Molecular Modeling of Bacterial Nanomachineries

THÈSE N^{O} 5404 (2012)

PRÉSENTÉE LE 22 JUIN 2012
À LA FACULTÉ DES SCIENCES DE LA VIE
UNITÉ DU PROF. DAL PERARO
PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Matteo Thomas DEGIACOMI

acceptée sur proposition du jury:

Prof. H. Lashuel, président du jury
Prof. M. Dal Peraro, directeur de thèse
Prof. S. Bernèche, rapporteur
Prof. P. De Los Rios, rapporteur
Prof. O. Michielin, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2012

Try to learn something about everything, and everything about something.
— Thomas Henry Huxley

Acknowledgements

I wish to thank Prof. Matteo Dal Peraro for offering me to work on this project, trusting me, and allowing me freedom to explore, some of my apparently weird ideas.

I owe a lot to all my lab mates for both their scientific support and advices, and the warm, friendly atmosphere they contributed to create in LBM. Therefore, I heartly thank Dr. Marco Stenta for his enthusiasm and willingness to spread his extraordinary knowlege to all of us, Dr. Davide Alemani and Enrico Spiga for helping the computer scientist I was to grasp the physicist's view on molecular modeling, and Christophe Bovigny, for devoting so much time and efforts to the smooth functioning of our lab's hardware and software. Finally, a special mention goes to Thomas Lemmin, who shared all my joys and struggles during this work and, after spending four years side by side, ended up knowing everything about it as well.

It has been an honor to collaborate with Prof. Gisou Van der Goot and Ioan Iacovache (now doctor!). I'm thankful to them in number of ways but, if I had to pick just one, I'd simply say thank you for believing in me, and in what molecular simulation has to offer to biology.

When I first joined LBM, we just had two offices and number of boxes containing brand new computers to be installed. If it hadn't been for Sebastien Ferrara and Nicholas Michel, members of the Life Sciences IT crew, I'd still be under a desk, tangled in some cables, wondering how to recompile that Linux kernel.

I'm grateful to my parents Ulderica and Giovanni, and my grandparents Berny, Eris, Virgilio and Ermanno. They played a central role in my education, and life in general, and always supported me in any possible way.

Finally I've been blessed to have good friends supporting me during hard times, and enduring hearing me talking about this work way too much. I wish I could mention all of them here, but I'd need way, way more space. So, well, I'll thank you personally, guys!

Lausanne, 21 may 2012 M. T. D

Abstract

Proteins have the ability to assemble in multimeric states to perform their specific biological function. Unfortunately, characterizing experimentally these structures at atomistic resolution is usually difficult. For this reason, *in silico* methodologies aiming at predicting how multiple protein copies arrange to form a multimeric complex would be desirable.

We present Parallel Optimization Workbench (POW), a swarm intelligence based optimization framework able to deal, in principle, with any optimization problem. We show that POW can be applied to biologically relevant problems such as prediction of protein assemblies and the parameterization of a Coarse-Grained force field for proteins. By combining POW optimizations, Molecular Dynamics simulations, Poisson-Boltzmann calculations and a variety of experiments, we subsequently study two bacterial nanomachieries: *Aeromonas hydrophila*'s pore-forming toxin aerolysin, and *Yersinia enterocolitica* injectisome. These structures are challenging both for their size, and for the timescales involved in their functioning.

Aerolysin is a pore-forming toxin secreted as an hydrophilic monomer. By means of large conformational changes, the protein heptamerizes on the target cell's surface, and finally inserts a β-barrel into its lipid bilayer, causing cell death. The main hurdle in the study of this structure is the complexity of the mode of action, which spans timescales currently unreachable by classical molecular dynamics. We show that aerolysin C-terminal region has the dual role of preventing premature oligomerization and helping the folding of tertiary structure, qualifying therefore as an intramolecular chaperone. We study the transmembrane β-barrel properties and compare them with those of the homologous protein α-hemolysin. We show that aerolysin's barrel is more rigid than α-hemolysin's, and should be anion selective. We present models for aerolysin heptamer both in prepore and, for the first time, in membrane-inserted conformation. Our results are validated experimentally, and are consistent with known biochemical and structural data.

The injectisome is an example of a type III secretion system. Its most striking feature is probably its size: hundreds of proteins assemble in a unique structure spanning the Gram-negative bacterial double membrane, and protruding outside the cell as a needle for tenth of nanometers. Obtaining an atomistic representation of this massive structure, and therefore some insights about its mode of action, is one of the greatest challenges. We show that the final length of injectisome's needle is determined by the secondary structure content of a ruler protein located inside its cavity during assembly. Using POW, we also produce the first model for *Yersinia* injectisome's basal body, highlighting the flexibility of this region in adapting between the inner and outer membranes.

As a whole, this work demonstrates that a synergy of dry and wet experiments can provide precious insights into macromolecular structure and function.

keywords: macromolecular assembly, Particle Swarm Optimization, optimization, molecular dynamics, biomolecular modeling, aerolysin, pore-forming toxin, injectisome

Résumé

Les protéines ont la capacité à s'assembler de façon multimérique pour effectuer des taches biologiques spécifiques. Malhereusement, leur charactérisation à résolution atomique reste atypique. C'est pourquoi, des méthodologies *in silico* ayant pour objectif la prédiction d'arrangement multimérique de protéines apporteraient un appui non négligeable à la partie expérimentale.

De ce fait, un nouveau environement basé sur du *swarm intelligence*, Parallel Optimization Workbench (POW), à été développé. Celui-ci est capable de résoudre ce type de problème, tout aussi bien que n'importe quel autre problème d'optimisation. Dans ce travail nous démontrons que POW résoud de façon optimale des structures biologiques, tout aussi bien que des questions concernant la paramétrisation de champs de force *Coarse-Grain* pour des protéines. En combinant POW avec des simulations de dynamique moléculaire, des calculs Poisson-Boltzmann, et diverses d'expériences *in vitro*, on a étudie deux nanomachines bactériennes : la toxine aérolysine, et l'injectisome de *Yersinia enterocolitica*. Ces deux systèmes réprésentent un défi à la fois par leur dimention et par le temps nécessaire à leur fonctionnement.

Aérolysine est une toxine sécrètée par *Aeromonas hydrophilprobla* sous une forme monomérique hydrophile. Par le biais de changements conformationels majeurs, la protéine heptamérise sur la surface de la cellule cible, et insère un β-barrel à travers la membrane, ce qui entraine la mort de la cellule. La plus grande difficulté dans l'étude de cette structure est la complexité de son méchanisme d'action, qui a lieu dans des temps inaccessibles par la classique dynamique moléculaire. On montre que la région C-terminale d'aérolysine joue un double rôle : prévenir une oligomérisation prématurée, et faciliter le repliement dans la structure tertiaire, ce qui le désigne comme un chaperon intramoléculaire. Une séconde étude s'est portée sur les propriétés du β-barrel transmembranaire, et les comparaisons avec celles de sa protéine homologue α-hémolysine. On démotre que le barrel d'aérolysine est plus rigide, et qu'il devrait être sélectif aux anions. On présente deux models de aérolysine en conformation heptamérique en état de prépore et, pour la première fois, inserée dans une membrane. Nos résultats sont validés expérimentalement, et sont cohérents avec des donnés structurales et biochimiques connues.

L'injectisome est un exemple de *type III secretion system*. La charactéristique la plus remarquable est probablement sa dimension : des centaines de protéines s'assemblent dans une structure unique traversant la double membrane Gram-négative, et s'étendant dans le milieu extracellulaire sous un forme d'aguille longue des dizaines de nanomètres. Un des plus grands défis consiste dans l'obtention d'une réprésentation atomique de cette structure et, ainsi, l'aquisition d'informations sur son fonctionnement. On prouve que la longeur finale de l'aguille extracellulaire est déterminée par la quantité de structure secondaire d'une protéine petite règle localisée à l'intérieur de la cavité lors de son assemblage. En utilisant POW, le premier model du corps basal de l'injectisome de *Yersinia* est modélisé.

En conclusion, ce travail démontre que la synergie entre simulation et expériment permet d'obtenir de précieuses informations concernant la structure et la fonction de macromolécules.

mots clés : assemblage macromoléculaire, Particle Swarm Optimisation, optimisation, dynamique moléculaire, modeling biomoleculaire, aérolysine, toxine, injectisome

Riassunto

Le proteine hanno la capacità di assemblarsi in stati multimerici per assolvere una specifica funzione biologica. Sfortunatamente, caratterizzare sperimentalmente queste strutture ad una risoluzione atomistica è spesso arduo. Per questa ragione, l'esistenza di metodologie *in silico* aventi come obiettivo di predire come diverse proteine possano aggregarsi in un complesso sarebbe di grande beneficio.

In questo lavoro presentiamo Parallel Optimization Workbench (POW), una piattaforma basata su *swarm intelligence* capace, in principio, di risolvere un qualsiasi problema di ottimizzazione. Mostriamo come POW possa essere applicato a problemi di rilevanza biologica come la predizione di assemblaggi di proteine e la parametrizzazione di campi di forza *coarse-grained* per proteine. In seguito, combinando POW con simulazioni di dinamica molecolare, calcoli Poisson-Boltzmann ed una varietà di esperimenti, studiamo due nanomacchine batteriche: aerolisina, una tossina secreta da *Aeromonas hydrophila*, ed il *type III secretion system* di *Yersinia enterocolitica*. Queste strutture rappresentano una sfida sia per le loro dimensioni che per i tempi necessari al loro funzionamento

Aerolisina è una tossina secreta sotto forma di monomero idrofilico. Attraverso grandi cambiamenti conformazionali, la proteina ettamerizza sulla superficie della cellula obiettivo e inserisce un β*-barrel* nella sua membrana lipidica, causandone la morte. Il maggiore ostacolo nello studio di questa struttura è la complessità del suo meccanismo di azione, che ha luogo in tempi inaccessibili alla dinamica molecolare classica. Mostriamo che la regione C-Terminale di aerolysin ha un doppio ruolo: prevenire una prematura oligomerizzazione, e aiutare il ripiegamento della struttura terziaria, agendo in tal modo come chaperone intramolecolare. Studiamo le proprietà del β*-barrel* transmembranale, e le compariamo con quelle della proteina omologa α-emolisina. Mostriamo che il barrel di aerolisina è più rigido di quello di α-emolisina, e che questo dovrebbe essere selettivo agli anioni. Presentiamo due modelli di aerolisina in conformazione ettamerica sia in stato di preporo che, per la prima volta, inserita in membrana. I nostri risultati sono validati sperimentalmente, e sono coerenti con dati strutturali e biochimici noti.

L'Injectisome è un esempio di *type III secretion system*. La sua caratteristica più rilevante è probabilmente la sua dimensione: centinaia di proteine si assemblano in un'unica struttura in grado di attraversare la doppia membrana Gram-negativa, ed estendersi all'esterno della cellula sotto forma di ago per decine di nanometri. Una delle maggiori sfide consiste nell'ottenere una rappresentazione atomica di questa enorme struttura, e dunque preziose informazioni sul suo funzionamento. Mostriamo che la lunghezza finale dell'ago extracellulare è determinata dalla quantità di struttura secondaria di una proteina righello localizzata all'interno della cavità durante il suo assemblaggio. Usando POW, produciamo in seguito il primo modello del corpo basale dell'Injectisome di *Yersinia*.

Nel complesso, questo lavoro dimostra come una sinergia tra simulazione ed esperimento permetta di ottenere preziose informazioni sulla struttura e la funzione di macromolecole.

parole chiave: assemblaggio macromolecolare, PSO, ottimizzazione, dinamica molecolare, modeling biomolecolare, aerolisina, tossina, injectisome

Contents

Acknowledgements **v**

Abstract (English/Français/Italiano) **vii**

List of figures **xv**

List of acronyms **xxiii**

1 Introduction **1**
- 1.1 Biology *in silico* . . . 1
- 1.2 The Role of Minimization . . . 2
- 1.3 Objectives . . . 3

2 Methods **5**
- 2.1 Molecular Simulations . . . 5
- 2.2 Binding Free Energy Calculations . . . 11

3 Parallel Optimization Workbench (POW) **15**
- 3.1 The engine: Particle Swarm Optimization (PSO) . . . 15
- 3.2 Architecture . . . 19
- 3.3 Function Minimization . . . 21
- 3.4 Macromolecular assembly . . . 24
  - 3.4.1 Search Space Definition and Data Manipulation . . . 25
  - 3.4.2 Fitness Function . . . 29
  - 3.4.3 Postprocessing . . . 30
  - 3.4.4 Benchmarks . . . 30
- 3.5 Parameters fitting for Coarse-Grained force fields in molecular simulations . . 37
  - 3.5.1 Search Space Definition and Data Manipulation . . . 38
  - 3.5.2 Fitness Function . . . 38
- 3.6 Conclusion . . . 39

4 Pore-forming toxin aerolysin from *Aeromonas hydrophila* **41**
- 4.1 Biological Background . . . 41
- 4.2 Role of the C-Terminal Peptide . . . 46
  - 4.2.1 Covalent bonding is not required for binding of the CTP to aerolysin . . 46

4.2.2 Identification of key residues for CTP-aerolysin binding 48
4.2.3 *In silico* mutation F457G affects the stability of both CTP and Domain 4 49
4.2.4 *In silico* removal of the CTP affect aerolysin's folding 50
4.2.5 Mutation or removal of the CTP prevents folding of aerolysin *in vitro* and *in vivo* 52
4.2.6 The CTP promotes folding and prevents heptamerization 55
4.2.7 Discussion 56
4.2.8 Material and Methods 58
4.3 Macromolecular assembly of Y221G soluble pore 60
4.3.1 Mutation Y221G prevents Domain 4 unfolding and loop extraction . . . 61
4.3.2 Assembly of aerolysin Y221G heptamer 63
4.3.3 Materials and Methods 65
4.4 Modeling of the wild-type heptamer 69
4.4.1 Loop extraction affects domains arrangement 70
4.4.2 Model of the aerolysin WT heptamer 72
4.4.3 Materials and Methods 76
4.5 Properties of transmembrane β-barrel 78
4.5.1 Aerolysin's pore model 79
4.5.2 Pores electrostatic profile 80
4.5.3 Ionic current with imposed voltage 83
4.5.4 Pores Flexibility 83
4.5.5 Discussion 87
4.5.6 Material and Methods 88
4.6 Conclusion 90

5 *Yersinia enterocolitica* Injectisome 93
5.1 Biological Background 93
5.2 The mechanism behind YscP molecular ruler 95
5.2.1 YscP helical content modulates needle length 95
5.2.2 Needle can accomodate folded YscP and YscF 98
5.2.3 Materials and Methods 99
5.3 Modeling of the MS-ring at the basal body 100
5.3.1 The periplasmic domain of YscD is elastic 100
5.3.2 Assembly of the MS-ring 101
5.3.3 Material and Methods 101

A Appendix 105
A.1 POW User Guide 105
A.2 DockSymmCircle module 110
A.3 POW prediction using various constraints 117
A.4 Aerolysin disorder prediction 120
A.5 YscP secondary structure prediction 121

Bibliography **135**

Curriculum Vitae **137**

List of Figures

2.1 *A ligand is depicted in red, a receptor in blue. In order to compute the binding free energy between the two binding partners, the following thermodynamics cycle can be adopted.* ΔG_{solv} *represents the energy difference between solvated and desolvated state, and* ΔE_{gas} *the binding free energy of the binding partners* in vacuo 11

3.1 *Schematic of the Optimizer architecture. Every box represents a class. Classes highlighted in blue are common to any optimization problem, and can be considered as a black box for by the user. Classes in the yellow area change depending on the problem being solved. We call a module a file containing a definition for these classes aimed at solving a specific problem. Input is provided as a text file containing keywords with associated values.* 20

3.2 *Two dimensional sine (left) and Rastrigin (right) functions.* 21

3.3 *Result on multidimensional Sine (left row) and Rastrigin (right row) using PSO (top row) and PSO-KaR (bottom row). Every result is an average of 10 independent optimization runs. The "Kick and Reseed" procedure improves fitness convergence when hard fitness functions are optimized, whereas it has no effect when optimizing easy functions.* 23

3.4 *Result on multidimensional Rastrigin function using different velocity thresholds for PSO-KaR optimization. The final fitness achieved with different threshold values is shown for runs on the 100D Rastrigin (left) and 50D Rastrigin (right). Error bars indicate the standard deviation over 10 runs. In both cases, a too high thresholds introduces too much noise, and a too low one does not perturb the swarm sufficiently. Best results are obtained, in both cases, with a velocity threshold equal to 0.01.* 24

3.5 *POW pipeline as implemented in the* `DockSymmCircle` *module.* 26

3.6 *Best symmetrical rigid assembly predictions (in yellow) superposed to known X-ray crystal structures (in blue). In panel A, red spheres highlight the position of residues displaying a high cross-linking efficiency.* 36

3.7 *In a coarse-grained representation, atoms are grouped in pseudo-atoms also called "beads".* 37

4.1 *PFT mode of action. The protein is secreted as as soluble monomer which, upon binding to specific receptors on the target cell, gets activated by cleavage of a specific region. Activated toxins multimerize according to a circular symmetry and insert into the lipid bilayer.* 41

4.2 *The only Pore-Forming Toxins having been crystallized in both monomeric and membrane-inserted conformation are nowadays the β-PFTs α-hemolysin from* Staphilococcus aureus *(top left) and Cytolysin A from* Salmonella enterica *(top right), and the α-PFT Cytolysin from* Vibrio choleare *(bottom).* 42

4.3 *Aerolysin monomer. Docking to GPI anchor owns to Domains 1 (cyan) and 2 (red). Residues having a role in binding are depicted in licorice (W45, I47, M57, Y61, K66 on Domain 1 and Y162, W324, H332 on Domain2). Domain 3's (green) loop region (tan) is responsible of membrane insertion. Domain 4 (yellow) is connected to the C-Terminal Peptide (CTP, in blue) via a loop (dashed blue line) being too flexible to be crystallized. CTP cleavage leads to aerolysin activation.* 43

4.4 *On top, loop region switching from water-soluble to membrane inserted conformation. Residues realign in order to form a β-hairpin having an alternating hydrophobic-hydrophilic pattern. On bottom images, model of the heptameric β-barrel (images from [57]).* 44

4.5 *WT proaerolysin harboring a six-histidine tag at the C-terminus was produced. Both proaerolysin and activated aerolysin (proaerolysin processed with trypsin) were incubated with Nickel beads. Both remained attached to the beads and could be eluted with imidazole. Aerolysin only could be released from the beads with urea.* 47

4.6 *Equilibrated model of IAEDANS tag. The molecule perturbs the interactions within CTP and Domain 4, helping CTP detachment as soon as proteolytic cleavage takes place.* 48

4.7 *MM/PBSA alanine scanning on CTP residues. Hydrophobic residues L441, F457 and L462 are the main responsible of CTP binding to Domain 4.* 49

4.8 *In MD, WT CTP remains steadily connected to Domain 4 (top left). Conversely, F457G CTP quickly loses secondary structures, and starts disconnecting from Domain 4 (top right). CTP unfolding in F457G simulations also affects Domain 4 fold. (bottom) secondary structure evolution of Domain 4 and CTP in F457G and WT simulations.* 50

4.9 *Effects of CTP removal. Without CTP, Domain 4 RMSF increases in Domain 4(top), and unfolding takes place along simulation (bottom).* 51

4.10 *Cumulated results of eight disorder prediction algorithms. The segment 268-282 is consistently predicted as intrinsically disordered. This strand unfold both in F457G mutation and upon CTP removal.* 52

4.11 *The amount of toxin present in cell extracts as well as in the periplasmic fraction were quantified for 3 independent experiments (n = 3). Error bars represent standard deviations. Proaerolysins F457A/G were essentially undetectable in the periplasmic fraction, and proaerolysin S453P was barely detectable.* 53

4.12 *(A) WT and mutant proaerolysins were processed with soluble trypsin, added to erythrocytes and the transmitted light at 600 nm of the sample was followed at room temperature as a function of time. Plots represent the percentage of transmitted light as a function of time. This is a representative experiment out of 4 independent experiments. (B) WT and mutant proaerolysins were processed with soluble trypsin, then subjected to a serial dilution (1/2) in a 96 well plate and incubated with erythrocytes. The number of wells lysed after 60 min at room temperature was determined. Error bars represent standard deviations (n = 3). (C) WT and mutant proaerolysins were unfolded in 4 M urea, then diluted in a urea free medium. After 10 min, samples were treated with trypsin for 10 min. Erythrocytes were added and the transmitted light at 600 nm of the samples was followed at room temperature as a function of time. This is a representative experiment out of 5 independent experiments.* 54

4.13 *(A) Circular dichroism spectra were acquired for proaerolysin before (blue) and after processing with trypsin (red, labeled aerolysin-CTP). Proteins were unfolded in 4 M urea and refolded by 10 fold dilution in a urea free buffer. The spectra of Refolded proaerolysin (yellow) and Refolded aerolysin (which has lost its peptide by unfolding) (green) were then acquired. Secondary structure could be recovered only by proaerolysin (B) Proaerolysin was subjected to proteolysis (red) or not (blue), unfolded in between 0 M and 6.3 M urea and allowed to refold in urea free buffer before the hemolytic activity was assessed. Proaerolysin was processed with trypsin after refolding. The results are the mean of 3 independent experiments. Error bars represent the standard deviation. Only proaerolysin was active after refolding.* 55

4.14 *Heptameric aerolysin Y221G cryo-EM map at 13.5Å resolution [137]* 60

4.15 *Comparison within WT and Y221G secondary structure during MD simulation. When CTP is removed, in WT unfolding propagates from the top of Domain 4 towards the loop region. In Y221G, unfolding is blocked in correspondence of L277, which anchors its strand to the rest of Domain 4. In the inset, in mutation Y221G L277 rearranges, pointing inside the neighboring hydrophobic pocket.* 61

4.16 *Comparison of Ramachandran plot of L277 and S276 residues in WT and Y221G simulation. In Y221G, S276 maintains an helical conformation along the whole trajectory due to L277 rearrangement.* 62

4.17 *First and second principal component of WT and Y221G MD simulations. Proteins explore different regions of the conformational space.* 63

4.18 *Projection of WT and Y221G simulations in the same eigenspace. The two proteins explore different regions of the eigenspace.* 64

4.19 *Top and side view of Y221G heptamer atomistic model superposed to Y221G cryo-EM map (13.5 Å resolution) produced by Tsitrin et al. [137]. Yellow spheres show interfacing residues that, mutated to alanine, lead to delayed heptamerization. Green spheres show residues responsible of binding to GPI-anchors. In the inset, the loop region could reach the membrane by sliding through the hole underneath it.* 66

4.20 *(A) Classes extracted from a negative staining experiment on aerolysin harboring a disulfide bridge on its loop region. Two types of associations can be observed, prepore-pore (highlighted in blue) and pore-pore (in red). (B) 3D reconstruction of a prepore-pore dimer.* 70

4.21 *Setup for aerolysin's loop steering. Loop region (yellow), is rotated around the E236. The rotation plane is identified within the* C_α *atoms of residues E235, G250 and N261 (Van der Waals spheres)* 71

4.22 *Comparison of Domain 4 flexibility in WT aerolysin upon CTP removal and loop extraction. Both CTP and loop affect domains arrangement.* 72

4.23 *Top view (left) and side view (right) of aerolysin's modeled* β*-barrel. Dashed lines indicate the approximate position of a bilayer's polar heads. Several polar residues are exposed to the pore lumen, and an aromatic belt is favorably placed next to lipids' heads, contributing to the pore's steady hooking.* 73

4.24 *Top and side view of WT heptamer atomistic model superposed to the density map produced in van der Goot Lab.* 75

4.25 *(*Cl^- *(left, in cyan) and* Na^+ *(right, in yellow) ion positions along MD simulations of aerolysin (top) and* α*-hemolysin (bottom). In aerolysin, no cation populates the pore along the simulation, while* α *hemolysin allows the transit of both anions and cations.* 80

4.26 *Pore radius and PMF measurements for aerolysin (1, left) and* α*-hemolysin (2, right). (A) accessible surface radius along the pore, averaged on the performed MD simulation (grey area being the standard error). (B) aerolysin and* α*-hemolysin* β*-barrels. Dashed lines represent the position of lipid bilayer's phosphate groups. (C) PMF for* Ca^{2+}*,* Na^+ *and* Cl^- *along the pore axis averaged on the performed MD simulation (colored areas being the standard error; notice the difference in PMF scale).* 81

4.27 *Comparison of Radial Distribution Function of water around chlorine ions in the bulk and in the constriction of aerolyin's* β*-barrel. When crossing the constriction, ions get partially desolvated.* 82

4.28 *In the first row, (A) the density distribution along the z axis and (B) the cumulated count of transits through the constriction point with 4 different imposed voltages are shown for* Na^+ *ions. In the second row, (C) the density distribution along the z axis and (D) the cumulated count of transits through the constriction point are shown for* Cl^- *ions. Image E shows aerolysin* β*-barrel. The location of its constriction point, as well as the direction of z axis are highlighted. In image F, the I/V graph for anions and cations is shown.* 84

4.29 *Hairpins of aerolysin (top left) and α-hemolysin (top right). Fluctuation measures have been performed in correspondence of colored regions. (center) Deformation probability of aerolysin and α-hemolysin in three different points (colored regions above). Aerolysin is more stable, and fluctuation patterns are the same along the whole pore. Conversely, α-hemolysin is less stable, and the region neighboring the hairpin's turns shows broader fluctuations. (bottom) Deformation probability of aerolysin and α-hemolysin chimeric pores next to cytosolic mouth. α-hemolysin can be stabilized by mutating its bottom ring to aerolysin bottom ring, or inserting a rivet structure.* 86

5.1 *Assembly process of injectisome's needle. Proteins constituting needle's building blocks travel inside the growing structure assembling into needle's distal end, while another protein is responsible of measuring the structure's length. In* Yersinia enterocolitica *these proteins are YscF and YscP, respectively. The basal body consists of three rings: a large cytosolic ring (C-ring), a ring spanning the plasma membrane and peptidoglycan (MS-ring) and a ring traversing the outer membrane (OM-ring). Image courtesy of Cornelis lab, Basel.* 94

5.2 *Effect of amino acids substitutions in YscP on the needle length. A. Schematic representation of YscP. Secondary structure predicted by SIMPA96. Blue bars = α-helices; red bars = β-sheets; purple bars = unstructured. B. Table showing the substitutions introduced between residues 138 and 380. Glycines were also introduced into predicted α-helices. C. The histograms of length measurements. s.d., standard deviation; n, number of measured needles.* 96

5.3 *Correlation between the YscP number of residues and the length of needles present in various different wild-type strains. The needle length is plotted against the total number of residues in YscP.* 97

5.4 *Molecular modeling of YscP protein and comparison with experimental needle length. Computed lengths of YscP wild type and mutants are reported in circles (top), and are compared with the experimental measures of the needle (triangles at the bottom). Results from explicit solvent MD simulations are also reported (empty squares at the bottom). Correlation between needle length and number of prolines substitutions is shown in the lower slope (triangles). All the values were compared with the wild-type value by the t-test. The stars indicate the probability of difference to wild type.* $NS = P > 0.05$ (not significant); $*0.05 > P > 0.01$ (significant); $**0.01 > P > 0.001$ (very significant); $***P < 0.001$ (extremely significant). 98

5.5 *Maximal (solid line) and minimal (dashed line) width of YscP36-380. The average maximum (*0.95 *nm) and minimum (*0.79 *nm) values are represented by constant lines. At the bottom, a cartoon represents the secondary structure: grey boxes represent predicted helices, while the oval represents the position of a globular domain called signal 2.* 99

5.6 *Extended (A) and bent (B) conformation of YscD monomer* 100

5.7 *Top and side view of injectisome's MS ring, supposing that both YscJ (inside, in yellow) and YscD (outside, in blue) assemble into a 24-mer.* 102

A.1 *POW graphical interface for* `Function` *module* allowing the user to save, edit, validate and launch POW input files. 110
A.2 *Aerolyin disorder prediction from eight different predictors. Disordered residues are indicated with "*", ordered ones with "-"* 120
A.3 *YscP secondary structure prediction from seven different prediction algorithms. "H" letters indicate an helical region, "-" random coil. In bold, regions selected for mutations to Alanine or Proline.* 121

List of Acronyms

AA	All-Atom
CG	Coarse-Grained
CCC	Cross-Correlation Coefficient
GPI	GlycolsysPhosphatidilInositol
MD	Molecular Dynamics
MM/PBSA	Molecular Mechanics/Poisson Boltzmann/Surface Area
NMA	Normal Modes Analysis
PCA	Principal Components Analysis
PFT	Pore Forming Toxin
PMF	Potential Mean Force
PME	Particle Mesh Ewalds
POW	Parallel Optimization Workbench
PSO	Particle Swarm Optimization
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SASA	Solvent Accessible Surface Area
VDW	Van der Waals
WT	Wild Type

1 Introduction

1.1 Biology *in silico*

Biomolecules are dynamic structures, responding to the changing cellular environmental conditions in order to accomplish a specific biological function. Molecular modeling techniques aim at capturing their structural and dynamic characteristics by reproducing the ensemble of physical forces acting at the atomic level. In latest years, the role of modeling in biology has become increasingly important. Nowadays, *in silico* studies can not only explain known experimental results, but also correctly predict biomolecular properties that can be subsequently validated *in vitro*. This is achieved by exploiting a large variety of experimental inputs, from genetic sequences to crystal structures, from cryo-EM maps to functional studies. By bridging theory to experiment in biology, molecular modeling constitutes therefore a valuable tool to understand Nature at its finest detail.

Still, describing precisely the nature of atomistic interactions requires complex and finely calibrated physical models. As an additional hurdle, the evaluation of all the physical terms included in such models usually comes with a very high computational cost. For molecular simulation to be an effective and reliable tool, two main objectives must therefore be pursued: performance and accuracy. Data production should indeed be as fast, but nevertheless as precise as possible.

One of the available computational techniques is molecular simulation, aiming at describing the evolution of atomic systems along time by iteratively computing the forces acting on every individual atom. Since its foundations, in early 1960s, the accessible timescales and system sizes have not stopped growing. Remarkably, while initially few tenth of atoms could be studied for timescales in the picosecond range, nowadays large macromolecular assemblies can be simulated for microseconds, and more. The factors affecting this evolution can be mainly identified in the continuous refinement of existing physical models, the development of new computational paradigms, and the steady, concurrent progress in hardware architectures and algorithmic efficiency.

Novel computational paradigms typically aim at either reducing the systems complexity, or speeding up their conformational sampling. An example of the first approach is coarse-grained (CG) modeling which, by adopting a simplified version of the studied system, both reduces its size and increases the usable simulation's timestep. This is achieved by grouping individual atoms into bigger pseudo-atoms. As a consequence, larger molecular systems are accessible for longer timescales. Other techniques do not modify the way the system is represented, but how the molecules conformational space is explored instead. These "enhanced sampling" schemes, such as metadynamics or umbrella sampling, aim at making more likely the observation of rare events in simulation via the addition of specific biasing forces.

Clearly, a pivotal role is also played by technological improvements, leading to the construction of increasingly powerful computers. According to Moore's Law, an empirical observation done in 1965 by Intel co-founder Gordon E. Moore, transistor density doubles every 2 years. Similarly, Kryder's Law predicts that data storage doubles annually. Beyond all original expectations, newly developed computers still respect these trends. Recently, a major technological breakthrough has been accomplished by assembling computational nodes exploiting Graphical Processing Units (GPU). This development has large implications for molecular simulation. Indeed, GPU revealed to be much more efficient than the classic Central Processing Units (CPU) in the computation of atomic interactions, with a comparable precision. The great advantage of GPU is that molecular simulations are not bound to supercomputers, but can also be performed on simple workstations. Codes performing molecular simulation have been constantly adapted to exploit available computational resources at their best. With no exception, new software performing molecular simulations on GPU is currently emerging.

It is remarkable how the steady progress of molecular simulations observed during the last 50 years heirs from fruitful interactions and collaborations of scientists having different backgrounds, from physics to chemistry, biology, mathematics and computer science.

1.2 The Role of Minimization

Combinations of *in vivo*, *in vitro* and *in silico* experiments can nowadays tackle biological problems of great complexity, such as the prediction of structural and functional properties of biological nanomachines. Still, despite all the improvements in algorithms and computing resources, lots remains beyond the limits of modern molecular simulation. Technical improvements are continuously extending its ranges of application, but developments of alternative computational paradigms are still needed. Or, in other terms, to reach a destination faster and safer one needs a good car, but also a good street.

On one hand better, more refined physical models are still needed. Force fields describing molecular interactions are still being tuned and extended in order to generate simulations reproducing exactly experimentally measurable macroscopic features. This can be done both by introducing new functional forms, and by optimizing the parameters controlling these functions. CG force fields are no exception. In fact, while there is a general agreement about the functional forms describing all atom interactions, the description of CG interactions is still matter of debate.

Instead of improving an existing street, some particularly challenging problems might require the creation of a new one. Within these, we find the prediction of protein folding, and how folded proteins assemble into multimeric complexes. Often monomeric proteins arrange in a multimeric structure in order to achieve a specific task. As the folded state of single proteins, the stable conformation of these assemblies is also unique. From the experimental side, owing to their size and complexity, multimers are difficult to crystallize. In principle, knowing the structure of the individual subunits it would be possible to predict *in silico* the structure of the whole assembly. Unfortunately, these subunits can change conformation when multimerizing, and these conformational changes might span timescales way beyond the limits of any current simulation capability. *In silico* prediction could be a precious source of information, but at present the use of molecular simulations in this context is possible only for a very limited number of cases.

The parameterization of a force field and the prediction of protein assembly seem to be different problems, but they share in fact an important characteristic: both require *minimization*. When parameterizing a CG force field, one tries to minimize the difference within a CG simulation and given data, either experimental, either computational (e.g. All Atom simulations).When predicting a protein assembly, one tries to find the multimeric conformation having unique structure and minimal energy. Unfortunately, problems related to biological systems often imply very large and complex search spaces. In this context, minimization algorithms based on classical mathematical approaches (such as steepest descent or conjugate gradient) will fail to converge to the global minimum, and are therefore unsuitable. For this reason, tackling these problems with a minimization technique robust to local minimum is of capital importance. It is also important to point out that (at least in some cases) we are not blind in our search: experimental results can, and should be used as a guide.

1.3 Objectives

With a combination of known and new computational techniques, we study two biological systems being challenging both for their size, and for the timescales involved in their functioning: (i) *Aeromonas hydrophila*'s pore-forming toxin Aerolysin, and (ii) *Yersinia enterocolitica* injectisome.

- Aerolysin is a pore-forming toxin secreted as an hydrophilic monomer. By means of some conformational changes, the protein oligomerizes on the target cell's surface, and finally inserts into its lipid bilayer, causing cell death. The main hurdle in the study of this structure is the complexity of this mode of action, which spans timescales unreachable by classical molecular dynamics.
- The injectisome is an example of a type III secretion system. Its most striking feature is probably its size: hundreds of proteins assemble in a unique structure spanning the Gram-negative bacterial double membrane, and protruding outside the cell as a needle for tenth of nanometers. Obtaining an atomistic representation of this massive structure, and therefore some insights about its mode of action, is one of the greatest challenges.

These macromolecular structures, which qualify as nanomachines, are studied by means of molecular dynamics techniques, Poisson-Boltzmann calculations, and a newly developed minimization framework called Parallel Optimization Workbench (POW). The latter is used to predict the most likely assembly of monomeric proteins on the basis of known experimental restraints. Our computational results are compared and validated with *in vitro* experiments, and used to guide further experimental investigation.

This thesis is structured as follows. First, the main computational techniques we adopted are briefly described (Chapter 2). Subsequently, POW implementation and benchmarks are detailed (Chapter 3). Finally, computational results on aerolysin (Chapter 4) and the injectisome (Chapter 5), as well as their comparison with experimental data, are presented.

2 Methods

The main computational techniques adopted in this work are based on molecular simulation and estimation of binding free energy. In this chapter we describe the physical foundations of these techniques, as well as their implications for the study of biological systems.

2.1 Molecular Simulations

Molecular Mechanics Force Fields

In molecular simulation, one describes atoms as being immersed in an energy potential field generated by the physical characteristics of its environment. In a All-Atom (AA) representation, each atom is represented as a charged point mass interacting via bonded and non-bonded interactions with its neighbors. Every interaction is represented using a potential having a predetermined functional form. With the assumption that all the contributions are simply additive, the potential acting on one atom can be represented with the following sum:

$$U = \underbrace{U_{bond} + U_{angle} + U_{dihedral} + U_{improper}}_{bonded} + \underbrace{U_{vdW} + U_{coulomb}}_{non-bonded} \tag{2.1}$$

U_{bond} represents the covalent bond within two neighboring atoms. For small displacements with respect of an equilibrium distance, the potential of a single bond can be approximated by an harmonic term:

$$U_{bond}(r) = k_b(r - r_0)^2 \tag{2.2}$$

where k_b is the spring constant defining the force at which the two atoms are restrained around the equilibrium distance r_0.
U_{angle} represents the angle potential within three consecutive covalently bonded atoms, and

is also approximated with a following harmonic term (summing over all the angles α):

$$U_{angle}(\alpha) = k_a(\alpha - \alpha_0)^2 \tag{2.3}$$

where k_a is the spring constant defining the force at which the three atoms are restrained around the equilibrium angle α_0.

$U_{dihedral}$ describes the dihedral angles within 4 consecutive covalently connected atoms. It is a periodic potential represented as the following sum of cosines (summing over all the dihedrals multiplicity n):

$$U_{dihedral}(\psi) = \sum_n k_d \left[1 + cos(n\psi - \psi_0)\right] \tag{2.4}$$

where k_d is the spring constant defining the force at which the four atoms are restrained around the equilibrium angle ψ_0. The value of n affects the potential periodicity, and depends on the nature of interacting atoms.

$U_{improper}$ describes the out of plane rotation of one atom, and is usually used in order to enforce planarity (summing over all the impropers i):

$$U_{improper}(\phi) = k_i(\phi - \phi_0)^2 \tag{2.5}$$

where k_i is the spring constant defining the force at which the improper angle within the four atoms is restrained around the equilibrium angle ϕ_0.

U_{vdW} describes the non-bonded van der Waals interaction within two atoms. Typically, this interaction is computed only for atoms not being covalently connected within each other by a bond (1-2 interaction) or angle (1-3). Sometimes, the Van der Waals interaction within atoms having a 1-4 distance is also ignored or, if accounted, scaled. U_{vdW} has usually the form of a 12-6 Lennard-Jones potential:

$$U_{vdW}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.6}$$

This energy term has a single minima located at σ, and converges to zero at infinite. The depth of the potential well is determined by the value of ϵ. Other functional forms, such as the Buckingham or Morse potential, can be used. However, the 12-6 Lennard-Jones is usually favoured for its lighter computational weight. In order to further decrease the computational cost of U_{vdW}, a cutoff distance is usually imposed. Only interactions of atoms being closer than a predefined threshold c_t are added to the U_{vdW} sum. This approximation is acceptable since U_{vdW}, which converges to 0 with r^{-6}, is considered as short-ranged. Typically, a switching function scales the U_{vdW} value for distances greater than a predefined threshold c_{switch}, so that U_{vdW} is equal to 0 at c_t. The switching function preserves the continuity of the potential as well as that of its first derivative.

$U_{Coulomb}$ describes the electrostatic interaction within non-bonded atoms. Again, 1-2, 1-3 and sometimes 1-4 interactions are ignored:

$$U_{Coulomb}(r) = \frac{1}{4\pi\epsilon_0\epsilon_r}\frac{q_1 q_2}{r} \tag{2.7}$$

Where q_1 and q_2 are the charges of the two atoms. The computation of long-ranged $U_{coulomb}$ is extremely costly, for this reason two main approaches are typically used to lessen its computational weight. The first approach, is to simply apply a cutoff distance as previously described for U_{vdW}. The main disadvantage is that in highly polar systems the potential at cutoff distance could fluctuate due to entry and exit of charged atoms in the exclusion zone. Results of simulations performed with different cutoff values could even vary. To lessen this phenomenon one should select a very large cutoff value, which however rapidly leads to more computationally expensive calculations. Indeed, the complexity of the cutoff method is $O(N^2)$. The second approach, called Particle Mesh Ewald (PME) is usually more suitable since it is not only insensitive to boundary effects, but has a complexity of $O(N.logN)$ [32]. In PME, the electrostatic potential is decomposed in a sum of two terms, a short-range term computed in real space, and a long range term computed in reciprocal space via Fast Fourier Transform.

Additional terms can be added to the sum 2.1. These can be as different as cross-correlation terms defining correlation within bonded potentials, external biasing forces or distance restraints keeping specific atoms at a desired distance. In some cases a bonded potential within atoms at 1-3 distance, called Urey-Bradley potential, is also added to the functional form. The coefficients in the equations above depend on the nature of involved atoms. The functional form of the interactions, as well as their parameterization, determines what is called a *force field*. In this work we mainly used the Amber99SB [21] and CHARMM27 [86] force fields. Within all the force fields available for biomolecular simulations, these have been shown to be within the most accurate in reproducing experimental results. Remarkably, Amber99SB is reported to better reproduce structural features of proteins solved both via NMR [75] and X-ray crystallography [23] as compared to other popular force fields such as OPLS or Gromos.

Unfortunately, at present the official Amber force field does not feature any parameterization for lipids. For this reason, in order to simulate eukaryotic membrane models, we adopted to CHARMM27 force field, which features an accurate parameterization for POPC and POPE lipids.

Molecular dynamics

Molecular Dynamics aims at simulating the evolution of an atomic ensemble over time. This is done by computing in an iterative way the forces acting on every atom on the basis of a Molecular Mechanics force field. Let $X(t)$ the position of an ensemble of atoms at time t. The potential $U_{MM}(X(t))$ generates a force on every atom:

$$F(X(t)) = -\nabla U(X(t)) \tag{2.8}$$

The force is calculated and evaluated on the atoms of a molecular structure in an iterative manner, usually for steps being in the femtosecond scale. To do so, Newton's law of motion is converted into an iterative scheme. This scheme, called Verlet, is defined as follows:

$$\begin{aligned} X_{n+1} &= 2X_n - X_{n-1} + \Delta t^2 M^{-1} \nabla U_n(X(t)) && (2.9) \\ V_n &= \frac{(X_{n+1} - X_{n-1})}{2\Delta t} && (2.10) \end{aligned}$$

Where X_n and V_n are atom's position and velocity at time timestep n, Δt the timestep size and M the atoms mass. Therefore, in Verlet integration scheme two system positions are needed to update system's velocity. Three main variations of this scheme exist: leapfrog, position and velocity Verlet, the latter being the most popular. The integration timestep should be sufficiently small to sample the phenomena having the highest frequency, that is the fluctuation of a covalent bond within a hydrogen and a heavy atom. This fluctuation has a frequency in the order of 10^{14} Hz, and can be sampled with a timestep of 1 fs. After position X_{n+1} is computed, algorithms such as SHAKE [124] or LINCS [50] can modify it by imposing distance constraints. Constraining the oscillation of hydrogen distance to its heavy atom and bonds within heavy atoms allows the timestep to be increased to 2 fs. This timsetep site is sufficient to sample the oscillation of a covalent bond within two heavy atoms.

Molecular dynamics simulations of biomolecules typically take place inside a cubic domain, which we call unit cell. The cell is usually cubic, but other geometries (triclinic, truncated octahedron,...) are also possible. The cell is periodically replicated in every direction. In order to avoid an atom to interact with both another atom and its image in a neighboring cell, the minimum image convention is applied: every atom interacts only with the closest image of other atoms.

Real biological systems evolve in specific temperature (around 300-310 Kelvin) and pressure (around 1 atmosphere) ranges. In order for a MD simulation to reproduce faithfully their characteristics, having a control on these quantities is therefore required. Simulation in a specific thermodynamic ensemble can be performed by controlling pressure and temperature via a barostat and a thermostat. Berendsen [15] and Nose-Hoover [104, 52] thermostats control the system's temperature by modifying the equations of motion so that atoms velocities $V(n)$ are scaled. As such, a coupling to a thermal bath is reproduced. Differenly, Langevin thermostat [125] acts via additional forces, a dissipative friction term and a random term, applied on every atom.

A barostat controls atoms positions $X(n)$ so that the system is maintained around a desired pressure. As a consequence, this will also affects the unit cell's size. In order to evaluate a system's internal pressure, the virial theorem can be exploited:

$$P = \frac{2}{3V}(E_k - vir) \tag{2.11}$$

This theorem relates the pressure P to the system's volume V, kinetic energy E_k and virial vir. Several equations reproducing the effect of a barostat have been proposed. Berendsen and Nose-Hoover thermostats, for instance, can be reinterpreted in order to act as a barostat. Finally, the Parrinello-Rahman [111] barostat extends Nose-Hoover by also allowing the simulation of anisotropic systems (i.e. allowing each system dimension to vary independently).

Given a sampling time of a system's phase space (positions and velocities) tending to infinite, the ergodic theorem states that the time average of a measured quantity will be equal to its ensemble average. MD simulations cannot however sample a phase space for an infinite time. Still, by simulating a system for a time much longer than the timescales required to observe a desired phenomenon, one can suppose that the ergodic theorem is satisfied. This "ergodic hypothesis" is, in fact, at the basis of MD simulations. The time needed to sufficiently sample a relevant biological event is related to the number of atoms present in the system being simulated. MD being extremely computationally expensive, sampling techniques artificially enhancing the phase space sampling have been developed. Within these, we will mention metadynamics [74] and umbrella sampling [134].

Molecular Dynamics codes running on multiple processors are available. In this work, we used

NAMD [116]. NAMD features a very good scaling on multiple processors, and can simulate molecular systems parameterized with both Amber and CHARMM force fields.

The ensemble of successive atomic conformations obtained via MD, also called trajectory, are analyzed in order to assess relevant structural features. Within the most direct quantities measured in this work we find the following:

- Root Mean Square Deviation (RMSD). Let X_1 and X_2 the atomic positions of two systems containing the same amount of atoms. RMSD quantifies how similar the two systems are:

  $$RMSD(X_1, X_2) = \sqrt{\frac{\sum_{i=1}^{n}(X_{1,i} - X_{2,i})^2}{n}} \tag{2.12}$$

  Tracking the RMSD of atoms configurations of a system simulated in MD with respect of a reference structure is useful to determine whether the system has reached equilibrium. Indeed, at equilibrium the system's RMSD is indeed expected to converge to a constant value.

- Root Mean Square Fluctuation (RMSF). RMSF quantifies how much an atom position X fluctuates around a mean point along time:

  $$RMSF(X) = \sqrt{\frac{\sum_{t=1}^{T}(X_t - \bar{X})^2}{T}} \tag{2.13}$$

  where X is an ensemble of measures and $\bar{X}$ their mean value in time T. In a protein, the more an amino acid's RMSF is low, the more it is stable, and viceversa. In fact RMSF is related to beta-factor measured in X-ray crystallography as follows:

  $$\beta(X) = \frac{3}{8}\pi^2 RMSF(X)^2 \tag{2.14}$$

- amino acids' secondary structure. Amino acids in a polypeptide chain arrange locally according to their intrinsic nature and to the characteristics of the environment. This local arrangement, called "secondary structure" is determined by the values of amino acid backbone's dihedral angles ϕ and ψ. Specific couples of these angles can lead to polypeptide chains being totally extended, coiled in an helical conformation or simply randomly arranged. Secondary structure can be classified in seven different categories according to the Dictionary of Protein Secondary Structure (DSSP) [63]. Tracking the secondary structure evolution of a protein in a MD simulation is a useful tool to highlight conformational changes.

Additional specific analysis techniques of MD trajectories are described in dedicated sections in the following of this work.

2.2 Binding Free Energy Calculations

In order to compute the binding free energy of two binding partners we adopted the MM/PBSA (Molecular Mechanics/Poisson-Boltzmann/Surface Area) method [133], readily available in Amber Tools suite [21]).

According to the thermodynamics cycle shown in picture 2.1,the binding free energy ΔG of two molecules, ligand and receptor, can be decomposed in the following sum of molecular mechanics, solvation and entropic contributions:

$$\Delta G = -\left(\left\langle \Delta G_{solv}^{ligand} \right\rangle + \left\langle \Delta G_{solv}^{receptor} \right\rangle\right) + \langle \Delta E_{gas} \rangle + \left\langle \Delta G_{solv}^{complex} \right\rangle - T \langle \Delta S_{solute} \rangle \tag{2.15}$$

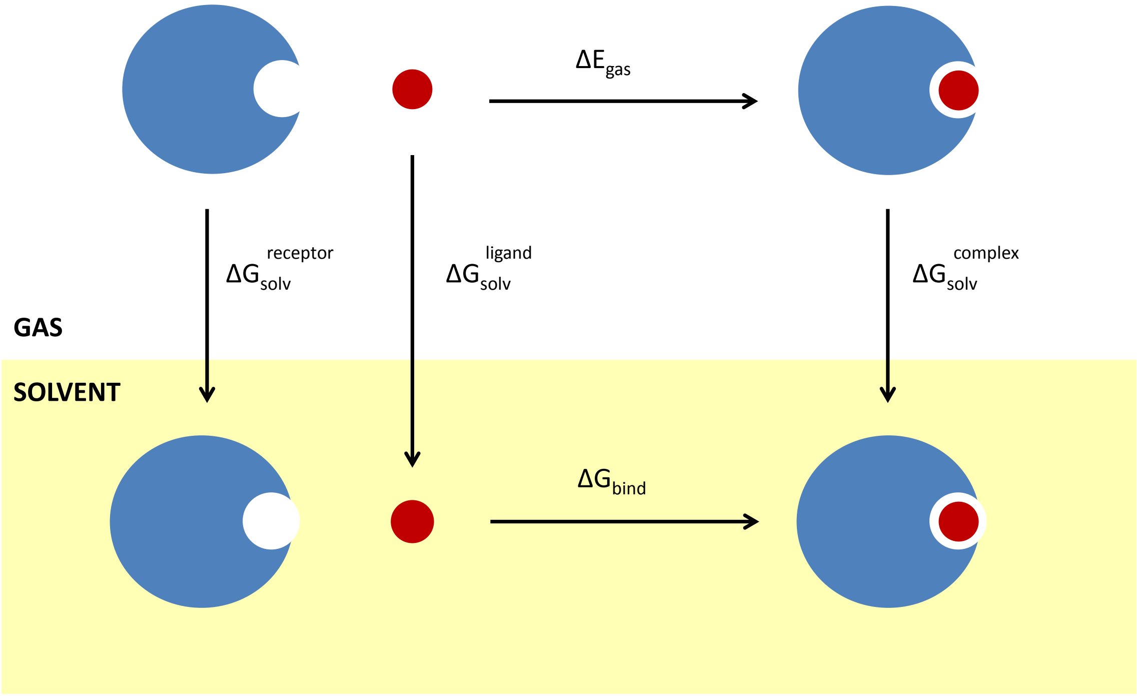


Figure 2.1: *A ligand is depicted in red, a receptor in blue. In order to compute the binding free energy between the two binding partners, the following thermodynamics cycle can be adopted.* ΔG_{solv} *represents the energy difference between solvated and desolvated state, and* ΔE_{gas} *the binding free energy of the binding partners* in vacuo

ΔE_{gas} defines the energy difference within separated binding partners and their complex in gas phase. This contribution is simply determined by electrostatic, Van Der Waals and internal terms computed via molecular mechanics (see previous section):

$$U_{MM} = U_C + U_{vdW} + U_{int} \tag{2.16}$$

The solvation $G_{solvation}$ of every state is constituted by a polar and a nonpolar contribution. The Polar Solvation contribution G_{PS} is computed via the Poisson-Boltzmann (PB) equation:

$$\vec{\nabla} \cdot \left[\epsilon(\vec{r})\vec{\nabla}\Psi(\vec{r})\right] = -\rho^f(\vec{r}) - \sum_i c_i z_i q \lambda(\vec{r}) exp\left[-\frac{z_i q \Psi(\vec{r})}{k_B T}\right] \tag{2.17}$$

Where ϵ is the dielectric, $\Psi(\vec{r})$ the electrostatic potential, $\rho^f(\vec{r})$ the solute's charge density, z_i and c_i the charge and concentration of ion i far from the solute, q the charge of a proton, $\lambda(\vec{r})$ a function determining the accessibility of ions to every point in space and k_B the Boltzmann constant. The solution of this differential equation with respect to $\Psi(\vec{r})$ constitutes the polar contribution to solvation energy. Since computing $\Psi(\vec{r})$ with the Poisson-Boltzmann equation is computationally expensive, models approximating it have been developed. Within these we find the general Born (GB) model [45]. Adopting the GB model leads to a very large gain in terms of computational time, the resulting measurement will however be less accurate.

One of the most popular ways to approximate a solute's nonpolar solvation term G_{NS}, i.e. the energy needed for the formation of a cavity, is by means of a linear correlation with its solvent accessible surface area (SASA):

$$\Delta G_{NP} = \gamma SASA + b \tag{2.18}$$

Where γ is the surface tension coefficient and b the nonpolar solvation energy of a point solute. These adjustable parameters are usually set on the basis of experimental measures of solvation energies of small molecules.

The term $T\Delta S_{solute}$ is also added in order to account for the entropy change of the system at a given temperature T *in vacuo*. In order to estimate it, Normal Modes Analysis (NMA) or Principal Component Analysis (PCA) of the given trajectory can be performed. These calculations are however computationally expensive, and are therefore often omitted. Still, entropic contribution in large molecules can be important.

The ensemble averages on all the presented terms should be performed on a set of decorrelated conformations generated via a molecular mechanics simulations. Two approaches exist, namely multiple and single trajectory. In the first case, MM/PBSA calculation is performed on the basis of molecular dynamics simulations independently run for ligand, receptor and complex. In the second case a single trajectory of the complex is computed, and trajectories for ligand and receptor are generated by simply extracting their coordinates from the complex. The first technique is usually more precise, but is however more computationally expensive.

Precision of computation in ΔG within large macromolecules such as proteins is often limited. The size of these systems, with the effect of possible conformational changes upon binding, can indeed be hard to estimate. Conversely, the computation of $\Delta\Delta G$ (for instance the difference in binding free energy within a Wild Type and mutated molecule to the same receptor) can be highly accurate. Higher accuracy is mainly due to a cancelation of the entropic contributions. In this context, Alanine Scanning can lead to accurate predictions. By mutating an amino acid to alanine, most of its interactions with the environment are canceled. Therefore, the difference in binding free energy within Wild Type and alanine mutated protein defines the importance of the mutated amino acid's sidechain for its binding.

3 Parallel Optimization Workbench (POW)

Optimization problems related to the macromolecular assembly of biological systems often imply very large and complex search spaces. In this context, traditional minimization methods tend to be inefficient. To tackle this problem, we developed Parallel Optimization Workbench (POW), a general optimization framework based on Particle Swarm Optimization.

3.1 The engine: Particle Swarm Optimization (PSO)

Let a function $f(\vec{x})$, where $\vec{x} \in F^n \subset \mathbb{R}^n$. We will call f *fitness function*, and *search space* the multidimensional real space F^n in which the function is defined. We want to find a point $\vec{x}_{min} \in F^n$ such that $y_{min} = f(\vec{x_{min}})$ is the function's global minima.

PSO is a particularly robust heuristic optimization technique aiming at finding the point $\vec{x_{min}}$ having the lowest fitness. This technique, invented in 1995 by James Kennedy and Russell C. Eberhart, represents the search process as a model of birds' social behavior when flocking [67]. To do so, an ensemble of solutions (also called particles `p`) have their position `x(p)` and velocity `v(p)` randomly initialized in the multidimensional search space. At every discrete timestep, the velocity of every particle is updated. This, in turn, is used to compute a new position, where the fitness `f` will have to be evaluated. Every particle keeps track of the value `f_best(p)` and position `x_best(p)` of its best found solution, and also compares it to the best solution found by neighboring particles (solution located at position `x_best'`). These quantities will be used to update the swarm state. Algorithm 1 represents a typical PSO implementation.

Typically, within 20 and 60 particles are initialized. Their velocity in the search space is affected by 3 factors. `w` is called inertia, and determines how the chosen particle trajectory is preserved along time. This value varies within 0 and 1. It has been shown that improved performance can be achieved by starting inertia at high values and gradually reducing it while optimization proceeds [128]. A high inertia value produces a more "turbulent" swarm, which is ideal for an

Algorithm 1 Particle Swarm Optimization

```
for every timestep t do:
    for every particle p do:
        $inertia \leftarrow w * v(p, t-1)$
        $personal \leftarrow cp * rand(0,1) * (x(p, t-1) - x_{best}(p))$
        $global \leftarrow cn * rand(0,1) * (x(p, t-1) - x'_{best})$
        $v(c, t) \leftarrow inertia + personal + global$
        if $|v(p, t)| \geq size(space)$ then
            $v(p, t) \leftarrow norm(v(p, t)) * size(space)$
        end if
        $x(p, t) \leftarrow x(p, t-1) + v(p, t)$
        if $f(x(p, t)) \leq f_{best}(p)$ then
            $f_{best}(p) \leftarrow f(x(p, t))$
            $x_{best}(p) \leftarrow x(p, t)$
        end if
    end for
end for
```

initial exploratory phase. Reducing the inertia has a "cooling" effect, which is more suitable when areas of interest have been discovered. `cp` scales the influence of knowledge of best found solution by the current particle, whereas `cn` scales the influence of best solution found by neighbors. While in literature values for initial and final `w` to 0.9 and 0.4 are usually accepted, values of `cn` and `cp` are subject to debate. In this context, meta-optimization approaches (optimization of parameters to improve PSO performance for a specific problem) have been proposed [114, 91].

A particle's neighborhood can be defined either as indexed, either as geographic. In the first case, an index is assigned to every particle, and a predefined connectivity is set. This can go from a fully connected graph (so called *gbest* neighborhood) to a ring in which particles having consecutive indexes are considered as neighbors (*lbest*). In geographic neighborhood, a particle will select as neighbors only particles being close in the search space. Either the n first neighbors, either all the neighbors in a predefined cutoff can be kept into account. The nature of neighborhood relationship can have an effect on PSO performance. Interestingly, it has been observed that a fully connected swarm would converge faster, but perform poorly than a partially connected one, being more sensitive to local minima [66, 68].

Search space boundary conditions can be enforced in several ways, the most common being periodic and reflexive. In periodic boundary conditions, the unit cell containing the search space is replicated periodically in every direction. A particle leaving the unit cell, will reappear in the neighboring cell. in reflexive boundary conditions, cell borders are considered as hard walls. Particles would therefore bounce elastically against them. An upper threshold for particles' velocity corresponding to the search space size is usually set.

At the beginning of every new timestep every particle is updated about the swarm status (current position of all particles, and their respective best found solution value and location). Subsequently, their position and velocity are updated independently. After having been updated about the swarm's state at the beginning of a new timestep, particles act as independent agents. For this reason PSO can be considered as embarrassingly parallel.

As a halting condition, a maximal number of timesteps can be set. Alternatively, the best found fitness value can be tracked, and PSO search can be stopped once a predefined convergence criteria is met. Finally, measures of swarm position, velocity, or cognitive diversity can be exploited [26]. Position diversity tracks the spread of particles in the search space. Velocity diversity assesses the amount of particles movement, which is expected to drop when search converges. Finally, cognitive diversity defines the spread in best solutions stored in particles local memory. Measuring a swarm's diversity can be seen as a way to assess the swarm's "exploration vs. exploitation" balance. Initially, the particles roam randomly in the search space (exploration) but, as the search proceeds, they subsequently focus more and more on specific regions of interest (exploitation). One does not want the swarm either to explore excessively (which would correspond to a random search) either to exploit excessively (which would resemble to a gradient descent, prone to lead to the discovery of suboptimal solutions). Importantly, this criterion also holds for other distributed stochastic optimization techniques such as Genetic Algorithm.

Applicative studies show that this algorithm is both highly robust to local minima and usually converges as fast, and in some cases even faster, than other heuristic approaches such as Genetic Algorithm or Simulated Annealing [16, 38, 102, 11, 2]. However, it has also been shown that when the fitness function's profile becomes extremely rough, the search might still terminate with a premature convergence, that is, with all particles stagnating in local minima [11]. In the following, we briefly review the main variations proposed to increase PSO robustness and convergence rate.

Some approaches try to avoid early convergence by controlling particles velocities or positions. FATPSO (Fuzzy Adaptive Turbulent PSO [2]) does so by perturbing particles' velocity when this drops below a threshold adaptively tuned via a Fuzzy Logic controller. Similarly, Gregarious PSO [113] randomizes particles velocities when these get stuck close to the global minima. VBR-PSO [17] tracks the average swarm velocity, and reinitializes all the particles when this drops below a predefined threshold. M. Clerc proposes to restart particles positions according to a No-Hope criterion: when there is no more hope to find the optimal solution, particles are reinitialized around the best found solution by keeping into account the estimated local fitness function shape. The gravity center of the Swarm, called the Queen, is also tested [28].

Differently, instead of directly controlling particles velocities or positions, other approaches modify PSO coefficients. This is the case of HPSO-TVAC, [119] which sets time varying `cn` and `cp` coefficients, `w` to zero, and randomly accelerates particles when these get stuck. The usage of appropriate PSO coefficient has also been the aim of Clerc and Kennedy [29]. By analyzing the behavior of an individual particle in the swarm, the authors proposed a criteria to select them, and also introduced a constriction coefficient. The latter allows to control the balance within exploration and exploitation, and removes the need for a threshold to maximal particle velocity.

Algorithm 2 Particle Swarm Optimization using Kick and Reseed approach

```
for every timestep t do:
    for every particle p do:
        $inertia \leftarrow w * v(p, t-1)$
        $personal \leftarrow cp * rand(0,1) * (x(p, t-1) - x_{best}(p))$
        $global \leftarrow cn * rand(0,1) * (x(p, t-1) - x'_{best})$
        $v(c, t) \leftarrow inertia + personal + global$
        if $|v(p, t)| \geq size(space)$ then
            $v(p, t) \leftarrow norm(v(p, t)) * size(space)$
            $x(p, t) \leftarrow x(p, t-1) + v(p, t)$
        else if $|v(p, t)| \leq v_{min}$ $and$ $f(x(p, t)) \geq f_{min}$ then
            $v(p, t) \leftarrow rand(0,1) * v_{min}$
            $x(p, t) \leftarrow x(p, t-1) + v(t)$
        else if $|v(p, t)| \leq v_{min}$ $and$ $f(x(p, t)) \leq f_{min}$ then
            $v(p, t) \leftarrow rand(0,1) * v_{min}$
            $x(p, t) \leftarrow rand(0,1) * space$
        else
            $x(p, t) \leftarrow x(p, t-1) + v(p, t)$
        end if
        if $f(x(p, t)) \leq f_{best}(p)$ then
            $f_{best}(p) \leftarrow f(x(p, t))$
            $x_{best}(p) \leftarrow x(p, t)$
        end if
    end for
end for
```

We implement here a new PSO flavor which we call *kick and reseed*. This method is meant to avoid early convergence and lead to an increased sampling. To do so, particles velocities are constantly monitored. When a particle slows below a predefined threshold velocity v_{min}, two possible actions are taken. If the current fitness value is above a predefined threshold f_t, the particle velocity is randomly reinitialized. Conversely, if the current fitness is below said threshold, the particle is also randomly reseeded in a new position of the search space, and its memory about its personal best erased. The latter behavior is meant to increase the PSO sampling. Algorithm 2 describes our variation of the original PSO, called *PSO Kick and Reseed* (PSO-KaR).

A minimal `v_min` velocity is defined. Particles being too slow receive a random kick if their fitness is not low enough (above a threshold `f_min`). If slow particles also have a good fitness, they are randomly kicked and restarted in a new position in search space. The effect of our modification on PSO performance is studied in section 3.3.

When the fitness function is particularly hard to optimize, PSO-KaR might still converge too slowly to identify the global minima before the maximal number of timesteps is met. In this case, multiple runs can be performed. In PSO-KaR, we propose a second improvement meant to improve sampling in this context. We suppose that every time a particle is repelled or reseeded, the region was thoroughly sampled by the particle. For this reason, we want to avoid other particles to explore the same space. To do so, we introduce a repelling force (in the following referred as *repeller*), pushing particles away from explored regions.

Algorithm 3 PSO-KaR velocity update variation, using repellers

```
$bias \leftarrow 0$
for every repeller r do:
    $bias \leftarrow -\frac{s}{(x(p,t)-r)^2}$
end for
```

The repelling force is generated by a simple $f(x) = x^{-2}$ potential. Repellers list is incremented along a PSO-KaR execution, and is inherited by the following runs. Let r a list of points repulsion points, the contribution of the repelling potential on particles velocities is therefore calculated as shown in Algorithm 3. `s` is a user defined scaling value, determining how strong and long range the repulsion force is.

3.2 Architecture

We implemented PSO in a framework allowing the resolution of virtually any optimization problem via the addition of a specific module (POW, Parallel Optimization Workbench). This object oriented code is developed in Python, and supports parallel computation by exploiting MPI libraries. The architecture of our framework is represented in figure 3.1. Every box corresponds to a specific class. Classes highlighted in blue are common to any optimization problem, and can be considered as a black box by the user. Classes in the yellow area change depending on the problem being solved. We will call a *module* a file containing an implementation for these classes aiming at solving a specific problem. In order to use POW, a user has to provide two information: the module name, and a parameterization file.

The parameterization file contains a set of keywords associated to one or more values. Some keywords are standard for any optimization problem (for instance, those defining the behavior of PSO-KaR), whereas others are problem specific. The classes `DefaultParser` (for standard keywords) and `Parser` (for custom ones) are in charge of reading the input file. Once the parameters are parsed POW loads, if needed, specific data structures required by the user. This operation is performed by the class `Data`. Since this class is part of a module, depending on how this class is implemented, any data structure can be manipulated. Subsequently, POW defines the problem's search space. Every dimension of the search space is defined by upper and lower boundaries, as well as by specific boundary conditions. Creation of the search space is problem specific, and is managed by the `Space` class. Conversely, management of boundary conditions is the same for any optimization problem, and is implemented in the `DefaultSpace` class. The class `PSO` implements the PSO-KaR algorithm as described in section 3.1. The user is free to set them at will using specific keywords in the parameterization file. As default values, we set initial and final `w` to 0.9 and 0.4 respectively, `cp` to 1.4 and `cn` to 1.2. Along the optimization run, every measure performed by every particle is stored in a log file. In order to extract useful information, postprocessing this log file is necessary. The class `Postprocess` is in charge of this. Useful functions the user might need are preimplemented in the `DefaultPostprocess` class.

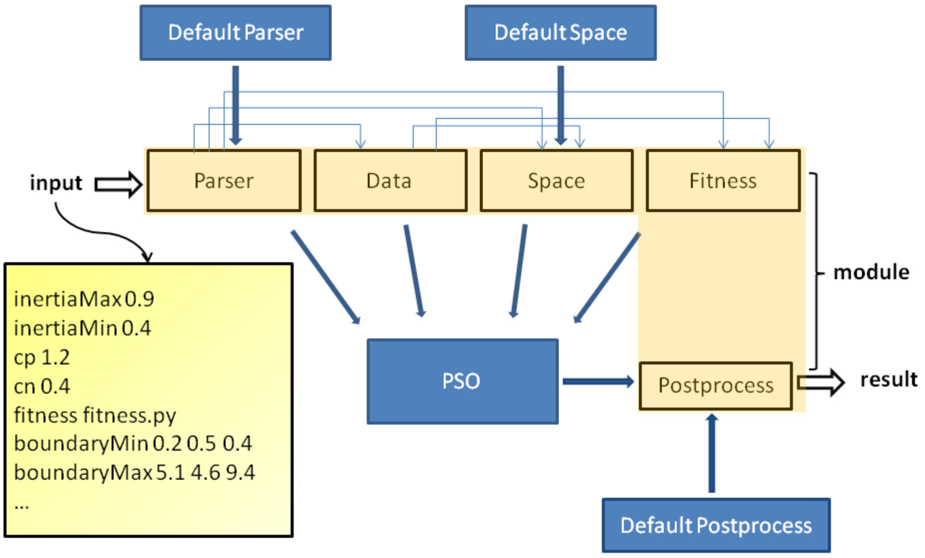


Figure 3.1: *Schematic of the Optimizer architecture. Every box represents a class. Classes highlighted in blue are common to any optimization problem, and can be considered as a black box for by the user. Classes in the yellow area change depending on the problem being solved. We call a module a file containing a definition for these classes aimed at solving a specific problem. Input is provided as a text file containing keywords with associated values.*

POW has been conceived so that the creation of a new module (i.e. a specific implementation of the `Parser`, `Data`, `Space`, `Fitness` and `Postprocess` classes) is trivial even for a user unaware of its internal architecture. The following modules are already available:

- `Function`: Find the global minima of a defined function. This is the simplest module, since no data structure has to be processed.
- `CGmatch`: perform a force, potential or property matching of a Coarse-Grain force field given an All Atom reference
- `ProteinProtein`: rigidly assemble a heterodimer provided known experimental constraints
- `DockSymmCircle`: assemble a homo multimer in a circular symmetry according to known stoichiometry and experimental constraints. Monomers can be treated both as rigid of flexible objects, and a docking receptor can be kept into account.

In the next section these modules will be described. A user manual are available in Appendix A.1.

3.3 Function Minimization

The module `Function` is the simples application of POW. This module does indeed not require manipulation of any data structure, and is adapted to the minimization of functions. In order to assess the capabilities of PSO-KaR and compare them with the standard PSO, we ran a set of tests using this module.

Two classic multidimensional benchmark functions were used: sine and Rastrigin (Figure 3.2).

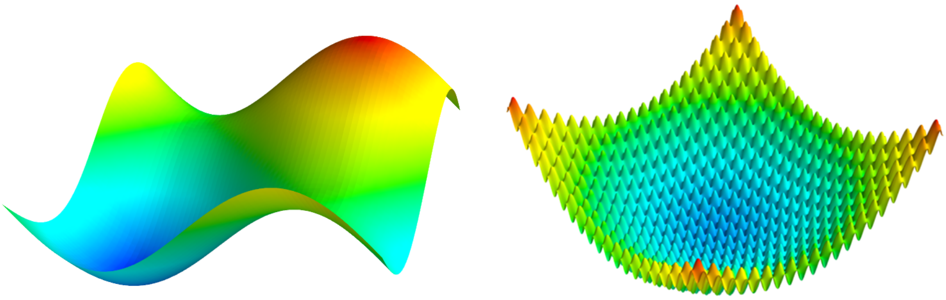


Figure 3.2: *Two dimensional sine (left) and Rastrigin (right) functions.*

The sine function is defined as follows:

$$f(\vec{x}) = 1 + \sum_n \frac{sin(x_n)}{n} \tag{3.1}$$

This function has been used as a trivial test. In the interval $[0, 2\pi]$ one unique minima exists, $f(\frac{3}{2}\pi) = 0$.

The Rastrigin function is defined as follows:

$$f(\vec{x}) = 10 + \sum_n x_n^2 - 10 * cos(2\pi x_n) \tag{3.2}$$

This function is particularly hard to optimize. In the interval $[-5.12, 5.12]$ it contains a large number of local minima, that get increasingly deep the closer they are to the unique global minima located in $f(0) = 0$.

1, 2, 3, 5, 5, 10, 20, 30, 40, 50 and 100 dimensional Rastrigin and Sine function were submitted to PSO and PSO-KaR optimization. For PSO-KaR, the minimal velocity threshold was set to 0.01. For both PSO and PSO-KaR, every of these functions was optimized in 10 independent runs with 1000 steps and 80 particles. In both cases particles neighbors were defined as the particles having preceding and following indexes. The average fitness at every optimization step are shown in Figure (3.3). We notice that, in every case, the more a function dimensionality increases, the more fitness convergence rate decreases. Increasing the size of the search space has the effect of reducing the particles density in it. Therefore, finding the funnel leading to the global minima, and the right trajectory to find its lowest point, becomes increasingly difficult. On Rastrigin function, early convergence is immediately observable when using PSO. This phenomenon becomes more dramatic when dimensionality increases. This is due both to the increased number of almost equivalent local minima, and by the already mentioned reduced particles density in the search space. For simple fitness functions (such as sine or low dimensional Rastrigin) no relevant difference can be observed within the PSO and PSO-KaR procedures. However, when increasing the fitness function complexity, the effect the "kick and reseed" procedure becomes more and more relevant. This shows that, while being comparable to PSO for easy fitness functions, PSO-KaR prevents early convergence in hard optimization problems.

To better assess the effect of the "kick and reseed" procedure, we tested different velocity thresholds while running PSO-KaR optimizations on the two hardest fitness function used

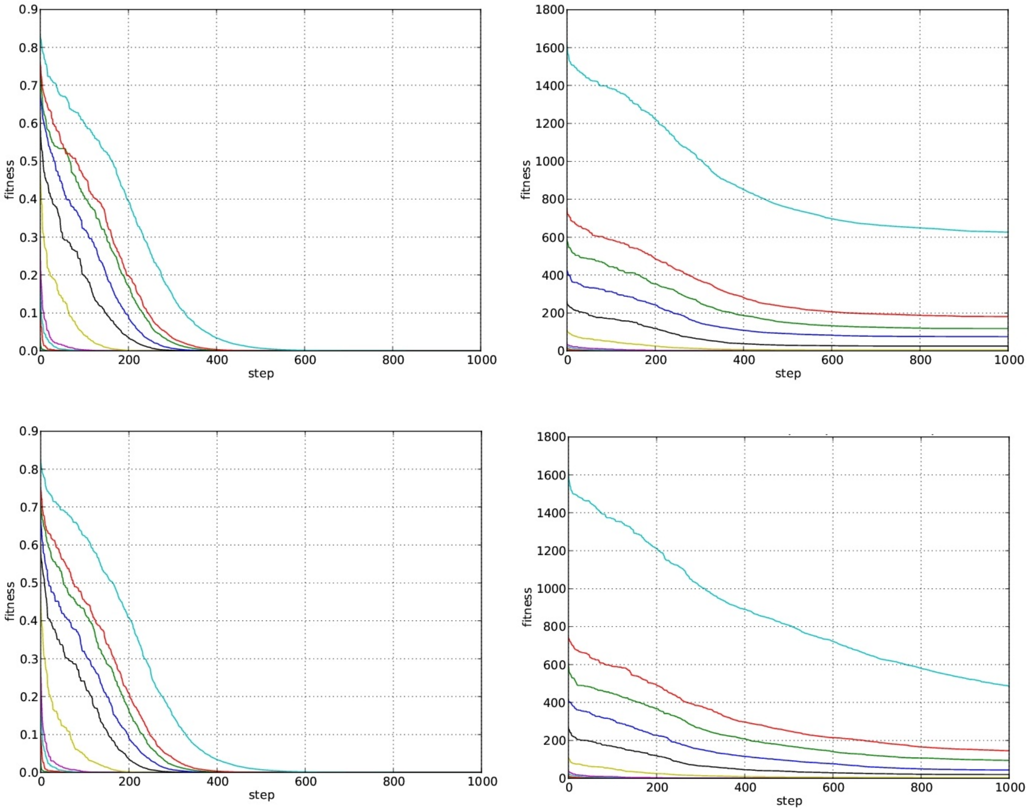


Figure 3.3: *Result on multidimensional Sine (left row) and Rastrigin (right row) using PSO (top row) and PSO-KaR (bottom row). Every result is an average of 10 independent optimization runs. The "Kick and Reseed" procedure improves fitness convergence when hard fitness functions are optimized, whereas it has no effect when optimizing easy functions.*

above, i.e. the 50 and 100 dimensional Rastrigin functions. Values equal to 1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001 and 0 (equivalent to the standard PSO) were tested. Averages of 10 independent runs with 1000 optimization steps and 80 particles are visible in Figure 3.4. For both the 50 and the 100 dimensional Rastrigin, a threshold value equal to 1 turns out to be too high. High thresholds introduce excessive noise in the swarm, which in turn converges with more difficulty. Conversely, thresholds being too low (such as 0.0001), do not perturb the swarm sufficiently to have a real positive effect in its search process. The best result, leading to improved performances with respect of the standard PSO, was obtained in both tests by a velocity threshold equal to 0.01. This value was therefore set as default in POW implementation. Importantly, the KaR procedure does not affect negatively the precision of search process. Indeed, despite the addition of noise in the search process, the standard deviation in PSO-KaR does not increase with respect of the standard PSO. By comparing the results of the 100 and 50 dimensional Rastrigin, we observe that in the second case a broader range of threshold values leads to similar performances. Previously, we also observed that for very simple functions (sine) the KaR procedure has no effect. This indicates that the more the fitness function becomes complex, the more a good choice of velocity threshold becomes important.

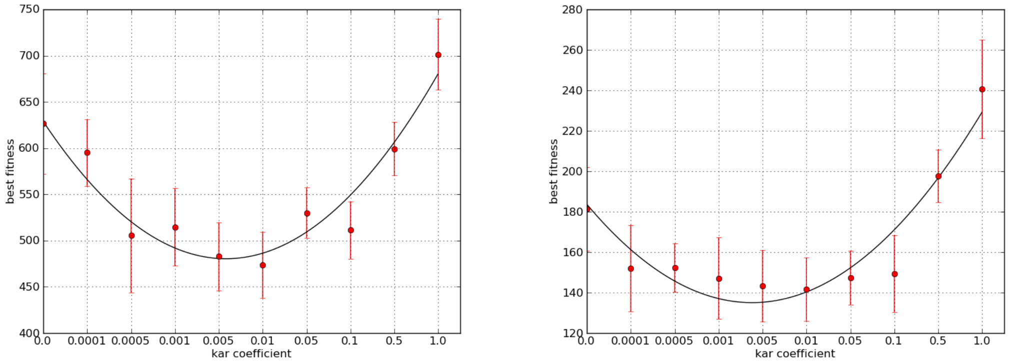


Figure 3.4: *Result on multidimensional Rastrigin function using different velocity thresholds for PSO-KaR optimization. The final fitness achieved with different threshold values is shown for runs on the 100D Rastrigin (left) and 50D Rastrigin (right). Error bars indicate the standard deviation over 10 runs. In both cases, a too high thresholds introduces too much noise, and a too low one does not perturb the swarm sufficiently. Best results are obtained, in both cases, with a velocity threshold equal to 0.01.*

3.4 Macromolecular assembly

Proteins have the ability to assemble in multimeric states to perform their specific biological function. As the native folded state of single proteins, the most stable conformation of these macromolecular assemblies corresponds to an arrangement that is unique. Unfortunately, it is usually difficult to characterize the structure of multimeric assemblies at atomistic resolution. This is due to both their size and complexity, which make the production of sufficiently pure crystal for X-ray crystallography challenging. Moreover, if the assembly is amphipathic such as in the case of transmembrane assemblies, crystallization is even more difficult. For this reason, *in silico* methodologies aiming at predicting how multiple protein copies arrange to form a multimeric complex would be desirable. In principle, knowing the structure of an individual subunit should be sufficient to reconstruct the structure of the whole assembly. Importantly, by exploiting the fact that multimers often respect a certain symmetry, the reconstruction process can be simplified. However, proteins are intrinsically dynamic objects and can often undergo conformational changes when multimerizing, contributing to make the prediction of a macromolecular assembly structure an extremely challenging task.

The prediction of protein assembly can be interpreted as a minimization problem having an extremely large and complex search space. To tackle it, several solutions have been proposed to date. Some approaches, as in SymmDock [126], M-ZDOCK [117], or MolFit [14], reduce the search space by imposing a specific symmetry and subsequently rigidly dock the binding partners so that a predefined energy function is minimized. All these schemes are *ab initio*, i.e. they do not exploit any previous knowledge about the system being studied. Still, it is important to point out that, while producing an X-ray structure of a macromolecular assembly is challenging, low resolution data are usually more accessible. Thus, these structural information can provide important geometric restraints that the final assembly should respect. One of the major efforts in this context is represented by IMP (Integrative Modeling Platform), which is able to deal with a variety of experimental restraints and predict very large macromolecular assemblies via a Monte-Carlo and Conjugate Gradient search [123]. Successful examples of this approach are represented by models of the Nuclear Pore Complex [9] and the 26S Proteasome [76]. Another Monte-Carlo based approach, Rosetta, has been shown to precisely predict the multimeric arrangement according to several symmetries, keeping into account both backbone and sidechain flexibility via a mutistep refinement procedure [10]. In order to better reproduce protein flexibility, some programs resort to molecular simulations. This is the case of HADDOCK [34], which can assemble up to six monomers according to given experimental constraints by first docking them rigidly, and subsequently refine them via simulated annealing. Differently, instead of directly performing molecular simulations, 3D-DOCK exploits the relative ensemble of produced structures. Finally, given a good structural starting point, the molecular dynamics-based MDFF protocol [135] can flexibly dock and refine monomers inside a cryo-EM map of their multimeric structure.

Our aim is to predict multimeric arrangements using the structural and dynamic information of the monomeric state and low-resolution spatial restraints for the final assembly. Our approach exploits specific modules implemented for POW to model likely assemblies so that protein's natural flexibility is taken into account, and that known experimental constraints are respected. PSO has been already used successfully for docking small molecules in protein's active sites. Examples of such an approach are pso@autodock [97] and ParaDocks [89]. In our knowledge, however, our implementation is the first example of PSO applied to protein-protein interaction.

Two modules are available for the prediction of protein-protein interactions: `ProteinProtein` (dealing with heterodimers) and `DockSymmCircle` (dealing with homo multimers assembled according to a circular symmetry). Image 3.5 represents the workflow adopted in the `DockSymmCircle` for the prediction of protein-protein assembly on the base of a given symmetry and experimental constraints.

3.4.1 Search Space Definition and Data Manipulation

In the module `ProteinProtein`, one of the two proteins (the receptor) is kept fixed, whereas the other (the ligand) is freely displaced and rotated. The search space is therefore six-dimensional (ligand translation and rotation around the x, y and z axis). When assembling a dimer, the ligand is first translated at the origin, and rotated around three rotation angles

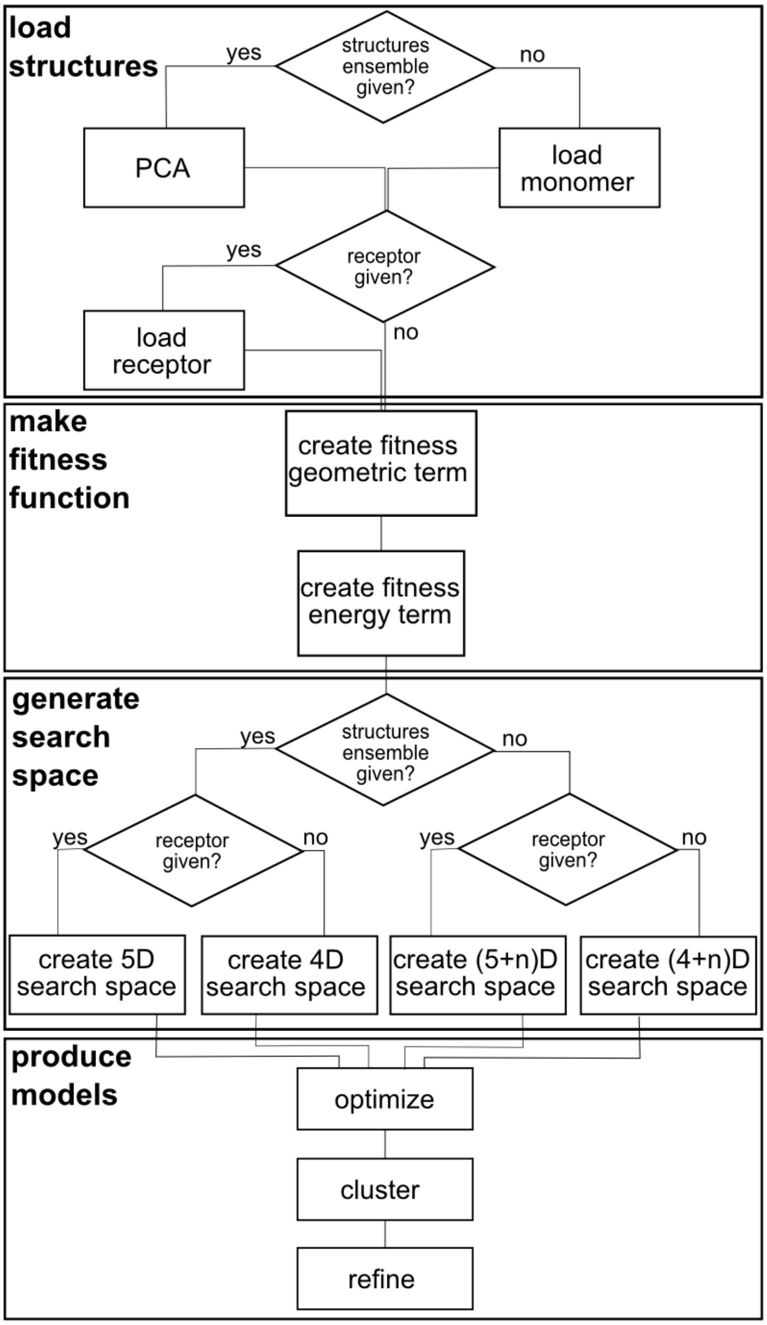


Figure 3.5: *POW pipeline as implemented in the `DockSymmCircle` module.*

(α, β, γ) according to the following transformation matrix:

$$\begin{aligned} R &= R_z(\gamma) \cdot R_y(\beta) \cdot R_x(\alpha) && (3.3) \\ &= \begin{pmatrix} cos(\gamma) & -sin(\gamma) & 0 \\ sin(\gamma) & cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} cos(\beta) & 0 & sin(\beta) \\ 0 & 1 & 0 \\ -sin(\beta) & 0 & cos(\beta) \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\alpha) & -sin(\alpha) \\ to0 & sin(\alpha) & cos(\alpha) \end{pmatrix} && (3.4) \end{aligned}$$

The center of mass of the rotated ligand is subsequently translated to a desired (x, y, z) coordinate. It has to be pointed out that both ligand and receptor are in this case treated as rigid bodies.

If an ensemble of ligand structures is available (typically obtained from a MD trajectory), flexibility can be introduced as set of further dimensions in the search space. To do so, Principal Component Analysis (PCA) is initially performed on the ensemble. Let T a matrix $3n$ by s containing the structures ensemble, where n is the number of atoms in the system and s the number of different structures. Let $C = cov(T)$ the $3n$ by $3n$ covariance matrix of structures ensemble T. Compute the C matrix diagonalization:

$$E^{-1}CE = V \tag{3.5}$$

Matrix E will contain $3n$ eigenvectors, matrix V is a diagonal matrix containing the corresponding eigenvalues, ranked from the biggest to the smallest. Subsequently, the eigenvector's cumulative energy g is computed:

$$g(m) = \sum_{q=0}^{m} V(q,q), \forall m \in [0, 3n] \subset \mathbb{N} \tag{3.6}$$

The minimal number of eigenvectors m_{min} representing a user defined percentage of protein's motion p is then extracted by selecting the first m respecting the condition:

$$\frac{g(m)}{g(3n)} \geq p \tag{3.7}$$

Finally, the projection value P of every structure in the ensemble along the first m_{min} eigenvectors is computed:

$$\text{P}= E(0 \leq m \leq m_{min}, 3n)^T \dot{T} \tag{3.8}$$

The obtained projection values P, also called fluctuations, will be used as an index on the structures ensembles, which we can consider as a conformation database. The search space is therefore characterized by three rotations, three translations, and n fluctuations. In order to produce a complete assembly, the ligand in the database having its eigenvector projection being the closest to the desired fluctuation values is first extracted. Subsequently, the rigid geometric operations described in previous paragraph are performed. The main advantage of using a MD trajectory with this approach is that the protein conformations used to assemble the multimer will respect protein's natural (and physically plausible) flexibility.

In order to predict a circularly symmetric assembly according to given stoichiometry and experimental constraints, the module `DockSymmCircle` is available (see workflow on Figure 3.5). The conformational space of rigid assemblies having a circular symmetry is defined by the three rotation angles (α, β, γ) of a single monomer with respect of a center of symmetry aligned along the z axis, and a displacement r with respect to it, which represents the radius of the assembly in its narrowest point. In detail, the assembly of a single multimer is performed as follows. First, the monomer's center of mass is moved at origin, and rotated according to the transformation matrix R shown in equation 3.3. Via the computation of a bounding box, the atom located the furthest along the positive direction of the x coordinate is identified. The protein is then moved so that the found atom will be located at the origin. After these operations, the monomer will be placed in a position such that the point of symmetry will be located at a coordinate $(r, 0, 0)$. A number of monomers equivalent to the desired stoichiometry is subsequently generated and displaced so that a circle is equally partitioned. A symmetrical structure having the desired radius around the z axis is therefore generated. It has to be pointed out that this docking is based on a purely geometric assembly of rigid bodies.

If an ensemble of protein structures is available, flexible assembly can be performed. This will add additional dimensions in the search space as previously described for the `ProteinProtein` module. MD-based flexible docking around a rigid receptor can be also performed. In this case two additional degrees of freedom, i.e. rotation and translation of the whole assembly along the z axis, are also kept into account.

3.4.2 Fitness Function

Both `ProteinProtein` and `DockSymmCircle` modules adopt the same fitness function evaluation procedure. This function, which defines an assembly quality, depends on two factors: geometry and energy. For geometric contribution, specific measures of the current multimer are compared to values being experimentally known. The aim is to minimize the difference within obtained multimer and desired measures. Target measures t can be as diverse as assembly width or height obtained from cryo-EM maps to atomic distances obtained with FRET or cross-linking experiments. Measures to be performed on a multimer m are user provided via a function $c(m)$. The geometric score of a specific multimer $G(m)$ is determined by the euclidean distance within obtained and target measures as follows:

$$G(m) = \sqrt{(t - c(m)) \cdot (t - c(m))} \tag{3.9}$$

In order to avoid clashes, an energy contribution is also performed. Two energy functions are available. The first is "minimalistic", it is indeed just constituted by a 9-6 Lennard-Jones potential within all the C_α and C_β atoms of two neighboring monomers extracted from the assembly:

$$E(m) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^9 - \left(\frac{\sigma}{r}\right)^6 \right] \tag{3.10}$$

Where r are all the distances within couples of atoms being at a distance smaller than 12 Å , $\epsilon = 1$ and $\sigma = 4.7$. The values of these constants correspond to a the Coarse Grain parameterization for C_α atoms in Martini force field [88].

A second energy function, more precise but also more computationally expensive, is available when just two binding partners (ligand and receptor) are docked. This is based on the estimation of the molecular mechanics contribution to the binding free energy of two monomers via a Coarse-Grained representation of the complex. For every structure attempted by PSO, the total CG non-bonded energy (Coulomb and van der Waals) of the complex is first estimated after 50 steepest descend minimization steps. The two binding partners are subsequently extracted from the minimized complex, and their non-bonded energy computed via a simple single point. The energy functions is finally estimated via the following equation:

$$E(m) = (E^{complex}_{Coulomb} + E^{complex}_{VdW}) - (E^{ligand}_{Coulomb} + E^{ligand}_{VdW}) - (E^{receptor}_{Coulomb} + E^{receptor}_{VdW}) \tag{3.11}$$

Estimation of the coarse-grained representation of the binding partners is performed via a VMD [54] tcl script, whereas the computation of systems non-bonded energy is performed via a LAMMPS [118] molecular dynamics engine run. The LAMMPS executable has been customized in our laboratory.

The final fitness function f is computed by means of the following weighted sum:

$$f(m) = c * E(m) + (1 - c) * G(m) \tag{3.12}$$

where c is a real value within 0 and 1 (by default set to 0.2). After preliminary tests, we found that results are not sensitive to variations of this value. The rough energy function in equation 3.10 only avoids clashed of subunits, and at the current stage is not sufficiently precise to allow a blind docking, i.e. a docking where no geometric restraints are provided. However, work in the development of more accurate energy functions to be included in the fitness function is currently ongoing. We expect this will enhance the capabilities for the broad problem of protein-protein recognition.

3.4.3 Postprocessing

All fitness evaluations obtained during PSO are collected, and solutions having a fitness lower than a predefined threshold are retained. Since several solutions usually represent similar conformations, clustering is subsequently performed. Two *ad hoc* clustering approaches able to determine automatically the number of required clusters are available. The first uses solutions distance in search space as clustering criteria, the second uses the RMSD within different solutions. Distance clustering is performed according to the code shown in Algorithm 4. The fitness value associated to a specific centroid fitness is equal to the fitness of the solution being the closest to it. RMSD clustering is almost identical to distance clustering, the clustering criteria being the RMSD within structures instead of distance within solutions. Clusters coordinates are ranked according to their fitness, and returned to the user in an output file. Assemblies corresponding to these solutions are generated as an ensemble of PDB files.

3.4.4 Benchmarks

We benchmarked our method to large protein complexes exploiting information usually accessible, like their stoichiometry and relative produced symmetry. We selected a extensive set of complexes spanning several symmetry classes, namely Acyl Carrier Protein Synthase (C_3 symmetry), Chorismate Mutase (C_3), Lumazine Synthase (C_5), SM Archeal Protein (C_7) and EscJ (C_{24}). It should be noted that some of these proteins have been also used to benchmark

Algorithm 4 clustering of best POW solutions according to their relative distance in the search space

```
$P \leftarrow [p1, p2, ...pn]$
$C \leftarrow []$
$PC \leftarrow [0, 0, ...]$
$clusternb \leftarrow 0$
while True do
    if exist n with PC(n)==0 then
        $cluster \leftarrow cluster + 1$
        $PC(n) \leftarrow cluster$
        $C(cluster) \leftarrow P(n)$
    else
        $break$
    end if
    while True do
        $newaddition \leftarrow False$
        for all m with P(m)==0 do
            if $distance(P(m), C(clusternb)) \leq threshold$ then
                $newaddition \leftarrow True$
                $P_C(m) \leftarrow cluster$
                $pos \leftarrow 0$
                $cnt \leftarrow 0$
                for all k with PC(k)==cluster do
                    $pos \leftarrow pos + P(k)$
                    $cnt \leftarrow cnt + 1$
                end for
                $C(clusternb) \leftarrow pos/cnt$
            end if
        end for
        if $newaddition == False$ then
            $break$
        end if
    end while
end while
```

other methods (e.g. Rosetta's symmetrical predictions [10]). In all the cases, we extracted a single monomer from the available crystal structure, and attempted to reconstruct the original multimer imposing various combinations of experimentally plausible geometric restraints. In order to verify the influence of restraint choice on the final prediction, several combinations of restraints were used. All the tests were performed running 3 PSO repetitions of 200 steps with 80 particles using 4 processors on an Intel AMD64 dual-quad core machine. In postprocessing, structures having a fitness smaller than zero were selected. Such a value indicates that its corresponding structure respects all geometric restraints and does not feature any clashing C_α or C_β. Selected solutions having an RMSD smaller than 1 were subsequently clustered. All reported execution times also include postprocessing (solution filtering, clustering and generation of corresponding pdb files). In the following, all the tests cases are detailed. In the Discussion section 3.4.4, POW results (with relative computation timing) are summarized (Table 3.1). Best structures superposed to the known crystal structure are shown in Figure 3.6. This section is meant to show how POW performs in a variety of situations imposing different stoichiometry and restraints. However, further tests were performed in order to assess the influence of restraints on the overall result. In fact, similar final results could be obtained with a variety of different set of restraints (see Annex A.3).

PhoQ dimer

We first tested our protocol for the general case of protein-protein interactions when some experimental restraints are known and no information about the symmetry of the complex are used. The test case we chose is the periplasmic sensor domain (a.a. 45-186) of PhoQ, a two-component system histidine kinase responsible of detection of divalent cations at the inner bacterial membrane. Two X-ray structures of this domain are however available: one from *Salmonella enterica* (pdb: 1YAX), and one from *Escherichia coli* (pdb: 3BQ8). The crystals being different, Goldberg et al. [46] determined via cross-linking experiments on the dimer's interface (α-helix 45-62) that the structure from *E.coli* shows the most physiologically relevant arrangement. From this complex, two individual monomers were extracted and defined as ligand and receptor respectively, while five cross-linking measures displaying high efficiency (Thr47, Leu51, Gly54, Asn57, Leu58, see Figure 2A) were used as geometric restraints. Since cross-linking efficiency does not provide a high-resolution information about residues distance, we considered that the five most efficient cross-links simply indicate that residues' C_α atoms are 9 ± 3Å apart, consistently with disulfide bond distances. For this reason, we implemented a "soft constraint". Given a multimer m, the simplest example of such a constraint within two atoms $a1$ and $a2$ can be represented with the following equation:

$$c(m) = \begin{cases} 9 & \text{if } 6 < d(a1, a2) < 12 \\ \text{d(a1,a2)} & \text{otherwise} \end{cases} \tag{3.13}$$

where $d(a1, a2)$ is the euclidean distance within atoms $a1$ and $a2$. The target value for this constraint is 9, i.e. the average distance.

Selected structures having a RMSD smaller than 5 Å within themselves were finally clustered. The total execution time was 3 minutes and 32 seconds. 6 different solutions were finally produced, the best one having a backbone RMSD equal to 1.85 Å (Figure 3.6 A). Its interfacing residues, located on α-helix 46-62, had a backbone RMSD of 0.76 Å. Importantly, all the solutions respected the given geometric restraints and presented no steric clashes, possibly being representative conformations of alternative signaling states accessible to this sensor domain.

Chorismate Mutase

Chorismate Mutase is involved in phenylalanine and tyrosine synthesis pathway by converting chorismate to prephenate. Trimeric *Clostridium thermocellum* Chorismate Mutase was crystallized at a resolution of 2.2 Å (pdb: 1XHO) [143]. Chain A was extracted as representative structure. Assembly radius was bound to be within 0 and 3 Å. The height (47 ± 2 Å) and width (49 ± 2 Å) of the final assembly were adopted as unique geometric constraints. The total execution time was 4 minutes and 50 seconds. 8553 solutions were accepted, and subsequently clustered in 22 different structures. The best model, shown in Figure 3.6 B, had a backbone RMSD of 1.52 Å.

Acyl Carrier Protein Syntase

Acyl carrier protein synthase (AcpS) contributes to lipids and fatty acid biosynthesis by catalyzing holo-ACP formation. This protein assembles into a trimer, containing three binding sites at protein interfaces. Trimeric *Streptococcus pneumoniae* AcpS was crystallized with 3'5'-ADP docked in two of its three active sites at a resolution of 1.9 Å (pdb: 1FTH) [27]. Chain A was extracted as representative structure. Assembly radius was bound to be 0 ± 2 Å(a negative radius leads to multimers crossing the center of symmetry). To set a geometric constraint, we first selected two residues from different subunits being part of protein's binding site. ASP10 and HIS105 bind the same phosphate group. For this reason, we set the distance within these two residues to be of 9 ± 2 Å. 11110 solutions accepted, and subsequently clustered in 3 different structures. The best model, shown in Figure 3.6 C, had a backbone RMSD of 1.91 Å.

Lumazine Syntase

Lumazine Synthase, responsible 6,7-dimethyl-8-ribityllumazine condensation in riboflavine synthesis, is present in archea, bacteria, fungi and plants. Circularly symmetric pentameric

Lumazine Synthaze from *Saccharomyces cerevisiae* was crystallized at a resolution of 1.85 Å (pdb: 1EJB) [90]. Since the first 16 N-Terminal residues show a different coil conformation, we excluded them from our test. At this point, we supposed that all the five monomers could be considered as identical. Assembly prediction was rigidly assembling Lumazine Synthaze chain A. In this test, we combined four different geometrical constraints. The first two, height and width, should be equal to 75 ± 2 and 49 ± 2 Å respectively. The two others were defined as distances within atoms. We set the distance within two ASP103 residues facing each other as 3.4 ± 2 Å, and the distance within ASP103 and a neighboring HIS107 as 9.4 ± 2 Å. The function describing these two last constraints had the same form of equation 3.13. Assembly radius was set to 5 ± 1 Å. The total execution time was 3 minutes and 53 seconds. 7269 solutions were accepted, and subsequently clustered in 6 different assemblies. The best solution, shown in Figure 3.6 D, had a backbone RMSD equal to 1.89 Å.

SM Archeal Protein

SM Archeal Protein is part of various mRNA processing machineries, and comprises a central cationic pore. Heptameric *Pyrobaculum aerophilum* SM Archeal Protein was crystallized at a resolution of 1.75 Å (pdb: 1I8F) [101]. We considered all seven monomers as identical, and extracted chain A as a representative monomeric structure. Assembly radius was bound to be 4 ± 2 Å, which is a reasonable size for a cationic channel (diameter within 4 and 12 Å in the narrowest point). Optimization was run on the base of a single, experimentally plausible geometric constraint: we imposed residue D29 to be exposed to the pore lumen. For this reason, ASP29 distance from the pore central axis was set to be smaller than 6 Å (i.e. the largest possible pore radius) using a functional form like equation 3.13.The total execution time was 2 minutes and 30 seconds. 6900 solutions accepted, and subsequently clustered in 6 different structures. For all of these, we computed the backbone RMSD with respect of the crystal within residues 15 to 80 (i.e. the residues common to all the seven subunits in the crystal). The best model, shown in Figure 3.6 E, had a backbone RMSD of 0.95 Å.

EscJ

Escherichia coli 24-mer EscJ is part of type III secretion system's basal body. The protein is exposed to the periplasm, and encircles another24-mer protein, EscD. EscJ had a 4-mer basic unit crystallized (pdb: 1YJ7 [148]). We set the assembly radius to be 38 ± 2 Å and imposed three different restraints. We set assembly's width to be 176 ± 2 Å, its height to 55 ± 2 Å and residue Pro99 to face inwards (distance from the origin smaller than 40, the largest acceptable radius). The total execution time was 8 minutes and 59 seconds. 10176 solutions were accepted, and subsequently clustered in 9 different assemblies. From the obtained 24-mer models predicted by POW we extracted a 4-mer basic unit which was compared to the available crystal. The best model, shown in Figure 3.6 F, had a backbone RMSD of 2.04 Å.

Discussion

We observe that the backbone RMSD of the best results obtained above was in any case very low (always smaller than 2 Å). In general, despite producing clash-free models consistent with imposed restraints, the fitness function in its current (simple) implementation is not able to rank obtained structures. On the other side, this limitation is balanced by the very limited set of solutions produced by our method, which allows a direct (e.g. *in silico* and/or *in vitro*) assessment of the biological significance of the ensemble. It has also to be noted that in this set of tests a single subunit was extracted from the known assembly and used to reconstruct the multimer. Since docking was rigid at this stage, i.e. proteins were not deformed during optimization, part of RMSD difference is explained by the small differences within assemblies' subunits. Execution time is affected by the protein size and the complexity of desired geometric restraints. The largest contributor to execution time is however the postprocessing phase, which is affected by the amount of final solutions that need to be generated. In these examples, a small number of possible assemblies were produced, in most cases in less than five minutes. Larger execution times can, however, be expected when more restraints on bigger systems are added. The number of final solutions is affected by how stringent the defined restraints are. By imposing more severe filtering and the clustering thresholds, a smaller number of solutions can however be obtained.

Protein	symm.	time	solutions	bRMSD
(A) PhoQ	-	3m00s	6	1.85 Å
(B) Chorismate Mutase	C3	2m16s	20	1.52 Å
(C) Acyl Carrier	C3	2m59s	3	1.91 Å
(D) Lumazine Synthaze	C5	3m53s	6	1.89 Å
(E) SM Archeal Protein	C7	2m30s	6	0.95 Å
(F) EscJ	C24	8m52s	9	2.04 Å

Table 3.1: *Summary of POW assembly prediction using various experimental restraints. The number of independent solutions after clustering is indicated. bRMSD refers to protein's backbone best RMSD.*

The examples shown here are to be considered benchmark cases for exploring the general performance of our method. Still, the final assembly of a multimeric system is often unknown, and considering the intrinsic flexibility of the monomeric subunit can be crucial for sampling conformations more favorable to form the final assembly. We applied our method to the prediction of the heptameric conformation of *Aeromonas hydrophila* pore-forming toxin aerolysin. For this protein, flexibility cannot be ignored when attempting to predict its multimeric conformation. In fact, this class of toxins is known to undergo large conformational changes in order to assemble into a transmembrane complex. Details about aerolysin multimer modeling are presented in Sections 4.3 and 4.4.

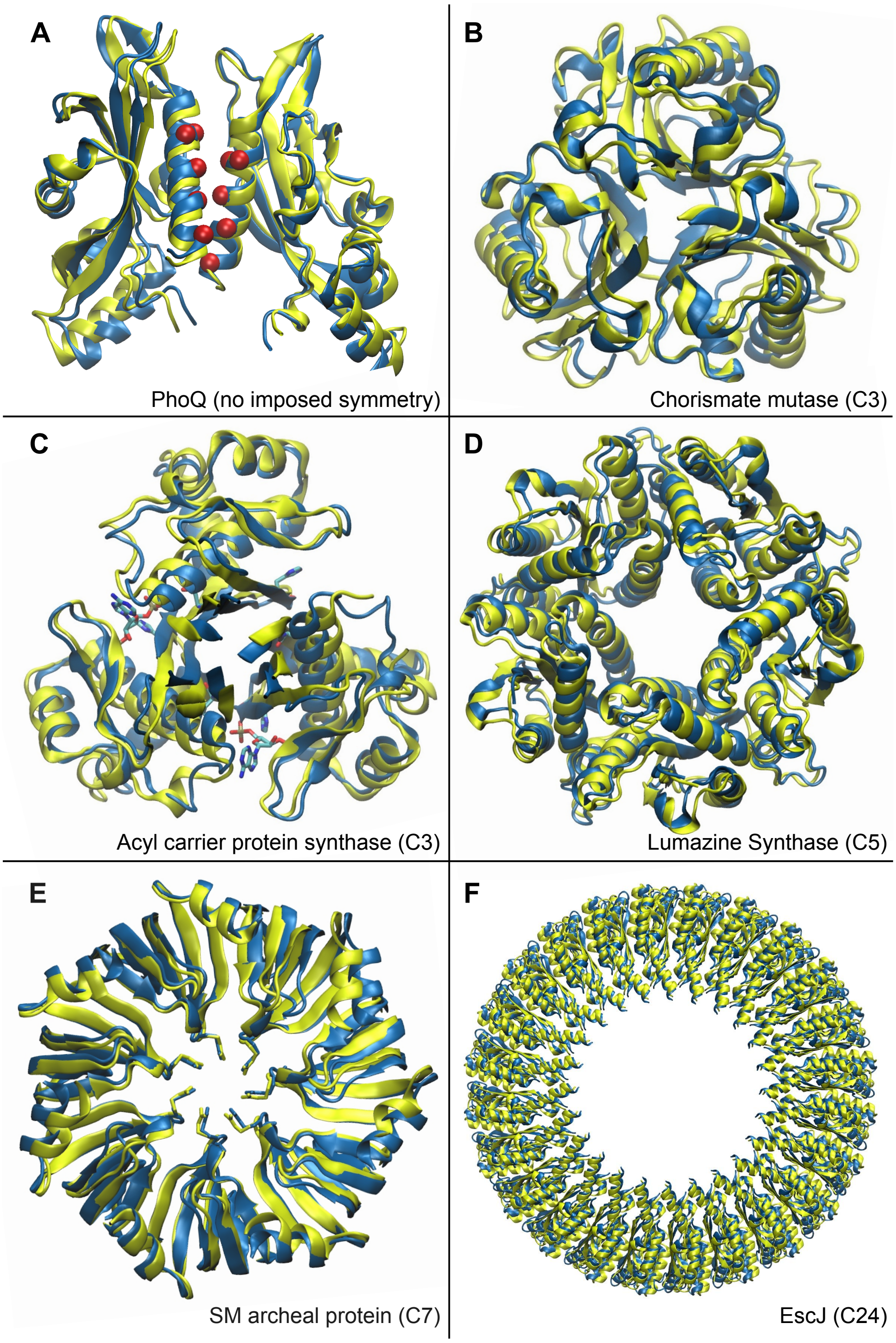


Figure 3.6: *Best symmetrical rigid assembly predictions (in yellow) superposed to known X-ray crystal structures (in blue). In panel A, red spheres highlight the position of residues displaying a high cross-linking efficiency.*

3.5 Parameters fitting for Coarse-Grained force fields in molecular simulations

The computational complexity of a molecular system can be reduced by modeling it via a coarse-grained (CG) representation instead of an all-atom (AA) one. In this representation, atoms are grouped in pseudo-atoms which we call *beads* (see Figure 3.7). A CG model, as an AA one, should describe with a energy potential the way beads interact. Such potential can have a numerical or analytical form. A numerical potential is typically provided in the form of a hashtable: for every distance or angle measure, the corresponding force is returned. In an analytical potential, an appropriate functional form describing the shape of the potential for a type of interaction is defined. This function will depend on a set of parameters which will vary depending on the nature of the beads involved in the interaction.

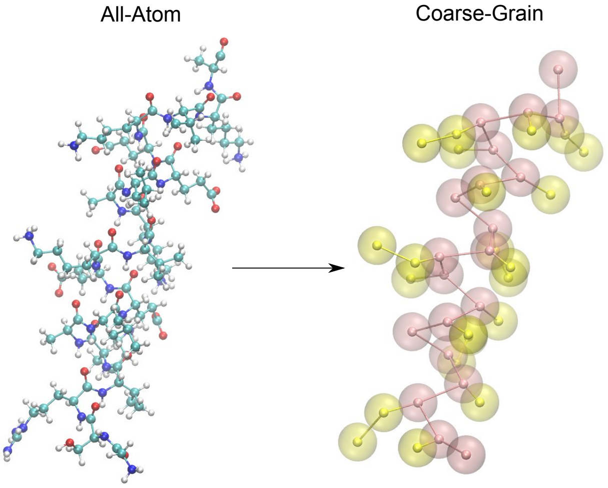


Figure 3.7: *In a coarse-grained representation, atoms are grouped in pseudo-atoms also called "beads".*

The values chosen for these parameters is crucial, since they will determine how accurately the CG potential reproduces the characteristics of an AA simulation. In order to determine these parameters, three main approaches can be identified. In *Force Matching*, one searches for a set of parameters generating forces on CG beads equal to the ones computed via AA simulations. In *Potential Matching* instead, the objective is to match the AA potential. Finally, in *property matching*, the CG force field should reproduce a given macroscopic characteristic computable with a AA simulation or a real experiment. Any physical property can be used, from Radial Distribution Function (RDF) to solvation free energy. Any of these matching approaches should be applied to an ensemble of reference structures, i.e. the matching should be in average good on all the available structures. This is important in order to avoid any bias (overfitting) towards a specific AA configuration.

3.5.1 Search Space Definition and Data Manipulation

The module `CGMatch` is in charge of performing force, potential, and property matching using the molecular dynamics engine LAMMPS [118]. The user should provide an ensemble of CG structures in separate files. For every structure, a second file should contain the value of the desired matching criteria as estimated via an AA simulation. The user can decide which parameters of the force field should be parameterized, and which can be kept fixed. Every parameter having to be optimized constitutes a dimension in the search space. In order to test a set of parameters, a LAMMPS MD simulation has to be launched, and its results parsed and compared with the user provided target values.

3.5.2 Fitness Function

Given an ensemble of CG parameters $\vec{p}$, for any of the mentioned matching methods the fitness function f is defined as an average of euclidean distances as follows:

$$f(\vec{p}) = \frac{1}{d \cdot s} \sum_{n} \sqrt{(AA_n - CG_n(\vec{p})) \cdot (AA_n - CG_n(\vec{p}))} \tag{3.14}$$

Where d is the number of degrees of freedom in the system, n the number of AA structures used for matching, AA_n the target AA value for the n_{th} structure, and $CG_n(\vec{p})$ the measured value on that same structure using a CG force field parameterized with parameters $\vec{p}$.

It is possible to combine several matching criteria via a simple weighted sum:

$$f(\vec{p}) = \sum_{k} a_k \cdot f_k(\vec{p}) \tag{3.15}$$

where a_k are user-defined weighting coefficients, and $f_k(\vec{p})$ any fitness criteria.

This tool constitutes an easy and effective way to parameterize the forces acting on a protein in a CG representation, so that the error with respect of a atomistic simulation is minimized. Other lab members are currently using this tool in order to develop a CG force field for proteins.

3.6 Conclusion

We have implemented Parallel Optimization Workbench (POW), a flexible and easily customizable optimization environment allowing the user to deal with virtually any optimization problem. POW was successfully used for function minimization, docking two given binding partners, assembling homo multimers according to a predefined circular symmetry and All-Atom to Coarse-Grained force/potential/property matching. Within the additional applications we foresee, we find the the molecular mechanics parameterization of molecules starting from a quantum mechanics description. In fact, when minimizing a structure via QM calculations, one looks for a combination of bonds, angles and dihedrals leading to a structure having a minimal energy. This is usually done via a gradient descent. In this context, we believe that POW could peform this minimization more efficiently.

We proposed a variation of the classical PSO optimization, the PSO "Kick and Reseed" method (PSO-KaR). We show that our approach improves the performance of PSO search when the the fitness function becomes rough and highly dimensional. PSO was chosen given both its reported good performances compared to other optimizators, as well as for its simplicity in implementation. POW architecture allows however the insertion of other optimizators in the pipeline. New optimizers are constantly being developed, and some are reported to perform extraordinarily well. Within these we can mention the Firefly algorithm [144, 145] (shown to be even more efficient than PSO [82]), Cuckoo search [146] and Viability Evolution (ViE) [42]. Interestingly, ViE also offers the possibility of dealing with protein-protein geometric constraints in a more elegant and natural way: instead of being added to the fitness function via a weighted sum, constraints determine the possibility of survival of a given solution.

We showed that our protocol based on swarm-intelligence optimization is able to predict, exploiting a limited set of low-resolution experimental spatial restraints, the conformation of homo-multimers, according to a predefined circular symmetry, or general protein-protein complexes. The major strength of our algorithm is that it can quickly return a small set of possible structures respecting the imposed restraints and presenting no severe clashes. These models can be readily used to generate further working hypotheses on the biological function of the assembly and to steadily advance toward the resolution of high-resolution structures.

POW predictions can be guided by virtually any possible experimental results providing insights about proteins structure. Both macroscopic measures (such as height and width extracted from cryo-EM or SAXS experiments) and atomic distances (obtained for instance from cross-linking experiments or FRET measurements) can effectively lead to a correct assembly prediction. At present, cryo-EM maps provide to our protocol only information about the assembly general shapes (height, width, concavity, pore radius, etc.). PSO can, however, deal with any kind of fitness function, thus a natural extension of our protocol will be the direct assembly a multimer into a provided electron density map.

Our method for assembly prediction can also take native protein flexibility into account by automatically extracting relevant conformation using a principal components analysis of a provided structural ensemble. Using aerolysin and YscD from *Yersinia* injectisome as real test cases (see Chapters 4 and 5, respectively), we demonstrated indeed that exploiting a structural ensemble generated via MD simulations greatly improved the prediction outcome. As a matter of fact, such an ensemble could be generated via any other technique allowing the exploration of a macromolecule's conformational space. In this context, one of the possible extensions is the possibility of automatically performing a normal modes analysis on a single provided structure. Exploration of the conformational space constituted by a linear combination of the main discovered modes would subsequently take place. The immediate advantage of this approach is clearly its extremely affordable computational cost, while, on the other side, flexibility extracted from an harmonic approximation of protein molecular interactions is limited to a specific equilibrium state and cannot access further conformational arrangements.

It has to be pointed out that our approach is based on a purely geometric optimization, sidechain arrangement is therefore not refined at the moment. Our aim was at this stage, in fact, to quickly generate a small ensemble of reasonable protein arrangements. Even though we do not produce a refined *de novo* prediction, aerolysin's example shows that our results can provide important insights into multimeric arrangement and guide the design of new experiments. The current energetic contribution to the fitness function is simply constituted by a coarse potential on the protein scaffold used to avoid steric clashes. Energy scoring based on the contribution of the monomer-to-monomer binding energy via a molecular mechanics and/or coarse-grained representation of the complex would likely help to fish for the best solutions and to address completely experimentally blind assembly searches. The goal is to obtain a lightweight but more accurate energetic contribution, better describing van der Waals interactions and accounting for electrostatic contributions. On this respect, taking care of sidechains refinement could also lead to significant improvements.

We believe we just scratched the surface of the capabilities of this novel approach. In fact, since the PSO engine is not sensitive to the kind of imposed symmetry, implementing in POW other common symmetries (such as helical or icosaedral) is trivial, and is part of our future development plans. Moreover, improvements on (i) the fitness function with the inclusion of a broader set of geometric restraints, and/or (ii) the energy scoring with the use of more accurate molecular mechanics potentials will certainly contribute to enhance the quality of the final predicted assemblies and eventually address the prediction of protein-protein interactions in large macromolecular networks.

4 Pore-forming toxin aerolysin from *Aeromonas hydrophila*

4.1 Biological Background

Pore forming toxins (PFT) are a complex virulence factor common to several bacterial families. These proteins are secreted as water-soluble but, upon binding to specific receptors on the target cell's surface, they transform into a transmembrane complex via oligomerization and important conformational changes. This results in the creation of a pore, which induces a reactions cascade usually leading to membrane lysis. A schematics of the complete mode of action of a typical PFT is shown in Figure 4.1.

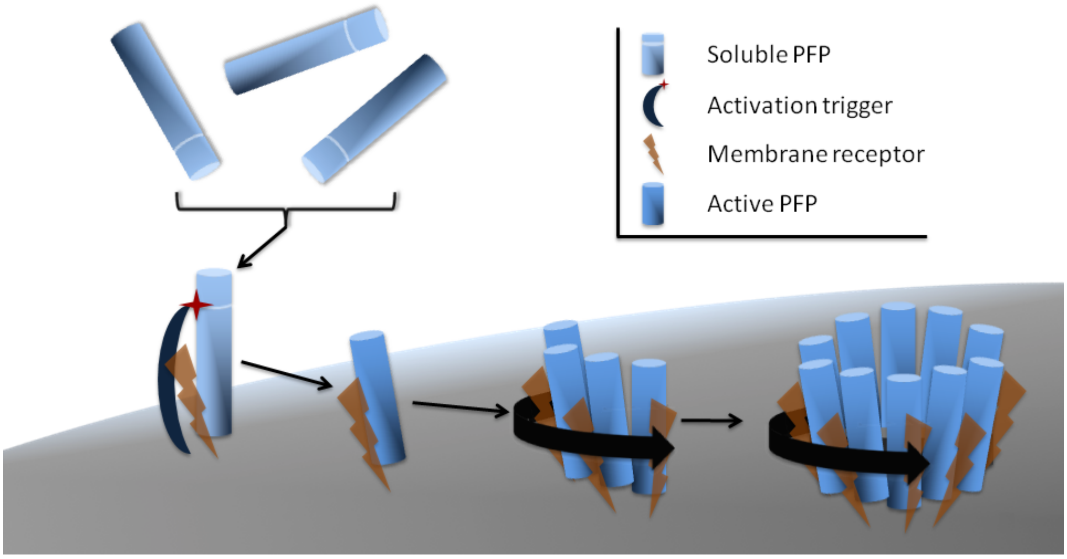


Figure 4.1: *PFT mode of action. The protein is secreted as as soluble monomer which, upon binding to specific receptors on the target cell, gets activated by cleavage of a specific region. Activated toxins multimerize according to a circular symmetry and insert into the lipid bilayer.*

Depending on the secondary structure of their transmembrane element PFT are classified as

α or β. The number of proteins participating to oligomerization, as well as the dimensions of the produced pore, can vary. In particular, β-PFTs span from heptameric complexes punching 2 nm holes, up to 50 nm pores produced by the assembly of around 50 monomers. The latter pores are produced by members of the largest family of PFT, the Cholesterol-Dependent Cytolysins (CDC), and allow the transit of fully folded proteins.

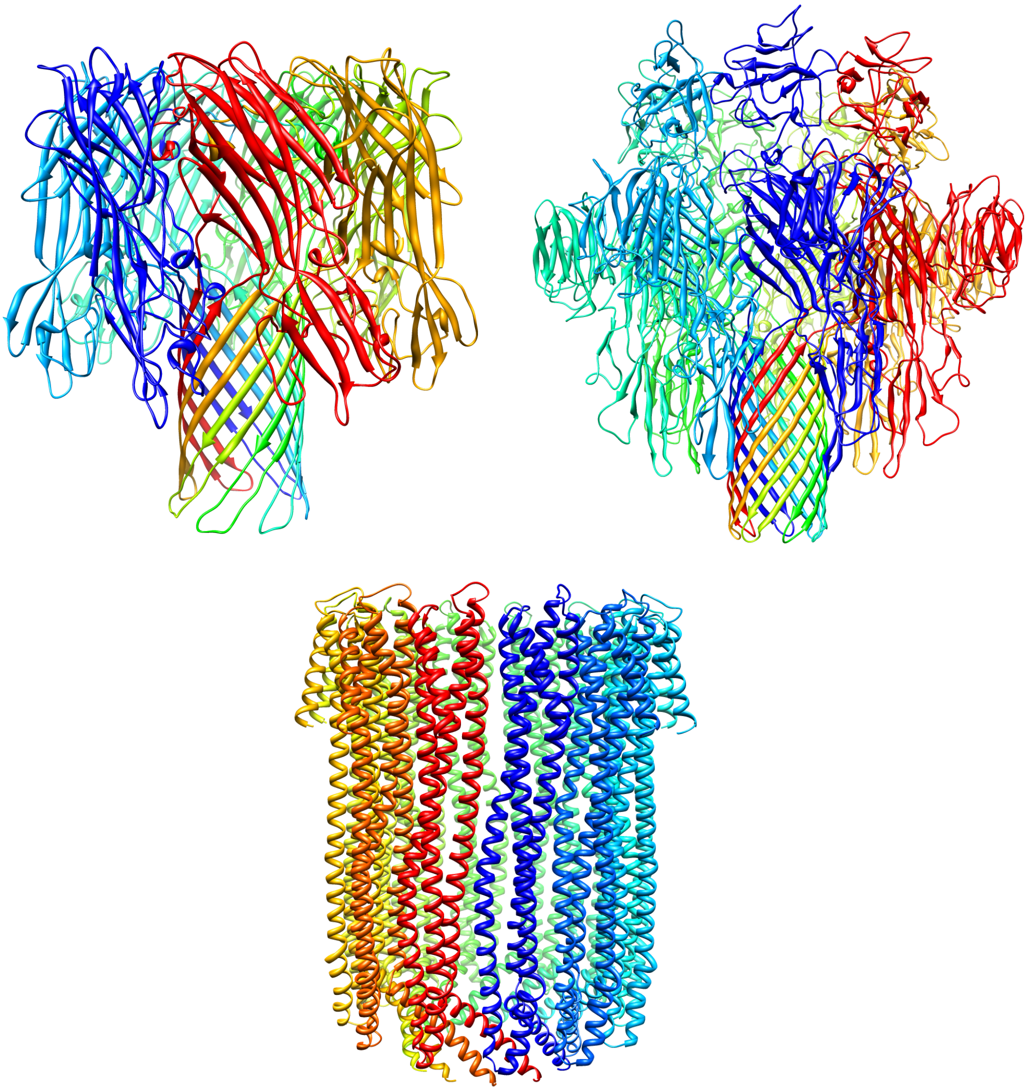


Figure 4.2: *The only Pore-Forming Toxins having been crystallized in both monomeric and membrane-inserted conformation are nowadays the β-PFTs α-hemolysin from* Staphilococcus aureus *(top left) and Cytolysin A from* Salmonella enterica *(top right), and the α-PFT Cytolysin from* Vibrio choleare *(bottom).*

Owing to the amphipathic nature of these assemblies, obtaining a crystal structure of their membrane-inserted state is particularly challenging. Nowadays, just three PFT have been crystallized in a multimeric state: the dodecameric α-PFT Cytolysin A from *Salmonella enterica*

[99], and the heptameric β-PFTs α-hemolysin from *Staphilococcus aureus* [131] and Cytolysin from *Vibrio choleare* [33] (see Figure 4.2). Interestingly, the conformations of α-hemolysin and Cytolysin show that the transmembrane element is a β-barrel constituted of seven β-hairpins. It can be noticed that this barrel can form upon an important conformational change in every monomer, extracting the β hairpin region from its original position.

Aerolysin is the main β-PFT secreted by several *Aeromonas* bacteria, which are involved in food-borne infections such as gastroenteritis. The protein, structurally related to α-hemolysin, is structured in two lobes. As indicated in Figure 4.3, four domains have been defined [110]. The smaller lobe has been defined as Domain 1, whereas the bigger lobe has been divided in 3 functionally distinct domains. An abundance of β-sheets can be noticed. One of those is particularly long, around 93 Å, and spans the whole major lobe while also undergoing a 180 degrees twist.

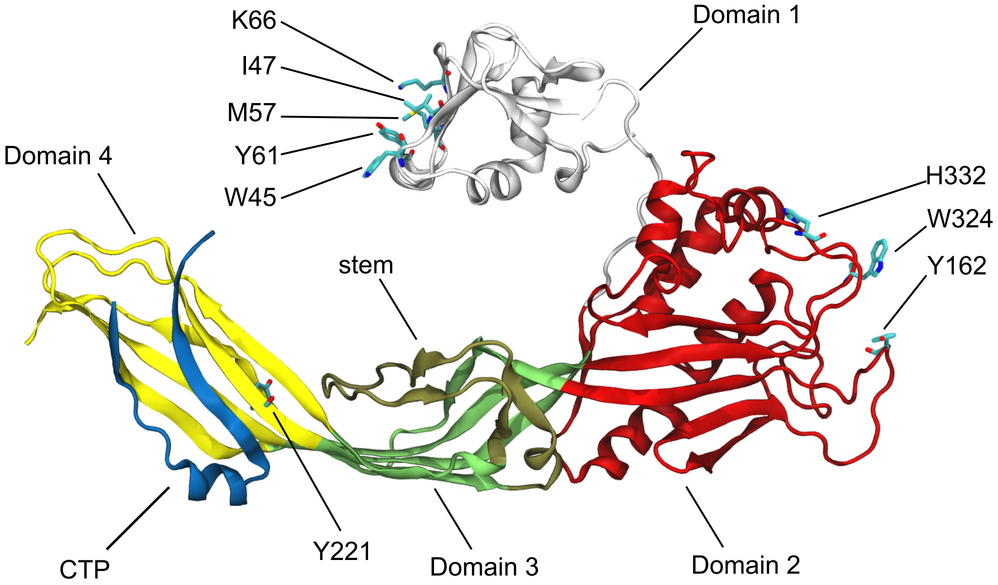


Figure 4.3: *Aerolysin monomer. Docking to GPI anchor owns to Domains 1 (cyan) and 2 (red). Residues having a role in binding are depicted in licorice (W45, I47, M57, Y61, K66 on Domain 1 and Y162, W324, H332 on Domain2). Domain 3's (green) loop region (tan) is responsible of membrane insertion. Domain 4 (yellow) is connected to the C-Terminal Peptide (CTP, in blue) via a loop (dashed blue line) being too flexible to be crystallized. CTP cleavage leads to aerolysin activation.*

Aerolysin is released via Type II secretion system as a soluble precursor, Proaerolysin (see Figure 4.3). While at high concentration aerolysin forms dimers, at physiological concentration aerolysin is monomeric [40]. Monomeric Proaerolysin binds to high-affinity receptor Glycosilphosphatidilinositol (GPI) anchor on the target cell [4]. It has been shown that proaerolysin can bind to GPI-anchors only if an anchored protein is present [6]. Proaerolysin binds to a large variety of GPI-anchored proteins, such as thy-1, contactin, N-CAM120, semaphorin 7 and CD14 [35, 103, 41]. At present just two exceptions are known, CD59 and gp63 [6]. Two binding regions have been detected, located on Domain 1 and Domain 2 respectively [85] (see Figure 4.3). Domain 2 binds with high-affinity to GPI-anchor's conserved core while, since all proteins mentioned above lack of sequence homology but are all glycosylated, Domain 1 binds to these N-linked sugars [6]. Several reasons could explain why aerolysin fails to bind to CD59 and pg63, one of which could be that sugars on GPI-anchored protein must be placed in a favorable position to allow subsequent binding of Domain 2 to GPI-anchor's glycan core.

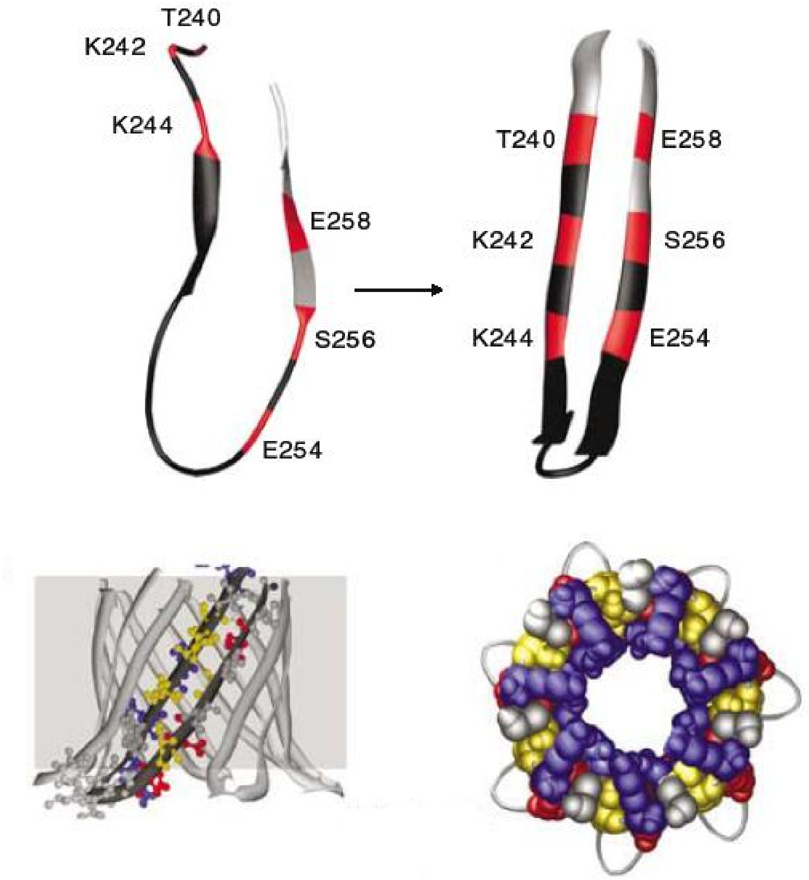


Figure 4.4: *On top, loop region switching from water-soluble to membrane inserted conformation. Residues realign in order to form a β-hairpin having an alternating hydrophobic-hydrophilic pattern. On bottom images, model of the heptameric β-barrel (images from [57]).*

Upon binding, Proaerolysin is processed into Aerolysin either by soluble digestive enzymes or by transmembrane protein Furin [3]. This results in a cleavage of the 40 residues C-terminal peptide located on the tip of Domain 4 (see Figure 4.3, in blue. In the following of this report we will address to this region as *CTP*). Mature Aerolysin undergoes circular oligomerization into a heptameric complex, which is its channel forming configuration. This process is favored by transient associations with cholesterol-rich micro-domains (so called lipid rafts), leading to local increases in toxins concentration [5]. The respective position of the seven proteins forming the heptamer is still subject of debate. Nevertheless, the transmembrane element has been identified in a β-barrel constituted by seven loop regions of Domain 3 [57]. As shown in Figure 4.4, it has been proved that a rearrangement of the loop region gives rise to a β-hairpin having a characteristic alternating hydrophobic-hydrophilic pattern.

Upon oligomerization, a conformational change is believed to extract the loops from their original positions, so that they point the underlying membrane. Heptamer is subsequently believed to insert spontaneously into the bilayer with a mechanism that is still poorly understood. Given its hydrophobic nature, after insertion the tip of the hairpin would fold back to the bilayer's core in a rivet-like fashion, anchoring this way the whole transmembrane barrel to the membrane [57]. The resulting assembly extraordinarily stable, withstanding in fact proteolysis at high temperatures, incubation with 8M urea or 1% SDS [77].

Channel formation of plasma membrane leads to a selective permeabilization to small ions. It has been observed that in the presence of Aerolysin, membrane permeability to calcium ions is increased, and a depolarization of the cellular membrane occurs [70]. Subsequently various cellular responses can be triggered, such as release of calcium from the Endoplasmic Reticulum (ER) in granulocytes, apoptosis in T-cells or vacuolation of the ER in epithelial cells.

Interestingly, specific point mutations can have dramatic effects on Aerolysin mode of action. Y221G mutation, for instance, gives rise to a soluble pore, whose cryo-EM map has been obtained [137]. Under the light of the rivet-like model introduced above, it is speculated that this mutation does not insert into the membrane because the β-barrel region is not formed.

By means of *in silico* studies coupled with *in vitro* and *in vivo* experiments (all carried out in the vander Goot lab, EPFL), we study several aspects of aerolysin's structure and function. In detail, the following points will be addressed

- What is the role of the C-Terminal Peptide in protein activation, and what are the consequences of its removal
- Which mechanism triggers the extraction of the transmembrane loop from its resting place, and how mutation Y221G hinders it
- What are the structures, at an atomistic resolution, of aerolysin Y221G and WT in their heptameric conformation
- Which are the properties of the transmembrane β-barrel in terms of accessible area, electrostatics and ion conductance, and how do they compare with other PFTs.

4.2 Role of the C-Terminal Peptide

PFT are produced as soluble proteins that diffuse and bind to target cells via specific receptors. Many subsequently assemble into ring-like structures, undergoing a conformational change with consequent exposure of hydrophobic surfaces. This drives spontaneous membrane insertion, leading to the formation of water filled pores. This peculiarity raises two interesting questions. The first is: since PFTs can adopt two quite different conformations, how is the folding reaction during biogenesis directed towards obtaining the soluble fold? The second question is: what mechanisms prevent pore-formation from occurring in the producing cell? We decided to address these related questions by studying the role of the C-Terminal Peptide (CTP) of aerolysin in its folding and activation process.

Our aim was to address the precise role of the CTP by combining computational techniques with site-directed mutagenesis, structural analysis, and functional assays performed by Van der Goot lab at EPFL. Our collaboration reveals that the CTP drives the protein into the soluble state during biogenesis, protecting proaerolysin from aggregation possibly by promoting folding, a quite unexpected observation considering the C-terminal location of the peptide. Interestingly, mutagenesis of specific residues in the CTP not only affected the efficiency of proaerolysin folding both *in vitro* and *in vivo*, but also reduced the capacity of the CTP to prevent premature assembly of the heptamer, highlighting the dual role of the CTP in both preventing aggregation of the newly synthesized protein possibly by assisting folding, and controlling the quaternary assembly of the active complex. In the following sections, our computational results as well as their relationship with *in vitro* and *in vivo* experiments will be described [56].

4.2.1 Covalent bonding is not required for binding of the CTP to aerolysin

To characterize the molecular interactions between the CTP and Domain 4, we performed classical molecular dynamics (MD) simulations. To remain as close as possible to experimental *in vivo* conditions, we performed in MD simulations at room temperature (27 degrees Celsius) and atmospheric pressure (1 atm), and the proteins were solvated by water molecules at physiological salt concentration (0.15 M of NaCl). MD simulations reported here were all based on the X-ray structures of wild type (WT) proaerolysin (for details about the structures used, see Methods 4.2.8). The loop connecting the CTP to the rest of the molecule is however not visible in any reported crystal structure of proaerolysin, probably due to its flexibility. Thus proaerolysin was *de facto* modeled in a situation mimicking a cleaved proaerolysin state (here after termed aerolysin-CTP). During the 200 ns of MD, the CTP remained firmly bound to the protein. Native hydrogen bonds and salt bridges were preserved along the entire trajectory, as were secondary structure elements, both in the CTP and in Domain 4. The mean conservation of secondary structure in the system, i.e. the percentage of residues in a β-sheet conformation along the MD simulation with respect to the initial crystal structure, was 86±5% over the last 100 ns.

Since the MD simulations were performed with the absence of a covalent bond between the CTP and Domain 4, these observations pointed to a strong binding affinity of CTP for Domain 4. By computing electrostatic, van der Waals and solvation contributions to the binding of CTP to Domain 4, we estimated their binding energy to be 115 kcal/mol. Our MD-based observations thus suggest that the CTP remains bound to aerolysin upon proteolytic activation of the protoxin.

This was confirmed using two independent experimental approaches. First, a dimeric X-ray structure of the proteolytically processed form of an aerolysin mutant unable to form heptamers was determined, namely H132N (PDB entry 3G4O). Not only was the observed structure very similar to that of wild-type (WT) proaerolysin (RMSD of 0.74 Å for subunits A and 0.92 Å for subunits B in the dimer), it also contained the CTP, in an essentially identical conformation.

As a second approach to investigate whether the CTP remains bound to the mature toxin following proteolysis, a WT proaerolysin having a six-histidine tag at the C-terminus, i.e. at the end of the CTP was produced. When proaerolysin was incubated with Nickel beads, it remained attached to the beads both before ad after trypsin processing (see Figure 4.5). This shows that the CTP had not been released upon proteolysis. Consistently with the non-covalent interaction between the mature toxin and the CTP, aerolysin could be released from the beads with 4M urea.

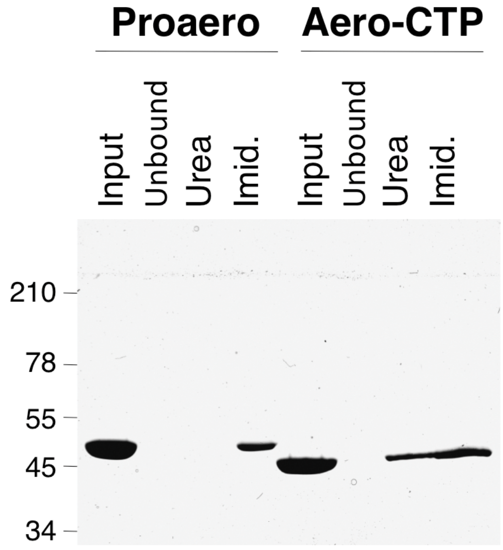


Figure 4.5: *WT proaerolysin harboring a six-histidine tag at the C-terminus was produced. Both proaerolysin and activated aerolysin (proaerolysin processed with trypsin) were incubated with Nickel beads. Both remained attached to the beads and could be eluted with imidazole. Aerolysin only could be released from the beads with urea.*

It had been previously reported that processing of proaerolysin with trypsin leads to the release of the CTP from aerolysin [139]. This conclusion was based on the observation that fluorescence energy transfer was lost between a fluorescent probe, IEADANS, attached to an engineered cysteine on the CTP at position 445 and T203 in Domain 4. Our current findings suggest that the previously observed release of the CTP was artefactually induced by the mutation and/or labeling of the cysteine at position 445. Indeed, in WT proaerolysin, I445 on the CTP is buried within a hydrophobic pocket in Domain 4 and labeling of C445 with the bulky and polar IAEDANS fluorophore must have triggered a severe perturbation at the CTP-Domain 4 interface, leading to premature release of the CTP upon trypsin cleavage (see Figure 4.6).

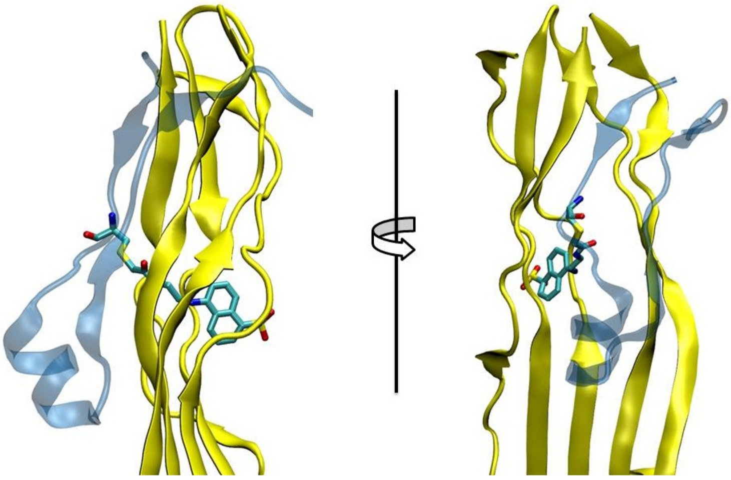


Figure 4.6: *Equilibrated model of IAEDANS tag. The molecule perturbs the interactions within CTP and Domain 4, helping CTP detachment as soon as proteolytic cleavage takes place.*

4.2.2 Identification of key residues for CTP-aerolysin binding

Both X-ray structures of WT proaerolysin and H132N aerolysin-CTP show the presence of a similar complex network of interactions between the CTP and Domain 4 composed of H-bonds (10 in subunit A, and 16 in subunit B), salt bridges (D207 with R442 and K198 with E451), and hydrophobic interactions. To identify key residues responsible for binding of the CTP to Domain 4, we performed in silico alanine scanning on most of the CTP. In silico mutation of a given CTP residue to alanine has the effect of removing most of the native non-bonded interactions (i.e., electrostatic and van der Waals contributions) with the local environment. By comparing the binding free energy of the WT species and its alanine mutant, it is possible to estimate the individual contribution of a given CTP residues to the binding affinity with Domain 4. The greater the variation, the more the residue has a relevant role in the steady binding of the CTP to Domain 4.

As expected, mutation of solvent exposed residues showed little variation in the binding free energy. A low but significant variation was observed for certain polar residues, such as N458, which forms a hydrogen bond with D222, D448, which forms a salt bridge with K198, and

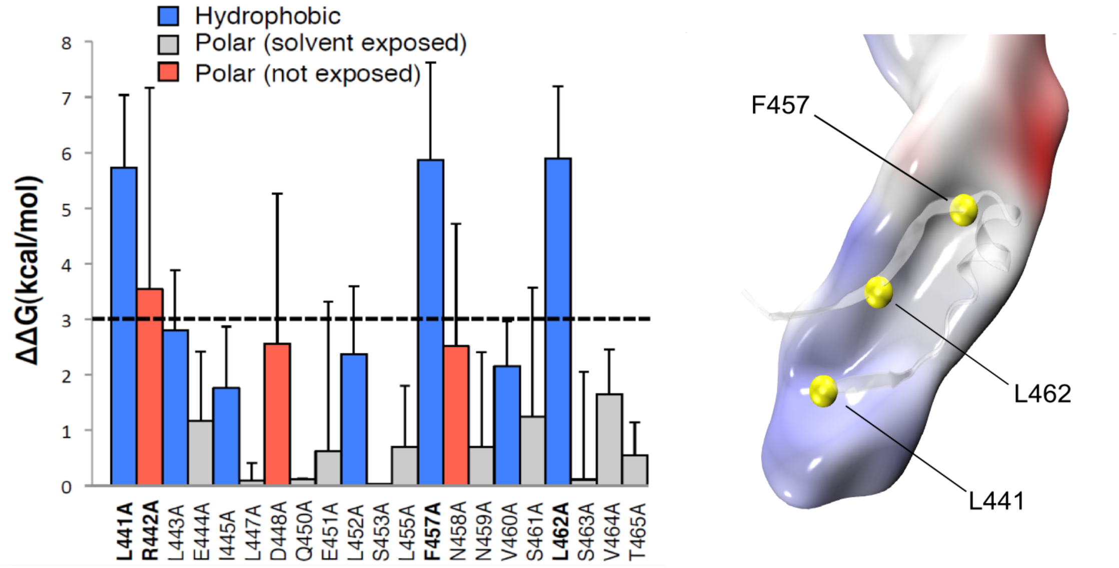


Figure 4.7: *MM/PBSA alanine scanning on CTP residues. Hydrophobic residues L441, F457 and L462 are the main responsible of CTP binding to Domain 4.*

especially R442, which forms a salt bridge with D207. The most dramatic variations in the binding free energy (about 6 kcal/mol) were observed for three hydrophobic residues: L441, F457 and L462. All three residues point inside a hydrophobic pocket in Domain 4 underlying the CTP. More specifically, L441 interacts with V285, A204, P283 on Domain 4 and L443 on the CTP, L462 interacts with V217, L219, and I296 on Domain 4 and I414 and L443 on the CTP, and finally F457 points straight into Domain 4, and is blocked by steric hindrance with V197 and L297 on the Domain 4 and A411 and L452 on the CTP.

4.2.3 *In silico* mutation F457G affects the stability of both CTP and Domain 4

Since F457 on the CTP points straight into Domain 4, we investigated the effect of mutating this residue to glycine *in silico* using an MD setup similar to the one adopted for the WT species. The mean conservation of the secondary structure of the CTP drastically dropped from 76±12% for the wild type to 17±4% for the F457G mutant. The CTP structure remaining after the simulation was a portion of the α-helix, which we determined to be the most stable structural element in an MD simulation of just the CTP in water. Interestingly, the F457G mutation also affected the underlying Domain 4. Indeed, after 100 ns of MD, the mean conservation of the secondary structure of Domain 4 (including CTP) was 86±5% for the wild type and 67±5% for the F457G mutant (see Figure 4.8). By computing electrostatic, van der Waals and solvation contributions to the binding of the mutated CTP (F457G) to domain 4, we estimated their binding free energy to be 75 kcal/mol. This represents a significant reduction with respect to the 115 kcal/mol previously estimated from the aerolysin-CTP MD simulation.

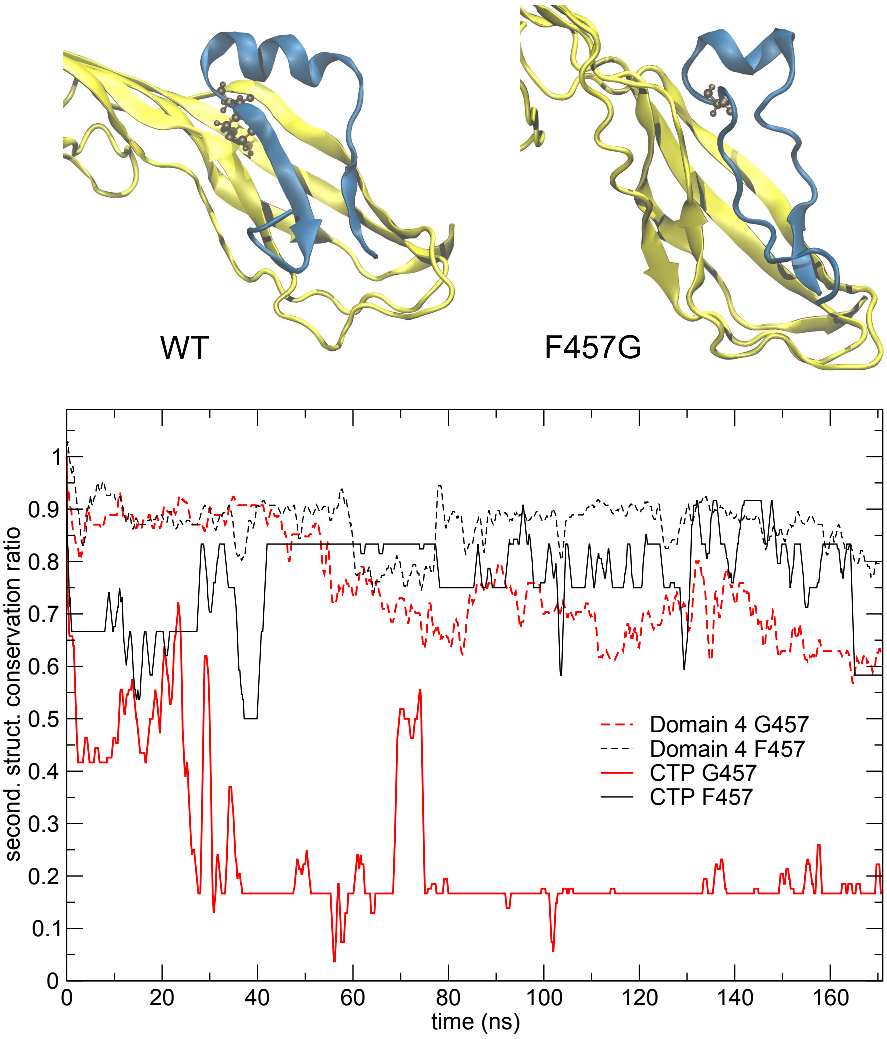


Figure 4.8: *In MD, WT CTP remains steadily connected to Domain 4 (top left). Conversely, F457G CTP quickly loses secondary structures, and starts disconnecting from Domain 4 (top right). CTP unfolding in F457G simulations also affects Domain 4 fold. (bottom) secondary structure evolution of Domain 4 and CTP in F457G and WT simulations.*

4.2.4 *In silico* removal of the CTP affect aerolysin's folding

The effects produced by a mutated CTP on Domain 4 prompted us to compare the structural features of aerolysin with and without its CTP. *In silico*, we removed the CTP from the proaerolysin crystal structure, and 200 ns MD simulation was performed. Simulations performed in the presence and absence of the CTP were subsequently compared. The structural flexibility of each residue was quantified by calculating the root mean square fluctuation (RMSF) of the residue along the MD trajectory. Removing the CTP had no significant effect on the structure of Domain 2 and 3 (Domain 1 was omitted from the simulation since it is known to act as an independent folding unit [77]). In contrast, removal of the CTP led to an average increase of 6.8±3.2 Å of the RMSF for a given residue in Domain 4, suggesting that the CTP stabilizes the structure of Domain 4. The CTP also had an influence on the secondary structure of Domain 4. This was assessed by tracking the percentage of secondary structure conservation along the two simulations, i.e. the percentage of residues adopting a β-sheet conformation in the absence of CTP as compared to crystal structure of proaerolysin. In the absence of the CTP, the secondary structure conservation of Domain 4 was 67±10% , compared to 86±5% in the presence of CTP, and the root mean square deviation (RMSD) after 200 ns of MD was 8.8 Å, compared to 3.6 Å in the presence of CTP (see Figure 4.9). *In silico* removal of the CTP led to the unfolding of the β-strand encompassing residues S272 to S280 in Domain 4.

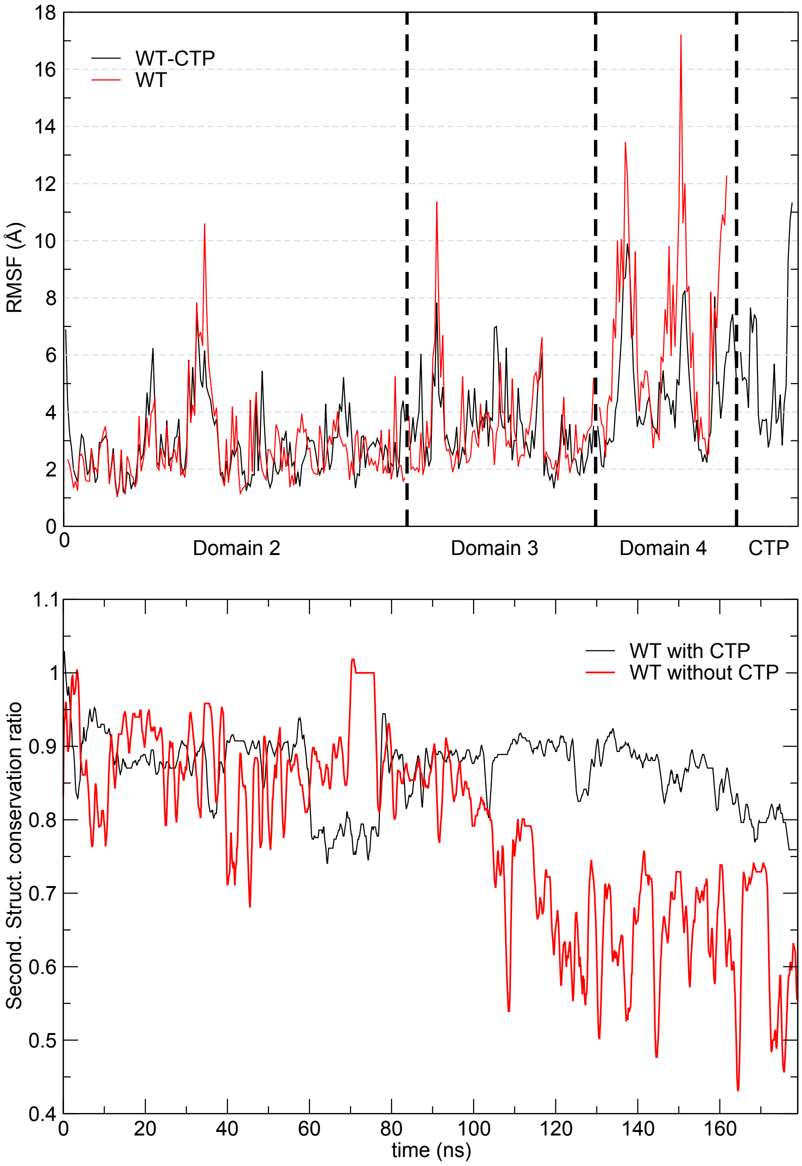


Figure 4.9: *Effects of CTP removal. Without CTP, Domain 4 RMSF increases in Domain 4(top), and unfolding takes place along simulation (bottom).*

Interestingly, a further sequence-based analysis using order prediction algorithms identified the 268-282 segment as the most disordered region of Domain 4 (see Figure 4.10, for algorithms used see Appendix A.4), raising the possibility that the β structure observed for this segment in the proaerolysin crystal structure is in fact imposed by the CTP. It is interesting to note that induced folding of intrinsically unstructured elements often involves hydrophobic, rather than polar, interactions [95] as observed here for the CTP-Domain 4 interface.

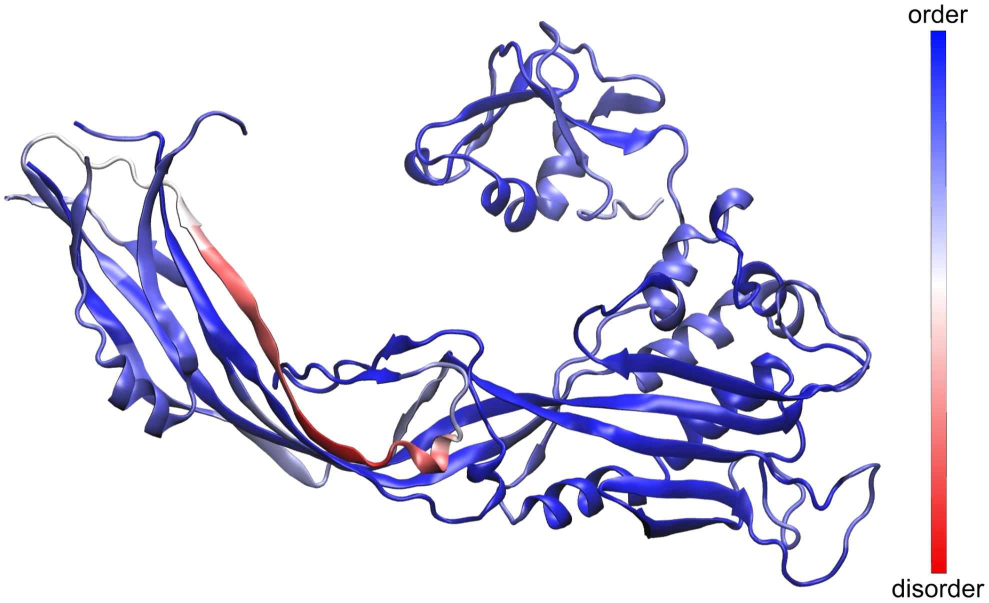


Figure 4.10: *Cumulated results of eight disorder prediction algorithms. The segment 268-282 is consistently predicted as intrinsically disordered. This strand unfold both in F457G mutation and upon CTP removal.*

4.2.5 Mutation or removal of the CTP prevents folding of aerolysin *in vitro* and *in vivo*

In silico alanine scanning analysis predicted that mutation of L441, F457 and L462, and to a lesser extent R442, to alanine should affect binding of the CTP to Domain 4. To test these predictions, constructs expressing these mutants in the *E. coli* periplasm were generated. We also sought a mutation that would affect the secondary structure of the CTP but not the binding. We chose to change S453 to proline since this residue localizes to the middle of the α-helix of the CTP and does not make contacts with Domain 4. In agreement, *in silico* mutation of S453 to alanine did not lead to a significant variation in the binding free energy. Due to the folding of its side chain back onto the protein backbone, proline imposes severe constraints to the backbone geometry leading to helix breaking.

All proaerolysin mutants were detected in bacterial extracts showing that they were synthesized and not degraded to any significant extent (see Figure 4.11). Proaerolysins L441A, R442A and L462A were recovered in significant amounts in the periplasmic fraction. Proaerolysin S453P was barely detectable in the periplasmic fraction, but following purification low amounts of the protein could be obtained. Proaerolysins F457A/G were essentially undetectable in the periplasmic fraction and neither could be recovered following purification on Nickel columns. These observations show that mutating S453P or F457G induced aggregation of proareolysin in the bacterial periplasm, either due to the exposure of a hydrophobic patch or improper folding of part of the protein. The small amounts of toxin that could be purified for all mutants was however properly folded as indicated by the WT-like hemolytic activity of the mutants following trypsin cleavage (see Figure 4.12 A and B).

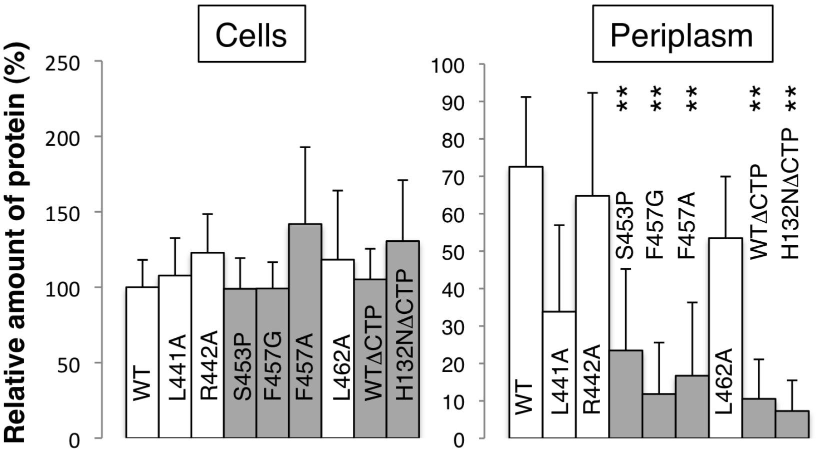


Figure 4.11: *The amount of toxin present in cell extracts as well as in the periplasmic fraction were quantified for 3 independent experiments (n = 3). Error bars represent standard deviations. Proaerolysins F457A/G were essentially undetectable in the periplasmic fraction, and proaerolysin S453P was barely detectable.*

The *in vitro* folding of the mutant proaerolysins was subsequently investigated. For this, proaerolysins, WT and mutants, were unfolded in urea. All four proaerolysin mutants showed very similar urea unfolding curves. Following unfolding in 4M urea, refolding of proaerolysins was triggered by dilution in a urea-free buffer. The efficiency of folding was indirectly monitored by measuring the hemolytic activity of the refolded proaerolysins after proteolysis with trypsin. Hemolysis was followed as a function of time. Under these conditions, refolded L441A and S453P systematically showed a delayed hemolytic activity (see Figure 4.12 C). These experiments reveal that *in vitro* folding into a soluble state of the L441A and S453P proaerolysin mutants was impaired and the extent correlated with the ability of the mutants to fold into a soluble state *in vivo*.

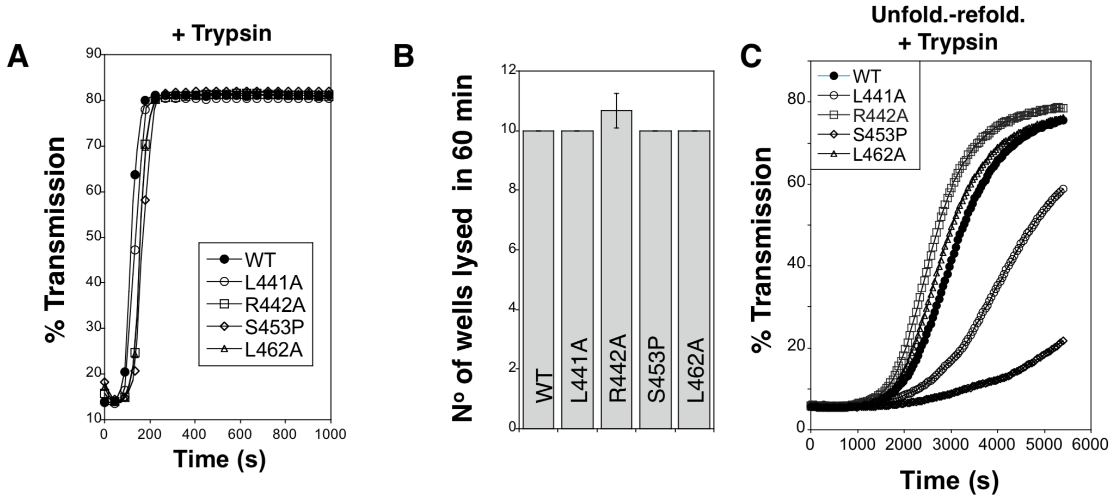


Figure 4.12: *(A) WT and mutant proaerolysins were processed with soluble trypsin, added to erythrocytes and the transmitted light at 600 nm of the sample was followed at room temperature as a function of time. Plots represent the percentage of transmitted light as a function of time. This is a representative experiment out of 4 independent experiments. (B) WT and mutant proaerolysins were processed with soluble trypsin, then subjected to a serial dilution (1/2) in a 96 well plate and incubated with erythrocytes. The number of wells lysed after 60 min at room temperature was determined. Error bars represent standard deviations (n = 3). (C) WT and mutant proaerolysins were unfolded in 4 M urea, then diluted in a urea free medium. After 10 min, samples were treated with trypsin for 10 min. Erythrocytes were added and the transmitted light at 600 nm of the samples was followed at room temperature as a function of time. This is a representative experiment out of 5 independent experiments.*

Since aerolysin without the CTP could not be purified from bacteria, we studied the folding of aerolysin by cleavage of proaerolysin with trypsin followed by unfolding in urea. Unfolding transitions occurred at similar concentrations of chaotropic agents whether proaerolysin was processed by trypsin or not, suggesting proaerolysin and aerolysin-CTP have similar stabilities. Refolding was initiated by dilution into a urea free buffer, and secondary structure was subsequently monitored by circular dichroism (CD) in the far UV. Upon unfolding and refolding, the spectrum of proaerolysin was to a large extent recovered (Figure 4.13 A). In contrast, the spectrum of aerolysin under refolding conditions showed typical features of a random coil, with a strong negative ellipticity in the 200 nm spectral region possibly due to protein aggregation. Thus, aerolysin was unable to reach a soluble state *in vitro* confirming the *in vivo* experiments. That aerolysin failed to reach a native conformation in vitro was confirmed by the lack of hemolytic activity after refolding from >4 M urea, in contrast to proaerolysin which refolded properly even from >6 M urea (Figure 4.13 B).

Altogether, these computational analyses and experimental observations indicate that the CTP is required for proper folding of aerolysin and that both the structure of the CTP and its binding affinity to Domain 4 are important for proaerolysin to reach a soluble active state.

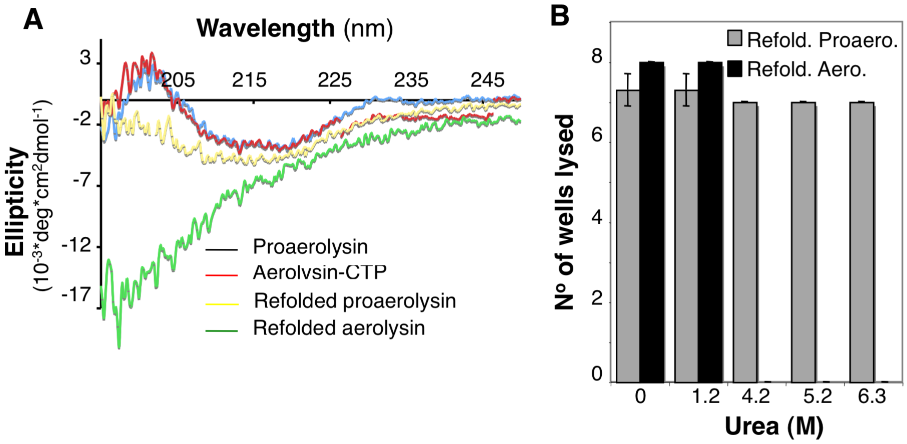


Figure 4.13: *(A) Circular dichroism spectra were acquired for proaerolysin before (blue) and after processing with trypsin (red, labeled aerolysin-CTP). Proteins were unfolded in 4 M urea and refolded by 10 fold dilution in a urea free buffer. The spectra of Refolded proaerolysin (yellow) and Refolded aerolysin (which has lost its peptide by unfolding) (green) were then acquired. Secondary structure could be recovered only by proaerolysin (B) Proaerolysin was subjected to proteolysis (red) or not (blue), unfolded in between 0 M and 6.3 M urea and allowed to refold in urea free buffer before the hemolytic activity was assessed. Proaerolysin was processed with trypsin after refolding. The results are the mean of 3 independent experiments. Error bars represent the standard deviation. Only proaerolysin was active after refolding.*

4.2.6 The CTP promotes folding and prevents heptamerization

Finally, it was investigated whether the CTP could also act when added in *trans* during in vitro refolding of aerolysin. It was found that addition of 5-fold molar excess of synthetic CTP led to a significant and reproducible recovery in hemolytic activity of aerolysin, whereas addition of an irrelevant peptide did not. The fact that rescue was only partial is not surprising since having a covalently bound CTP, as in proaerolysin, greatly increases the effective concentration of the peptide, a situation that cannot be mimicked by ectopic addition of excess CTP. A corollary of the observation that the CTP inhibits oligomerization is that the CTP must be displaced from the mature protein for the process to occur and thus that weaker CTP binding should promote oligomerization. It was first tested that the S453P CTP is be released more readily than the WT CTP. Indeed, apparent k_{off} rate of S453P was estimated as being 10 times higher than WT. That the S453P CTP has a lower affinity for the mature toxin was confirmed by the observation that upon binding of S453P aerolysin-CTP to Nickel beads, about 40% of the total aerolysin was recovered in the unbound fraction, i.e. it was released from the bead-bound CTP, whereas less than 10% of the aerolysin was released when performing a similar experiment with the WT toxin. Importantly, the CTP-free aerolysin fraction recovered from the S453P-treated beads had the same hemolytic activity as WT proaerolysin treated with trypsin. Three important conclusions can be drawn from this observation: 1) it is confirmed that the CTP is not required for pore formation [139] 2) CTP-free aerolysin does not unfold, but might change conformation, since it retains its full activity; 3) CTP-free aerolysin does not undergo unproductive aggregation, thus the role of the CTP is not merely to prevent aggregation of monomers.

Upon trypsin cleavage of S453P proaerolysin, oligomerization occurred faster than for WT. This is consistent with the fact that, if CTP release is necessary for oligomerization, then oligomerization should be accelerated when CTP binding is weaker. To our surprise, we found that S453P actually already showed some hemolytic activity even in the absence of trypsin cleavage, which is never observed for WT proaerolysin. This activity was some 15 fold lower than upon trypsin cleavage, yet significant and reproducibly detectable. Hemolytic activity in the proaerolysin form was also observed for the other CTP mutants L441A, R442A and L462A, but to a lesser extent. These observations show that cleavage of the loop linking the CTP to Domain 4 is not essential and that peptide displacement is sufficient.

4.2.7 Discussion

Guided by a combination of molecular simulations and *in silico* mutagenesis analysis and using a combination of structural and functional assays on WT and mutant toxins, we show that the CTP is essential for the folding of aerolysin into a soluble toxin. Due to the fact that it promotes folding but is not part of the final active conformation of the protein, i.e. the transmembrane heptameric pore, the CTP qualifies as a chain-linked molecular chaperone [25]. Chaperones comprise both proteins that favor the folding reaction of substrate proteins and proteins that control the quaternary assembly of multisubunit complexes. These two distinct roles can also be found in chain-linked, or intramolecular, chaperones and have been termed type I (folding) and type II (assembly) intramolecular chaperones [25]. Chain-linked chaperones can be short peptides (around 40 residues) or independent folding units. They are often found in secreted or transmembrane proteins, a situation that requires the protein to be translocated across the plasma membrane in prokaryotes (as for proaerolysin) or the ER membrane in eukaryotes. As discussed below, due to the directionality of membrane translocation coupled to protein synthesis, type I intramolecular chaperones are found at the N-terminus of proteins. However, exceptions, such as aerolysin, exist. Indeed, an N-terminal chaperone prevents misfolding a priori, while a C-terminal chaperone would act a posteriori. In contrast, most documented type II intramolecular chaperones are C-terminal. Irrespective of their localization, chain-linked chaperones should not be part of the final structure, which implies that they must be cleaved off at some point.

One of the earliest and best-characterized examples of a protein with an N-terminal intramolecular chaperone is *Bacillus subtilis* subtilisin, in which the 77 first amino acids fold into a well defined domain promoting the folding of the next 275 residues, acting as a type I chaperone, and is subsequently cleaved off by autoproteolysis [58].

C-terminal intramolecular chaperones have also been described. They are, however, generally of the type II, playing a role in controlling the quaternary assembly of proteins such as tail spikes of bacteriophages or fiber forming collagen [25]. Examples of type I C-terminal chaperones are rare and evidence is circumstantial [106, 138, 122, 61]. Aerolysin thus appears to be the first example of a protein bearing a C-terminal chain-linked chaperone promoting the formation of soluble monomeric subunits and controlling assembly of the active complex, i.e. both type I and type II. The present studies indeed show that the aerolysin CTP acts as a type II chaperone in controlling the onset of heptamerization, a role consistent with its C-terminal location. More unexpectedly, we found that the CTP drives formation of soluble proaerolysin. Mutations in the CTP that affects its structure (S453P) or its binding to Domain 4 (L441A, F457G) indeed lead to aggregation of proaerolysin both in vivo and in vitro. Moreover aerolysin, devoid of CTP, also aggregated. Importantly, addition in trans of a 5 fold molar excess of synthetic CTP allowed partial recovery of activity. Upon CTP release, the trigger for which remains to be established, aerolysin remains folded, possibly with a somewhat different conformation, as illustrated by the full hemolytic activity of CTP-free aerolysin obtained from the S453P mutant. The unaltered activity of CTP-free aerolysin also indicates that the CTP plays a role in the biogenesis of the toxin and does not prevent unproductive aggregation of protein once folded. Altogether, these observations thus classify the aerolysin CTP as a chain-linked intramolecular chaperone.

Our observations clearly indicate that the CTP prevents aggregation of proaerolysin during biosynthetic folding. As mentioned above, and as supported by the ability of the CTP to promote recovery of hemolytic activity upon *in vitro* folding of aerolysin, the CTP appears to do more than merely preventing aggregation as also suggested by the molecular dynamics studies. Confirming that the CTP promotes the folding of aerolysin and how it does so will require further investigation. Since proaerolysin is translocated from N- to C-terminus when crossing the inner *Aeromonas* membrane, the CTP appears last. In particular, it appears some 250 residues later than some of the residues it interacts with. What is also puzzling is that the CTP is not an independent folding unit that could guide folding of the rest of the protein, as is the case for most type I intramolecular chaperones. Our MD simulations suggest that when released from the protein, as mimicked by the F457G mutation, or when free in solution, the CTP is largely unfolded. Therefore the CTP might stabilize folding intermediates. It has been proposed that, as a protein follows its folding landscape, the chaperone domain binds to, stabilizes and increases the population of molecules with native conformations. Thus, as opposed to general chaperones, which are thought to lack any structural information about the protein they fold, dedicated chaperones and possibly the aerolysin CTP could promote folding via conformational selection [84, 136, 72] and thus convey steric information. This hypothesis is consistent with the observation that one segment of Domain 4 with which the CTP interacts in the final structure, residues 269-279, is predicted to be unstructured. Even though largely unfolded, such segments are likely to fluctuate between multiple folded states during short times, one of which could be stabilized by the CTP. A prediction from the conformational selection model for CTP-mediated proaerolysin folding is that folding should be affected by mutations in the CTP. This is indeed what we observed for the mutants suggested by in silico alanine scanning mutagenesis and in particular for the S453P and F457A/G mutations.

As mentioned above, the CTP appears to force segment 268-272 into a β-strand conformation. Importantly, this segment is directly followed by the loop in Domain 3 that is to form one of the amphipathic β-hairpins of the heptameric transmembrane β-barrel pore. The ability of the CTP to control the folding state of the underlying β-strands (note that Domain 4 shares multiple β-strands with Domains 3 and 2) suggests that the peptide also acts as a switch to control the pore formation process. Our observations indeed show that CTP release promotes oligomerization and that the CTP is not part of the final pore. Future studies will address what triggers release of the CTP. Our preliminary observations indicate that specific detergents can displace the CTP, consistent with the importance of hydrophobic interactions in CTP binding and suggesting that the target cell membrane may play a role. Future studies will also address whether CTP release triggers partial unfolding of Domain 4 and whether these changes propagate to Domain 3 helping overcome the energy barrier that leads to formation of the heptamer, the most thermodynamically stable conformation [77].

4.2.8 Material and Methods

Structural models

WT proaerolysin in its dimeric form has been crystallized with a resolution of 2.8 Å (entry 1PRE in protein databank). In this crystal structure, two loops located on top of Domain 4 proved too flexible to be crystallized, namely residues 207 to 211 and 423 to 439. A crystal structure of dimeric proaerolysin mutant Y221G has been obtained with a higher resolution of 2.2 Å (entry 3C0N in protein databank). In this crystal, residues 207 to 211 could be mapped in the crystal structure but loop 423 to 439 is still missing. This loop connects the CTP to the rest of the protein, and contains the site where cleavage takes place during aerolysin activation (420-427).

A model of wild-type aerolysin with the propeptide bound (labeled CTP-WT) to use in molecular dynamics simulations was obtained by using 3C0N structure, and mutating residue 221 back to tyrosine using 1PRE as a structural template (wild-type rebuilding did not caused any steric problem since 3C0N and 1PRE were virtually identical). We assumed that this model would mimic the aerolysin structure after cleavage, i.e. C-terminal propeptide no longer covalently connected to the protein, but still bound to it. In fact, this model is structurally equivalent to the cleaved aerolysin H132N X-ray structure showed in this work. We modeled active aerolysin (labeled WT) by removing the propeptide from the previous model. Mutation F457G has been performed by removing the F457 side-chain.

Aerolysin contains six histidines. Their protonation state at physiological pH has been defined by the presence of proton donors and acceptors in their neighborhood in the crystal structure. We concluded that in H107, H121, H132, H186 and H332 Nϵ atom is protonated, whereas in H341 Nδ is protonated.

Molecular Dynamics Simulations

Structural model systems were solvated in a rectangular box of pre-equilibrated TIP3P water molecules, and their total charge was neutralized by the addition of Na+ and Cl- counterions. Molecular dynamics has been performed for aerolysin with CTP (labeled Aero-CTP), without CTP (labeled Aero) and mutation F457G using the Amber99SB force field [21] on NAMD molecular dynamics engine [116], using the SHAKE algorithm on all the bonds, and Particle-mesh Ewald for treating the electrostatic interactions in periodic boundary conditions [13]. We used an integration time step of 2 fs. The systems were energy-minimized by means of 1000 conjugate gradient steps, and subsequently gradually heated from 0 to 300 K in 1 ns at 1 atm. Simulations were run in the NPT ensemble at 1 atm and 300 K. Temperature was controlled by means of Langevin forces, using a damping constant of 1 ps^{-1}.

Preliminary results confirmed that Domain 1 is bulky and, being connected by a long random coil to the large lobe, extremely flexible with respect to the rest of the protein. Since this resulted in no influence on the structure of the other domains, and that Domain 1 is known to act as an independent folding unit [77], we decided to remove it in order to reduce the system size and therefore speed up the remaining computation. All simulations were run for at least 200 ns. RMSD of MD simulations showed that every system equilibrated in around 10 ns. Analysis of MD trajectories, as well as rendering of protein structures, has been performed using VMD [54].

Alanine Scanning

Alanine scanning was performed on the single 200 ns molecular dynamics trajectory of CTP-WT system. A subset of 200 decorrelated frames (one every 10 ns) was extracted. On this subset, we calculated binding free energies of Ala mutant species using the MM/PBSA method, as implemented in the AMBER molecular dynamics package [21]. The Poisson-Boltzmann method was used to compute the electrostatic contribution to the solvation free energy. Ionic strength molarity was set to 0.1 M, the protein dielectric constant to 1, and the solvent to 80. Every residue being part of the CTP, excluding glycines and prolines, was scanned. These residues play a major role in the determination of strand flexibility, thus the alanine scanning is known to perform poorly. The MM/PBSA method was also used to estimate the binding free energy of WT and F457G mutated CTP to domain 4. For both these measures, 200 decorrelated frames extracted from aero-CTP and F547G MD simulations were used, respectively.

4.3 Macromolecular assembly of Y221G soluble pore

It has been found that a single point mutation, Y221G has a dramatic effect on aerolysin's functionality [137]. Remarkably, the toxin can still heptamerize but the resulting assembly, unable to insert into the lipid bilayer, is locked in a hydrophilic conformation. Authors suggest that mutation Y221G impairs the conformational changes which are known to take place once the protein activates (CTP removal and loop region displacement). As a support to this finding, a cryo-EM of Y221G's soluble heptamer at 13.5 Å resolution has been obtained (see Figure 4.14). The density map shows seven monomers assembled according to a circular symmetry. The multimeric structure is 150 Å wide, 80 Å large, and comprises 10 Å large central pore. No β-barrel can be observed.

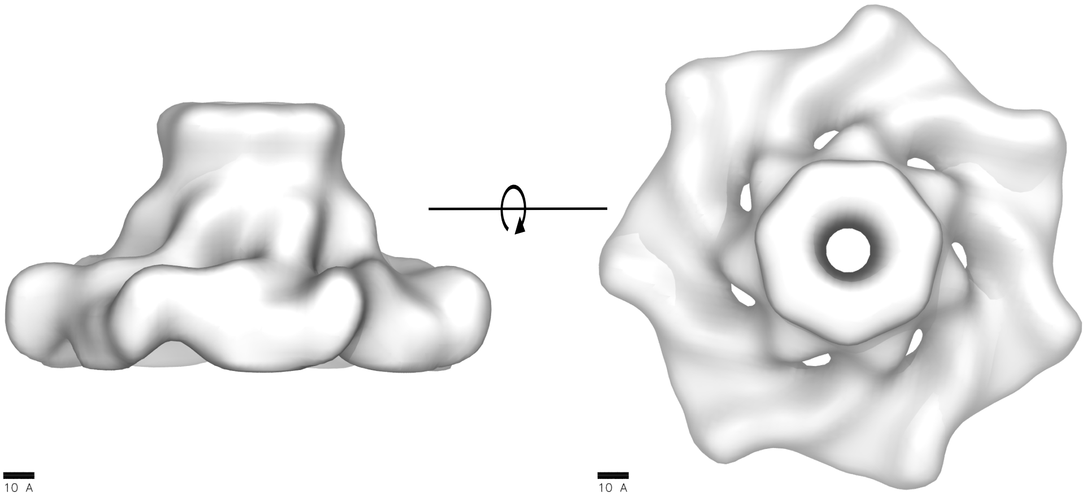


Figure 4.14: *Heptameric aerolysin Y221G cryo-EM map at 13.5Å resolution [137]*

A tentative atomistic model for Y221G heptameric conformation fitted onto the available cryo-EM has been proposed [137]. The model had a good cross-correlation coefficient with respect of the cryo-EM map (0.78), but was based on the assumption that aerolysin's major lobe is rigid. The mushroom-shaped model was oriented with its "stem", constituted by Domain 4, pointing towards the membrane. Given the large hydrophobic patch located underneath the CTP [110] and an abundance of aromatic residues forming an aromatic belt, Domain 4 was in fact pinpointed as a good candidate for a transmembrane region. Iacovache et al. later demonstrated that loop region in Domain 3 assembles into a β-barrel becoming, in fact, the transmembrane region [57]. The question of how the multimeric structure is oriented with respect to the membrane remains, therefore, open.

By exploiting the available X-ray structure and cryo-EM map of aerolysin Y221G, we present here an atomistic model of its heptameric assembly. The model is built by keeping into account the protein's natural flexibility, and is finally used to predict specific point mutations impairing aerolysin's assembly process.

4.3.1 Mutation Y221G prevents Domain 4 unfolding and loop extraction

In order to characterize the toxin's flexibility and pinpoint important differences within WT and relevant mutations, MD was performed on representative crystal structures. Crystals of dimeric proaerolysin WT and dimeric mutation Y221G are available in the protein databank (PDB databank entries 1PRE and 3C0N, respectively). We initially removed Domain 1 from both monomers, since preliminary results indicated that this domain does not affect the dynamics of aerolysin's major lobe constituted by Domains 2, 3 and 4. First, both monomers were simulated with the CTP non-covalently attached (simulations labeled CTP-WT and CTP-Y221G, respectively). After 100 ns, protein activation was mimicked by manually removing the CTP from both equilibrated structures. Subsequently, 200 more ns were simulated (labels WT and Y221G, respectively).

After 100 ns, both CTP-WT and CTP-Y221G simulations were stable. This is not surprising since, as seen in Section 4.2.4, the CTP preserves Domain 4 from unfolding. However, the β-sheet content in CTP-Y221G simulation was slightly higher. This was mainly due to a different arrangement of residue L277. In fact, in simulation CTP-WT, L277 is solvent exposed, since the access to Domain 4's hydrophobic core is hindered by the side chain of residue Y221, located in the opposite strand. Conversely, in simulation CTP-Y221G, the absence of Y221 leaves L277 space for rotating its side chain, which locks in a hydrophobic pocket constituted by residues L219, I295, L297, A411 and CTP residues F457 and V460. With the removal of the polar Y221 and the access of L277, Domain 4's hydrophobic core ends up being more rigid.

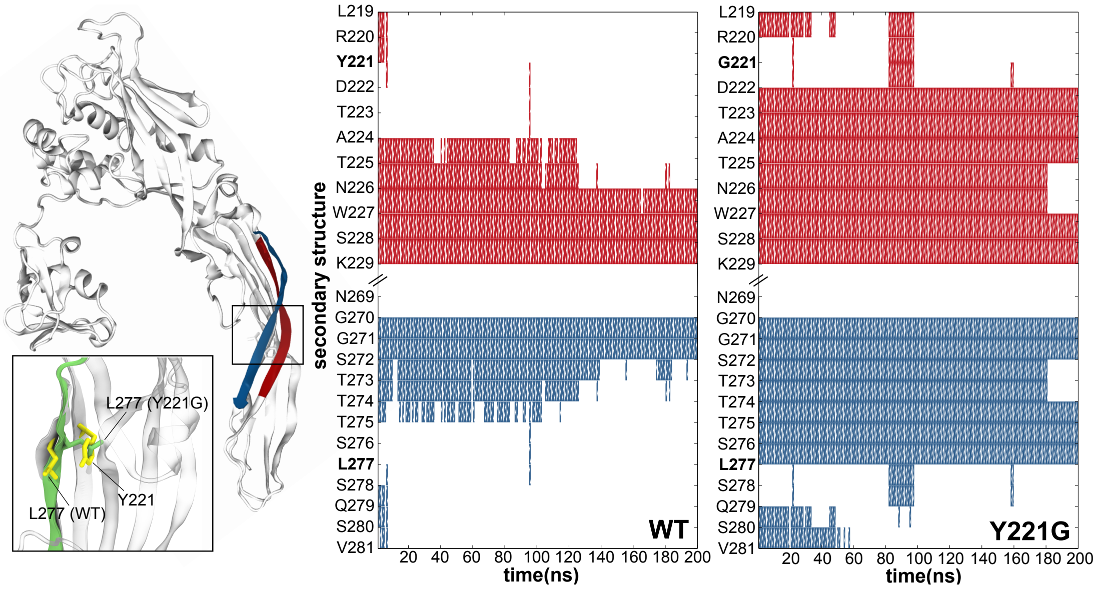


Figure 4.15: *Comparison within WT and Y221G secondary structure during MD simulation. When CTP is removed, in WT unfolding propagates from the top of Domain 4 towards the loop region. In Y221G, unfolding is blocked in correspondence of L277, which anchors its strand to the rest of Domain 4. In the inset, in mutation Y221G L277 rearranges, pointing inside the neighboring hydrophobic pocket.*

Upon CTP removal, Domain 4 of WT simulation began unfolding. The strands being the most affected were L219 to K229 and N269 to V281. Interestingly, unfolding initiated in the neighborhood of Y221, and propagated towards the loop region. Finally, after 200 ns, just 33% of the initial secondary structure was preserved. Conversely, CTP removal affected way less the secondary structure of simulation Y221G. Indeed, after 200 ns, 59% of the initial secondary structure was preserved (Figure 4.15). Interestingly, after CTP removal, L277 holds its position in Y221G hydrophobic pocket, while neighboring residues also preserve their initial β-sheet secondary structure (see Figure 4.16).

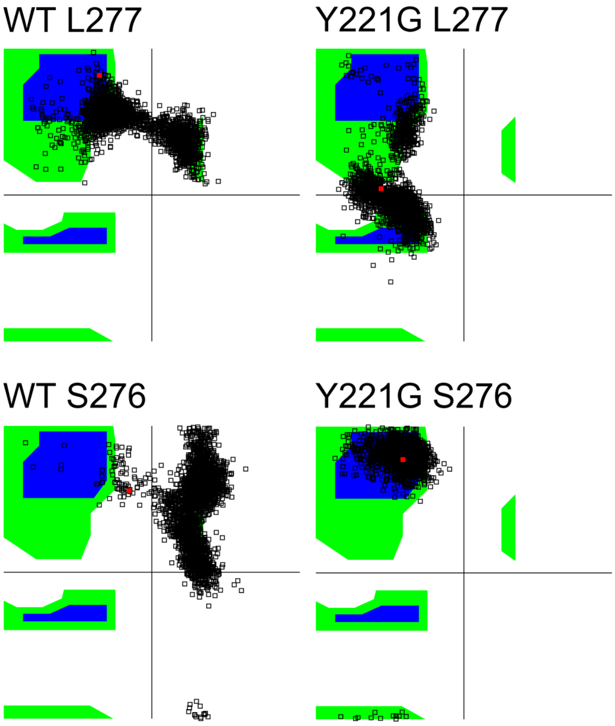


Figure 4.16: *Comparison of Ramachandran plot of L277 and S276 residues in WT and Y221G simulation. In Y221G, S276 maintains an helical conformation along the whole trajectory due to L277 rearrangement.*

In the absence of Y221G, residue L277 plays therefore a relevant role on the increased order in Y221G Domain 4. The peculiar position of Y221 in the hydrophobic pocket located underneath the CTP could eventually serve as a trigger for conformational change, once the CTP is removed. Computing the RMSD of the last frames of simulations WT and Y221G with respect to aerolysin's crystal structure, underlines how WT went through larger conformational changes: WT RMSD was 6.12 Å, whereas in Y221G it was 4.4 Å. We conclude that mutation Y221G has a stabilizing effect on aerolysin. It is tempting to speculate that strand unfolding in WT might ultimately propagate in the loop region, affecting its binding stability.

4.3.2 Assembly of aerolysin Y221G heptamer

The differences within WT and Y221G simulations suggested that the two proteins explored a different conformational space. In order to highlight the most relevant degrees of freedom of both proteins, we performed a Principal Components Analysis (PCA) on the C_α atoms of their trajectories (see Figure 4.17).

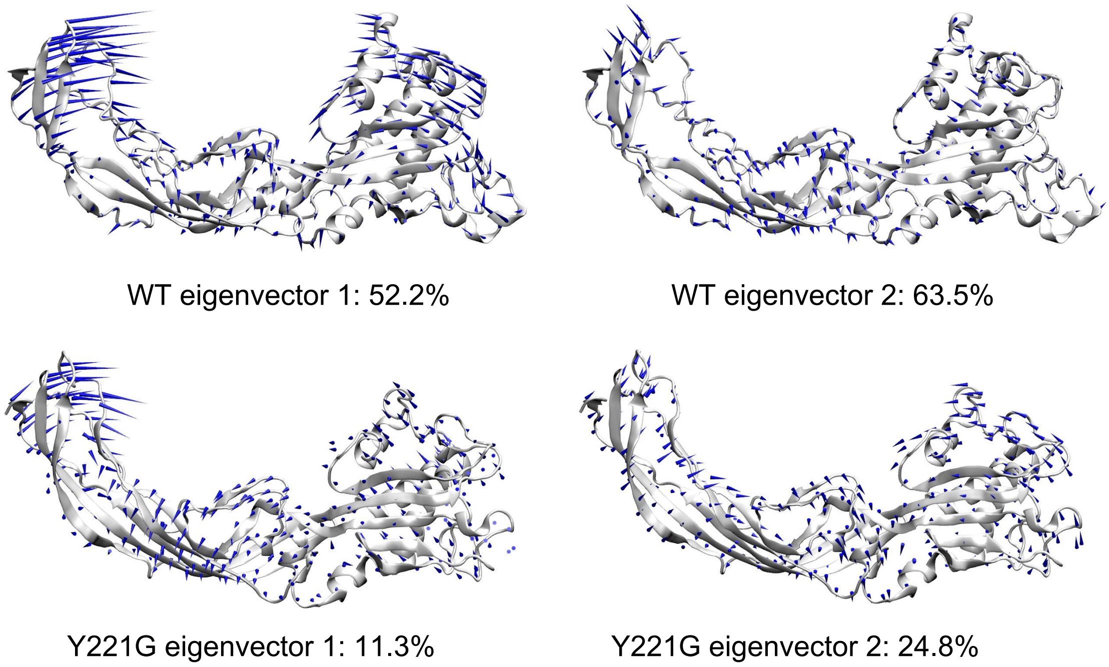


Figure 4.17: *First and second principal component of WT and Y221G MD simulations. Proteins explore different regions of the conformational space.*

By observing the most relevant modes, we notice that Y221G and WT main movements differ. Additionally, for Y221G simulation just three eigenvectors are sufficient to represent most protein motions (a total of 94.5%), whereas in order to describe more than 90% of WT protein movements, 12 eigenvectors are needed. In order to directly compare the two simulations, the two simulations were p projected in the same eigenspace (see Figure 4.18). Results clearly indicate that aerolysin WT and Y221G explore different regions of the eigenspace. This analysis clearly indicates that mutation Y221G not only affects the secondary structure of aerolysin, but also its overall flexibility. PCA unveiled that in both simulations the most relevant movement was related to the flexibility of Domain 4 with respect of Domain 3. This degree of freedom cannot be ignored when attempting to predict aerolysin's multimeric conformation.

Having obtained information about Y221G flexibility, we tackled therefore the problem of predicting its multimeric conformation on the base of the available cryo-EM structure. To do so, we initially extracted one frame every 100 ps from our MD Y221G simulation, and used

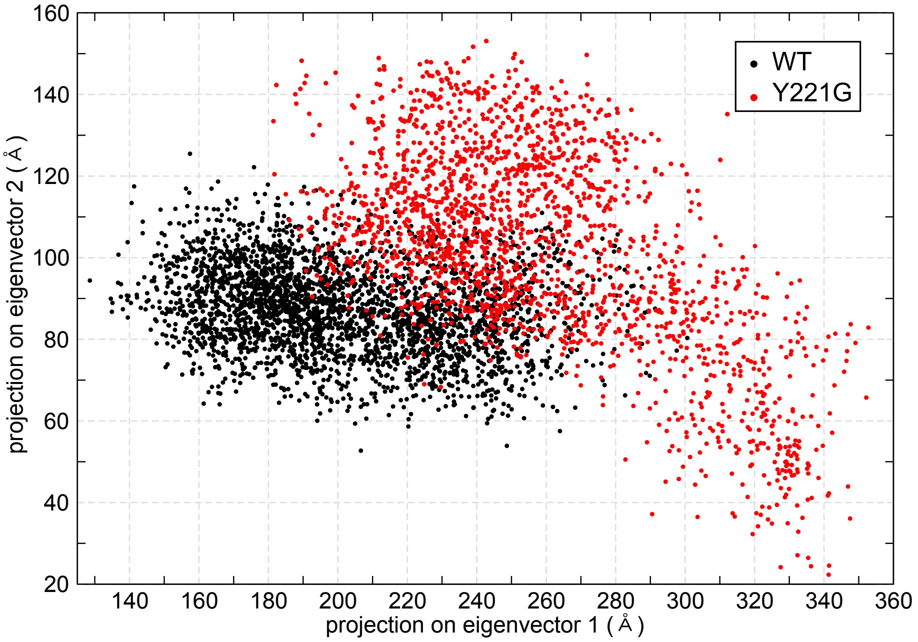


Figure 4.18: *Projection of WT and Y221G simulations in the same eigenspace. The two proteins explore different regions of the eigenspace.*

them as a structure database. Every structure was indexed by the values of the projections on trajectory's 3 main eigenvectors (representing 94.5% of protein movement). The heptamers' conformational space was defined as the three rotation angles and a translation of a monomer with respect of the center of symmetry of the assembly. Having defined the correct orientation of a monomer, a complete heptamer can be simply assembled by extracting a structure from the database and applying a rototranslation matrix so that a circle is equally partitioned in seven sectors. To score the quality of an assembly, also called its fitness, we kept into account two contributions: one geometric and one energetic. Our task is to find the assembly having the lowest fitness by exploring the space of possible alignments of 7 monomers extracted from our conformations database. The time necessary to find the best solution in such an optimization problem by means of a brute force approach is directly proportional to the size of the search space and to the granularity of the search grid. Such an approach becomes inefficient when the search space becomes large and complex as in the case of protein-protein docking. To tackle our problem, we adopted a Particle Swarm Optimization technique (as implemented in POW software, see Chapter 3). POW generated six models, which were rigidly docked in the available cryo-EM map using Situs [142], and ranked according to their cross-correlation coefficient (CCC). The best model obtained with this procedure had a CCC of 0.72. This model was further optimized by means of a protocol based on Molecular Dynamics Flexible Fitting (MDFF). This technique aims at docking flexibly all atom structures in density maps using MD simulations. The result was a model having a CCC of 0.92 (see Figure 4.19).

Conversely, when only a single monomer extracted from aerolysin's X-ray structure was used to assemble a multimer, none of the solutions explored met the solutions filtering criteria. This indicates that no structure satisfying the given geometric restraints could be found. We docked the best solution (fitness equal to 4.75) into the Y221G cryo-EM map obtaining a CCC of 0.57, value significantly lower than what obtained using the flexible docking protocol and consistent with the several inconsistencies of the model with known experimental data. This comparison shows that using an ensemble of structures representative of monomer flexibility leads to improved performances in assembly prediction and to the generation of more biologically sound models.

The position of Domain 1 and 2 GPI-anchor binding sites prompts us to locate the membrane in contact with these regions. The hairpin M140-G153, also supposed to be responsible of binding to the membrane [92], is by consequence also in contact with it. In our model, we are therefore turning aerolysin's structure upside down with respect to of all the other models proposed up to now [110, 137]. Domain 3 acts as a hinge and Domain 4 stems upwards, constituting therefore the pore mouth. Interestingly, Domain 4 strands L219 to K229 and N269 to V281, connected to the loop region, are aligned vertically. Polar and charged residues K198, D207, D222, R288 and K290 are exposed to the lumen. The loop region itself, is favorably located above a hole in the density map. By means of a sliding movement, the loop could easily reach the underlying bilayer via a sliding movement.

In order to validate this model experimentally, we focused our attention on polar residues located in the neighborhood of a protein-protein interface. 14 residues that could have an active role in binding two aerolysin monomers were identified. Mutating these to alanine should reduce the binding affinity of two monomers, as such affecting the multimerization process. Mutagenesis results, although preliminary, are encouraging. Mutations K198A, D216A, R282A (Domain 4 - Domain 4 interface) K369A, E367A (Domain 2 - Domain 1 interface) and K351A (Domain 2 - Domain 3 interface), lead indeed to delayed or even impaired heptamerization.

Another ongoing experiment aims at detecting, using nanogold labeling, residues located on top of domain 4 (pointing therefore upwards in our model). For this reason, we suggested four possible mutations: A287C, L422C, T210C and R288C. These residues are all solvent exposed on top of Domain 4, and are not involved in any relevant interaction within monomers. A successful labeling in this experiment will validate, at the same time, our model and its orientation with respect to the lipid bilayer.

4.3.3 Materials and Methods

Structural Models

WT proaerolysin in its dimeric form has been crystallized with a resolution of 2.8 Å (entry 1PRE in protein databank). In this crystal structure two loops located on top of Domain 4 proved too flexible to be crystallized, namely residues 207 to 211 and 423 to 439. A crystal structure of dimeric proaerolysin mutant Y221G has been obtained with a higher resolution of

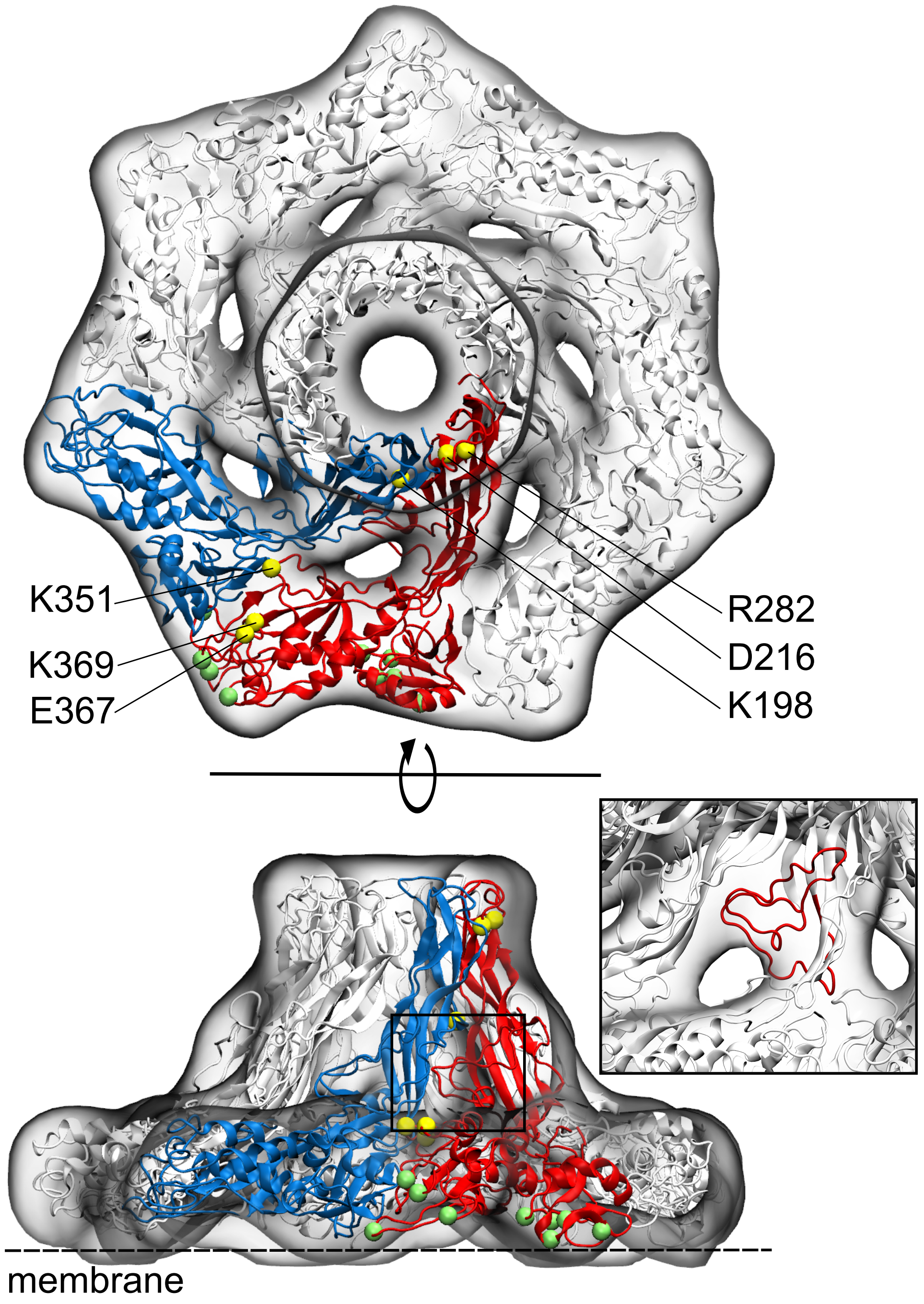


Figure 4.19: *Top and side view of Y221G heptamer atomistic model superposed to Y221G cryo-EM map (13.5 Å resolution) produced by Tsitrin et al. [137]. Yellow spheres show interfacing residues that, mutated to alanine, lead to delayed heptamerization. Green spheres show residues responsible of binding to GPI-anchors. In the inset, the loop region could reach the membrane by sliding through the hole underneath it.*

2.2 Å (entry 3C0N in protein databank). In this crystal, residues 207 to 211 could be mapped in the crystal structure but loop 423 to 439 is still missing. This loop connects the CTP to the rest of the protein, and contains the site where cleavage takes place during aerolysin activation (420-427). We obtained a model of WT aerolysin with the CTP bound (labeled CTP-WT) by using 3C0N, and mutating residue 221 back to tyrosine using 1PRE as a template. We assumed that this model would mimic the aerolysin structure after cleavage, i.e. CTP no longer covalently connected to the protein, but still bound to it. Aerolysin contains six histidines. Their protonation state has been defined by the presence of proton donors and acceptors in their neighborhood in the crystal structure. We concluded that H107, H121, H132, H186 and H332 have the NE2 atom protonated, whereas in H341 atom NE1 is protonated.

Molecular dynamics simulation protocol

CTP-WT and CTP-Y221G systems were solvated in a rectangular box of pre-equilibrated TIP3P water, and their total charge was neutralized by the addition of Na^+ and Cl^- ions. Molecular dynamics has been using the Amber99SB force field [21] on NAMD molecular dynamics engine [116], with SHAKE algorithm on all the bonds, and Particle-mesh Ewald treating the electrostatic interactions in periodic boundary conditions. We chose an integration step of 2 fs. The systems had their energy initially minimized by means of 1000 conjugate gradient steps, and have subsequently been gradually heated from 0 to 300 K in 1 ns at 1 Atm. Simulations were run in the nPT ensemble at 1 atm and 300K. Temperature has been controlled by means of Langevin forces, using a damping constant of 1 ps^{-1}. Preliminary results confirmed that Domain 1 is bulky and, being connected by a long random coil to the bigger lobe, extremely flexible with respect to the rest of the protein. Since this resulted in no influence on the structure of the other domains, and that Domain 1 is known to act as an independent folding unit [77], we decided to remove it from every system in order to reduce the system size and therefore speedup the computation. Simulations were run during 100 ns. RMSD of every simulation showed that every system equilibrated in around 10 ns. At the end of simulations aerolysin's structures were extracted and the CTP was manually removed. We assumed that this would mimic the effect of proteolytic cleavage. The two newly obtained systems, labeled WT and Y221G respectively, were equilibrated and simulated during 200 ns using the same simulation protocol detailed above. Analysis of molecular dynamics trajectories, as well as rendering of protein structures, has been performed with VMD provided tools [54].

Principal Component Analysis

From Y221G and WT trajectory we extracted one frame every 100 ps containing only the coordinates of the C_α atoms of the protein. 2000 uncorrelated frames were therefore obtained, for every simulation. PCA was performed on the obtained conformations ensembles. The first three eigenvectors were representative of 94.5% and 72.9% of protein movement, respectively. By computing the dot product of the structures with all the selected eigenvectors we obtained the fluctuation associated to every conformation with respect to the main modes. In order to project WT and Y221G simulations in the same eigenspace, 2000 structures (C_α only) were first extracted from both simulations. These were subsequently concatenated in a unique trajectory, on which PCA was run.

POW protocol

PSO search space was 7-dimensional. Three dimensions defined the rotation angles of a monomer with respect of the center of symmetry of the multimer, and their value was allowed to vary within 0 and 360 degrees. One dimension represented the translation with respect of the center, and varied within 4 and 6 Å, i.e. the expected pore radius as measured in Y221G cryo-EM map. Finally, three values represented the desired fluctuation of the protein with respect of its main eigenvectors defined via PCA. These values varied within the minimal and maximal values in the eigenspace. Periodic boundary conditions were applied on all dimensions. To generate an assembly corresponding to a position on the search space, the structure in the database being the closest to the desired fluctuations values is initially extracted. The monomer is subsequently rotated and translated, and a 7-fold symmetry is finally imposed.To score the quality of an assembly, a sum of two contributions, energetic and geometric, was adopted (see equation 3.12). The geometric contribution f_c is meant to select only the assemblies matching observed data. Y221G cryo-EM map indicates that the assembly should be 85 ± 5 Å high and 150 ± 5 Å wide, and have a central pore being 10 ± 2 Å large. To assess f_c, we computed the Euclidean distance within the width w_m and height h_m of our models and the desired target measures $w_t = 150 \pm 5$ and $h_t = 85 \pm 5$. The second contribution, energetic, is meant to avoid configurations containing severe clashes, and computed a 9-6 Lennard-Jones potential within the C_α of two neighboring monomers, applying a cutoff at 12 Å (see equation 3.10).

For PSO protocol we used 80 particles, 200 iterations and 3 repetitions. All the measures performed by every particle were logged and only coordinates having fitness smaller than zero were retained as potential good solutions. A solution having fitness smaller than zero respects all geometrical constraints and has a good crystal packing (no clash). Given that several solutions found by our POW were sampling the same region of the search space, we clustered solutions having an RMSD smaller than 5 Å and considered the 6 obtained clusters as representatives of the detected assemblies.

Model refinement

The assemblies corresponding to the 6 solutions obtained with POW were rigidly docked into Y221G cryo-EM using Situs [142]. Resulting docks were ranked by computing their cross-correlation coefficient (CCC). It has to be noted that these assemblies were obtained by a simulation of aerolysin's major lobe only. For this reason, using HEX [120] we docked Domain 1 within two monomers extracted from the assembly having the highest CCC. A new complete model was obtained by applying a 7-fold symmetry to Domain 1's best scoring pose.

Available cryo-EM had a resolution of 13.5 Å. The best model obtained with PSO search, composed of 44205 atoms, has been optimized by means of a Molecular Dynamics Flexible Fitting (MDFF) protocol as implemented in NAMD [135]. In MDFF, a biasing potential is added to the standard MD force field in order to guide the protein towards region having a higher density. In order to avoid overfitting, proteins' secondary structure was restrained. 5 ns simulation was performed *in vacuo* at 300 K and 1 Atm, with SHAKE restraining all bonds and a dielectric constant of 80. 5000 conjugate gradient minimization steps were finally performed. In order to compare the all atom assembly to the target cryo-EM, a density map of the all atom assembly having the same resolution of the cryo-EM (i.e. 13.5 Å) was generated. The CCC within the two maps was subsequently computed.

4.4 Modeling of the wild-type heptamer

Upon heptamerization, WT aerolysin undergoes large conformational changes transforming it into an amphipathic assembly. In the process, the CTP gets removed from domain 4, and the loop region gets displaced, crossing the membrane and forming a transmembrane β-barrel. Wilmsen et al. observed that aerolysin heptamers inserted into lipid vesicles form almost flat discs being 150 Å wide and within 30 and and 40 Å thick [141]. Interestingly, these aerolysin discs appear to be mostly separated from the membrane. This observation is strikingly different from the cryo-EM map of Y221G heptamer, which appears to be mushroom-shaped [137] and way taller (around 88 Å).

Recently, new density maps of aerolysin heptamers were obtained via a negative staining in van der Goot lab (see Figure 4.20). This result was possible by producing aerolysin mutations harboring a disulfide bridge in the loop region via the double point mutation K246C and E258C. This mutation hinders the complete deployment of the loop in the shape of a β-barrel (a "partial barrel" is formed), and produces aerolysin heptamers dimerizing in two possible ways. In most cases, a heptamer in prepore state dimerizes with an heptamer having its barrel already formed (in the following, we will simply refer to this state to as "pore"). In this case, pore's β-barrel (visible in some cases) locks inside the cavity of the prepore. Interestingly, while the prepore appears similar to the cryo-EM of Y221G, the pore appears distinctively disc-shaped.

This result constitutes the first density map of aerolysin having its barrel at least partially formed. Consistently with the first observations of Wilmsen et al., it hints that prepore (Y221G-like) and pore state are dramatically different. In this study, we aim at modeling at an atomistic resolution the structure of aerolysin in its pore conformation. We will first characterize the conformational changes triggered by the extraction of the loop region from its resting place. Subsequently, the transmembrane β-barrel and the disc-shaped aerolysin assembly will be modeled and assembled. This will be done by exploiting the newly available density map of aerolysin in its pore state.

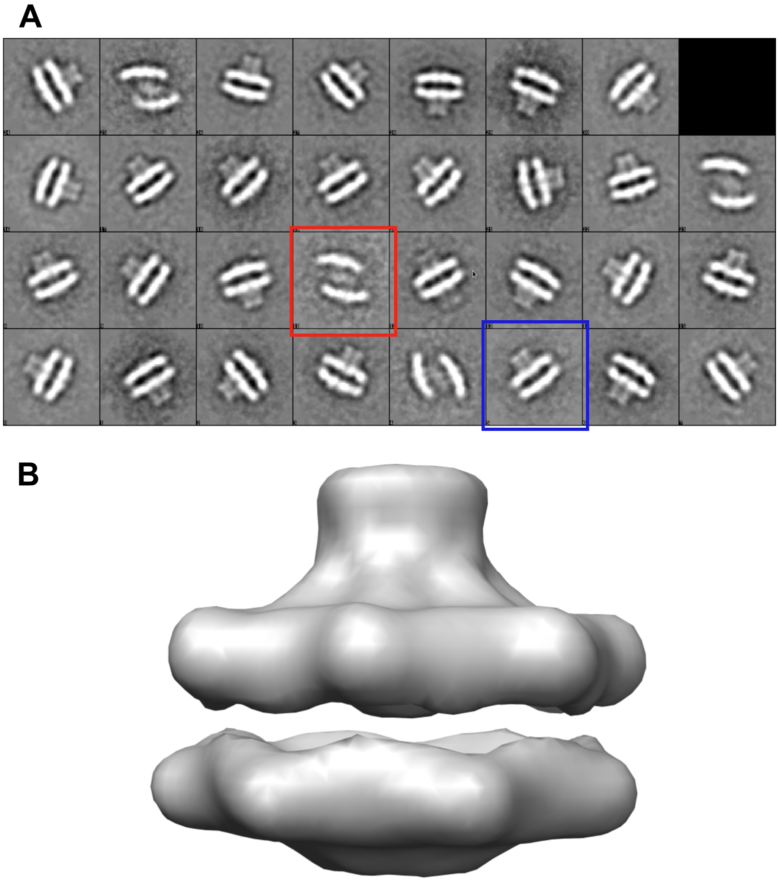


Figure 4.20: *(A) Classes extracted from a negative staining experiment on aerolysin harboring a disulfide bridge on its loop region. Two types of associations can be observed, prepore-pore (highlighted in blue) and pore-pore (in red). (B) 3D reconstruction of a prepore-pore dimer.*

4.4.1 Loop extraction affects domains arrangement

The most striking difference within aerolysin WT and Y221G is that only the first is able to produce a transmembrane β-barrel via a conformational change of the loop region located on Domain 3. In our Y221G model, the loop region could reach the membrane surface by sliding away from its resting place. This prompted us to determine the effects of such a rearrangement on the toxin's structure. For this reason we equilibrated in MD new aerolysin WT (labeled WT+CTP) and and WT having its CTP removed (labeled WT) systems and subsequently steered the loops away from their resting place. This was done by applying a biasing force on loop's C_α atoms so that a rotation at constant velocity around a hinge, which we selected as residue E236, was obtained (see Figure 4.21). These newly obtained systems, labeled WT_loop_extracted and WT+CTP_loop_extracted respectively, were simulated for 200 ns.

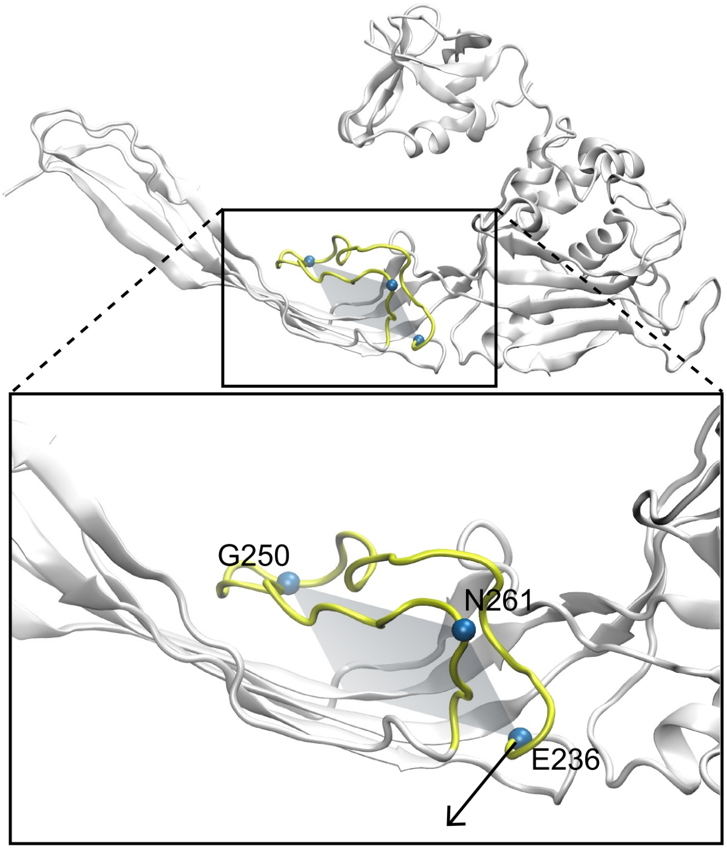


Figure 4.21: *Setup for aerolysin's loop steering. Loop region (yellow), is rotated around the E236. The rotation plane is identified within the C_α atoms of residues E235, G250 and N261 (Van der Waals spheres)*

Interestingly, in both systems steering the loop led Domain 4 to rearrange with respect of Domain 3, assuming a totally new configuration (see Figure 4.22). In order to quantify this rearrangement, we first of all considered aerolysin domains as rigid bodies flexing one respect to each other. All trajectories were initially aligned on Domain 3 (in green in Figure 4.22 inset). Subsequently, for every simulation frame the Principal Components of Domain 4 (in yellow) were calculated. Since Domain 4 is elongated, the first principal component corresponds to the vector connecting Domain 3 to the tip of Domain 4. For every frame, this vector was extracted and converted in spherical coordinates. The obtained ϕ angle represents how parallel Domain 3 and 4 are, whereas the ψ angle displays Domain 4 side twist. We observe that in simulation CTP-WT Domain 4 angles fluctuate around an equilibrium point corresponding to crystal structure arrangement. Upon loop extraction, Domain 4 twists sidewise and flattens with respect of Domain 3. When both CTP is removed and loop is extracted, Domain 4 twists and completely flattens with respect of Domain 3. This result shows that both CTP and loop region affect aerolysin's domains arrangement.

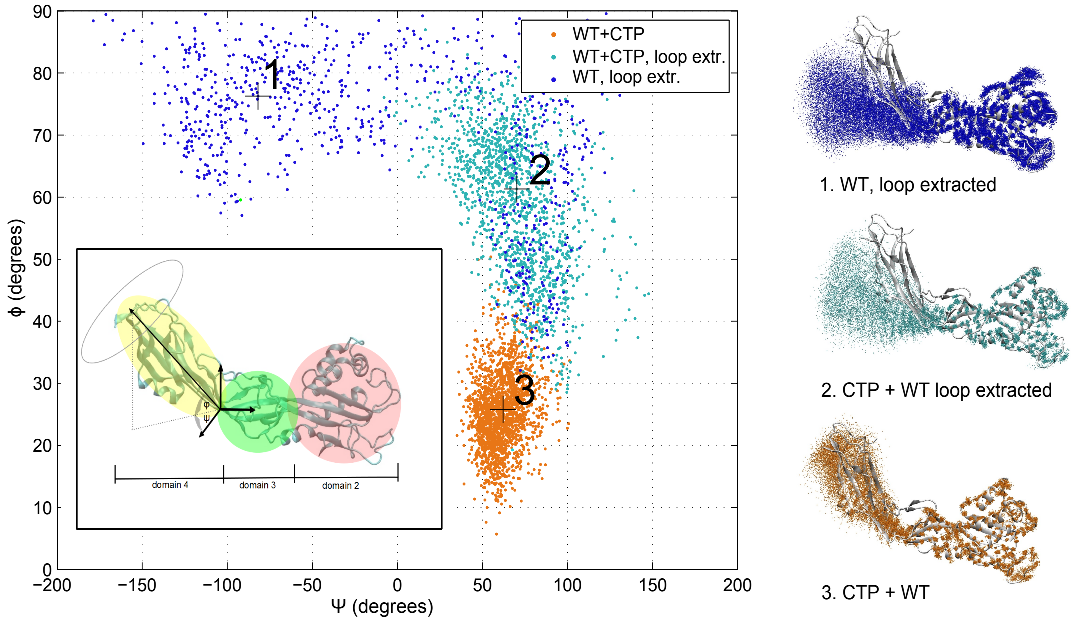


Figure 4.22: *Comparison of Domain 4 flexibility in WT aerolysin upon CTP removal and loop extraction. Both CTP and loop affect domains arrangement.*

4.4.2 Model of the aerolysin WT heptamer

On the basis of the density map of aerolysin WT heptamer produced in Van der Goot lab, we aimed at modeling the membrane inserted assembly's structure at atomistic resolution. To do so, we distinguished two parts which were modeled independently: the *transmembrane β-barrel* (constituted by residues part of the loop region), and the *outer assembly* (all the residues, loop region excluded). The available density map does not provide a clear view of the first part, but it clearly shows the second.

Since no structure of the transmembrane β-barrel exists for aerolysin, we modeled it on the basis of its homologue α-hemolysin, using as seed the first model produced by Iacovache et al [57] (hereafter "rivet model"). This was done by assuming that aerolysin's barrel should have a length similar to the one of α-hemolysin. For this reason, we extended the rivet model so that the length of α-hemolysin barrel is matched (from residue K229 to residue N269, see Figure 4.23). Strikingly, unlike α-hemolysin, the resulting model exposes a large amount of charged residues on its lumen (E237, K238, K242, K244, K246, E252, E254, E258). Also, interestingly, an alternating hydrophobic-hydrophilic pattern, peculiar of transmemberane β-barrels, is present on a length of about 25 Å. This length matches the average thickness of the hydrophobic region of in an eukaryotic bilayer. This alternating pattern terminates on a ring composed of aromatic residues W265 and Y233, facing outwards. This aromatic belt would provide an ideal anchoring to the bilayer's external surface. Just one polar residue, E237, is facing outwards, in contact with the membrane. However, its short distance from lipids polar heads would permit it to snorkel. It should also be pointed out that, as already noted by

Iacovache et al., aerolysin's barrel could hook to the internal part of the bilayer in a peculiar rivet-like conformation. The model was finally minimized and equilibrated in a POPC bilayer.

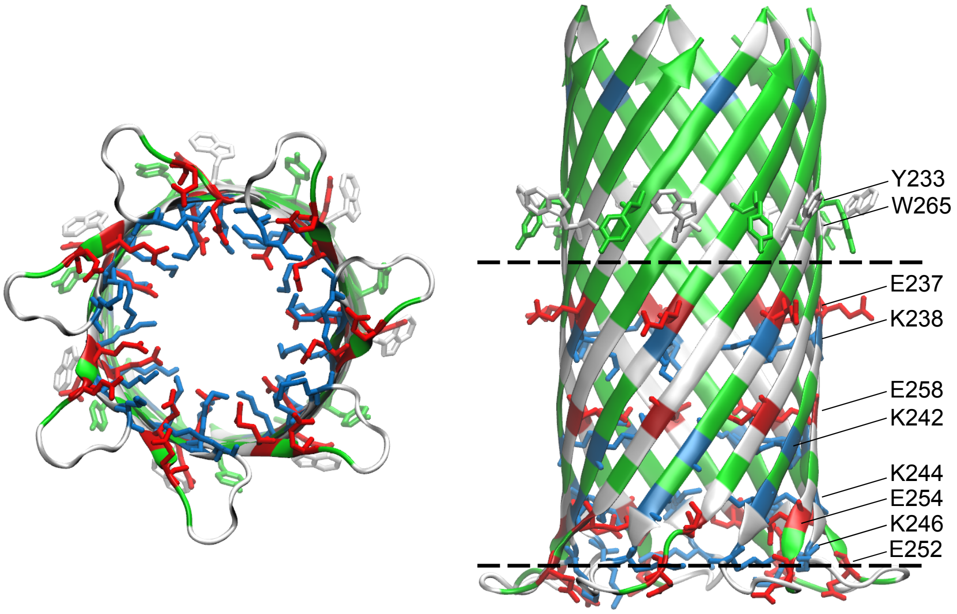


Figure 4.23: *Top view (left) and side view (right) of aerolysin's modeled β-barrel. Dashed lines indicate the approximate position of a bilayer's polar heads. Several polar residues are exposed to the pore lumen, and an aromatic belt is favorably placed next to lipids' heads, contributing to the pore's steady hooking.*

In order to properly model the outer assembly, aerolysin cannot be considered as a rigid body. In fact, as previously seen, loop extraction and CTP removal trigger large conformational changes both in secondary and tertiary structure. In order to keep into account this flexibility, we exploited the information obtained via our MD simulations of loop extraction reported in Section 4.4.1. From the simulation labeled WT_loop_extracted we extracted therefore 2000 different protein conformations. In every extracted snapshot the loop is a flexible, unstructured region being solvent exposed. For this reason, residues already modeled in the β-barrel were removed from the snapshots. This new collection of different conformations was used to model the outer assembly by means of POW, which was requested to produce a heptamer having a circular symmetry, matching height (less than 45 Å) and width (150 ± 5 Å) or the pore structure (disc shaped) in the known density map. As an additional restraint, we imposed G270 to be as close as possible to the pore's center. This restraint was adopted to ensure that residue G270 would be favorably placed to connect with the previously modeled β-barrel. We obtained a model having a cross-correlation coefficient (CCC) with respect of the available density map equal to 0.74. The structure was further optimized by means of a protocol based on Molecular Dynamics Flexible Fitting (MDFF), which generated a structure having a CCC equal to 0.89.

As a final step, we aligned both the β-barrel and the outer assembly so that their symmetry axis is at the origin, aligned along the z axis. Subsequently, the barrel was shifted along the z axis so that its aromatic belt would be located at the same height of the outer assembly's binding sites. The two structures were finally connected by binding residues S228 and G270 of the outer assembly with residues K229 and N269 of the modeled barrel. Not surprisingly, the bond within the two structures was extremely stretched. For this reason, we relaxed the strands 268-283 and 214-231 with 1000 conjugate gradient minimization steps followed by a simulated annealing protocol (see Material and Methods 4.4.3). The obtained model is represented in Figure 4.24.

Importantly, this model is consistent with all the structural features already known about aerolysin heptamer, and contributes to the understanding of some unexplained observations. The binding sites of Domain 1 and 2 are located on the outside of the assembly disc, and are oriented towards the membrane. This position is perfectly suitable for the binding of a GPI-anchor. Interestingly, the protonation state of residue H132, not far from Domain 2 binding site, has been shown to influence ability of aerolysin to heptamerize [19]. It was though that this residue had a role in the local arrangement of neighboring sidechains, since it was not expected to be in contact with any neighboring monomer. In in our model, however, H132 is located just in front of Domain 1 binding site, the possibility that this residue influences the binding of two adjacent monomer or the binding of a monomer to a GPI-anchor cannot therefore be ruled out. As seen in Section 4.2.1, proteolytic cleavage is not sufficient to remove the CTP from its location on top of Domain 4. In previous models it was hypothesized that the hydrophobic patch located under the CTP was the transmembrane region; removal of CTP could therefore be driven by specific interactions with the lipid bilayer. In our model, however, the hydrophobic patch is completely covered by a neighboring monomer. This raises the interesting possibility that the CTP on a monomer could be removed upon heptamerization. This would be possible by assuming that the incoming binding partner (Domain 4) has a higher binding affinity to the hydrophobic patch than the CTP. The β-barrel displays a peculiar hydrophobic-hydrophilic pattern for a length consistent with a typical bilayer. The pattern terminates on an aromatic belt constituted by residues Y233 and W265, which stabilize the membrane inserted complex. The barrel is connected via two short loops to the outer assembly, giving the assembly itself some flexibility with respect of the barrel. Akiba et al. noted that aerolysin, as its homologue protein parasporin, displays a characteristic Ser/Thr track, which is expected to interact with the membrane [7]. In our model, several residues of the Ser/Thr track (T225, S228, S272, T273) are located just above the β-barrel aromatic belt, on the loops connecting it to the outer assembly. These residues could act as a guide for the newly extracted loop, helping its correct positioning with respect of the membrane, subsequently contributing to stabilize the membrane-inserted complex. The hairpin M140-G153, also supposed to interact with the membrane [92], is favorably oriented with respect to it. Interestingly, during our annealing protocol, the strands 268-283 and 214-231 were prone to structure in a β-sheet conformation. Interestingly, in some regions the sheets of neighboring monomers associated in a larger sheet. This hints the attractive possibility that these strands could assembly in an

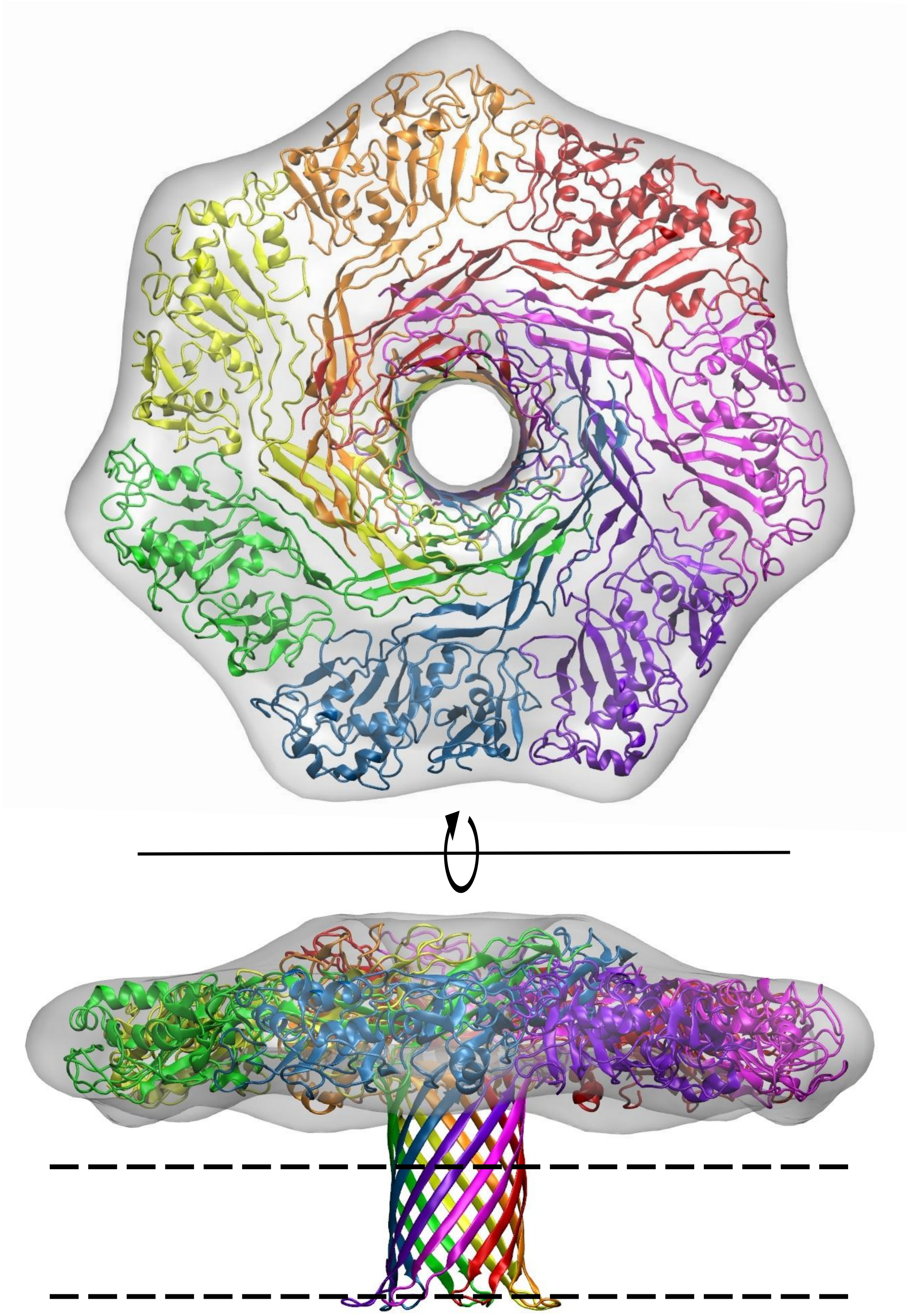


Figure 4.24: *Top and side view of WT heptamer atomistic model superposed to the density map produced in van der Goot Lab.*

additional barrel, which could confer higher robustness to the complex.

When inserting into the lipid bilayer, aerolysin's multimer rearranges from a mushroom (prepore) to a disk (pore) shape. By assuming Y221G heptamer as a good representative of aerolyin's prepore conformation, both states have now been modeled at an atomic resolution (see Y221G heptamer model in Section 4.3). In order to characterize the possible conformational changes involved in the transition from the prepore to the membrane inserted state, we morphed our Y221G heptamer into the WT one. Remarkably, the transition is possible without relevant clashes, and is mainly due to a rearrangement of Domains 4. In fact, in prepore state these regions are aligned almost perpendicularly with respect of the lipid bilayer. By twisting sidewise in a shutter-like fashion, Domains 4 ends up lying on on top top of each other. By doing so, the holes observed in Y221G cryo-EM map get sealed, which as a consequence increases the contact surface of adjacent monomers.

As presented in Section 4.3.2, mutations K198A, D216A, R282A, K369A, E367A and K351A lead to proteins displaying a delayed or even impaired heptamerization. As in the prepore model, in the pore model these residues are located at a protein-protein interface. Mutations K185A, D188A, K290A, E307A and K309A did not affect aerolysin heptamerization. Interestingly, while in the prepore model these residues are located at protein interface, in the pore model they are solvent exposed. This can indicate that these residues are unrelevant for a proper binding within two monomers, and further validate the models of the prepore and pore structure.

In conclusion our results show that aerolysin undergoes a large conformational change when switching from a prepore to a pore conformation. The prepore conformation detected via negative stain experiment presents a large similarity with the cryo-EM map, indicating that the latter could indeed represent aerolysin locked into a prepore conformation. By describing rearrangements at the domain level, specially within Domain 3 and 4, we could model the disc shaped structure of aerolysin heptamer. The arrangement of proteins in our model, different with respect of other β pore-forming toxins such as *Staphilococcus aureus* α-hemolysin or *Vibrio cholerae* cytolysin, is consistent with a variety of known biochemical and structural data.

4.4.3 Materials and Methods

Steered molecular dynamics protocol

WT and CTP-WT were assembled and equilibrated according to the same protocol detailed in the previous section. Subsequently, rotational constraints as implemented in NAMD 2.6 were applied on C_α atoms in region E236 to A265. The rotation axis was identified as the normal of the plane identified within residues S236, N262 and G251. S235 was selected as the rotation hinge. A biasing force was applied so that the selected atoms would rotate of 180 degrees around the rotation axis fixed on the hinge in 1 ns. The obtained systems were simulated for 200 ns according to the same protocol detailed in section 4.3.3.

Pores models

In order to construct aerolysin's transmembrane barrel, the rivet model was adopted as initial seed [57]. This model, based on cysteine scanning experiments, is constituted by seven hairpins containing residues K230 to A260. By using α-hemolysin barrel as template (pdb: 7AHL), the rivet model was elongated to match α-hemolysin barrel's length. This was done by first aligning the two barrels, transferring the backbone coordinates of α-hemolysin to aerolysin barrel for every missing residue, and finally reconstructing aerolysin's side chains on the newly built backbone (using psfgen tool in VMD). The final model, comprising the region S228 to G271, was minimized with 1000 steepest descent steps, inserted into a POPC bilayer, solvated, and equilibrated at 300k and 1 Atm for 1 ns using NAMD and CHARMM force field.

Outer assembly model

From the simulation labeled WT_loop_extracted one frame every 100 ps was extracted, which gave us a total of 2000 different snapshots. Since residues K229 to N269 were separately modeled in the β-barrel region, we removed them from every snapshot. We adopted a POW setup identical to what described in Section 4.3.3. No model matched the filtering criteria (fitness smaller than 0). The best model (fitness equal to 47.9) respected all the provided geometric constraints, but presented few mild backbone clashes, which explain the fitness value greater than zero.

It has to be noted that the assembly was obtained from snaphots containing aerolysin's major lobe only (Domains 2, 3 and 4). For this reason, using HEX [120], we docked Domain 1 within two monomers extracted from the assembly. A new complete model was obtained by applying a 7-fold symmetry on Domain 1's best scoring pose.

Our outer assembly model was subsequently optimized by means of a Molecular Dynamics Flexible Fitting (MDFF) protocol as implemented in NAMD [135], using CHARMM force field. This protocol was meant to improve the fit in the available density map. In order to avoid overfitting, proteins' secondary structure was restrained. 1 ns simulation was performed *in vacuo* at 300 K and 1 Atm, with SHAKE restraining all bonds and a dielectric constant of 80. 5000 conjugate gradient minimization steps were finally performed. In order to compare the all atom assembly to the target cryo-EM, a density map of the all atom assembly having the same resolution of the cryo-EM was generated.

Model refinement

After having connected the β-barrel with the outer assembly as described in main text, atoms in regions 268-283 and 214-231 were first minimized with 1000 conjugate gradient minimization steps using NAMD. All the other atoms were kept fixed, and the computation of the forces within fixed atoms was excluded. This minimization phase ensured that the bonds S228-K229 and N269-G270 were not anymore overstretched. Subsequently, a simulated annealing protocol in implicit water (dielectric equal to 80) was performed on region 268-283 and 214-231. Again, the rest of the protein was kept frozen. Simulation was run in the NVT ensemble, having temperature controlled with Langevin dynamics, with 1 fs timestep. The system was brought to 10 to 310K in 300 ps, increasing the temperature by 2K every 2 ps, and subsequently kept constant for 100 ps. The system being equilibrated at 300K, the temperature was subsequently raised to 410 K in 50 ps, kept constant for 50 ps, and decreased back to 310 K in 100 ps. This annealing cycle was repeated 10 times. At the end of this cycle, the system was equilibrated at 300K for 1 additional nanosecond.

Heptamers morphing

Morphing Y221G heptamer into WT one was done by means of Chimera's "Morph Conformations" tool [115]. 100 intermediate structures were produced using the corkscrew interpolation method. In order for the two maps to contain exactly the same atoms, Y221 sidechain was removed from WT heptamer model.

4.5 Properties of transmembrane β-barrel

Upon heptamerization, aerolysin forms a transmembrane heptameric β-barrel. Even though crystallization of the multimeric membrane-inserted conformation of aerolysin has not, up to now, been successful, residues being exposed to the pore lumen could be pinpointed via cysteine scanning. On the basis of these results, Iacovache et al. produced a model for aerolysin's β barrel [57] (hereafter "rivet model"). The model was created by exploiting structural homologies within aerolysin and *Staphilococcus aureus* α-hemolysin, which heptameric structure has been solved at 1.9 Å resolution [131]. As α-hemolysin, aerolysin's modeled pore consists of seven hairpins. However, aerolysin exposes much more charged residues (lysines and glutamic acids) to the inner lumen, and hooks to the intracellular side of the membrane via loops in a peculiar rivet-like conformation. It has been observed that aerolysin's multimeric form is more stable than α-hemolysin's one. Indeed, aerolysin withstands boiling in SDS [24]. It is also known that, upon aerolysin insertion, membrane permeability to calcium ions increases [70]. It is not proven, however, whether calcium flows directly through the aerolysin's pore, or is imported into the cell via another mechanism. In fact, some results even indicated that the pore could be anionic [24].

While little is known about aerolysin's pore properties, α-hemolysin's has been extensively studied since the second half of '80s both *in vitro* and *in silico*. It has been found that α-hemolysin is water filled and weakly anion selective at neutral pH. The channel displays however a voltage and pH gating mechanism [93]. It has been shown that the protein's selectivity can be altered by adding a specific noncovalent adapter, cyclodextrin, at the entrance of the β-barrel. Depending on the nature of this adapter, the pore can be made more anionic or cationic [49]. Importantly, residence time of cyclodextrin in the cavity can be dramatically increased by specific point mutations [48]. This feature, coupled with the protein stability, makes

α-hemolysin suitable as a single-molecule sensor device. Remarkaby, DNA has been shown to cross the pore [65], α-hemolysin is therefore a promising tool for fast DNA sequencing (see [18] for a review). Interestingly, a first attempt in the context of single-molecule detection (transport of unfolded proteins) has been recently proposed for aerolysin as well [112].

α-Hemolysin pore's properties have been studied by a number of computational studies exploiting a variety of techniques. The current-voltage relationship has been studied with MD simulations biased by an electric field [8], one dimensional Nernst-Planck analysis [96], Grand canonical Monte Carlo Brownian dynamics and three dimensional Poisson-Nernst-Plank electrodiffusion algorithms [105, 107]. These works, generally consistent with conductance experiments, highlighted how the asymmetric distribution of charges determines the pore's asymmetric conductance. The role of of cyclodextrin was studied both with standard MD (by restraining protein's backbone) [129], with multiscale MD [83] and with grand canonical Monte Carlo Brownian dynamics [37]. The two latter studies found that in the presence of cyclodextrin ions get partially desolvated, which increases the strength of the electrostatic interactions with polar residues at the pore mouth. MD simulations also noted that the β-barrel cross-section fluctuates around an elliptical conformation [8].

In this section we present the first computational study on aerolysin's pore. By means of MD simulations and Poisson-Boltzmann calculations we compare aerolysin and α-hemolysin pores electrostatic properties with respect to Na^+, Ca^{2+} and Cl^- ions. Via a simulated electrophysiology experiment, we estimate pore's conductivity for positive and negative charges. Finally, the flexibility of the two pores is compared.

4.5.1 Aerolysin's pore model

Unlike α-hemolysin, nowadays no structure of the transmembrane β-barrel is available. For this reason, in this study we adopted the aerolysin β-barrel model produced in Section 4.4.2. In order to characterize aerolysin's β-barrel properties and highlight the differences with α-hemolysin's, we superposed the backbones of the two structures, aligned them along their z axis, inserted them into a modeled POPC bilayer patch, solvated the whole systems with explicit water molecules and ran two 150 ns molecular dynamics (MD) simulations. Using an ensemble of structures extracted from these simulations, we assessed average pores' internal radius (Figure 4.26 A1). Aerolysin's pore results to be narrower owing to the rather long side chains exposed to its lumen. Its constriction point is located at z=-2 Å, and corresponds to residue K424. Differently, α-hemolysin constriction point is located right below the barrel's external mouth, and corresponds to residue M133 (Figure 4.26 A2).

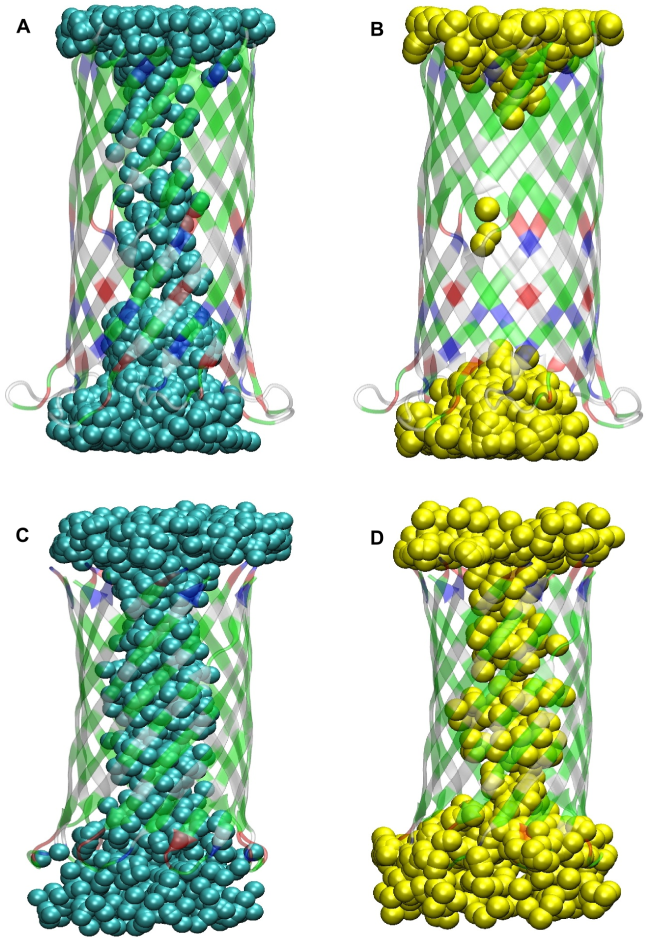


Figure 4.25: *(Cl^- (left, in cyan) and Na^+ (right, in yellow) ion positions along MD simulations of aerolysin (top) and α-hemolysin (bottom). In aerolysin, no cation populates the pore along the simulation, while α hemolysin allows the transit of both anions and cations.*

4.5.2 Pores electrostatic profile

By observing the aerolysin and α-hemolysin simulations performed above, we notice that no positive charge (Na^+) crosses aerolysin's pore along the trajectory, whereas several negative charges (Cl^-) do it (Figure 4.25). Conversely, in α-hemolysin simulation both anions and cations populate the pore. This observation makes us conclude that an energy barrier is blocking the diffusion of positive charges in aerolysin, whereas α-hemolysin seems to be less selective. The much more charged nature of aerolysin's pore lumen, therefore, affects ion's flow.

In order to quantify the electrostatic effects of the pores side chains on ions, we performed an ensemble of Poisson-Boltzmann calculations. This allowed us to estimate the Potential Mean Force (PMF) profiles of aerolysin and α-hemolysin pores for Ca^{2+}, Na^+ and Cl^- ions. Results for aerolysin (Figure 4.26 C1), indicate that an important barrier for positive charges is located in correspondence of the constriction point, at z = -2 Å (Figure 4.26 A). This corresponds to the location of the sidechains of K242 and K244. Even though negatively charged residues E254 and E258 are also present in the barrel, their electrostatic effect is screened by neighboring lysine residues.

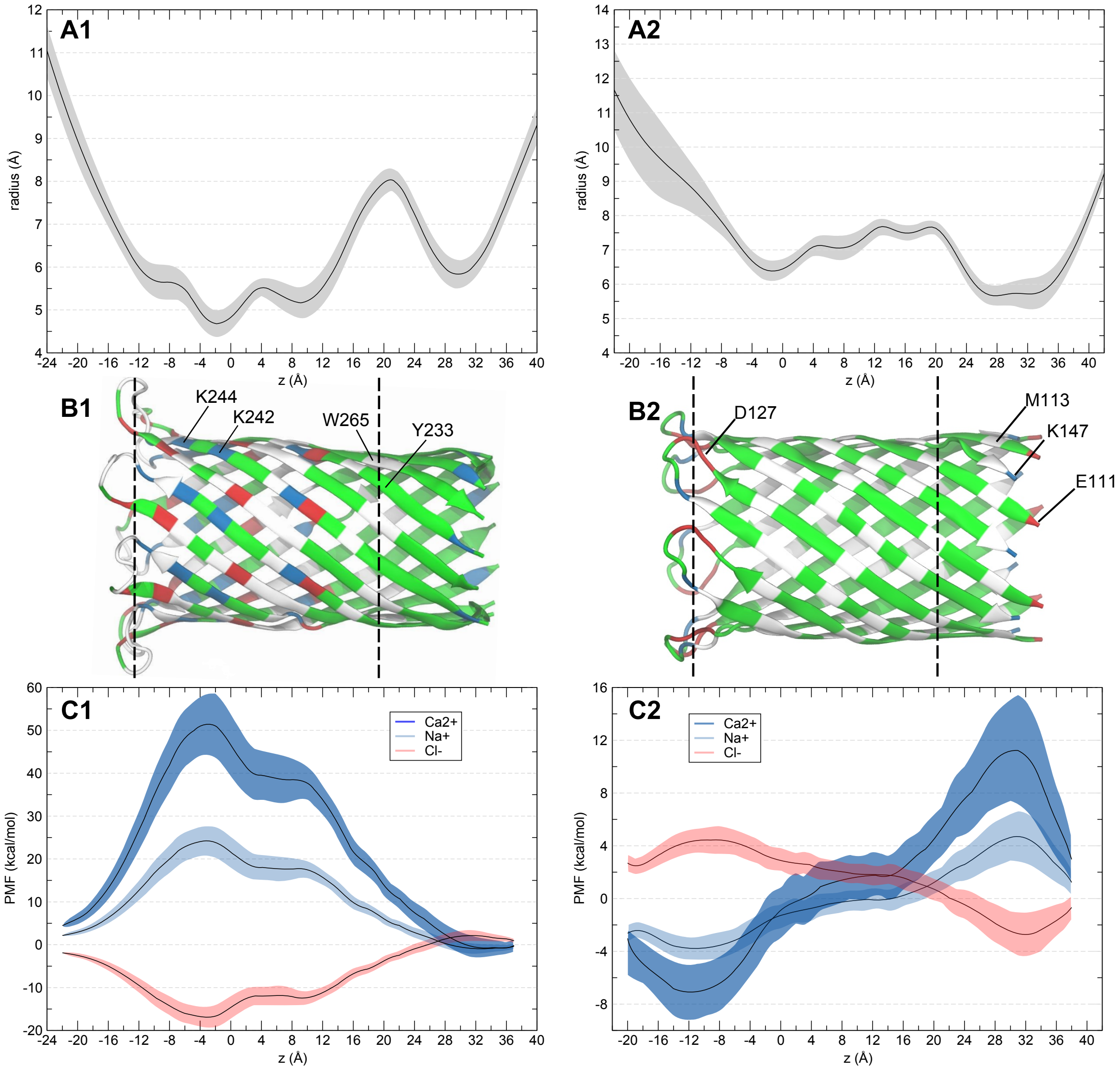


Figure 4.26: *Pore radius and PMF measurements for aerolysin (1, left) and α-hemolysin (2, right). (A) accessible surface radius along the pore, averaged on the performed MD simulation (grey area being the standard error). (B) aerolysin and α-hemolysin β-barrels. Dashed lines represent the position of lipid bilayer's phosphate groups. (C) PMF for Ca^{2+}, Na^{+} and Cl^{-} along the pore axis averaged on the performed MD simulation (colored areas being the standard error; notice the difference in PMF scale).*

Results for α-hemolysin are strikingly different (see Figure 4.26 C2): the potential felt by every ion is indeed low along the whole pore lumen, increasing only at its distal ends. As previously seen, α-hemolysin's pore external mouth also represents its constriction point. Just above this region, residues E111 and K147 face the pore entrance. In this area, lysine's longer side chain produces a mild barrier for cations. Still, this constriction point is less selective of aerolysin's one owing to the presence of neighboring residue E111 and to the slightly larger pore radius (about 1.5 Å larger in average). On the opposite side, the pore's cytosolic mouth hosts the D127 belt, which constitutes a mild barrier for anions. Even though residue K131 is located right below, little screening takes place. This is due to the larger pore radius in this region, to the fact that K131 is located on a loop (undergoing therefore larger fluctuations), and to the presence of residue D128, with which K131 forms transient salt bridges.

To assess the effect of crossing the β-barrel on the solvation shells of one ion, we selected all Cl^- ions located next to aerolysin's constriction point (-4 Å<z<0 Å) along the whole trajectory (no Na^+ ions reaches this region) and computed their water Radial Distribution Function (see Figure 4.27). We can observe that Cl^- ions get partially desolvated when crossing the β-barrel. The rather large pore radius allows however some water to surround the ion. This rules out the possibility that pore selectivity might be related to ion's radius. Still, this partial desolvation decreases water screening effect, enhancing therefore the electrostatic effect of lysine residues located in the constriction.

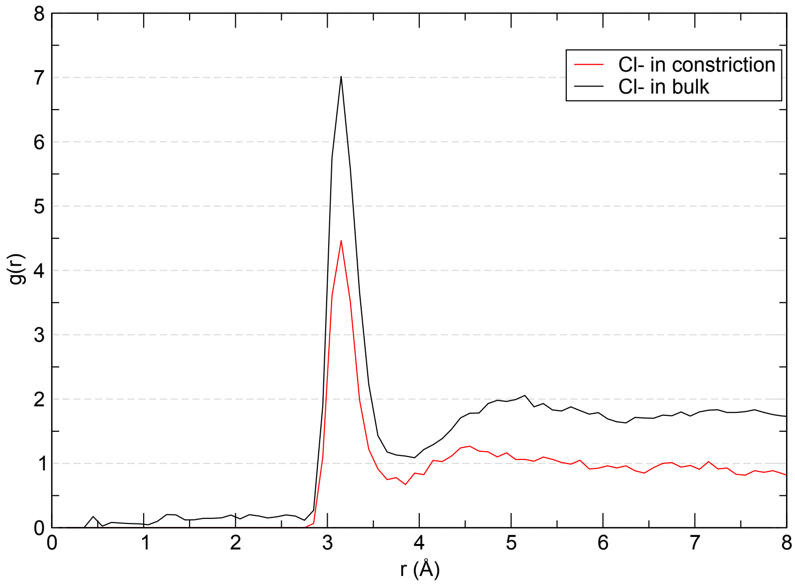


Figure 4.27: *Comparison of Radial Distribution Function of water around chlorine ions in the bulk and in the constriction of aerolyin's β-barrel. When crossing the constriction, ions get partially desolvated.*

4.5.3 Ionic current with imposed voltage

In vivo, ionic concentrations imbalance within the two sides of the bilayer generate an electric potential. Concentration and potential gradient play a role in the transport of ions through a channel. In order to determine whether or not an ion will cross aerolysin's β-barrel, it is therefore important to keep these gradients into account. For this reason, we simulated an electrophysiology experiment, with the aim of determining aereolysin's barrel conductance. This was done by simulating for 50 ns aerolysin's barrel while applying different intensities of constant electric fields along the z axis. We run simulations with -1, -0.5, 0, 0.5 and 1 $Volt/L_z$ (L_z being the box's z size) were run, and tracked all ions' positions.

In Figure 4.28 A the probability distribution of Na^+ positions along the z axis is shown, while the cumulate count of ions crossing the constriction point along time (z=-4 Å, see Figure 4.28 E) is shown in Figure 4.28 B. We can immediately observe that with positive voltages, Na^+ ions cumulate at the cytosolic mouth of the pore but just with a very high voltage, i.e. $1V$ few of them manage to overcome the constriction point. When negative Voltages are imposed, Na^+ can travel through the pore, and finally cumulate at the opposite side of the constriction point. Again, with a very high voltage only cations manage to cross the whole barrel.

Conversely, Cl^- can easily cross the whole barrel. In Figure 4.28 C, Cl^- density distribution indicate that anions can populate the whole pore, getting attracted by the constriction region. Ion count on this point (see Figure 4.28 D) shows that the ion flow increases linearly with the imposed voltage. With ± 1 V hundreds of crossing occurrences are recorded during the 50 ns of simulated time. Interestingly, a slight preference for Cl^- traveling from cis to trans can also be observed.

Having observed that the cumulated ion count grows fairly linearly along time, we could estimate the value of ion's current by 10 ns windows along the whole simulation, and produce a I/V graph (4.28 F). It can be immediately observed that a large resistance is present for the transit of cations, whereas the resistance is one order of magnitude lower for anions. These observations indicate that aerolysin is most likely to be an anionic channel.

4.5.4 Pores Flexibility

s already observed in α-hemolysin by Aksimentiev et al. [8], we noticed that both aerolysin and α-hemolysin pores were fluctuating along our simulations. In fact, the assemblies were flexible and tended to assume an elliptic conformation. However, interestingly, aerolysin looked stiffer than α-hemolysin. This observation is supported by our measurements of pores radius previously performed. In fact, a way larger standard error can be observed in the internal mouth region of α-hemeolysin, as compared to aerolysin's. To characterize this behavior, we aligned the two pores and sliced them at three different heights (three different colored regions in Figure 4.29). We subsequently fitted to every slice an ellipse, using as reference the position of the C_α atoms. The difference within the long and the short axis of the

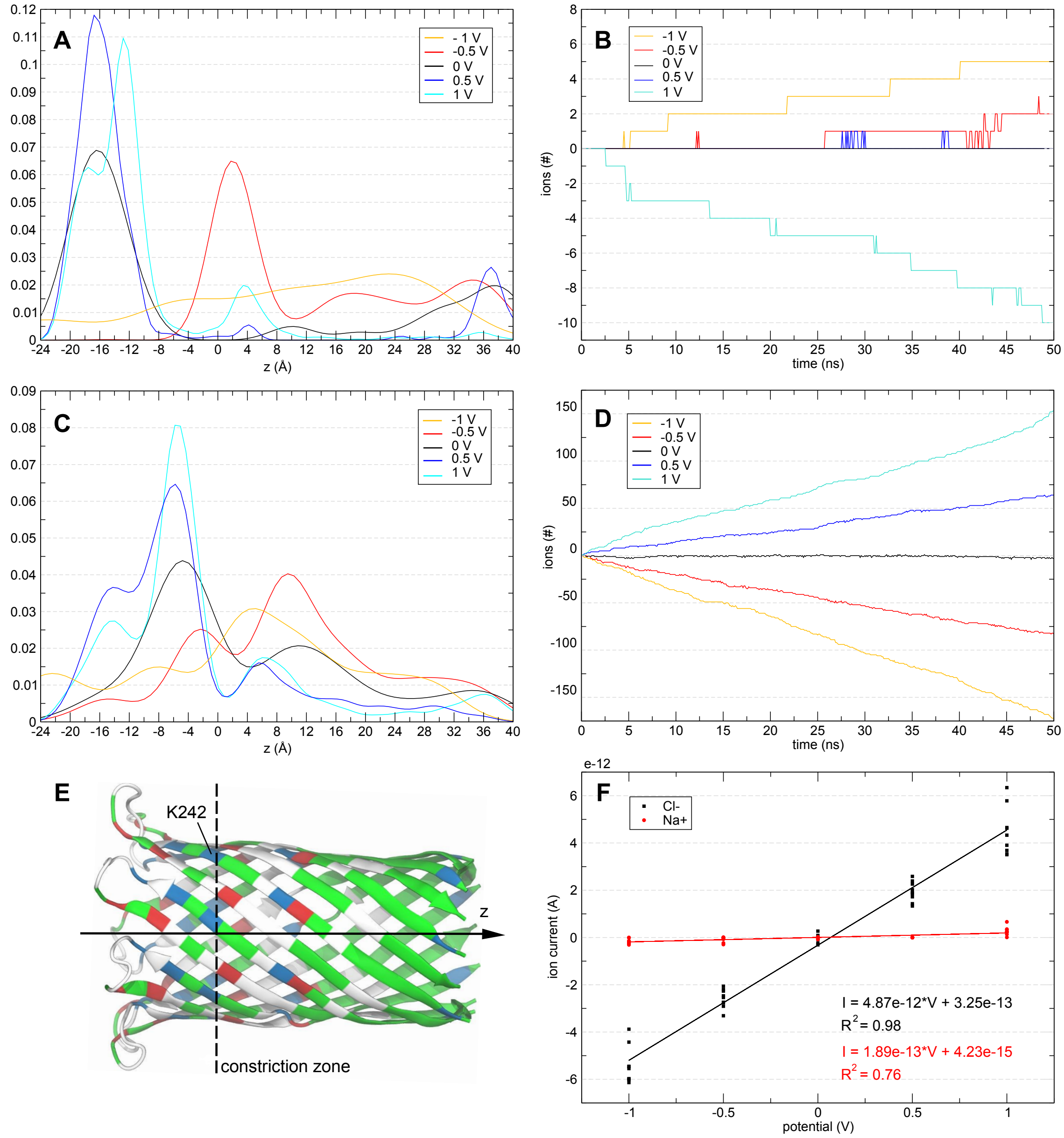


Figure 4.28: *In the first row, (A) the density distribution along the z axis and (B) the cumulated count of transits through the constriction point with 4 different imposed voltages are shown for* Na^+ *ions. In the second row, (C) the density distribution along the z axis and (D) the cumulated count of transits through the constriction point are shown for* Cl^- *ions. Image E shows aerolysin* β*-barrel. The location of its constriction point, as well as the direction of z axis are highlighted. In image F, the I/V graph for anions and cations is shown.*

fitted ellipse would serve as an indication of the barrel's deformation. The smaller this value is to zero, the more the barrel is circular. Averaging this measure along the simulations allowed us to obtain a statistic about the pores fluctuations. Results clearly indicate that aerolysin fluctuates less and tends to be more circular than α-hemolysin (Figure 4.29). Also, the slices taken at three different heights display a similar behavior. Conversely, α-hemolysin slices taken at the top and center of the pore resulted equally broader and shifted towards an higher deformation. Importantly, the measure taken at the pore exit resulted even more flexible.

These findings prompted us to determine the reason of such a difference. We hypothesize that higher rigidity in aerolysin is due to its the longer side chains in correspondence of the slicing points. The side chains would reduce the effective pore radius and hinder the pore deformation. The even higher flexibility of α-hemolysin in correspondence of the pore exit might also be dependent on the absence of a rivet structure, present in aerolysin.

ID	SIMULATION DESCRIPTION
A.1	aerolysin WT from model
A.2	aerolysin K244T + E254G
A.3	aerolysin with α-hemolysin loop (G126 DDTGKI)
B.1	α-hemolysin WT from crystal (pdb: 7AHL [131])
B.2.	α-hemolysin T125K + G133E
B.3	α-hemolysin with aerolysin's rivet (F245 KWPLVGET)

Table 4.1: *Summary of the simulated chimeric and WT β-barrels*

To validate this hypothesis, we performed *in silico* mutagenesis experiments. To assess the effect of side chains length on pore flexibility, we produced two chimeric pores by exchanging the residues of aerolysin and α-hemolysin in correspondence of the slicing point displaying the highest fluctuation difference. This is the slice closest to the pore's cytosolic mouth, highlighted in green in figure 4.29. To assess the effect of the loops conformation, we produced two more models exchanging aerolysin's rivet with α-hemolysin's loops. All these structures, along with new WT aerolysin and α-hemolysin pores, were aligned, inserted in the same POPC membrane patch, solvated, and simulated for 80 ns. A summary of performed simulations is presented in Table 4.5.4.

Results confirmed our hypotheses. Double point mutations transformed the fluctuation patterns of α-hemolysin into the aerolysin ones and viceversa (see figure 4.29 and Table 4.5.4). The longer side chains exposed in aerolysin's pore lumen contribute therefore to the stiffness of the whole barrel. Additionally, it appeared that the rivet of aerolysin does affect the pore stability as well. Indeed, α-hemolysin could be stabilized by replacing its turn region with aerolysin's rivet. Viceversa, aerolysin was destabilized by replacing its rivet with α-hemolysin's turn.

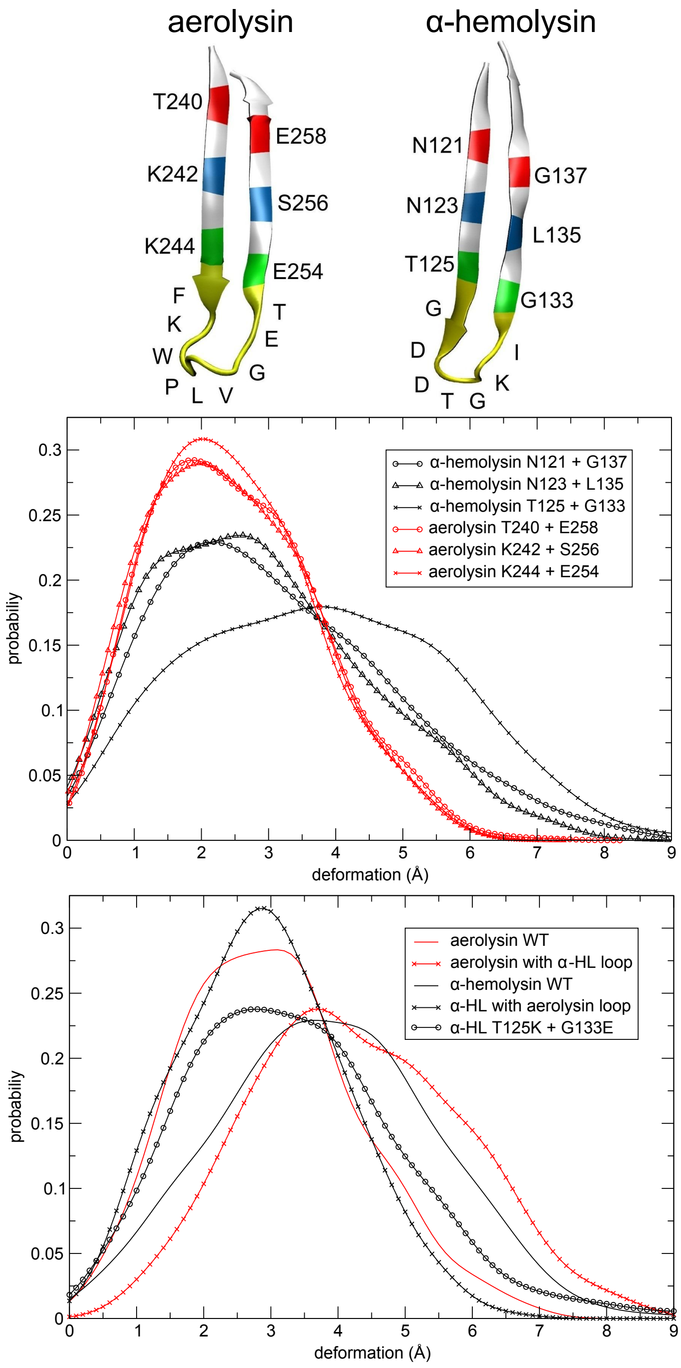


Figure 4.29: *Hairpins of aerolysin (top left) and α-hemolysin (top right). Fluctuation measures have been performed in correspondence of colored regions. (center) Deformation probability of aerolysin and α-hemolysin in three different points (colored regions above). Aerolysin is more stable, and fluctuation patterns are the same along the whole pore. Conversely, α-hemolysin is less stable, and the region neighboring the hairpin's turns shows broader fluctuations. (bottom) Deformation probability of aerolysin and α-hemolysin chimeric pores next to cytosolic mouth. α-hemolysin can be stabilized by mutating its bottom ring to aerolysin bottom ring, or inserting a rivet structure.*

	A.1	A.3	B.1	B.3	B.2.3
A.1	1	0.53	0.62	**0.99**	**0.98**
A.3	0.53	1	**0.98**	0.56	0.64
B.1	0.62	**0.98**	1	0.65	0.74
B.3	**0.99**	0.56	0.65	1	**0.97**
B.2.3	**0.98**	0.64	0.74	**0.97**	1

Table 4.2: *Correlation coefficients within all the fluctuations of performed simulations. Switching residues or loops within α-hemolysin and aerolysin barrels switches also the fluctuation patterns with very high correlation (greater than 95%)*

4.5.5 Discussion

According to obtained results, positive charges have to overcome a relevant energy barrier in order to cross aerolysin's pore from side to side. As shown, ions are not stripped of their solvation shell while crossing the pore. This indicates that ion selectivity should be due to electrostatics interactions only. Interestingly, a study on several pore forming toxins shown that their specific selectivity is substantially due to charge and location of residues in the channel [94]. It has to be pointed out that the Poisson-Boltzmann calculations here performed do not indicate whether an ion will cross the pore or not, but only what kind of energy profile the ion will have to face while doing it. The rate of ions diffusion is indeed also determined by their concentrations on the two sides of the membrane, as well as the membrane's electric potential gradient. Our simulated electrophysiology experiment, however, also highlighted how an important voltage has to be imposed in order to allow positive charges to flow through aerolysin's barrel, while negative charges flow easily through it. As a whole, our results indicate that it is unlikely that Ca^{2+} could cross the pore by means of passive transport and that, therefore, the pore should be anionic. On this point, this work supports the observations of Chakraborty et al [24]. The increased permeability to calcium ions observed by Krause et al [70], might rely on a more complex mechanism. Interestingly, our results indicate that aerolysin should have a much higher selectivity than α-hemolysin. Electrophysiology experiments could provide further insights into aerolysin's pore properties.

Similarly to α-hemolysin, aerolysin's barrel cross-section is also elliptical. This feature, possibly common to all membrane-inserted pores, should be kept into account in any computational model studying pores' ionic conductance. Pore deformation and side chain flexibility might indeed have an important effect in this context. Our mutagenesis experiments made us conclude that side chains length affects pore stiffness. Indeed, we observe that, the longer the side chains, the less the pore will fluctuate. Increased stability is also provided by aerolysin's hooking mechanism to the inner membrane. Since aerolysin's heptamer has been shown to be more resistant to boiling than α-hemolysin's one, it is tempting to hypothesize that pore stiffness affects pore stability, which in turn is reflected by a higher resistance to boiling. Producing *in vitro* some of our chimeric pores could serve as a validation.

4.5.6 Material and Methods

Pores models

In order to construct aerolysin's transmembrane barrel, the original rivet model was adopted as initial seed [57]. This model, based on cysteine scanning experiments, is constituted by seven hairpins containing residues K230 to A260. By using α-hemolysin barrel as template (pdb: 7AHL) [131], the rivet model was elongated to match α-hemolysin barrel's length. The final model, comprising the region S228 to G271, was minimized with 1000 steepest descent steps. α-hemolysin pore, i.e. residues K110 to Y148, was extracted from crystal structure 7AHL. In order to produce the chimeric pores described in Table 4.5.4, VMD mutator tool [54] was used. In order to exchange the loop regions of the two barrels, WT aerolysin and α-hemolysin were first aligned on their backbone atoms. Atomic coordinates of loop regions were subsequently switched within the two. Every produced structure was minimized by means of 1000 steepest descent steps.

All produced β-barrels were inserted at the center of a 80 Å by 80 Å POPC bilayer. Systems have been solvated in a rectangular box of TIP3 water and ionized with 0.15 molar Na^+ and Cl^-. Systems had an average size of about 40.000 atoms. Periodic boundary conditions have been defined such that the membrane would form a continuous surface with its neighboring patches.

Molecular Dynamics Simulations

All MD simulations have been performed using the CHARMM27 force field [86] on NAMD molecular dynamics engine [116], with SHAKE algorithm on all the bonds, and PME treating the electrostatic interactions in periodic boundary conditions. We chose an integration step of 2 fs. All the simulated systems have been first minimized with 2000 conjugate gradient steps, and subsequently equilibrated. Water has been firstly equilibrated for 100 ps at 300 K (non-water atoms' movement restrained). We then equilibrated the membrane, by restraining just the protein's atoms at 300K for 100 ps. Finally, the systems has been warmed up from zero to 300K in 400 ps. From this equilibrated state, WT aerolysin and α-hemolysin were simulated for 150 ns in the NPT ensemble using Langevin dynamics to set the system at 300 K and 1 atmosphere. Using the same conditions, chimeric pores were simulated for for 100 ns. Simulations of WT aerolysin and α-hemolysin were also performed by imposing an additional biasing electric field E aligned along the z axis:

$$E = \frac{V}{L_z} \tag{4.1}$$

Where V is the voltage and L_z the dimensions of the simulation box along the z axis after equilibration. For both proteins, simulations with a voltage equal to -1,-0.5, 0, 0.5 and 1 were run for 60 ns.

Barrel Radius

One protein conformation per nanosecond was extracted from the last 100 ns of unbiased WT aerolysin and α-hemolysin simulations. 100 structures per simulation were therefore obtained. These were aligned along their z axis, and superposed according to the backbone atoms of β-sheet residues facing the pore lumen (β-sheet residues according to DSSP [64] algorithm applied on the initial model). In order to asses the barrel's accessible surface, a HOLE [130] run was performed on every structure, which measured the barrel radius every 1 Å. This lead to an ensemble of measuring points scattered along the pore z axis. Results were averaged via a sliding window having 1 Å size. For every window position, mean and standard error were computed.

Ions' Potential Mean Force

In order to assess the electrostatic potential felt by a specific ion along the pore, we performed Poisson-Boltzmann calculations using APBS [12]. One single measure requires three independent calculations: one for the barrel alone (pore), one for the ion alone (ion) and one for a protein plus ion complex (complex). The potential felt by an ion in a specific location is calculated as follows:

$$E = E_{complex} - E_{ion} - E_{pore} \tag{4.2}$$

We obtained a complete energy profile for an ion located along the central axis of a pore by repeating the above calculation on a set of measure points. Ions were therefore placed along the pore center as defined by HOLE at intervals of 1 Å. This method was applied to all the 100 barrels previously extracted. The 100 obtained profiles were finally averaged in order to obtain a mean energy profile. Such profiles (mean and standard error of every measure point) have been computed for $Na+$, Ca^{2+} and Cl^{-} ions.

Ion Density

Density of Cl^{-} and Na^{+} ions inside the lumen in pores simulated with a Voltage bias imposed was assessed by first extracting every 100 ps the z coordinates of all ions located at less than 7 Å from the protein. Ions distributions were subsequently obtained by running a gaussian kernel density on all extracted coordinates.

Ionic current

For every voltage-biased simulation, ionic current was assessed in correspondence of the barrel's constriction point as determined by HOLE measurements. The last 50 ns of all simulations were first aligned on the position of the constriction point. In aerolysin this point corresponds to residue K242, in α-hemolysin to M133.
Ions flowing through the constriction point were subsequently counted, using a timestep of 100 ps. Ion counter Δq would be incremented of one unit if an ion flows in the positive direction of z axis within two adjacent frames, and decreased otherwise. Ionic current was assessed in time frames of 10 ns, superposed of 5 ns. In these time frames, having a mostly linear behavior, the steepness of a linear regression was computed as to assess the local ionic current. A total of 8 current measurements were therefore obtained for every imposed voltage. Every current estimation was finally plotted against its corresponding voltage, and a linear regression was computed in order to assess pore's resistance.

Barrel fluctuation measurement

In order to assess the WT and chimeric pores deformation, we extracted the coordinates of C_α atoms corresponding to three different horizontal slices (see Figure 4.29). For every slice, 14 coordinates were extracted, i.e. one measure per β-sheet. In order to avoid biasing due to membrane equilibration, the first 20 ns of every simulation were excluded. Statistics were therefore performed only on the last 80 ns of simulation. The atomic coordinates extracted from every slice were fitted with an ellipse. The absolute difference within the long and short radius of the fitted ellipse provides an indication of pore deformation: the more the two radii are equal, the more the pore is circular and viceversa. This assessment of pore deformation was computed for every performed fit, and deformation probability distributions were finally determined by applying a gaussian kernel on the measures ensemble.

4.6 Conclusion

We have shown that aerolysin's C-Terminal peptide (CTP), which is 40 residues long, has a double function. On one hand, it helps protein folding in a water soluble form (qualifying therefore as an intramolecular chaperone), on the other it prevents premature oligomerization. Via *in silico* alanine scanning we highlighted that the CTP is mainly bound via hydrophobic interactions, and correctely predicted point mutations that would destabilize its binding to aerolysin's major lobe.

We have studied the role of mutation Y221G on aerolysin's structure. Experimentally, this mutation was known to lead to the formation of soluble heptamers, but the reason behind this effect was unknown. We propose that Y221G has a stabilizing effect on Domain 4. This region is highly dynamic, and locked in a specific inactive conformation by the CTP. CTP removal allows Domain 4 reorganization, and leads to Domain 3 loop extraction via strand unfolding. Mutation Y221G disables strand unfolding by anchoring the strand N269-S278 to aerolysin's core. Interestingly, this is the only strand consistently predicted as disordered by eight different disorder prediction algorithms.

By coupling POW with biased molecular dynamics simulations, we propose a model of Y221G soluble heptamer. The model is built on the basis of the available Y221G cryo-EM and molecular dynamics simulations of Y221G monomer allowing us to keep into account the protein's natural flexibility. The model has a high cross-correlation coefficient with respect of the known cryo-EM map (0.91), and is consistent with known biochemical and structural data. The model is validated by correctly predicting interfacing residues subsequently validated *in vitro*. We show that loop extraction from its resting place in Domain 3 affects domains arrangement. The direct consequence is that Y221G heptamer model cannot be considered as a reliable model for WT heptamer. We subsequently propose a model for WT heptamer on the basis of domain flexibility highlighted by our observations on loop extraction effects. Again, we obtain a model having a high cross-correlation coefficient (0.89) and consistency with experimental data. Together, Y221G and WT heptamer models shed a new light on the conformational changes aerolysin has to undergo in order to insert into the lipid bilayer, and draw a new paradigm on aerolysin orientation at the membrane surface.

We finally compare a modeled aerolysin's β-barrel with barrels from homologous protein α-hemolysin. We highlight that aerolysin's pore is narrower than α-hemolysin's, and that its large amount of charged residues facing the pore lumen generates a relevant electrostatic field. A simulated electrophysiology experiment subsequently highlighted how aerolysin's pore blocks the transit of positive charges while allowing a smooth transit of negative charges. On the basis of these observations, we predict that aerolysin should be anion selective, and that selectivity owns to purely electrostatic interactions. Finally, we observe that aerolysin's β-barrel fluctuates less than α-hemolysin's owing both to the presence of longer side chains exposed to pore lumen and to the anchoring effect of the rivet region. This highlights the importance of keeping into account the shape of transmembrane β-barrel's cross section when attempting to predict relevant interactions within the pore and molecules crossing it.

In conclusion, our studies exhaustively characterized aerolysin's mode of action, from the activation of a single monomer to the final heptameric assembly. Our observations, backed up by *in vitro* and *in vivo* experiments, can be instrumental for pharmaceutical and biotech applications. On the first aspect, we foresee the possibility of disabling aerolysin's activation process by means of specific molecules mimicking the role of the CTP. On the second aspect, we believe that aerolysin, as α-hemolysin, could be effectively exploited as a device for single-molecule experiments.

5 *Yersinia enterocolitica* Injectisome

5.1 Biological Background

The injectisome is a nanomachine evolutionary related to the flagellum [87], allowing specific bacteria to inject effector proteins across eukaryotic cell membranes, a process called type III secretion (T3S). More than twenty five different species of Gram-negative bacteria that interact with live animals, plants, nematodes or insects, either as pathogens or as symbionts are endowed with this expo export apparatus. The activity of T3S effectors allows bacteria to invade non-phagocytic cells or to inhibit phagocytosis by phagocytes, to downregulate or to promote pro-inflammatory responses, to induce apoptosis, to prevent autophagy, or to modulate intracellular trafficking [98].

Injectisomes have evolved into seven different families [47, 109], however the injectisomes found in most free-living animal pathogens belong to only three families. The Ysc injectisome of *Yersinia* represents the archetype of one of these families. Similar injectisomes are found in *Pseudomonas aeruginosa* [121] and in the fish pathogen *Aeromonas salmonicida* [20] for instance. The injectisomes from *Shigella flexneri* and *Salmonella enterica Typhimurium SPI-1* represent archetypes of a second family, which is largely distributed among animal pathogens. The archetypes of the third family are found in enteropathogenic and enterohemorrhagic *E. coli.*

The injectisome is made of about 25 different proteins, and consists of a basal body surmounted by a hollow stiff needle that projects from the bacterial surface into the exterior milieu [71]. The basal body consists of three rings: a large cytosolic ring (C-ring), a ring spanning the plasma membrane and peptidoglycan (MS-ring) and a ring traversing the outer membrane (OM-ring). The MS-ring contains five integral membrane proteins forming the export channel across the plasma membrane while the C-ring is thought to contain the essential ATPase complex. The needle is hollow and stiff, and is composed of proteins arranged into a helical symmetry. During morphogenesis the needle components travel through the growing structure and polymerize at its distal end [79]. In the tail of bacteriophages, it appears that one protein acts as a tape measure or a molecular ruler [55]. The molecular ruler, in its elongated state, determines the number of subunits of another protein that are allowed to polymerize to create the tubular structure. The same concept has been proposed to determine injectisome's needle length, but in this case the situation is more complex since the needle is assembled outside from the bacterial cytosol and the exact mechanism remains a matter of debate (see Figure 5.1).

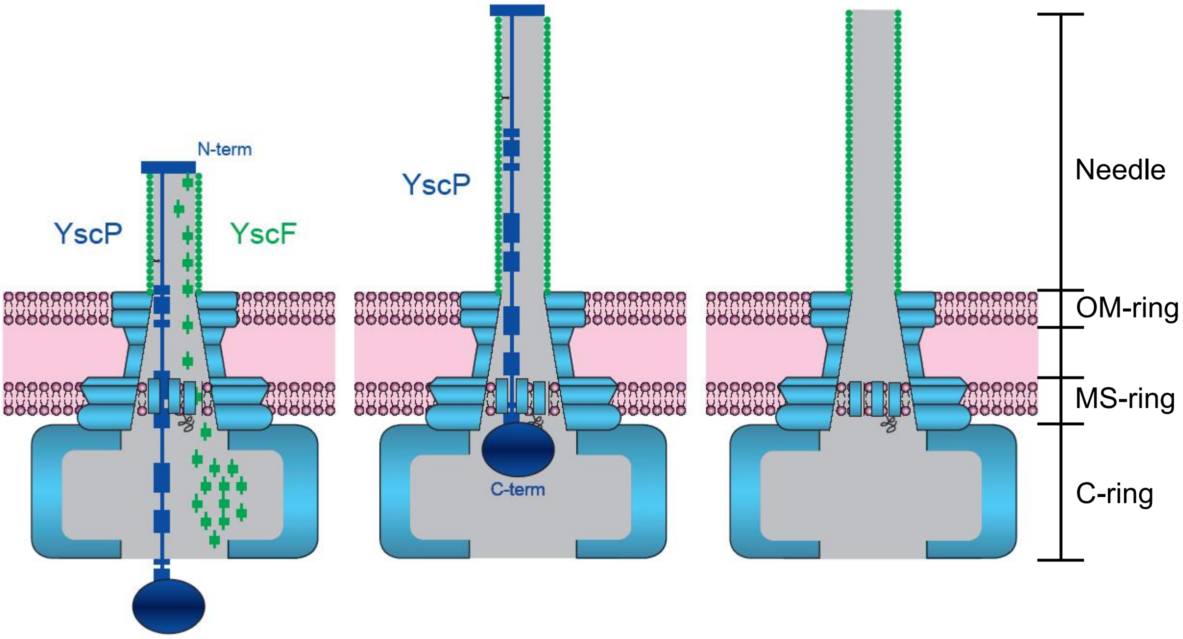


Figure 5.1: *Assembly process of injectisome's needle. Proteins constituting needle's building blocks travel inside the growing structure assembling into needle's distal end, while another protein is responsible of measuring the structure's length. In* Yersinia enterocolitica *these proteins are YscF and YscP, respectively. The basal body consists of three rings: a large cytosolic ring (C-ring), a ring spanning the plasma membrane and peptidoglycan (MS-ring) and a ring traversing the outer membrane (OM-ring). Image courtesy of Cornelis lab, Basel.*

The assembly of the injectisome requires a complex and tightly regulated assembly process. The final structure is extremely large, and spans the double Gram-negative bilayer. This makes obtaining structural and functional information at atomistic resolution extremely difficult. Here, we will study *Yersinia* injectisome, pursuing two objectives:

- study the mechanism through which the "ruler" protein determines the length of the injectisome's needle
- model the entire MS ring

5.2 The mechanism behind YscP molecular ruler

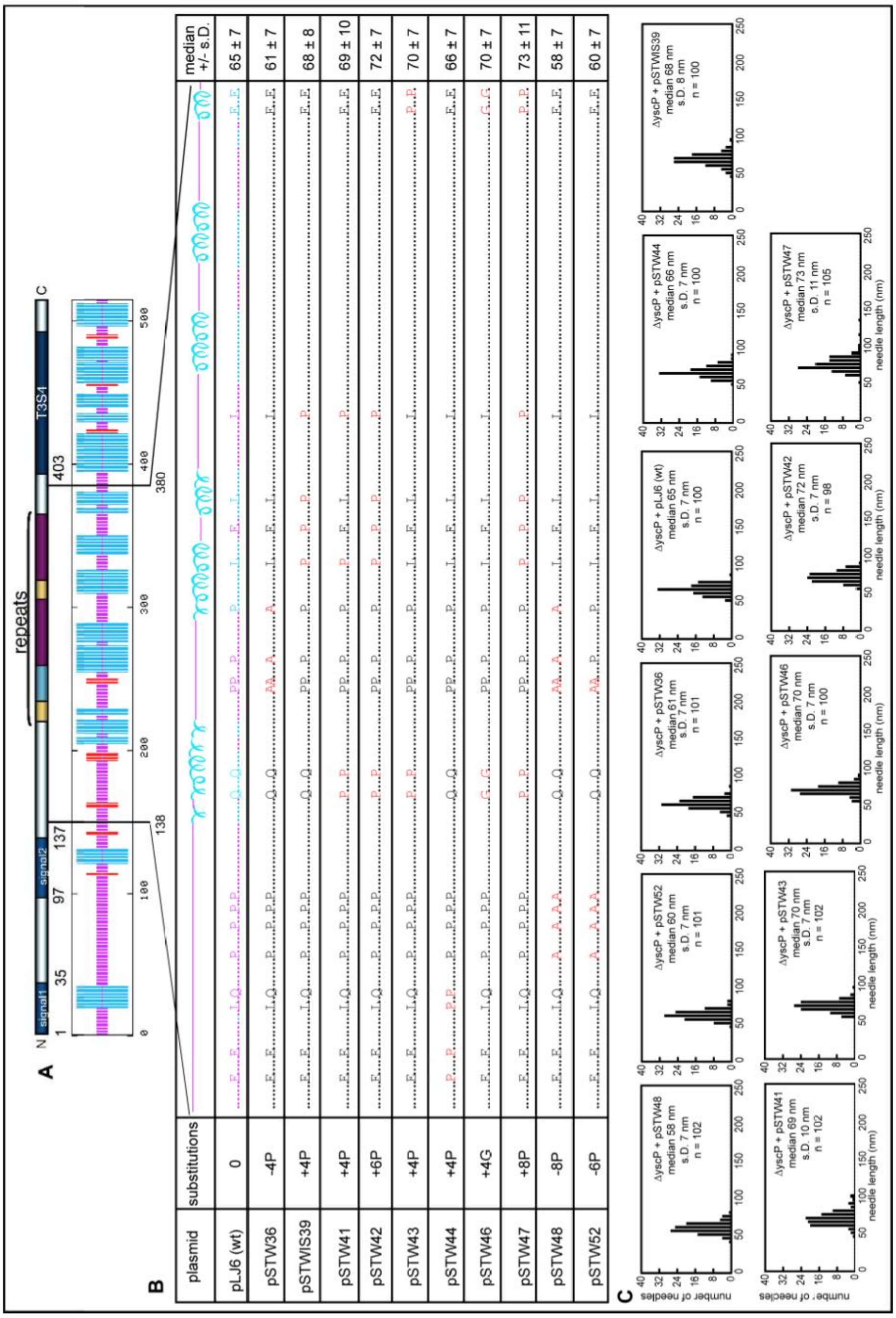
The *Yersinia enterocolitica* Ysc injectisome terminates with a 65-nm-long needle, made of about 140 copies of the YscF protein [51]. It is known that the switch to export late substrates is triggered by a protein, which is itself exported [22]. In *Yersinia enterocolitica*'s injectisome this protein is called YscP. This switch function has been assigned to a specific domain of YscP, predicted to have a globular structure and called T3S4 for Type 3 Secretion Substrate Specificity Switch [22]. Residues deletions and insertions in YscP (outside the T3S4 domain) lead to shorter and longer needles, respectively, with a linear correlation between the size of YscP and the needle length [62]. This led to a model where YscP acts as a molecular ruler measuring the needle. According to the model, export of needle subunit proteins would be allowed until the length of the needle reaches the length of the extended YscP ruler domain after which, the T3S4 domain of YscP would signal the secretion apparatus to stop exporting needle subunits. YscP is thus a protein with a dual function, ruler and substrate specificity switch.

We aim at proving that not only the number of residues, but also the secondary structure of YscP plays a role in determining the needle length. In particular, we demonstrated that the predicted length of various YscP variants in their functional state approximates to the actual needle length [140].

Experimental studies performed in the lab of Guy Cornelis at the Biozentrum in Basel were based on the creation of nine YscP mutations. Said mutations were different to each other in terms of their secondary structure: by mutating to a proline (known to be an helix breaker) residues being part of an helix, the helical content of YscP was reduced, whereas substituting prolines with alanine (an helix friendly residue) the helical content was increased (see Figure 5.2). The lengths of needles generated with these mutations were subsequently measured by electron microscopy. Experimentalists have however to face a challenge: since YscP is located inside the growing needle, no direct observation of the process of needle growth is possible. Numerical simulations can therefore provide precious information related to YscP secondary structure upon stretching that could not be gathered in any other experimental way.

5.2.1 YscP helical content modulates needle length

An *in silico* molecular dynamics approach was used to simulate the pulling of YscP starting from a random initial conformation, which would mimic the extension of YscP upon needle growth. We modeled the secondary structure of YscP wild type and the same nine mutants produced by experimentalists, and ran a series of independent steered MD simulations by fixing the N-Terminal region of YscP while pulling the C-Terminal at constant velocity. Given the extended nature of YscP, pulling the protein in explicit water requires a very large simulation box. For this reason, all the mutants were pulled and measured 20 times *in vacuo*. In order to assess the difference within a steering experiment *in vacuo* and in explicit water, we pulled



plasmid	substitutions	median +/- s.D.
pLJ6 (wt)	0	65 ± 7
pSTW36	-4P	61 ± 7
pSTWIS39	+4P	68 ± 8
pSTW41	+4P	69 ± 10
pSTW42	+6P	72 ± 7
pSTW43	+4P	70 ± 7
pSTW44	+4P	66 ± 7
pSTW46	+4G	70 ± 7
pSTW47	+8P	73 ± 11
pSTW48	-8P	58 ± 7
pSTW52	-6P	60 ± 7

Figure 5.2: *Effect of amino acids substitutions in YscP on the needle length. A. Schematic representation of YscP. Secondary structure predicted by SIMPA96. Blue bars = α-helices; red bars = β-sheets; purple bars = unstructured. B. Table showing the substitutions introduced between residues 138 and 380. Glycines were also introduced into predicted α-helices. C. The histograms of length measurements. s.d., standard deviation; n, number of measured needles.*

YscP WT and two of its mutants in explicit water as well.

From the analysis of YscP extension under constant velocity pulling, we found that the helical regions are conserved within the ns timescale, and upon pulling YscP models extend maintaining their initial secondary structure. Approaching a fully linear extended conformation, α-helices abruptly start to break under mechanical stress, eventually reaching a completely extended β-strand-like conformation.

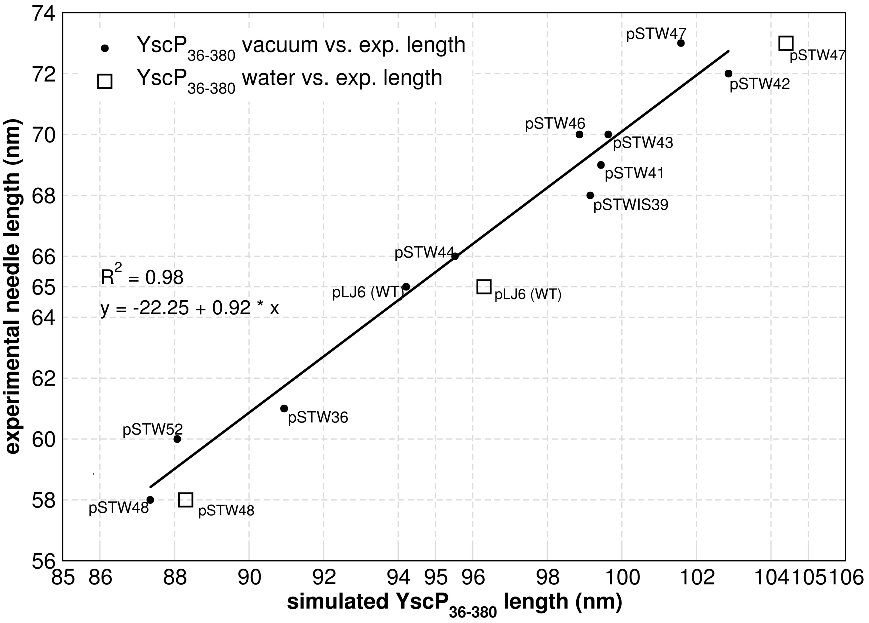


Figure 5.3: *Correlation between the YscP number of residues and the length of needles present in various different wild-type strains. The needle length is plotted against the total number of residues in YscP.*

Interestingly, we found that when the initial helical content of YscP is preserved, the length of the wild-type protein (93.5nm) along with the mutant conformations is strikingly correlated with the measured needle length, showing a constant difference of about 29 nm between the length of YscP and the length of the needle (see Figure 5.3). Importantly, no relevant differences within pulling *in vacuo* and in explicit water could be observed. Figure 5.4 plots the length values observed in MD simulations before the unfolding of any α-helix of YscP against the relative number of prolines (upper slope), whereas the lower slope shows the needle length measured experimentally. It can be noticed that the distance within the two slopes is constant (29 nm). This value corresponds approximately to the distance between the cytoplasm and the surface of the basal body (26 nm). Hence, YscP is properly commensurate to the length of the injectisome from the inner side of the plasma membrane to the needle tip. The molecular modeling analysis thus confirms that the functional form of YscP consists in a succession of α-helices that modulates the needle length.

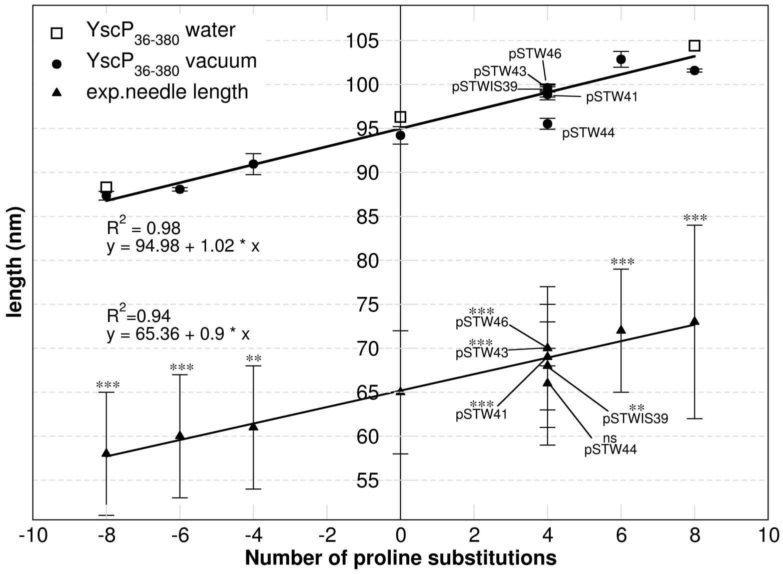


Figure 5.4: *Molecular modeling of YscP protein and comparison with experimental needle length. Computed lengths of YscP wild type and mutants are reported in circles (top), and are compared with the experimental measures of the needle (triangles at the bottom). Results from explicit solvent MD simulations are also reported (empty squares at the bottom). Correlation between needle length and number of prolines substitutions is shown in the lower slope (triangles). All the values were compared with the wild-type value by the t-test. The stars indicate the probability of difference to wild type.* $NS = P > 0.05$ (not significant); $*0.05 > P > 0.01$ (significant); $**0.01 > P > 0.001$ (very significant); $***P < 0.001$ (extremely significant).

5.2.2 Needle can accomodate folded YscP and YscF

The model of the molecular ruler, as presented earlier, proposes that the ruler is located inside the growing needle. Based on the hypothesis that YscP is fully extended conserving its helical content, the width of the wild-type ruler was evaluated using our computational models for YscP (Fig. 5.5). The average min-max width of the ruler goes from $0.79 \pm 0.10nm$ to $0.95 \pm 0.10nm$, and it never reaches values higher than $1.3nm$.

Given that the needle is known from cryo-electron microscopy experiments to have a section of $2-3nm$, the YscP protein can be easily accommodated within the needle cavity, although presenting conserved helical motifs. Moreover, on the needle channel there would be enough space to accommodate at the same time both the YscP protein and the YscF building blocks of the needle. Based on this structural analysis, non-specific interactions between YscP and YscF proteins can likely happen during the construction of the needle.

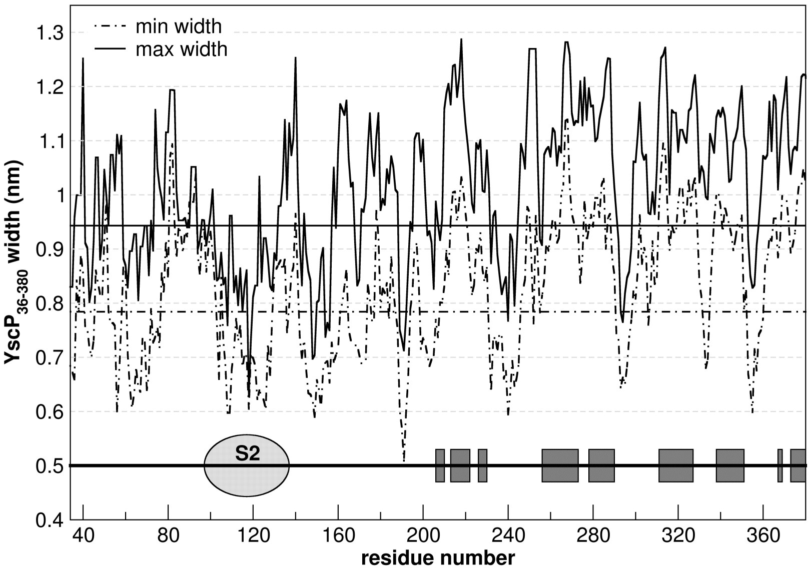


Figure 5.5: *Maximal (solid line) and minimal (dashed line) width of YscP36-380. The average maximum (*0.95 *nm) and minimum (*0.79 *nm) values are represented by constant lines. At the bottom, a cartoon represents the secondary structure: grey boxes represent predicted helices, while the oval represents the position of a globular domain called signal 2.*

5.2.3 Materials and Methods

We obtained information about YscP secondary structure for the wild type using several secondary structure predictors (see Appendix A.5) and subsequently sculpted structural models representing a partially extended protein with intact helices intercalated by unstructured coiled regions.

Wild Type model, as well as models of 9 mutants have been pulled 20 times at constant velocity, using steered Molecular Dynamics as implemented in NAMD molecular dynamics engine [60]. In each simulation, pulling force, helical content (percentage of residues being part of an helix) and strand length are evaluated. The protein length upon unfolding of its helices was assessed by measuring the distance between the two terminal C_α atoms. Unfolding was determined both by a decrease of helical content determined by the DSSP algorithm [63] and by peaks in the force required to pull the protein at constant speed.

5.3 Modeling of the MS-ring at the basal body

The injectisome's basal body is formed by several ring-shaped protein complexes. In *Yersinia*, a ring composed by 12 copies of YscC spans the outer bacterial membrane and interacts at the periplasmic side with the MS ring, formed by the YscJ and YscD proteins. Both YscD and YscJ feature one transmembrane helix spanning the internal membrane, but only YscD has a cytoplasmic domain. Cross-linking experiments on the *Salmonella* and *Shigella* orthologues revealed a close association between YscJ and YscD. However, the stoichiometry of the MS ring is still a matter of debate. While a model for *Salmonella* MS ring homologue has been recently produced [127], no atomistic structure for *Yersinia* MS ring is known. For this reason, we aim here at modeling this region.

5.3.1 The periplasmic domain of YscD is elastic

The periplasmic YscD component (152-346) is constituted by three compact $\alpha\beta$ domains connected by short flexible loops. While a recent X-ray of the the periplasmic region of YscD revealed an extended arrangement of the three domains (unpublished data from Heinz lab at the Helmholtz Centre for Infection Research), the structure of PrgH (the *Salmonella* homologous, pdb: 3GRO [132]) features a bent conformation (see Figure 5.6). Taken together, the two structures suggest the existence of several degenerate conformational states ranging from fully extended to partially bent arrangement. To address this point, MD simulations were run on both fully elongated (from YscD X-ray) and bent (constructed using PrgH X-ray as a template) YscD monomer embedded in a box of water molecules. The simulations indicated that both bent and extended states are stable in the simulation time (100ns), and that the flexible loops between domains act as hinges. Principal Component Analysis on both trajectories highlighted a concerted domain movement consistent with a state switchover.

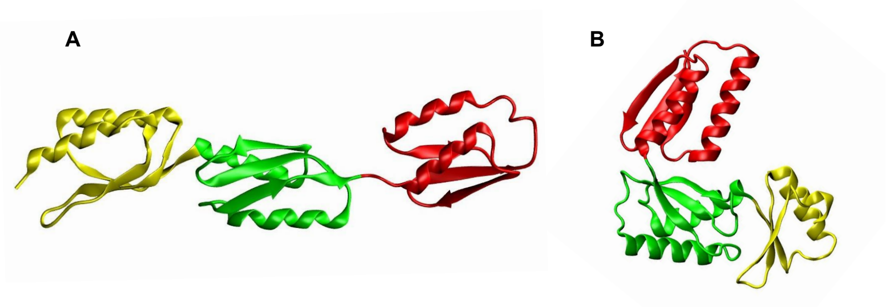


Figure 5.6: *Extended (A) and bent (B) conformation of YscD monomer*

Importantly, this result indicates that the periplasmic domain of YscD is naturally elastic. Tertiary structure elasticity, defined as the rearrangement of tertiary structure in response to mechanical force, represents the first mode of elastic response to external stimuli. To address the tertiary structure elasticity of the periplasmic domain of YscD, we applied a steered molecular dynamics (SMD) protocol to the bent conformation of YscD previously produced. Starting from an equilibrated structure, The C_α of D152 was kept frozen while the C_α of N346 was slowly pulled. During the SMD simulation (160 ns) the system undergoes a complete stretching from the bent PrgH-like conformation to the fully extended conformation observed in the YscD X-ray structure. The most prominent degrees of freedom during the extension process, are the two bending degrees of freedom between neighboring domains pairs. The measured forces are consistent with the capability of YscD of easily getting longer or shorted in response of external stimuli.

5.3.2 Assembly of the MS-ring

A model for a 24-mer YscJ assembly (periplasmic domains) was simply derived by homology modeling based on the X-ray structure of the *E. coli* EscJ homologous (PDB: 1YJ7 [148]) using Modeller [39]. The YscJ 24-mer features a compact protomeric arrangement of circular shape. The surface in contact with the outer leaflet of the inner membrane exposes charged and polar residues suitable for interaction with the lipid phosphate head-groups (see Figure 5.7, in blue).

Unlike for the case of the YscJ ring, no suitable structural template exists for the YscD ring. In order to create an atomistic model of the whole MS ring, POW was therefore exploited. The program was requested to assemble 24 YscD units according to a circular symmetry, using the YscJ model as substrate. In order to take into account YscD structural flexibility, an ensemble of structures extracted from the previously performed SMD was provided.

POW produced an ensemble of MS ring assemblies respecting the given restraints. All models were then scored on the basis of the matching between the external size (height, width) of the YscD 24mer and a new available cryo-tomograph of the full injectisome *in situ* of the *Yersinia* basal body (unpublished data from Stalhberg lab in Basel). The best model is shown in Figure 5.7. Interestingly, just the YscD extended conformation, which is consistent with the novel X-ray structure, led to models having a good matching with the available cryo-electron tomograph. This result is particularly important since for the first time the flexibility of the basal body and the MS-ring specifically is shown to be crucial for the anchoring and adaption of the injectisome complex at both the outer and inner membrane bilayers in bacterial cells. This result also shows how POW can be effectively applied to the case of a symmetrical, flexible assembly on a target substrate.

5.3.3 Material and Methods

YscD modeling

The initial bent conformation was constructed on the basis of PrgH (PDB: 3GRO; UniProtKB: P41783) by performing independent structural alignment between each domain (loops

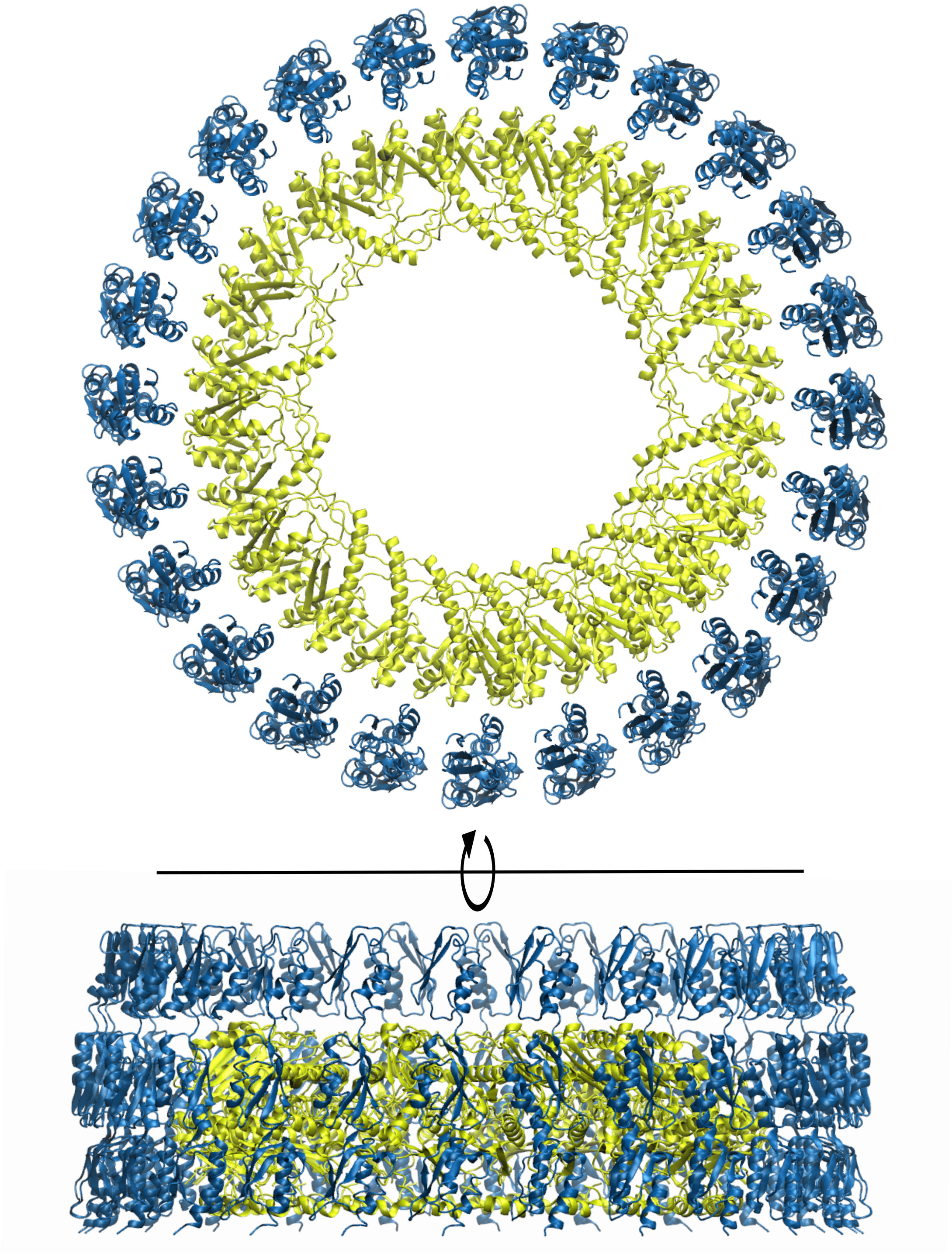


Figure 5.7: *Top and side view of injectisome's MS ring, supposing that both YscJ (inside, in yellow) and YscD (outside, in blue) assemble into a 24-mer.*

excluded) of YscD and PrgH; then Modeller [39] was used to splice together the three YscD fragments and to add and relax the loops.

YscD molecular dynamics and steering

Extended and bent YscD were both embedded in a box of water molecules and subjected to geometry optimization and molecular dynamics. After 10 ns equilibration, the systems were subjected to 70 ns equilibrium simulation in the NPT ensemble. Starting from the final geometry of the MD simulation, the bent YscD system was extended according to a standard SMD procedure. The C_α of D152 was kept frozen while the C_α of N346 was slowly pulled at the constant velocity of 1 Å ns^{-1} to reduce the effects of hydrodynamic drag force [53]. A constant stretching force of 5 kcal/ mol, resulting in a thermal noise deviation of 0.35 Å, was employed to pull the C_α of N346 along a fixed direction. The force opposing the stretching oscillates around an average value of 1.9 kcal/mol/Å(standard deviation 1 kcal/mol/Å).

POW setup

POW was run using an ensemble of YscD structures extracted from the previously described SMD simulation. Our YscJ model was set to be the docking substrate. As constraints, height (100 Å) and width (260 Å) of the available tomography map were used. 80 particles, 200 iteractions and 3 repetitions were run. Solutions having a fitness smaller than 0 were selected, and clustered according to their RMSD (threshold equal to 1) 20 models were finally produced, and visually assessed.

A Appendix

A.1 POW User Guide

Requirements

POW requires the following python (>=2.5) packages to be installed:

- numpy
- scipy
- MDAnalysis
- mpi4py

The execution of parallel calculation will also require the installation of OpenMPI

Provided Files

The compressed folder *POW.tar.gz* containing all the needed files is downloadable at lbm.epfl.ch/resources. This file unpacks in a folder called POW, which can be placed anywhere in your computer. The folder contains the following files:

- Assembly.py : data structure for heterodimers assembly
- Default.py : classes common to any POW implementation
- DockDimer: dock two proteins
- DockSymmCircle : rigid/flexible assembly of n monomers according to a circular symmetry, possibly around a given receptor
- Function: generic function optimization

- flexibility.py: functions for Principal Components Analysis
- parse.py: performs just the postprocessing, without running PSO. Usage goes as follow: `./parse.py module input_file [logfile]`
- POW.py: main executable
- Protein.py: PDB parser
- PSO.py: parallel implementation of Particle Swarm Optimization

Launching

POW is launched in the console by means of the following command:

```
mpiexec -n 4 $INSTALLATION_PATH/POW module input.txt
```

It is advised to create an alias, in order to make POW execution easier. The following lines create a default call using 4 processors:

```
export NPROC=4
alias pow="mpiexec -n $NPROC $POW_DIR/POW.py
```

An execution becomes now as simple as:

```
pow module input.dat
```

This call will launch POW on 4 processors. A proper execution requires the user to provide two arguments to the call: the desired optimization module `module` and a parameterization file `input.dat`. The parameterization file describes with a series of keywords how POW should behave. The input file providing all the parameterisations for the search should be passed as parameter. The file is structured as a succession of keywords having one or more correspondent values. Keywords are case sensitive, and their order is irrelevant. Some keywords are necessary for any kind of optimization procedure, whereas other are module specific.

Default keywords

The following keywords are typical to any optimization problem, and are therefore accessible by any module:

- steps $< number\ of\ steps\ to\ perform >$
  Acceptable values: positive integer

Default value: 100
Description: The number of steps that will be computed in the *PSO*.

- particles < $number\ of\ particles$ >
  Acceptable values: positive integer
  Default value: 40
  Description: The number of particles that will be used in each step of the *PSO*.

- repeat < $number\ of\ repetition$ >
  Acceptable values: positive integer
  Default value: 1
  Description: Repeat can be used to repeat the *PSO* multiple times. The idea behind repetition is that particles tend to focus on one solution corresponding to a good fitness. Repetition restart the system and consequently increase the chance to find other solutions.

- repulsion < $activate\,|\,desactivate$ >
  Acceptable values: on | off
  Default value: off
  Description: Repulsion can be seen as a flag on position of good solution where particles should be repulsed. With this option activated particles that found good solutions are forced to discover other solution. This is particularly interesting when there are more than one repetition, because flags are kept between repetition, consequently two distinct repetitions will tend to discover different solution. This option is currently being tested and will be released in future version.

- neighborType < $type\ of\ neighbor$ >
  Acceptable values: indexed | geographic
  Default value: geographic
  Description: NeighborType set the kind of neighborship between particles. Indexed is good because a neighbor of a particle will stay the same during the whole execution, however it is not natural. The typical example used to explain *PSO* is about a group of birds searching for food in a big field. Birds communicate between them where they found a lot of food. It is more natural for a bird to communicate to neighboring birds that are geographically close than with birds that are close in terms of index. Consequently, geographic is the *neighborType* set by default. However, geographic suffer from the fact that neighbor are not always the same and change frequently. This involves more computation than for the indexed type.

- neighborSize < $number\ of\ neighbor$ >
  Acceptable values: positive integer
  Default value: 1
  Description: NeighborSize is considered only for the geographic neighbor type. It specifies the number of particles to consider as neighbor for a particle. In future release, this option will also apply to indexed neighbor type.

- boundaryMin < *min boundary for each dimension* >
  Acceptable values: list of lower boundary for each dimension, separated by spaces.
  Default value: module dependent
  Description: It is the minimum boundary for each dimension of the space. The first three values correspond to the rotations of the monomer on x, y and z axis respectively. The last one is the value specified by the *radius* keyword. In case you did not use the *radius* keyword, you MUST specify a minimum radius here.

- boundaryMax < *max boundary for each dimension* >
  Acceptable values: list of upper boundary for each dimension, separated by spaces.
  Default value: module dependent
  Description: It is the maximum boundary for each dimension of the space. The first three values correspond to the rotations of the monomer on x, y and z axis respectively. The last one is the value specified by the *radius* keyword. In case you did not use the *radius* keyword, you MUST specify a maximum radius here.

- boundaryType < *type of the boundary* >
  Acceptable values: 0 | 1
  Default value: module dependent
  Description: For each dimension it is possible to define the boundary condition. *0* and *1* stands for periodic and repulsive boundary conditions respectively.

- inertiaMax < *max inertia of particles* >
  Acceptable values: float 0-1
  Default value: 0.9
  Description: It is the maximum inertia of particles. Between steps of the *PSO* the inertia is decreased until *inertiaMin.*

- inertiaMin < *min inertia of particles* >
  Acceptable values: float 0-1
  Default value: 0.4
  Description: It is the minimum inertia of particles. Between steps of the *PSO* the inertia is decreased until *inertiaMin.*

- cp < *influence of local best solution* >
  Acceptable values: float
  Default value: 1.2
  Description: It is the influence on a particle of the best solution found by that particle.

- cp < *influence of global best solution* >
  Acceptable values: float
  Default value: 1.4
  Description: It is the influence on a particle of the best position found by neighbors of that particle.

- kar_threshold < *threshold for kar execution* >
  Acceptable values: float > 0
  Default value: 0.01
  Description: When a particle is being slower than this threshold, the kick and reseed procedure (KaR) will be triggered. The particle will receive a random kick that will reaccelerate it. If, moreover, the particle's current fitness is smaller than filter_threshold, it will be also reseeded in a random location. This avoids early convergence and forces the swarm to explore further the search space. Notice that setting kar_threshold to 0 disables KaR.

- filter_threshold < *fitness value to accept* >
  Acceptable values: float
  Default value: 0
  Description: An ensemble of solution is found, but just some of these will be good. This variable sets a threshold on the solutions fitness function.

- output < *text file* >
  Acceptable values: UNIX filename
  Description: The text file will be used to store results.

Function Module

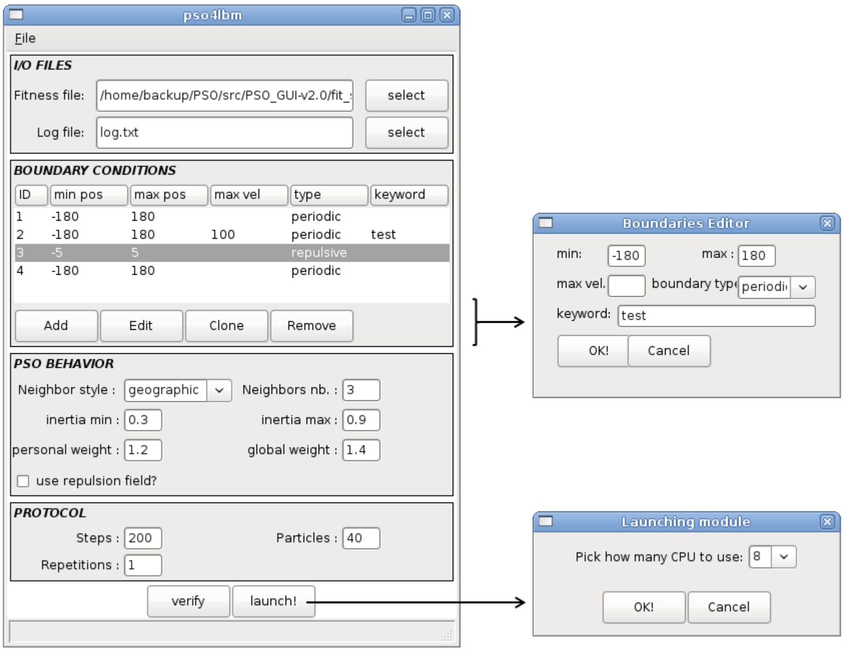
The `Function` module allows the minimization of any function not requiring manipulation of any data structure. The file containing the fitness function to be evaluated is passed to POW via the following keyword:

- fitness < *fitness extraction file* >
  Acceptable values: UNIX filename
  Default value: fit_multimer
  Description: This file contains the implementation for the Fitness class, and should have the following form:

```
class Fitness:
        def __init__(self,data,params):
    pass
        def evaluate(self, num, pos):
           #num: PSO particle index
           #pos: array containing the particle's position in search space
            #compute fitness on the base of pos values
            return fitness
```

The `Function` module can also be operated via a graphical interface invoked with the command `pso_gui.py` (see Figure A.1). The interface allows the user to create, edit and save a

POW input file, validate it, and launch a POW run on multiple processors.



Fitness file: /home/backup/PSO/src/PSO_GUI-v2.0/fit_

Log file: log.txt

ID	min pos	max pos	max vel	type	keyword
1	-180	180		periodic	
2	-180	180	100	periodic	test
3	-5	5		repulsive	
4	-180	180		periodic	

Neighbor style : geographic

Neighbors nb. : 3

inertia min : 0.3

inertia max : 0.9

personal weight : 1.2

global weight : 1.4

☐ use repulsion field?

Steps : 200

Particles : 40

Repetitions : 1

min: -180

max : 180

max vel.

boundary type periodic

keyword: test

Pick how many CPU to use: 8

Figure A.1: *POW graphical interface for* `Function` *module* allowing the user to save, edit, validate and launch POW input files.

A.2 DockSymmCircle module

Additionally to default POW keywords A.1, the following keywords are defined:

- radius $< fixed\ radius\ of\ the\ multimer >$
  Acceptable values: float
  Description: If you know precisely the radius you can use the *radius* keyword. If the precise radius is not known, the user should limit as much as possible the range of values for the radius in *boundaryMin* and *boundaryMax*. A drawback when specifying a range of value is that the search space increase proportionally to the length of the range. The problem complexity increase and the chance to get a good solution decrease.

- degree $< number\ of\ monomer >$
  Acceptable values: positive integer
  Description: It is the number of monomer that compose the multimer.

- target $< list\ of\ measures >$
  Acceptable values: list of float separated by spaces
  Description: The list of *target* measure will be used by the system to compute the fitness.

This list MUST have the same the same schema as the list computed from the *constraint* file.

- constraint < *constraint file* >
  Acceptable values: UNIX filename
  Description: The system generates a multimer corresponding to the particle position in the space and passes it to the constraint file. See the section A.2 for details about the structure of this file. The list of measure you return will be compared to the list of *target's* measure and output a fitness value that will be written to the output file. The list of *target* measure MUST have the same order as the list of measure computed in the constraint file.

- style < *type of assembly* >
  Acceptable values: flexible | rigid
  Default value: rigid
  Description: Define the type of assembly to perform. If rigid is chosen, the monomer keyword must be defined as well. If flexible is chosen, at least topology and trajectory keywords must be defined.

- monomer < *monomer PDB file* >
  Acceptable values: UNIX filename
  Description: PDB file containing the monomer. Requires style keyword set to rigid.

- trajectory < *coordinates of a MD trajectory* >
  Acceptable values: path to a dcd or crd file
  Description: Enesemble of protein structures. Requires style keyword set to flexible.

- topology < *topology of a MD trajectory* >
  Acceptable values: path to a charmm or amber topology
  Description: Topology of provided trajectory (see trajectory keyword). Requires style keyword set to flexible.

- trajSelection < *atom selection in MDAnalysis format* >
  Acceptable values: MDAnalysis AtomSelect
  Default value: protein
  Description: Select a subset of atoms from provided trajectory. If align keyword is set to yes, trajectory will also be aligned on this selection. PCA and subsequent assembly will only take these atoms into account. Requires style keyword set to flexible.

- projection < *projection of MD trajectory on main eigenvectors* >
  Acceptable values: path to a projections file
  Description: If provided, Principal Components Analysis will not be performed, and this file providing projections on main eigenvectors will be used instead. This file should consist of a number of lines matching the number of atoms in the provided trajectory, and a number of columns corresponding to the desired number of eigenvectors used for projection. Requires style keyword set to flexible.

- align < $define\ whether\ to\ align\ the\ given\ trajectory$ >
  Acceptable values: yes | no
  Default value: yes
  Description: If set to yes, the provided trajectory will be aligned on the protein. Taken into account only if style keyword is set to flexible.

- ratio < $energy\ represented\ by\ eigenvectors$ >
  Acceptable values: float 0-1
  Default value: 0.8
  Description: After having performed PCA, symmetryMaker selects a number of representative eigenvector. These will represent at least a certain percentage of the trajectory's energy. Taken into account only if style keyword is set to flexible.

- detectClash < $clash\ detection\ switch$ >
  Acceptable values: on, off
  Default value: on
  Description: define whether a 9-6 Lennard-Jones function should be computed to assess the system's energy.

- mixingWeight < $weight\ energetic\ vs\ geometric\ contributions$ >
  Acceptable values: float 0-1
  Default value: 0.2
  Description: fitness function is computed via the equation $f = c * energy + (1 - c * distance)$, where c is the value of mixingWeight.

- receptor < $clustering\ distance\ within\ solutions$ >
  Acceptable values: UNIX filename
  Description: PDB file containing a receptor around which the assembly will be built.

- z_padding < $assembly\ vertical\ displacement$ >
  Acceptable values: float > 0
  Default value: 5
  Description: the whole assembly is displaced along the z axis with respect of the receptor. Boundary conditions are defined by a lower and higher boundary. These are computed around the size of the receptor. `z_padding` adds an additional displacement to the computed boundaries. Should be defined only if `boundaryMinReceptor` and `boundaryMinReceptor` are undefined, and if `receptor` is given.

- boundaryMinReceptor < $min\ boundary\ for\ receptor\ dimensions$ >
  Acceptable values: list of lower boundary for each dimension, separated by spaces.
  Default value: min_receptor 0 0 -360/(2*degree)
  Description: It is the minimum boundary for each dimension of the space. The first three values correspond to the rotations of the monomer on x, y and z axis respectively. The last one is the value specified by the *radius* keyword. In case you did not use the *radius* keyword, you MUST specify a minimum radius here.

- boundaryMaxReceptor < $max\ boundary\ for\ receptor\ dimensions$ >
  Acceptable values: list of upper boundary for each dimension, separated by spaces.
  Default value: max_receptor+z_pad 360 360 360/(2*degree)
  Description: It is the maximum boundary for each dimension of the space. The first three values correspond to the rotations of the monomer on x, y and z axis respectively. The last one is the value specified by the *radius* keyword. In case you did not use the *radius* keyword, you MUST specify a maximum radius here.

- cluster_threshold < $clustering\ distance\ within\ solutions$ >
  Acceptable values:float > 0
  Default value: 5
  Description: Similar solutions will be clustered in a unique solution. If RMSD clustering is chosen, a value smaller or equal to 5 Åis advised. If distance clustering is used, a number around 15 is suggested.

Note that the Default keywords `boundaryMin` and `boundaryMax` (see Default keywords section A.1) should include the following quantities in the following order:
α, β, γ, $radius$

Constraint File

The constraint file is user provided, and contains a python function containing user defined measure on the generated multimer. This script consists of one function that must be declared like this:

```
def constraint_check(multimer):
    #######################
    #user defined measures#
    #######################
    return measure1 measure2
```

The user can define various measures inside this function, and return them. The return order is significant, it should indeed match the order of target measures provided with the target keyword in input file. The *multimer* parameter is a *Multimer* object. This object provides the following functions for measurment of the structure:

- multimer.get_width(), returns the assembly width
- multimer.get_height(), returns the assembly heigth
- multimer.atom_select(unit,chain,resid,name), returns the a numpy 2D array containing all the coordinates of atoms matching the selection.
- multimer.get_center_of_geometry(), , returns the center geometric center of the protein

Examples

The minimal set of keywords for a POW input script are as follows:

```
monomer input.pdb
constraint constraint.py
degree 5
radius 10
target 10 20
```

This will rigidly assemble 5 monomers from file input.pdb so that the circular radius is 10. constraint.py file will be used as constraint. This file will compute two measures, that should be compared with the target measures 10 and 20.

A complete example showing how to perform a rigid assembly is as follows:

```
steps 150
particles 50
repeat 3

boundaryMin 0 0 0 8
boundaryMax 360 180 360 12

assembly_style rigid
monomer protein.pdb

constraint constraint.py
degree 7
target 85 150

cluster_style rmsd
filter_threshold 0
cluster_threshold 5
```

In this example a calculation protocol with 150 iterations, 50 particles and 3 repetitions has been chosen. boundaryMin and boundaryMax keyword define a multimer with a radius varying from 8 to 12 Å. The provided monomer (protein.pdb) will be treated as a rigid body, and assembled in a heptameric structure (7-fold simmetry) being constrained by constrain.py function. In postprocessing, only solutions having a fitness smaller than 0 will be retained, and solutions having an RMSD smaller than 5 within themselves will be clustered.

By replacing the `monomer` keyword of previous example with what follows, it's possible to perform a flexible assembly.

```
assembly_style flexible
topology proten.prmtop
trajectory trajectory.dcd
align yes
ratio 0.80
```

Flexible assembly requires a trajectory (in crd or dcd format) and a topology (pdb or psf). If the protein in the trajectory is not aligned, symmetryMaker can do this for you by means of the align keyword. This done PCA is performed on C_α atoms. Notice that the number of degrees of freedom (3*N, where N is the number of carbons) must be greater than the number of frames in the simulation. A number of eigenvector representing more than 0.8 (80%) of the system's energy will be extracted and treated as protein's degrees of freedom.

Aligning the trajectory and performing a PCA can take quite a lot of time. However, preprocessing phase, will generate an aligned trajectory (`aligned.dcd`) and a file containing eigenvectors projection (proj_coordinates.dat). You indicate symmetry maker to use these file to avoid repeting the preprocessing. This can be done in this way:

```
assembly_style flexible
topology proten.prmtop
trajectory aligned.dcd
align no
projection proj_coordinates.dat
```

Creation of a new module

A module contains an implementation for `Parser`, `Data`, `Space`, `Fitness` and `Postprocess` classes. The following lines represent a module skeleton.

```
from Default import Parser as R
from Default import Space as S
from Default import Postprocess as PP

#import other packages here

class Parser(P):
   def __init__(self,infile):
    #parse more params if needed

class Data:
    def __init__(self,params):
    #if needed, load files using parsed information contained in params object

class Space(S):
    def __init__(self,params,data):
        #build search space using params and loaded data objects by defining:
        #self.low = low boundaries
        #self.high = high boundaries
        #self.boundary_type = int array (0 = periodic, 1 = reflexive boundaries)

class Fitness:
    def __init__(self,data,params):
        #load data here if needed (e.g. target measures,...)

    def evaluate(self,num,pos):
        #return fitness value

class Postprocess(PP):
    def __init__(self,params,data):
        #load params and data structure
    def run(self):
        #parse logfile and postprocess
```

A.3 POW prediction using various constraints

In this section, additional benchmarks of the POW DockSymmCircle module are reported. Results are obtained by adopting different geometric restraints, but using the same experimental setup stated in main text (Chapter 3). Good results (bRMSD smaller than 2 Å from known crystal) can be obtained, for every protein, with a variety of restraints. Small differences (order of 0.2 Å) within different trials are mainly due to the clustering procedure. Not surprisingly, the amount of obtained valid models usually increases the less stringent the used geometric restraints are, and viceversa. In general, the more precise (and reasonable) restraints are provided, the better the result (i.e. a smaller amount of possible models is produced).

SM Archeal Protein

Heptameric *Pyrobaculum aerophilum* SM Archeal Protein was crystallized at a resolution of 1.75 Å (pdb: 1I8F) [101]. We extracted chain A as a representative monomeric structure. Residues 15 to 80 of all chains (i.e. the residues common to all the seven subunits in the crystal) were selected to compute the backbone RMSD within the PSO-generated solutions and the known crystal. Table A.1 reports results using a variety of different geometric restraints. We tested the effect of making the same geometric condition less stringent, using a global measures (width and height), imposing a residue to face outside the complex (Pro80) or two residues in the pore center (Arg29 and Asp30) to be in contact.

restraint	time	sol.	bRMSD
w=66 ± 2 Å, h=40 ± 2 Å	2m16s	2	1.52 Å
d(Arg29,(0,0,0))<10 Å, w=66 ± 2 Å, h=40 ± 2 Å	2m53s	3	1.63 Å
d(Arg29,(0,0,0))<10 Å	2m53s	15	1.28 Å
d(Pro80,(0,0,0))>60 Å	2m53s	45	1.61 Å
d(Arg29,Asp30)<4 Å	2m56s	27	1.52 Å

Table A.1: Summary of POW predition results on SM Archeal Protein heptamer using several geometric restraints.

Chorismate Mutase

Trimeric *Clostridium thermocellum* Chorismate Mutase was crystallized at a resolution of 2.2 Å (pdb: 1XHO) [143]. Chain A was extracted as representative structure. Backbone RMSD was computed on all residues common to all the three chains, i.e. residues 2 to 113. Table A.2 reports results using a variety of different geometric restraints. We tested the effect of making the same global geometric condition more stringent, and also applied geometric restraints on specific atoms. On this aspect, we constrained the N-Terminal residues (Val2) as well as residue Met74 (in protein's core) to be close. We notice that, as soon as restraints become more stringent (for instance by restraining the distance within two atoms), the amount of produced multimers drops.

restraint	time	sol.	bRMSD
w=49 ± 2 Å, h=42 ± 2 Å	2m20s	20	1.89 Å
w=49 ± 2 Å, h=42 ± 2 Å, d(Met74,Met74)=3.5 ± 1 Å	2m20s	3	1.72 Å
w=49 ± 4 Å, h=42 ± 4 Å, d(Met74,Met74)=3.5 ± 1 Å	2m26s	6	1.94 Å
d(Met74,Met74)=3.5 ± 1 Å, d(Val2,Val2)=8 ± 1 Å	1m50s	2	1.59 Å
w=49 ± 3, h=42 ± 3, d(Val2,Val2)=8 ± 1 Å	1m55s	7	2.49 Å

Table A.2: Summary of POW predition results on Chorismate Mutase trimer using several geometric restraints.

Acyl Carrier

Trimeric *Streptococcus pneumoniae* AcpS was crystallized with 3’5’-ADP docked in two of its three active sites at a resolution of 1.9 Å (pdb: 1FTH) [27]. Chain A was extracted as representative structure. Backbone RMSD was computed taking into account all residues common to all three chains, i.e. residues 3 to 68, 75 to 99 and 101 to 118. Table A.3 reports results using a variety of different geometric restraints. The influence of using global measures (width and height), other residues part of the binding site (i.e. Lys64 and Asp55) and various mixes of these quantities were tested.

restraint	time	sol.	bRMSD
w=60 ± 2 Å, h=47 ± 2 Å	2m12s	9	1.80 Å
w=60 ± 2 Å, h=47 ± 2 Å, d(Asp10,His105)=9 ± 2 Å	1m59s	3	1.82 Å
w=60 ± 4 Å, h=47 ± 4 Å, d(Asp10,His105)=9 ± 4 Å	1m59s	8	1.98 Å
d(Lys64,Asp55)=10 ± 4 Å	2m14s	8	1.86 Å
d(Lys64,Asp55)=10 ± 4 Å, d(Asp55,His105)=9 ± 4 Å	2m20s	6	1.95 Å

Table A.3: Summary of POW predition results on Acyl Carrier trimer using several geometric restraints.

EscJ

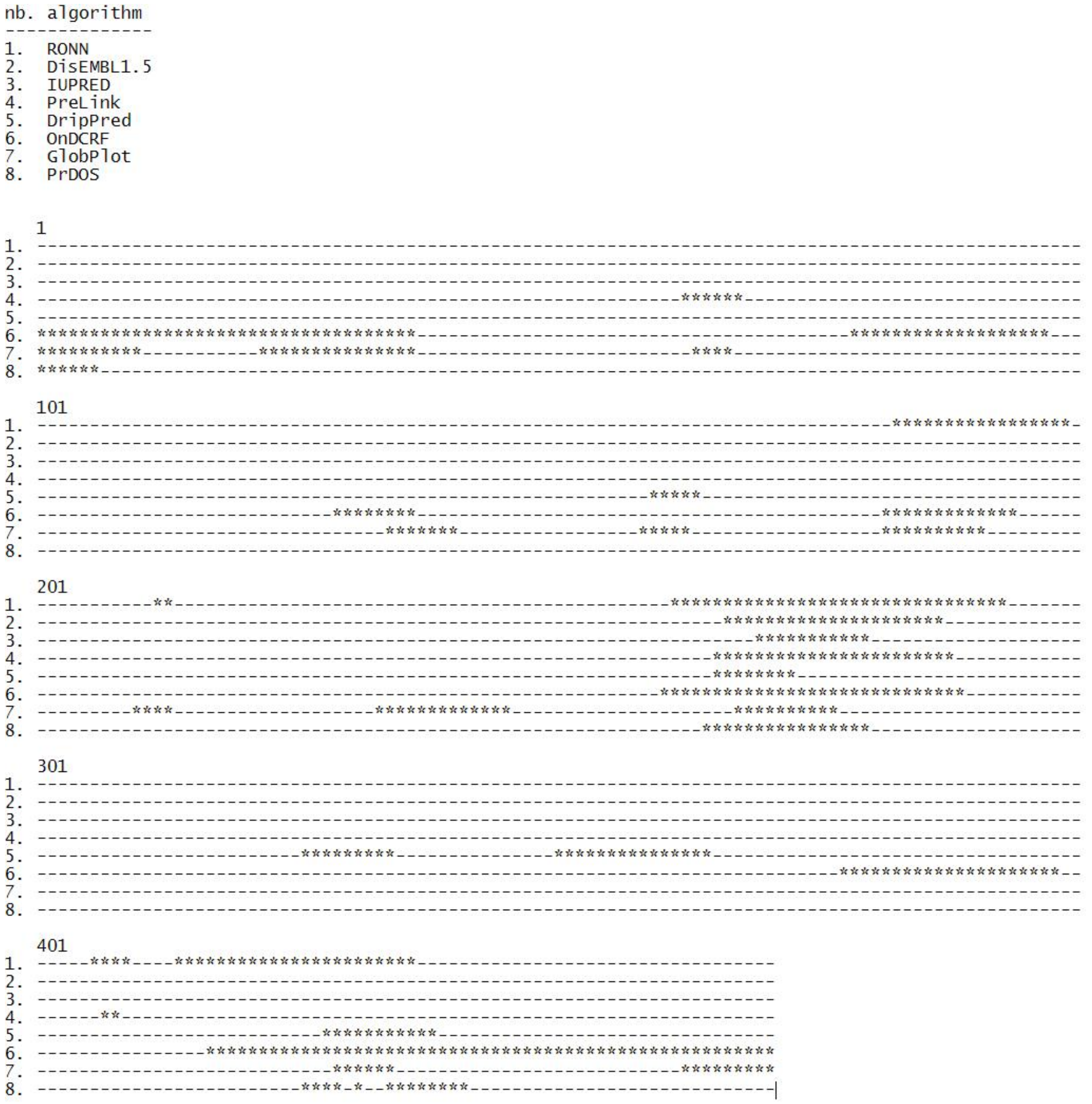
restraint	time	sol.	bRMSD
w=176 ± 2 Å, h=180 ± 2 Å	6m52s	7	2.58 Å
w=176 ± 4 Å, h=180 ± 4 Å, d(Pro99,(0,0,0))<40 Å	12m53s	43	1.79 Å
w=176 ± 4 Å, h=180 ± 4 Å, top Lys178	13m21s	41	2.24 Å
w=176 ± 4 Å, h=180 ± 4 Å, top Lys178, d(Pro99,(0,0,0))<40 Å	12m50s	42	0.43 Å
Lys178 on top, d(Pro99,(0,0,0))<40	8m40s	10	2.06 Å

Table A.4: Summary of POW predition results on EscJ using several geometric restraints.

24mer EscJ, part of type III secretion system’s basal body, had a 4-mer basic unit crystallized (pdb: 1YJ7 [148]). Chain A was extracted as representative structure. Backbone RMSD was computed with respect of the tetrameric basic unit taking into account all residues common to all three chains, i.e. residues 21 to 91, 98 to 133 and 141 to 186. Table A.4 reports results using a variety of different geometric restraints. On this test case, we tested the effect of using a unique, stringent set of global measures (height and width). We also coupled such measures with residues specific restraints. In particular, we tested the effect of imposing residue Lys178 to face upwards, and to combine this restraint with others.

A.4 Aerolysin disorder prediction

Figure A.2 summarizes the results obtained by eight different disorder prediction algorithms on aerolysin aminoacid sequence (UniProt entry P09167 (AERA_AERHY)). In detail, we used RONN [147], DisEMBL1.5 [80], IUPRED [36], PreLink [30], DripPred [1], OnDCRF [73], GlobPlot [81] and PrDOS [59]. Figure 4.10 has been obtained by coloring every residue according to the number of algorithms predicting it as disordered. Thus, red regions are predicted as disordered by all seven algorithms, whereas blue ones are considered as ordered by all of them. Region GLN268 to ARG282 is the only region predicted as disordred by at least 6 algorithms.

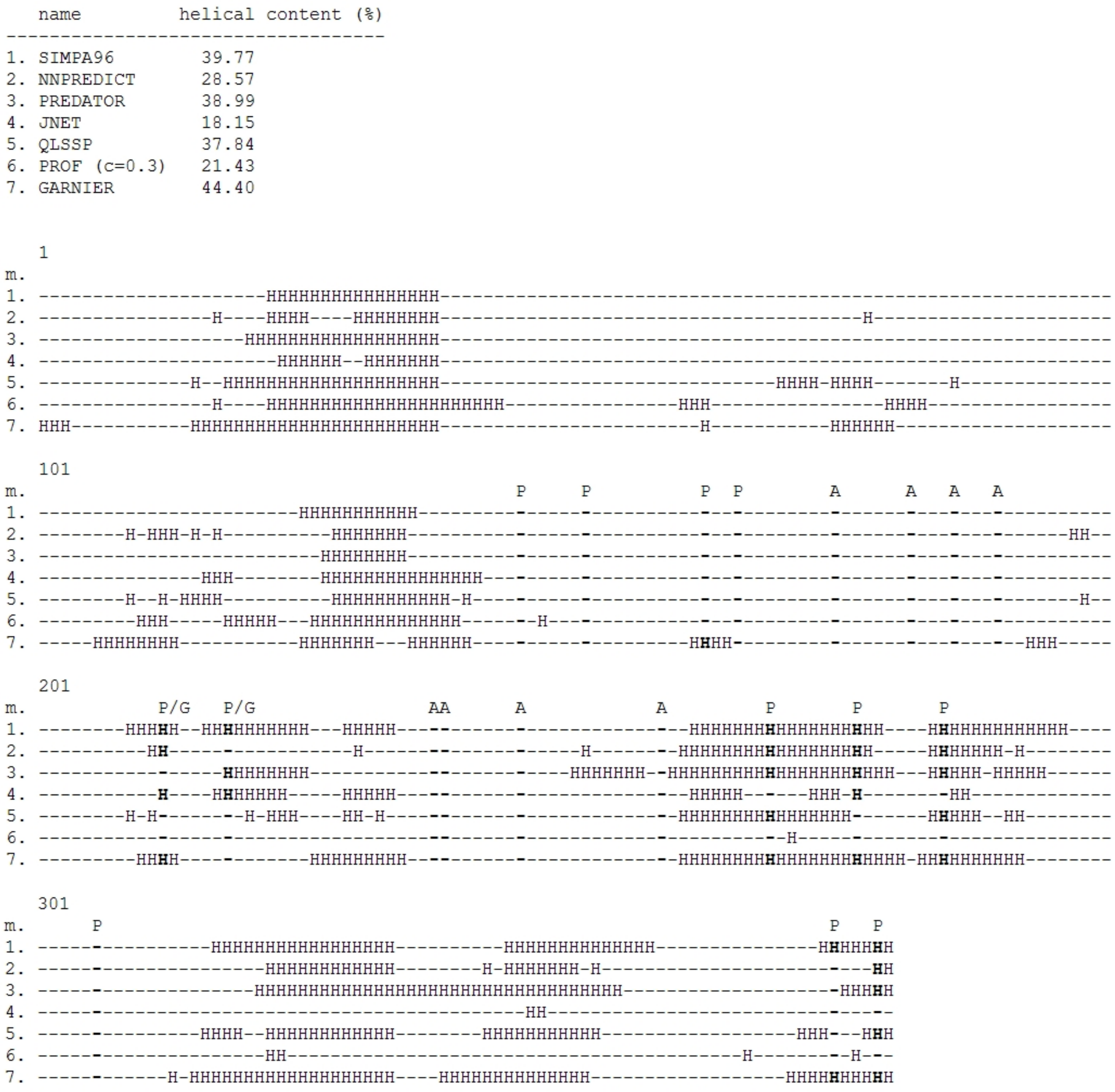


```
nb. algorithm
--------------
1.  RONN
2.  DisEMBL1.5
3.  IUPRED
4.  PreLink
5.  DripPred
6.  OnDCRF
7.  GlobPlot
8.  PrDOS


   1
1. ----------------------------------------------------------------------------------------------------
2. ----------------------------------------------------------------------------------------------------
3. ----------------------------------------------------------------------------------------------------
4. --------------------------------------------------------------******--------------------------------
5. ----------------------------------------------------------------------------------------------------
6. ************************************------------------------------------------*******************---
7. **********-----------***************---------------------------****---------------------------------
8. ******----------------------------------------------------------------------------------------------

   101
1. ----------------------------------------------------------------------------------*****************-
2. ----------------------------------------------------------------------------------------------------
3. ----------------------------------------------------------------------------------------------------
4. ----------------------------------------------------------------------------------------------------
5. -----------------------------------------------------------*****------------------------------------
6. ----------------------------********---------------------------------------------*************------
7. ---------------------------------*******------------------*****------------------**********---------
8. ----------------------------------------------------------------------------------------------------

   201
1. -----------**------------------------------------------------********************************-------
2. ------------------------------------------------------------------*********************-------------
3. ---------------------------------------------------------------------***********--------------------
4. -----------------------------------------------------------------***********************------------
5. -----------------------------------------------------------------********---------------------------
6. ------------------------------------------------------------*****************************-----------
7. ---------****-------------------*************----------------------**********-----------------------
8. ----------------------------------------------------------------****************--------------------

   301
1. ----------------------------------------------------------------------------------------------------
2. ----------------------------------------------------------------------------------------------------
3. ----------------------------------------------------------------------------------------------------
4. ----------------------------------------------------------------------------------------------------
5. -------------------------*********----------------***************-----------------------------------
6. -----------------------------------------------------------------------------*********************--
7. ----------------------------------------------------------------------------------------------------
8. ----------------------------------------------------------------------------------------------------

   401
1. -----****----***********************-----------------------------------
2. -----------------------------------------------------------------------
3. -----------------------------------------------------------------------
4. ------**---------------------------------------------------------------
5. ---------------------------***********---------------------------------
6. ----------------*******************************************************
7. ----------------------------******----------------------------*********
8. -------------------------****-*--********------------------------------|
```

Figure A.2: *Aerolyin disorder prediction from eight different predictors. Disordered residues are indicated with "*", ordered ones with "-"*

A.5 YscP secondary structure prediction

Figure A.3 summarizes the results obtained by seven different secondary structure prediction algorithms on YscP aminoacid sequence (UniProt entry Q93KT6 (Q93KT6_YEREN)). In detail, we used SIMPA96 [78], NNPREDICT [69], PREDATOR [43], JNET [31], QLSSP [100], PROF [108] and GARNIER [44].



```
   name         helical content (%)
-----------------------------------
1. SIMPA96        39.77
2. NNPREDICT      28.57
3. PREDATOR       38.99
4. JNET           18.15
5. QLSSP          37.84
6. PROF (c=0.3)   21.43
7. GARNIER        44.40


   1
m.
1. ---------------------HHHHHHHHHHHHHHHH---------------------------------------------------------------
2. ----------------H----HHHH----HHHHHHHH----------------------------------------H----------------------
3. -------------------HHHHHHHHHHHHHHHHHH---------------------------------------------------------------
4. ----------------------HHHHHH--HHHHHHH---------------------------------------------------------------
5. --------------H--HHHHHHHHHHHHHHHHHHHH--------------------------------HHHH-HHHH-------H--------------
6. ----------------H----HHHHHHHHHHHHHHHHHHHHHH-----------------HHH----------------HHHH-----------------
7. HHH-----------HHHHHHHHHHHHHHHHHHHHHHH-------------------------H-----------HHHHHH--------------------

   101
m.                                              P     P          P  P        A      A   A   A
1. ------------------------HHHHHHHHHHH-----------------------------------------------------------------
2. --------H-HHH-H-H----------HHHHHHH--------------------------------------------------------------HH--
3. --------------------------HHHHHHHH------------------------------------------------------------------
4. ---------------HHH--------HHHHHHHHHHHHHHH-----------------------------------------------------------
5. --------H--H-HHHH----------HHHHHHHHHHH-H---------------------------------------------------------H--
6. ---------HHH-----HHHHH---HHHHHHHHHHHHHH--------H----------------------------------------------------
7. -----HHHHHHHH-----------HHHHHHH---HHHHHH---------------------HHHH---------------------------HHH-----

   201
m.            P/G   P/G                AA       A            A         P       P       P
1. --------HHHHH--HHHHHHHHHH---HHHHH----------------------------HHHHHHHHHHHHHHHHHH----HHHHHHHHHHHHH----
2. ----------HH-----------------H---------------------H--------HHHHHHHHHHHHHHHHHH-----HHHHHHH-H--------
3. -----------------HHHHHHHH-------------------------HHHHHHH--HHHHHHHHHHHHHHHHHHHHH---HHHHH-HHHHH------
4. -----------H----HHHHHHH-----HHHHH----------------------------HHHHH------HHH-H--------HH-------------
5. --------H-H--------H-HHH----HH-H----------------------------HHHHHHHHHHHHHHHH-------HHHHH--HH--------
6. ----------------------------------------------------------------------H-----------------------------
7. ---------HHHH------------HHHHHHHHH--------------------------HHHHHHHHHHHHHHHHHHHHH-HHHHHHHHHH--------

   301
m.      P                                                                    P   P
1. ----------------HHHHHHHHHHHHHHHHH-----------HHHHHHHHHHHHHH---------------HHHHHHH
2. ---------------------HHHHHHHHHHHH---------H-HHHHHHH-H-------------------------HH
3. --------------------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH--------------------HHHHH
4. ----------------------------------------------HH--------------------------------
5. ---------------HHHH--HHHHHHHHHHHH---------HHHHHHHHHHH------------------HHH---HHH
6. ---------------------HH-------------------------------------------H---------H---
7. ------------H-HHHHHHHHHHHHHHHHHHH-----HHHHHHHHHHHHHH------------------HHHHHHHHHH
```

Figure A.3: *YscP secondary structure prediction from seven different prediction algorithms. "H" letters indicate an helical region, "-" random coil. In bold, regions selected for mutations to Alanine or Proline.*

Bibliography

[1] http://www.forcasp.org/paper2127.html.

[2] A. Abraham and H. Liu. Turbulent particle swarm optimization using fuzzy parameter tuning. *Foundations of Computational Intelligence Volume 3*, pages 291–312, 2009.

[3] L. Abrami, M. Fivaz, E. Decroly, N.G. Seidah, F. Jean, G. Thomas, S.H. Leppla, J.T. Buckley, and F.G. van der Goot. The pore-forming toxin proaerolysin is activated by furin. *J Biol Chem*, 273(49):32656–32661, 1998. Department of Biochemistry, University of Geneva, 1211 Geneva, Switzerland.

[4] L. Abrami, M. Fivaz, P. E. Glauser, R. G. Parton, and F. G. van der Goot. A pore-forming toxin interacts with a GPI-anchored protein and causes vacuolation of the endoplasmic reticulum. *J Cell Biol*, 140(3):525–40, 1998. Department of Biochemistry, University of Geneva, 1211 Geneva, Switzerland.

[5] L. Abrami and F.G. van der Goot. Plasma membrane microdomains act as concentration platforms to facilitate intoxication by aerolysin. *J. Cell Biol.*, 147:175–184, 1999.

[6] L. Abrami, M.C. Velluz, Y. Hong, K. Ohishi, A. Mehlert, M. Ferguson, T. Kinoshita, and F. Gisou van der Goot. The glycan core of gpi-anchored proteins modulates aerolysin binding but is not sufficient: the polypeptide moiety is required for the toxin–receptor interaction. *FEBS letters*, 512(1):249–254, 2002.

[7] T. Akiba, Y. Abe, S. Kitada, Y. Kusaka, A. Ito, T. Ichimatsu, H. Katayama, T. Akao, K. Higuchi, E. Mizuki, et al. Crystal structure of the parasporin-2 bacillus thuringiensis toxin that recognizes cancer cells. *Journal of molecular biology*, 386(1):121–133, 2009.

[8] A. Aksimentiev and K. Schulten. Imaging [alpha]-hemolysin with molecular dynamics: ionic conductance, osmotic permeability, and the electrostatic potential map. *Biophysical journal*, 88(6):3745–3761, 2005.

[9] F. Alber, S. Dokudovskaya, L.M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, et al. Determining the architectures of macromolecular assemblies. *Nature*, 450(7170):683–694, 2007.

[10] Ingemar Andre, Philip Bradley, Chu Wang, and David Baker. Prediction of the structure of symmetrical protein assemblies. *Proceedings of the National Academy of Sciences*, October 2007.

[11] P. Angeline. Evolutionary optimization versus particle swarm optimization: Philosophy and performance differences. In *Evolutionary Programming VII*, pages 601–610. Springer, 1998.

[12] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A*, 98(18):10037–10041, August 2001.

[13] P.F. Batcho, D.A. Case, and T. Schlick. Optimized particle-mesh ewald/multiple-time step integration for molecular dynamics simulations. *The Journal of Chemical Physics*, 115:4003, 2001.

[14] A. Berchanski, D. Segal, and M. Eisenstein. Modeling oligomers with cn or dn symmetry: application to capri target 10. *Proteins: Structure, Function, and Bioinformatics*, 60(2):202–206, 2005.

[15] H.J.C. Berendsen, J.P.M. Postma, W.F. Van Gunsteren, A. DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684, 1984.

[16] D. Besozzi, P. Cazzaniga, G. Mauri, D. Pescini, and L. Vanneschi. A comparison of genetic algorithms and particle swarm optimization for parameter estimation in stochastic biochemical systems. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 116–127, 2009.

[17] K.J. Binkley and M. Hagiwara. Balancing exploitation and exploration in particle swarm optimization: velocity-based reinitialization. *Information and Media Technologies*, 3(1):103–111, 2008.

[18] D. Branton, D.W. Deamer, A. Marziali, H. Bayley, S.A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, et al. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10):1146–1153, 2008.

[19] J.T. Buckley, U. Wilmsen, C. Lesieur, A. Schulze, F. Pattus, M.W. Parker, and F.G. van der Goot. Protonation of histidine-132 promotes oligomerization of the channel-forming toxin aerolysin. *Biochemistry*, 34(50):16450–16455, 1995.

[20] S.E. Burr, T. Wahli, H. Segner, D. Pugovkin, J. Frey, et al. Association of type iii secretion genes with virulence of aeromonas salmonicida subsp. salmonicida. *Diseases of aquatic organisms*, 57(1):167–171, 2003.

[21] DA Case, TA Darden, TE Cheatham III, CL Simmerling, J. Wang, RE Duke, R. Luo, RC Walker, W. Zhang, KM Merz, et al. Amber 11. *University of California, San Francisco*, 2010.

[22] Agrain Celine, Callebaut Isabelle, Journet Laure, Sorg Isabel, Paroz Cecile, Mota L. Jaime, and R. Cornelis. Characterization of a type iii secretion substrate specificity switch (t3s4) domain in yscp from yersinia enterocolitica. *Molecular Microbiology*, 56(1):54–67, April 2005.

[23] D.S. Cerutti, P.L. Freddolino, R.E. Duke Jr, and D.A. Case. Simulations of a protein crystal with a high resolution x-ray structure: evaluation of force fields and water models. *The Journal of Physical Chemistry B*, 2010.

[24] T. Chakraborty, A. Schmid, S. Notermans, and R. Benz. Aerolysin of aeromonas sobria: evidence for formation of ion-permeable channels and comparison with alpha-toxin of staphylococcus aureus. *Infection and immunity*, 58(7):2127, 1990.

[25] Y.J. Chen and M. Inouye. The intramolecular chaperone-mediated protein folding. *Current opinion in structural biology*, 18(6):765–770, 2008.

[26] S. Cheng and Y. Shi. Diversity control in particle swarm optimization. In *Swarm Intelligence (SIS), 2011 IEEE Symposium on*, pages 1–9. IEEE, 2011.

[27] N.Y. Chirgadze, S.L. Briggs, K.A. McAllister, A.S. Fischl, and G. Zhao. Crystal structure of streptococcus pneumoniae acyl carrier protein synthase: an essential enzyme in bacterial fatty acid biosynthesis. *The EMBO journal*, 19(20):5281–5287, 2000.

[28] M. Clerc. The swarm and the queen: towards a deterministic and adaptive particle swarm optimization. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 3. IEEE, 1999.

[29] M. Clerc and J. Kennedy. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *Evolutionary Computation, IEEE Transactions on*, 6(1):58–73, 2002.

[30] Poupon A. Coeytaux K. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, (21):1891–1900, 2005.

[31] J.A. Cuff and G.J. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 40(3):502–511, 2000.

[32] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *Journal of Chemical Physics*, 98:10089–10089, 1993.

[33] S. De and R. Olson. Crystal structure of the vibrio cholerae cytolysin heptamer reveals common features among disparate pore-forming toxins. *Proceedings of the National Academy of Sciences*, 108(18):7385, 2011.

[34] S.J. de Vries, A.D.J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, and A.M.J.J. Bonvin. Haddock versus haddock: new features and performance of

haddock2. 0 on the capri targets. *Proteins: structure, function, and bioinformatics*, 69(4):726–733, 2007.

[35] D.B. Diep, K.L. Nelson, S.M. Raja, E.N. Pleshak, and J.T. Buckley. Glycosylphosphatidylinositol anchors of membrane glycoproteins are binding determinants for the channel-forming toxin aerolysin. *Journal of Biological Chemistry*, 273(4):2355, 1998.

[36] Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, 347(4):827–839, April 2005.

[37] B. Egwolf, Y. Luo, D.E. Walters, and B. Roux. Ion selectivity of α-hemolysin with β-cyclodextrin adapter. ii. multi-ion effects studied with grand canonical monte carlo/brownian dynamics simulations. *The Journal of Physical Chemistry B*, 114(8):2901–2909, 2010.

[38] E. Elbeltagi, T. Hegazy, and D. Grierson. Comparison among five evolutionary-based optimization algorithms. *Advanced Engineering Informatics*, 19(1):43–53, 2005.

[39] N. Eswar, B. Webb, M.A. Marti-Renom, MS Madhusudhan, D. Eramian, M.Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using modeller. *Curr Protoc Protein Sci*, 2(12):15–32, 2007.

[40] M. Fivaz, M.C. Velluz, and F.G. van der Goot. Dimer dissociation of the pore-forming toxin aerolysin precedes receptor binding. *Journal of Biological Chemistry*, 274(53):37705–37708, 1999.

[41] M. Fivaz, F. Vilbois, S. Thurnheer, C. Pasquali, L. Abrami, P.E. Bickel, R.G. Parton, and F.G. Van Der Goot. Differential sorting and fate of endocytosed gpi-anchored proteins. *The EMBO journal*, 21(15):3989–4000, 2002.

[42] D. Floreano and C. Mattiussi. *Bio-inspired artificial intelligence: theories, methods, and technologies.* The MIT Press, 2008.

[43] D. Frishman and P. Argos. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*, 9(2):133, 1996.

[44] J. Garnier, J.F. Gibrat, and B. Robson. Gor method for predicting protein secondary structure from amino acid sequence. *Methods in enzymology*, 266:540–553, 1996.

[45] H. Gohlke, C. Kiel, and D.A. Case. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the ras-raf and ras-ralgds complexes. *Journal of molecular biology*, 330(4):891–913, 2003.

[46] S.D. Goldberg, C.S. Soto, C.D. Waldburger, and W.F. DeGrado. Determination of the physiological dimer interface of the phoq sensor domain. *Journal of molecular biology*, 379(4):656–665, 2008.

[47] U. Gophna, E.Z. Ron, and D. Graur. Bacterial type iii secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene*, 312:151–163, 2003.

[48] L.Q. Gu, S. Cheley, and H. Bayley. Prolonged residence time of a noncovalent molecular adapter, β-cyclodextrin, within the lumen of mutant α-hemolysin pores. *The Journal of General Physiology*, 118(5):481, 2001.

[49] L.Q. Gu, M. Dalla Serra, J.B. Vincent, G. Vigh, S. Cheley, O. Braha, and H. Bayley. Reversal of charge selectivity in transmembrane protein pores by using noncovalent molecular adapters. *Proceedings of the National Academy of Sciences*, 97(8):3959, 2000.

[50] B. Hess, H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[51] E. Hoiczyk and G. Blobel. Polymerization of a single protein of the pathogen yersinia enterocolitica into needles punctures eukaryotic cells. *Proc Natl Acad Sci U S A*, 98(8):4669–4674, April 2001.

[52] W.G. Hoover et al. Canonical dynamics: equilibrium phase-space distributions. *Physical Review A*, 31(3):1695–1697, 1985.

[53] J. Hsin and K. Schulten. Improved resolution of tertiary structure elasticity in muscle protein. *Biophysical Journal*, 100(4):22, 2011.

[54] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[55] Katsura I. and Hendrix RW. Length determination in bacteriophage lambda tails. *Cell*, 5651(39):691–8, December 1984.

[56] I. Iacovache, M.T. Degiacomi, L. Pernot, S. Ho, M. Schiltz, M. Dal Peraro, and F.G. van der Goot. Dual chaperone role of the c-terminal propeptide in folding and oligomerization of the pore-forming toxin aerolysin. *PLoS pathogens*, 7(7):e1002135, 2011.

[57] I. Iacovache, P. Paumard, H. Scheib, C. Lesieur, N. Sakai, S. Matile, M.W. Parker, and F.G. Van der Goot. A rivet model for channel formation by aerolysin-like pore-forming toxins. *The EMBO Journal*, 25:457–466, 2006.

[58] H. Ikemura, H. Takagi, and M. Inouye. Requirement of pro-sequence for the production of active subtilisin e in escherichia coli. *Journal of Biological Chemistry*, 262(16):7859, 1987.

[59] T. Ishida and K. Kinoshita. Prdos: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Ress*, (35), 2007.

[60] Barry Isralewitz, Jerome Baudry, Justin Gullingsrud, Dorina Kosztin, and Klaus Schulten. Steered molecular dynamics investigations of protein function. *Journal of Molecular Graphics and Modelling*, 19(1):13–25, February 2001.

[61] R. Jacob, B. Pürschel, and H.Y. Naim. Sucrase is an intramolecular chaperone located at the c-terminal end of the sucrase-isomaltase enzyme complex. *Journal of Biological Chemistry*, 277(35):32141, 2002.

[62] L. Journet, C. Agrain, P. Broz, and R. Cornelis. Characterization of a type iii secretion substrate specificity switch (t3s4) domain in yscp from yersinia enterocolitica. *Science*, 5651(302):1757–60, December 2003.

[63] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[64] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[65] J.J. Kasianowicz, E. Brandin, D. Branton, and D.W. Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770, 1996.

[66] J. Kennedy. Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 3. IEEE, 1999.

[67] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 4, pages 1942–1948. IEEE, 1995.

[68] J. Kennedy and R. Mendes. Population structure and particle swarm performance. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 2, pages 1671–1676. IEEE, 2002.

[69] DG Kneller, FE Cohen, and R. Langridge. Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology*, 214(1):171–182, 1990.

[70] K. H. Krause, M. Fivaz, A. Monod, and F. G. van der Goot. Aerolysin induces G-protein activation and Ca2+release from intracellular stores in human granulocytes. *J Biol Chem*, 273(29):18122–9, 1998.

[71] T. Kubori, A. Sukhan, S. I. Aizawa, and J. E. Galán. Molecular characterization and assembly of the needle complex of the salmonella typhimurium type iii protein secretion system. *Proc Natl Acad Sci U S A*, 97(18):10225–10230, August 2000.

[72] S. Kumar, B. Ma, C.J. Tsai, N. Sinha, and R. Nussinov. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Science*, 9(1):10–19, 2000.

[73] Wang L. and Sauer U.H. Ond-crf: predicting order and disorder in proteins using conditional random fields. *Bioinformatics*, April 2008.

[74] A. Laio and F.L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71:126601, 2008.

[75] O.F. Lange, D. Van der Spoel, and B.L. De Groot. Scrutinizing molecular mechanics force fields on the submicrosecond timescale with nmr data. *Biophysical journal*, 99(2):647–655, 2010.

[76] K. Lasker, F. Förster, S. Bohn, T. Walzthoeni, E. Villa, P. Unverdorben, F. Beck, R. Aebersold, A. Sali, and W. Baumeister. Molecular architecture of the 26s proteasome holocomplex determined by an integrative approach. *Proceedings of the National Academy of Sciences*, 109(5):1380–1387, 2012.

[77] C. Lesieur, S. Frutiger, G. Hughes, R. Kellner, F. Pattus, and F.G. van der Goot. Increased stability upon heptamerization of the pore-forming toxin aerolysin. *Journal of Biological Chemistry*, 274(51):36722, 1999.

[78] J.M. Levin, B. Robson, and J. Garnier. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS letters*, 205(2):303–308, 1986.

[79] Chun-Mei Li, Ian Brown, John Mansfield, Conrad Stevens, Tristan Boureau, Martin Romantschuk, , and Suvi Taira. The hrp pilus of the pseudomonas syringae elongates from its tip and acts as a conduit for translocation of the effector protein hrpz. *EMBO*, 8(21):1909–1915, April 2002.

[80] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell. Protein disorder prediction: implications for structural proteomics. *Structure*, 11(11):1453–1459, November 2003.

[81] R. Linding, R.B. Russell, V. Neduva, and T.J. Gibson. Globplot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research*, 31(13):3701–3708, 2003.

[82] S. Łukasik and S. Żak. Firefly algorithm for continuous constrained optimization tasks. *Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems*, pages 97–106, 2009.

[83] Y. Luo, B. Egwolf, D.E. Walters, and B. Roux. Ion selectivity of α-hemolysin with a β-cyclodextrin adapter. i. single ion potential of mean force and diffusion coefficient. *The Journal of Physical Chemistry B*, 114(2):952–958, 2009.

[84] B. Ma, S. Kumar, C.J. Tsai, and R. Nussinov. Folding funnels and binding mechanisms. *Protein engineering*, 12(9):713, 1999.

[85] C.R. MacKenzie, T. Hirama, and J.T. Buckley. Analysis of receptor binding by the channel-forming toxin aerolysin using surface plasmon resonance. *Journal of Biological Chemistry*, 274(32):22604–22609, 1999.

[86] A.D. MacKerell Jr, D. Bashford, M. Bellott, R.L. Dunbrack Jr, JD Evanseck, MJ Field, S. Fischer, J. Gao, H. Guo, S. Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[87] R. M. Macnab. How bacteria assemble flagella. *Annu Rev Microbiol*, 57:77–100, 2003.

[88] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, and A.H. De Vries. The martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[89] R. Meier, M. Pippel, F. Brandt, W. Sippl, and C. Baldauf. ParaDockS: A Framework for Molecular Docking with Population-Based Metaheuristics. *J. Chem. Inf. Model*, 50(5):879–889, 2010.

[90] W. Meining, S. Mörtl, M. Fischer, M. Cushman, A. Bacher, and R. Ladenstein. The atomic structure of pentameric lumazine synthase from saccharomyces cerevisiae at 1.85 å resolution reveals the binding mode of a phosphonate intermediate analogue1. *Journal of molecular biology*, 299(1):181–197, 2000.

[91] M. Meissner, M. Schmuker, and G. Schneider. Optimized particle swarm optimization (opso) and its application to artificial neural network training. *Bmc Bioinformatics*, 7(1):125, 2006.

[92] J.A. Melton-Witt, L.M. Bentsen, and R.K. Tweten. Identification of functional domains of clostridium septicum alpha toxin. *Biochemistry*, 45(48):14347–14354, 2006.

[93] G. Menestrina. Ionic channels formed bystaphylococcus aureus alpha-toxin: Voltage-dependent inhibition by divalent and trivalent cations. *Journal of Membrane Biology*, 90(2):177–190, 1986.

[94] G. Menestrina. Ion channels and bacterial infection: the case of [beta]-barrel pore-forming protein toxins of staphylococcus aureus. *FEBS Letters*, 552(1):54–60, September 2003.

[95] B. Mészáros, P. Tompa, I. Simon, and Z. Dosztányi. Molecular principles of the interactions of disordered proteins. *Journal of molecular biology*, 372(2):549–561, 2007.

[96] M. Misakian and JJ Kasianowicz. Electrostatic influence on ion transport through the ahl channel. *Journal of Membrane Biology*, 195(3):137–146, 2003.

[97] Garrett M. Morris, David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. Automated docking using a lamarckian genetic

algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, January 1999.

[98] L.J. Mota and G.R. Cornelis. The bacterial injection kit: type iii secretion systems. *Annals of medicine*, 37(4):234–249, 2005.

[99] Marcus Mueller, Ulla Grauschopf, Timm Maier, Rudi Glockshuber, and Nenad Ban. The structure of a cytolytic ±-helical toxin pore reveals its assembly mechanism. *Nature*, May 2009.

[100] P.J. Munson, V. Di Francesco, and R. Porrelli. Protein secondary structure prediction using periodic-quadratic-logistic models: Statistical and theoretical issues. In *System Sciences, 1994. Vol. V: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference on*, volume 5, pages 375–384. IEEE, 1994.

[101] C. Mura, D. Cascio, M.R. Sawaya, and D.S. Eisenberg. The crystal structure of a heptameric archaeal sm protein: Implications for the eukaryotic snrnp core. *Proceedings of the National Academy of Sciences*, 98(10):5532, 2001.

[102] V. Namasivayam and R. Günther. PSO@ Autodock: A fast flexible molecular docking program based on swarm intelligence. *Chemical Biology & Drug Design*, 70(6):475–484, 2007.

[103] K.L. Nelson, S.M. Raja, and J.T. Buckley. The glycosylphosphatidylinositol-anchored surface glycoprotein thy-1 is a receptor for the channel-forming toxin aerolysin. *Journal of Biological Chemistry*, 272(18):12170, 1997.

[104] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81:511, 1984.

[105] S.Y. Noskov, W. Im, and B. Roux. Ion permeation through the [alpha]-hemolysin channel: theoretical studies based on brownian dynamics and poisson-nernst-plank electrodiffusion theory. *Biophysical journal*, 87(4):2299–2309, 2004.

[106] Y. Ohnishi and S. Horinouchi. Extracellular production of a serratia marcescens serine protease in escherichia coli. *Bioscience, biotechnology, and biochemistry*, 60(10):1551–1558, 1996.

[107] J. O'Keeffe, I. Cozmuta, D. Bose, and V. Stolc. A predictive md-nernst-planck model for transport in alpha-hemolysin: Modeling anisotropic ion currents. *Chemical Physics*, 342(1-3):25–32, 2007.

[108] M. Ouali and R.D. King. Cascaded multiple classifiers for secondary structure prediction. *Protein Science*, 9(6):1162–1176, 2000.

[109] M.J. Pallen, S.A. Beatson, and C.M. Bailey. Bioinformatics, genomics and evolution of non-flagellar type-iii secretion systems: a darwinian perpective? *FEMS microbiology reviews*, 29(2):201–229, 2005.

[110] M.W. Parker, J.T. Buckley, J.P.M. Postma, A.D. Tucker, K. Leonard, F. Pattus, and D. Tsernoglou. Structure of The Aeromonas toxin proaerolysin in its water-soluble and membrane-channel states. *Nature*, 367:292–295, January 1994.

[111] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[112] M. Pastoriza-Gallego, L. Rabah, G. Gibrat, B. Thiebot, F.G. van der Goot, L. Auvray, J.M. Betton, and J. Pelta. Dynamics of unfolded protein transport through an aerolysin pore. *Journal of the American Chemical Society*, 2011.

[113] S. Pasupuleti and R. Battiti. The gregarious particle swarm optimizer (g-pso). In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 67–74. ACM, 2006.

[114] M.E.H. Pedersen and A.J. Chipperfield. Simplifying particle swarm optimization. *Applied Soft Computing*, 10(2):618–628, 2010.

[115] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. Ucsf chimera – a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[116] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with namd. *J Comput Chem*, 26(16):1781–1802, December 2005.

[117] B. Pierce, W. Tong, and Z. Weng. M-zdock: a grid-based approach for cn symmetric multimer docking. *Bioinformatics*, 21(8):1472–1478, 2005.

[118] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[119] A. Ratnaweera, S.K. Halgamuge, and H.C. Watson. Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients. *Evolutionary Computation, IEEE Transactions on*, 8(3):240–255, 2004.

[120] D.W. Ritchie, D. Kozakov, and S. Vajda. Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational fft generating functions. *Bioinformatics*, 24(17):1865–1873, 2008.

[121] A. Roy-Burman, R.H. Savel, S. Racine, B.L. Swanson, N.S. Revadigar, J. Fujimoto, T. Sawa, D.W. Frank, and J.P. Wiener-Kronish. Type iii protein secretion is associated with death in lower respiratory and systemic pseudomonas aeruginosa infections. *Journal of Infectious Diseases*, 183(12):1767–1774, 2001.

[122] R. Rozenfeld, L. Muller, S.E. Messari, and C. Llorens-Cortes. The c-terminal domain of aminopeptidase a is an intramolecular chaperone required for the correct folding, cell

surface expression, and activity of this monozinc aminopeptidase. *Journal of Biological Chemistry*, 279(41):43285, 2004.

[123] D. Russel, K. Lasker, B. Webb, J. Velázquez-Muriel, E. Tjioe, D. Schneidman-Duhovny, B. Peterson, and A. Sali. Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biology*, 10(1):e1001244, 2012.

[124] J.P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of< i> n</i>-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[125] T. Schneider and E. Stoll. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B*, 17:1302–1322, Feb 1978.

[126] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H.J. Wolfson. Patchdock and symmdock: servers for rigid and symmetric docking. *Nucleic acids research*, 33(suppl 2):W363–W367, 2005.

[127] O. Schraidt and T.C. Marlovits. Three–dimensional model of salmonella's needle complex at subnanometer resolution. *Science*, 331(6021):1192, 2011.

[128] Y. Shi and R. Eberhart. Parameter selection in particle swarm optimization. In *Evolutionary Programming VII*, pages 591–600. Springer, 1998.

[129] I.Y. Shilov and M.G. Kurnikova. Energetics and dynamics of a cyclic oligosaccharide molecule in a confined protein pore environment. a molecular dynamics study. *The Journal of Physical Chemistry B*, 107(29):7189–7201, 2003.

[130] O.S. Smart, J.G. Neduvelil, X. Wang, BA Wallace, and M.S.P. Sansom. Hole: a program for the analysis of the pore dimensions of ion channel structural models. *Journal of molecular graphics*, 14(6):354–360, 1996.

[131] Langzhou Song, Michael R. Hobaugh, Christopher Shustak, Stephen Cheley, Hagan Bayley, and J. Eric Gouaux. Structure of Staphylococcal alpha -Hemolysin, a Heptameric Transmembrane Pore. *Science*, 274(5294):1859–1865, 1996.

[132] T. Spreter, C.K. Yip, S. Sanowar, I. André, T.G. Kimbrough, M. Vuckovic, R.A. Pfuetzner, W. Deng, A.C. Yu, B.B. Finlay, et al. A conserved structural motif mediates formation of the periplasmic rings in the type iii secretion system. *Nature structural & molecular biology*, 16(5):468–476, 2009.

[133] J. Srinivasan, T.E. Cheatham III, P. Cieplak, P.A. Kollman, and A. David. Continuum solvent studies of the stability of dna, rna, and phosphoramidate-dna helices. *Journal of the American Chemical Society*, 120(37):9401–9409, 1998.

[134] GM Torrie and JP Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[135] Leonardo G. Trabuco, Elizabeth Villa, Kakoli Mitra, Joachim Frank, and Klaus Schulten. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, 16(5):673–683, May 2008.

[136] C.J. Tsai, B. Ma, and R. Nussinov. Folding and binding cascades: shifts in energy landscapes. *Proceedings of the National Academy of Sciences*, 96(18):9970, 1999.

[137] Y. Tsitrin, C. J. Morton, C. el Bez, P. Paumard, M. C. Velluz, M. Adrian, J. Dubochet, M. W. Parker, S. Lanzavecchia, and F. G. van der Goot. Conversion of a transmembrane to a water-soluble protein complex by a single point mutation. *Nat Struct Biol*, 9(10):729–33, 2002.

[138] R. Uellner, M.J. Zvelebil, J. Hopkins, J. Jones, L.K. MacDougall, B.P. Morgan, E. Podack, M.D. Waterfield, and G.M. Griffiths. Perforin is activated by a proteolytic cleavage during biosynthesis which reveals a phospholipid-binding c2 domain. *The EMBO journal*, 16(24):7287–7296, 1997.

[139] F.G. van der Goot, KR Hardie, MW Parker, and J.T. Buckley. The c-terminal peptide produced upon proteolytic activation of the cytolytic toxin aerolysin is not involved in channel formation. *Journal of Biological Chemistry*, 269(48):30496–30501, 1994.

[140] Stefanie Wagner, Isabel Sorg, Matteo Degiacomi, Laure Journet, Matteo Dal Peraro, and Guy R. Cornelis. The helical content of the yscp molecular ruler determines the length of the yersinia injectisome. *Molecular Microbiology*, 71(3):692–701, 2009.

[141] HU Wilmsen, KR Leonard, W. Tichelaar, JT Buckley, and F. Pattus. The aerolysin membrane channel is formed by heptamerization of the monomer. *The EMBO journal*, 11(7):2457, 1992.

[142] W. Wriggers. Using situs for the integration of multi-resolution structures. *Biophysical reviews*, 2(1):21–27, 2010.

[143] H. Xu, C. Yang, L. Chen, I.A. Kataeva, W. Tempel, D. Lee, J.E. Habel, D. Nguyen, J.W. Pflugrath, J.D. Ferrara, et al. Away from the edge ii: in-house se-sas phasing with chromium radiation. *Acta Crystallographica Section D: Biological Crystallography*, 61(7):960–966, 2005.

[144] X.S. Yang. Firefly algorithms for multimodal optimization. *Stochastic algorithms: foundations and applications*, pages 169–178, 2009.

[145] X.S. Yang. *Nature-inspired metaheuristic algorithms*. Luniver Press, 2011.

[146] X.S. Yang and S. Deb. Cuckoo search via lévy flights. In *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*, pages 210–214. IEEE, 2009.

[147] Z. R. Yang, R. Thomson, P. Mcneil, and R. M. Esnouf. Ronn: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, 21(16):3369–3376, August 2005.

[148] C.K. Yip, T.G. Kimbrough, H.B. Felise, M. Vuckovic, N.A. Thomas, R.A. Pfuetzner, E.A. Frey, B.B. Finlay, S.I. Miller, and N.C.J. Strynadka. Structural characterization of the molecular platform for type iii secretion system assembly. *Nature*, 435(7042):702–707, 2005.

MATTEO DEGIACOMI

Rue du Lac 20
1020 Renens (VD)
Switzerland
Phone: +4179 305 17 27
E-Mail: matteothomas.degiacomi@epfl.ch

23.12.1981
Swiss

Academic Background

2012 PhD in Biotechnology and Bioengineering, Laboratory for Biomolecular Modeling, School of Life Sciences, Swiss Federal institute of Technology Lausanne (EPFL), Switzerland. *Molecular Modeling of Bacterial Nanomachineries*

2007 - Master of Science MSc in Computer Science (EPFL), Specialized in Biocomputing Average 5.4/6
- Master Project in bio-inspired robotics (BIRG laboratory, EPFL) *Robotics applications of vision-based action selection*

2006 Semester Project in bio-inspired robotics (BIRG laboratory, EPFL) *Locomotion exploiting body dynamics*

2005 Bachelor of Science BSc in Computer Science (EPFL)

2001 School leaving certificate (scientific orientation), Liceo Lugano 2

Working Experiences

2011 - teaching assistant for *Bioinformatics* and *Biomolecular Structure and Mechanics* classes (EPFL)
- Administrator of Laboratory for Biomolecular Modeling's cluster and workstations

2007 - Collaborator of the BIRG laboratory (EPFL). Technician and speaker during robot demonstrations at NextFest's technological exposition (Los Angeles)

2006 - Internship in Paul Scherrer Institut (PSI), Switzerland. Programming of optimisation and image processing algorithms in C++ and PHP
- Assistant for the *Models of Biological Sensory-Motor Systems* class (EPFL)

2005 - Teacher for a mathematics support class in Canobbio's secondary school.
- Creation of a PHP/MySQL web site (www.bandadicanobbio.ch) and of a multimedia CD for the Canobbio Music Band (Ticino).

Languages

Italian (mother language), **French** (level C2 of the European Standard of Languages), **English** (Cambridge Certificate in Advanced English, level C1), **German** (level A2)

Information Technologies

Operating Systems	Linux (Redhat, Debian, Suse), Windows (7 and previous)
Software	Matlab, VMD, PyMol, Chimera, Inkscape, Flash MX
Molecular Dynamics	NAMD, Amber, Gromacs, LAMMPS
Databases	MySQL, Oracle
Programming/Scripting	Python , Perl, Shell, PHP, Java, C++

Memberships

Biophysical Society
Swiss Chemical Society

Skills

- fast learning
- precise in work and in the respect of deadlines
- good interpersonal relations and ability to work in a team
- good teaching skills

Personal Interests

Music	classic and electric guitar, trombone
Sport	rollerblades, alpine ski, football (goalkeeper) and beach volley

Publications

M.T.Degiacomi, M.Dal Peraro, *Macromolecular assembly prediction using swarm intelligence*, submitted

I. Iacovache, M. T. Degiacomi and F. G. van der Goot, *“Pore Forming Toxins”*, in: Comprehensive Biophysics, Vol 5, Membranes, 2012

I. Iacovache, M. T. Degiacomi, L. Pernot, M. Schiltz, M. Dal Peraro and F. G. van der Goot, *Dual intramolecular chaperone role of the aerolysin C-terminal propeptide in folding and heptamerization*. PLoS Pathogens, 2011

S. Wagner, I. Sorg, M. Degiacomi, L. Journet, M. Dal Peraro, and G. R. Cornelis. *The helical content of the YscP molecular ruler determines the length of the Yersinia injectisome.* Molecular Microbiology, 2009

S. Gowal, M. T. De Giacomi, and J.-Y. Le Boudec. *Comment on: A Validated Model of Cell-Mediated Immune Response to Tumor Growth*. Cancer Research*,* 2007

Conferences, Workshops and Seminars

Pores 2012, Prato, Italy, 2012
Presentation: *“The unexpected mechanism of aerolysin pore formation”*

Biophysical Society, San Diego, USA, 2012
Poster: “*Unraveling the assembly of macromolecular machines: the case of the pore-forming toxin aerolysin*”

Durham University, UK, 2011
Presentation: “*a complex form of bacterial attack: the pore-forming toxin Aerolysin*”

CECAM Workshop “Combining Experimental and Computiational Techniques to Study Protein Behavior”, Lugano, Switzerland, 2011
Poster: *Activation and Assembly Process of the Pore-Forming Aerolysin*

WATOC Ninth Triennial Congress, Santiago de Compostela, Spain, 2011
Poster: *Dual intramolecular chaperone role of the aerolysin C-terminal propeptide in folding and heptamerization*

Lausanne Biomolecular Modeling Seminar Series, EPFL, 2011
Presentation: *A complex form of bacterial attack: the pore-forming toxin aerolysin*

International Workshop on Coarse-Grained Biomolecular Modeling, Levi, Finland, 2010
FEBS Congress, Gothenburg, Sweden, 2010
Poster: *The role of the C-terminal peptide in the assembly and activation of aerolysin poreforming toxin*

CCP5 Summer School in Molecular Simulation, Sheffield, UK, 2009
6th International NCCR Symposium on New Trends in Structural Biology, Zurich, Switzerland, 2008

ISQBP President’s Meeting, Pushing the Boundaries of Biomolecular Simulation, Ascona, Switzerland, 2008