.2) (Bruker Biospin) and/or with NMRPipe [35] and analysed with CCPN analysis 2.2 software [36]. The ^{1}H, ^{13}C and ^{15}N chemical shifts of backbone side and chains resonances have been assigned as described previously [25]. Proton–proton distance constraints were obtained from ^{15}N and ^{13}C separated 3D NOESY spectra employing 100 ms mixing times. ^{3}J(HN, Ha) coupling constants were measured to derive restraints for ϕ dihedral angles. Additional ϕ/ψ restraints were obtained from backbone chemical shifts using TALOS+ [37]. Hydrogen bond restraints were defined from slow-exchanging amide protons identified after H_2O/D_2O exchange . ^{1}H–^{15}N residual dipolar couplings were measured in isotropic and anisotropic phases obtained using polyacrylamide gel (8.5% (w/v). Heteronuclear {^{1}H} ^{15}N nuclear Overhauser enhancements (hetNOEs), longitudinal and transversal ^{15}N relaxation rates (T_1, T_2) were measured using standard 2D methods [38], duty-cycle heating compensation were used for both T_1 and T_2 relaxation experiments [39]. T_1 and T_2 decay curves were sampled at 12 different time points (50–2000 ms and 12–244 ms, respectively) collected in random order, with 2.5 seconds recovery delay. The {^{1}H}^{15}N NOEs were measured recording HSQC [38] spectra with and without proton saturation in an interleaved fashion using a 4 seconds recycle delay. T_1, T_2 and hetNOE values have been obtained using the NMRView fitting routine [40].

Structure calculation

Structures were calculated using ARIA 2.3.1 [41] in combination with CNS [42] using the experimentally derived restraints (Table 1). All NOEs were assigned manually and calibrated by ARIA 2.3.1. A total of eight iterations (20 structures in the first six iterations) were performed: 100 structures were computed in the last iteration. The ARIA2.3.1 default water refinement was performed on the 15 best structures of the final iteration. Structural quality was assessed using PROCHECK-NMR [43] and CING [44]**.** The family of the 15 lowest energy structures has been deposited in the Protein Data Bank with the accession code 2mpe**.**

Circular dichroism measurements

CD spectra (20°C) were acquired on a Jasco J-815 using a rectangular quartz cuvette (1 mm path length, Hellma). Each spectrum was averaged over 4 scans collected in 0.1 nm

intervals with an average time of 0.5 s. BPSL1050 concentration was 10 μM in 10 mM NaF, 20 mM NaH_2PO4/Na_2HPO4 buffer (pH 7). Thermal denaturation curves were obtained monitoring the ellipticity at 222 nm and at 200 nm from 20 °C to 95 °C at a constant scan rate of 1 C°/min.

Molecular Dynamics Simulations and Signal Propagation Analysis

Three structures extracted from the NMR bundle where used for three independent MD simulations (100 ns each). Simulations and analyses were performed using the GROMACS 4.51 software package [45], GROMOS96 force field [46] and the SPC water model [47]. Simulations details are described in Supplemental Information.

The main structural fluctuations of the protein were analyzed through Principal Component Analysis of the trajectories [28]). Residue-pair Communication Propensity (CP) is calculated as a function of the fluctuations of interresidue distances [48]. Details are described in Supplemental Information.

***In silico/in vitro* epitope prediction, mapping and synthesis**

All details for computational epitope prediction, epitope mapping with murine sera, peptide synthesis and conjugation to the carrier protein are described in Supplemental Information.

Rabbit immunizations and generation of Abs.

New Zealand white rabbits were subcutaneously vaccinated in multiple sites in hind quarter locations with purified BPSL1050 in a 1 mL volume emulsion of complete Freud's adjuvant (Sigma Chemicals co., St Louis, MO, USA). Boosters were administered at days 14, 28, 42, 56, 70, 84 and 98 in a 1 mL volume emulsion of incomplete Freud's adjuvant. Blood samples were collected from the central auricular artery in between immunizations and exsanguinations took place on day 107 (Harlan, Derby, UK).

Serum antibody concentration

The concentration of antigen specific antibody in the terminal immune sera was determined by ELISA. Experimental details are summarized in Supplemental Information.

Production and purification of anti-L1050#1 and anti-L1050#2 antibodies.

Rabbit polyclonal antibodies were generated by Primm srl, Milano Italy. Antisera were immunopurified against peptide, chemically linked to Cyanogen Bromide Activated Sepharose (Sigma-Aldrich), following the manufacturer's instructions.

Indirect ELISA to detect antibodies to *B. pseudomallei*.

Antibody recognition of the two synthesized peptides (L1050#1 and L1050#2) and the full BPSL1050 was detected using indirect ELISA using 96-well microtitre plates (Nunc Maxisorp) that were uncoated or coated with 1 µg/mL of crude *B. pseudomallei* extract as a control, 10 µg/ml of BPSL1050 protein, or 3 µg/mL peptides for L1050#1 and L1050#2 in 0.1 M carbonate-bicarbonate buffer (pH 9.6), incubating at 37°C for 3 h.. The results were normalized with O.D. of uncoated well, and represented as Absorbance index = ($O.D._{450nm}$ test - $O.D._{450nm}$ uncoat) / $O.D._{450nm}$ uncoat.

Agglutination of *B. pseudomallei* after exposure to antibodies raised against BPSL1050, L1050#1 and L1050#2

RFP-expressing *B. pseudomallei* K96243 [49] or *E. coli* 29522 (ATCC), as negative control, were sub-cultured from overnight broths into LB and grown to log phase. Bacteria (1 x 108) were incubated with either 1 µg BPSL1050 antisera or 1 µg antibody towards BPSL1050 epitope 1 (L1050#1) or epitope 2 (L1050#2) reconstituted in PBS. Controls were pre-bleed from the rabbit before immunization, *B. pseudomallei* capsule antibody 4V1H12 [50] and PBS alone. Incubation was performed static at 37 °C for 30 minutes. A 30 µL aliquot of culture was placed onto a glass cover slip and allowed to air dry before fixing in 1% paraformaldehyde for 10 min. The cover slips were washed three times in PBS and mounted onto glass microscope slides using VECTORshield hardset mounting media with DAPI (Vector laboratories, Burlingame, CA). Bacterial agglutination was observed under epi-fluorescence microscopy (Zeiss, Oberkochen, Germany).

Acknowledgements

This project was supported by Fondazione CARIPLO (Progetto Vaccini, contract number 2009-3577), by Fondazione Telethon (TDGM00307TU), and by the Italian Ministry of Education and Research through the Flagship (PB05) "InterOmics.

References

1. Currie, B. J., Dance, D. A., and Cheng, A. C. (2008) The global distribution of burkholderia pseudomallei and melioidosis: An update. *Trans. R. Soc. Trop. Med. Hyg. 102 Suppl 1*, S1-4.
2. Wiersinga, W. J., van der Poll, T., White, N. J., Day, N. P., and Peacock, S. J. (2006) Melioidosis: Insights into the pathogenicity of burkholderia pseudomallei. *Nat. Rev. Microbiol. 4*, 272-282.
3. Aldhous, P. (2005) Tropical medicine: Melioidosis? never heard of it.. *Nature. 434*, 692-693.
4. Bondi, S. K., and Goldberg, J. B. (2008) Strategies toward vaccines against burkholderia mallei and burkholderia pseudomallei. *Expert Rev. Vaccines. 7*, 1357-1365.
5. Peacock, S. J., Limmathurotsakul, D., Lubell, Y., Koh, G. C., White, L. J., Day, N. P., and Titball, R. W. (2012) Melioidosis vaccines: A systematic review and appraisal of the potential to exploit biodefense vaccines for public health purposes. *PLoS Negl Trop. Dis. 6*, e1488.
6. Harley, V. S., Dance, D. A., Drasar, B. S., and Tovey, G. (1998) Effects of burkholderia pseudomallei and other burkholderia species on eukaryotic cells in tissue culture. *Microbios. 96*, 71-93.
7. Gan, Y. (2005) Interaction between burkholderia pseudomallei and the host immune response: Sleeping with the enemy? *Journal of Infectious Diseases. 192*, 1845-1850.
8. Dowling, A. J. (2013) Novel gain of function approaches for vaccine candidate identification in burkholderia pseudomallei. *Front. Cell. Infect. Microbiol. 2*, 139.
9. Sarkar-Tyson, M., and Titball, R. W. (2010) Progress toward development of vaccines against melioidosis: A review. *Clin. Ther. 32*, 1437-1445.
10. Cheng, A. C., and Currie, B. J. (2005) Melioidosis: Epidemiology, pathophysiology, and management. *Clin. Microbiol. Rev. 18*, 383-416.
11. Harris, P. N. A., Williams, N. L., Morris, J. L., Ketheesan, N., and Norton, R. E. (2011) Evidence of burkholderia pseudomallei-specific immunity in patient sera persistently nonreactive by the indirect hemagglutination assay. *Clin Vaccine Immunol. 18*, 1288-1291.
12. Sirisinha, S., Anuntagool, N., Dharakul, T., Ekpo, P., Wongratanacheewin, S., Naigowit, P., Petchclai, B., Thamlikitkul, V., and Suputtamongkol, Y. (2000) Recent developments in laboratory diagnosis of melioidosis. *Acta Trop. 74*, 235-245.
13. Sirisinha, S. (1991) Diagnostic value of serological tests for melioidosis in an endemic area. *Asian Pac. J. Allergy Immunol. 9*, 1-3.
14. Hara, Y., Chin, C. Y., Mohamed, R., Puthucheary, S. D., and Nathan, S. (2013) Multiple-antigen ELISA for melioidosis--a novel approach to the improved serodiagnosis of melioidosis. *BMC Infect. Dis. 13*, 165-2334-13-165.
15. Mora, M., Veggi, D., Santini, L., Pizza, M., and Rappuoli, R. (2003) Reverse vaccinology. *Drug Discov. Today. 8*, 459-464.
16. Serruto, D., Bottomley, M. J., Ram, S., Giuliani, M. M., and Rappuoli, R. (2012) The new multicomponent vaccine against meningococcal serogroup B, 4CMenB: Immunological, functional and structural characterization of the antigens. *Vaccine. 30 Suppl 2*, B87-97.
17. De Gregorio, E., and Rappuoli, R. (2014) From empiricism to rational design: A personal perspective of the evolution of vaccine development. *Nat. Rev. Immunol. 14*, 505-514.
18. Dormitzer, P. R., Grandi, G., and Rappuoli, R. (2012) Structural vaccinology starts to deliver. *Nat. Rev. Microbiol. 10*, 807-813.
19. Dormitzer, P. R., Ulmer, J. B., and Rappuoli, R. (2008) Structure-based antigen design: A strategy for next generation vaccines. *Trends Biotechnol. 26*, 659-667.
20. Nuccitelli, A., Cozzi, R., Gourlay, L. J., Donnarumma, D., Necchi, F., Norais, N., Telford, J. L., Rappuoli, R., Bolognesi, M., Maione, D., Grandi, G., and Rinaudo, C. D. (2011) Structure-based approach to rationally design

a chimeric protein for an effective vaccine against group B streptococcus infections. *Proc. Natl. Acad. Sci. U. S. A. 108*, 10278-10283.
21. Gourlay, L. J., Peri, C., Ferrer-Navarro, M., Conchillo-Sole, O., Gori, A., Rinchai, D., Thomas, R. J., Champion, O. L., Michell, S. L., Kewcharoenwong, C., Nithichanon, A., Lassaux, P., Perletti, L., Longhi, R., Lertmemongkolchai, G., Titball, R. W., Daura, X., Colombo, G., and Bolognesi, M. (2013) Exploiting the burkholderia pseudomallei acute phase antigen BPSL2765 for structure-based epitope discovery/design in structural vaccinology. *Chem. Biol. 20*, 1147-1156.
22. Lassaux, P., Peri, C., Ferrer-Navarro, M., Gourlay, L. J., Gori, A., Conchillo-Sole, O., Rinchai, D., Lertmemongkolchai, G., Longhi, R., Daura, X., Colombo, G., and Bolognesi, M. (2013) A structure-based strategy for epitope discovery in burkholderia pseudomallei OppA antigen. *Structure. 21*, 167-175.
23. Felgner, P. L., Kayala, M. A., Vigil, A., Burk, C., Nakajima-Sasaki, R., Pablo, J., Molina, D. M., Hirst, S., Chew, J. S., Wang, D., Tan, G., Duffield, M., Yang, R., Neel, J., Chantratita, N., Bancroft, G., Lertmemongkolchai, G., Davies, D. H., Baldi, P., Peacock, S., and Titball, R. W. (2009) A burkholderia pseudomallei protein microarray reveals serodiagnostic and cross-reactive antigens. *Proc. Natl. Acad. Sci. U. S. A. 106*, 13499-13504.
24. Fiorucci, S., and Zacharias, M. (2010) Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophys. J. 98*, 1921-1930.
25. Gaudesi, D., Quilici, G., and Musco, G. (2013) H, C, N backbone and side chain NMR resonance assignments of BPSL1050 from burkholderia pseudomallei. *Biomol. NMR Assign.*
26. Holm, L., and Rosenstrom, P. (2010) Dali server: Conservation mapping in 3D. *Nucleic Acids Res. 38*, W545-9.
27. Nair, M., Adinolfi, S., Pastore, C., Kelly, G., Temussi, P., and Pastore, A. (2004) Solution structure of the bacterial frataxin ortholog, CyaY: Mapping the iron binding sites. *Structure. 12*, 2037-2048.
28. Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993) Essential dynamics of proteins. *Proteins. 17*, 412-425.
29. Soriani, M., Petit, P., Grifantini, R., Petracca, R., Gancitano, G., Frigimelica, E., Nardelli, F., Garcia, C., Spinelli, S., Scarabelli, G., Fiorucci, S., Affentranger, R., Ferrer-Navarro, M., Zacharias, M., Colombo, G., Vuillard, L., Daura, X., and Grandi, G. (2010) Exploiting antigenic diversity for vaccine design: The chlamydia ArtJ paradigm. *J. Biol. Chem. 285*, 30126-30138.
30. Koehler, C., Carlier, L., Veggi, D., Balducci, E., Di Marcello, F., Ferrer-Navarro, M., Pizza, M., Daura, X., Soriani, M., Boelens, R., and Bonvin, A. M. (2011) Structural and biochemical characterization of NarE, an iron-containing ADP-ribosyltransferase from neisseria meningitidis. *J. Biol. Chem. 286*, 14842-14851.
31. Pal'tsyn, A. A., Kolokol'chikova, E. G., Badikova, A. K., Chervonskaia, N. V., and Grishina, I. A. (1999) The role of agglutination during bacterial infection. *Biull. Eksp. Biol. Med. 127*, 4-8.
32. Bull, C. G. (1915) The agglutination of bacteria in vivo. *J. Exp. Med. 22*, 484-491.
33. Gori, A., Longhi, R., Peri, C., and Colombo, G. (2013) Peptides for immunological purposes: Design, strategies and applications. *Amino Acids. 45*, 257-268.
34. Kodadek, T. (2014) Chemical tools to monitor and manipulate the adaptive immune system. *Chem. Biol. 21*, 1066-1074.
35. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR. 6*, 277-293.
36. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins. 59*, 687-696.
37. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR. 44*, 213-223.
38. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G., Shoelson, S. E., Pawson, T., Forman-Kay, J. D., and Kay, L. E. (1994) Backbone dynamics of a free and phosphopeptide-complexed src homology 2 domain studied by 15N NMR relaxation. *Biochemistry. 33*, 5984-6003.
39. Yip, G. N., and Zuiderweg, E. R. (2005) Improvement of duty-cycle heating compensation in NMR spin relaxation experiments. *J. Magn. Reson. 176*, 171-178.
40. Johnson, B. A., and Blevins, R. A. (1994) NMR view: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR. 4*, 603-614.
41. Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T. E., and Nilges, M. (2007) ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics. 23*, 381-382.
42. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr. 54*, 905-921.

43. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR. 8*, 477-486.
44. Doreleijers, J. F., Sousa da Silva, A. W., Krieger, E., Nabuurs, S. B., Spronk, C. A., Stevens, T. J., Vranken, W. F., Vriend, G., and Vuister, G. W. (2012) CING: An integrated residue-based structure validation program suite. *J. Biomol. NMR. 54*, 267-283.
45. Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput. 4*, 435-447.
46. van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D. P., Glattli, A., Hunenberger, P. H., Kastenholz, M. A., Oostenbrink, C., Schenk, M., Trzesniak, D., van der Vegt, N. F., and Yu, H. B. (2006) Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem. Int. Ed Engl. 45*, 4064-4092.
47. Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987) The missing term in effective pair potentials. *J. Phys. Chem-Us. 91*, 6269-6271.
48. Morra, G., Verkhivker, G., and Colombo, G. (2009) Modeling signal propagation mechanisms and ligand-based conformational dynamics of the Hsp90 molecular chaperone full-length dimer. *PLoS Comput. Biol. 5*, e1000323.
49. Wand, M. E., Muller, C. M., Titball, R. W., and Michell, S. L. (2011) Macrophage and galleria mellonella infection models reflect the virulence of naturally occurring isolates of B. pseudomallei, B. thailandensis and B. oklahomensis. *BMC Microbiol. 11*, 11-2180-11-11.
50. Cuccui, J., Milne, T. S., Harmer, N., George, A. J., Harding, S. V., Dean, R. E., Scott, A. E., Sarkar-Tyson, M., Wren, B. W., Titball, R. W., and Prior, J. L. (2012) Characterization of the burkholderia pseudomallei K96243 capsular polysaccharide I coding region. *Infect. Immun. 80*, 1209-1221.

Figures and Tables

Prediction method	Name	Epitope sequence
MLCE	MD1-E1	$_{47}$IQAGL$_{51}$
MLCE	MD2-E1	$_{45}$D $_{47}$IQAG$_{50\ 54}$GGQTG$_{58}$
MLCE	MD3-E1	$_{50}$GLIIGGQT$_{58}$
MLCE	MD1-E2	$_{78}$SVG$_{80\ 83}$AGAQS$_{87}$
MLCE	MD2-E2	$_{78}$SVGLQAGAQSK$_{88}$
MLCE	MD3-E2	$_{77}$LSVG$_{80\ 83}$AG$_{84}$
MLCE	MD1-E3	$_{110}$AAGADA$_{115\ 117}$V $_{122}$MGANGAIDTTTATA$_{135}$
MLCE	MD2-E3	$_{122}$MGANGAIDT$_{130\ 132}$TATA$_{135}$
MLCE	MD3-E3	$_{110}$A $_{112}$GADASVA$_{118\ 121}$KMGANGAIDTTTATAP$_{136}$
EDP	EDP-1	$_{44}$PDVIQAGLIIGGQTGN$_{59}$
EDP	EDP-2	$_{76}$SLSVGLQAGAQSK$_{88}$
Epitope mapping	EXP-1	$_{39}$GVLVFPDVIQAGLIIGGQTGNGALR$_{63}$
Synthetic peptide	Lipo#1	$_{40}$VLVFPDVIQAGLIIGGQTGNGALRV$_{64}$
Synthetic peptide	Lipo#2	$_{76}$SLSVGLQAGAQSK$_{88}$
Synthetic peptide	Lipo#3	$_{110}$AAGADASVALVKMGANGAIDTTTATAPVE$_{137}$

Table 2: List of computationally (MLCE; EDP) predicted epitopes and synthesized peptide

Table 1: Summary of conformational constraints and statistics for BPSL1050

Experimental distance restraints[a,b]	
All	1861
Sequential (|i-j| = 1)	353
Medium range (1< |i-j| ≤ 4)	234
Long range (|i-j| > 4)	282
Intraresidual	937
Hydrogen bonds	55
$^1D_{NH}$	49
Dihedral angles (φ)	72
Dihedral angles (ψ)	72
Deviation from idealized covalent geometry	
Bonds (Å)	0.0030 ± 0.00006
Angles (°)	0.445 ± 0.008
Coordinate Precision (Å)	
N, C^{α}, C' / all heavy atoms (residues 8-129)[c]	0.920 ± 0.145/ 1.145 ± 0.132
N, C^{α}, C' / all heavy atoms 2nd structure[d]	0.368 ± 0.074/ 0.650 ± 0.064
Structural quality	
Procheck	
% residues in most favored region of Ramachandran plot	89.0
% residues in additionally allowed region	10.5
% residues in generously allowed region	0.2
% residues in disallowed region[e]	0.2
iCING[f]	
Red, (%, #)	11, 14
Orange, (%, #)	13, 17
Green, (%, #)	76, 99

a) No distance restraint was violated by more than 0.5Å.

b) No dihedral angle restraints was violated by more than 5°.

c) R.m.s. deviation between the ensemble of structures and the mean structure .

d) R.m.s. deviation between the ensemble of structures and the mean structure calculated on residues 8-29, 34-38, 47-53, 66-71, 74-81, 85-107, 109-110, 115-116 and 120-124.

e) Residues in disallowed regions are located on the flexible loops L1 and L2.

f) interactive Common Interface for NMR structure Generation ROG (Red, Orange, Green) factor: % residues percentage, # number of residues. Residues in red are located on the flexible loops L1a nd L2.

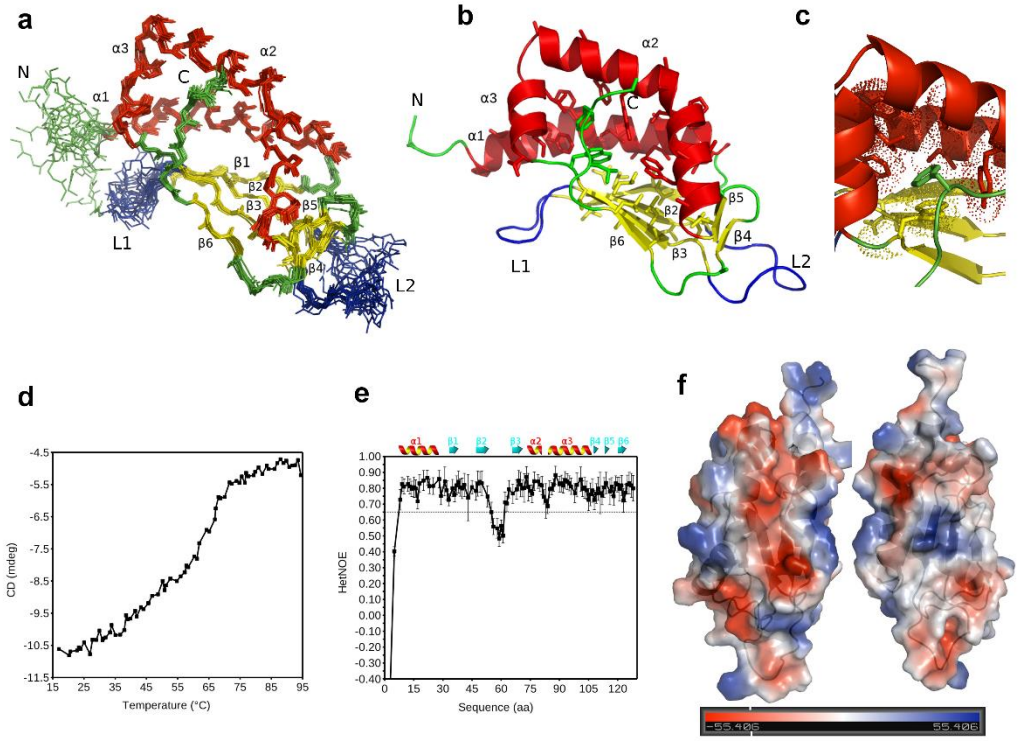


Figure 1: Solution structure of BPSL1050. (A) Superposition of BPSL1050 of the best 15 BPSL1050 structures, α-helices, ß-strands and mobile loops (L1 and L2) are represented in red, yellow and blue, respectively. (B) Cartoon representation of BPSL1050 showing the hydrophobic side-chains (in sticks) stabilizing the protein fold. (C) Zoom into the hydrophobic cluster around the C-terminal Phe128 (green). (D) Thermal denaturation curve of BPSL1050 (10 µM in 10 mM NaF, 20 mM NaH2PO4/Na2HPO4 buffer, pH 7) monitored at 222 nm by far-UV circular dichroism. (E) Backbone dynamics of BPSL1050; residues of L2 display heteronuclear NOEs <0.65 (gray line). (F) Electrostatic surface representation of BPSL1050 showing a patch of negatively charged residues on one side of the protein, the image on the right is rotated by 180° around the z axis.

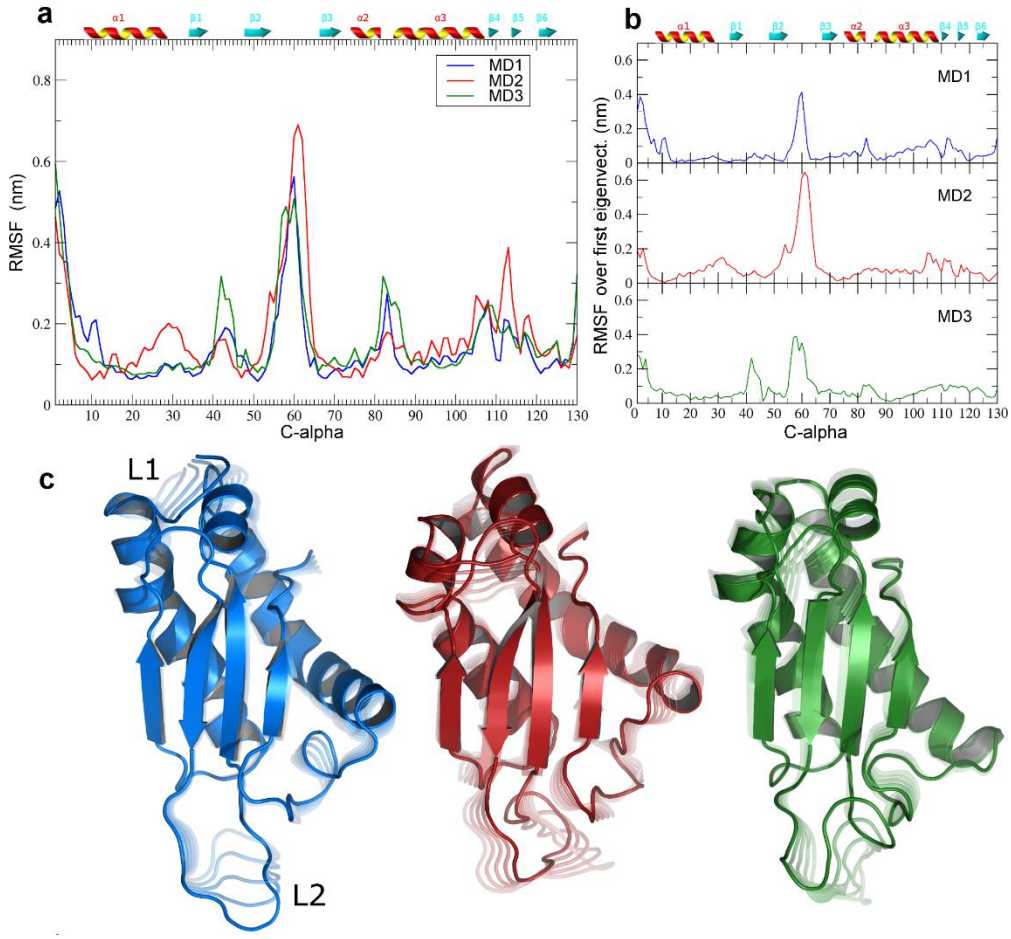


Figure 2: Visual representation of the principal motions of BPSL1050 according to MD1-3. (A) Root Mean Square Fluctuation (RMSF) of the residue positions with respect to the NMR starting structure calculated over the 3 combined MD trajectories. (B) Residue-based RMSF calculated after projecting each trajectory along the main ED eigenvectors. (C) Structures representing the extreme conformations corresponding to a principal motion (solid color and most transparent structure). The coordinates of the two extremes were interpolated into three intermediate positions resulting in a blurry effect. The blur amplitude indicates the protein regions involved in the motion. Blue, red and green structures are representative of the principal component (eigenvector) of simulations MD1, MD2 and MD3, respectively.

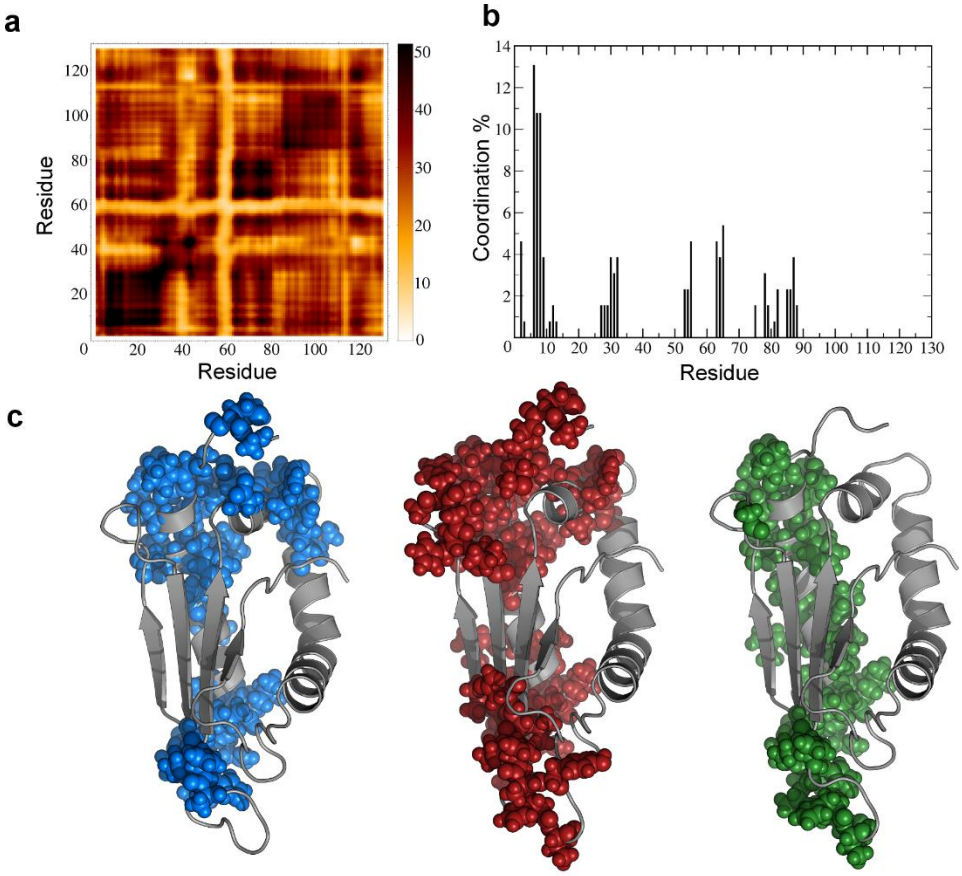


Figure 3: Communication propensities (CP) on BPSLI050. (A) Distance Fluctuation (DF) Matrix calculated on the timeframes of MDI. The non-dimensional color intensity indicates the coordination between any pair of residues of the protein sequence. (B) The Coordination Propensity (CP) profile is obtained from the DF Matrix after diagonalization and the selection of all coordinated pairs beyond a fixed range distance retaining a DF intensity higher than local average. The percentage of coordination of these residues is displayed in the profile. (C) Spheres representation highlights on the 3D structures the coordinated residues as reported by the profiles of MD1, MD2 and MD3 (in blue, red and green respectively). The network connecting two distal regions of BPSLI050 by means of helix α1 is visible in all three simulations.

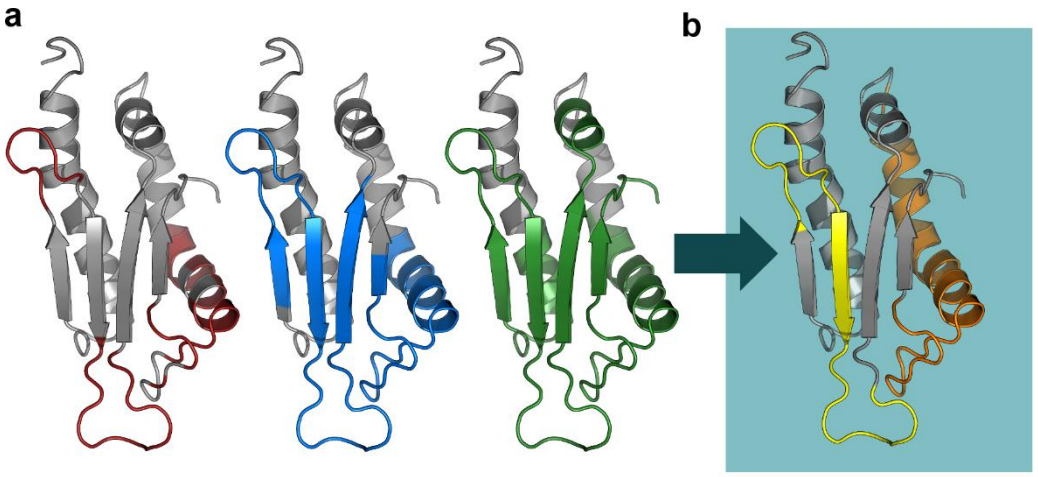


Figure 4: Cartoon and color representation of the predicted, mapped and synthesized epitopes. (A) All the antibody-binding regions predicted from three MD simulations are displayed in red (MLCE) and blue (EDP). The sequences captured by experimental epitope mapping are shown in green. (B) Based on the predictions and experimental indications, two epitopes have been chosen for synthesis, in yellow and orange for L1050#1 and L1050#2, respectively.

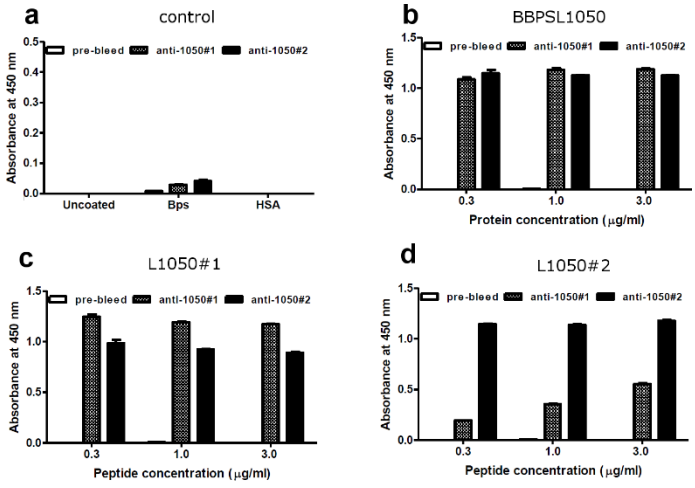


Figure 5: Recognition of BPSL1050 protein/peptides by anti-L1050#1 and anti-L1050#2 antisera (A) Crude Bps antigen, (B) BPSL1050 protein, BPSL1050 peptides (C) L1050#1 and (D) L1050#2 were individually coated into 96-well polystyrene plates. Then, 0.5 µg/ml of either un-immunised rabbit sera or sera from rabbits immunised with L1050#1 (hatched bars) or L1050#2 (solid bars) were tested in duplicate for binding to the plate-immobilised antigens. Rabbit antibodies that bound were detected using a suitable HRP-anti-hIgG conjugate.

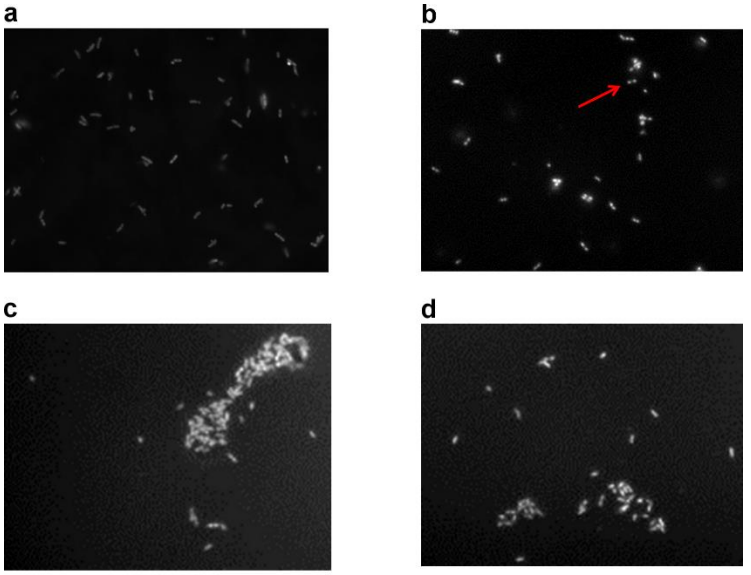
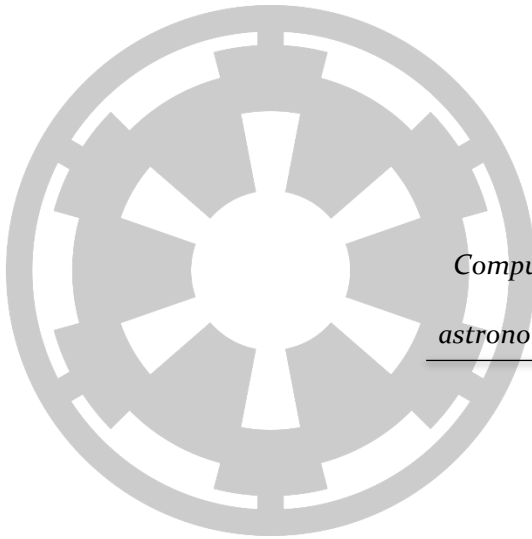


Figure 6: Agglutination of B. pseudomallei after exposure to antibodies raised against BPSL1050 or L1050#1 and L1050#2. RFP-expressing B. pseudomallei was incubated with 1 µg of antibodies for 30 min at 37°C. Bacterial agglutination was observed under fluorescence microscopy. Panels show (A) RFP-expressing B. pseudomallei incubated with PBS as negative control, (B) antibodies raised against BPSL1050 whole protein where red arrow indicates agglutination, (C) antibodies raised against L1050#1 and (D) antibodies raised against L1050#2. Magnification = X100.

PART III - SUPPLEMENTARY MATERIALS.



Computer science is no more about computers than astronomy is about telescopes

Edsger Dijkstra

The final part gathers additional information and data regarding works still unpublished. All data is presented in form of figures and tables. The methods employed (MD simulations and predictions) are consistent with the procedures described in part II.

Figure SI: Sequence alignment of all *in silico* and *in vitro* B-cell epitope prediction and mapping techniques, compared to final, synthesized T-cell/B-cell peptides. $FliC_{Bp}$ full-length sequence illustrating epitope residues predicted by MLCE (blue), EDP (green) based on the $FliC_{Bp}$ crystal structure and MD simulations; the consensus of three (BepiPred, BCPred and AAP) B-cell sequence-based prediction servers (purple), experimentally mapped peptides (red) and the three synthesized peptides (orange) that represent the final choice of putative B-cell epitopes, coinciding with sequence-based predictions. Grey shaded boxes indicate residues shared between two or more independent identification methods. The N and C-terminal residues visible in the electron density of the structure are underlined. (The synthesized peptides include also sequence 213-231, selected as a putative T-cell epitope)

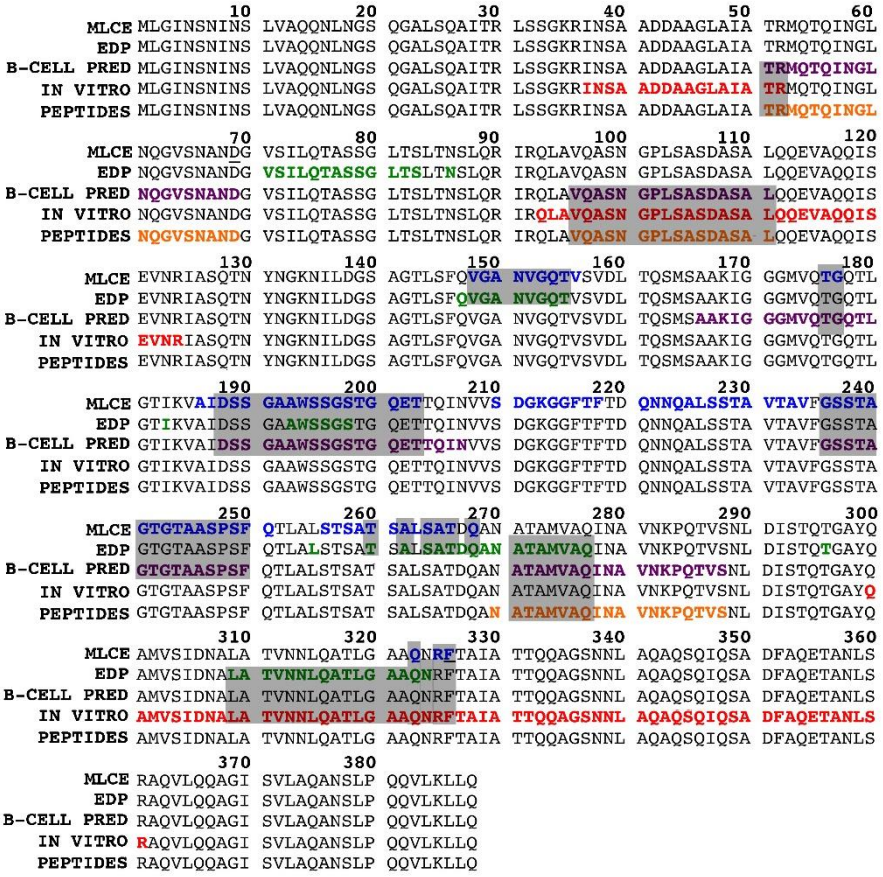
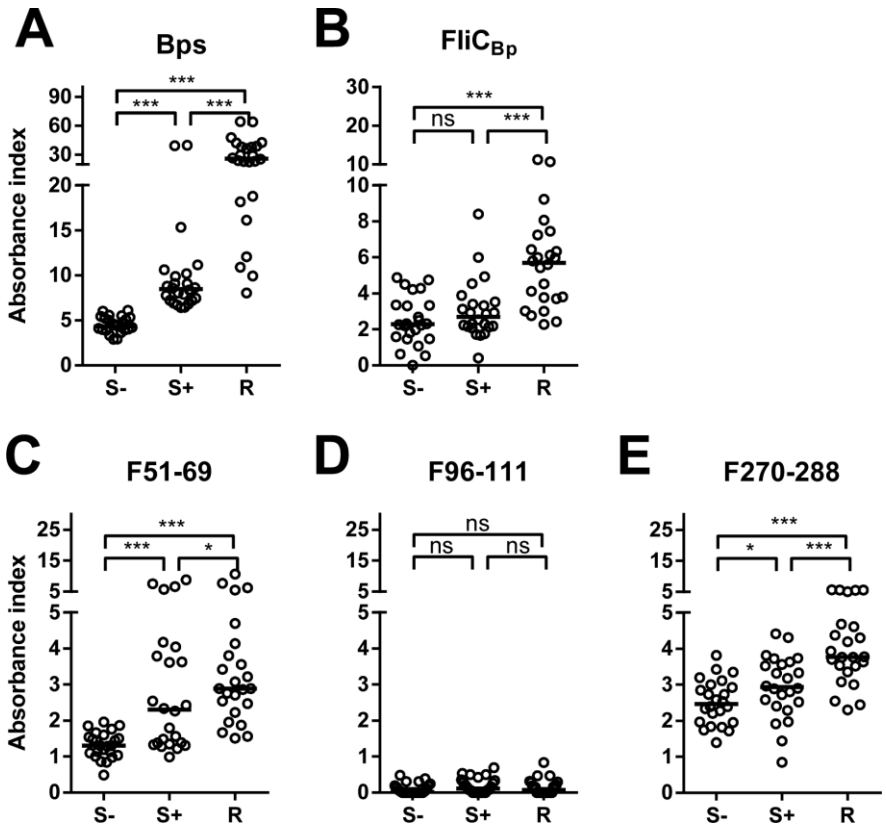


Figure S2: Distribution of human antibody against *B. pseudomallei* related proteins and peptides among seronegative (S- ; n = 24), seropositive (S+ ; n = 24) and melioidosis recovered individuals (R ; n = 24) detected by Indirect ELISA. A) Crude *B. pseudomallei* antigens **B)** recombinant $FliC_{Bp}$ **C)** synthesized epitope F51-69 **D)** F96-111 **E)** F270-288. The samples were coated into 96-well polystyrene plate and then probed with diluted plasma samples of healthy and recovered individuals and quantified by indirect ELISA. The results were represented by Absorbance index ($O.D._{test}$ – $O.D._{uncoated}$ / $O.D._{uncoated}$). Experiments were performed in duplicate and results represent the mean of the Absorbance index ± SE. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns = not significant compared between plasma sample groups using one-tailed Mann-Whitney U test.



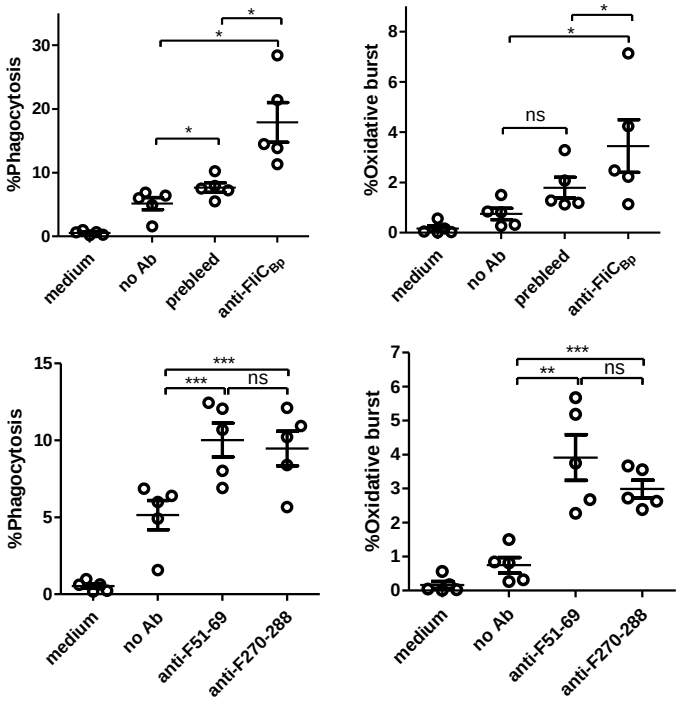


Figure S3: Rabbit anti-FliC$_{Bp}$ predicted peptide antibodies enhance phagocytosis and oxidative burst against *B. pseudomallei* in purified human PMNs (N = 5). Rabbit anti-FliC$_{Bp}$, prebleed antisera at dilution 1:50 or anti-predicted peptide antibodies at 40 μg/ml were used for opsonization. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns, not significant compared between with or without antibody groups using one-tailed paired t test.

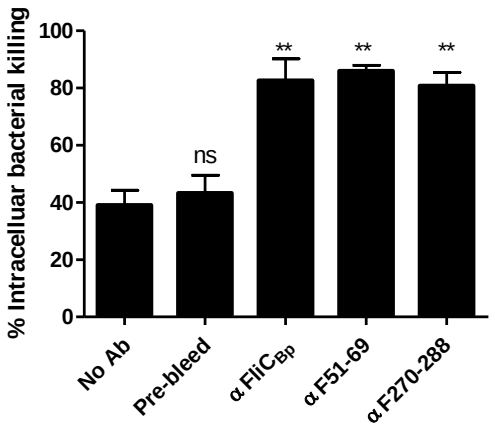


Figure S4: Rabbit anti-FliC$_{Bp}$ protein, anti-F51-69 and anti-270-288 enhance *B. pseudomallei* intracellular killing activity in purified human PMNs (N = 2). Paired t test was applied for testing between each antibody conditions and no antibody. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns, not significant.

Figure S5: $FliC_{Bp}$ peptides induced human IFN-γ and IL-10 production from PBMCs. PBMCs at 5 x 10^5 cells/well from 20 seropositive healthy donors were stimulated with fixed Bps (PBMCs : organism = 1:30), 3 µg/ml of phytohaemagglutinin (PHA), 10 µg/ml of $FliC_{Bp}$ protein, and 50 µg/ml of $FliC_{Bp}$ peptides (F51-69, F96-111, F270-288 and F213-231) for 48 h, level of IFN-γ and IL-10 were quantified by ELISA. Vertical lines are median and dash lines are representing for a limit of detection. The comparisons between stimuli were performed by Friedman test and post tested by Dunn's Multiple Comparison Test. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns, not significant.

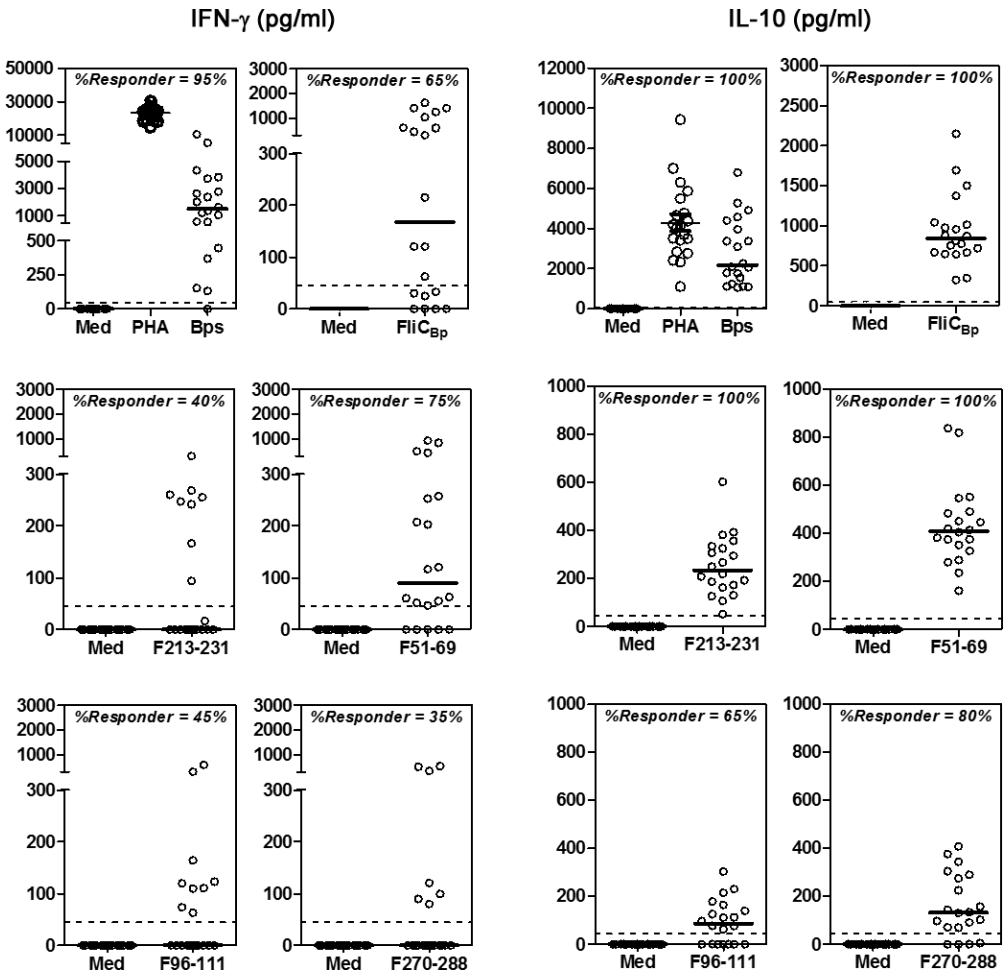
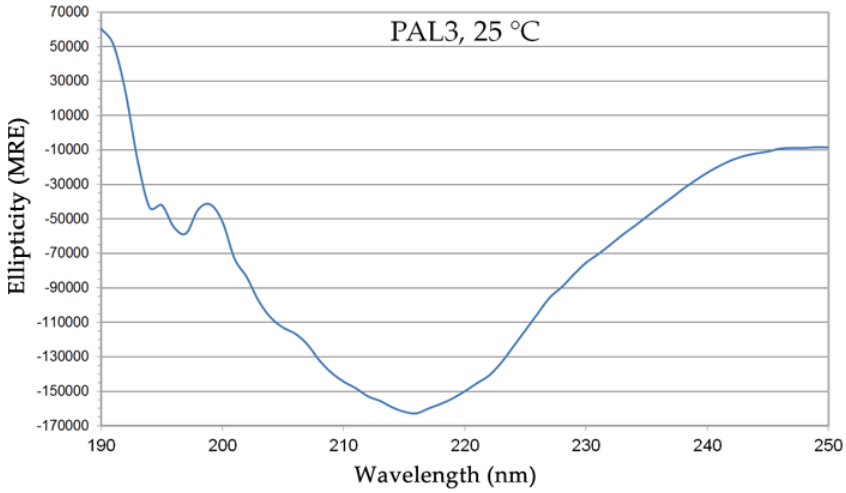


Figure S6: CD measurements of peptides PAL3 and PAL3b. Circular Dicroism spectra acquired for **A**) peptide PAL3, 25mM aqueous phosphate buffer, 0.1 mg/mL concentration (25°C, 0.1 cm quartz cell). **B**) PAL3B at different temperatures (5°C-95°C) and PB 0.2 mM, NaF 1.5 mM, pH 6.24. The CD spectra were plotted as molar ellipticity (degree *cm2/dmol) versus wavelength (nm). **C**) The table reports the percentage of helix composition for peptide PAL3B at nearly physiological temperature (40°C).

A



C

CD 40°C	PAL3B
α Helix221nm %	47.916
α Helix209nm %	55.000
α Helixmean %	51.458
R-factor (MRE221/MRE209)	0.871

B

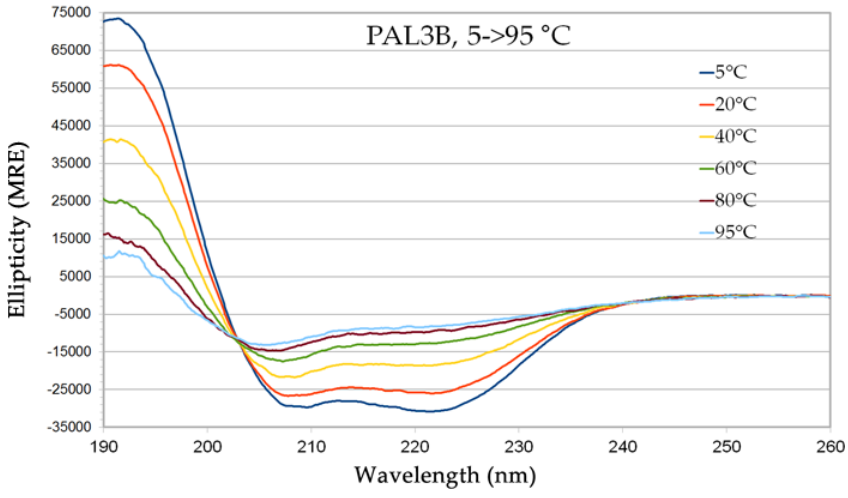


Figure S7: Study of the prophylactic effect of anti-PAL3 and anti PAL3B antibodies ip administered against an acute in infection with Bps576 in C57BL6 mice. Passive immunization with anti-PAL3 antibodies (n=8) and anti PAL3B antibodes (n=8) was administered intraperitoneal route 6 hours before challenge, 1mg/ mouse of the corresponding antibody dissolved in PBS 7.2pH (2.5 mg/ml). Administration volume of 20ml/kg. Control mice were injected PBS 7.2pH (n=6), and a control antibody (n=6). Mice infection intranasal route, 50µL/mouse.

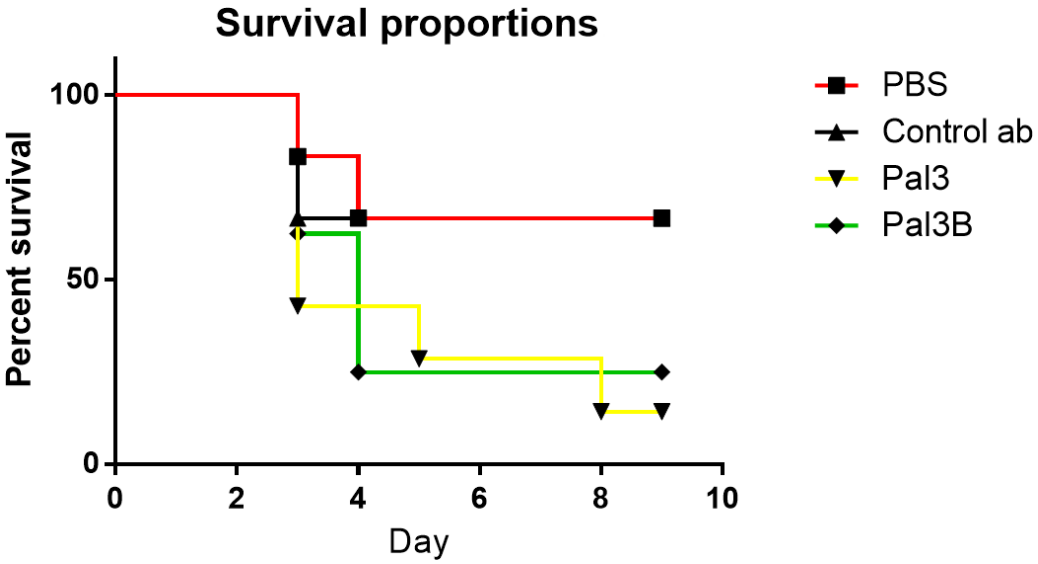


Figure S8: Sequence and predicted secondary structure of Pal_{Bp} mimic model before MD simulations. The sequence at step B represents the excised folding unit connected by Pro 45 and Gly 46, while mimic at step C features a N-ter reduction (with the addition of Ile 4) and the addition of the helix stabilizer, here simplified as "residue" X at position 28. A black line is connecting the disulphide bridge positions, Cys 34 and Cys 49. Helices are depicted red, while strands are colored blue.

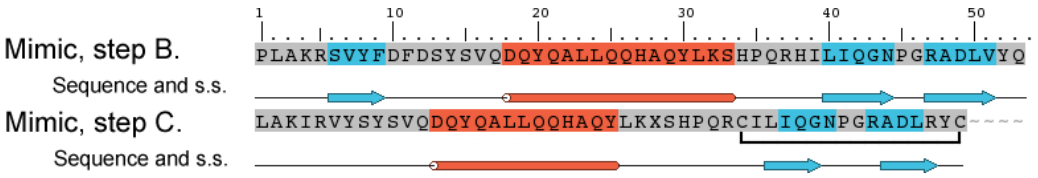


Figure S9: Time evolution of the secondary structure of Pal_{Bp} mimics across MD simulations. The stability of Pal_{Bp} mimics in solution have been assessed as a measure of the secondary structure elements throughout 100 ns of MD simulation. A) The upper panels represent two replicas of PalBp mimic, step C, deprived of the helix stabilizer fragment. The pronounced stability of the three β elements is opposed to the partial (upper replica) and total (lower replica) instability of the α helix. B) Two replicas of PalBp mimic, step C, with the helix stabilizer connected in cis orientation. The α helix is fully stable throughout the simulations, but the stability of the β strands is impaired.

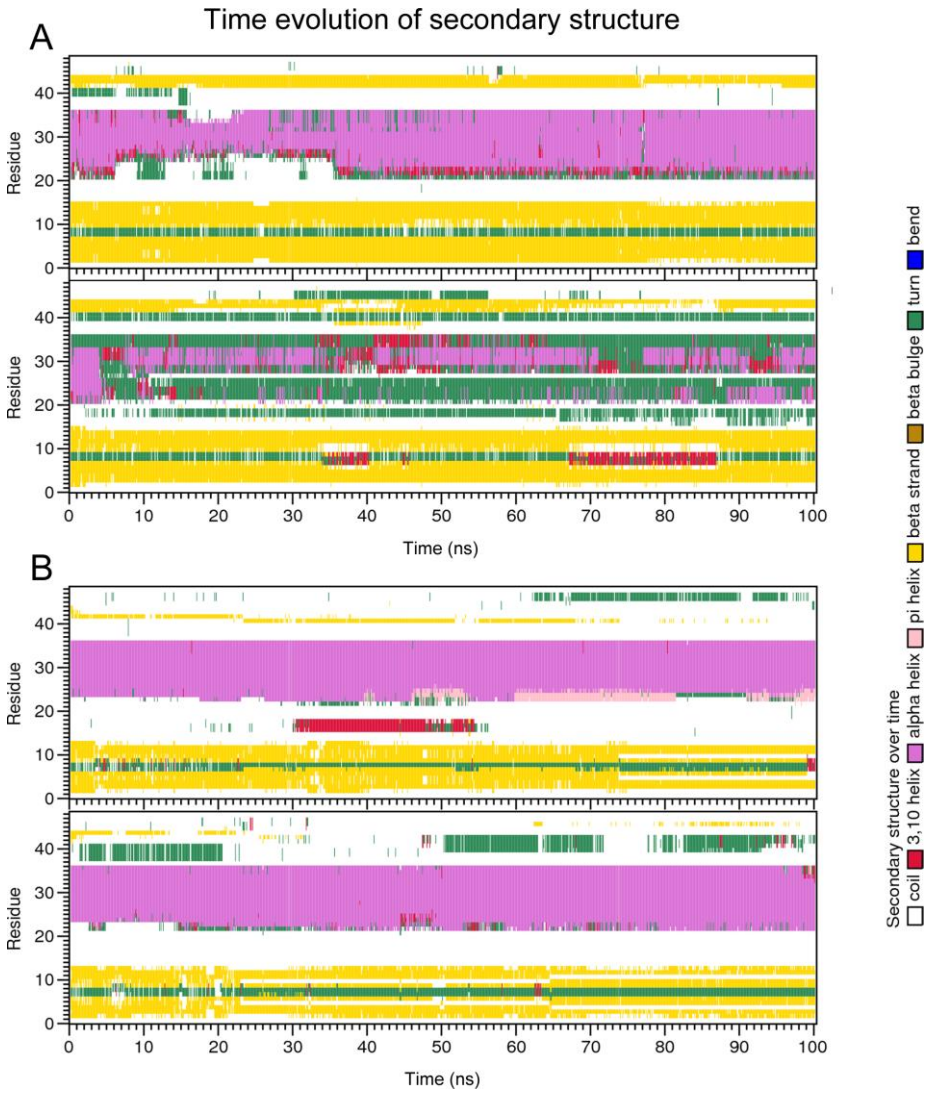


Figure S10: Overall occurrence propensity of each amino acid into BLUESTRIPE results. The percentage of occurrence of each amino acid in the results of BLUESTRIPE analysis is displayed together with the occurrence of each amino acid on the surface of the protein dataset. As it appears clear, the propensities are very similar comparing monomer and complex analyses, and both are consistent with the surface propensity control for the majority of residues. This result indicate that the analysis, and possibly the *blue stripe* formation, may not be dependent on the chemical peculiarity of the amino acids involved. Exceptions include Gly and Pro (higher propensity). The difference is most likely an artefact, since the energy matrix approximation relies on a local definition of energetic coupling. Glycine is lacking a full side chain, and Proline is not a strong interactor, making the two amino acids constitutively uncoupled and more easily selected by the analysis. The large Standard Deviations depicted by the error bars is due to the heterogeneous nature of the sample, purposely selected to account for different types of proteins and transient interactions.

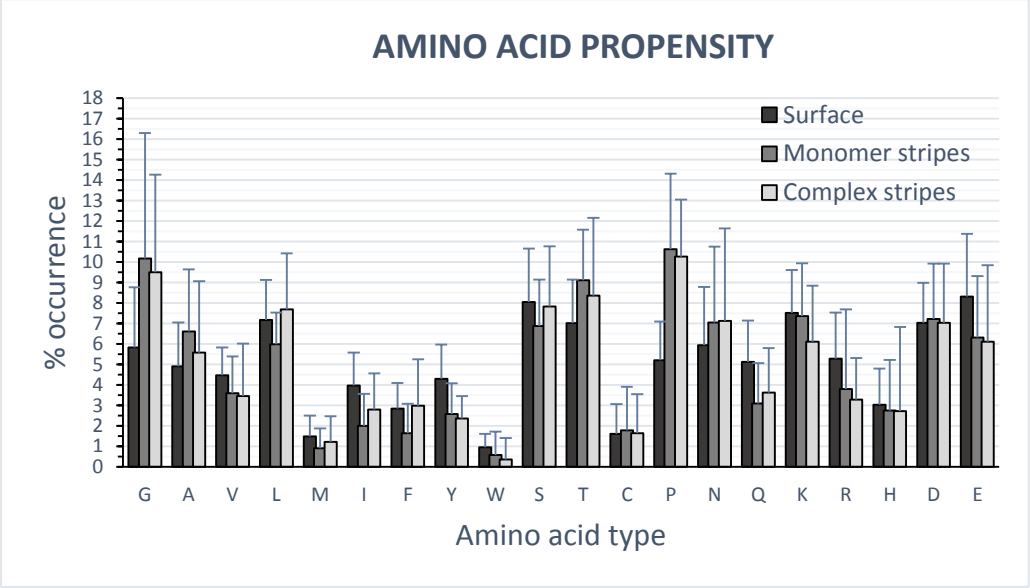


Table SI: List of computationally and experimentally determined putative epitopes of protein BPSLI445. B) MLCE predictions are subdivided into individual epitopes (E1-E3) spotted on each MD replica (MD1-MD3). The sequences identified across the different replicas are largely overlapping. EDP predictions highlighted two putative epitopes (EDP1-EDP2), while one sequence was identified by experiental epitope mapping (EXP-1). The overall consensus epitopes are represented by sequences Lipo#1 – Lipo#3.

Prediction method	Name	Epitope sequence
MLCE	MD1-E1	$_{47}$IQAGL$_{51}$
MLCE	MD2-E1	$_{45}$D $_{47}$IQAG$_{50\ 54}$GGQTG$_{58}$
MLCE	MD3-E1	$_{50}$GLIIGGQT$_{58}$
MLCE	MD1-E2	$_{78}$SVG$_{80\ 83}$AGAQS$_{87}$
MLCE	MD2-E2	$_{78}$SVGLQAGAQSK$_{88}$
MLCE	MD3-E2	$_{77}$LSVG$_{80\ 83}$AG$_{84}$
MLCE	MD1-E3	$_{110}$AAGADA$_{115\ 117}$V $_{122}$MGANGAIDTTTATA$_{135}$
MLCE	MD2-E3	$_{122}$MGANGAIDT$_{130\ 132}$TATA$_{135}$
MLCE	MD3-E3	$_{110}$A $_{112}$GADASVA$_{118\ 121}$KMGANGAIDTTTATAP$_{136}$
EDP	EDP-1	$_{44}$PDVIQAGLIIGGQTGN$_{59}$
EDP	EDP-2	$_{76}$SLSVGLQAGAQSK$_{88}$
Epitope mapping	EXP-1	$_{39}$GVLVFPDVIQAGLIIGGQTGNGALR$_{63}$
Synthetic peptide	Lipo#1	$_{40}$VLVFPDVIQAGLIIGGQTGNGALRV$_{64}$
Synthetic peptide	Lipo#2	$_{76}$SLSVGLQAGAQSK$_{88}$
Synthetic peptide	Lipo#3	$_{110}$AAGADASVALVKMGANGAIDTTTATAPVE$_{137}$

PDB code	Mapped epitope	Epitope data source	test
1CB0	218-HEEAVSVDRVL-230	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
1D3B	52-RVAQLEQVYI-63	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
1EA3	0-SLLTEVETYVL-12	Cell bound MHC - Fluorescence	Association (or direct binding)
	1-LLTEVETYVL-12	Cell bound MHC - Fluorescence	Association (or direct binding)
	11-SIIPSGPLK-21	Cell bound MHC - Radioactivity	Competition (or eq. binding)
	115-LSYSAGAL-124	Cell bound MHC - Fluorescence	Association (or direct binding)
	123-ASCMGLIY-132	Lysate - Radioactivity	Association (or direct binding)
	126-MGLIYNRM-135	Cell bound MHC - Fluorescence	Association (or direct binding)
	132-RMGAVTTEV-142	Cell bound MHC - Fluorescence	Association (or direct binding)
	138-TEVAFGLV-147	Cell bound MHC - Fluorescence	Association (or direct binding)
	143-GLVCATCEQIA-155	Cell bound MHC - Fluorescence	Association (or direct binding)
	15-SGPLKAEIAQRLEDV-31	Purified MHC - Radioactivity	Association (or direct binding)
	25-RLEDVFAGK-35	Purified MHC - Radioactivity	Association (or direct binding)
	31-AGKNTDLEVLMEWLKTRPIL-52	Purified MHC - Fluorescence	Competition (or eq. binding)
	45-KTRPILSPLTK-57	Cell bound MHC - Radioactivity	Competition (or eq. binding)
	48-PILSPLTKGI-59	Cell bound MHC - Fluorescence	Association (or direct binding)
	55-KGILGFVFTLTVPSER-72	Purified MHC - Radioactivity	Association (or direct binding)
	56-GILGFVFTL-66	Cell bound MHC - Fluorescence	Association (or direct binding)
	94-AVKLYRKL-103	Cell bound MHC - Fluorescence	Association (or direct binding)
1HA0	297-PKYVKQNTLKLAT-311	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
1I7Z[a]	205-LSSPVTKSF-215	null	null
1OVA	123-LYRGGLEPI-133	Purified MHC - Radioactivity	Competition (or eq. binding)
	175-NAIVFKGL-184	Purified MHC - Radioactivity	Competition (or eq. binding)
	24-ENIFYCPI-33	Purified MHC - Radioactivity	Competition (or eq. binding)
	256-SIINFEKL-265	Cell bound MHC - Fluorescence	Association (or direct binding)
	257-IINFEKLTEWTSSNVM-274	Cell bound MHC - Fluorescence	Association (or direct binding)
	26-IFYCPIAI-35	Purified MHC - Radioactivity	Competition (or eq. binding)
	270-NVMEERKIKVYLPRM-286	Purified MHC - Radioactivity	Competition (or eq. binding) approx. KD
	272-MEERKIKVYLPRMKME-289	Cell bound MHC - Fluorescence	Dissociation
	306-SSSANLSGISSAESLKISQA-327	Purified MHC - Radioactivity	Competition (or eq. binding) approx. KD
	322-ISQAVHAAHAEINEAGR-340	Cell bound MHC - Fluorescence	Competition (or eq. binding)
	95-VYSFSLASRL-106	Purified MHC - Radioactivity	Competition (or eq. binding)
2GIB	32-AQFAPSASA-42	Purified MHC - Radioactivity	Competition (or eq. binding)
	35-APSASAFFGM-46	Purified MHC - Radioactivity	Competition (or eq. binding)
	37-SASAFFGMSR-48	Purified MHC - Radioactivity	Competition (or eq. binding)
	44-MSRIGMEVTPSGTWL-60	Purified MHC - Radioactivity	Competition (or eq. binding) approx. KD
	51-VTPSGTWLTY-62	Purified MHC - Radioactivity	Competition (or eq. binding)
	55-GTWLTYHGAIKLDDK-71	Purified MHC - Radioactivity	Competition (or eq. binding) approx. KD
	70-DPQFKDNVILLNKHI-86	Purified MHC - Radioactivity	Competition (or eq. binding) approx. KD
	78-ILLNKHIDA-88	Cell bound MHC - Fluorescence	Association (or direct binding)
	81-NKHIDAYKTFPPTEP-97	Purified MHC - Radioactivity	Competition (or eq. binding) approx. KD
	88-KTFPPTEPK-98	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
2JK2	18-GELIGTLNAAKVPAD-34	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
2VB1	10-AMKRHGLDNYRGYSL-26	Purified MHC - Fluorescence	Dissociation
	113-RCKGTDVQAWIRGCRL-130	Purified MHC - Radioactivity	Competition (or eq. binding)
	19-YRGYSLGNWVCAAKFE-36	Purified MHC - Fluorescence	Dissociation
	45-NTDGSTDYGILQINSR-62	Cell bound MHC - Fluorescence	Association (or direct binding)
	9-AAMKRHGLDNYRGY-24	Purified MHC - Fluorescence	Competition (or eq. binding)

PDB code	Mapped epitope	Epitope data source	test
2WA0	134-GVYDGREHTV-145	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
3BZH	43-ILGPPGSVY-53	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
3FEY	135-GAVDPLLAL-145	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
	237-ELPIVTPAL-247	Purified MHC - Fluorescence	Association (or direct binding) approx. KD
3HLA	102-VGSDWRFLRGYHQYA-118	Purified MHC - X-ray crystall.	Structure (crystal, NMR, etc.)
	102-VGSDWRFLRGYHQYAYDG-121	Purified MHC - Fluorescence	Competition (or eq. binding)
	27-VDDTQFVRFDSDAASQRMEPR-49	Purified MHC - Fluorescence	Association (or direct binding)
	29-DTQFVRFDSDAASQRMEP-48	Purified MHC - Fluorescence	Association (or direct binding)
	59-WDGETRKVKAHSQTHRVDLGTLRGY-85	Cell bound MHC - T cell response	Competition (or eq. binding)

Table S2: List of proteins and MHC-II epitopes in the dataset: from left to right, we show for each protein the PDB code structure used, the list of IEDB mapped epitopes, the data source and the experimental means employed for mapping. Shaded rows highlight such epitopes whose structure has been deposited in Protein Data Bank (in complex with MHC molecules).

[a] This antigen protein is not featured in IEDB repository.

Contingency Table	POSITIVES	NEGATIVES
TRUE	TP	TN
FALSE	FP	FN

SENSITIVITY (RECALL)	$\frac{TP}{TP+FN}$
SPECIFICITY	$1-\frac{FP}{TN+FP}$
PRECISION (PPV)	$\frac{TP}{TP+FP}$
ACCURACY (Q2)	$\frac{TP+TN}{TP+TN+FP+FN}$

Matthews Corr. Coeff
$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$

Table S3. Formal description of the statistical parameters used in the evaluation of MLCE and BLOCKS as epitope predictors. Starting from the classification of results into four categories expressed in the contingency tables, sensitivity, specificity, precision, random precision, accuracy and MCC are calculated as measures of the predictive performance.

Table S4. Comparison of *BLUEPRINT* analysis on individual monomers vs full complexes. The number of residues identified by BLUEPRINT analysis (performed on the monomers as well as on the complexes) is reported next to the ID of each pair of interacting proteins. The cells highlighted in blue correspond to proteins forming a *blue stripe* (14/15 for monomers, 12/15 for complexes). The percentage of identity is a quantitative measure of the number of residues in common between monomer and complex analysis. The number of residues identified in both conditions is very similar, and the overall identity is considerably high, indicating a general propensity over the prediction of the same protein areas.

PDB ID			No. of residues identified		
COMPLEX	MONOMERS		MONOMER	COMPLEX	Identity (%)
1acb	1egl	2cga	63	61	36.1
1ay7	1a19	1rgh	53	49	46.9
1b6c	1d6o	1ias	90	82	57.3
1buh	1hcl	1dks	70	87	61.4
1dfj	9rsa	2bnh	152	125	64
1f51	1ixm	1srr	80	61	62.3
1gpw	1thf	1k9v	102	95	61.1
1ira	1ilr	1g0y	99	104	55.6
1k74	1mzn	1zgy	125	100	65
1kac	1nob	1f5w	75	61	49.2
1ktz	1tgk	1m9z	59	51	43.1
1kxp	1ijj	1kw2	174	151	65.6
1pxv	1x9y	1nyc	63	45	48.9
1r0r	2grk	1scn	73	63	66.7
1tmq	1jae	1b1u	109	82	64.6

11.3 SUPPLEMENTARY METHODS

The references for the supplementary methods are listed at the end of the section.

MD simulation protocol

The majority of the computational investigations performed throughout this thesis are made upon MD simulations of the structural data.

The following protocol has been used for epitope prediction purposes, and has been employed also for BLUESTRIPE analysis and for the dynamic characterization of antigens and peptides. Each replica simulation was performed using the GROMACS 4.5.1 or 4.5.4 software package[1], GROMOS 96A6 force field[2] and the SPC water model[3]. The charge state of ionizable groups was chosen to mimic physiological pH conditions. The systems were solvated in cubic boxes large enough to contain 1 to 1.2 nm of solvent around the protein solute. The systems were subsequently energy minimized with a steepest descent method for 1000 steps, followed by another 1000 steps of steepest descent restraining the backbone and 1000 steps of conjugated gradient with no restraints. After minimization, MD simulations were carried out using the following set-up: the calculation of electrostatic forces utilized the PME implementation of the Ewald summation method. The constraining of all bond lengths was performed with the LINCS algorithm[4]. A dielectric permittivity, c=1, and a time step of 2 fs were used. All atoms were given an initial velocity obtained from a Maxwellian distribution at the temperature of 300 K. The density of the system was adjusted by weak coupling to a bath of constant pressure (P_0=1 bar, coupling time τ_P=1 ps)[5]. The temperature was maintained close to the intended values using Berendsen thermostat with a coupling constant of 0.2 ps[5]. The proteins and the rest of the systems were coupled separately to the temperature bath. The runs were carried out using NPT conditions for a variable simulation time, usually 5ns for BLUESTRIPE analysis and 50 ns for antigen characterization and epitope prediction. The first portion of each trajectory (in the aforementioned cases 1 and 10 ns respectively) was not used in the subsequent analyses in order to minimize convergence artifacts.

RMSD and RMSF measures were used to assess the convergence of the simulation as well as the overall protein flexibility.

Essential Dynamics Analysis

The main structural fluctuations of the protein were analyzed through Principal Component or Essential Dynamics (ED) Analysis of the covariance matrix of atomic positions obtained from the trajectories. This allows the relevant motions to be separated from the background fluctuation noise[6]. The principal eigenvectors are in general associated with the slow modes, which are responsible for protein functions. Translational and rotational degrees of freedom are eliminated and the average atomic coordinates, $x_{i,ave}$, are calculated along the MD trajectories. The essential directions of correlated motions were then calculated by diagonalizing the covariance matrix C_{ij} of atomic positions from MD. The ED analysis has been performed with GROMACS 4.5I and GROMACS 4.54 software package.

Signal Propagation Analysis

The analysis of signal propagation is derived from the Distance Fluctuations (DF) analysis, which can be used along with various related quantities to characterize the salient internal dynamics properties of a protein undergoing structural fluctuations {{14;20;44}}. Signal transduction events in proteins are directly related to the fluctuation dynamics of atoms, defining the communication propensities (CP) of a pair of residues as a function of the fluctuations of inter residue distances {{74}}. Residues whose Cα-Cα distance fluctuates with a relatively small intensity during the trajectory are supposed to communicate more efficiently than residues whose distance fluctuations are large. The Communication Propensity (CP) of any two residues is defined as the mean-square fluctuation of the inter residue distance (d_{ij}) defined as distance between the Cα atoms of residue *i* and residue *j*.

$$CP = \left\langle (d_{ij} - d_{ij,ave})^2 \right\rangle \qquad (1)$$

By projecting these quantities on the 3D structures of the proten, it is possible to identify the regions affected in their motion by the mechanic propagation of the fluctuation signal.

Computational analysis of S2 order parameter

The S2 order parameter is a generalized descriptor of the spatial aspects of reorientational motion of N-H_N vectors of heteronuclear NMR relaxation of ^{15}N-labeled proteins. It represents a rich source of dynamic and thermodynamic information, ranging in the time scales of nanoseconds and sub-nanoseconds. An analytical estimation of NMR S2 order parameter may be calculated from the backbone of high resolution protein structures, in the form introduced by Brüschweiler *et al.* {{75}} (2) and adapted to be evaluated on a bundle of time frames collected from a protein simulation trajectory.

$$S_i^2 = tanh(0.8 \sum_k (exp(-r_{i-1,k}^o / 1\mathring{A})) + 0.8(exp(-r_{i,k}^H / 1\mathring{A}))) + b \qquad (2)$$

$r^0_{i\text{-}1k}$ is the distance between the carbonyl oxygen of amino acid $i - 1$ to heavy atom k and $r^H_{i,k}$ is the distance between the amide proton H^N and heavy atom k. The parameter b is set to –0.1, which takes into account that order parameters of rigid protein regions typically lie around 0.9. The sum ranges over all heavy atoms k that do not belong to amino acids i and $i - 1$.

1. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput* 4 (3), 435-447.
2. van Gunsteren WF, Daura X, Mark AE (2006) Gromos force field. *Encyclopedia of Computational Chemistry* 2: 1211–1216.
3. Berendsen HJC, Grigera JR, Straatsma PR (1987) The missing term in effective pair potentials. *J Phys Chem* 91: 6269–6271.
4. Hess B, Bekker H, Fraaije JGEM, Berendsen HJC (1997) A linear constraint solver for molecular simulations. *J Comp Chem* 18: 1463–1472.
5. Berendsen HJC, Postma JPM, van Gunsteren WF, Di Nola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81: 3684–3690.
6. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. Proteins 17: 412-425.
7. 48. Morra, G., Verkhivker, G., and Colombo, G. (2009) Modeling signal propagation mechanisms and ligand-based conformational dynamics of the Hsp90 molecular chaperone full-length dimer. *PLoS Comput. Biol. 5*, e1000323.
8. Morra, G.; *et al*. Dynamics-Based Discovery of Allosteric Inhibitors: Selection of New Ligands for the C-terminal Domain of Hsp90. *J. Chem. Theory Comput*. 2010, 6: 2978–2989.
9. Torella, R.; et al. Investigating dynamic and energetic determinants of protein nucleic acid recognition: analysis of the zinc finger zif268-DNA complexes. *BMC Struct. Biol.* 2010, 10:42
10. Morra, G.; Potestio, R.; Micheletti, C.; Colombo, G. Corresponding Functional Dynamics across the Hsp90 Chaperone Family: Insights from a Multiscale Analysis of MD Simulations. *PLOS Comp. Biol.* 2012, 8(3):e1002433.
11. Zhang, F; Brüschweiler, R. Contact model for the prediction of NMR N-H order parameters in globular proteins. *J. Am. Chem. Soc.* 2002, 124(43):12654-5.