University of Udine

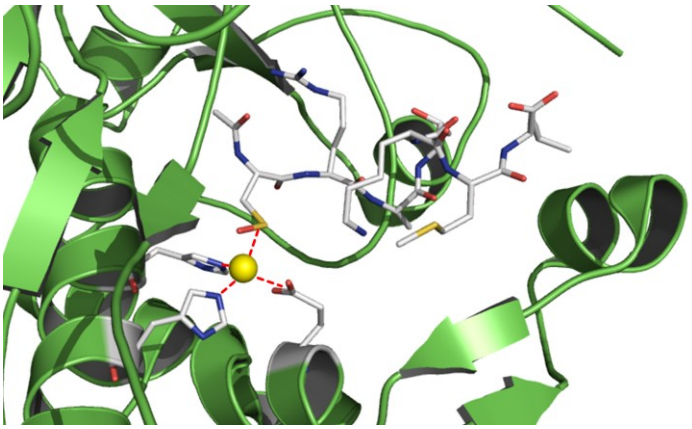
Doctorate course in

Biomedical sciences and biotechnology

CYCLE XXVIII

Research doctorate thesis

Genomic and virtual characterization of Italian group I *Clostridium botulinum* lineages.



Candidate:

Ferdinando Spagnolo

Tutors:

Professor Brancolini Claudio

Col. Florigio Lista

ACADEMIC YEAR

2015/2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

Francesca, your gorgeous support, in every aspect of my life, is a sweet, endlessly, love proof.

GLOSSARY

Next Generation Sequencing A high-throughput sequencing method which parallelizes the sequencing process, producing thousands or millions of sequences at once.
Deep Sequencing Techniques of nucleotide sequence analysis that increase the range, complexity, sensitivity, and accuracy of results by greatly increasing the scale of operations and thus the number of nucleotides, and the number of copies of each nucleotide sequenced.
Paired-End Sequencing Sequence both ends of the same fragment and keep track of the paired data.
Adapter Short oligonucleotides which are attached to the DNA to be sequenced. An adapter can provide a priming site for both amplification and sequencing of the adjoining, unknown nucleic acid.
Library A collection of DNA fragments with adapters ligated to each end.
Bridge Amplification Generation of in situ copies of a specific DNA molecule on an oligo-decorated solid support.
Emulsion PCR A method for bead-based amplification of a library. A single adapter-bound fragment is attached to the surface of a bead, and an oil emulsion containing necessary amplification reagents is formed around the bead/fragment component. Parallel amplification of millions of beads with millions of single strand fragments produces a sequencer-ready library.
Alignment Mapping of sequence reads to a known reference sequence
Reference sequence/genome A fully assembled version of a genome that can be used for mapping short DNA sequence reads for comparisons of genomes from various individuals
Coverage Depth The number of nucleotides from reads that are mapped to a given position of reference genome.
Specificity The percentage of sequences that map to the intended targets out of total bases per run.
Uniformity The variability in sequence coverage across target regions.
Homopolymer Uninterrupted stretch of a single nucleotide type (e.g., TTT or GGGGGG)
InDel InDel stands for Insertion or deletion. A form of structural variation in which a DNA segment is either deleted or inserted.
SNP SNP stands for Single Nucleotide Polymorphism. A single base difference found when comparing the same DNA sequence from two different individuals.

ABSTRACT

Clostridium botulinum is a taxonomic designation that comprehends a broad variety of spore forming, Gram-positive bacteria producing the botulinum neurotoxin (BoNT). *C. botulinum* is the etiologic agent of botulism, a rare but extremely severe neuroparalytic syndrome. Fine-resolution genetic characterization of *C. botulinum* isolates of any BoNT type is relevant for epidemiological studies, forensic microbiology and medical treatment. In this research we sequenced a set of Group I *C. botulinum* genomes. These new sequences where included in a yet published *C. botulinum* genome dataset in order to describe the resulting phylogeny. The genetic characterization allowed us to identify a new BoNT subtype, BoNT/F8. Moreover, BoNT genes were used to model nine diverse botulin neurotoxins, which were characterized together with the published experimental dataset in order to describe the interaction between BoNT catalytic domain structures and a set of ten peptide inhibitors.

RÉSUMÉ

It is well known that, in a warfare environment, one of the most harmful weapon derives from the use of biological agents. In addition, it is dangerously growing the possible biological threat from terroristic attacks. Moreover, nowadays, the worlds' globalized communication system not only allows a quick transportation mean, but it is also a vehicle to foster the spread of highly contagious diseases (e.g. *ebola*). For these reasons it is essential to obtain fast and detailed tools to promptly detect & identify the biological element on the field and to hold a robust technique in order to assess correctly any attribution of blame. My research, therefore, is focused on the analysis of genetic variation of bacteria by means of Next Generation Sequencing (NGS) techniques to build a genetic reference database useful for definitive identification and attribution. NGS is a powerful tool that permits labs to be able to: firstly, a high throughput threat identification of any sampled specimen; secondly, the amount of data produced is a valuable source of information to be exploited in order to better understand not only the level of risk due to a specific threat but also all the remedies (from the defense policy to the clinical strategy).

The main target of the research is the retrieval of whole genomes genetic information in order to produce phylogenetic graphs for variability analysis. Hence my skills are used for data mining and processing by means of bioinformatics tools. Our facilities comprehend two main NGS platforms: Illumina MiSeq and Roche 454 flex. The flow of operations begins from the data experimentally obtained (in this case in form of reads) contigs are produced and aligned to reference genomes in order to obtain a whole genome sequence. The starting point of the research has been the testing of open source bioinformatics tools. A well-known worked example was chosen as a test case: the outbreak of hemolytic uremic syndrome caused by *E. Coli* in Germany in 2011. In that specific epidemic, it was very important to retrieve the source of the infection, therefore the huge amount of sequencing data that had been produced is now available. The test consisted in assembly, ordering of contigs, annotation, genome comparison and benchmarking. It has been confirmed that the adopted assembly tools were robust and

efficient. Abyss was chosen as assembler for the ease of use (even though other software based on the same algorithm might be faster – *e.g.* Velvet). The exploration of the ordered assembly was performed best by means of visualization. Mauve had been chosen because of its robust built-in progressive alignment routine and, mainly, because its graphical interface allows both a visual inspection of the results and annotation tools. Phylogenetic analysis was performed by means of Clonal Frame and Clonal Origin suite of programs. The summarized methods were validated reproducing *E. Coli* test case and the described flow of procedures, was used to perform a whole genome sequencing of *Brucella Abortus* gene 1250, submitted to NGS platform. *Brucella* Chromosomes were added to the previously obtained phylogeny.

Genome recombination has been studied by means of Clonal Frame and Clonal Origin, moreover, through Python scripts were performed for data mining purposes and for the implementation of a routine for Clonal Frame output processing capable to describe the rate of variation within a set of genomes. Outputs are provided as data matrix and heat-map graph.

Previously, a wide panel of bacteria and viruses were sequenced and mapped to a reference genome, or de novo sequenced (*Brucella, Bacillus, Neisseria, Yersinia, Chikungunya*…). In the past we used phylogenetic trees to infer a clonal evolution of the core genomes. In this work we constructed phylogenetic trees based on single nucleotide polymorphisms (SNPs). Since there is no ready-to-use pipeline to generate phylogenetic trees from SNPs, we developed software capable, exploiting pre-existing canonical SNPs matrixes, to directly spot SNPs to a brand new sequence. The code, written in Python, follows these steps:

1. Get the SNPs matrix (creating it from a set of genomes if required);
2. Map the SNPs to a selected reference genome;
3. Pairwise alignment (multiple a. in refinement) & check;
4. SNPs labelling in the new sequence;
5. New SNPs set insertion in new ranked matrix.

This new tool provides us the possibility to produce a phylogenetic tree directly.

Finally, the genomic information obtained with Next Generation Sequencing platforms has been exploited to model the corresponding structures. A total of nine virtual *Botulin* Neurotoxin models were build using homology modelling procedures applying PyMod and MODELLER suite of software. All structures underwent quality assessment and reached quality scores comparable to experimentally obtained ones. Moreover, we inserted the virtual catalytic domains in a set of over 60 experimentally resolved catalytic structures in order to perform docking calculations across selected peptide inhibitors from deposited BoNT-peptide-inhibitor complexes. The catalytic zinc-coordinated function was build using Autodock4$_{Zn}$ docking scripts while its metal-tailored parameters and scoring function and algorithm provided ligand-receptor affinity estimates. Even if the theoretical methods used here remained at the mechanical level, it has been possible to produce experimental-level quality metal-coordinated models for all genomic dataset. Most of the obtained docking poses were coherent with experimental homologous complex structures and docking estimates of binding affinity allowed us to cluster the theoretically obtained model in the expected experimental serotype feature. Even though *ab initio* minimisations of each Zn-coordinated virtual structure or *ab initio* molecular dynamics calculations would enhance the accuracy of the results, the resulted virtual models are yet valuable for exploitation and the retrieved quantity of data can be used to improve the applied scoring function. Eventually, the virtual structure of the new Botulin neurotoxic subtype F8 - descripted only at genomic level – was produced and its catalytic function has been tested with a set of known peptide inhibitors through molecular docking showing a native behaviour towards BoNT/F specific peptide inhibitors.

CHAPTER 1
INTRODUCTION

1.1 Clostridium botulinum

Clostridium botulinum is a pathogenic Gram positive, sporulating, anaerobic bacterium, triggering various forms of botulism, a disease occurring in various forms: foodborne, infant, adult form of infant botulism and iatrogenic. This bacterium is classified into four distinct groups (I-IV) depending on the capability of the microorganism to digest complex proteins, Table 1.

Table 1: Clostridium botulium classification.

	Group I	Group II	Group III	Group IV
Toxin Types	**A, B, F**	**B, E, F**	**C, D**	**G**
Proteolysis	**+**	**-**	**weak**	**-**
Saccharolysis	**-**	**+**	**-**	**-**
Host	**human**	**human**	**animal**	**-**
Toxin gene	chromosome/plasmid	chromosome/plasmid	bacteriophage	plasmid
Close relatives	*C. sporogenes, C. putrificum*	*C. butyricum, C. beijerinickii*	*C. haemolyticum, C. novyi type A*	*C. subterminale, C. haemolyticum*

Group I is proteolytic and causes human botulism. The bacteria belonging to the *Clostridium* genus - *Clostridium tetani Clostridium difficile*, *Clostridium perfringens* and *Clostridium sordelli* - are worldwide distributed in the anaerobic regions of the intestines of animals in the form of spores (Figure 1).

This form allows *Clostridium* bacteria to survive for long periods even in extreme physical / chemical conditions until advantageous conditions are reached, germinating into the vegetative form. The switch between the two forms (vegetative and sporigenous) is the surrounding habitat: presence of nutrients combined with humidity and

anaerobiosis drives the vegetative form, while a poorly enriched but oxygenated habitat provokes the metamorphosis into the spore.

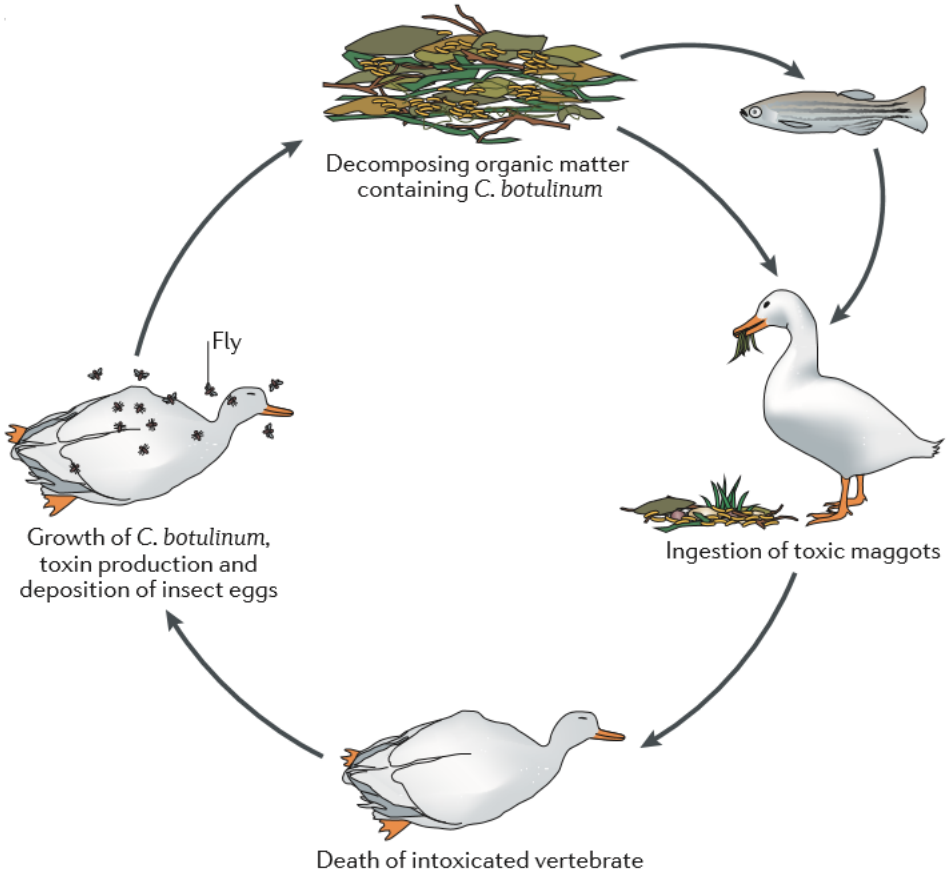


Figure 1: Clostridium botulinum wild life cycle (Nat Rev Microbiol. 2014 Aug;12(8):535-49).

A variety of *Clostridia* are pathogenic due to the production of protein toxins. Few *Clostridia* like *C. tetani*[1] and *C. botulinum* secrete neurotoxin that act in the nervous system

[1] Tetanus affects skeletal muscle, a type of striated muscle used in voluntary movement. The other type of striated muscle, cardiac, or heart muscle, cannot be tetanized because of its intrinsic electrical properties. The tetanus toxin initially binds to peripheral nerve terminals. It is transported within the axon and across synaptic junctions until it reaches the central nervous system. There it becomes rapidly fixed to gangliosides at the presynaptic inhibitory motor nerve endings, and is taken up into the axon by endocytosis. The effect of the toxin is to block the release of inhibitory neurotransmitters glycine and gamma-aminobutyric acid (GABA) across the synaptic cleft, which is required to check the nervous impulse. If nervous impulses cannot be checked by normal inhibitory mechanisms, the generalized muscular spasms characteristic of tetanus are produced. The toxin appears to act by selective cleavage of a protein component of synaptic vesicles, synaptobrevin II, and this prevents the release of neurotransmitters by the cells [162].

provoking a spastic (tetanus) and a flaccid paralysis (botulinum). Nevertheless, it is known that more than 40 different botulinum neurotoxins (BoNT) are produced by six distinct *Clostridia* [1] [2].

BoNTs quaternary structure is composed by four domains arranged into three functional units, two enabling the binding and translocation into the peripheral nerve terminals and the metalloprotease provoking the inhibition of the neurotransmitter (shown in Figure 5). On the one hand, the peculiar BoNTs neurospecificity and effectiveness makes them to be the most powerful toxins ever known, conferring them the primacy of being the most harmful potential bioterrorism warfare agents [3] [4]. On the other hand, their selectivity has been exploited for pharmaceutical purposes to treat human diseases characterized by an abnormal functionality of nerve terminals.

1.2 Botulism

Botulism is a disease that occurs to animals and human beings. It is reported that a botulism outbreak can lead to thousands of intoxications in few days [5]: in a natural environment an outbreak can be auto-boosting due to the fact that *C. botulinum* spores can easily germinate in decomposing organic material under anaerobiotic conditions such decomposing carcasses [6] [7]. Figure 1 depicts *C. botulinum* life cycle. Decomposing organic matter, infected by *C. botulinum* vegetative form and BoNTs, are eaten by vertebrates where, in their intestine, it turns into spore. BoNT insensitive invertebrates may contribute to disseminate the bacterium and its toxin. The ingested BoNT is provokes physiological dysfunctions that drive the animal to death. In the cadaver the spore germinates and multiply, producing the neurotoxin too. In nature, the toxigenicity of *clostridia* strains is not an absolute condition: it is demonstrated a high recombination rate along the evolution of C. botulinum species via vertical and horizontal transmission [8]. Human botulism occurs very rarely developing five different forms, depending on the way of access: food-borne botulism, infant botulism, adult form of infant botulism, iatrogenic botulism and inhalational botulism (Figure 2).

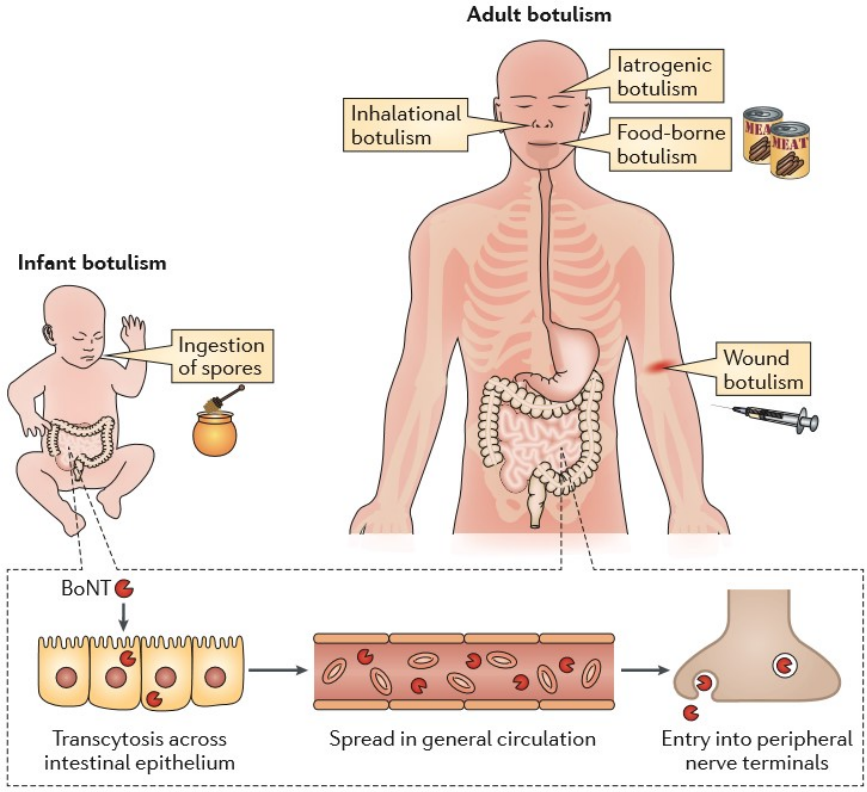


Figure 2: Human forms of botulism (Nat Rev Microbiol. 2014 Aug;12(8):535-49).

Food-borne botulism. It manifests after the ingestion of BoNT pre-contaminated food. If the toxin survives in the digestive proteolytic habitat, it is absorbed in the intestine from where it reaches its biological target.

Infant botulism. Is provoked by the ingestion of neuro-toxigenic spores. In the babies' intestine, the spores germinate producing the neurotoxin. Since the infant has not developed the whole microbiota, *C. botulinum* suffers a limited bacterial competition. This environment facilitates *botulinum* infection.

Adult form of infant botulism. Some pathological, morphological conditions or chemoprophylaxis may reproduce the peculiarities of infant microbiota exposing an adult to the intoxication. In the case of food-borne, infant and adult form of infant botulism, there are two key steps in the intoxication dynamics, BoNT passage through the digestive apparatus layers (intestinal mucus layer, polarized intestinal epithelial monolayer reaching the general circulation) and the arrival to the peripheral cholinergic nerve terminals (leading to the peculiar paralysis) [7] [9].

Iatrogenic botulism. In case of direct contact of the neuro-muscular plaque favored by intentional injection (pharmaceutical or cosmetic treatment) or wounded tissue exposure to BoNT, the intestine absorption is bypassed and is immediately delivered in the blood circulation from where it can be delivered to its cholinergic target [10] [11].

Inhalational botulism. Eventually BoNT may be inhaled. From the respiratory tract BoNT may reach blood stream and successively its target.

Finally, food-borne and infant botulism are the predominant forms of the disease while the other forms are rarely encountered (the Inhalational form almost negligible for the highly inefficient intoxication via aerosol [3] [12]).

It is still unknown the mechanisms that drive BoNT through the lymphatic system and the blood circulation. However, it is proved that this peptide neurotoxin can last days in the hosts' circulation [13] [14]. Moreover, it is proved that BoNT are not able to reach the central nervous system via blood or lymphatic streams [10]. It is than remarkable the extreme BoNT specificity to the peripheral skeletal ad autonomic cholinergic nerves (Paragraph 1.3.1). As stated, BoNT induce a flaccid paralysis blocking the cholinergic transmission. Botulism lethality is linked to the paralysis of the respiratory muscles and the inevitable impossibility to breathe. Botulism affected patients fully recover if mechanical artificial ventilation is guaranteed [1] [12]. The duration of the paralysis depends on the species affected and BoNT serotype [15] [16] [17]. It is still unknown how BoNTs are metabolized in the neuro-muscular junction, but it is assessed that *in vivo* their action is reversible and the tissue fully restores.

Pharmacological treatment. There is no approved pharmacological treatment. Basically there are three pioneering approaches: vaccines, antibodies and small molecules. *Vaccines.* Various BoNT HC domains (Paragraph 1.3.1) were expressed in *Pichia pastoris* and were shown to induce protective antibodies in animals [18] [19]. A recombinant vaccine composed of the HC domains from BoNT/A1 and BoNT/B1 has shown promising results in clinical trials, and vaccines for other serotypes are now under development. Other BoNT domains have been tested in animals [20]. *Antibodies.* Specific antitoxin antibodies might be used to prevent and treat botulism by eliminating circulating BoNTs. The

identification of BoNT serotypes is often pursued by means of antibodies. The delivery of BoNT-specific antibodies to the cholinergic neurons is far for being effective. However, sophisticated biotechnologies capable to produce high-affinity humanized monoclonal antibodies are now available and have been used to produce BoNT/A-, BoNT/B-, BoNT/E- and BoNT/F-specific antibodies [21] [22] [23]. Like other antibody engineering applications, another promising approach is to generate single-chain toxin-binding camelid-like antibodies, which have the potential for intracellular use (Selection of antibody and antibody fragments structure types schemes in Figure 3).

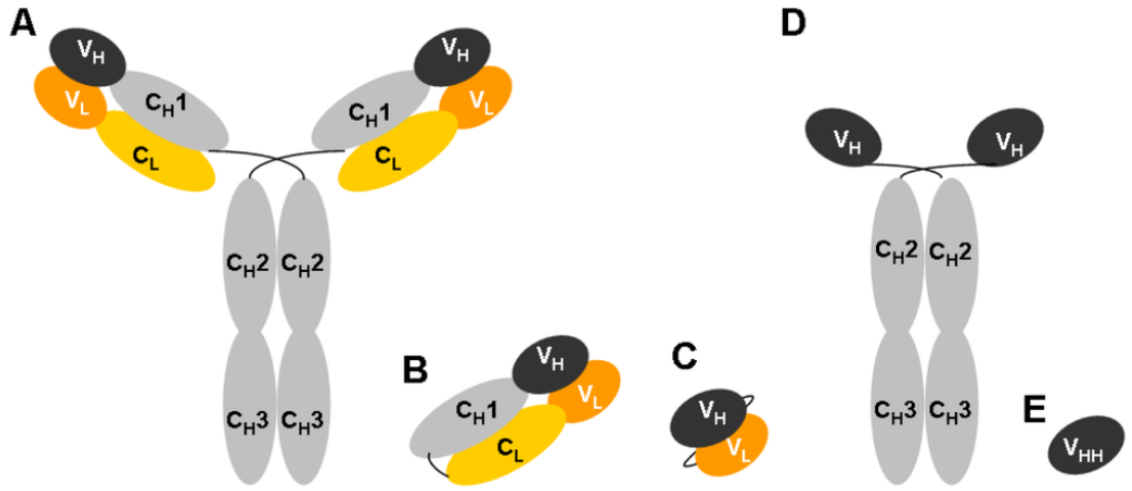


Figure 3: Selection of antibodies and antibody fragments. A) Whole human IgG; B) Fab (Fragment antigen binding); C) ScFv (Single-chain variable Fragment); D) Camelid HcAb (Heavy-chain only Antibody); E) VHH/Nanobodies (Variable domain of HcAb) (Figure from Int. J. Mol. Sci. 2012, 13(4), 4727-4794;).

Small molecules. The development of small molecules capable to block the BoNT metalloprotease activity in the cholinergic neuron is complicated, owing to the complex BoNT chain L-substrate interaction. Few molecules, mainly zinc chelating compounds, have passed the stage of inhibition of the toxins in cultured neurons and have therefore not yet reached the level of testing in animals [24] [25].

Nevertheless, beyond the yet familiar application of BoNTs in cosmetics or aesthetic medicine, a variety of BoNTs medical applications is proposed. The possibility to locally induce a flaccid paralysis may relieve the effects of an uncontrolled muscular contraction. For certain conditions, such as facial lacerations or disjointed bone fractures, a toxin that

has a short duration of action might be more useful in ameliorating the course and outcome of the illness. Moreover, from a technical perspective, the persistence of BoNT activity is important for their long duration of action, this implies the inoculation of fewer injections of very low doses. BoNT engineering can lead to a safe, long lasting, super specific bio-compound [26] [27] [28].

1.3 Botulin neurotoxin (BoNT)

There are seven known serotypes (A-G) of BoNT (up to 40 genetic variants, Table 2) [29] [30] [31]. BoNTs are conventionally classified into subtypes depending on their amino acidic sequences. *bont* genes are encoded within mobile genetic elements that allow a horizontal transfer among different isolates [29] [30] [32]. Since BoNT strains can be present in both botulinum chromosome and plasmids, and a high recombination rate is demonstrated, the expressed neurotoxin is very mutable, the collection of *C. botulinum* isolates of any BoNT type is relevant for both epidemiological studies and forensic microbiology.

Table 2: botulin neurotoxin classification.

Clostridial species	Proteolytic C. botulinum group I	Non proteolytic C. botulinum group II	C. botulinum group III	C. argenitense (group IV)	C. butiricum	C. baratii
Type	**A; B; F; H**	**B; E; F**	**C; D**	**G**	**E**	**F**
Subtype	A1; A2; A3; A4; A5; A6; A7; A8; A9; A10; B1; B2; B3; B5 (Ba); B6; B7; A(B); Ab; Af; Af84; Bf; F1; F2; F3; F4; F5	B4; E1; E2; E3; E6; E7; E8; E9; E10; E11; F6	C; D; CD; DC		E4; E5	F7

The synthesis of BoNTs starts with the expression of a single polypeptide chain of ~ 150 kDa. This precursor is cleaved by proteases at a loop characterized by a disulphide bond generating a light and a heavy chain (L chain ~ 50 kDa; H chain ~ 100 kDa) still linked together by the disulphide bond and a peptide belt with non-covalent interaction [33] [34] [35]. Furthermore BoNT consists of a total of four domains, each one owning a peculiar

function in the mechanism of toxicity: L chain encodes for the zinc metalloprotease that specifically cleaves SNARE (soluble N-ethyl-maleimide-sensitive factor attachment protein receptor) proteins that are necessary for neurotransmitter exocytosis; the HN domain (the N terminus of the H chain) is required for translocation of the L chain across the membrane of endocytic vesicles into the neuronal cytosol; and the HC domain (the C terminus of the H chain) is responsible for presynaptic binding and endocytosis and consists of two sub domains that have different folding and binding properties.

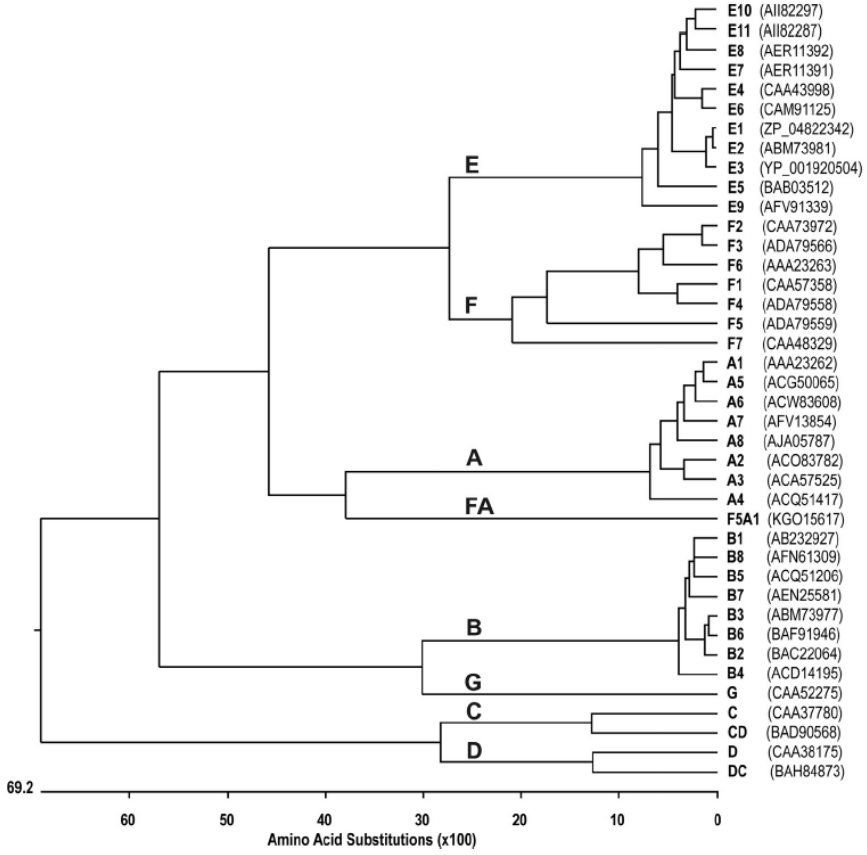


Figure 4: Phylogenetic tree of BoNT sero and subtypes. 40 BoNT variants are clustered into eight clades largely corresponding to the seven serotypes (Toxicon 107 (2015) 9e24)**.**

Genetically, *bont* gene is located next to a non-toxic gene: nonhaemagglutinin (*ntnha*) as shown in Figure 6. *ntnha* encodes a multi domain protein, NTNHA, that, once expressed, forms a heterodimer with BoNT [36].

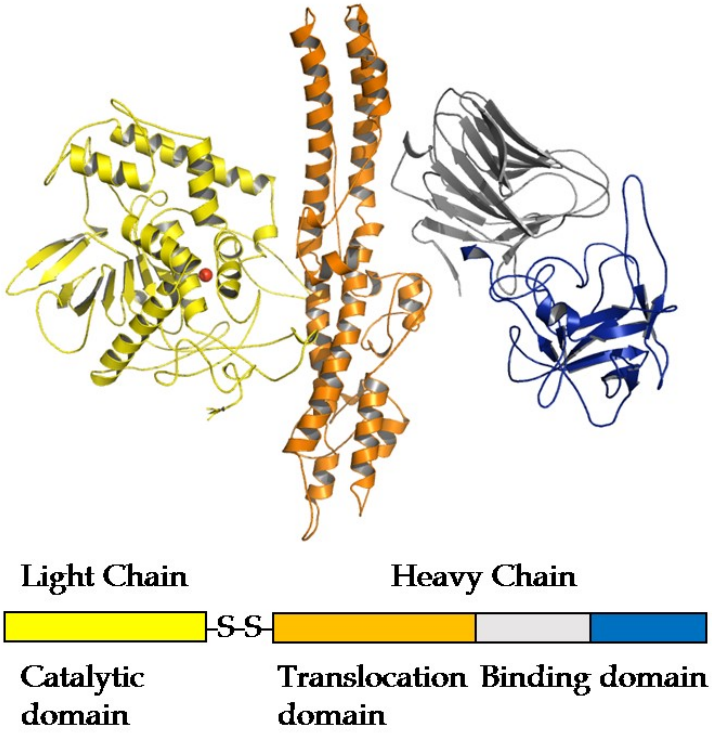


Figure 5: Botulinum neurotoxin functional organization.

BoNT and NTNHA show complementary motifs that favor multiple protein-protein contacts, suggesting that, from an evolutionary point of view, *ntnha* and *bont* might be the result of a duplication event. The result of such structural relationship is a stabilization of BoNT structure that is preserved from pH denaturation, protease activity and protein modifying agents, by NTNHA, acting as a protective scaffold to allow BoNT to survive in the habitat of release [29] [30] [36]. The *bont* and *ntnha* genes lie in close proximity to genes that encode either haemagglutinin or OrfX proteins, thought to have a protective role too. *orfX* and haemagglutinin operon are present in specific subtypes (*orfX* BoNT subtypes A1, A2, A3, A4, E1–E11 and F1–F6; the operon in strains that produce BoNT/A1, A5, B1–B7, C, D and G9). The protein products of the haemagglutinin operon (HA17, HA33 and

HA70) interact with the NTNHA–BoNT/A heterodimer, generating large oligomers that are known as progenitor toxin complexes (PTCs) [37] [38].

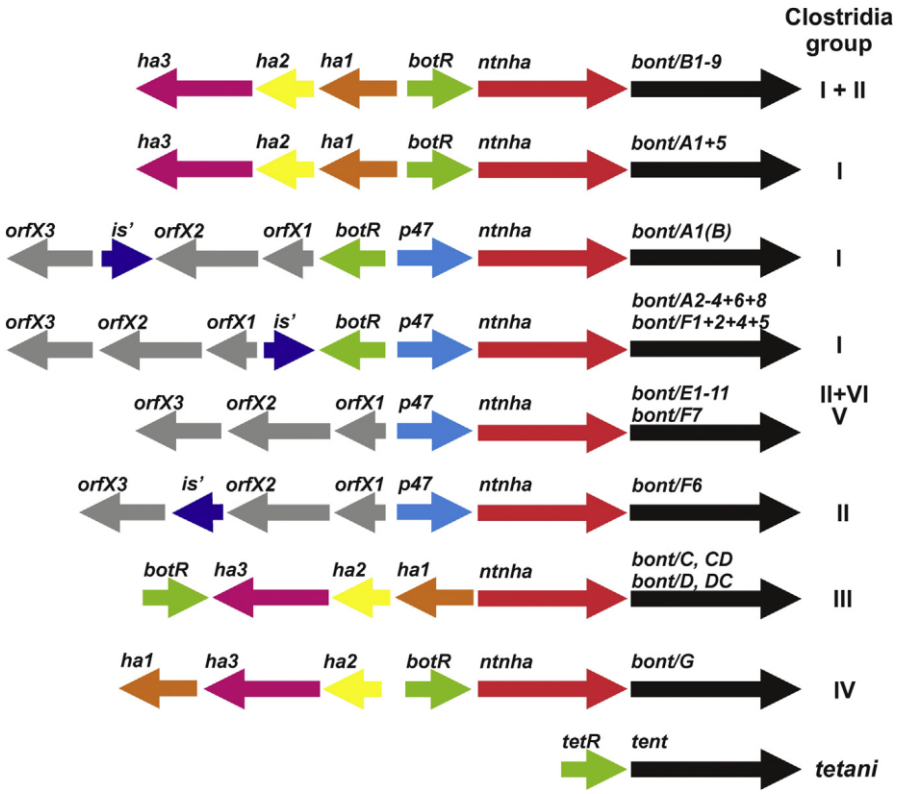


Figure 6: *bont* toxin gene cluster. *ntnha* and *bont* are encoded contiguously, allowing a fast NTNHA-BoNT interaction after expression (Infection, Genetics and Evolution 30 (2015) 102–113).

The corresponding PTCs of BoNT/B and BoNT/E have also been structurally characterized [39]. The haemagglutinin proteins of PTCs provide nine potential carbohydrate-binding sites [38], whose suggested main role of the haemagglutinin complex is to facilitate trans-epithelial absorption of the toxin [40] [41] [42]. Thus, it is possible that haemagglutinins function as adhesins and attach to the mucus layer, epithelial cells or other cells in the intestinal layer, such as M cells and neuroendocrine crypt cells [43]. A complex entry route has been suggested, in which the PTC is proposed to cross the epithelial barrier, followed by its release on the basolateral side. The haemagglutinin complex then dissociates from PTC and disrupts the epithelial barrier by loosening E-cadherin-mediated cell–cell adhesion, which opens the paracellular route to the toxin [42].

1.3.1 Mechanism of action

BoNT peculiar evolution of target selectivity brought a unique binding mode based, not only on low affinity but selective interactions, but mostly on two different receptors, polysialoganglioside (PSG) and a protein receptor in the lumen of synaptic vesicles [44] [45] [46] [47] [48] [49]. This unique binding mode encompasses two phases, an initial binding to a presynaptic receptor and a functional binding.

Initial binding. PSG is the first presynaptic BoNT receptor [50] [51]. On the presynaptic membrane, PSGs reach high densities and are organized in micro domains including glycoproteins, and their oligosaccharide portion (BoNT-binding moiety). In addition, PSGs can generate very specific interactions with target proteins and influence trans membrane signalling, endocytosis and vesicle trafficking making PSGs role as pivotal for BoNT to pass the cellular membrane [52] [53]. PSGs negative charge and BoNT dipole nature assure a rapid binding that morphs as this complex approaches the membrane, accelerating BoNT-membrane interaction [10] [54]. There are two PSG-binding sites is located on the surface of the carboxy-terminal subdomain of the HC-C domain one specific for BoNT/A, BoNT/B, BoNT/E, BoNT/F and BoNT/G serotypes and one for BoNT/C, BoNT/DC and BoNT/D (Figure 7). For the second group a secondary PSG-binding site, involving the binding to carbohydrate receptors and neurons in culture is described [55] [56] [20] [57] [58].

Functional binding. Following attachment to PSG, BoNT/B1, BoNT/DC and BoNT/G bind to segment 40–60 of the synaptic vesicle luminal domain of synaptotagmin (Syt) via a binding site in the HC-C domain that is close to the PSG-binding site [59] [60] [61] [62]. By contrast, BoNT/A1 and BoNT/E1 bind specifically to two different segments of the fourth luminal loop of the synaptic vesicle transmembrane protein SV2 [63] [64] [65] [66]. Although isoform SV2C seems to be the main receptor that is involved in BoNT/A1 binding in vitro, via an interaction with the N-terminal and C- terminal subdomains of the HC domain, both SV2A and SV2B can also mediate BoNT/A1 entry, and all three isoforms are expressed on motor nerve terminals [63] [64]. Glycosylated residues are

present in the toxin-binding site of SV2 [67] and are potentially clinically relevant, but this requires further investigation. In fact, a different pattern of glycosylation among individuals would provide a simple explanation for the variable sensitivity of different patients to BoNT/A1 injection, which is often observed in clinical settings.

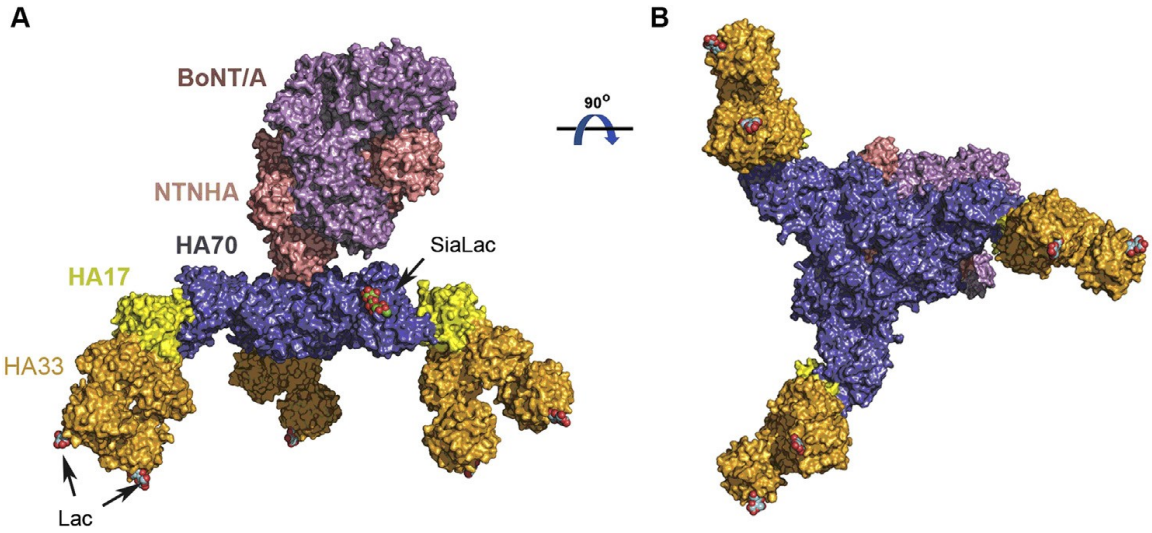


Figure 7: Molecular architecture of L-PTC of BoNT/A (L-PTC/A). A) L-PTC/A surface representation (Lee et al., 2013). BoNT/A and NTNHA-A (Gu et al., 2012) sits on top of the HA complex. The nLoop of NTNHA-A mediates the interaction with the centre of the symmetric HA70/A trimer. HA17/A mediates contact between D3 domain of HA70/A and two HA33/A molecules. Sialyllactose (SiaLac) bound to HA70/A and lactose (Lac) complexed to HA33/A (shown as sphere models) indicate the in summary nine carbohydrate binding sites mediating efficient intestinal adsorption of L-PTC/A; B) A 90 rotation of L-PTC/A about a horizontal axis (Taken from SI of (Lee et al., 2014b)).

Clearly, this variability might also be applicable to different vertebrate species. Syt and SV2 are integral proteins of the synaptic vesicle membrane and expose their BoNT-binding sites to the synaptic vesicle lumen. Therefore, unlike PSG, these protein receptors are not exposed on the nerve terminal surface and are not accessible to BoNT. However, they become available following the fusion of the synaptic vesicle with the presynaptic membrane, which exposes the synaptic vesicle lumen to the extracellular environment. Accordingly, BoNT binding to protein receptors occurs only after fusion of the synaptic vesicle to the presynaptic membrane, and this seems to facilitate the subsequent step of intoxication, which requires the endocytosis of BoNT. Nonetheless, it is possible that some

Syt molecules might be present on the presynaptic membrane following complete merging of the synaptic vesicle with the plasma membrane [68]. The protein receptors of other BoNTs have not been characterized in comparable detail so far, and conflicting results have been reported, which indicates that further characterization is needed. *Entrance into nerve terminals*. The second step of nerve terminal intoxication involves BoNT internalization (Figure 8). The complexation PSG - synaptic vesicle receptors (Syt or SV2, is serotype specific and increases the BoNT-membrane strength interactions [44]. Specifically, to BoNT/A1 it rapidly enters the synaptic vesicle lumen [69] [68]. The number of toxin molecules (one or two) correlates with the number of SV2 molecules in the synaptic vesicle membrane [70].

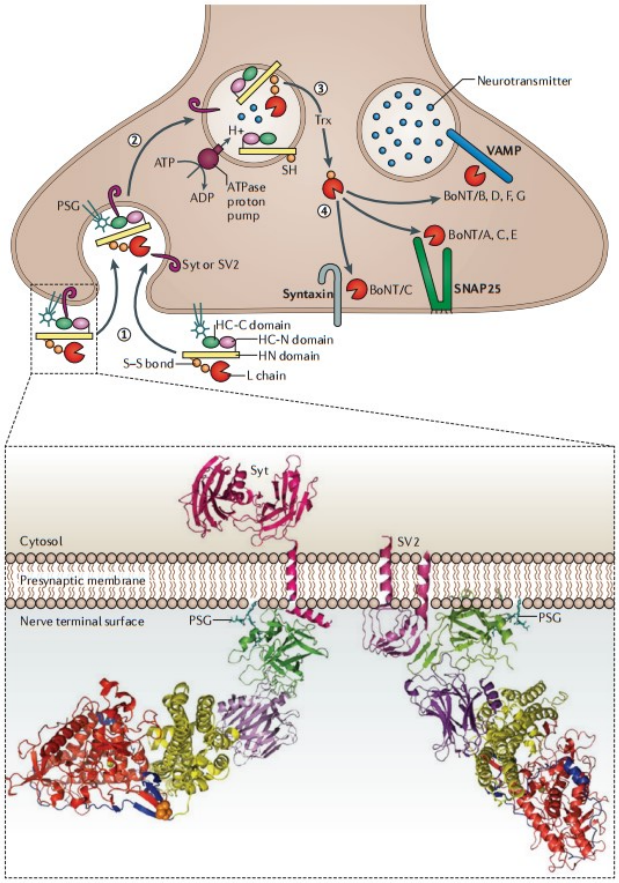


Figure 8: Binding and trafficking of botulinum neurotoxins inside nerve terminals (Nature Reviews Microbiology vol. 12, 535–549 (2014)).

BoNT/A1 rate of entry correlates with the rate of synaptic vesicle endocytosis and with the rate of paralysis of the mouse phrenic nerve hemi diaphragm [71], which is the standard NMJ that is used to test the potency of BoNTs [72] [73]. The mechanism of internalization of other BoNTs remains to be established. By contrast, in cultured CNS neurons, other vesicles and trafficking routes might contribute to entry, particularly at the very high toxin concentrations that are frequently used in the laboratory [74].

Release into the cytosol. In order to reach their target SNARE proteins in the cytosol of nerve cells, the catalytically active L chain must be translocated from the synaptic vesicle lumen into the cytosol. The main driving force for L-chain translocation is the transmembrane pH gradient that is generated by the vesicular ATPase proton pump, which drives the re-entry of neurotransmitter into the synaptic vesicle (along with H+ ions) after exocytosis74 (Figure 8). Thus, neurotoxigenic BoNTs evolved to exploit two major physiological events that occur at nerve terminals: synaptic vesicle endocytosis (to enter nerve terminals) and neurotransmitter refilling of the synaptic vesicle (to deliver the L chain metalloprotease into the cytosol). The molecular aspects of BoNT translocation across the synaptic vesicle membrane into the cytosol have been only partially elucidated, but studies that have been carried out in the past decade have provided considerable insights and have led to the proposal of a molecular model for this process. Translocation across the synaptic vesicle membrane. It has long been known that BoNTs form ion channels of low conductance in planar lipid bilayers at low pH, and this process is associated with translocation of the L chain and the cleavage of its target SNARE proteins [75]. Experimental approaches mimicking *in vivo* conditions suggest a direct influence of pH in this step of intoxication. Lowering the pH at the cis side of the membrane (facing synaptic vesicle lumen) induces the L chains of BoNTs to cross the membrane through a 15–20 Å wide channel [76] (this channel dimensions enable the passage of α-helices, suggesting that the L chain must unfold to pass through the channel). Stabilization of the L chain tertiary structure with antibodies prevents channel formation [77]. This is supported by the fact that cargo molecules, (unfolding at low pH), are transported into the neuronal cytosol when attached to the BoNT N terminus [78]. Further studies have suggested that the HN

domain alone is sufficient to form the transmembrane channel and that the peptide belt that links the L chain and the H chain regulates the formation of the HN channel [79] [80] [81] [82]. Residues that are present in all three BoNT domains are responsible for the pH sensitivity of translocation [83] [84]. The release of the L chain on the *trans* side (cytosolic side) of the membrane requires the inter-chain disulphide bond to be reduced: BoNTs that have a reduced inter-chain disulphide bond do not form channels [85]. On the basis of these data, a model for translocation has been proposed (Figure 9). This model posits that the low pH of the synaptic vesicle lumen induces a conformational change in the HN domain, which then inserts into the membrane and forms a translocation channel that chaperones the passage of the partially unfolded L chain from the luminal side to the cytosolic side of the synaptic vesicle membrane. The L chain remains attached to the synaptic vesicle until the inter-chain disulphide bond is reduced, which occurs at the end of this process. A revised model for BoNT translocation, that requires further experimental studies, propones an initial BoNT - PSG and SV2 or Syt binding inside the synaptic vesicle lumen, which has a neutral pH, immediately after endocytosis (Figure 9). The vesicular ATPase then pumps protons into the synaptic vesicle and the luminal pH becomes progressively more acidic. Notably, protons and other cations are attracted to the anionic membrane surface of the synaptic vesicle and their local concentration reduces the pH near the membrane to 1–1.5 units below that of the lumen (Figure 9). The amino acids histidine, glutamate and aspartate become protonated within the pH range (4.5–6) and are predicted to be involved in L-chain translocation. However, the actual pKa values of these residues depend on their molecular surroundings. BoNTs lack conserved histidine residues, except those that are in the active site, but they do contain conserved carboxylate residues that are predicted to have high pKa values [86] [87] [88]. Assuming that the residues that are important for the low pH-driven process are conserved, seven conserved carboxylates that have high pKa values are located in the HN domain, three are located in the L chain and one is located in the HC domain (Figure 9). The spatial distribution of these residues reinforces the suggestion that BoNTs contain more than a single pH sensor. The model posits that these carboxylates become partially or entirely

protonated in a sequential pKa depending manner as the pH of the synaptic vesicle lumen decreases (Figure 9). Even a partially protonated BoNT has a net positive charge that favors its interaction with the anionic membrane surface [89] [90] [91]. The BoNT surface that is involved in membrane interactions is suggested to be the surface that contains the inter-chain disulphide bond and the membrane-inserting segment (residues 637–688) (Figure 9); the opposite side of the BoNT molecule lacks carboxylates of appropriate pKa values. The predicted collapse of BoNT onto the membrane surface is not prevented by receptor interactions, as either binding is weakened by the low pH or the two receptors are flexible. BoNT is suggested to undergo a gross structural change that involves both the L chain and the HN domain and is facilitated by simultaneous changes in the conformation and organization of membrane lipids (Figure 9).

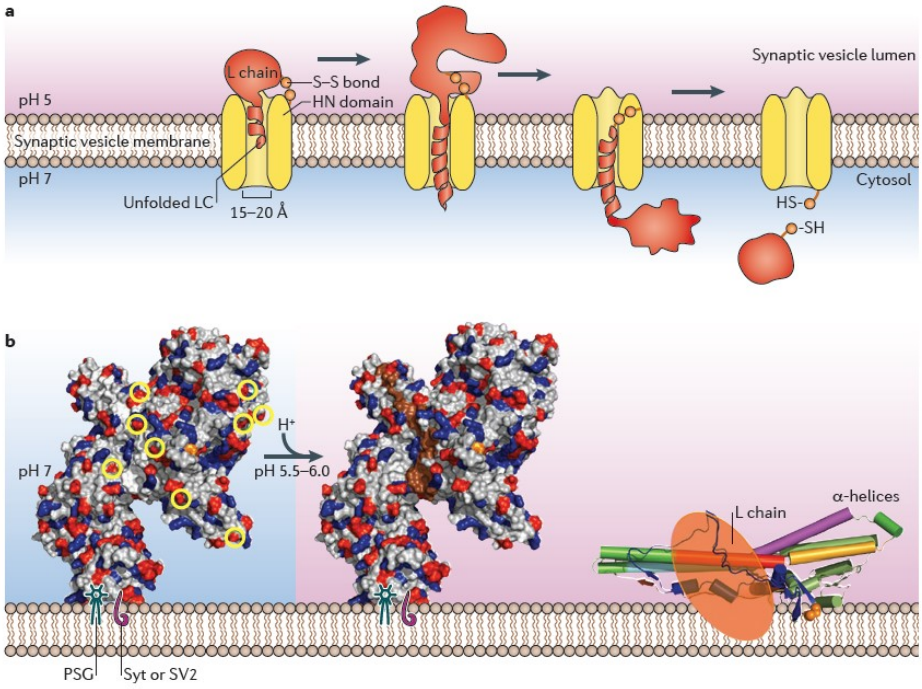


Figure 9: model for translocation (Nature Reviews Microbiology vol. 12, 535–549 (2014)).

Such changes are caused by the acidic pH of the lumen, but other factors that might contribute include ionic strength, the high Ca^{2+} concentration and the high negative curvature of the luminal synaptic vesicle membrane. The ensuing molecular events are currently unknown, but, on the basis of previous studies, it is suggested [92], that the L

chain becomes a "molten globule", which is a protein variant that retains native secondary structure but has increased hydrophobicity, to enable membrane insertion [93] [94] [95] [96]. The α-helices of the HN domain contain amphipathic segments and residues that have a low propensity to form a helical structure, which suggests that the long α-helices of the HN domain might break into shorter protein segments that insert into the membrane and thereby form an ion channel. However, whether this actually occurs is currently unknown, and clearly, more studies are needed to clarify this essential step of the BoNT intoxication process. Importantly, the reduction of the inter-chain disulphide bond at any stage before its exposure to the cytosol prevents L-chain translocation, so this domain must emerge on the cytosolic side before reduction takes place. The reduction of protein disulphide bonds is catalyzed in the cell cytosol by different enzymatic systems, including glutaredoxins, thioredoxins and other systems. Using a discriminating pharmacological approach, the redox system NADPH–thioredoxin reductase (TrxR)–thioredoxin (Trx) was found to have a major role in release of the L chain into the neuronal cytosol [97]. Following Trx-mediated reduction of the disulphide bond, L-chain translocation is irreversible and the toxin is now free to interact with its target proteins. The Trx tertiary fold is similar to that of ancestral chaperonins, so it is also possible that Trx functions as a chaperonin for L-chain translocation. Mechanism of BoNT-induced neuroparalysis The L chains of all known BoNTs are metalloproteases that are specific for one of the SNARE proteins: VAMP (vesicle-associated membrane protein; also known as synaptobrevin), SNAP25 (synaptosomal-associated protein of 25 kDa) or syntaxin. BoNT/C cleaves both SNAP25 and syntaxin, BoNT/B, BoNT/D, BoNT/F and BoNT/G only target VAMP and BoNT/A and BoNT/E cleave SNAP25. The fact that inactivation of any one of these three proteins inhibits neurotransmitter release is the strongest evidence that these three proteins form the core of the neuroexocytosis nanomachine. The SNARE family of proteins includes many isoforms of VAMP, SNAP25 and syntaxin, which are differentially expressed in many non-neuronal cells and tissues. Although several of these isoforms can be cleaved by BoNTs, these substrates are not accessible in vivo, as non-neuronal cells lack appropriate receptors for the toxin. The molecular basis

of the neuroparalytic activity of BoNTs has recently been reviewed in depth [98] [24]. With the exception of BoNT/A and BoNT/C, all BoNTs cleave isolated SNARE proteins by removing large cytosolic segments, which prevents the formation of the SNARE complex. BoNT/A and BoNT/C remove only a few residues from the C terminus of SNAP25, and this truncated form of SNAP25 can form a stable SNARE complex; thus, the molecular mechanism of BoNT/A- and BoNT/C induced neuroparalysis remains to be elucidated. It is possible that the core of the nanomachine is comprised of a SNARE supercomplex that is formed by several SNARE complexes and that the C terminus of SNAP25 is involved in protein–protein interactions among the individual SNARE complexes [99] [100] [101]. An alternative explanation is that BoNT/A cleaves another protein target essential for neurotransmitter release. However, such protein substrates have not yet been found, despite extensive searches, and they are unlikely to exist, owing to the unique mode of recognition of VAMP, SNAP25 and syntaxin by the L-chain metalloprotease. In fact, the SNARE-binding site of the metalloprotease is a long channel that is occupied by the peptide belt in the intact protein; however, when the L chain is released, this channel is vacated and the substrate can then insert into the channel. The L chain interacts extensively with the substrate and contacts several exosites of the protein in addition to the cleavage site.

1.4 Botulinum neurotoxin as biological weapon

The research for the use of BoNT as biological weapon began at the time of World War II**.** The countries that developed research programs to use BoNT as weapon were United States, Japan and, perhaps, Germany. The Japanese biological warfare group (Unit 731) performed experiments on humans, feeding cultures of *C. botulinum* to prisoners with lethal effect. In United States, the research on the weaponisation of BoNT continued after the war till 1970, when was ended by executive orders of the president Richard Nixon. Also Soviet Union developed similar researches, and BoNT was one of several agents tested in the research center Aralk-7, on Vozrozhdeniye Island in the Aral Sea. In the 1972, Biological and Toxin Weapons Convention prohibited research and production of

biological weapons. Four of the countries considered by US governments as "state sponsors of terrorism" (Iran, Iraq, North Korea and Syria) are believed to have produced biological weapon based on BoNT. In 1991, in the presence of the United Nations inspectors, the production and the loading of BoNT into military weapons (missiles and bombs) were admitted by Iraq. The use of weaponised BoNT was tried by terrorist Japanese sect "Aum Shinrikyo": aerosols were dispersed at multiple sites in Tokyo on at least three occasions between 1990 and 1995. These attacks failed, probably for faults in technical procedures [3].

CHAPTER 2
AIMS

Botulism is a medical emergency because of the severity of the provoked disease characterized by a very high mortality rate and the lack of adequate specific therapeutic facilities. For these reasons it is necessary to identify rapidly the source suspected to limit further exposure caused or natural. In the case of epidemics or deliberate release of botulinum toxins, [32] [102] their traceability with canonical tools is difficult and less discriminating than the genotyping of organisms that produced them. From the design and validation of a Next Generation Sequencing pipeline, the aims of this project of research focus to characterize both genetically and virtually *C. Botulinum* Group I strains. Secondly we will extend the neurotoxin experimental model dataset yet published with the creation of new virtual neurotoxin models obtained by means of homology modelling procedures. Eventually, we will evaluate the interaction of the whole set of the experimentally and virtually obtained neurotoxin models with a set of peptide inhibitors with *in silico* techniques. The 3D models will be though available for further studies such as virtual protein-protein interaction to improve the understanding of the different BoNTs toxigenic mechanism.

CHAPTER 3
MATERIALS AND METHODS

3.1 Bacterial strains

C. botulinum strains were provided by the National Reference Center for Botulism, Istituto Superiore di Sanità (ISS, Rome, Italy). This institution maintains a strains collection mainly isolated during the last three decades, from clinical cases occurred in various Italian regions. The strains analyzed in this study are reported in Table 7. All strains belong to group I, as verified by rRNA 16S gene sequence and biochemical assays.

3.2 Genomic approach

In Figure 10 is graphically reported the sequencing pipeline followed to obtain the genomic sequence of the *C. botulinum strains*. The sequencing was accomplished by means of two Next Generation Sequencing platforms (NGS): illumina MiSeq and Roche 454. From the bacteria specimens' DNA is extracted and prepared for being sequenced. DNA for genome sequencing was extracted from cellular lysate using the phenol-chloroform method [103].

Bioinformatics for assembly differ upon illumina or Roche sequencing technology:

a) illumina MiSeq. The obtained fastq files, sequencing procedure summarized in paragraph 3.2.1, were analyzed with fastQC[2], a quality control tool for high throughput

[2] **FastQC:** Modern high throughput sequencers can generate hundreds of millions of sequences in a single run. Before analysing this sequence to draw biological conclusions it is compulsory to perform quality control checks to ensure that the raw data are reliable before using them. Most sequencers generate a QC report as part of their analysis pipeline, but this is usually only focused on identifying problems which were generated by the sequencer itself. FastQC aims to provide a QC report which can spot problems originating either in the sequencer or in the starting library material. FastQC can be run in one of two modes: run as a standalone interactive application for the immediate analysis of small numbers of FastQ files, or run in a non-interactive mode where it would be suitable for integrating into a larger analysis pipeline for the systematic processing of large numbers of files.

sequence data [104] (in Figure 35 quality plots of the experimentally produced reads, in Figure 36 FastQC sequenced reads quality assessment). Poorly assessed reads were tuned by means of Sickle[3] [105]. Reads were than assembled into contigs using ABySS, a de novo, parallel, paired-end sequence assembler that is designed for short reads[4] [106].

The yet contigs were than mapped in a draft genome using Mauve[5], a system for constructing multiple genome alignments in the presence of large-scale evolutionary events such as rearrangement and inversion, in Figure 11 is shown a snapshot of a multiple alignment session [107]. Multiple genome alignments provide a basis for research into comparative genomics and the study of genome-wide evolutionary dynamics. If the DNA strain had been sequenced using Roche 454 platform too, the two obtained draft genomes were merged in order to resolve any gaps. The remaining gaps, originating mainly from the unresolved repeated regions, were left open. In order to avoid homopolymer length errors, a known bug that characterizes 454 technology, it has been necessary to compare 454 sequences to Illumina [108]. An *ad hoc* script developed in our laboratory based upon Figure 12 algorithm (APPENDIX C).

---

3 **Sickle:** Tool that uses sliding windows along with quality and length thresholds to determine when quality is sufficiently low to trim the 3'-end of reads and also determines when the quality is sufficiently high enough to trim the 5'-end of reads. It also discards reads based upon the length threshold. It takes the quality values and slides a window across them whose length is 0.1 times the length of the read. If this length is less than 1, then the window is set to be equal to the length of the read. Otherwise, the window slides along the quality values until the average quality in the window rises above the threshold, at which point the algorithm determines where within the window the rise occurs and cuts the read and quality there for the 5'-end cut. Then when the average quality in the window drops below the threshold, the algorithm determines where in the window the drop occurs and cuts both the read and quality strings there for the 3'-end cut.

4 **AbySS**: ABySS (Assembly By Short Sequences), a parallelized sequence assembler, was developed for performing an efficient assemble data from large-scale sequencing projects. The ABySS algorithm proceeds in two stages: 1) All possible substrings of length k (termed k-mers) are generated from the sequence reads; k-mer data set is then processed to remove read errors and initial contigs are built; 2) Mate-pair information is used to extend contigs by resolving ambiguities in contig overlaps.

5 **Mauve**: It is a genome comparison method that identifies conserved genomic regions, rearrangements and inversions in conserved regions, and the exact sequence breakpoints of such rearrangements across multiple genomes. Furthermore, it performs a traditional multiple alignment of conserved regions to identify nucleotide substitutions and small insertions and deletions (indels). By integrating previously separate analysis steps, Mauve provides additional ease-of-use and sensitivity over other systems when comparing genomes with significant rearrangements

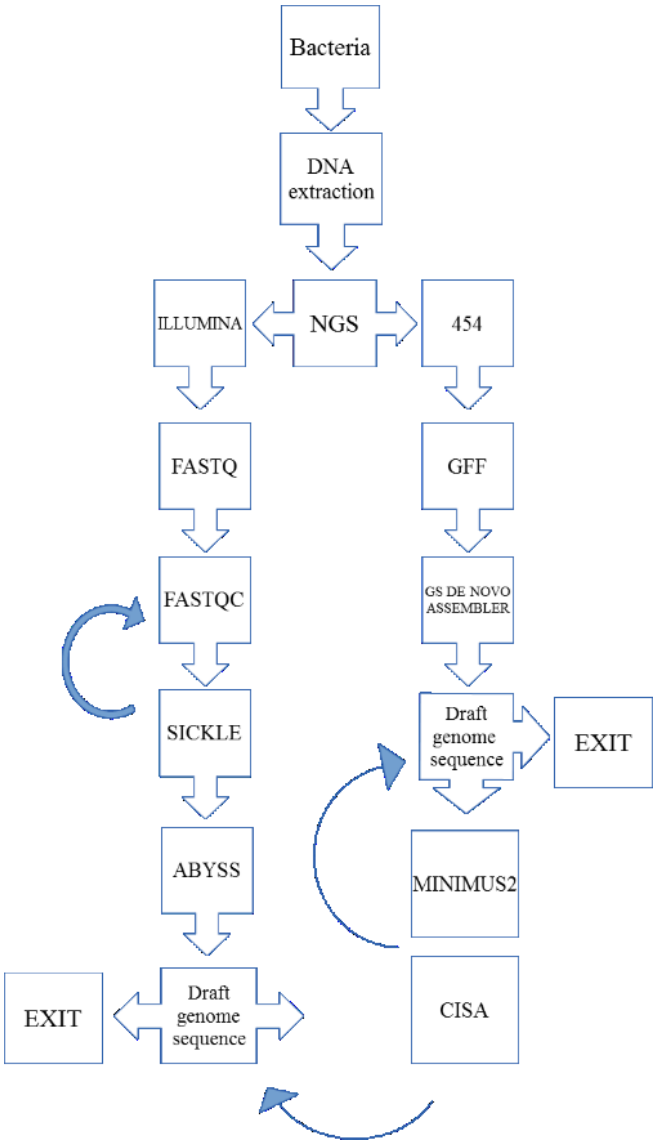


Figure 10: Adopted sequencing pipeline.

b) *Roche 454*. The obtained GFF files, sequencing procedure summarized in paragraph 3.2.2, are processed by GS de novo assembler, part of the Roche Newbler[6] software package, proprietary software which performs assembly of reads and generates contigs. [109]. The assembly was performed with the default parameter values of 90% for the

[6] **Newbler** is a software package for *de novo* DNA sequence assembly. It is designed specifically for assembling sequence data generated by the 454 GS-series of pyrosequencing 454 Life Sciences platforms.

minimum overlap identity and 40bp for the minimum overlap length. The contigs were then aligned with the reference genome A1 ATCC 3502 [110] deposited in GenBank [111] and reordered using the software Mauve.

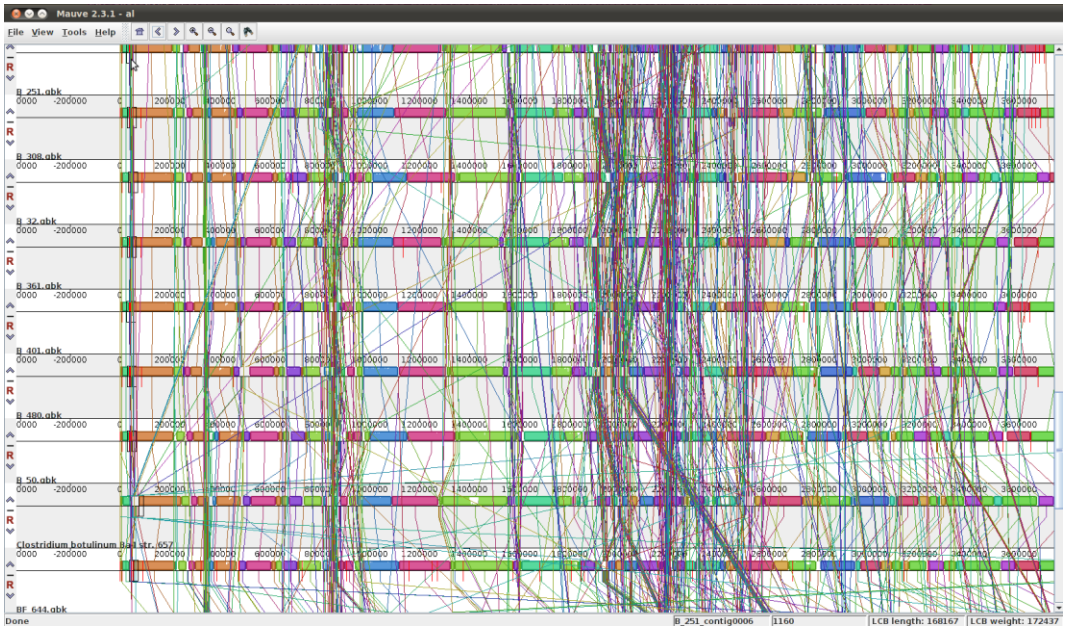


Figure 11: *C. botulinum* mapping through Mauve graphycal ionterface.

Then misassemblies constituted by collapsed repeats were detected using the GS assembler output files and BLAST searches [112]. The applied criteria were abnormally high coverage, correlated mismatches in the reads alignment, the presence of distinct groups of reads mapping to one end of the same contig that do not overlap in the part exceeding the end of the contig, and connect it to two different other contigs. The Illumina sequences were used to cover, when possible, the gaps between the 454 contigs. The remaining gaps between the contigs were covered with Sanger sequencing [113] [114]. Regard the gaps due to repeated sequences, in some cases the closure was possible resolving the different copies, when the differences were sufficiently frequent and separated by a distance in bps lower than the average reads length. In some cases, the reads alignment relative to a collapsed repeats contig did not show any mismatches correlated across overlapping reads, and it was assumed that the copies are identical, justifying the closure of different gaps with the same sequence. In all the other cases, the

gaps were left open, and the contigs containing the collapsed repeats were excluded from the draft assembly.

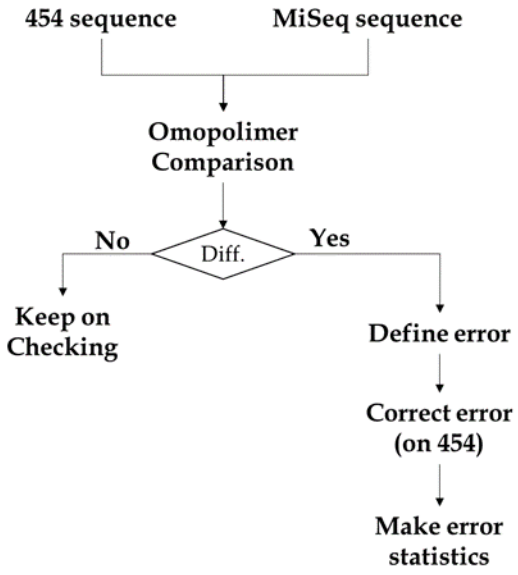


Figure 12: Roche 454 homopolymer correction algorithm.

Once obtained a satisfactory assembly, the genome annotation was automatically generated by the NCBI PGAP pipeline [115].

3.2.1 Illumina MiSeq

Illumina sequencing technology relies on sequencing by synthesis (SBS) methodology. This platform is characterized by fluorescently labeled reversible terminators imaged as each dNTP is added, and then cleaved to allow incorporation of the next base. Since all 4 reversible terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias. The method virtually eliminates errors and missed calls associated with strings of repeated nucleotides (homopolymers).

Illumina methodology can be presented as a three steps procedure: amplification, sequencing, and analysis. Purified DNA is fragmented into smaller sequences and specific adapters, indices, and other kinds of molecular modifications that act as reference points during amplification, sequencing, and analysis (shown in Figure 13 A and B)Figure 13:

illumina MiSeq platform workflow.. The amplification and sequencing take place into a chip where hundreds of thousands of oligonucleotides are anchored in order to be complementary with the DNA sequence terminals. Once the DNA fragments and the anchored oligonucleotides are complementary attached, the cluster generation begins. During this step each fragment of DNA is replicated. Next, primers and modified nucleotides enter the chip (shown in Figure 13 C).

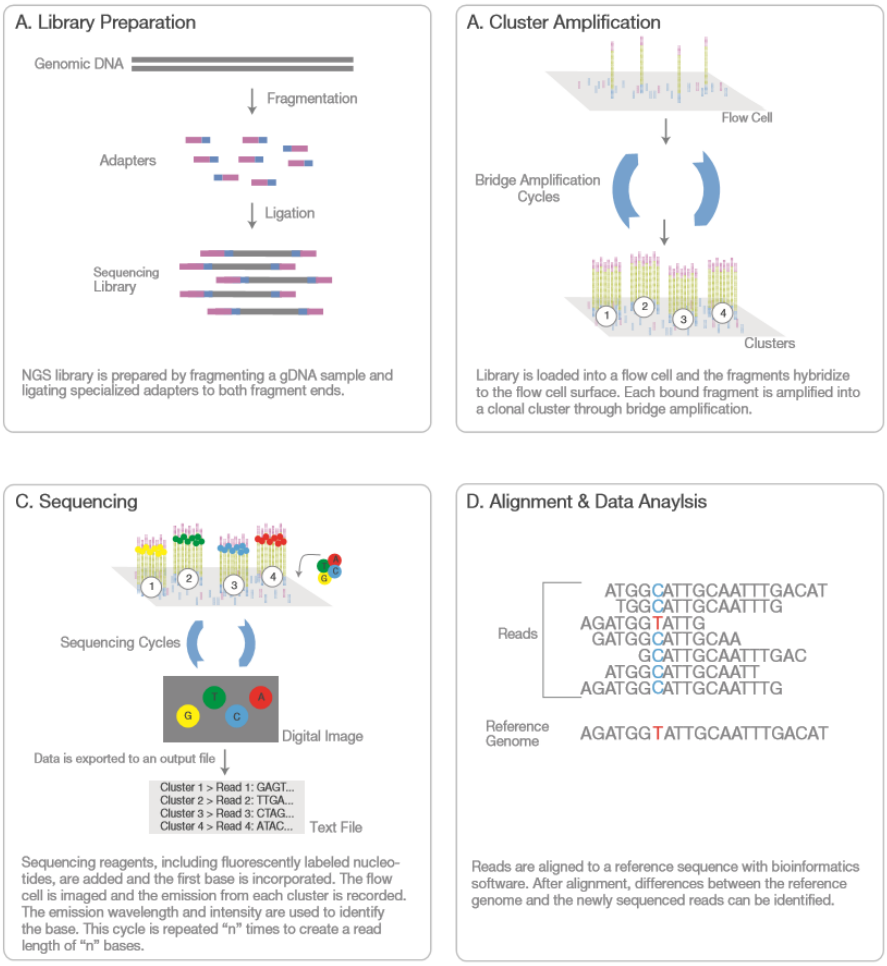


Figure 13: illumina MiSeq platform workflow. (http://www.illumina.com/systems/miseq.html)

These nucleotides have reversible 3' blockers that force the primers to add on only one nucleotide at a time as well as fluorescent tags. After each round of synthesis, a camera takes a picture of the chip. The fluorescence produced in the elongation is characterized by a specific wavelength and is recorded and acquired by the platform. A

chemical step is then used in the removal of the 3′ terminal blocking group and the dye in a single step. The process continues until the full DNA molecule is sequenced [116]. With this technology, thousands of places throughout the genome are sequenced at once via massive parallel sequencing. The third step is the production of the *fastq* file reporting both the sequence and the relative quality of each nucleotide (shown in Figure 13 D).

3.2.2 Roche 454

454 Sequencing uses a large-scale parallel pyrosequencing system capable of sequencing ~ 600 DNA megabases per run on the Genome Sequencer FLX with GS FLX Titanium series reagents (Figure 14) [117]. The platform fixes nebulized and adapter-ligated DNA fragments to small beads in a water-in-oil emulsion. The DNA beads is amplified by PCR. Each DNA-bead is placed into a micro well on a fiber optic chip. A mix of enzymes (DNA polymerase, ATP sulfurylase, and luciferase) is added into the well. The chip is finally introduced GS FLX sequencer. 454 is capable to sequence up to 600 million base pairs per run with ~ 500 base pair read lengths. The sequencing procedure imply two main stages: DNA library preparation and emPCR and Sequencing.

DNA library preparation and emPCR. Genomic DNA is fractionated into smaller fragments (~ 600 base pairs) and polished. Short adaptors are ligated onto the ends of the fragments to provide priming sequences for both amplification and sequencing of the sample-library fragments. One adaptor (Adaptor B) contains a 5'-biotin tag for immobilization of the DNA library onto streptavidin-coated beads. After nick repair, the non-biotinylated strand is released and used as a single-stranded template DNA (sstDNA) library. The sstDNA library is assessed for its quality and the optimal amount (DNA copies per bead) needed for emPCR is determined by titration [118]. The sstDNA library is immobilized onto beads containing a library consistent of single sstDNA molecule. The bead-bound library is emulsified with the amplification reagents in a water-in-oil mixture. Each bead is captured within its own micro reactor where PCR amplification occurs. This results in bead-immobilized, clonally amplified DNA fragments.

Sequencing. Single-stranded template DNA library beads are added to the DNA Bead DNA polymerase Incubation Mix and layered with sulfurylase and luciferase enriched onto the chip. Beads are fixed to the device vessels by centrifugation. The layer of Enzyme Beads ensures that the DNA beads remain positioned in the wells during the sequencing reaction. The bead-deposition process is designed to maximize the number of wells that contain a single amplified library bead. The yet prepared chip is inserted in the FLX sequencer. The fluidics sequencing buffers and nucleotides are added across the wells of the plate. The four DNA nucleotides are added sequentially in a fixed order across the chip device during a sequencing run.

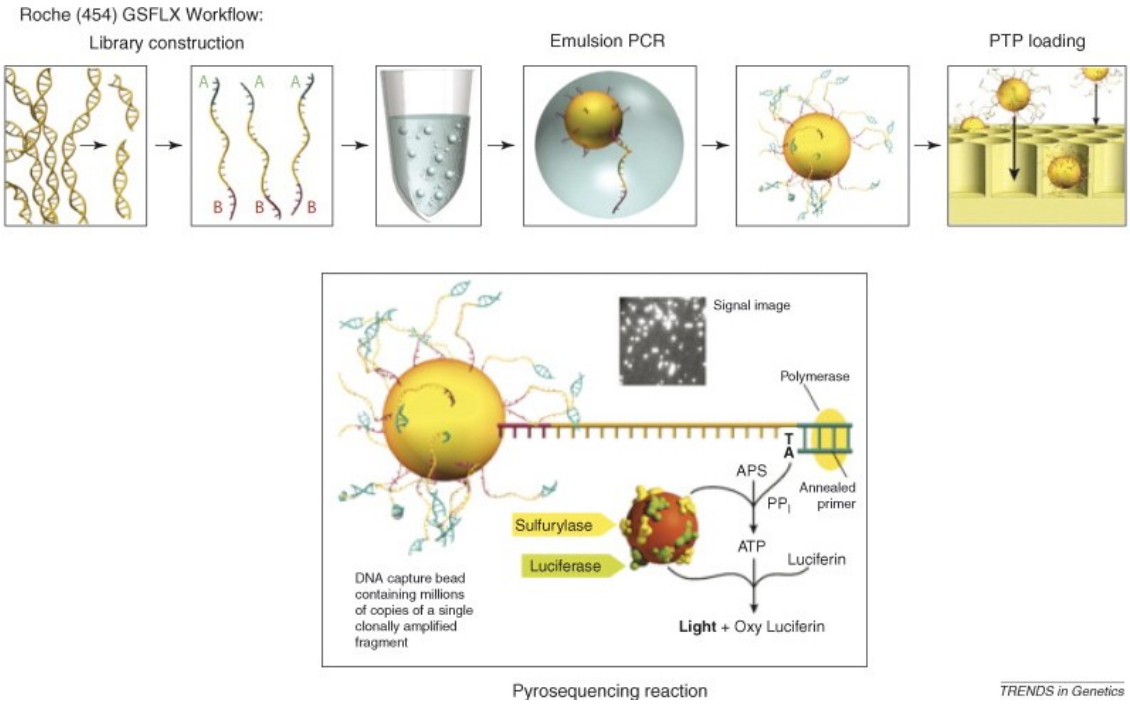


Figure 14: Roche 454 workflow scheme. (http://454.com/products/gs-flx-system/index.asp)

During the nucleotide flow, millions of copies of DNA bound to each of the beads are sequenced in parallel. When a nucleotide complementary to the template strand is added into a well, the polymerase extends the existing DNA strand by adding nucleotide(s). Addition of one (or more) nucleotide(s) generates a light signal that is recorded by the CCD camera in the instrument. This technique is based on sequencing-by-synthesis and is called pyrosequencing [119]. The signal strength is proportional to the

number of nucleotides. However, the signal strength for homopolymer stretches is linear only up to eight consecutive nucleotides after which the signal falls-off rapidly [120]. Data are stored in standard flowgram format (SFF) files for downstream analysis.

3.2.3 Whole genome phylogenetic analysis

A whole genome alignment of the 10 Italian draft assemblies (Table 7) and 10 group I genomes available in GenBank was generated with MAUVE (A1 ATCC3502 - NC_009495, A1 ATCC19397 - NC_009697, A1 Hall - NC_009698, A2 Kyoto - NC_012563, A3 Loch Maree - NC_010520, Ba4 strain 657 - NC_012658, B1 Okra - NC_010516, F Langeland - NC_009699, A5 H04402 065 - FR773526 and F strain 230613 - CP002011). From the whole genome alignment, using the StripSubsetLCB script (distributed with MAUVE), a file containing the core genome blocks longer than 500 bps was extracted, and then analyzed by the software ClonalFrame[7] [121]. Three independent runs of ClonalFrame were performed, each consisting of 10,000 burn-in MCMC (Markov Chain Monte Carlo) iterations and 10,000 collected MCMC iterations. The *bont* carrying plasmid sequences were analyzed independently with ClonalFrame following the same pipeline applied to the chromosome. Plasmid analysis included the *bont*-gene-carrying plasmids found in the new sequenced genomes and the plasmids NC_012654, NC_010379 and NC_010418, belonging to the genomes Ba4 strain 657, B1 Okra and A3 Loch Maree, respectively. Three independent runs for the plasmid consisted of 100,000 burn-in MCMC iterations and 100,000 after burn-in period MCMC iterations. In both cases, the Gelman-Rubin test [122] showed a satisfactory degree of convergence (chromosome: 1.00 for the parameter ϑ, 1.01 for R, 1.01 for ν, 1.00 for δ; plasmid: 1.00 for ϑ, 1.09 for R, 1.02 for ν, 1.03 for δ). The chromosomal core genome regions were analyzed with ClonalOrigin, following the procedure previously described [123]. The number of burn-in and post

[7] **ClonalFrame**: this software estimates the clonal relationships between the members of a sample, while also estimating the chromosomal position of homologous recombination events that have disrupted the clonal inheritance.

burn-in iterations was set for the two ClonalOrigin runs according the software's standard conditions. The average value of three parameters, ϑ (mutation rate), ρ (recombination rate) and δ (average length of recombination), determined in the first run, were: $\delta = 395$, $\vartheta = 0.0311$ and $\rho = 0.0233$. The results of the second ClonalOrigin run were elaborated by mean of an ad hoc script reported in APPENDIX C to determine the posterior distribution of the number of recombination events. Phylogenetic comparisons of 150 core-genome genes (Table 16) and *bont* sequences were performed with multiple alignments and dendrogram calculation using UPGMA algorithm, applying the BioNumerics software (BioNumerics created by Applied Maths NV. Available from http://www.applied-maths.com.). The presence/absence of 40 genes was studied by BLAST searches within the 20 analyzed genomes (Table 17). The results were imported in BioNumerics as binary character values and a dendrogram was constructed with DICE binary coefficient and UPGMA algorithm. The identity of the unique genes characterizing the newly sequenced genomes was investigated with BLAST searches on the GenBank database, using blastn and blastp programs. Manual recombination analysis was performed using the SimPlot[8] software [124].

3.3 *In silico* structural approach

In silico structural approach was designed in order to virtually design the 3D structures encoded in the *bont* genes here sequenced. After nucleotide – amino acidic sequence translation, the BoNT structures were built with homology modelling procedures. After manual curation and quality assessment the zinc coordination function of all models was designed. Finally, to better correlate the serotype-BoNT activity relationship, we extracted the catalytic domain to test the inhibitory potential of a set of known peptide inhibitors with docking calculations. Eventually a qualitative analysis of the obtained results was performed.

[8] **SimPlot**: calculates and plots the percent identity of the query sequence against a panel of reference sequences in a sliding window, which is moved across the alignment in steps.

3.3.1 *Sequence Data.*

The ten *Clostridium botulinum* strains used in this study for homology modelling procedures belong to ISS collection of Botulinum samples isolated in Italy as presented in Paragraph 3.1 (references reported in Table 7) and sequenced with the above mentioned pipeline (Paragraph 3.2). The Next generation sequences obtained were translated into the amino acid sequence of the ten putative protein primary structures by means of EXPASY translation utility [125].

3.3.2 *Molecular visualization and editing.*

PyMOL 1.7 [126] was used for molecular visualization. *PyMod* [127] and *autodock* plugins[9] [128] were embedded to PyMOL to perform respectively search of the experimental templates to include in homology modelling procedures and qualitative docking analysis.

3.3.3 *Homology modelling.*

Homology modelling of a protein structure is a reliable *in silico* technique to design a 3D virtual model of a protein structure inferred from the amino acid sequence [129] [130] [131]. The accuracy and quality of the obtained models is strictly dependent on the dissimilarity between the target and the template sequence. Moreover, the more populated is the set of templates; the more robust will be the model. Nowadays high-sensitivity tools are publicly available and offer built-in capacities to provide an almost *prêt-à-porter* model and complete quality estimate of the predicted structure. In this work, we followed a three steps PyMOD pipeline. From each protein sequence, we performed a preliminary database search with psi-BLAST algorithm [112] [132]. Secondly, for every

[9] **Autodock/Vina plugin for PyMOL:** This plugin is a PyMol implementation written in order to ease the portability of the autodock input preparation and analysis such: Defining binding sites and export to Autodock and VINA input files, Doing receptor and ligand preparation automatically, Starting docking runs with Autodock or VINA from within the plugin, Viewing grid maps generated by autogrid in PyMOL, Handling multiple ligands and set up virtual screenings and Set up docking runs with flexible sidechains

experimental template, we performed protein structure alignment by incremental combinatorial extension (CE) [133] using PyMODs' embedded routine. After structural templates alignment, we input the neurotoxin sequences templates and queries to the ClustalW [134] sequence alignment program including all sequences in the workspace, using default parameters. In Figure 15 is reported a sequence alignment as schemed with MODELLER9.15 plotting function. We kept the previously obtained structural alignment, since we wanted to keep the structural alignment *in-frame* (i.e., adding indels, when necessary, to both templates).

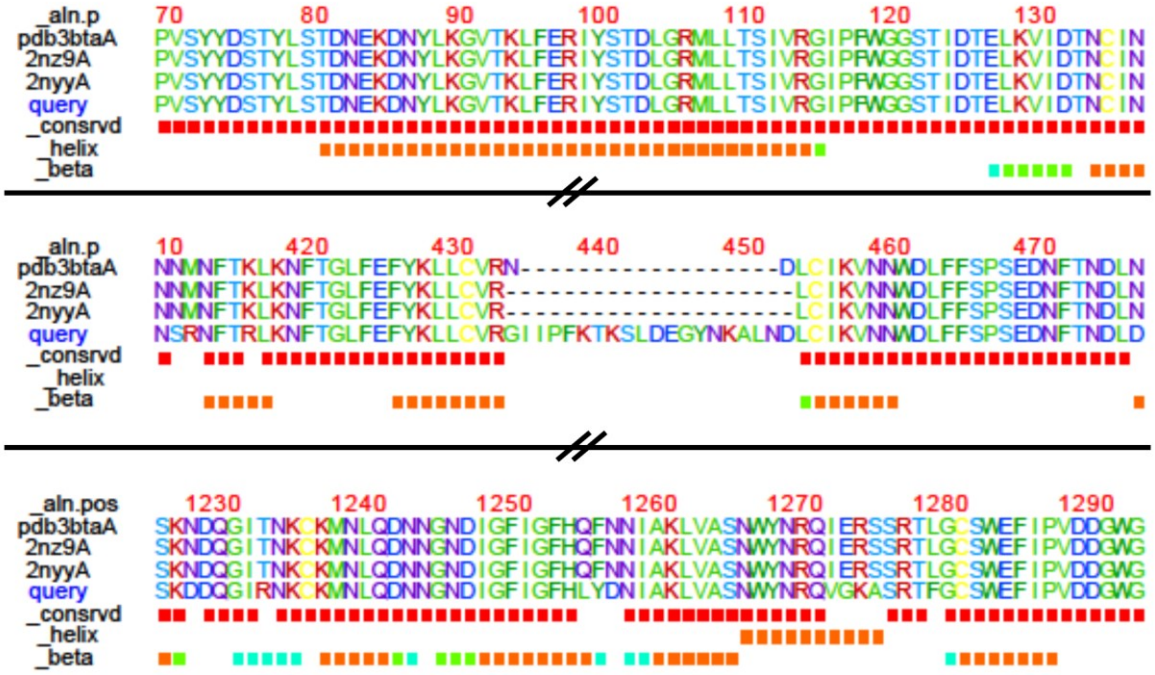


Figure 15: amino acidic sequence alignment. This snapshot represents the sequence alignment produced for *bont* sequence A2 117 (query) and the most similar experimentally resolved templates (pdb structures 3bta, 2nz9a and 2nyy).

Once we obtained the structural and sequence alignments merged, we manually checked the alignment to pinpoint potential misaligned regions.

MODELLER9 .15 We used MODELLER9.15 utilities to refine the row preliminary structures [135] [136] [137] [138]. The reliability of this homology-modelling platform is provided by the multidimensional set of built-in spatial restraints, namely: homology-

derived, stereo chemical, statistical preferences for dihedral angles and non-bonded inter-atomic distances that might be optionally biased by manually curated features.

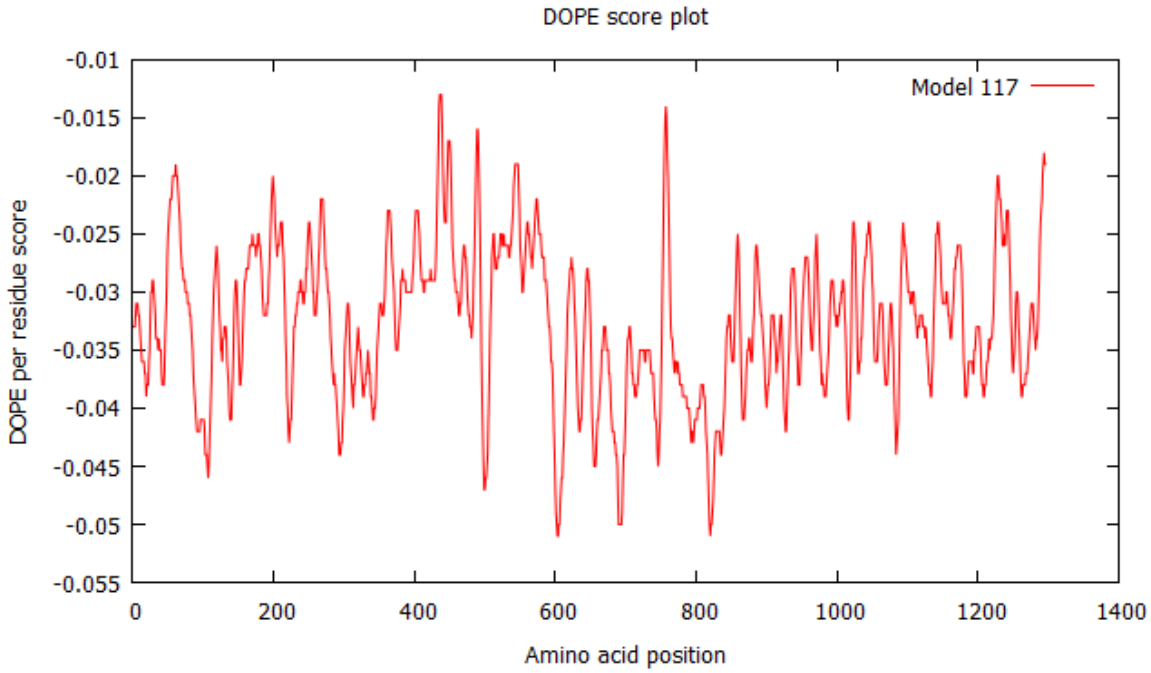


Figure 16: DOPE score plot. Discrete Optimized Protein Energy, is a statistical potential used to assess homology models in protein structure prediction.

Those restraints, presented as probability density functions, are optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing (used for loop refinement). Generally, we included in MODELLER inputs an alignment of a sequence with the identified templates and the atomic coordinates of the templates obtained through the above-described PyMOD pipeline. We identified the portions of the

amino acidic structures that needed a refinement using MODELLERs' discrete optimized protein energy (DOPE) scoring function[10].

DOPE results were plotted for a qualitative analysis with GNUplot 4.6[11] (DOPE plot in Figure 16). We generated up to five models per low-valued loops. Finally, we retained for quality assessment the model with the best DOPE value.

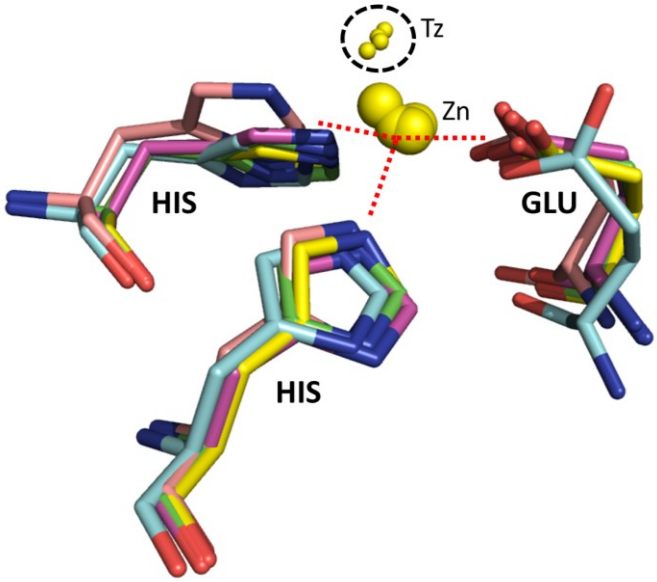


Figure 17: Superposition of the modelled zinc coordinated function. The coordinating aminoacidic side chain triade (HIS-HIS-GLY) of all modelled structures are superimposed. Zinc atom (Zn) is grafted from the experimentally resolved model template. The dummy atom (Tz) is modelled with Autodock4$_{Zn}$ scripting tools.

---

[10] **DOPE**, or Discrete Optimized Protein Energy, is a statistical potential used to assess homology models in protein structure prediction. DOPE is based on an improved reference state that corresponds to non-interacting atoms in a homogeneous sphere with the radius dependent on a sample native structure; it thus accounts for the finite and spherical shape of the native structures. It is implemented in the popular homology modeling program MODELLER and used to assess the energy of the protein model generated through many iterations by MODELLER, which produces homology models by the satisfaction of spatial restraints. The models returning the minimum molpdfs can be chosen as best probable structures and can be further used for evaluating with the DOPE score. The DOPE method is generally used to assess the quality of a structure model as a whole. Alternatively, DOPE can also generate a residue-by-residue energy profile for the input model, making it possible for the user to spot the problematic region in the structure model.

[11] **Gnuplot** is a portable command-line driven graphing utility for Linux, OS/2, MS Windows, OSX, VMS, and many other platforms.

Zinc coordinated function. For homology modelling purposes, since the experimentally resolved template structures of the zinc coordinated catalytic function were identical to the query, we grafted it the metal coordinated function to the model, Figure 17 [139].

3.3.4 Model quality evaluation.

Usually the quality of a homology model includes the evaluation of model geometry, stereochemistry and analysis of Ramachandran plots [140]. This may allow the identification of any residues with not permitted torsion angles. Procheck [141] was used for both detailed analysis of the torsion angles for all amino acids in the neurotoxins generating Ramachandran plots. QMEAN server [142] as utilized for model quality estimation. QMEAN6 scoring function [143] estimated the *degree of nativeness* of the structural features of our virtual models. More precisely, QMEAN6 scoring function is a linear combination of six structural descriptors (local geometry is analysed by a torsion angle potential over three consecutive amino acids, distance-dependent interaction potentials are used to assess long-range interactions - First, at a residue-level it is based on C_β atoms only, at the second level an all-atom potential is used -, solvation energy is calculated to investigates the burial status - accessibility to water - of the residues). Since QMEAN server quality check caps protein model structure at its 600^{th} amino acid, quality check was performed onto each models' domain, taking into account that each domain is linked together by disordered loops.

3.3.5 Experimental structure datasets.

All the deposited BoNT protein structures were retrieved (Table 3) from the Protein Data Bank (PDB) [144].

From the experimental dataset: 1) zinc-coordinated structures were retained for homology modelling and for docking calculations purposes 2) when a peptide-inhibitor ligand was present in the experimental complex structure, it was retained and enriched the selection of peptide BoNT inhibitor to be included in the docking ligand dataset

(properties of the ligands dataset reported in Table 4, ligand structures reported inTable 5).

Table 3: experimental BoNT structures.

	BoNT serotypes					
	A	**B**	**C**	**D**	**E**	**F**
pdb structures	1e1h, 2vu9, 3boo, 3c8a, 3ddb, 3k3q, 3qiz, 3qw6, 3v0a, 3zur, 4el4, 4iqp, 4ktx, 2nyy, 2vua, 3c88, 3c8b, 3dse, 3qix, 3qj0, 3qw7, 3v0b, 3zus, 4elc, 4jra, 4kuf, 2nz9, 2w2d, 3c89, 3dda, 3fuo, 3qiy, 3qw5, 3qw8, 3v0c, 4ej5, 4hev, 4ks6, 1xtf, 2g7k, 2g7p, 2ilp, 2imc, 3bon, 3bwi, 1xtg, 2g7n, 2g7q, 2imb, 3bok, 3bta, 3ds9	2etf, 2xhl, 4kbb, 1f31, 1g9a, 1g9c, 1i1e, 1s0c, 1s0e, 1s0g, 2nm1, 2np0, 3zuq, 1epw, 1f82, 1g9b, 1g9d, 1s0b, 1s0d, 1s0f, 1z0h	2qn0, 3de2, 3deb, 3n7k, 3r4s, 3r4u, 2qn0	3n7j, 3obr, 3obt, 3ogg, 3rmx, 3rmy, 2fpq	3ffz, 1t3a, 1t3c, 1zkw, 1zkx, 1zl5, 1zl6, 1zn3	2a8a, 3fie, 3fii, 2a97, 3fuq

Table 4: Description of the selected peptide inhibitors.

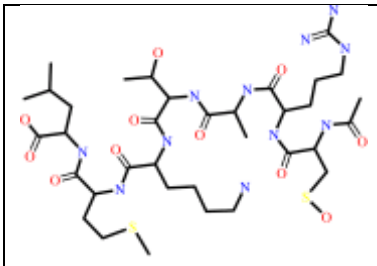
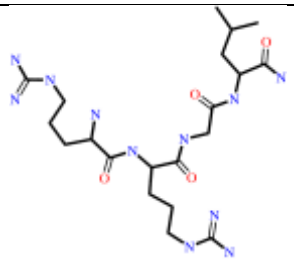
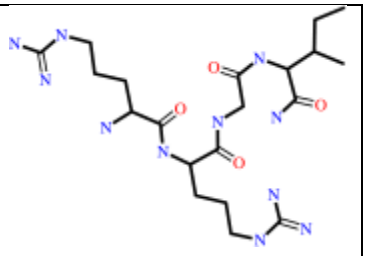
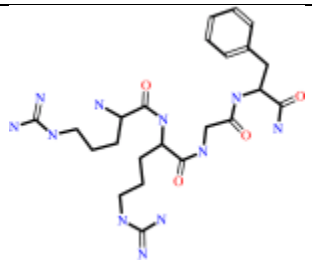
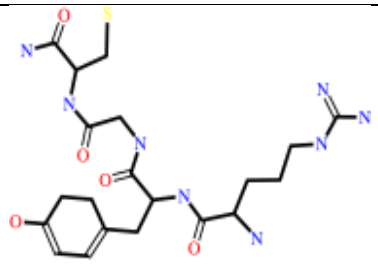
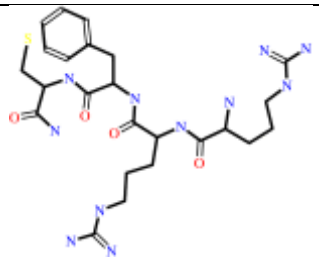
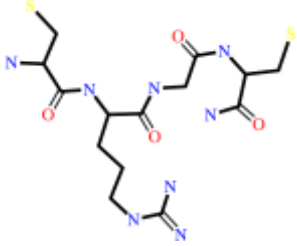
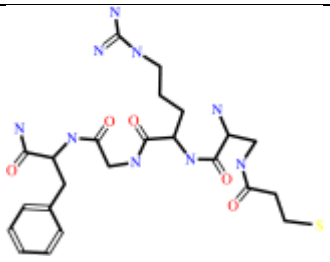
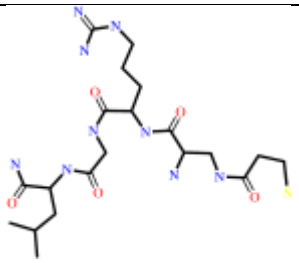
Ligand	peptidic composition	complex pdb id	BoNT serotype
lig01	ACE-CSO-ARG-ALA-THR-LYS-MET-LEU	3boo	A
lig02	ARG-ARG-GLY-LEU-NH2	3c8a	A
lig03	ARG-ARG-GLY-ILE-NH2	3c8b	A
lig04	ARG-ARG-GLY-PHE-NH2	3qw5	A
lig05	ARG-TYR-GLY-CYS-NH2	3qw6	A
lig06	ARG-ARG-PHE-CYS-NH2	3qw7	A
lig07	CYS-ARG-GLY-CYS-NH2	3qw8	A
lig08	MPT-DPP-DAR-GLY-DPN-NH2	4ks6	A
lig09	MPT-DPP-ARG-GLY-LEU-NH2	4ktx	A
lig10	DCY-ASP-ARG-GLU-LEU-VAL-LYS-NH2	3fie	F

3.3.6 Docking calculations.

Ligands preparation.

From BoNT-peptide inhibitor complex structures, ligand coordinates were retained and processed with AutoDockTools v.1.5.651 script *prepare_ligand4.py* [145] with default settings was used to add Gasteiger-Marsili charges [146] merge non-polar hydrogens, and assign atom types. All torsions were allowed to the ligands. All ligands were extracted from their original pdb complex structure with the exception of *lig10*.

Table 5: Scheme of the selected peptide inhibitors as summarized in Table 4.

 3boo ligand - Lig01	 3c8a ligand - Lig02	 3c8b ligand - Lig03
 3qw5 ligand - Lig04	 3qw6 ligand - Lig05	 3qw7 ligand - Lig06
 3qw8 ligand - Lig07	 3ks6 ligand - Lig08	 3ktx ligand - Lig09

Lig 10 preparation.

Pdb structure 3fie is a BoNT/F zinc catalytic domain complexed with INH01, a 12 peptide inhibitor [147]. INH01 interaction with 3fie BoNT/F zinc catalytic domain is very peculiar: the inhibitor interacts with its target inside the catalytic pocket coordinating the zinc atom function and hugs the receptor in a horseshoe fashion. Since we are interested in the pockets' dynamics in order to characterize the different BoNT subtypes on peptide ligand capability base, we cut the zinc coordinated peptide inhibitor head of INH01 at its 9th amino acid from obtaining lig10, then prepared as above described for docking calculations (lig10 shown in Figure 18).

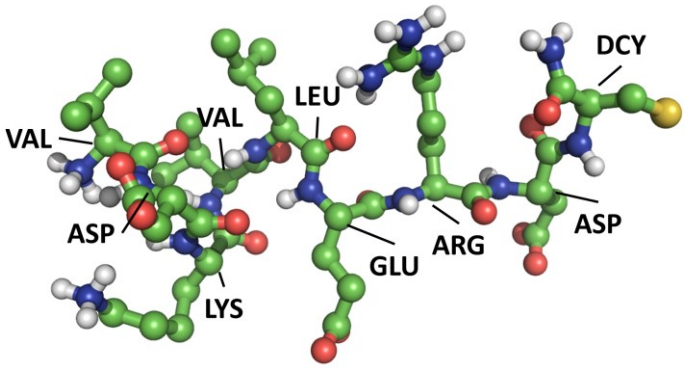


Figure 18: lig10 sticks and balls structure representation as prepared for docking calculations. Colour code: green carbon atoms, red oxygen atoms, blue nitrogen atoms, gold sulphur atom, white hydrogen atoms (after ligand preparation for docking through Autodock script prepare_ligand4.py options -A, -U).

Receptor dataset.

All BoNT experimental structures containing the zinc coordinated function deposited in the PDB were downloaded and inserted in the receptor dataset (Table 3). All the unique designed virtual structures were also added to the receptor dataset. Only the clostridial neurotoxin zinc protease domain was used for docking calculations.

Receptor preparation.

All receptor structures were pre-processed following the standard AutoDockTools script prepare_receptor4.py with default settings to remove all water molecules, to assign Gasteiger-Marsili charges, merge non-polar hydrogens and assign atom types. For dockings with AutoDock4$_{Zn}$ [148], the charge of the zinc ions was left to the default value (0.0) and TZ pseudo atoms added with the custom AutoDock4$_{Zn}$ Python script zinc_pseudo.py.

Table 6: flexible doking parameters.

pdbID	peptidic inhibitor	flexible residues	rigid residues	zinc coordinates
3boo	lig01	VAL70; ILE161; SER254; LEU256	GLU262 HIS223 HIS227	13.718; -5.630; 9.160
3c8a	lig02	GLU164; PHE369; ASP370	GLU262 HIS223 HIS227	27.198; 28.354; 55.0120
3c8b	lig03	ILE161; GLU164; PHE369; ASP370	GLU262 HIS223 HIS227	26.981; 28.367; 55.062
3qw5	lig04	GLU164; PHE194; PHE369; ASP370	HIS227 HIS223 GLU262	27.160; 28.353; 54.748
3qw6	lig05	GLU164; PHE194; PHE369; ASP370	HIS227 HIS223 GLU262	27.162; 28.416; 54.965
3qw7	lig06	GLU164; LEU256; PHE369; ASP370	HIS227 HIS223 GLU262	26.687; 28.665; 54.958
3qw8	lig07	GLU164; PHE369; ASP370	HIS227 HIS223 GLU262	27.069; 28.490; 54.881
4ks6	lig08	PHE163; GLU164; CYS165; PHE194; PHE369; ASP370	HIS227 HIS223 GLU262	7.427; 5.493; 35.088
4ktx	lig09	PHE163; GLU164; CYS165; PHE369; ASP370	HIS227 HIS223 GLU262	7.428; 5.403; 34.987
3fie	lig10	SER166; SER167; TYR168	HIS227 HIS231 GLU266	18.381; -0.474; -33.075

While rigid docking calculations run for all ligand-receptor couples, flexible chain docking was run for re-docking and virtual model-ligand couples (flexible docking features reported in Table 6).

Grid parameters.

The docking grid box was defined to include the catalytic pocket completely. *Autogrid4* generated receptor grids applying Autodock4$_{ZN}$ atom-types parameters in a 126

points width tri-dimensional grid spacing 0.2 Å. Grid input file included the AD4Zn.dat specific parameter file for zinc atom.

Docking calculations.

Autodock4 program was used for docking purposes. The *Autodock4*$_{Zn}$ scoring function, specifically parameterized for zinc metal-proteins was applied to all calculations for both rigid and flexible docking. Specific docking parameters were: Lamarckian genetic algorithm (LGA), with a maximum of 25000000 energy evaluations run for at most 270000 generations, population of 150 individuals, mutation rate 0.02 and crossover rate is 0.8 [149]. Docking outputs were used to rank the lowest energy BoNT-inhibitor complex. The experimental BoNT/inhibitor complexes were used as template to assess a topological qualitative "*consensus*" between the docked ligand poses in order to compare the zinc-ligand coordination within the ligand set.

CHAPTER 4
RESULTS AND DISCUSSION – GENOMIC CHARACTERIZATION

4.1 Whole genome sequencing

In order to provide a comprehensive pattern of the genetic variability of Italian C. botulinum group I population, 10 strains were selected for sequencing from the strains collection maintained at ISS (Istituto Superiore di Sanità), as representatives of different clusters obtained by MLVA-15. The complete list of the new sequenced genomes, with information related to the serotype, botulism form, geographical origin and the isolation year, is reported in (Table 7). All the new sequenced strains were isolated in Italy from clinical cases notified during the last three decades. The strain A2 117 was isolated from a large outbreak occurred in 1996, due to the consumption of contaminated Italian mascarpone cheese [150]. The clinical cases from which A2B3 87 and B2 450 were isolated were described previously [151] [152].

The 10 Italian strains were sequenced by means of two Next Generation Sequencing (NGS) platforms: 454GS FLX+ and Illumina MiSeq (Table 7). 454 sequencing coverage ranges from 15X to 69X (27X in average), while Illumina sequencing coverage from 47X to 234X (130X in average). After reads assembly, the number of resulting contigs varies from 10 to 15.

The total length of the 10 draft assemblies ranges between 3,797,055 bps and 4,070,655 bps (excluding plasmids). Overall, the total length of gaps has been estimated, with the reference sequence A1 ATCC 3502, between 50 Kb and 75 Kb (1.3 – 1.8 % of a C. botulinum group I average genome size 3,960,335bps). Data summarized in Table 8.

Table 7: Italian Group I Clostridium Botulinum genomes.

genome	serotype	BioProject	clinical case	geographical origin	isolation year	454 coverage	MiSeq coverage	n. contigs	n. plasmids[a]	all contigs length
267	B2	PRJNA213576	infant	Lazio	2003	69x	86x	12	2	3871108
433	B2	PRJNA213581	infant	Puglia	2009	15x	47x	12	2	3846238
128	B2	PRJNA213565	infant	Lazio	1998	20x	234x	10		3844467
117	A2	PRJNA213371	foodborne	Campania	1996	17x	71x	11		3808118
275	B2	PRJNA213573	wound	Lazio	2004	25x	92x	13		3978188
331	B2	PRJNA213567	infant	Sicilia	2004	24x	151x	10		3809031
357	F	PRJNA213614	foodborne	Trentino	2005	15x	118x	15	1	3832122
92	A2B7	PRJNA213606	foodborne	Sardegna	1993	29x	69x	14	1	3797005
87	A2B3	PRJNA213609	foodborne	Sicilia	1995	25x	246x	13	2	3847714
450	B2	PRJNA213608	wound	Sicilia	2009	28x	185x	10	1	4070655

Most of the gaps left in our sequences are due to unresolved repeats: the nine copies of the rRNA genes operon (total length of ~43 Kb), the two copies of β-N-acetylglucosamidase genes, the four copies of the beta-hydroxylase genes and other three genes with long internal repeats that were found in some genomes (a homolog of H04402_00311 CDS found in A2 117 and B2 275; a homolog of CLM_373-5 CDS found in F8 357; a homolog of CLK_3392 CDS found in A2B3 87). rRNA genes, that mainly constitute the missing sequences, are scarcely variable within the same species [153]. Therefore, the gaps should not affect phylogenetic analysis results. One or two plasmids, comprised in a single contig were found in six genomes (Table 8). B2 267 owns 2 small plasmids, both containing regions homologous to A1 ATCC 3502 plasmid pBOT3502 (NC_009496). F8 357 genome includes one small plasmid. The bont carrying plasmids of strains B2 433, A2B7 92, A2B3 87 and B2 450 show a clear homology with Ba4 strain 657, B1 Okra and A3 Loch Maree plasmids [154]. Moreover, B2 433 and A2B3 87 strains possess also a second smaller plasmid: B2 433 plasmid is homologous to the smaller B2 267 (sequence similarity 99%).

Table 8: Contigs statistics.

[1] (excluded plasmids) after 454 – illumina merging and sanger Gap closure; [2] remaining GAP due to not covered regions; [3] GAPs due to repeats.

Genomes	Roche (Newbler) contigs > 500 bps	tot. Bps contigs > 500 bps	Illumina (Abyss) contigs > 500 bps	tot. Bps contigs > 500 bps	n. contigs[1]	plasmids	remaining GAP due to not covered regions	GAPs due to repeats
B2 267	78	3898280	524	3824558	10	2	0	9
B2 433	72	4125175			10	2 (1 pl. Car.)	0	9
B2 128	67	3846884	87	3853909	10		0	9
A2 117	147	3806199	285	3798799	11		0	10
B2 331	207	3795475	86	3844089	10		0	9
A2B7 92	107	4058381	543	4003454	13	1 (pl. car.)	1	11
B2 275	137	3977447	251	3976409	13		0	12
F 357	319	3785437	146	3849714	14	1	1	12
A2B2 87	42	4175822	109	4175327	11	2 (1 pl. Car.)	0	10
B2 450	142	4303998	139	4323523	10	1 (pl. car.)	0	9

Interestingly, the A2B3 87 smaller plasmid (contig013) shows no homologies with any other sequenced *C. botulinum* plasmid. Such plasmid carries some typical phage genes, suggesting that it could be a non-integrated prophage (*e.g.*: gene at position 850-1647 bps could be a recombinase, containing the RecT recombination domain pfam03837; gene at 19642-22173 bps has the prophage endopeptidase tail domain pfam06605; gene at 29800-30822 bps owns the major capsid protein E domain pfam03864 and gene at 32745-33968 bps is characterized by the phage portal protein domain, SPP1 Gp6-like, pfam05133).

4.2 Relative location of bont and BoNT serotypes

As previously reported, the bont gene cluster was found either within the plasmid or within the chromosome [155] [154]. In strains A2B7 92, B2 433, A2B3 87, B2 450, the bont gene cluster is placed on the plasmid (Table 8), whereas, in B2 128, B2 275, B2 331, B2 267, A2 117 and F8 357 strains, it is located on the chromosome.

The genome locations of *bont* cluster in our genomes are the same already observed in the previously reported ones [156]: bont/B2 cluster of strains 128, 275, 331, 267 is located in *oppA/brnQ* operon, as bont/A1 cluster in ATCC 3502 and the bont/B cluster in A1(B) NCTC 2916 (Accession n. ABDO00000000.1) (ha cluster) [156]. In strains A2 117 and F8 357 the bont cluster is located in *arsC* operon, as in A2 Kyoto, in F Langeland and the bont/A cluster in A1(B) NCTC 2916 (*orfX* cluster) [156]. The bont/A2 cluster location in strain 117 is not completely identical to the A2 toxin gene position in Kyoto (between the two copies of *arsC* gene) but is about 10Kb downstream, corresponding to the Langeland and A1 NCTC 2916 bont cluster position [156]. These evidences suggest that bont/A2 cluster was probably inserted independently in the lineages of the two genomes. The position of bont/A and B clusters in A2B7 92, B2 433, A2B3 87 and B2 450 plasmids corresponds exactly to that observed in the plasmids of Ba4 strain 657, A3 Loch Maree and B1 Okra [156].

The BoNT subtype of the Italian genomes was determined performing blastp searches against the GenBank Protein database. The threshold to recognize different

subtypes is 2.6% difference of the amino-acid sequence [157]. With such criteria, almost all BoNTs of the new genomes can be classified in one of the previously characterized subtypes. BoNT of strain 357 BoNTs, which shows the highest amino-acid similarity with F1 Langeland (96%), can be proposed as a new F subtype (F8). Further studies are required to better characterize the biochemical properties of this new subtype.

4.3 Clonal phylogeny

Clonal phylogeny is constituted by the phylogenetic relations exclusively determined by the vertical heredity, the transmission of the genetic material from mother cell to daughter cell [158] [159]. Clonal phylogeny of the 20 genomes was calculated by ClonalFrame on the core-genome, the DNA regions shared among all the analyzed strains [158] [160]. In Figure 19, the dendrogram representing the clonal phylogeny is shown.

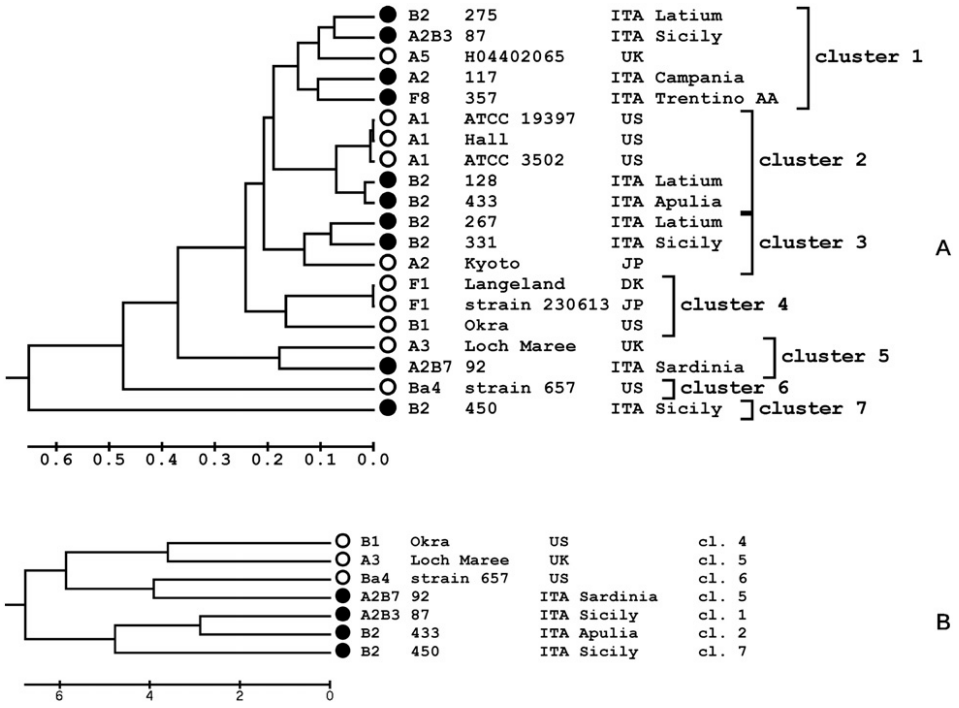


Figure 19: Clonal dendrograms. A) 20 C. botulinum strains - analysis of chromosomal sequences. B) Clonal dendrogram of the 7 *bont* carrying plasmids. Dendrograms produced with ClonalFrame runs on the core genome. X axis units = coalescent units (a unit of time normalized by population size - Degnan and Rosenberg, 2009).

Seven clusters were defined: cluster 1 contains the strains B2 275, A2B3 87, A5 H04402065, A2 117 and F8 357; cluster 2 includes A1 strains (ATCC 3502, ATCC 19697 and Hall) together with some B2 (433 and 128); in cluster 3, B2 267, B2 331 and A2 Kyoto are found; cluster 4 contains two very similar BoNT/F genomes (F1 Langeland and F1 strain 230613) and B1 Okra; cluster 5 groups Loch Maree along with A2B7 92; cluster 6 and 7 contain only one representative, Ba4 strain 657 and B2 450 respectively.

As illustrated in the clonal dendrogram, three branches, containing respectively B2 450, A3 Loch Maree and Ba4 strain 657, are fairly distant from all the other lineages. The remaining four lineages, characterized by the presence of Okra, Kyoto, ATCC 3502 and H04402065 respectively, are separated by a low genetic distance. A similar genealogy was obtained also in a study based on 25,555 SNPs (Single Nucleotide Polymorphisms) analysis comparing 17 C. botulinum group I genomes, where the groups containing A3 Loch Maree and Ba4 strain 657 are markedly distant from all the other groups [161].

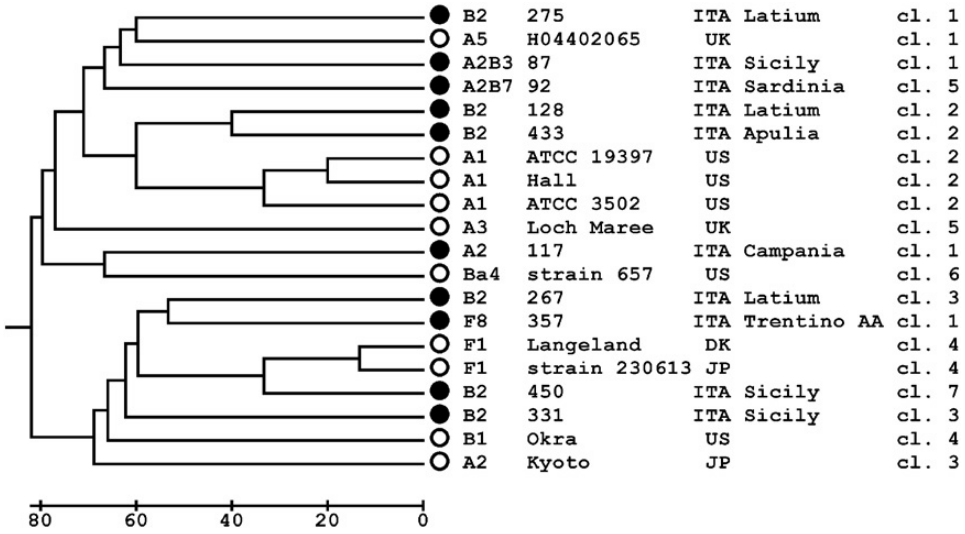


Figure 20: Phylogenetic relations of the 20 analysed genomes based on MLVA data. X axis units = % dissimilarity.

Interestingly, Italian strains do not form a monophyletic unit, but they are scattered in different lineages along with strains originating from different countries and even from different continents. This suggests a high phylogenetic diversity within the Italian C. botulinum group I population.

The topology of clonal dendrogram is quite different respect to that of MLVA dendrogram (Figure 20). On the one hand, very high mutation rate of tandem repeats makes MLVA highly discriminative for forensic purpose, while on the other hand, because of mutational saturation and frequent homoplasies, it is not very accurate to reconstruct deep phylogenetic relationships [162].

For the *bont* carrying plasmids, clonal phylogeny was evaluated separately. Core regions encompass 45,252 bps, about 1/5 of the average length of the considered plasmid. This relatively small common part is mainly due to the absence of an about 85 Kb region in the B1 *Okra* plasmid. The resulting clonal phylogeny is shown in Figure 19B. As in the chromosome clonal dendrogram (Figure 19A), the plasmids of the Italian strains are not grouped in a single clade. A2B7 92 plasmid forms a group with Ba4 strain 657 plasmid, another bivalent strain. B2 450 plasmid appears related to A2B3 87 and B3 433 plasmids, on the contrary B2 450 chromosome is highly divergent from all the other genomes, suggesting a recent acquisition of the plasmid in the B2 450 lineage.

4.4 Recombination

10 Italian strains were selected by MLVA among a large number of strains to be representative of different lineages, in order to avoid an over-representation of very similar genotypes. This is crucial for estimating a biologically meaningful recombination rate [163].

To assess the impact of recombination on genetic diversity of *C. botulinum* group I, ClonalFrame was launched to estimate the r/m ratio (r = rate of nucleotide substitution due to recombination, m = rate of nucleotide substitution due to point mutation [163]. The obtained value, 2.84, is considered a high value according to the scale [163], classifying *C. botulinum* group I as a high rate recombination species. This value confirms the findings obtained in previous studies [164] [165], in which the φ test [166] applied on MLST data showed statistically significant evidence of recombination in *C. botulinum* group I.

To analyze the recombination fluxes, the program ClonalOrigin was applied to core genome blocks longer than 500 bps. On the basis of the clonal dendrogram,

ClonalOrigin determined the branch of origin and destination and the number of the identified recombination events. The number of recombination events was calculated for the different clusters summing the recombination events of all the branches belonging to the same cluster. The results are reported in Table 3 where each cluster is characterized by a number as a donor of genetic material and a number as recipient.

Basically the recombination exchange is shown between each couple of clusters. It is probable that there is neither ecological separation between the clusters nor physiological constraints to the recombination flow. Moreover, the frequency of recombination within and between clusters is comparable.

The recombination was also checked calculating a dendrogram for each one of 150 core genes (Table 16) and observing if the obtained trees replicate the topology of the clonal dendrogram (Figure 19A). None of the 150 dendrograms has a topology equal to the clonal dendrogram. The conservation frequency of the clusters defined in clonal dendrogram is low: for example, cluster 1 is maintained in the dendrogram of 17 genes, cluster 2 is conserved in 76 genes, cluster 3 in 29, cluster 4 in 49, cluster 5 in 102. These inconsistencies may suggest a high recombination rate.

4.5 Bont phylogeny

Both chromosome and plasmid clonal dendrograms (Figure 20) do not reflect the BoNTs serotypes and subtypes based grouping. The discrepancy between BoNT type and chromosome phylogeny has already been observed in previous studies, and it has been explained supposing a frequent horizontal gene transfer of *bont* [8] [167] [168] [169].

For a more detailed characterization, a dendrogram was produced for each serotype (A, B and F), that compares the bont sequences of the 20 genomes analyzed in the present study and other bont sequences available in GenBank. In bont A dendrogram (Figure 21A), the three Italian strains are grouped with A2 Kyoto in a cluster of very similar A2 sequences, differently from chromosome clonal dendrogram (Figure 19A), where they are split in different clusters (A2 117 and A2B3 87 in cluster 1, A2 Kyoto in

cluster 3, A2B7 92 in cluster 5). A2 *bont* of strain 87 is more similar to the Kyoto one than to 117, despite the relationships of the three genomes in chromosome clonal dendrogram.

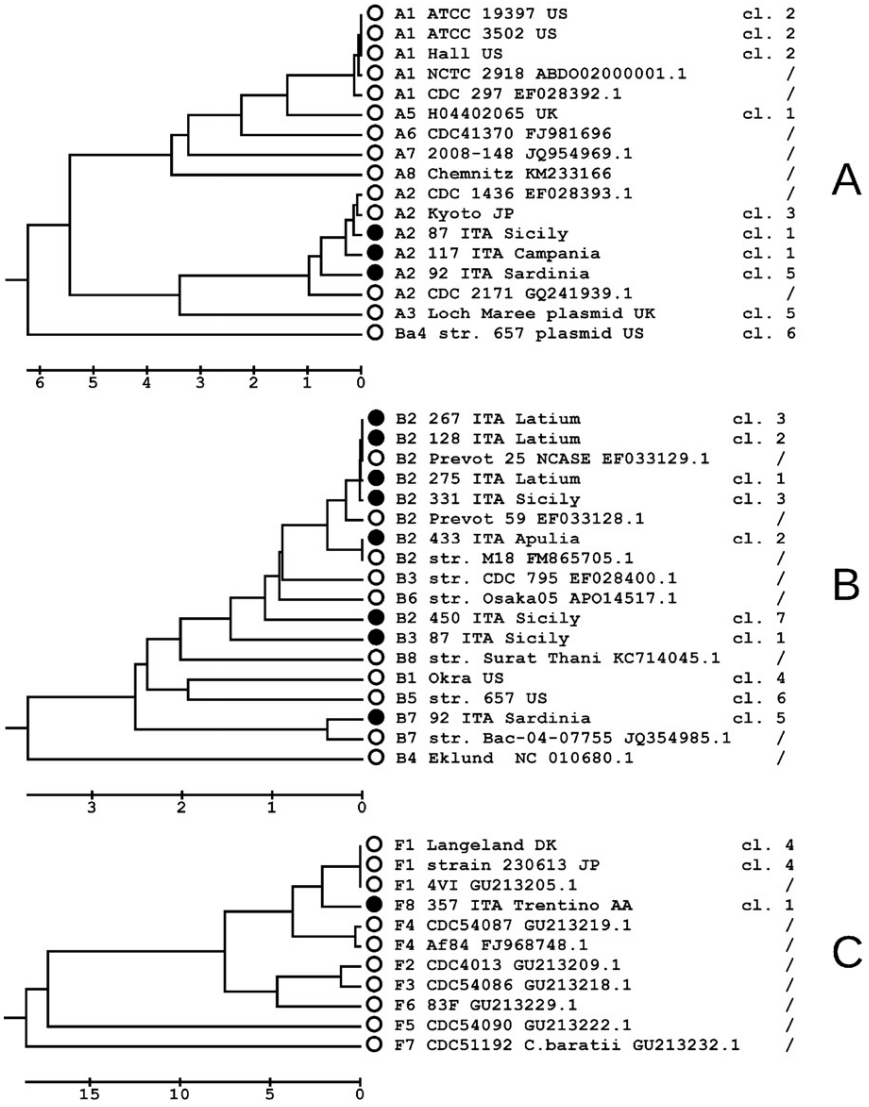


Figure 21: *bont* phylogeny. UPGMA dendrograms constructed on the nucleotide sequence alignment of: bont/A (A), bont/B (B), bont/F (C). Other GenBank *bont* selected sequences were included for comparison. X axis units=% dissimilarity.

Observing bont/B dendrogram (Figure 21B), we were also able to underline some discrepancies with whole genome phylogeny. For example, A2B3 87 has a bont sequence quite different from the 275 one, another genome included in cluster 1. Furthermore, some Italian B2 strains (275, 267, 331 and 128) show a very similar bont sequence: two B2 *bont* sequences (267 and 128) are identical to Prevot 25 NCASE, while those of B2 331 and B2

275 differ for only one nucleotide. 267, 128 and 275 have also identical toxin complex genes sequences (HA70, HA17, HA33 and NTNH). However, they do not appear highly correlated in the chromosome clonal phylogeny (Figure 19A), where they are located in different clusters. Probably, such bont gene complex sequence had a considerable lateral spread by recombination among different lineages. To confirm this assumption, evidences of recombination were searched in B2 128, that can be considered an advantageous candidate for recombination analysis because of the similarity with B2 433 genome but the different bont localization (plasmidic in 433 and chromosomic in 128).

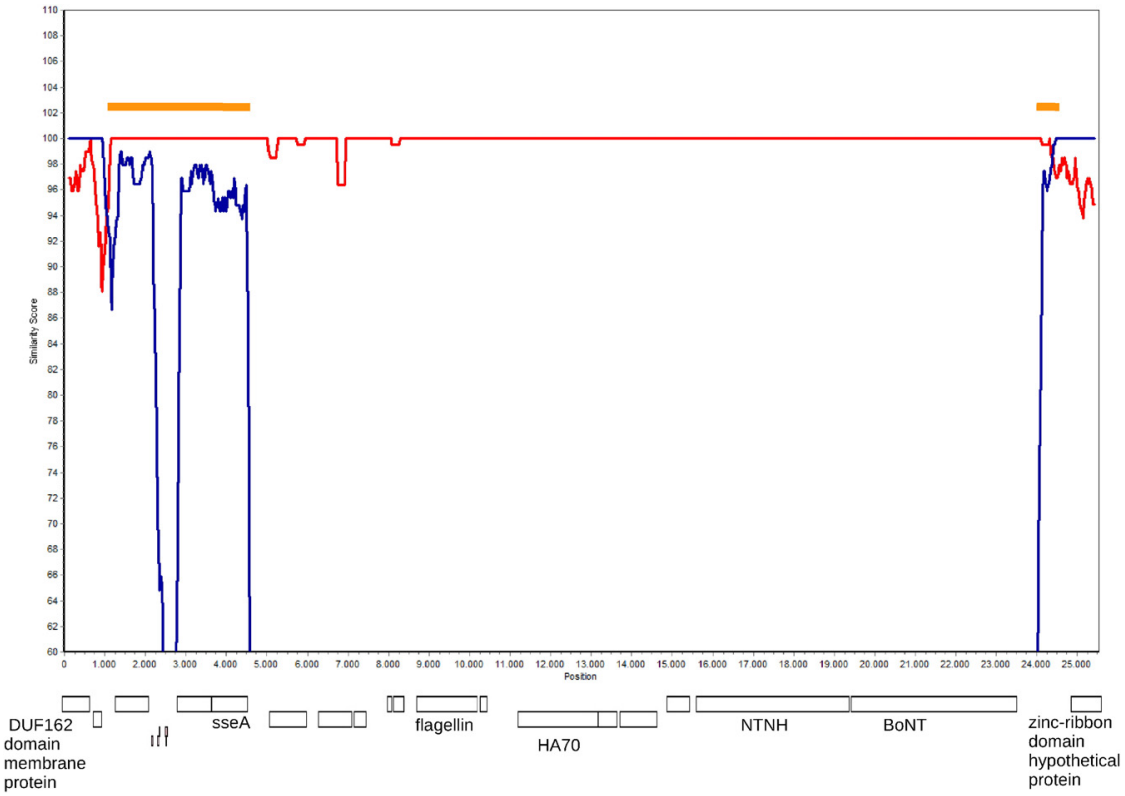


Figure 22: bont/B gene complex simplot output. Comparison of a 25 Kb sequence, (*bont/B*) and encompassing regions. Gene position and labels shown: B2 267= Red (contig008: 394,467–419,932 bps), B2 433= blue (contig008: 407,391–424,317 bps), B2 128= query(contig008: 407,507–432,961 bps).Window = 200 bps. Step= 20 bps. Orange strips highlights recombinant tracts.

SimPlot test [124] was performed to identify possible recombination breakpoints, using B2 128 as query, while B2 267 and B2 433 as the references. The tract analyzed by SimPlot spans between the core genes corresponding to CDSs CBO0788 and CBO0811 of

A1 ATCC 3502. In the SimPlot graph, flanking regions of B2 128 bont gene cluster show a chimerical pattern (constituted by 433-like distal and 267-like proximal portions).

This suggests a double recombination event (Figure 22), that could have been the vehicle for bont gene cluster insertion in the chromosome according the model described in Brigulla and Wackernagel 2010 [153] as "foreign DNA flanked by homologous sequences to a recipient chromosome".

About F phylogeny (Figure 19C), F 357 is clustered with the F1 strains, but, as previously observed, its BoNT sequence is quite different from this subtype, and amino acid similarity is low enough to consider F 357 as a distinct subtype.

4.6 Flexible genome characterization and typing

Total length of core-genome regions (>500 bps) is 2.62 Mbs, that is 66% of the length of A1 ATCC 3502, congruent with results obtained by DNA microarray [170]. As shown by whole genome DNA microarray, C. botulinum group I has a relatively stable genome, but it is not a closed pan-genome species [161] and the gene content has a moderate variability among the strains [170].

All genomes were characterized also considering the genes contained by a part of the strains in a species: the flexible genome [171]. Each of the 10 Italian sequenced genome was found to contain unique genes. For example, in B2 267 a probable ATP-dependent Lon protease (contig008; 762042-764072 bps) [122] was found, showing 80% amino acid sequence similarity with *Bacillus cereus* CDS WP_000389804.1. In B2 275, two adjacent genes (contig009; 1324153-1328225 bps) have 54% and 46% amino acid sequence similarity with the Vegetative Insecticidal Proteins Vip2A (ABR68092.1) and Vip1A (ABR68093.1) of *Bacillus thuringiensis*, respectively [172]. Three adjacent genes, present only in B2 128 (contig008; 1721577 – 1725673 bps), are probably involved in lantibiotics biosynthesis and transport [173], as suggested by amino acid sequence similarity with two *Peptoniphilus rhinitidis* genes (43% WP_010248479.1 and 54% WP_010248853.1). In A2B3 87, a series of genes (contig009; 393606 – 453348 bps) encodes for proteins containing the adenylation domain of nonribosomal peptide synthetases (NRPS) and/or a polyketide synthase

domain, and show homology with YP_002505315.1-5326.1 CDSs of *Clostridium cellulolyticum* H10 (amino acid sequence similarity ranging from 43% to 64%). This region probably constitutes an operon for biosynthesis of hybrid polyketide-amino acid metabolite [150].

Differences in gene content were used for a further characterization of the genomes.

Forty genes (Table 17), belonging to the flexible genome of *C. botulinum* group I, were chosen among those owned by at least two genomes. Binary data for presence/absence of the 40 genes were used to calculate the dendrogram in Figure 23. The grouping is similar to that obtained in clonal phylogeny (Figure 19A) cluster 2, 3 and 4 are conserved), with some exception that can be due to horizontal gene transfer, as the position of B2 450, that appears close to cluster 4, while in clonal phylogeny is divergent from all the others.

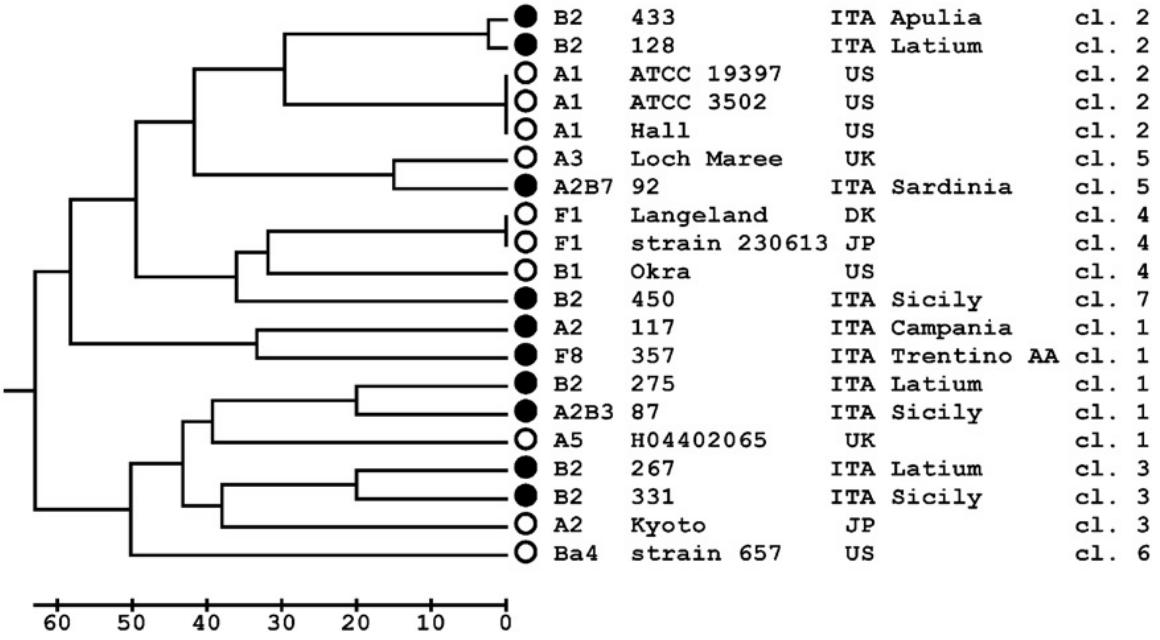


Figure 23: Flexible genome dendrogram. Dendrogram based on flexible genome analysis (40 genes). X axis units = % dissimilarity.

CHAPTER 5
RESULTS AND DISCUSSION – VIRTUAL CHARACTERIZATION

5.1 Model structures.

We run clustalO [174] multiple alignment to identify any phylogenically derived cluster. This allowed us to identify two identical amino acid sequences (B2_128 and B2_267) and cluster the set into three groups: 1) A2B7_92, A2B3_87, A2_117; 2) B2_450, B2_433, B2_331, B2_267, B2_128, B2_275 and 3) F_357, belonging to BoNT serotypes A, B and F respectively (phylograms of BoNTs and their four domains reported in Figure 24, identity matrixes reported in Table 20).

Since B2_128 and B2_267 sequences revealed to be identical, we retained only B2_128 in the modelling set. The analysis of the identity percentage within the four domains reveals identical motifs that might affect BoNT both serotyping and catalytic activity:

1) Clostridial neurotoxin zinc protease: A2B7_92, A2B3_87, A2_117 amino acidic sequences show an identity percentage higher than 98% B2_450, B2_433, B2_331, B2_267, B2_128, B2_275 are identical, F_357 percentage of identity ranges between 30 and 40%, the experimental structure 3bta, BoNT/A classified holds an identity percentage higher than 95% within A2B7_92, A2B3_87, A2_117 and around 33% within the rest of the dataset;

2) Translocation domain: A2B7_92, A2B3_87, A2_117 hold a high percentage of identity (around 98% within the group); 3bta percentage of identity within A cluster reaches 84% and within B cluster 52%, only 42% respect to F_357; B2_450, B2_433, B2_331, B2_267, B2_128 and B2_275 are reported to be identical, F_357 identity matrix percentage within the data set ranges from 42 to 48%;

3) N-terminal receptor binding: A2B3_87 and A2_117 are 100% identical, 99.48% with A2B7_92 and around 85% with 3bta; B2_267, B2_128 and B2_275 are 100% identical, and around 99% with B2_450, B2_433, and B2_331; F_357 identity matrix reaches 60% respect A cluster 48% respect B cluster;

4) C-terminal receptor binding: A2B7_92, A2B3_87, A2_117 hold a high percentage of identity (around 98% within the group); 3bta percentage of identity within A cluster reaches 91% and within B cluster 33%, while 43% respect to F_357; B2_331, B2_267 and B2_128 are reported to be identical, holding 96% identity with B2_433 and 89% with B2_450; F_357 identity matrix percentage within A group data set reach 32% and 44% within B group.

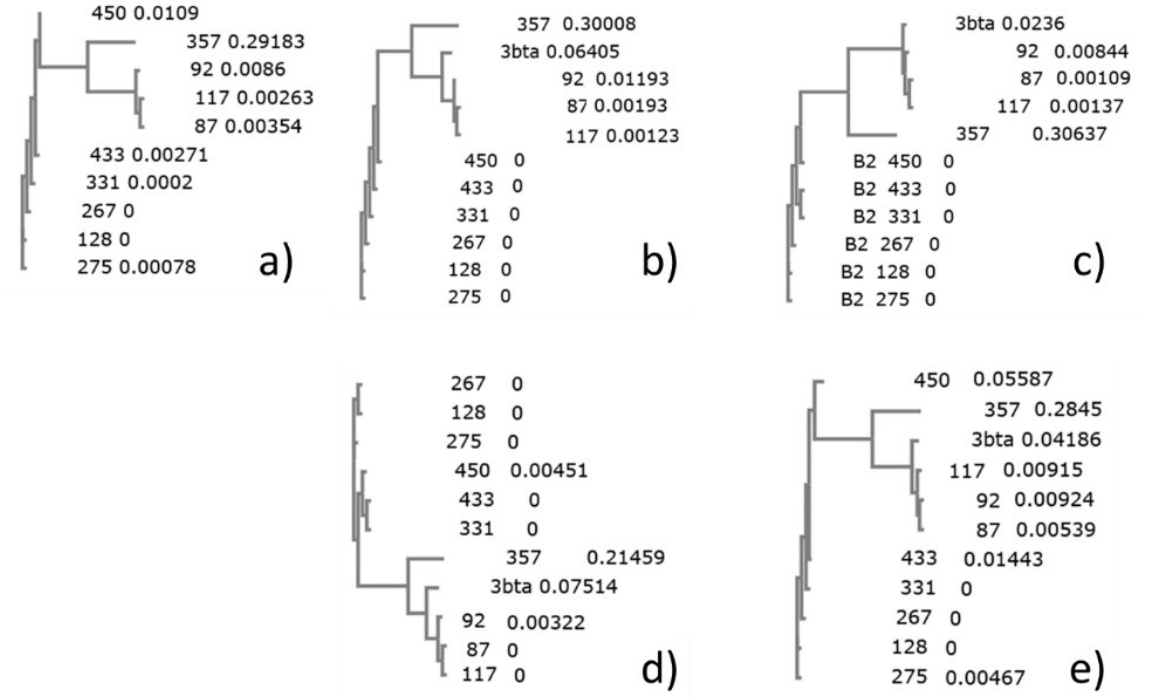


Figure 24: BoNT and its domains amino acid sequence phylogenies (clustalO output). phylogeny obtained computing: a) the whole BoNT amino acidic sequence; b) catalytic domain; c) translocation domain; d) N-terminal domain and e) C-terminal domain.

Finally, we modelled all nine non identical sequences as above described. Since in Pfam [175] are yet deposited whole BoNT/A annotated protein structures, we used as reference for domain characterization purposes the annotated amino acidic sequence of Botulinum neurotoxin type A EC=3.4.24.69, and pdb id 3bta [33] experimentally resolved structure as internal standard in order to identify the model putative domains and to cross check the homology models' quality (scheme of BoNT Pfam domains in Figure 25, chart of the domains in

Table 9 and details of quality estimation in APPENDIX B).

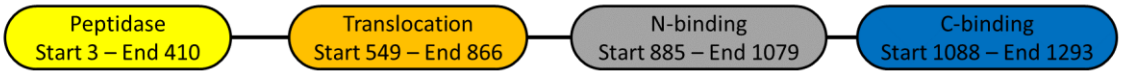


Figure 25: BoNT Pfam domains. Scheme of the reference annotated amino acidic sequence of Botulinum neurotoxin type A EC=3.4.24.69, source organism Clostridium botulinum (strain Hall / ATCC 3502 / NCTC 13319 / Type A) (NCBI taxonomy ID 441771), Length: 1296 amino acids.

Table 9: BoNT models domains characterization. The characterization relies to the clustalO multiple alignment to the reference annotated Hall amino acidic structure, Figure 25. The last line reports the templates used for modelling procedures.

Domain		Hall	92	87	117	450	433	331	128	275	357
1) Clostridial neurotoxin zinc protease	Start	3	4	4	4	4	4	4	4	4	4
	End	410	410	410	410	418	418	418	418	418	410
2) Translocation domain	Start	549	549	549	549	536	536	536	536	536	537
	End	866	866	866	866	853	853	853	853	853	863
3) N-terminal receptor binding	Start	885	885	885	885	872	872	872	872	872	882
	End	1079	1079	1079	1079	1066	1066	1066	1066	1066	1076
4) C-terminal receptor binding	Start	1088	1088	1088	1088	1075	1075	1075	1075	1075	1085
	End	1293	1293	1293	1293	1290	1290	1290	1290	1290	1280
pdb templates		//	2nyy, 3v0a, 3bta	2nyy, 2nz9 3v0a, 3bta	2nyy, 3v0a, 3bta	2np0, 1g9a, 1g9c	2np0, 1g9a, 1g9c	2np0, 1g9a, 1g9c	2np0, 1g9a, 1g9c	2np0, 1g9a, 1g9c	2nyy, 2nz9, 3fii, 3fuk

The nine models and the selected experimental protein structure choose for comparison, pdb id 3bta, underwent Ramachandran plot statistics with Procheck [141]. The values obtained from the virtual models were coherent with the experimental 3bta Ramachandran output, confirming the genuineness of the backbone geometry (Table 10 and Table 11).

More than 92% of the residues are located in the most favored regions, less than 8 % in additional allowed regions, less than 1 % in generously allowed regions and less than 0.1 % in disallowed regions (Table 10 and Table 11). Procheck provided the analysis of the observed distributions of φ-ψ, χ1- χ 2, χ -1, χ -3, χ -4 and ω values for each amino acid types and estimated as normal the global G-factor analysis [176]. Ramachandran plots for all virtually obtained models were produced by PDBsum utility [177] and qualitatively compared with the experimentally resolved structure 3bta Ramachandran plot (Figure 26). For all virtual models Ramachandran plots the backbone ψ - φ dihedral angles of the

amino acid residues are visualized in sterically allowed regions qualitatively coherent with the backbone (Figure 26 and Figure 27).

Once Ramachandran analysis revealed the genuinity of backbone features for all generated models, from each 3D structure we extracted the four domains and uploaded to QMEAN server [142] for model quality estimation. For all submitted theoretical models and domains QMEAN quality assessment provided values comparable to experimentally obtained protein structures (APPENDIX B).

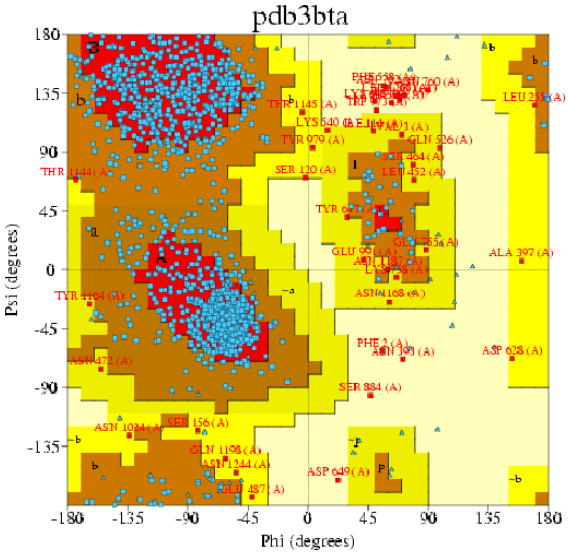


Figure 26: Ramachandran plot generated from BoNT/A (PDB ID 3bta). The red, brown, and yellow regions represent the favored, allowed, and "generously allowed" regions as defined by ProCheck for pdb code 3bta experimentally resolved model (*https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl* pdb code 3bta).

Since the QMEAN4[12] scores obtained for all homology models and their domains assessed structure qualities comparable or better than the experimentally reference (pdbid 3bta), we deposited all models in the PMDB Protein Model Database [178] (gene/PMDB code couples: A2_117/PM0080396; A2B3_87/PM0080398;

[12]**QMEAN4**: a linear combination of four structural descriptors: 1) local geometry is analyzed by a torsion angle potential over three consecutive amino acids; 2) two distance-dependent interaction potentials are used to assess long-range interactions: First, at a residue-level it is based on C-beta atoms only, at the second level an all-atom potential is used. 3) solvation energy is calculated to investigates the burial status (accessibility to water) of the residues.

A2B7_92/PM0080397; B2_267 (B2_128)/PM0080399; B2_275/PM0080400; B2_331/PM0080401; B2_433/PM0080402; B2_450/PM0080403; F_357/PM00804XX).

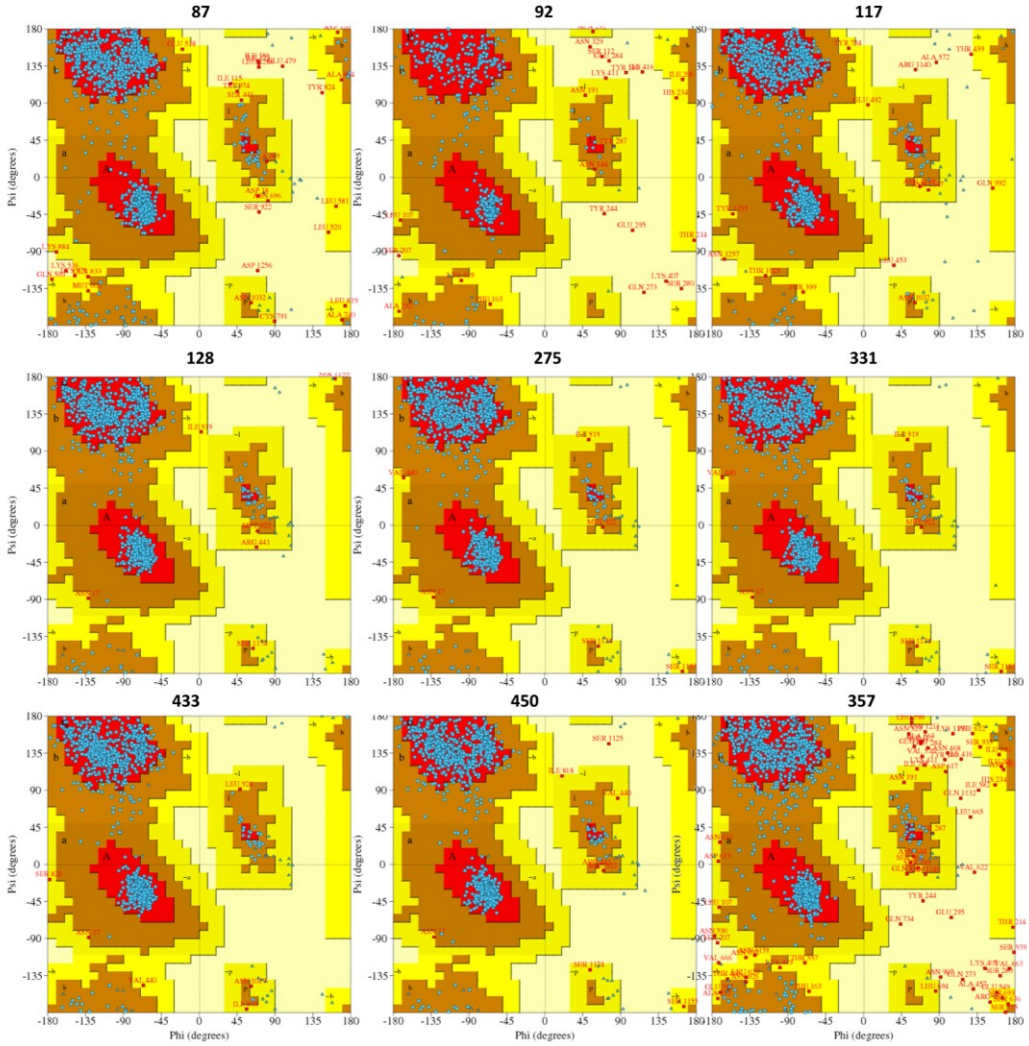


Figure 27: Virtual model Ramachandran plots. By means of Ramachandran plots it is possible to cluster the obtained models into the serotype corresponding cluster: BoNT/A virtual models (A2_117/PM0080396; A2B3_87/PM0080398 and A2B7_92/PM0080397), BoNT/B virtual models (B2_267-B2_128/PM0080399; B2_275/PM0080400; B2_331/PM0080401 and B2_433/PM0080402; B2_450/PM0080403) and BoNT/F virtual model (F_357/PM00804XX). For BoNT/F virtual model, Ramachandran plot reveals more unfavorable backbone geometries than the other virtual models. This might be due to the modelling template used for the catalytic domain: 3fii pdb structure. Overall, Ramachandran analysis is still comparable to the homologs experimentally resolved BoNT models.

Table 10: Ramachandran statistics.

Models	3bta		92		87		117		450	
	N. Res	%	N. Res	%	N. Res	%	N. Res	%	N. Res	%
Most favoured regions	1098	92.0%	1092	91.5%	1082	90.7%	1044	87.6%*	1097	92.0%
Additional allowed regions	87	7.3%	87	7.3%	90	7.5%	120	10.1%	90	7.5%
Generously allowed regions	7	0.6%	10	0.8%	12	1.0%	19	1.6%	6	0.5%
Disallowed regions	1	0.1%*	4	0.3%*	9	0.8%*	9	0.8%*	0	0.0%
Non-glycine and non-proline residues	1193	100.0%	1193	100.0%	1193	100.0%	1192	100.0%	1193	100.0%
End-residues (excl. Gly and Pro)	2		2		2		2		2	
Glycine residues	58		62		62		62		58	
Proline residues	38		39		39		40		38	
Total number of residues	1291		1296		1296		1296		1291	
Models	**433**		**331**		**128**		**275**		**357**	
	N. Res	%	N. Res	%	N. Res	%	N. Res	%	N. Res	%
Most favoured regions	1107	92.7%	1104	92.5%	1098	92.0%	1107	92.7%	1107	92.7%
Additional allowed regions	81	6.8%	83	7.0%	87	7.3%	81	6.8%	81	6.8%
Generously allowed regions	6	0.5%	6	0.5%	7	0.6%	6	0.5%	6	0.5%
Disallowed regions	0	0.0%	0	0.0%	1	0.1%*	0	0.0%	0	0.0%
Non-glycine and non-proline residues	1194	100.0%	1193	100.0%	1193	100.0%	1194	100.0%	1194	100.0%
End-residues (excl. Gly and Pro)	2		2		2		2		2	
Glycine residues	58		58		58		58		58	
Proline residues	37		38		38		37		37	
Total number of residues	1291		1291		1291		1291		1291	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20.0 a good quality model would be expected to have over 90% in the most favoured regions [A,B,L].

Table 11: Gfactors.

	3bta	92	87	117	450	433	331	267	275	357
Parameter	**Score**	**Score**	**Score**	**Score**	**Score**	**Score**	**Score**	**Score**	**Score**	**Score**
Dihedral angles:										
Phi-psi distribution	0.05	-0.04	-0.09	-0.27	0.04	0.04	0.05	0.05	0.04	0.04
Chi1-chi2 distribution	0.20	0.06	0.05	-0.11	0.16	0.17	0.19	0.20	0.17	0.17
Chi1 only	0.21	0.25	0.23	0.17	0.25	0.25	0.24	0.21	0.25	0.25
Chi3 & chi4	0.38	0.52	0.48	0.46	0.41	0.44	0.38	0.38	0.44	0.44
Omega	-0.10	-0.23	-0.17	-0.22	-0.12	-0.13	-0.11	-0.10	-0.13	-0.13
Average	0.08	0.00	-0.01	-0.12	0.07	0.07	0.08	0.08	0.07	0.07
Main-chain covalent forces:										
Main-chain bond lengths	-0.07	-0.10	-0.14	-0.31	-0.07	-0.12	-0.07	-0.07	-0.12	-0.12
Main-chain bond angles	-0.09	-0.27	-0.27	-0.45	-0.11	-0.15	-0.09	-0.09	-0.15	-0.15
Average	-0.08	-0.20	-0.22	-0.39	-0.09	-0.13	-0.08	-0.08	-0.13	-0.13
OVERALL AVERAGE	0.02	-0.07	-0.08	-0.22	0.01	0.00	0.02	0.02	0.00	0.00

G-factors provide a measure of how unusual, or out-of-the-ordinary, a property is.

Values below -0.5* - unusual

Values below -1.0** - highly unusual

Important note: The main-chain bond-lengths and bond angles are compared with the ref ideal values derived from small-molecule data. Therefore, structures refined using different restraints may show apparently large deviations from normality.

Table 12: QMEAN4 quality assessment values.

BoNT serotype	A				B					F
Model	**3bta**	**87**	**92**	**117**	**128**	**275**	**331**	**433**	**450**	**357**
Complete		-2.08	-2.27	-4.03	-1.47	-1.60	-1.66	-1.55	-1.55	
N-Terminal	-2.84	-2.99	-2.80	-3.47	-3.48	-3.56	-3.48	-3.58	-3.15	-2.29
C-Terminal	-1.40	-1.57	-1.53	-1.57	-0.66	-0.66	-0.56	-0.56	-0.72	-1.01
Traslocation	-2.36	-1.77	-3.50	-3.40	-1.23	-1.23	-1.23	-1.23	-1.23	-5.20
Zinc-protease	-2.20	-1.11	-1.35	-1.10	-0.39	-0.39	-0.39	-0.39	-0.39	-1.53

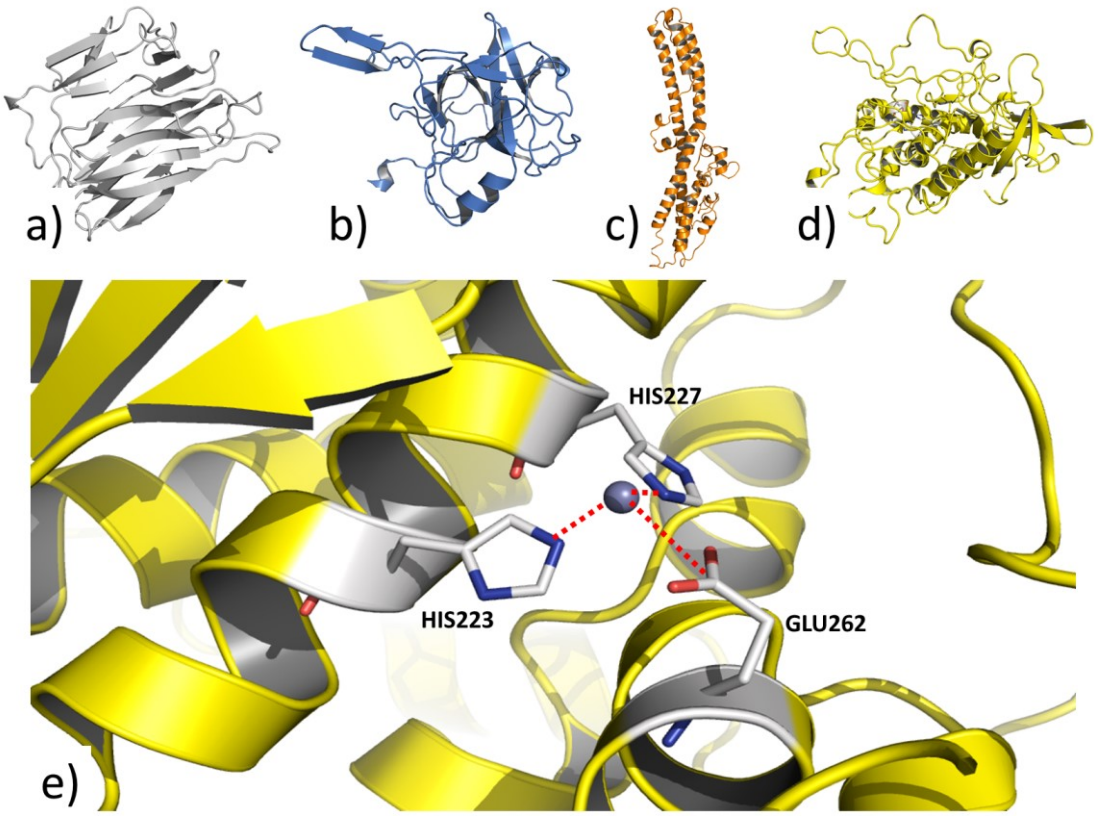


Figure 28: A2B7_92/PM0080397 homology model. Cartoon representation of A2B7_92/PM0080397 model sown in green cartoon: a) front view; b) top view; c) C-terminal domain; d) N-Terminal domain; e) Translocation domain; f) Zinc-protease domain; g) view of the zinc coordinated catalytic function. The model was built using three templates, all of them experimentally resolved (pdbid: 3nyy, 2nz9 and 3bta).

5.2 Docking calculations.

Since we unveiled five non-identical clostridial neurotoxin zinc peptidase domains, the zinc-catalytic metalloprotease of the theoretical models PM0080396, PM0080398, PM0080397, PM0080399 and PM00804XX were inserted in the receptor dataset (pdbid: 1epw; 1f31; 1f82; 1i1e; 1s0d; 1t3a; 1t3c; 1xtg; 1zkw; 1zkx; 1zl6; 1zn3; 2a8a; 2a97; 2fpq; 2g7p; 2g7q; 2ilp; 2imb; 2imc; 2nyy; 2nz9; 2qn0; 2w2d; 2xhl; 3bon; 3boo; 3bwi; 3c88; 3c89; 3c8a; 3c8b; 3dda; 3ddb; 3ds9; 3dse; 3ffz; 3fie; 3fii; 3qix; 3qj0; 3qw5; 3qw6; 3qw7; 3qw7; 3qw8; 3v0a; 3v0b; 3zus; 4ks6; 4ktx; 4kuf).

Re-docking. Re-docking check, with except of 3fie/lig10 complex, produced outputs superimposable with the experimental complex structures. In Figure 29 is presented the Lig01-3boo redocking. All 3fie re-docking poses produced output substantially qualitatively diverse from the experimental 3fie complex structure. Eventually lig10 poses did not even reproduce the experimental metal coordination. Since this ligand was prepared truncating INH01, it may have lost the native experimental complexing dynamics towards the experimentally obtained protein structure (Figure 30). For all other re-docking calculations the generated poses maintained ligand receptor complexation qualitatively coherent with the experimental receptor/zinc/ligand coordination in terms of ligand atom type coordination and zinc atom-ligand interacting atom type distance (Table 15). Results of all docking calculations are reported in Table 13 and Table 14.

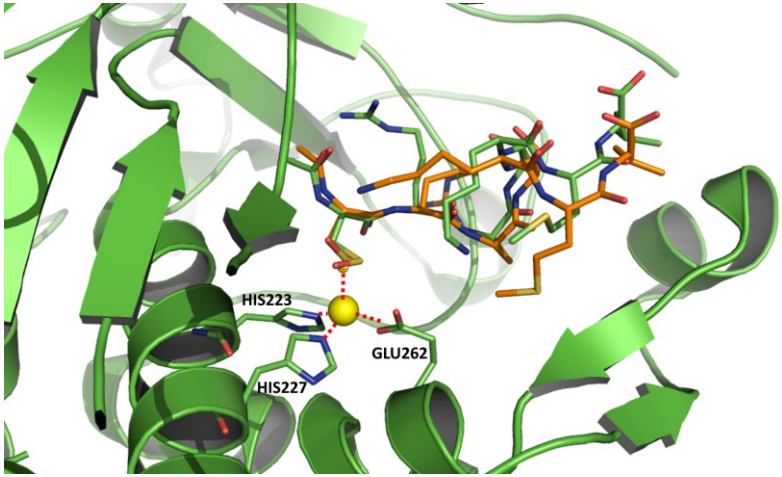


Figure 29: 3boo redock. Snapshot of the superimposed Lig01 redocked to its native receptor, pdbid 3boo. Color coding: 3boo complex in green cartoon and carbon-green sticks; Lig01 pose in carbon-orange sticks; Zinc atom and the dummy atom are depicted as yellow spheres; coordinating residues are labelled.

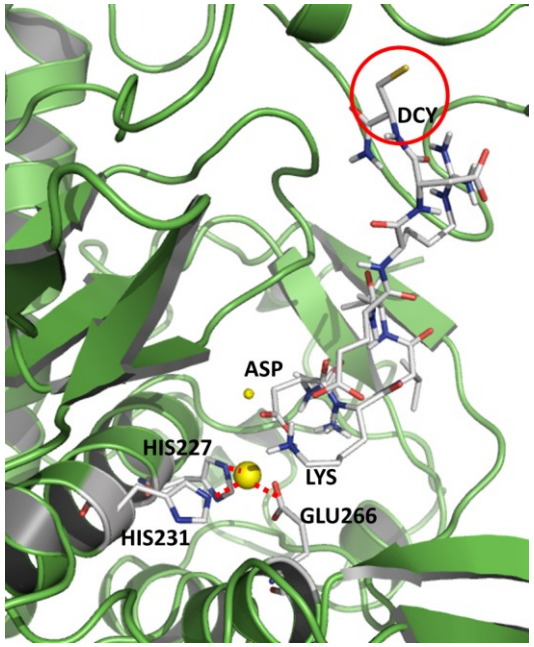


Figure 30: 3fie redock. Snapshot of the superimposed Lig10 redocked to its native receptor, pdbid 3fie. Color coding: 3fie BoNT zinc catalytic domain in green cartoon, zinc coordinating residues carbon-grey sticks; Lig10 pose in carbon- grey sticks; Zinc atom and the dummy atom are depicted as yellow spheres; coordinating residues are labelled. In this pose Lig10 appears upside-down respect to the complexing position of INH01 in 3fie complex. Highlighted with a red circle the INH01 Sulphur atom, in this case not even pointing to the metal coordinating atom.

Estimated Free Energy of Binding values globally range from -15.76 to -0.72 kcal/mol (excluding Estimated Free Energy of Binding values obtained from Lig10), Estimated Inhibition Constant, Ki[13], values range from 1.02 nM to 20.07 μM[14] (considering the lowest Estimated Inhibition Constant value for each ligand/receptor couple docking calculation, excluding Estimated Inhibition Constant values obtained from Lig10). Since the selected peptide inhibitors were extracted from BoNT/A-inhibitor complex dataset and the other serotypes show 33% of reciprocal identity, we expected that docking results would have respected at least two clusters: one characterized by a higher ligand-receptor affinity for BoNT/ A, C and E structures (since they share the same target, SNAP-25), and a second owning a poorer affinity for other BoNT serotypes B, D, F, G (sharing the VAMP target) and C targeting syntaxin.

[13] Ki is based on the principles of QSAR (quantitative structure-activity relationships) and is parameterized using a large number of protein-inhibitor complexes for known structure and inhibition constants (Ki). The user is encouraged to refer to the description of how this free energy function was derived in the original literature [149]

[14] Experimentally obtained BoNT/A inhibitor constant for N-acetyl-CRATKML peptide inhibitor is 330 nM.

Table 13: Docking affinity table.

BoNT zinc protease domains	Peptide inhibitors									
	Lig01	**Lig02**	**Lig03**	**Lig04**	**Lig05**	**Lig06**	**Lig07**	**Lig08**	**Lig09**	**Lig10**
flex 3boo	-3.88	-5.5	-8.13	-9.08	-9.14	-4	-6.62	-6.92	-4.08	-0.94
flex 3c8a	-5.96	-10.64	-13.01	-12.47	-12	-11.48	-10.37	-10.32	-11.34	-5.15
flex 3c8b	-5.21	-11.13	-10.74	-6.73	-9.85	-9.64	-8.84	-10.07	-10.48	-5.65
flex 3fie	-4.49	-9.49	-10.39	-11.13	-10.13	-10.95	-8.51	-9.6	-9.18	-0.6
flex 3qw5	-7.17	-12.35	-13.67	-13.35	-11	-10.29	-11.5	-12.66	-11.59	-5.77
flex 3qw6	-6.76	-11.67	-10.39	-10.52	-11.69	-13.05	-11.64	-10.67	-9.9	-3.34
flex 3qw7	-7.07	-14.52	-12.13	-11.15	-10.79	-11.12	-11.15	-10.38	-10.36	-2.39
flex 3qw8	-2.07	-7.58	-4.42	-6.64	-5.32	-11.54	-5.32	-5.17	-6.45	-0.54
flex 4ks6	-6.07	-10.17	-9.73	-11.08	-6.86	-8.98	-7.67	-8.43	-6.83	-1
flex 4ktx	-6.58	-15.76	-12.8	-15.56	-12.74	-15.21	-12	-12.23	-11.54	-1
H 117	-4.85	-8.79	-8.76	-11.05	-10.07	-8.42	-7.58	-7.25	-7.63	-0.47
H 87	-6.49	-10.82	-9.97	-11.77	-10.84	-10.82	-9.28	-9.41	-9.18	-4.57
H 92	-5.93	-9.86	-9.62	-10.26	-11.42	-13.42	-9.44	-9.08	-8.28	-3.58
H 450	-6.87	-11.88	-12.48	-11.1	-11.28	-11.9	-9.12	-8.58	-8.7	-1.08
H 357	-6.07	-10.66	-0.72	-7.74	-9.74	-9.38	-8.29	-9.3	-8.87	-9.81
A-3boo	-6	-8.56	-9.65	-9.88	-9.58	-9.69	-7.41	-7.56	-7.05	-1.14
A-3c8a	-5.28	-10.2	-11.14	-9.64	-8.57	-9.25	-8.3	-8.01	-9.29	0
A-3c8b	-6.59	-10.99	-10.54	-11.04	-11.01	-9.31	-8.32	-10.27	-8.93	0
A-3qw5	-5.55	-10.7	-10.73	-12.43	-10.43	-9.96	-8.34	-10.27	-10.26	-0.63
A-3qw6	-4.62	-10.2	-12.17	-11.03	-10.32	-11.62	-9.28	-10.05	-9.47	-1.09
A-3qw7	-4.52	-11.6	-11.18	-10.62	-10.04	-10.2	-8.34	-9.76	-9.43	0
A-3qw8	-5.22	-10.6	-11.37	-11.67	-9.65	-10.32	-9.42	-10.67	-10.3	0

Continued in next page

Continued from next page

BoNT zinc protease domains	Lig01	Lig02	Lig03	Lig04	Lig05	Lig06	Lig07	Lig08	Lig09	Lig10
A-4ks6	-8.41	-11.74	-12.93	-12	-10.51	-11.36	-10.56	-11.66	-11.36	-0.77
F-3fie	-4.84	-10.75	-11.03	-10.31	-10.14	-11.83	-8.07	-10.06	-9.95	-0.32
A-2nyy	-5.53	-9.81	-9.44	-8.64	-8.77	-9.3	-9.55	-9.91	-8	-3.62
A-2nz9	-6.54	-11.28	-11.56	-11.19	-9.76	-10.4	-9.12	-8.86	-8.86	-2.23
A-3v0a	-9.14	-11.17	-10.26	-10.71	-10.94	-13.77	-8.8	-10.33	-9.85	-4.86
A-3v0b	-6.29	-9.8	-9.68	-12.22	-11.05	-14.26	-8.81	-11.21	-9.56	-4.64
A-1xtg	-3.38	-8.14	-7.98	-7.69	-9.82	-9.59	-7.03	-7.37	-9.26	-1.79
A-2g7p	-5.02	-9.29	-8.93	-10.87	-9.41	-9.8	-8.56	-9.57	-6.39	-2.95
A-2g7q	-4.08	-9.6	-10.73	-10.99	-9.87	-9.21	-8.4	-8.85	-7.78	-5.17
A-2ilp	-5.47	-7.6	-6.66	-8.73	-8.84	-8.77	-6.72	-7.14	-6.63	0
A-2imb	-4.84	-8.47	-9.34	-10.2	-8.37	-8.56	-7.6	-7.33	-7.53	-2.9
A-2imc	-4.99	-8.77	-9.48	-11.61	-9.55	-8.65	-7.55	-7.02	-8.49	0
A-2w2d	-6.23	-10.62	-11.22	-12.34	-10.72	-10.27	-9.26	-11.14	-9.55	-3.38
A-3bon	-5.59	-7.97	-8.37	-7.14	-7.3	-9.19	-8.39	-9.77	-9.18	0.25
A-3bwi	-5.38	-11.38	-10.69	-10.67	-12.15	-10.54	-7.9	-9.3	-7.92	-1.45
A-3c88	-6.36	-10.49	-10.56	-11.21	-10.62	-10.12	-8.77	-10.13	-9.61	-2.5
A-3c89	-3.24	-9.6	-10.96	-10.94	-8.72	-9.43	-7.75	-7.87	-9.15	-2.19
A-3dda	-4.18	-10.14	-10.76	-11.32	-10.81	-8.42	-8.58	-9.54	-8.57	-0.53
A-3ddb	-4.81	-10.09	-9.98	-11.01	-10.3	-8.98	-8.61	-8.94	-8.89	-0.29
A-3ds9	-5.28	-9.82	-8.91	-10.46	-9.85	-11.91	-8.2	-8.8	-8.62	-0.48
A-3dse	-5.64	-11.06	-8.65	-9.71	-9.71	-8.76	-8.42	-8.78	-8.07	-1.93
A-3qix	-4.77	-9.65	-9.32	-9.39	-9.61	-9.63	-7.98	-8.47	-6.84	-0.73
A-3qj0	-6.77	-7.17	-8.11	-6.49	-10.32	-9.43	-7.25	-8.85	-6.17	-2.31
A-3zus	-7.36	-10.5	-11.6	-12.63	-11.86	-10.06	-10.25	-9.18	-9.75	-0.01
A-4kuf	0	-12.41	-13.1	-15.26	-10.66	-11.45	-10.9	-12.54	-11.3	0

Continued in next page

Continued from next page

BoNT zinc protease domains	Lig01	Lig02	Lig03	Lig04	Lig05	Lig06	Lig07	Lig08	Lig09	Lig10
B-1epw	-6.35	-10.09	-10.8	-11.81	-10.61	-12.23	-9.6	-10.13	-9.77	-3.27
B-1f31	-8.23	-12.82	-12.78	-13.01	-11.62	-12.57	-8.96	-9.72	-12.01	-4.25
B-1f82	-4.84	-10.89	-11.33	-10.42	-10.34	-9.28	-9.14	-8.98	-9.79	-2.2
B-1i1e	-7.49	-12.5	-11.94	-13.3	-10.9	-11.06	-10	-10.12	-9.46	-4.55
B-1s0d	-7.58	-13.21	-13.48	-13.76	-11.65	-13.51	-11.2	-10.86	-10.97	-3.09
B-2xhl	-7.74	-10	-11.39	-11.6	-11.06	-13.04	-9.34	-9.48	-9.18	-4.4
C-2fpq	0	-3.03	0	0	-5.88	-1.6	0	-4.08	-2.77	0
E-1t3a	-4.74	-10.88	-11.22	-10.06	-10.35	-9.98	-8.49	-10.38	-8	-1.62
E-1zkw	-5.42	-11.92	-11.62	-10.3	-9.26	-12.42	-9.47	-9.8	-9.13	-1.46
E-1zkx	-5.27	-8.1	-8.57	-8.33	-8.34	-9.54	-7.26	-5.9	-7.03	-0.4
E-1zl6	-7.26	-8.99	-9.84	-9.1	-8.44	-9.36	-7.57	-8.81	-8.2	-1.29
E-1zn3	-5.16	-9.13	-10.15	-9.57	-9.87	-10.25	-8.17	-8.52	-7.88	-0.58
E-3ffz	-7.18	-11.36	-11.22	-10.21	-10.28	-9.46	-8.71	-9.28	-9.57	-2.18
F-2a8a	-5.21	-10.87	-11.77	-10.01	-9.64	-9.6	-8.83	-8.63	-7.27	0
F-2a97	-6.33	-10.25	-10.9	-10.82	-10.89	-10.71	-8.95	-9.92	-8.58	-1.65
F-3fii	-4.48	-8.52	-10.06	-9.72	-10.99	-9.63	-9.12	-9.65	-9.65	-2.22

Table 14: EIK values

BoNT zinc protease domains	Lig01	Lig02	Lig03	Lig04	Lig05	Lig06	Lig07	Lig08	Lig09	Lig10
A-1e1h	48,63 nM	3,24 nM	1,11 nM	3,15 nM	1,02 nM	3,62 nM	1,42 nM	2,27 nM	////	1,70 nM
A-1xtf	////	2,02 nM	////	25,00 µM	1,20 nM	2,15 nM	1,78 nM	1,44 nM	////	////
A-1xtg	3,33 nM	1,08 µM	1,41 µM	62,98 nM	93,93 nM	6,990 µM	3,97 nM	0,16 µM	48,36 nM	2,30 µM
A-2g7n	1,17 nM	11,94 nM	1,14 nM	21,12 nM	15,71 nM	55,25 nM	0,22 µM	0,22 µM	0,90 µM	11,99 nM
A-2g7p	1,08 nM	0,15 µM	0,28 µM	0,12 µM	65,94 nM	0,52 µM	96,97 nM	20,70 µM	6,87 nM	10,76 nM
A-2g7q	1,03 nM	91,48 nM	13,74 nM	57,94 nM	0,17 µM	0,70 µM	0,32 µM	1,97 µM	2,14 nM	8,74 nM
A-2ilp	1,06 nM	1,19 nM	1,71 nM	0,33 µM	0,37 µM	11,94 µM	5,79 µM	1,42 nM	0,26 µM	0,39 µM
A-2imb	1,03 nM	0,61 µM	0,14 µM	0,73 µM	0,52 µM	2,71 nM	4,21 µM	3,04 µM	7,52 nM	33,34 nM
A-2imc	6,03 nM	0,37 µM	0,11 µM	99,47 nM	0,45 µM	2,90 nM	7,14 µM	0,60 µM	67,05 nM	3,09 nM
A-2nyy	1,41 nM	64,18 nM	0,12 µM	0,37 µM	0,15 µM	0,10 µM	54,66 nM	1,37 µM	2,23 nM	0,46 µM
A-2nz9	1,38 nM	5,42 nM	3,38 nM	70,55 nM	23,71 nM	0,20 µM	0,32 µM	0,31 µM	23,2 nM	6,29 nM
A-2w2d	1,22 nM	16,42 nM	5,95 nM	13,87 nM	29,80 nM	0,16 µM	6,81 nM	0,10 µM	3,35 nM	1,92 nM
A-3bon	1,65 nM	1,45 nM	0,73 µM	4,49 µM	0,18 µM	0,70 µM	69,48 nM	0,18 µM	////	2,70 nM
A-3boo	8,31 nM	0,53 µM	84,54 nM	94,49 nM	79,31 nM	3,73 nM	2,90 µM	6,79 µM	0,14 µM	57,73 nM
A-3bta	3,48 nM	4,13 nM	3,76 nM	1,87 nM	12,56 nM	0,45 µM	0,16 µM	0,38 µM	0,24 µM	1,56 nM
A-3bwi	1,30 nM	4,54 nM	14,64 nM	1,24 nM	18,85 nM	1,62 nM	0,15 µM	1,56 µM	86,42 nM	15,06 nM
A-3c88	1,32 nM	20,43 nM	18,09 nM	16,45 nM	38,45 nM	0,37 µM	37,33 nM	89,84 nM	14,65 nM	6,07 nM
A-3c89	4,21 nM	92,32 nM	9,22 nM	0,40 µM	0,12 µM	2,08 µM	1,69 µM	0,19 µM	24,95 nM	9,56 nM
A-3c8a	3,18 nM	33,65 nM	6,79 nM	0,51 µM	0,16 µM	0,83 µM	1,35 µM	0,15 µM	////	85,58 nM
A-3c8b	2,73 nM	8,82 nM	18,83 nM	8,46 nM	0,14 µM	0,79 µM	29,47 nM	0,28 µM	////	8,10 nM
A-3dda	1,38 nM	37,01 nM	12,87 nM	11,95 nM	0,67 µM	0,51 µM	0,10 µM	0,51 µM	0,41 µM	5,06 nM
A-3ddb	1,28 nM	40,06 nM	48,59 nM	28,38 nM	0,26 µM	0,49 µM	0,28 µM	0,30 µM	0,61 µM	8,55 nM
A-3ds9	4,23 nM	63,72 nM	0,29 µM	59,82 nM	1,86 nM	0,98 µM	0,35 µM	0,48 µM	0,44 µM	21,42 nM
A-3dse	5,48 nM	7,87 nM	0,45 µM	76,12 nM	0,38 nM	0,67 µM	0,36 µM	1,21 µM	38,21 nM	76,42 nM
A-3k3q	49,26 nM	17,8 nM	1,16 nM	1,41 nM	1,95 nM	1,17 nM	11,74 µM	2,32 nM	////	2,06 nM
A-3qix	2,00 nM	83,74 nM	0,14 µM	89,70 nM	87,94 nM	1,19 nM	0,62 µM	9,69 µM	0,29 nM	0,13 µM

Continued in next page

Continued from next page

BoNT zinc protease domains	Lig01	Lig02	Lig03	Lig04	Lig05	Lig06	Lig07	Lig08	Lig09	Lig10
A-3qj0	4,92 nM	5,55 nM	1,13 µM	27,14 nM	0,12 µM	4840 µM	0,32 µM	7,54 nM	20,37 nM	17,37 µM
A-3qw5	1,50 nM	14,41 nM	13,73 nM	22,79 nM	49,91 nM	0,76 µM	29,58 nM	30,4 nM	0,34 µM	1,11 nM
A-3qw6	1,50 nM	33,44 nM	1,20 nM	27,22 nM	3,03 nM	0,15 µM	43,20 nM	0,11 µM	0,15 µM	8,25 nM
A-3qw7	1,17 nM	3,13 nM	6,40 nM	43,90 nM	33,28 nM	0,76 µM	70,39 nM	0,12 µM	////	16,32 nM
A-3qw8	1,50 nM	17,04 nM	4,60 nM	84,89 nM	27,05 nM	0,12 µM	15,18 nM	28,42 nM	////	2,79 nM
A-3v0a	0,19 µM	6,49 nM	30,26 nM	6,01 nM	50,32 nM	0,35 µM	26,98 nM	59,98 nM	7,00 nM	10,59 nM
A-3v0b	1,18 nM	34,86 nM	79,84 nM	7,96 nM	2,90 nM	0,35 µM	6,07 nM	98,72 nM	0,16 µM	1,10 nM
A-3v0c	1,38 nM	35,69 nM	2,61 nM	11,51 nM	2,45 nM	1,65 µM	4,01 nM	1,09 µM	4,35 nM	7,80 nM
A-3zus	2,50 nM	20,18 nM	3,13 nM	2,01 nM	42,26 nM	30,46 nM	0,18 µM	71,59 nM	0,98 µM	2,14 nM
A-4ks6	3,20 nM	2,47 nM	1,05 nM	19,94 nM	4,73 nM	18,17 nM	2,83 nM	4,67 nM	0,27 µM	1,60 nM
A-4ktx	31,40 nM	4,97 nM	4,00 nM	2,88 nM	////	5,74 nM	4,11 nM	1,32 nM	////	1,44 nM
A-4kuf	////	2,66 nM	5,81 nM	15,37 nM	4,02 nM	10,3 nM	2,06 nM	5,22 nM	////	12,28 nM
B-1epw	1,68 nM	40,42 nM	12,16 nM	16,76 nM	1,08 nM	92,03 nM	37,64 nM	68,54 nM	4,00 nM	2,19 nM
B-1f31	2,06 nM	12,21 nM	1,10 nM	3,03 nM	1,35 nM	0,27 µM	75,55 nM	1,57 nM	2,17 nM	1,04 nM
B-1f82	1,50 nM	10,43 nM	4,92 nM	26,21 nM	0,15 µM	0,19 µM	0,26 µM	66,16 nM	24,42 nM	22,91 nM
B-1g9a	1,15 nM	1,56 nM	1,67 nM	27,83 nM	23,72 nM	0,16 µM	0,11 µM	0,22 µM	1,72 nM	1,83 nM
B-1g9b	1,78 nM	22,27 nM	2,45 nM	2,07 nM	1,92 nM	0,10 µM	0,46 µM	61,41 nM	5,28 nM	2,48 nM
B-1g9c	1,54 nM	9,99 nM	1,33 nM	5,93 nM	3,62 nM	77,58 nM	57,05 nM	0,24 µM	6,03 nM	1,51 nM
B-1g9d	2,18 nM	34,22 nM	6,35 nM	4,23 nM	5,79 nM	0,22 µM	59,00 nM	87,96 nM	1,33 nM	5,17 nM
B-1i1e	1,68 nM	2,14 nM	1,77 nM	10,15 nM	7,83 nM	46,97 nM	38,19 nM	0,11 µM	2,36 nM	1,81 nM
B-1s0d	2,77 nM	3,37 nM	1,82 nM	2,88 nM	1,37 nM	6,22 nM	10,93 nM	9,08 nM	5,42 nM	1,35 nM
B-1s0f	2,37 nM	5,79 nM	8,31 nM	16,33 nM	21,22 nM	0,38 µM	18,09 nM	0,15 µM	10,61 nM	2,12 nM
B-2xhl	1,39 nM	46,92 nM	4,50 nM	7,86 nM	1,13 nM	0,14 µM	0,11 µM	0,18 µM	0,13 µM	3,14 nM
B-2qn0	1,17 nM	31,26 nM	0,12 µM	////	3,58 nM	////	5,08 nM	////	////	10,73 nM
C-2fpq	52,50 nM	5,96 nM	31,36 nM	1,58 nM	5,13 nM	1,95 nM	1,02 nM	1,55 nM	////	2,96 nM
D-1t3a	1,28 nM	10,65 nM	5,94 nM	25,94 nM	48,01 nM	0,60 µM	7,61 nM	1,38 µM	64,54 nM	42,21 nM
E-1t3c	1,73 nM	1,25 nM	0,96 µM	1,38 nM	15,63 nM	////	2,01 nM	////	////	0,13 µM

Continued in next page

Continued from next page

BoNT zinc protease domains	Lig01	Lig02	Lig03	Lig04	Lig05	Lig06	Lig07	Lig08	Lig09	Lig10
E-1zl6	10,41 nM	0,25 µM	61,57 nM	0,65 µM	0,13 µM	2,84 µM	0,35 µM	0,98 µM	0,11 µM	0,21 µM
E-1zkw	1,31 nM	1,82 nM	3,04 nM	0,16 µM	1,28 nM	0,11 µM	65,91 nM	0,20 µM	85,46 nM	28,17 nM
E-1zkx	3,49 nM	1160 nM	0,52 µM	0,76 µM	0,10 µM	4,79 µM	2,43 nM	7,03 µM	0,50 µM	0,78 µM
E-1zn3	1,13 nM	0,20 µM	36,35 nM	58,13 nM	30,86 nM	1,02 µM	2,12 nM	1,66 µM	0,37 µM	1,23 nM
E-3ffz	1,83 nM	4,69 nM	5,95 nM	28,98 nM	0,11 µM	0,41 µM	0,15 µM	97,36 nM	25,32 nM	32,58 nM
F-2a8a	1,11 nM	10,75 nM	2,37 nM	85,31 nM	92,00 nM	0,33 µM	0,46 µM	4,67 µM	////	45,78 nM
F-2a97	1,61 nM	30,62 nM	10,30 nM	10,43 nM	14,18 nM	0,27 µM	53,61 nM	0,48 µM	61,45 nM	11,73 nM
F-3fie	1,02 nM	13,25 nM	8,18 nM	37,00 nM	2,13 nM	1,21 µM	42,54 nM	51,04 nM	0,57 µM	27,67 nM
F-3fii	3,81 nM	0,56 µM	1,37 nM	8,83 nM	86,82 nM	0,20 µM	84,86 nM	84,43 nM	23,50 nM	74,64 nM
H A_87	3,07 nM	11,75 nM	49,38 nM	11,35 nM	11,80 nM	0,15 µM	0,12 µM	0,18 µM	46,65 nM	2,37 nM
H A_92	1,19 nM	58,80 nM	89,51 nM	4,25 nM	2,35 nM	0,12 µM	0,21 µM	0,85 µM	2,39 nM	30,05 nM
H A_117	6,41 nM	0,36 µM	0,37 µM	41,63 nM	0,67 µM	////	////	////	0,45 µM	7,96 nM
H B_433	14,73 nM	80,65 nM	2,79 nM	10,12 nM	12,40 nM	0,23 µM	0,36 nM	13,34 nM	44,02 nM	1,48 nM
H F_357	12,75 nM	////	////	////	1,38 nM	0,44 µM	3,14 nM	1,18 nM	0,20 µM	////
flex 3c8a	13,42 nM	4,03 nM	1,36 nM	1,12 nM	3,85 nM	25,06 nM	27,45 nM	4,88 nM	10,92 nM	8,95 nM
flex 3c8b	2,38 nM	3,33 nM	13,50 nM	2,98 nM	27,07 nM	17,96 nM	1,59 nM	1,64 nM	0,27 µM	17,35 nM
flex 3fie	1,57 nM	0,11 µM	24,18 nM	37,62 nM	9,49 nM	0,57 µM	91,47 nM	0,18 µM	0,36 µM	6,94 nM
flex 3qw5	1,25 nM	2,27 nM	2,37 nM	8,58 nM	2,97 nM	3,71 nM	1,34 nM	3,18 nM	74,57 nM	1,98 nM
flex 3qw6	2,38 nM	2,79 nM	1,43 nM	2,68 nM	32,70 nM	2,74 nM	1,50 nM	55,42 nM	3,56 nM	3,69 nM
flex 3qw7	6,78 nM	4,22 nM	1,29 nM	12,27 nM	7,03 nM	6,68 nM	24,47 nM	25,47 nM	17,81 nM	6,76 nM
flex 3qw8	30,37 nM	6,53 nM	1,52 nM	1,18 nM	1,47 nM	2,67 nM	67,73 nM	1,98 nM	0,40 µM	2,16 nM
flex 4ks6	6,39 nM	3,33 nM	1,28 nM	27,97 nM	1,23 nM	2,49 nM	1,29 nM	1,32 nM	////	7,60 nM
flex 4ktx	1,44 nM	1,10 nM	4,31 nM	1,05 nM	10,53 nM	1,59 nM	1,08 nM	3,50 nM	////	1,02 nM

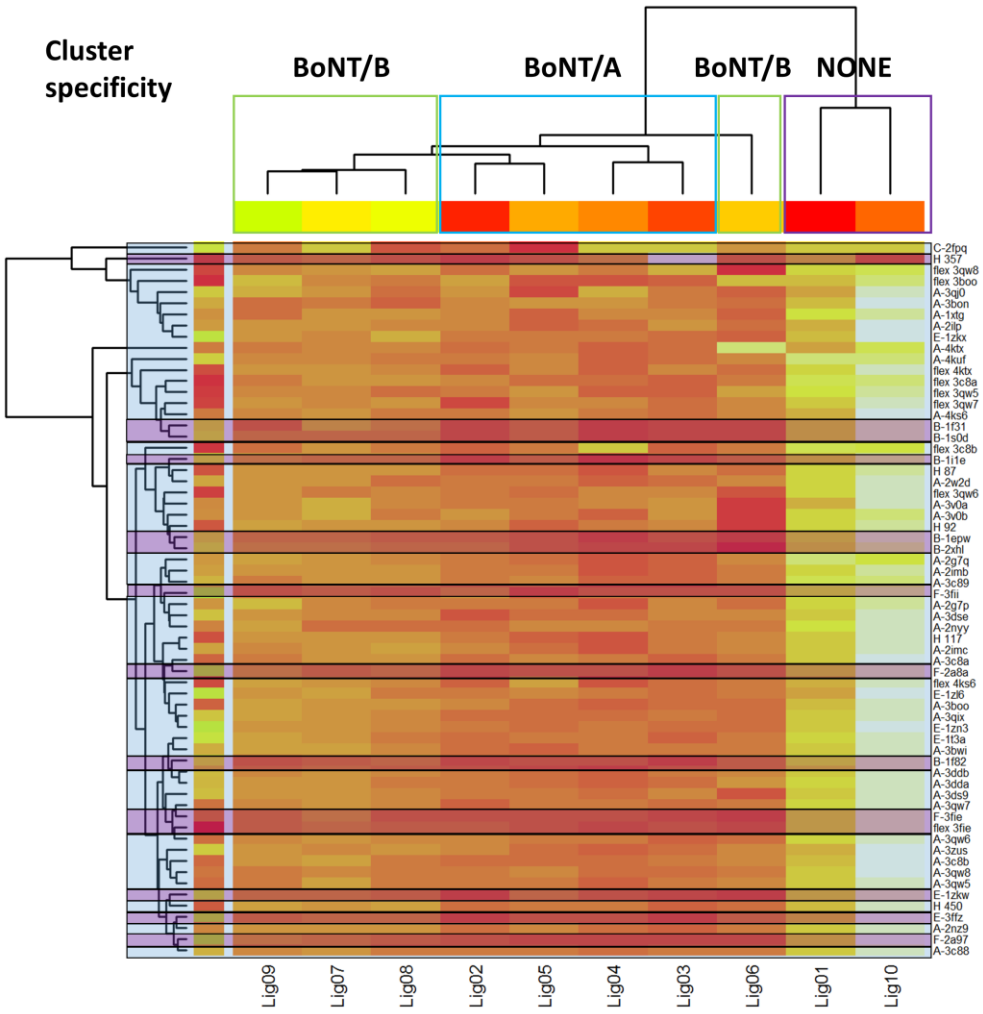


Figure 31: Estimated free energies of binding analysis heat map. Estimated free energies of binding are vertically and horizontally ordered. The horizontal dendrogram infers a ligand-oriented phylogeny (horizontally drown on heatmap top axis). The vertical dendrogram infers a BoNT-oriented phylogeny (vertically drown on heatmap left axis). From the ligand-oriented perspective ligands are grouped through their inhibitory potential in three main clusters holding a weak (purple top box) and BoNT/A (cyan top box) and BoNT/B (purple top box) affinity to the receptor. From the a BoNT-oriented phylogeny, even though there may be any coherence in BoNT serotyping, too many sub clusters are produced than BoNT types variety.

Estimated free energies of binding analysis. For all BoNT serotype structures, the estimated free energies of binding values are smoothly distributed throughout all ligand-receptor couples and it was not noticed a distinguished grouping.

Eventually it had been possible to infer a qualitative graphical clustering plotting the results' heatmap and dendrograms using R script (APPENDIX C). The produced outputs

are shown in Figure 31 and 24. Taking into account the estimated free energy of binding calculated for all generated poses, two dendrograms were produced, one to infer a ligand-oriented (horizontally drown on heatmap top axis) and a BoNT-oriented phylogeny (vertically drown on heatmap left axis). From the ligand-oriented perspective it is possible to group ligands inhibitory potential in two main clusters holding a weak and a high affinity to the receptor (highlighted in green, cyan and red boxes in Figure 31 and 24). The high affinity cluster shows a second level of clustering. The three groups hold a weak (red box in correspondence with lig01 and lig10 cells), mild (cyan box in correspondence with lig07, lig08 and lig09 cells) and high affinity for the receptor (red box in correspondence with lig02, lig03, lig04, lig05 and lig06 cells). When considering the zinc protease domain perspective, the phylogenies divides the structures into five main groups. In Figure 31 and 24 are color coded highlighted protein heatmap rows in relation to the biological target (green BoNT serotypes A, C, E targeting SNAP-25 and cyan BoNT serotypes B, D, F and G targeting VAMP).

All the predicted low affinity BoNT/inhibitor couples locate at the peripheral branch of each of the five main phylogenic clusters, suggesting, as expected a priori, that BoNT/A inhibitors bind selectively to BoNT/A related zinc protease domain structures.

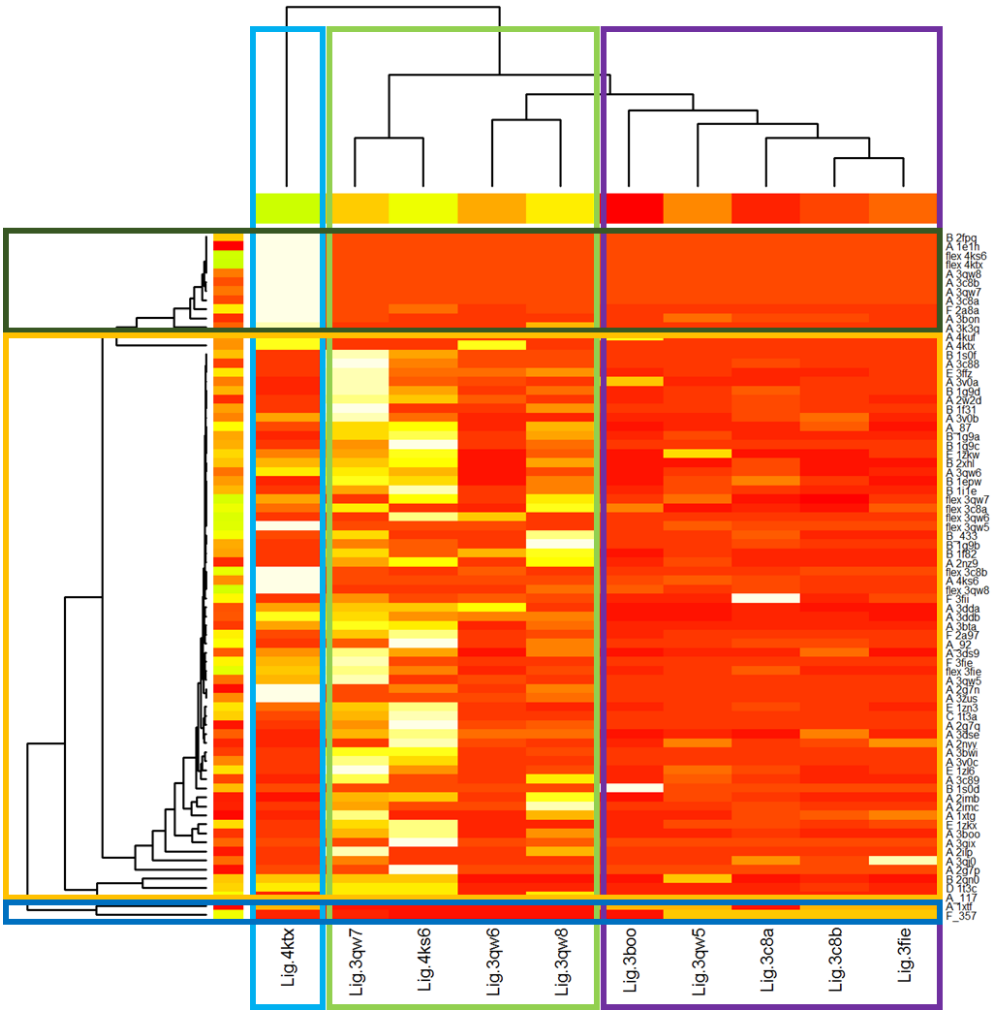


Figure 32: Estimated Inhibition Constant heatmap. A qualitatively evaluation of the Estimated inhibition constant can be made: *Horizontally*) three areas of inhibitory potential: low, medium high (respectively purple, light green and cyan boxes); Vertically three areas of receptor tolerance potential: low, medium high (respectively dark green, orange and blue boxes).

As stated, we performed flexible docking only to the re-docked BoNT/inhibitor couples since it was not noted substantial differences in comparison with rigid docking as concerns the ligand poses' zinc coordination features [148] (atom type – AT - involved in the coordination and Zn-AT distance reported in Table 15) the estimated values of free energies of binding (Table 13) and quality of the docking poses (Figure 33).

Table 15: Flexible docking ligand coordination chart. For every receptor-ligand couple are reported the ligand zinc-coordinating atom-type (AT) and the distance between the zinc coordinating atom and the catalytic zinc atom.

	lig01		lig02		lig03		lig04		lig05	
	d.	**AT**	**d.**	**AT**	**d.**	**AT**	**d.**	**AT**	**d.**	**AT**
3boo	1.85	N	0.81	N	0.92	N	0.50	N	1.10	O
3c8a	0.74	N	0.79	N	0.57	N	0.99	N	0.77	N
3c8b	0.86	N	0.54	N	0.76	N	1.02	N	0.65	N
3qw5	0.83	N	0.70	N	0.79	N	0.50	N	0.25	N
3qw6	0.58	N	0.79	N	0.68	N	1.07	N	0.38	N
3qw7	1.16	N	0.42	N	0.52	N	0.38	N	0.85	N
3qw8	0.83	N	1.02	N	0.95	N	1.09	N	0.65	N
4ks6	0.83	N	0.69	OA	0.62	OA	0.78	N	1.37	N
4ktx	1.20	N	0.59	N	0.93	N	0.96	N	1.25	N
3fie	1.18	OA	0.65	N	0.51	OA	1.00	N	1.21	OA
	lig06		**lig07**		**lig08**		**lig09**		**lig10**	
	d.	**AT**	**d.**	**AT**	**d.**	**AT**	**d.**	**AT**	**d.**	**AT**
3boo	0.30	N	0.90	N	1.20	OA	1.40	N	2.00	N
3c8a	0.40	N	0.44	N	1.36	OA	0.60	N	1.15	N
3c8b	1.18	N	0.95	N	0.71	N	0.73	N	1.98	OA
3qw5	0.29	N	0.80	OA	0.72	N	0.70	N	0.79	N
3qw6	0.34	N	0.86	N	1.08	N	0.68	N	1.67	N
3qw7	0.41	N	1.01	N	0.51	N	0.53	N	1.07	OA
3qw8	0.83	N	1.28	N	0.67	N	0.80	N	0.41	OA
4ks6	0.63	OA	0.49	OA	0.90	OA	0.63	N	//	//
4ktx	0.39	N	0.65	N	0.85	OA	1.01	N	//	//
3fie	0.40	N	1.28	N	1.17	N	0.87	OA	2.22	N

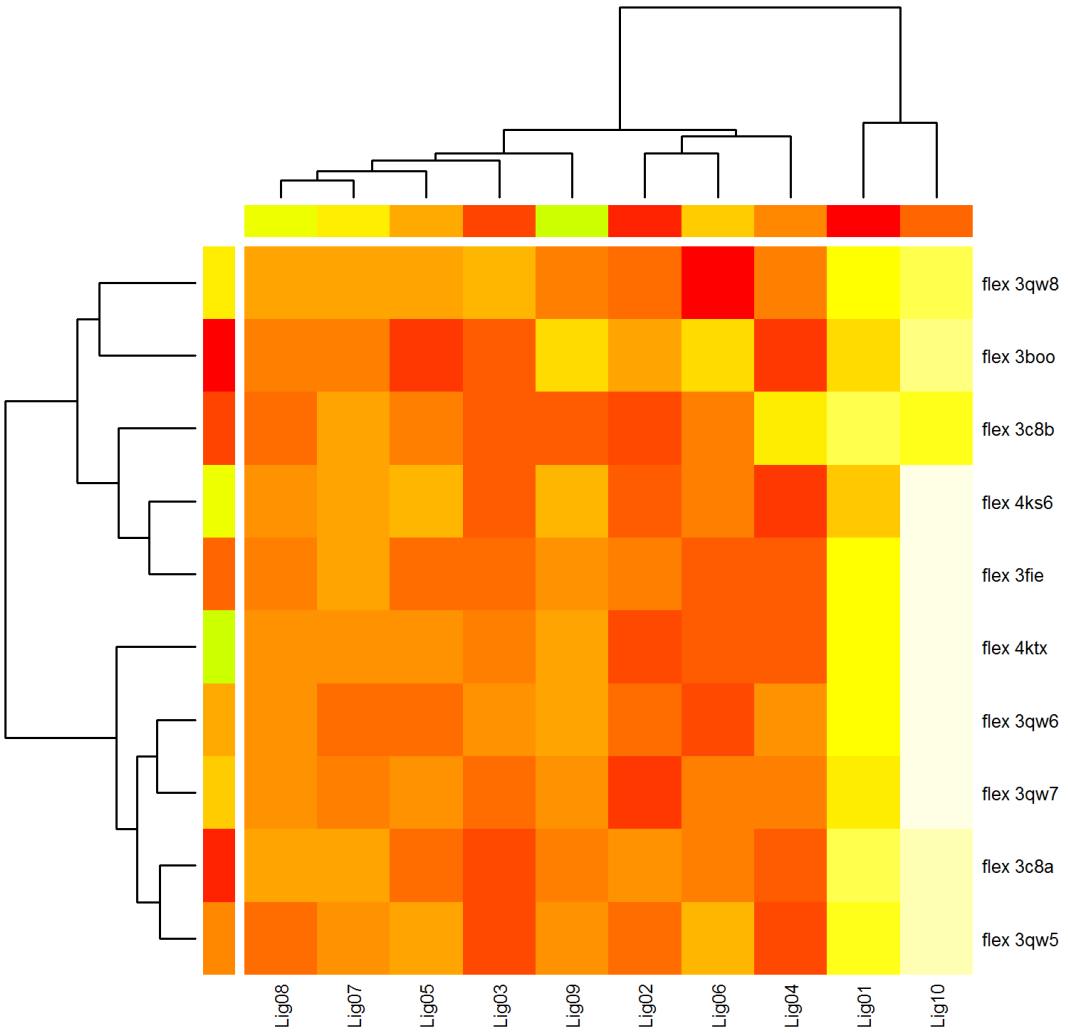


Figure 33: Flexible docking affinity energy heatmap.

As regards the virtually obtained zinc protease domains:

1) all docking calculations are in qualitative agreement with the once performed on the experimentally resolved structures (Table 13 and Table 14, Figure 31 and Figure 31);

2) all the five virtual structures respected the predicted clustering: PM0080396, PM0080398 and PM0080397 are correctly located in the high affinity BoNT/inhibitor cluster, while PM0080399 and PM00804XX in the low affinity BoNT/inhibitor cluster

(Figure 34 shows the 117 virtual model docked to the peptide inhibitors best ranked poses).

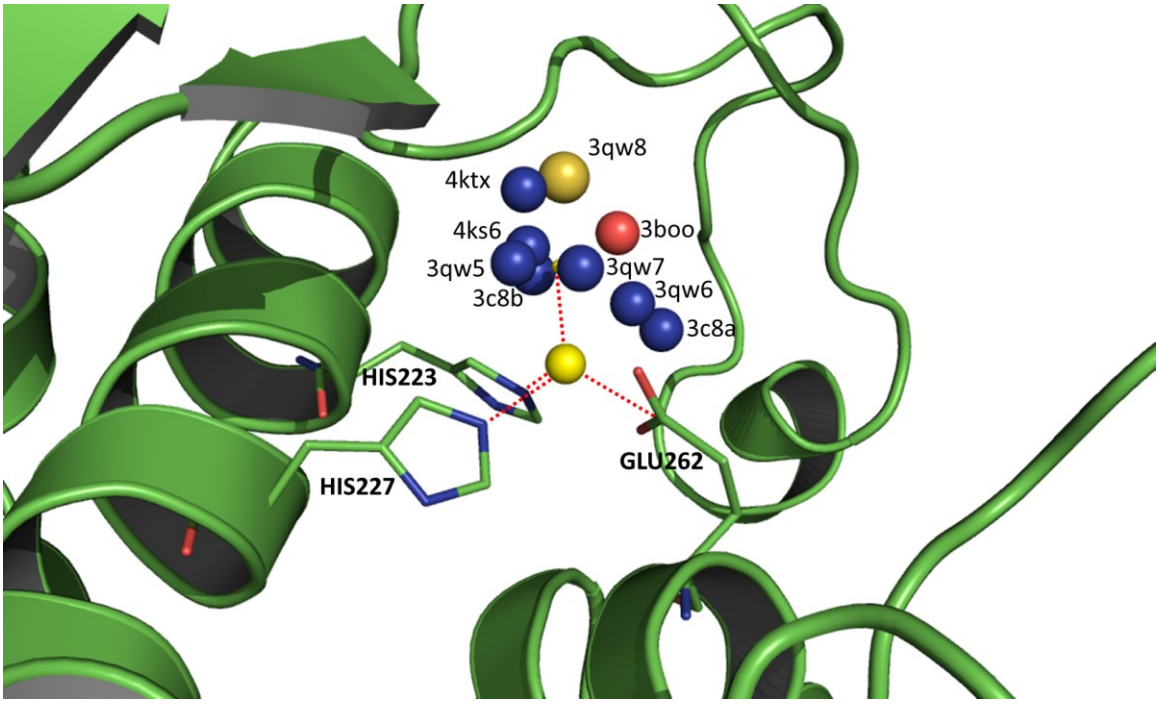


Figure 34: Docking poses of the catalytic domain of A117 virtual model. The zinc coordinating function is represented in sticks, (amino acidic side chains HIS 223, HIS227 and GLU263), spheres (yellow: Zinc and Tz dummy atoms; ligand coordinating atoms: blue for nitrogen, red for oxygen and gold for sulphur atom). The coordination is shown in red dotted lines. Docking qualitative results are coherent with the Autodock$_{ZN}$ reference publication **[148]**.

CHAPTER 6
CONCLUSIONS

Botulism is still a thoughtful medical issue, posing several concerns regarding the diagnosis and the therapy. For these reasons, it is valuable to develop and extended genetic characterization of different strains for epidemiology and forensics. To date genotyping studies of C. botulinum are limited in specific geographic areas and a broader survey is needed in order to gain a more robust description of C. botulinum genetic diversity. In the present study, to fill a gap in the current picture, we extended *C. botulinum* genome characterization to Italian representative strains, selected from previous MLVA typing studies. 10 Italian *C. botulinum* group I genomes were compared with 10 published genomes of the same group, none of which originating from Italy. Than we pursued the retrieval of whole genome genetic information of *C. botulinum* samples from ISS (Istituto Superiore di Sanità) collection in order to enrich the deposited genomic dataset to produce phylogenetic graphs for variability analysis. The extracted BoNT coding genes with Next Generation Sequencing methodology were compared with previously sequenced group I genomes, in order to genetically characterize the Italian population of *C. botulinum* group I and to investigate the phylogenetic relationships among different lineages. Genome sequencing of the strain 357 led us to identify a novel botulinum neurotoxin F8 subtype.

ClonalFrame results suggest that C. botulinum is a high recombining species, confirmed by the frequent conflicts in the dendrogram topologies of 150 different genes. Comparison between the clonal phylogeny and bont sequence phylogeny also hints frequent horizontal transfer of bont and in particular reveals a wide lateral diffusion of one B bont sequence among different lineages, probably by recombination. As reported in previous studies, bont and surrounding regions are frequently involved in recombination events and appear to be a hotspot of recombination. This implies that

genotyping based only on bont sequence is not sufficient to correctly represent the diversity among C. *botulinum* strains, but it is necessary to analyse and compare many genome elements. Further studies increasing the number of genomes could improve our current understanding of *C. botulinum* genetic dynamics in the future.

To date, less than 70 experimentally resolved BoNT structures (only 12 are composed by the four domains) are available for research purposes holding a resolution ranging from 2.0 to 4.0 Å. For this reason, a wider application of structural bioinformatics tools is a valuable investigative tool for broadening structurally derived knowledge in order to better understand the differences within Botulin serotype mechanism of action and target specificity, as Montecucco pointed out. That might lead to the identification of nouvelle inhibitors or the creation of monoclonal antibodies.

We modelled a total of nine unique BoNTs (3 A, 5 B and 1 F serotypes) holding five unique zinc domain protease (3 A, 1 B and 1 F serotypes). The new models hold a significantly different inter-serotype identity profile when considered in the entire structure and the zinc protease domain. This impacted to the binding affinity to known peptide inhibitors: as expected docking calculations clustered coherently within virtual and experimental zinc protease domains and the attributed serotype hypothetical behaviour. From the newly enriched BoNT virtual dataset, PM00804XX 3D structure modelled from a new BoNT/F8 subtype belonging to strain 357, showed a better affinity to lig10, part of the native inhibitor named INH1 of 3fie pdb structure (BoNT/F).

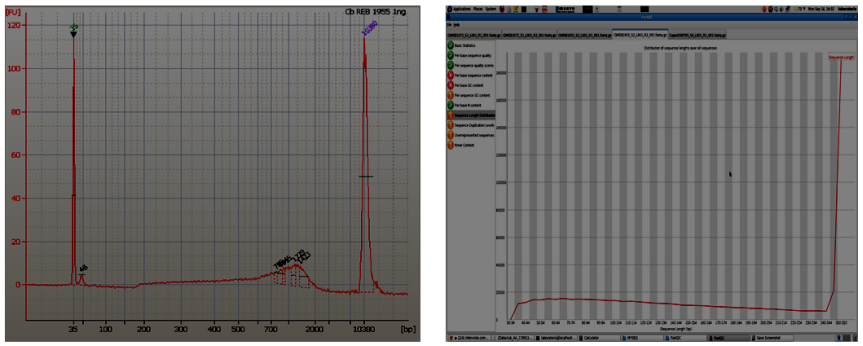
Moreover, we included all virtually designed BoNT catalytic domains in the set of the deposited experimentally resolved catalytic structures in order to perform docking calculations across selected peptide inhibitors from RCSB-PDB database deposited BoNT-peptide-inhibitor complex structures. The catalytic zinc-coordinated function was build using open source docking scripts tailored with specific zinc atom parameters, scoring function and ligand-receptor affinity estimate algorithm. The applied theoretical methods allowed us to produce experimental-level quality metal-coordinated models for all BoNT strains dataset. Most of the obtained docking poses were coherent with experimental homologous complex structures and docking estimates of binding affinity allowed us to

cluster the theoretically obtained model into the expected experimental serotype feature. Eventually, the virtual structure of the new Botulin neurotoxic subtype F8 was produced and its catalytic function has been tested with a set of known peptide inhibitors through molecular docking showing a native behavior towards BoNT/F specific peptide inhibitors.

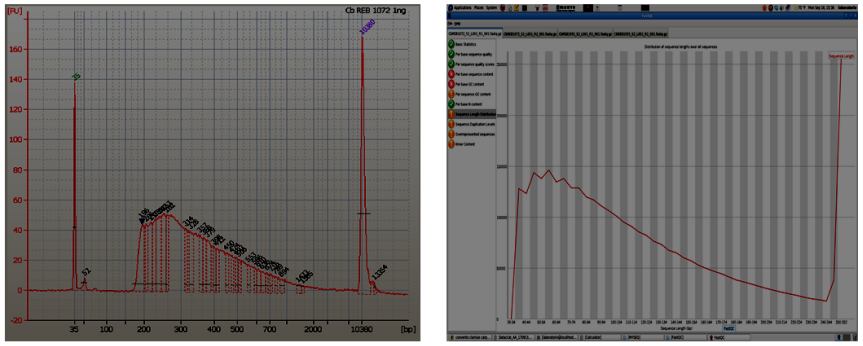
APPENDIX A

Supplementary Figures

C.botulinum_sample1 N. reads: 682'087



C.botulinum_sample2 N. reads: 3'409'043



C.botulinum_sample3 N. reads: 3'238'383

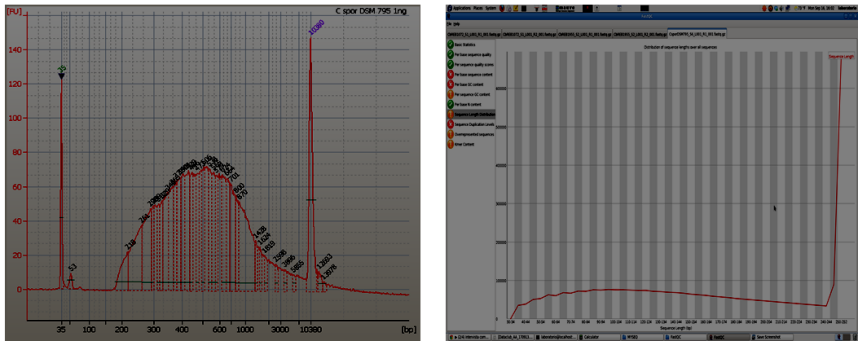


Figure 35: Quality reads.

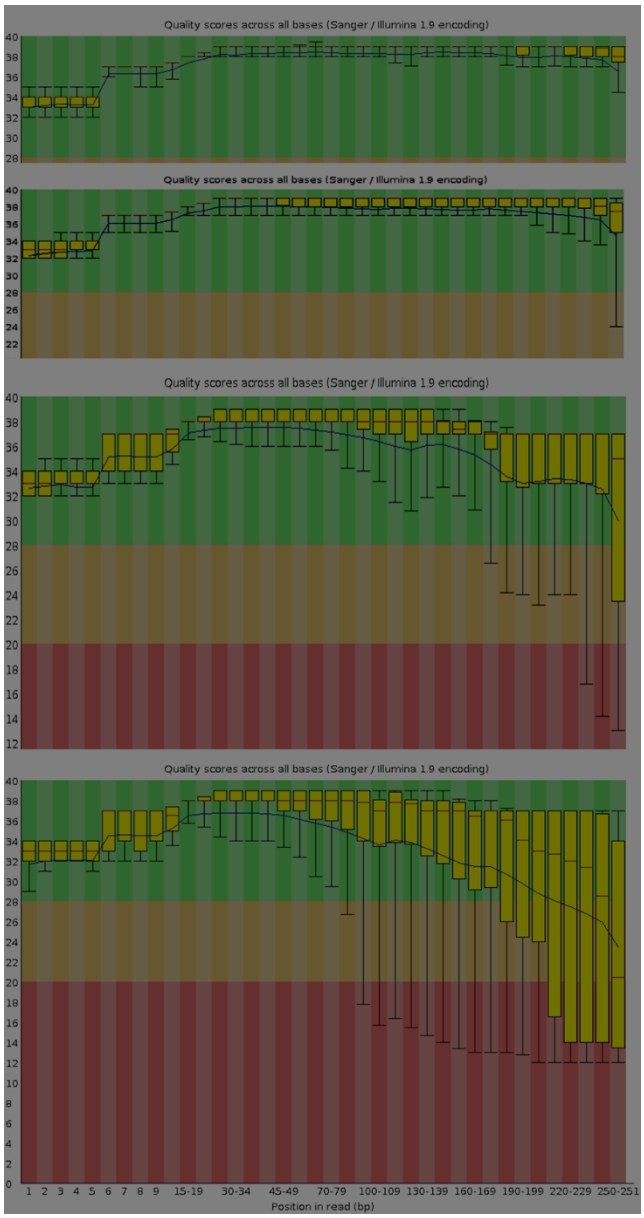


Figure 36: FastQC reads quality assessment.

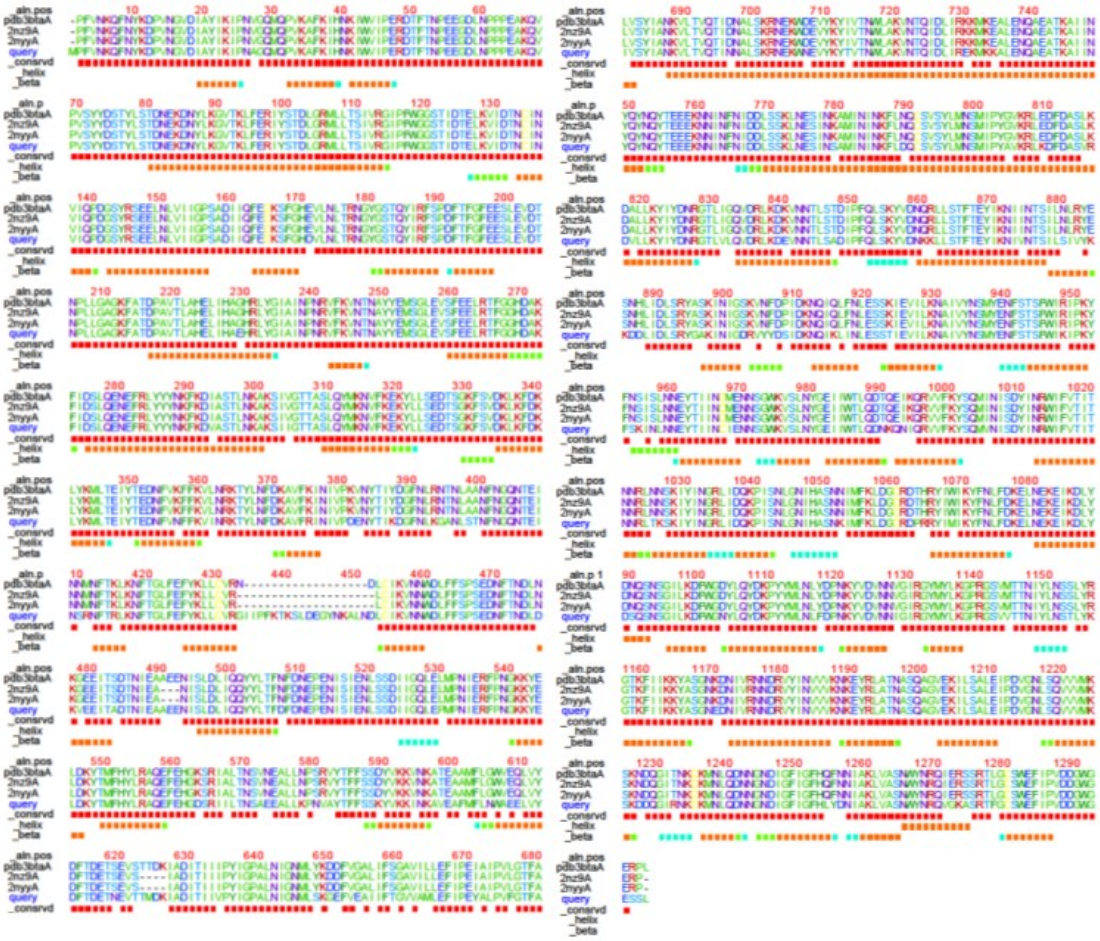


Figure 37: Sequence alignment for homology modelling procedures. The query, A2 117 BoNT aminoacidic sequence is aligned to the most similar experimetally resolved templates (pdbid 3bta, 2nz9, 2nyy).

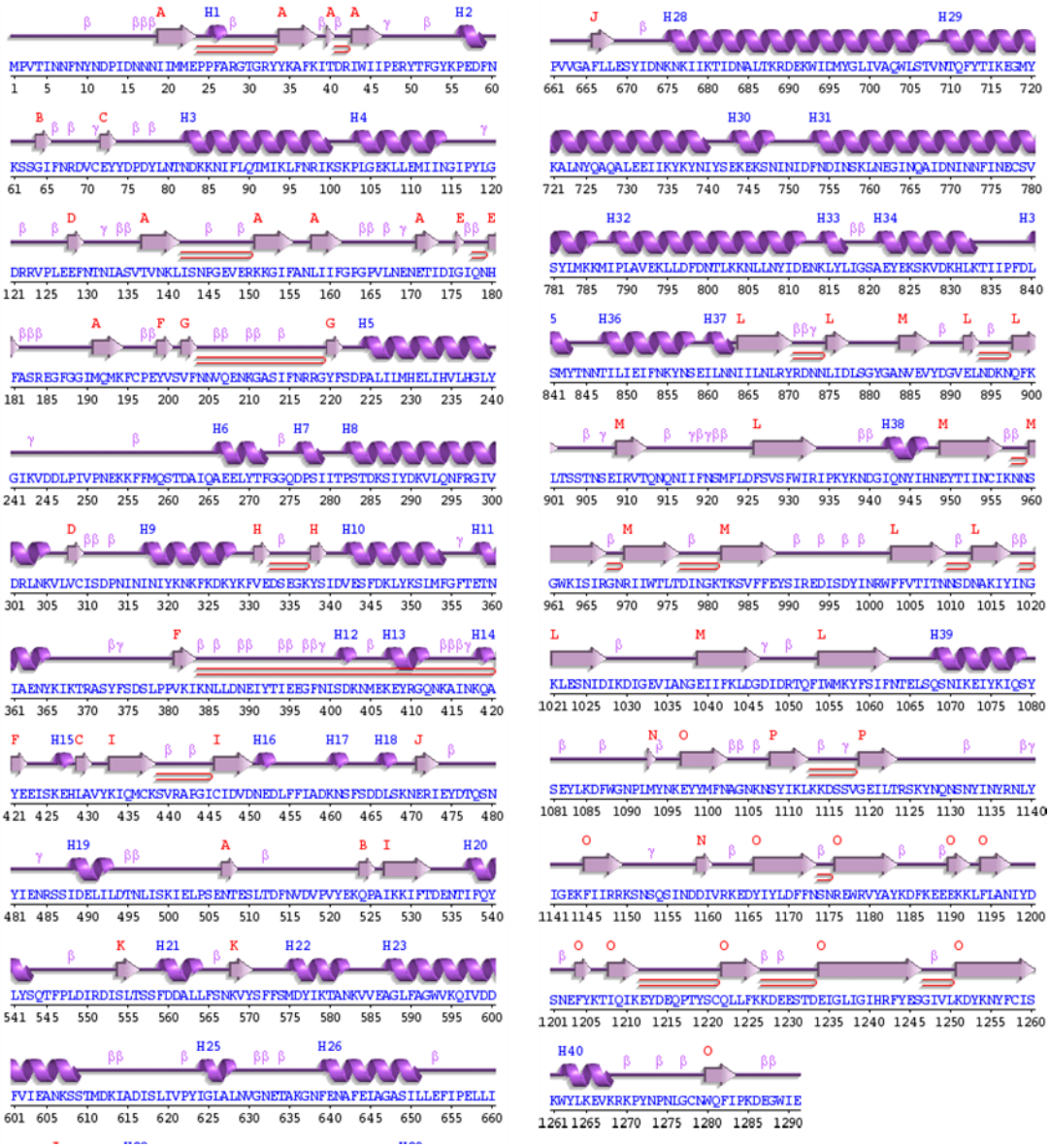


Figure 38: Predicted secondary structure of A2 117 BoNT aminoacidic sequence (output from ProCheck).

APPENDIX B

Supplementary Tables

Table 16: C. Botulinum genes.

	CDS is ATCC 3502	position in ATCC 3502 start	stop		gene name or family
1	**CBO0004**	2998	4101	*recF*	recombination protein
2	**CBO0016**	22071	23622	*serS-1*	seryl-tRNA synthetase
3	**CBO0053**	87028	88476		lysine decarboxylase
4	**CBO0065**	97129	98340	*ArcA1*	arginine deiminase
5	**CBO0070**	101129	101704	*spmA*	spore maturation protein
6	**CBO0099**	128105	131510	*ileS*	isoleucyl-tRNA synthetase
7	**CBO0150**	185799	186479	*atpA*	F0F1 ATP synthase
8	**CBO0173**	207254	209761	*secA*	preprotein translocase
9	**CBO0224**	264407	265780	*Rpo-N*	RNA polymerase factor sigma-54
10	**CBO0227**	268482	269678	*pgK*	phosphoglycerate kinase
11	**CBO0243**	287559	289178		vanadium haloperoxydase
12	**CBO0265**	311994	313391		sensor histidine kinase
13	**CBO0276**	325579	326793	*alr*	alanine racemase
14	**CBO0278**	329200	330525	*glvA-2*	maltose-6'-phosphate glucosidase
15	**CBO0294**	345062	346861	*ade-C*	adenine deaminase
16	**CBO0332**	384793	385392	*agrB*	Accessory gene regulator B
17	**CBO0368**	429001	430320	*rnfC*	electron transport complex, RnfABCDGE type,
18	**CBO0374**	434381	435469		cell surface protein
19	**CBO0399**	468193	470247		methyl-accepting chemotaxis protein
20	**CBO0419**	500538	500562	*aceK*	isocitrate dehydrogenase
21	**CBO0426**	508161	511613	*AddB*	ATP-dependent nuclease subunit B
22	**CBO0473**	559023	562553	*sbcC*	putative nuclease
23	**CBO0552**	649514	651505		methyl-accepting chemotaxis protein
24	**CBO0555**	653982	654383	*mscL*	large conductance mechanosensitive channel protein
25	**CBO0561**	660226	661308		spermidine-putrescine ABC transporter
26	**CBO0612**	710516	711655		membrane protein
27	**CBO0690**	792504	793253	*nifH*	nitrogenase iron protein
28	**CBO0698**	802053	803393	*gvcPA*	glycine dehydrogenase subunit 1
29	**CBO0700**	805041	806429	*lpdA*	dihydrolipoyl dehydrogenase
30	**CBO0727**	837449	839302	*chiA*	chitinase
31	**CBO0792**	894689	895528		sulfurtransferase family protein
32	**CBO0812**	919252	920553	*brnQ*	branched-chain amino acid transport system II carrier protein
33	**CBO0863**	967797	969239	*lipA*	secreted lipase
34	**CBO0914**	1019876	1021357	*cobQ*	cobyric acid synthase
35	**CBO0946**	1047486	1048202	*CbiL*	cobalt-precorrin-2 C(20)-methyltransferase
36	**CBO0972**	1071180	1071950	*FecE*	iron chelate uptake ABC transporter
37	**CBO1005**	1106422	1107945		L-lactate permease
38	**CBO1072**	1182001	1184183	*ArgS*	arginyl-tRNA synthetase
39	**CBO1081**	1190771	1192363		chitinase

Continued in next page

Continued from previous page

40	**CBO1095**	1204432	1205448	*gap2*	chitinase
41	**CBO1108**	1219729	1220595	*potB*	ABC transporter permease
42	**CBO1166**	1279602	1281245	*asdA*	aspartate aminotransferase
43	**CBO1192**	1306966	1310484		pyruvate ferredoxin oxidoreductase
44	**CBO1203**	1323006	1324634	*amyB*	beta-amylase
45	**CBO1230**	1349882	1350454	*cotJC*	spore coat protein
46	**CBO1231**	1350684	1351703		membrane protein
47	**CBO1263**	1376957	1378495	*grdC*	glycine reductase complex
48	**CBO1289**	1400879	1403074	*topB*	DNA topoisomerase III
49	**CBO1327**	1448902	1450932	*crr*	PTS system, glucose-specific, IIABC component
50	**CBO1354**	1478288	1480006		methyl-accepting chemotaxis protein
51	**CBO1355**	1480101	1480159		peptide/opine/nickel uptake ABC transporter
52	**CBO1379**	1502557	1503714		sensor histidine kinase
53	**CBO1380**	1503897	1504550		DNA-binding response regulator
54	**CBO1447A**	1579645	1580946		L-aspartate oxidase
55	**CBO1456**	1587887	1590067	*spoVD*	stage V sporulation protein D
56	**CBO1457**	1590297	1591748	*murE*	UDP-N-acetylmuramoylalanyl-D-glutamate--2
57	**CBO1473**	1604490	1605818	*aroA*	3-phosphoshikimate 1-carboxyvinyltransferase
58	**CBO1495**	1626744	1627766	*mtbaA*	methylcobalamin
59	**CBO1545**	1677739	1680375		cation transport ATPase
60	**CBO1583**	1713301	1714032		antibiotic ABC permease
61	**CBO1604**	1737074	1738738	*malL*	glycosyl hydrolase, family 13
62	**CBO1634**	1772528	1773919	*aco-L*	TPP-dependent acetoin dehydrogenase complex
63	**CBO1764**	1872833	1874212		peptide transporter
64	**CBO1773**	1880679	1882682		methyl-accepting chemotaxis protein
65	**CBO1778**	1888167	1889432	*rhlE*	ATP-dependent RNA helicase, DEAD/DEAH box family
66	**CBO1800**	1916663	1919461	*mutS*	DNA mismatch repair protein
67	**CBO1830**	1949755	1951479		hydrogenase
68	**CBO1872**	1998002	1997406	*spo0A*	sporulation transcription factor
69	**CBO1881**	2008527	2010395	*dxs*	1-deoxy-D-xylulose-5-phosphate synthase
70	**CBO1895**	2019747	2020670	*spoIIIAA*	stage III sporulation protein
71	**CBO1919**	2039662	2041371		methyl-accepting chemotaxis protein
72	**CBO1933**	2060032	2060715	*moeB*	molybdopterin biosynthesis protein
73	**CBO1967**	2100060	2101499		sensor histidine kinase
74	**CBO1976**	2110541	2111632	*gerAB2*	spore germination protein
75	**CBO1991**	2128609	2130006		PTS system, fructose family, IIBC component
76	**CBO2053**	2193170	2194309	*degT*	aminotransferase
77	**CBO2061**	2199690	2200979		secreted lipase
78	**CBO2108**	2248007	2249605	*eutE*	acetaldehyde dehydrogenase
79	**CBO2127**	2265748	2266167	*Vmra*	multidrug efflux pump
80	**CBO2149**	2292784	2294877		PTS system L-ascorbate specific IIABC sub
81	**CBO2162**	2304858	2306264		oligopeptide transport protein

Continued in next page

Continued from previous page

82	**CBO2194**	2344083	2345132	*etf*	electron transfer flavoprotein
83	**CBO2232**	2390454	2391902		aminoacyl-histidine dipeptidase
84	**CBO2244**	2403331	2404779		oxidoreductase
85	**CBO2264**	2420812	2422209		aminoacid-peptide transporter
86	**CBO2285**	2442057	2442764		DNA-binding response regulator
87	**CBO2405**	2532627	2533685	*recA*	recombinase
88	**CBO2418**	2548492	2550558	*infB*	translation initiation factor IF-2
89	**CBO2464**	2596125	2598239	*prd 1*	proline reductase complex 1
90	**CBO2480**	2614478	2616586	*prd 2*	proline reductase complex 2
91	**CBO2532**	2670357	2671130	*sigG*	sporulation sigma factor
92	**CBO2535**	2673034	2674143	*ftsZ*	cell division protein
93	**CBO2540**	2678553	2679301	*aroE*	shikimate dehydrogenase
94	**CBO2564**	2699897	2702536	*alaS*	alanyl-tRNA synthetase
95	**CBO2633**	2784118	2785317		membrane protein
96	**CBO2637**	2790985	2791776	*flgG*	flagellar basal body rod protein
97	**CBO2641**	2793525	2794253	*sigD*	RNA polymerase sigma factor for flagellar operon
98	**CBO2643**	2794941	2795801	*flhG*	flagellar biosynthesis protein
99	**CBO2646**	2799112	2800941	*FlhB*	bifunctional flagellar biosynthesis protein
100	**CBO2652**	2803750	2804562	*MotA*	chemotaxis protein
101	**CBO2661**	2811433	2812446	*fliG*	flagellar motor switch protein
102	**CBO2666**	2815880	2816734		Flagellin
103	**CBO2674**	2821503	2822936		cardiolipin synthase
104	**CBO2689**	2837248	2837925		capsular polysaccharide biosynthesis protein
105	**CBO2744**	2894735	2895730	*fliM*	flagellar motor switch protein
106	**CBO2748**	2897265	2899340	*cheA*	chemotaxis protein
107	**CBO2752**	2901730	2902176	*CheW*	chemotaxis protein
108	**CBO2770**	2930662	2932617	*gyrB*	gyrase B
109	**CBO2781**	2943254	2945320		AraC family transcriptional regulator
110	**CBO2811**	2976074	2978134		methyl-accepting chemotaxis protein
111	**CBO2832**	3005005	3007224		chitinase
112	**CBO2839**	3012881	3014323	*nagE*	N-acetylglucosamine-specific PTS system
113	**CBO2869**	3051223	3053868	*mgtA*	magnesium-translocating P-type ATPase
114	**CBO2911**	3102194	3103507	*ThiC*	thiamine biosynthesis protein
115	**CBO2938**	3126246	3127331	*rpoD*	RNA polymerase sigma factor
116	**CBO2952**	3143268	3143696	*gatB*	GatB/Yqey domain protein
117	**CBO2975**	3168726	3170633	*selB*	selenocysteine-specific translation elongation factor
118	**CBO3088**	3274552	3274974	*spoIIAB*	anti-sigma F factor
119	**CBO3094**	3280208	3281632	*wzx*	Membrane protein for export of O-antigen and teichoic acid
120	**CBO3095**	3282030	3283586	*MviN*	integral membrane protein
121	**CBO3096**	3283605	3284708		glycosyl transferase
122	**CBO3098**	3286060	3287292	*wzy*	exopolysaccharide biosynthesis family protein
123	**CBO3100**	3288615	3289673		UDP-N-acetylglucosamine 2-epimerase

Continued in next page

Continued from previous page

124	**CBO3106**	3295785	3296531		acetyltransferase
125	**CBO3112**	3307383	3308279	*gtaB*	UTP-glucose-1-phosphate uridylyltransferase
126	**CBO3149**	3350684	3351613	*oppB*	oligopeptide/dipeptide ABC transporter, permease protein
127	**CBO3163**	3363739	3365028	*folC*	folylpolyglutamate synthase/dihydrofolate synthase
128	**CBO3182**	3387637	3388443	*proC*	pyrroline-5-carboxylate reductase
129	**CBO3199**	3407583	3408722	*bcd*	butyryl-CoA dehydrogenase
130	**CBO3236**	3446235	3447131	*pyrD*	dihydroorotate dehydrogenase 1B
131	**CBO3298**	3508049	3509674	*groEL*	Thermic stress chaperonin
132	**CBO3308**	3520467	3521231		DNA-binding response regulator
133	**CBO3329**	3546922	3548511	*prfC*	peptide chain release factor 3
134	**CBO3358**	3588958	3591687		sigma-54 dependant transcriptional regulator
135	**CBO3359**	3591812	3593104	*celB*	PTS system, lactose/cellobiose family, IIC component
136	**CBO3373**	3607422	3608381	*pfkA*	6-phosphofructokinase
137	**CBO3420**	3660438	3661787	*glmM*	phosphoglucosamine mutase
138	**CBO3444**	3685798	3687414		Na/Pi-cotransporter family protein
139	**CBO3460**	3696382	3697032	*adk*	adenylate kinase
140	**CBO3461**	3697056	3698333	*SecY*	preprotein translocase
141	**CBO3466**	3699880	3700422	*rplF*	ribosomal protein L6
142	**CBO3488**	3716347	3720045	*rpoB*	DNA-directed RNA polymerase subunit beta
143	**CBO3497**	3725380	3725952	*sigH*	RNA polymerase factor sigma-70
144	**CBO3523**	3765668	3767341	*fhs*	Formate-tetrahydrofolate ligase
145	**CBO3542**	3787385	3788794		sensor histidine kinase
146	**CBO3543**	3788795	3789481		DNA-binding response regulator
147	**CBO3558**	3803553	3805349	*SonO*	methyl-accepting chemotaxis protein – nitric oxide sensor
148	**CBO3596**	3837995	3839750	*accC*	acetyl-CoA carboxylase
149	**CBO3616**	3857195	3858481	*purA*	Adenylosuccinate synthetase
150	**CBO3634**	3875071	3876126	*ytvI*	sporulation integral membrane protein

Table 17: 40 genes.

CDS	genome	gene name or family
CLJ_B0108	*Ba4 strain 654*	membrane spanning protein
CLD_0528	*B1 Okra*	bacitracin transport ATP-binding protein
CLD_0505	*B1 Okra*	vancomycin histidine protein kinase
CLM_0451	*A2 Kyoto*	putative cell wall-binding protease
CBO0432	*A1 ATCC 3502*	pyridine nucleotide-disulfide oxidoreductase family
CBO0463	*A1 ATCC 3502*	putative lipoprotein
CLK-3689	*A3 Loch Maree*	transmembrane NLP/P60 family protein
CLM_0623	*A2 Kyoto*	efflux ABC transporter permease
CBO0570	*A1 ATCC 3502*	CAAX amino terminal protease family protein
CBO626	*A1 ATCC 3502*	LytR family transcriptional regulator
CLD_0055	*B1 Okra*	ABC transporter permease family
CLK_0173	*A3 Loch Maree*	ABC transporter
CLK_0403	*A3 Loch Maree*	putative anion ABC transporter (modA)
CLJ_B1032	*Ba4 strain 654*	periplasmic sensor signal transduction histidine kinase
CLK_0481	*A3 Loch Maree*	M24 family peptidase
CLD_3223	*B1 Okra*	DNA binding response regulator
CBO1510	*A1 ATCC 3502*	peptide ABC transporter ATP binding protein
CBO1666	*A1 ATCC 3502*	methil accepting chemotaxis protein
CBO1775	*A1 ATCC 3502*	YbaK/prolyl-tRNA synthetase domain protein
CLM_2001	*A2 Kyoto*	carB
CLJ_B2157	*Ba4 strain 654*	hydrolase
CLK_1451	*A3 Loch Maree*	esterase/lipase
CLJ_B2234	*Ba4 str. 654*	lantibiotic ABC transporter
CBO2034	*A1 ATCC 3502*	GntR family transcriptional regulator
CBO2585	*A1 ATCC 3502*	putative phosphoenolpyruvate synthase
CLM_3205	*A2 Kyoto*	ABC transporter ATP-binding protein
CLJ_B3078	*Ba4 strain 654*	Sorbose-specific PTS system transporter subunit 2C
CLD_1717	*B1 Okra*	PTS system beta-glucoside-specific family
CLM_3249	*A2 Kyoto*	peptidase family S8/S53
CLD_1670	*B1 Okra*	spore germination protein
CLD_1637	*B1 Okra*	Sigma 54 dependant transcriptional regulator
CBO2921	*A1 ATCC 3502*	ribB
CBO3167	*A1 ATCC 3502*	ABC transporter ATP-binding protein
CBO3247	*A1 ATCC 3502*	ABC nitrate transporter
CLM_3791	*A2 Kyoto*	putative lipoprotein
CLM_3815	*A2 Kyoto*	putative flavoredoxin
CLM_3821	*A2 Kyoto*	NAD-dependent epimerase/dehydratase
CLK_3050	*A3 Loch Maree*	glycine/sarcosine/betaine reductase, component B, subunit...
CLD_3006	*B1 Okra*	ABC transporter, ATP-binding protein
CLD_0539	*B1 Okra*	MepB family protein

Table 18: Roche 454 assembly (Newbler).

Genome	contigs	contigs > 500 bps	N50	tot. Bps contigs > 500 bps
B2 267	95	78	122.783	3.898.280
B2 433	86	72	128.719	4.125.175
B2 128	87	67	106.454	3.846.884
A2 117	166	147	47.761	3.806.199
B2 331	237	207	35.260	3.795.475
A2B7 92	120	107	80.531	4.058.381
B2 275	160	137	64.044	3.977.447
F 357	345	319	19.314	3.785.437
A2B2 87	56	42	254.455	4.175.822
B2 450	163	142	78.650	4.303.998
Cb89	289	265	31.140	4.143.155
Cb129	109	89	128.782	4.022.915
Cb130	349	325	21.891	3.991.088
Cb222	253	231	29.803	3.836.259
Cb270	607	559	11.317	3.649.984
Cb279	107	90	78.102	3.845.690
Cb356	156	132	64.638	4.072.906
Cb659	149	132	66.608	3.818.631
cbut190	222	197	52.988	4.642.924
Cbut 86	198	171	61.216	4.612.014

Table 19: Illumina Myseq assembly (Abyss).

Genome	contigs	contigs > 500 bps	N50	tot. Bps contigs > 500 bps
B2 267	1.264	524	11.883	3.824.558
B2 128	374	87	100.819	3.853.909
A2 117	594	285	24.808	3.798.799
B2 331	297	86	80.929	3.844.089
A2B7 92	1.365	543	12.087	4.003.454
B2 275	616	251	27.757	3.976.409
F 357	489	146	53.369	3.849.714
A2B2 87	449	109	76.563	4.175.327
B2 450	497	139	64163	4323523
Cb89	1.651	1.043	5.874	3.782.927
Cb129	1.191	838	7.594	3.811.321
Cb130	379	142	73.222	4.035.466
Cb270	1.163	164	41.569	3.763.489
Cb356	504	111	77.032	4.090.761
cb1-3b	301	55	210.877	4.107.312
cb104	1.204	784	7.619	3.741.255
cb130	379	142	73.222	4.035.466
cb356	504	111	77.032	4.090.761
cb644	1.235	874	7.666	4.016.569
cb656	599	104	80.358	3.830.406
cb89 AB	1.651	1.043	5.874	3.782.927
cbibt2267	280	65	153919	3835628
cbibt2272f	330	122	63.449	3.947.537
cbibt2293b	947	178	42.020	3.986.751
cbibt2299b	598	312	22.927	4.066.216
cbreb1072	619	211	37.909	3.938.957
cbreb1955	4.046	1.437	7.018	4.579.594
cbreb83b	1.475	954	6.086	3.650.078
cbspor885	330	93	90.037	4.111.125
cbspor925	2148	374	62591	4088323
cbsporC1	1.325	845	7.221	3.845.240
cbsporDSM795	433	102	89.401	4.105.331
cbut109	1.459	1.053	6.546	4.364.189
cbut145	340	177	55.814	4.523.893
cbut182	465	291	29.148	4.476.780
cbut184	446	154	54.709	4.613.878
cbut190	379	182	65.911	4.668.945

Table 20: Identity matrix.

Whole BoNT

	450	**433**	**331**	**267**	**128**	**275**	**357**	**3BTA**	**92**	**117**	**87**
450	**100**	98.06	97.99	97.91	97.91	97.83	40.4	40.32	39.33	39.25	39.1
433	98.06	**100**	99.46	99.38	99.38	99.3	40.56	40.4	39.41	39.33	39.18
331	97.99	99.46	**100**	99.92	99.92	99.85	40.48	40.4	39.41	39.33	39.18
267	97.91	99.38	99.92	**100**	**100**	99.92	40.56	40.48	39.49	39.41	39.25
128	97.91	99.38	99.92	**100**	**100**	99.92	40.56	40.48	39.49	39.41	39.25
275	97.83	99.3	99.85	99.92	99.92	**100**	40.48	40.4	39.41	39.33	39.18
357	40.4	40.56	40.48	40.56	40.56	40.48	**100**	40.96	40.76	40.61	40.45
3bta	40.32	40.4	40.4	40.48	40.48	40.4	40.96	**100**	89.58	89.96	89.81
92	39.33	39.41	39.41	39.49	39.49	39.41	40.76	89.58	**100**	97.92	98.3
117	39.25	39.33	39.33	39.41	39.41	39.33	40.61	89.96	97.92	**100**	99.38
87	39.1	39.18	39.18	39.25	39.25	39.18	40.45	89.81	98.3	99.38	**100**

CATHALYTIC DOMAIN

	3bta	**92**	**87**	**117**	**450**	**433**	**331**	**267**	**128**	**275**	**F357**
3bta	**100**	95.09	95.58	95.58	33.09	33.09	33.09	33.09	33.09	33.09	32.91
92	95.09	**100**	98.28	98.03	33.33	33.33	33.33	33.33	33.33	33.33	32.66
87	95.58	98.28	**100**	99.75	33.09	33.09	33.09	33.09	33.09	33.09	32.41
117	95.58	98.03	99.75	**100**	33.09	33.09	33.09	33.09	33.09	33.09	32.41
450	33.09	33.33	33.09	33.09	**100**	**100**	**100**	**100**	**100**	**100**	39.15
433	33.09	33.33	33.09	33.09	**100**	**100**	**100**	**100**	**100**	**100**	39.15
331	33.09	33.33	33.09	33.09	**100**	**100**	**100**	**100**	**100**	**100**	39.15
267	33.09	33.33	33.09	33.09	**100**	**100**	**100**	**100**	**100**	**100**	39.15
128	33.09	33.33	33.09	33.09	**100**	**100**	**100**	**100**	**100**	**100**	39.15
275	33.09	33.33	33.09	33.09	**100**	**100**	**100**	**100**	**100**	**100**	39.15
CS	32.91	32.66	32.41	32.41	39.15	39.15	39.15	39.15	39.15	39.15	**100**

TRANSLOCATION DOMAIN

	357	**450**	**433**	**331**	**267**	**128**	**275**	**3bta**	**92**	**87**	**117**
357	**100**	48.26	48.26	48.26	48.26	48.26	48.26	42.9	43.22	42.27	42.27
450	48.26	**100**	**100**	**100**	**100**	**100**	**100**	52.68	49.84	49.21	49.21
433	48.26	**100**	**100**	**100**	**100**	**100**	**100**	52.68	49.84	49.21	49.21
331	48.26	**100**	**100**	**100**	**100**	**100**	**100**	52.68	49.84	49.21	49.21
267	48.26	**100**	**100**	**100**	**100**	**100**	**100**	52.68	49.84	49.21	49.21
128	48.26	**100**	**100**	**100**	**100**	**100**	**100**	52.68	49.84	49.21	49.21
275	48.26	**100**	**100**	**100**	**100**	**100**	**100**	52.68	49.84	49.21	49.21
3bta	42.9	52.68	52.68	52.68	52.68	52.68	52.68	**100**	84.54	84.23	84.54
92	43.22	49.84	49.84	49.84	49.84	49.84	49.84	84.54	**100**	96.85	97.16
87	42.27	49.21	49.21	49.21	49.21	49.21	49.21	84.23	96.85	**100**	99.68
117	42.27	49.21	49.21	49.21	49.21	49.21	49.21	84.54	97.16	99.68	**100**

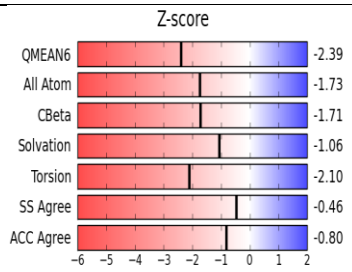
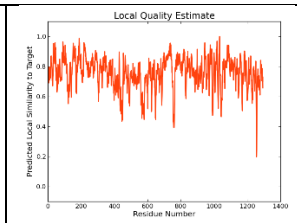
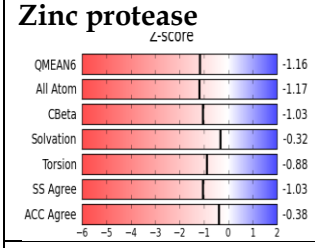
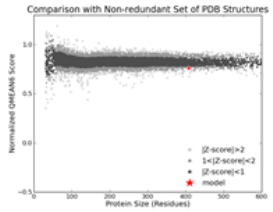
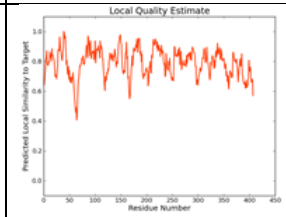
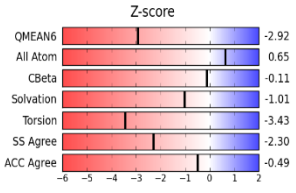
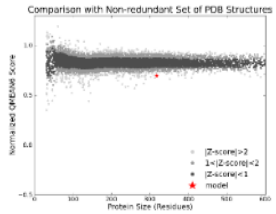
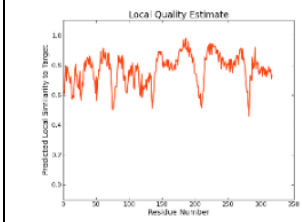
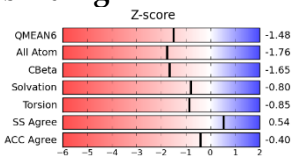
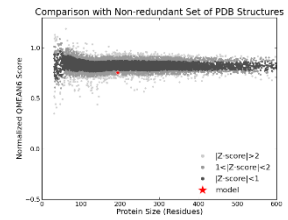
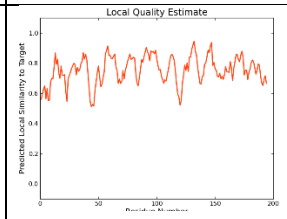
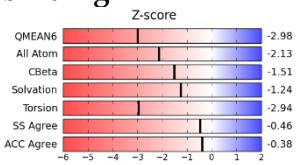
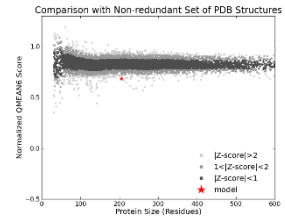
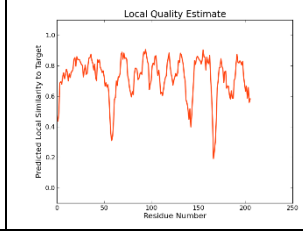
N-TERMINUS DOMAIN

	267	**128**	**275**	**450**	**433**	**331**	**357**	**3bta**	**92**	**87**	**117**
267	**100**	**100**	**100**	98.97	99.48	99.48	48.42	51.31	51.83	51.83	51.83
28	**100**	**100**	**100**	98.97	99.48	99.48	48.42	51.31	51.83	51.83	51.83
275	**100**	**100**	**100**	98.97	99.48	99.48	48.42	51.31	51.83	51.83	51.83
450	98.97	98.97	98.97	**100**	99.48	99.48	47.89	50.79	51.31	51.31	51.31
433	99.48	99.48	99.48	99.48	**100**	**100**	47.89	50.79	51.31	51.31	51.31
331	99.48	99.48	99.48	99.48	**100**	**100**	47.89	50.79	51.31	51.31	51.31
357	48.42	48.42	48.42	47.89	47.89	47.89	**100**	60.62	59.59	60.1	60.1
3bta	51.31	51.31	51.31	50.79	50.79	50.79	60.62	**100**	85.05	85.57	85.57
92	51.83	51.83	51.83	51.31	51.31	51.31	59.59	85.05	**100**	99.48	99.48
87	51.83	51.83	51.83	51.31	51.31	51.31	60.1	85.57	99.48	**100**	**100**
117	51.83	51.83	51.83	51.31	51.31	51.31	60.1	85.57	99.48	**100**	**100**

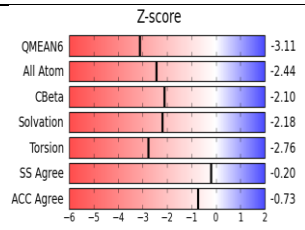
C-TERMINUS DOMAIN

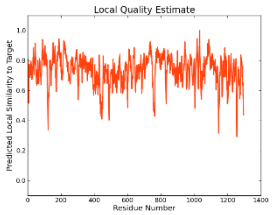
	450	**433**	**331**	**267**	**128**	**275**	**357**	**3bta**	**117**	**92**	**87**
450	**100**	89.77	89.3	89.3	89.3	88.84	31.58	33.67	33.67	32.16	32.66
433	89.77	**100**	96.74	96.74	96.74	96.28	33.16	33.67	33.67	32.16	32.66
331	89.3	96.74	**100**	**100**	**100**	99.53	32.63	33.67	33.67	32.16	32.66
267	89.3	96.74	**100**	**100**	**100**	99.53	32.63	33.67	33.67	32.16	32.66
128	89.3	96.74	**100**	**100**	**100**	99.53	32.63	33.67	33.67	32.16	32.66
275	88.84	96.28	99.53	99.53	99.53	**100**	32.11	33.17	33.17	31.66	32.16
CS	31.58	33.16	32.63	32.63	32.63	32.11	**100**	43.46	44.5	44.5	43.98
3bta	33.67	33.67	33.67	33.67	33.67	33.17	43.46	**100**	91.71	90.73	90.73
117	33.67	33.67	33.67	33.67	33.67	33.17	44.5	91.71	**100**	96.59	97.56
92	32.16	32.16	32.16	32.16	32.16	31.66	44.5	90.73	96.59	**100**	98.54
87	32.66	32.66	32.66	32.66	32.66	32.16	43.98	90.73	97.56	98.54	**100**

Quality estimate tables

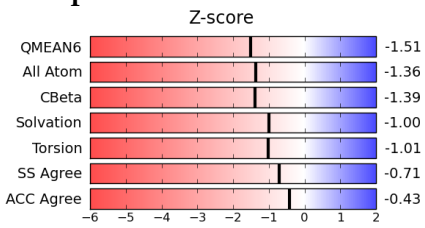
117 **GMQE** 0.97 **Ligand**: Zn in contact with H223, H227, E262		
Zinc protease 		
traslocation domain 		
N-terminal receptor binding 		
C-terminal receptor binding 		

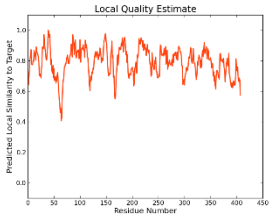
A2B3_87
GMQE 0.96
Ligand: Zn in contact with H223, H227, E262



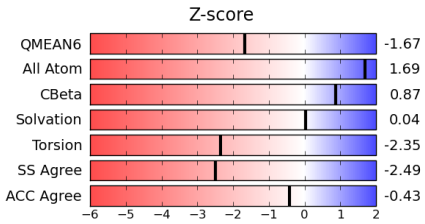


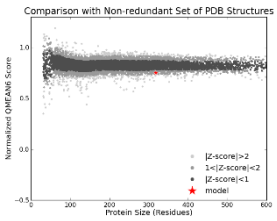
Zinc protease

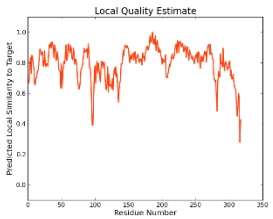




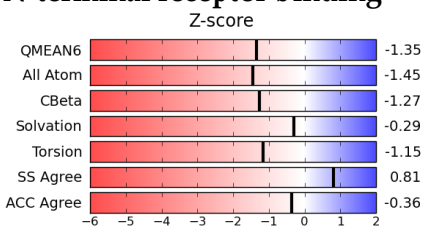
traslocation domain

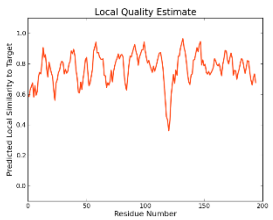




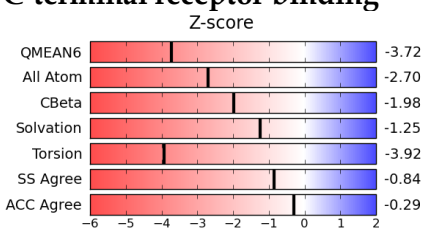


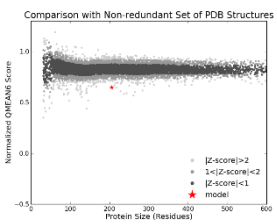
N-terminal receptor binding

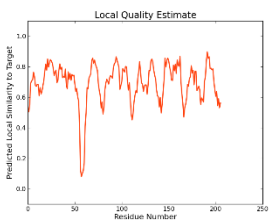


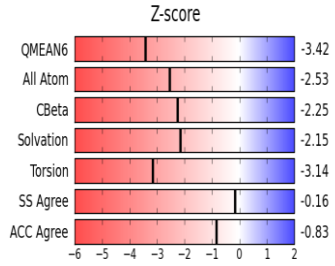
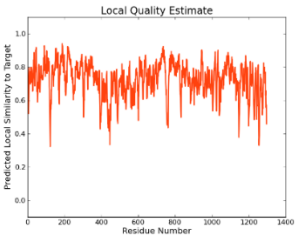
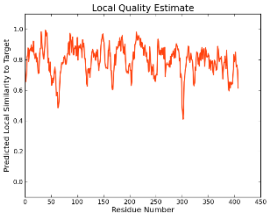
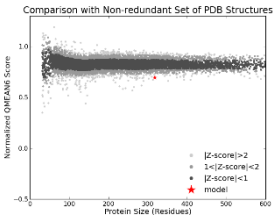
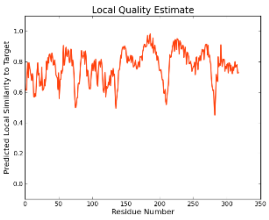
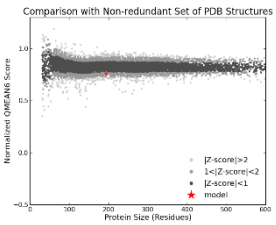
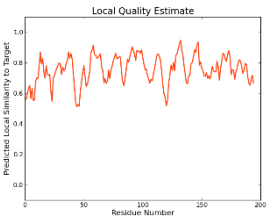
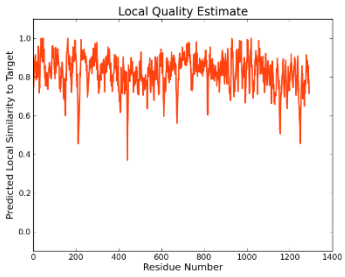
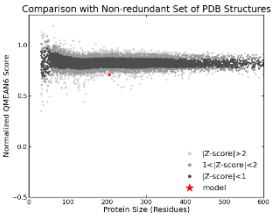
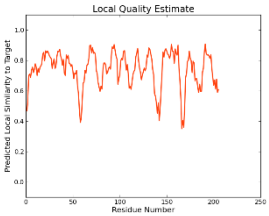
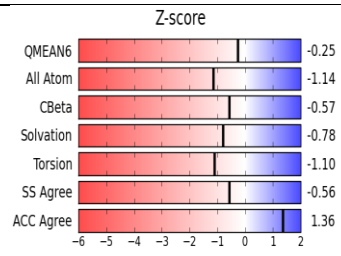
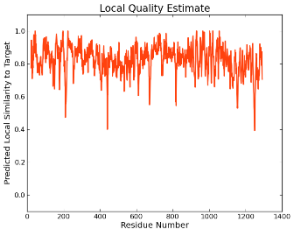


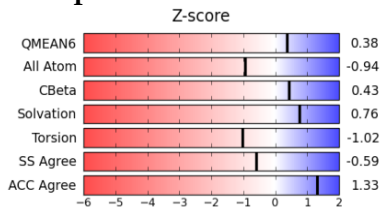
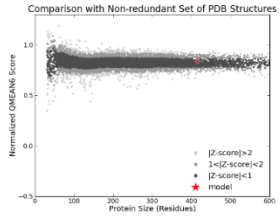
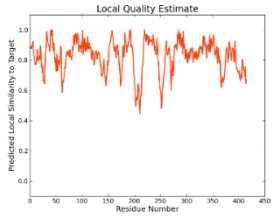
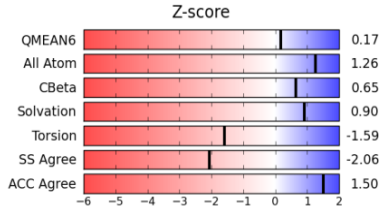
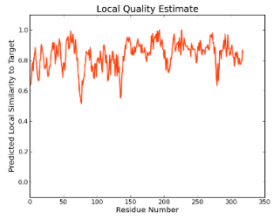
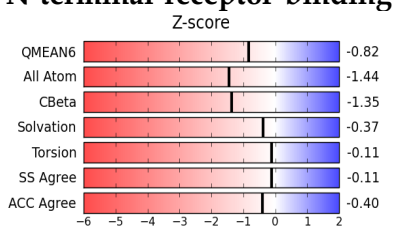
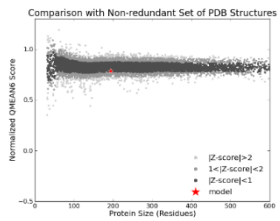
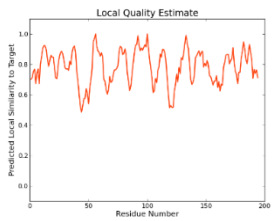
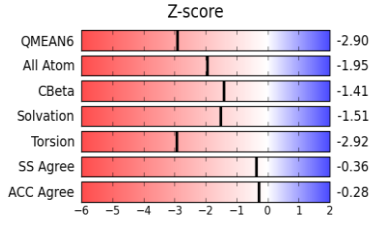
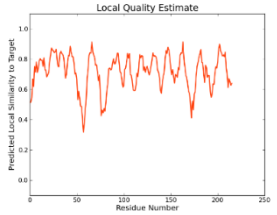
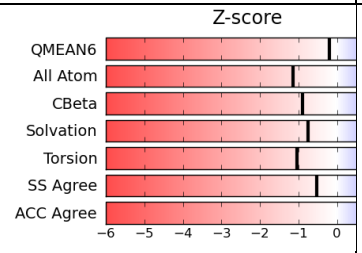
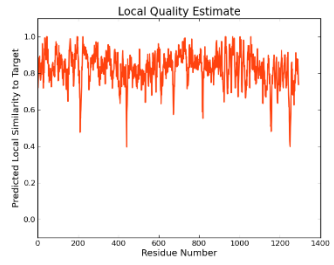
C-terminal receptor binding

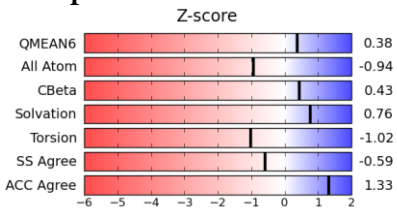
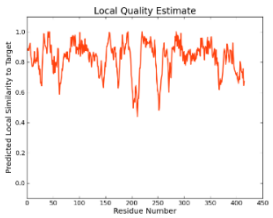
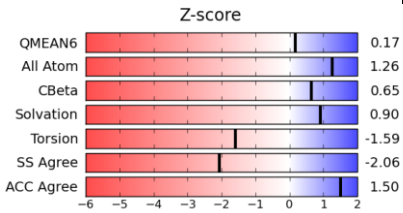
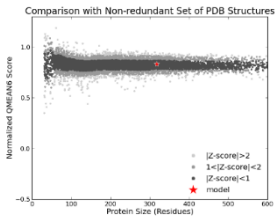
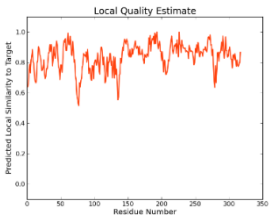
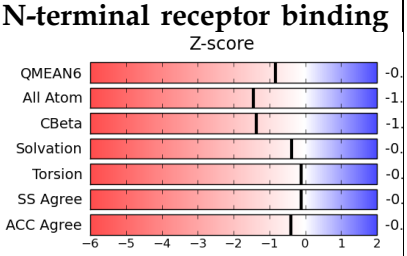
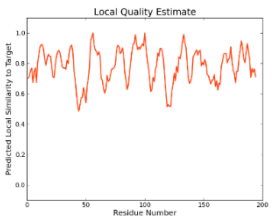
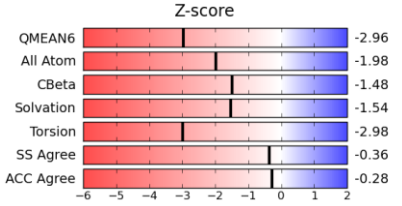
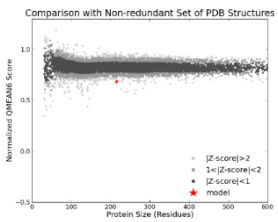
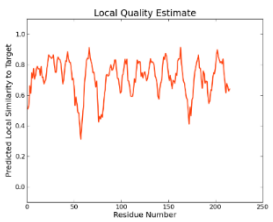


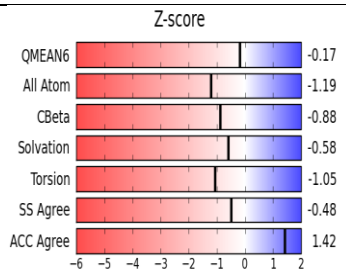
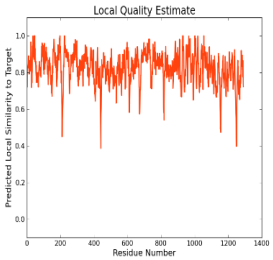
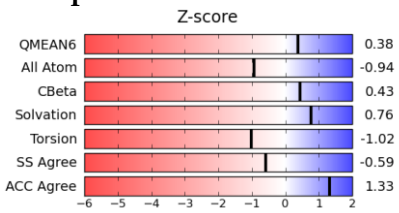
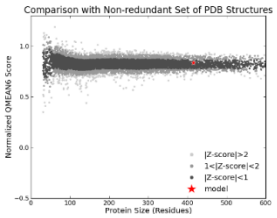
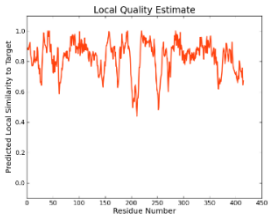
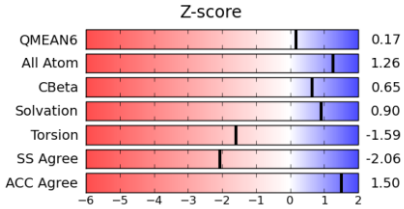
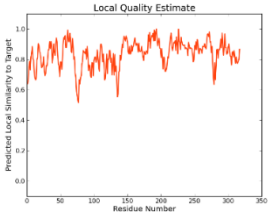
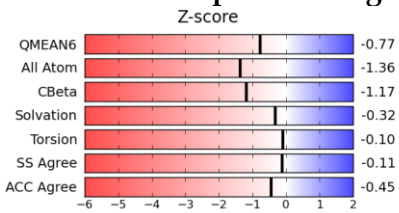
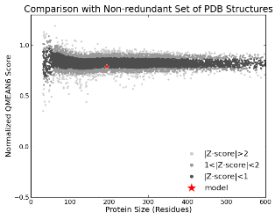
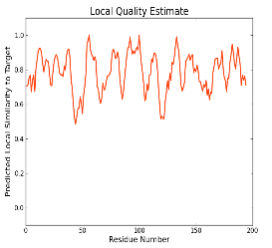
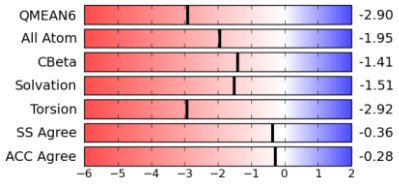
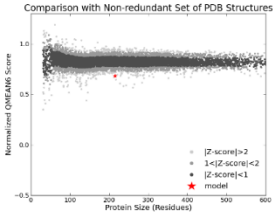
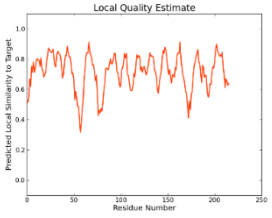


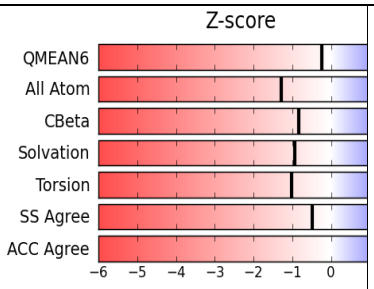
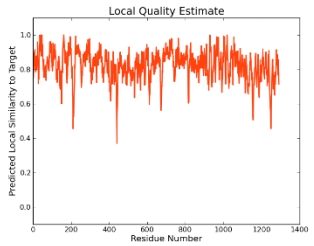
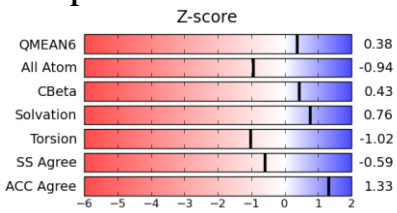
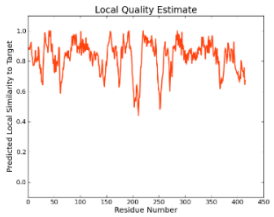
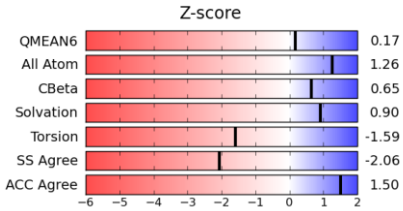
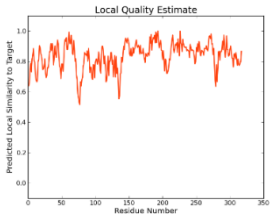
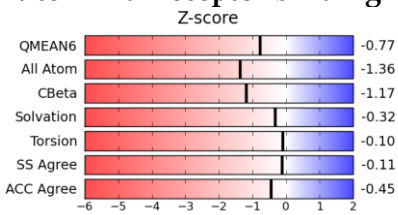
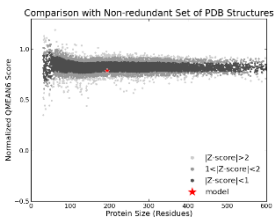
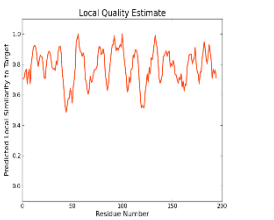
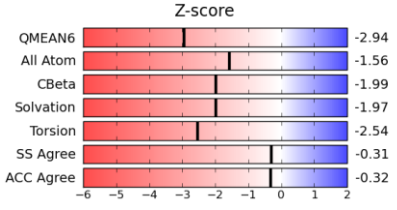
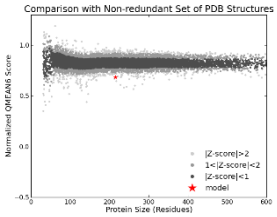
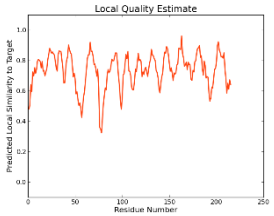


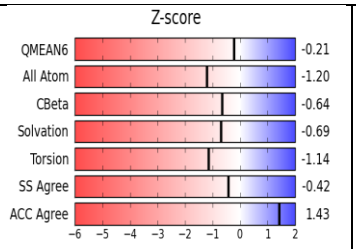
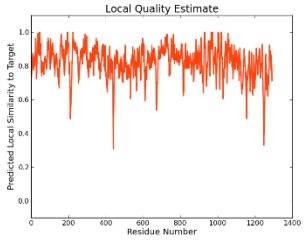
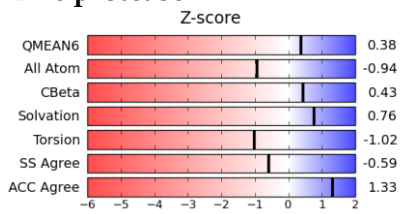
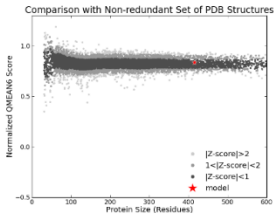
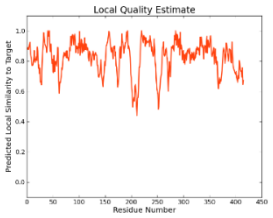
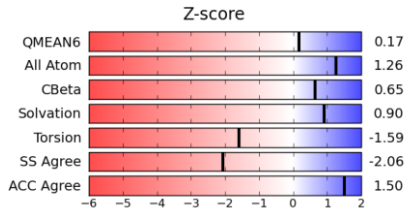
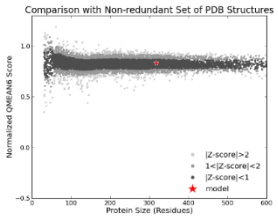
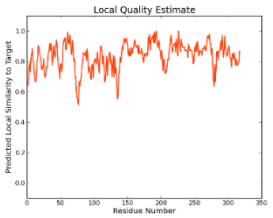
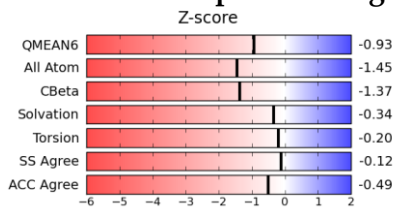
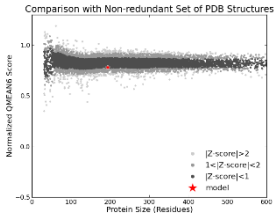
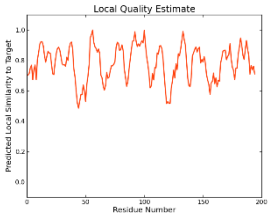
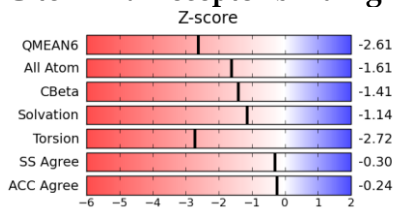
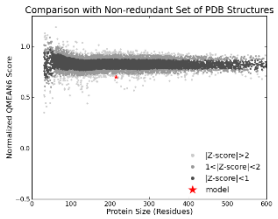
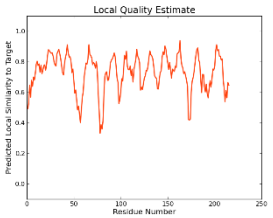
AA2B7_92 **GMQE** 0.96 **Ligand**: Zn in contact with H223, H227, E262		
Zinc protease		
traslocation domain		
N-terminal receptor binding		
C-terminal receptor binding		
		
B2_267 **GMQE** 0.99 **Ligand**: Zn in contact with H230, H234, E268 **Identities: B2_128**		

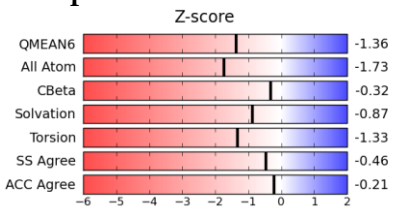
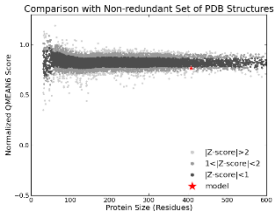
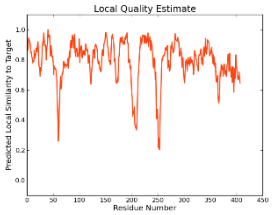
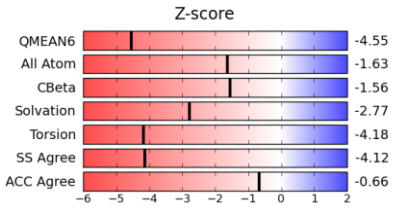
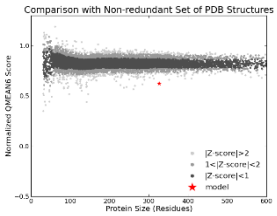
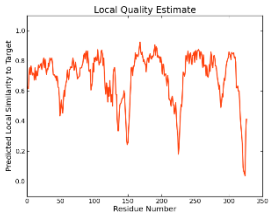
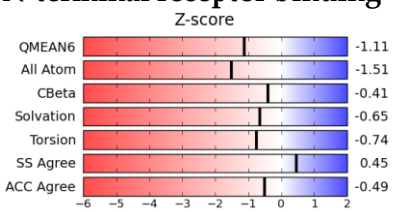
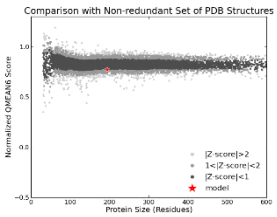
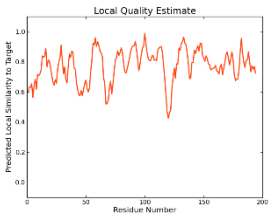
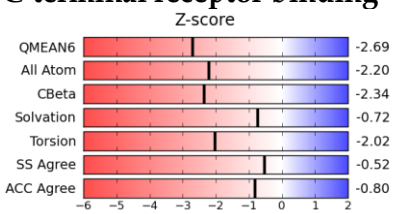
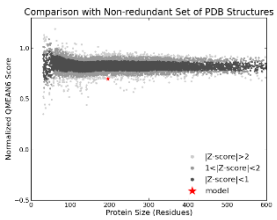
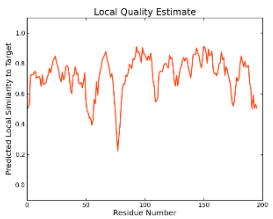
Zinc protease  **Identities: B2_450, B2_433, B2_331, B2_128, B2_275**		
traslocation domain  **Identities: B2_450, B2_433, B2_331, B2_128, B2_275**		
N-terminal receptor binding  **Identities: B2_128, B2_275**		
C-terminal receptor binding  **Identities: B2_331, B2_128**		
B2_275 **GMQE** 0.99 **Ligand**: Zn in contact with H230, H234, E268		

Zinc protease  **Identities: B2_450, B2_433, B2_331, B2_128, B2_267**		
traslocation domain  **Identities: B2_450, B2_433, B2_331, B2_128, B2_267**		
N-terminal receptor binding  **Identities: B2_128, B2_267**		
C-terminal receptor binding 		

B2_331 **GMQE** 0.99 **Ligand**: Zn in contact with		
Zinc protease  **Identities: B2_450, B2_433, B2_275, B2_128, B2_267**		
traslocation domain  **Identities: B2_450, B2_433, B2_275, B2_128, B2_267**		
N-terminal receptor binding  **Identities: B2_433**		
C-terminal receptor binding  **Identities: B2_128, B2_267**		

B2_433 **GMQE** 0.99 **Ligand**: Zn in contact with H230, H234, E268		
Zinc protease  **Identities: B2_450, B2_433, B2_331, B2_128, B2_275**		
traslocation domain  **Identities: B2_450, B2_433, B2_331, B2_128, B2_275**		
N-terminal receptor binding  **Identities: B2_331**		
C-terminal receptor binding 		

B2_450 **GMQE** 0.99 **Ligand**: Zn in contact with H230, H234, E268		
Zinc protease  **Identities: B2_267, B2_433, B2_331, B2_128, B2_275**		
traslocation domain  **Identities: B2_267, B2_433, B2_331, B2_128, B2_275**		
N-terminal receptor binding 		
C-terminal receptor binding 		

F_357 **GMQE** 0.99 **Ligand**: Zn in contact with H230, H234, E268		
Zinc protease 		
traslocation domain 		
N-terminal receptor binding 		
C-terminal receptor binding 		

APPENDIX C

Scripts

R script. Heatmap creation

```
activity_data <- read.table("ACTIVITY_FLEX.tsv", header=TRUE, row.names=1)
require(graphics); require(grDevices)
x  <- as.matrix(activity_data)
rc <- rainbow(nrow(x), start = 0, end = .2)
cc <- rainbow(ncol(x), start = 0, end = .2)
png(file="HEATMAP_ACTIVITY_FLEX.png", width = 5*300,height = 5*300, res = 300,
pointsize = 8)
hv <- heatmap(x,  col = heat.colors(20), scale = "row",
           RowSideColors = rc, ColSideColors = cc )
dev.off()
```

Python script. Recombination matrix generation

```
import os, sys
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
from matplotlib import cm

files = os.listdir(os.getcwd())

labels =
['1','2','3','4','5','6','7','8','9','10','11','12','13','14','15','16','17','18','19','20','21','22','23','24','25','26',
'27','28','29','30','31','32','33','34','35','36','37','38','39','40']

for i in files:
	if '.xml' in i:
		print 'parsing del file ',i
		max = 0
		min = 0
		n_iterazioni = 0
		testo = open(i).readlines()
		os.system('grep number '+i+'> temp')
		numbers = open('temp').readlines()
		number = numbers[-1][8:]
		n_iterazioni = int(len(numbers)/2)
		str_number = number.split('<')
		max = int (str_number[0])
		os.system('rm -rf temp')
		iterazioni = np.zeros((n_iterazioni,40,40),dtype=int)
		iter = np.zeros((40,40),dtype=int)
		count_iter = 0
```

```
#               matrix = [['' for i in range(40)] for j in range(40)]
                for line in testo[::-1]:
                        if '<recedge>' in line:
                                nod_x = line.split('efrom')
                                n_x = nod_x[1][1:-2]
                                nod_y = line.split('eto')
                                n_y = nod_y[1][1:-2]
#                               aaa = n_x+','+n_y
                                iter[int(n_x)-1,int(n_y)-1] += 1
#                               matrix[int(n_x)-1][int(n_y)-1] = aaa
#                               print aaa, matrix[int(n_x)-1][int(n_y)-1]
                        if '<number>' in line:
#                               matrix.tofile('matrix', sep='\t', format="%s")
#                               out_matrix = open('matrix','w')
#                               for el in matrix:
#                                   for EL in el:
#                                        out_matrix.write(EL+'\t')
#                                                print EL
#                                   out_matrix.write('\n')
#                               out_matrix.close()
#                               sys.exit()
#                               out = open('pippo'+str(count_iter)+'.tsv','w')
                                count_iter += 1
#                               print iter
#                               print count_iter
#                               for el in iter:
#                                       for EL in el:
#                                               out.write(str(EL)+'\t')
#                                       out.write('\n')
#                               out.close()
                                iterazioni[count_iter-1] = iter
                                if count_iter == n_iterazioni:
                                        iter = np.zeros((40,40),dtype=int)
                                        break
                risultato = np.zeros((40,40), dtype=int)
                for jk in iterazioni:
                        risultato += jk
                medie = risultato/n_iterazioni
                fig, ax = plt.subplots()
                matplotlib.use('Agg')
                cax = ax.imshow(medie, interpolation='nearest', cmap=cm.coolwarm)
                titolo = 'Conteggio relazioni tra nodi\ndati da file '+i
                ax.set_title(titolo)
```

```
                cbar = fig.colorbar(cax )
                ax.set_xlabel('Nodo donatore')
                ax.set_ylabel('Nodo accettore')
                ax.grid(True)
                ax.xaxis.set_label_position('bottom')
                ax.set_yticks(np.arange(40)+0.5, minor=False)
                ax.set_xticks(np.arange(40)+0.5, minor=False)
                ax.set_xticklabels(labels, minor=False, size=8)
                ax.set_yticklabels(labels, minor=False, size=8,  rotation=90)
                fname = i+'_plot.png'
                plt.savefig(fname)
#               plt.show()
                out = open('medie'+i+'.tsv','w')
                for el in medie:
                        for EL in el:
                        out.write(str(EL)+'\t')
                out.write('\n')
        out.close()
                print 'numero intarazioni', count_iter
                medie = np.zeros((40,40),dtype=int)
#               sys.exit()
```

Python script. Correction homopolymer

```
#!/usr/bin/python
__author__ = 'Ferdinando Spagnolo'
import sys, numpy
conta_A = 0
conta_T = 0
conta_C = 0
conta_G = 0
nucleotidi = {'A': 0, 'T': 1, 'C': 2, 'G': 3}
non_nucleotidi = '-N'
basi = "ATCG"
max_lunghezza_oligo = 20
intestazione_tabella = "\t\tlunghezza nucleotidi\n"
termine_tabella = "\t\tL: Conteggio oligonucleotide totale,\n\t\tR: conteggio correzioni Roche,\n\t\tI: Conteggio correzioni Illumina"
count = 0
for i in range(max_lunghezza_oligo):
 intestazione_tabella = intestazione_tabella+str(count)+"\t#\t"
 count += 1
intestazione_tabella = intestazione_tabella + "\n" + max_lunghezza_oligo*"(L,R,I)\t#\t"
total = len(sys.argv)
```

```
cmdargs = str(sys.argv)
print ("Numero totale degli argomenti passati al programma: %d " % total)
print ("Lista argomenti: %s " % cmdargs)
print ("Nome del programma: %s" % str(sys.argv[0]))
print ("Primo argomento argument: %s" % str(sys.argv[1]))
print ("Secondo argomento: %s" % str(sys.argv[2]))
f_in = open(sys.argv[1]).readlines()
head_r = "1_p1_A2B2_87_contig029.fsa "
head_i = "2_cb87_pl_ill.fa           "
str_r  = ""
str_i  = ""
for i in f_in:
 if i.startswith(head_r):
  ii = i.split()
  str_r += ii[1]
for i in f_in:
 if i.startswith(head_i):
  ii = i.split()
  str_i += ii[1]
#  inizializzo la matrice degli omopolimeri.
#  4 righe per ogni nucleotide A, T, C, G
#  ogni colonna e' composta da tre interi: il primo e'il conteggio dell'omopolimero, il secondo il conteggio di correzioni r, il terzo il conteggio di correzioni i
#  correzioni r: su omopolimero roche e' stato aggiunto un nucloeotide, il gap '-' su sequenza Roche e'sostituito dal nucleotide
#  correzioni i: su omopolimero roche e' stato tolto un nucloeotide, il gap '-' su sequenza Illumina e corrispondente nucleotide in Roche sono rimossi
#  per semplicita'di codice le prime due colonne sono zero per rispettare gli indici
m_omopolimeri = numpy.zeros((4,max_lunghezza_oligo), dtype=[('f0', '>i4'), ('f1', '>i4'), ('f2', '>i4')])
log_001 = "Controllo sequenze:\n"
print log_001
log_002 = "Controllo lunghezza:\n"
print log_002
if len(str_r)  != len(str_r):
 log_003 = "Sequenze di lunghezza diversa:\nLunghezza Roche: "+str(len(str_r))+" Lunghezza Illumina: "+str(len(str_i))+"\nPer l'esecuzione del programma accertarsi che le due sequenze Roche ed Illumina siano allineate.\nUscita del programma."
 print log_003
 sys.exit()
else:
 log_004 = "Sequenze di lunghezza uguale:\nLunghezza Roche: "+str(len(str_r))+" Lunghezza Illumina: "+str(len(str_i))+"\nSi procede con la correzione.\n"
 print log_004
 for i in str_r:
  if i == 'A': conta_A += 1
  if i == 'T': conta_T += 1
```

```
 if i == 'C': conta_C += 1
 if i == 'G': conta_G += 1
print "Tot A: ", conta_A
print "Tot T: ", conta_T
print "Tot C: ", conta_C
print "Tot G: ", conta_G
print "Tot nt:", conta_A + conta_T + conta_C + conta_G , "\n"

## individuazione errori oligonucleotidi ROCHE

print "Inizio correzione ROCHE\n"
oligonucleotide = str_r[0]
count = 3
start = 2
stop  = 3

for i in str_r[3:-1]:
 if i in oligonucleotide:
  oligonucleotide += i
  stop = count + 1
 else:
  if oligonucleotide[-1] not in non_nucleotidi:
   m_omopolimeri[nucleotidi[oligonucleotide[-1]]][len(oligonucleotide)][0] += 1
  if len(oligonucleotide)>1:
   if oligonucleotide[-1] not in non_nucleotidi:
#  cerco errore su 3'
    for k in range(1,4):
     nt_prec = str_r[start-k]
     if nt_prec == '-':
      if str_i[start-k] in oligonucleotide:
#       print "ricerca su 3'"
#       print "\ntrovato gap"
#       print "originale  ", oligonucleotide,oligonucleotide[-1], " esteso ", str_r[start-5:stop+5]
#       print "illumina   ", str_i[start-5:stop+5]
       m_omopolimeri[nucleotidi[oligonucleotide[-1]]][len(oligonucleotide)][1] += 1
       str_r = str_r[:start-k]+k*oligonucleotide[-1]+str_r[start:]
#       print "sostituito ",oligonucleotide, oligonucleotide[-1], " esteso ", str_r[start-5:stop+5]
      else:
       pass
     else:
      break
#  cerco errore su 5'

    for k in range(1,4):
     nt_succ = str_r[stop+k-1]
     if nt_succ == '-':
      if str_i[stop+k-1] in oligonucleotide:
```

```
#       print "ricerca su 5'"
#       print "\ntrovato gap"
#       print "originale  ", oligonucleotide,oligonucleotide[-1], " esteso ", str_r[start-5:stop+5]
#       print "illumina   ", str_i[start-5:stop+5]
       m_omopolimeri[nucleotidi[oligonucleotide[-1]]][len(oligonucleotide)][1] += 1
#########################
#       print str_r[start-5:stop], oligonucleotide[-1], str_r[stop+1:stop+5]
       str_r = str_r[:stop]+k*oligonucleotide[-1]+str_r[stop+k:]
#       print "sostituito ",oligonucleotide, k*oligonucleotide[-1], " esteso ", str_r[start-5:stop+5]
      else:
       pass
     else:
      break
  oligonucleotide = i
  start = count
 count += 1

## individuazione errori oligonucleotidi ILLUMINA

print "Inizio correzione ILLUMINA"

oligonucleotide = str_r[0]
count = 3
start = 2
stop  = 3

for i in str_i[3:-1]:
 if i in oligonucleotide:
  oligonucleotide += i
  stop = count + 1
 else:
  if len(oligonucleotide)>1:
   if oligonucleotide[-1] not in non_nucleotidi:
#  cerco errore su 3'
#    print "\ncerca errori su 3'\n"
    for k in range(1,4):
     nt_prec = str_i[start-k]
     if nt_prec == '-':
      if str_r[start-k] in oligonucleotide:
#       print oligonucleotide, str_r[start-k], str_i[start-k]
       m_omopolimeri[nucleotidi[oligonucleotide[-1]]][len(oligonucleotide)][2] += 1
#       print "conteggio errore effettuato"
#       print "gap esteso ",str_r[start-3:stop+3]
#       print "gap esteso ",str_i[start-3:stop+3]
       str_r = str_r[:start-k]+k*"#"+str_r[start:]
       str_i = str_i[:start-k]+k*"#"+str_i[start:]
#       print "gap esteso ",str_r[start-3:stop+3]
```

```
#       print "gap esteso ",str_i[start-3:stop+3]
      else:
       pass
     else:
      break
#  cerco errore su 5'
#    print "\ncerca errori su 5'\n"
#    print oligonucleotide
    for k in range(3):
     nt_succ = str_i[stop+k]
     if nt_succ == '-':
      if str_r[stop+k] in oligonucleotide:
#       print oligonucleotide, str_r[stop+k], str_i[stop+k], stop, stop+k, k
       m_omopolimeri[nucleotidi[oligonucleotide[-1]]][len(oligonucleotide)][2] += 1
#       print "conteggio errore effettuato"
       str_r = str_r[:stop+k]+"#"+str_r[stop+k+1:]
       str_i = str_i[:stop+k]+"#"+str_i[stop+k+1:]

      else:
       break
     else:
      break
  oligonucleotide = i
  start = count
 count += 1
print "Lunghezza stringhe: ", len(str_r), len(str_i)

numpy.savetxt("conteggi", m_omopolimeri, fmt='%s', delimiter='\t#\t', newline='\n',
header=intestazione_tabella, footer=termine_tabella, comments='#')

f=open("controllo.txt","w")

conta = 0
a = ''
b = ''
aa = 'ROCHE       '
bb = 'ILLUMINA    '
for i in range(len(str_r)):
 a += str_r[i]
 b += str_i[i]
 conta +=1

 if conta == 60:

  riga = aa+a+"\n"+bb+b+"\n\n"
  f.write(riga)
  conta = 0
```

```
  a = ''
  b = ''

f.close()

f=open("roche_illumina.fa","w")

conta = 0
a = ''
b = ''
aa = '> ROCHE     corretto   \n'
bb = '> ILLUMINA  corretto   \n'

f.write(aa)
for i in str_r:
 a += i
 conta +=1
 if conta == 60:
  f.write(a+"\n")
  conta = 0
  a = ''

conta = 0

f.write(bb)
for i in str_i:
 a += i
 conta +=1
 if conta == 60:
  f.write(b)
  conta = 0
  b = ''

f.close()
```

BIBLIOGRAPHY

[1] M. Cherington, «Clinical spectrum of botulism.,» *Muscle and nerve,* vol. 21, n. 6, pp. 701-10, Jun 1998.

[2] G. Schiavo, M. Matteoli e C. Montecucco, «Neurotoxins affecting neuroexocytosis.,» *Physiological reviews,* vol. 80, n. 2, pp. 717-66, Apr 2000.

[3] S. S. Arnon, R. Schechter, T. V. Inglesby, D. A. Henderson, J. G. Bartlett, M. S. Ascher, E. Eitzen, A. D. Fine, J. Hauer, M. Layton, S. Lillibridge, M. T. Osterholm, T. O'Toole, G. Parker, T. M. Perl, P. K. Russell, D. L. Swerdlow e K. Tonat, «Botulinum toxin as a biological weapon: medical and public health management.,» *JAMA,* vol. 285, n. 8, pp. 1059-70, 28 Feb 2001.

[4] «Possession, use, and transfer of select agents and toxins; biennial review. Final rule.,» *Federal register,* vol. 77, n. 194, pp. 61083-115, 5 Oct 2012 .

[5] L. D. S. Smith e H. Sugiyama, «Botulism: the Organism, its Toxins, the Disease,» pp. -, 1988.

[6] P. Aureli, M. Di Cunto, A. Maffei, G. De Chiara, G. Franciosa, L. Accorinti, A. M. Gambardella e D. Greco, «An outbreak in Italy of botulism associated with a dessert made with mascarpone,» *European journal of epidemiology,* vol. 16, n. 10, pp. 913-8, 2000.

[7] R. Koepke, J. Sobel e S. S. Arnon, «Global occurrence of infant botulism, 1976-2006,» *Pediatrics,* vol. 122, pp. e73-e82, 2008.

[8] K. K. Hill, G. Xie, B. T. Foley, T. J. Smith, A. C. Munk, D. Bruce, L. A. Smith, T. S. Brettin e J. C. Detter, «Recombination and insertion events involving the botulinum neurotoxin complex E strains.,» *BMC biology,* vol. 7, p. 66, 2009.

[9] P. Aureli, «Two cases of type E infant botulism caused by neurotoxigenic Clostridium butyricum in Italy,» *J. Infect. Dis.,* vol. 154, pp. 207-211, 1986.

[10] L. L. Simpson, «The life history of a botulinum toxin molecule,» *Toxicon,* vol. 68, pp. 40-59, 2013.

[11] T. N. Wenham, «Botulism: a rare complication of injecting drug use,» *Emerg. Med. J.,* vol. 25, pp. 55-56, 2008.

[12] E. A. Johnson e C. Montecucco, «Handbook of Clinical Neurology,» pp. 333-368, 2008.

[13] A. N. Sheth, «International outbreak of severe botulism with prolonged toxemia caused by commercial carrot juice,» *Clin. Infect. Dis.,* vol. 47, pp. 1245-1251, 2008.

[14] R. P. Fagan, J. B. McLaughlin e J. P. Middaugh, «Persistence of botulinum toxin in patients' serum: Alaska, 1959-2007,» *J. Infect. Dis.,* vol. 199, pp. 1029-1031, 2009.

[15] C. B. Shoemaker e G. A. Oyler, «Persistence of botulinum neurotoxin inactivation of nerve function,» *Curr. Top. Microbiol. Immunol.,* vol. 364, pp. 179-196, 2013.

[16] R. C. Whitemarsh, «Characterization of botulinum neurotoxin a subtypes1 through 5 by investigation of activities in mice, in neuronal cell cultures, and in vitro,» *Infect. Immun.,* vol. 81, pp. 3894-3902, 2013.

[17] M. Naumann, «Evidence-based review and assessment of botulinum neurotoxin for the treatment of secretory disorders,» *Toxicon,* vol. 67, pp. 141-152, 2013.

[18] M. P. Byrne e L. A. Smith, «Development of vaccines for prevention of botulism,» *Biochimie,* vol. 82, pp. 955-966, 2000.

[19] L. A. Smith, «Botulism and vaccines for its prevention,» *Vaccine,* vol. 27, pp. D33-D39, 2009.

[20] A. P.-A. Karalewitz e J. T. Barbieri, «Vaccines against botulism,» *Curr. Opin. Microbiol.,* vol. 15, pp. 317-324, 2012.

[21] C. Garcia-Rodriguez, «Molecular evolution of antibody cross-reactivityfor two subtypes of type A botulinum neurotoxin,» *Nature Biotech.,* vol. 25, pp. 107-116, 2007.

[22] J. Lou, «Affinitymaturation of human botulinum neurotoxin antibodies by light chain shuffling via yeast mating,» *Protein Eng. Des. Sel.,* vol. 23, pp. 311-319, 2010.

[23] L. W. Cheng, L. H. Stanker, T. D. Henderson, J. Lou e J. D. Marks, «Antibody protection against botulinum neurotoxin intoxication in mice,» *Infect. Immun.,* vol. 77, pp. 4305-4313, 2009.

[24] S. Pantano e C. Montecucco, «The blockade of the neurotransmitter release apparatus by botulinum neurotoxins,» *Cell. Mol. Life Sci.,* vol. 71, pp. 793-811, 2014.

[25] T. Binz, «Clostridial neurotoxin light chains: devices for SNARE cleavage mediated blockade of neurotransmission,» *Curr. Top. Microbiol. Immunol.,* vol. 364, pp. 139-157, 2013.

[26] J. Guo, X. Pan, Y. Zhao e S. Chen, «Engineering clostridia neurotoxins with elevated catalytic activity,» *Toxicon,* pp. 158-166, 2013.

[27] L. Ma, «Single application of A2 NTX, a botulinum toxin A2 subunit, prevents chronic pain over long periods in both diabetic and spinal cord injury-induced neuropathic pain models,» *J. Pharmacol. Sci.,* vol. 119, pp. 282-286, 2012.

[28] S. Chen e J. T. Barbieri, «Engineering botulinum neurotoxin to extend therapeutic intervention,» *Proc. Natl Acad. Sci. USA,* vol. 106, pp. 9180-9184, 2009.

[29] M. R. Popoff e P. Bouvet, «Genetic characteristics of toxigenic Clostridia and toxin gene evolution,» *Toxicon,* vol. 75, pp. 63-89, 2013.

[30] K. K. Hill e T. J. Smith, «Genetic diversity within Clostridium botulinum serotypes, botulinum neurotoxin gene clusters and toxin subtypes,» *Curr. Top. Microbiol. Immunol.,* vol. 364, pp. 1-20, 2013.

[31] S. E. Maslanka, C. Luquez, J. K. Dykes, W. H. Tepp, C. L. Pier, S. Pellett, B. H. Raphael, S. R. Kalb, J. R. Barr, A. Rao e E. A. Johnson, «A Novel Botulinum Neurotoxin, Previously Reported as Serotype H, Has a and F and Is Neutralized With Serotype A Antitoxin.,» *The Journal of infectious diseases,* 10 Jun 2015.

[32] F. S. A. A. P. A. F. A. G. B. A. T. D. C. A. S. F. P. V. A. F. A. B. D. M. D. L. Giordani F, «Genomic characterization of Italian Clostridium botulinum group I strains,» *Infection, Genetics and Evolution,* n. 36, pp. 62-71, 2015.

[33] D. B. Lacy, W. Tepp, A. C. Cohen, B. R. DasGupta e R. C. Stevens, «Crystal structure of botulinum neurotoxin type A and implications for toxicity.,» *Nature structural biology,* vol. 5, n. 10, pp. 898-902, Oct 1998.

[34] S. Swaminathan e S. Eswaramoorthy, «Structural analysis of the catalytic and binding sites of Clostridium botulinum neurotoxin B,» *Nature Struct. Biol.,* vol. 7, pp. 693-699, 2000.

[35] D. Kumaran, «Domain organization in Clostridium botulinum neurotoxin type E is unique: its implication in faster translocation,» *J. Mol. Biol.,* vol. 386, pp. 233-245, 2009.

[36] S. Gu, «Botulinum neurotoxin is shielded by NTNHA in an interlocked complex,» *Science,* vol. 335, pp. 977-981, 2012.

[37] I. Ohishi e G. Sakaguchi, «Oral toxicities of Clostridium botulinum type C and D toxins of different molecular sizes,» *Infect. Immun.,* vol. 28, pp. 303-309, 1980.

[38] K. Lee, «Structure of a bimodular botulinum neurotoxin complex provides insights into its oral toxicity,» *PLoS Pathog.,* vol. 9, pp. e1003690-, 2013.

[39] D. A. Benefield, S. K. Dessain, N. Shine, M. D. Ohi e D. B. Lacy, «Molecular assembly of botulinum neurotoxin progenitor complexes,» *Proc. Natl Acad. Sci. USA,* vol. 110, pp. 5630-5635, 2013.

[40] P. F. Bonventre, «Absorption of botulinal toxin from the gastrointestinal tract,» *Rev. Infect. Dis.,* vol. 1, pp. 663-667, 1979.

[41] Y. Sugawara, «Botulinum hemagglutinin disrupts the intercellular epithelial barrier by directly binding E-cadherin,» *J. Cell Biol.,* vol. 189, pp. 691-700, 2010.

[42] Y. Fujinaga, Y. Sugawara e T. Matsumura, «Uptake of botulinum neurotoxin in the intestine,» *Curr. Top. Microbiol. Immunol.,* vol. 364, pp. 45-59, 2013.

[43] A. Couesnon, J. Molgo, C. Connan e M. R. Popoff, «Preferential entry of botulinum neurotoxin A H domain through intestinal crypt cells and targeting to cholinergic neurons of the mouse intestine,» *PLoS Pathog.,* vol. 8, pp. e1002583-, 2012.

[44] C. Montecucco, «How do tetanus and botulinum toxins bind to neuronal membranes?,» *Trends Biochem. Sci.,* vol. 11, pp. 314-317, 1986.

[45] A. Rummel, «Double receptor anchorage of botulinum neurotoxins accounts for their exquisite neurospecificity,» *Curr. Top. Microbiol. Immunol.,* vol. 364, pp. 61-90, 2013.

[46] Q. Chai, «Structural basis of cell surface receptor recognition by botulinum neurotoxin B,» *Nature,* vol. 444, pp. 1096-1100, 2006.

[47] R. Jin, A. Rummel, T. Binz e A. T. Brunger, «Botulinum neurotoxin B recognizes its protein receptor with high affinityand specificity,» *Nature,* vol. 444, pp. 1092-1095, 2006.

[48] R. P. Berntsson, L. Peng, M. Dong e P. Stenmark, «Structure of dual receptor binding to botulinum neurotoxin B,» *Nature Commun.,* vol. 4, pp. 2058-, 2013.

[49] C. Montecucco, O. Rossetto e G. Schiavo, «Presynaptic receptor arrays for clostridial neurotoxins,» *Trends Microbiol.,* vol. 12, pp. 442-446, 2004.

[50] W. E. Van Heyningen, «Tentative identification of the tetanus toxin receptor in nervous tissue,» *J. Gen. Microbiol.,* vol. 20, pp. 310-320, 1959.

[51] L. L. Simpson e M. M. Rapport, «The binding of botulinum toxin to membrane lipids: sphingolipids, steroids and fattyacids,» *J. Neurochem.,* vol. 18, pp. 1751-1759, 1971.

[52] K. Simons e D. Toomre, «Lipid rafts and signal transduction,» *Nature Rev. Mol. Cell Biol.,* vol. 1, pp. 31-39, 2000.

[53] A. Prinetti, N. Loberto, V. Chigorno e S. Sonnino, «Glycosphingolipid behaviour in complex membranes,» *Biochim. Biophys. Acta,* vol. 1788, pp. 184-193, 2009.

[54] J. D. Black e J. O. Dolly, «Interaction of 125I-labeled botulinum neurotoxins with nerve terminals. II. Autoradiographic evidence for its uptake into motor nerves by acceptor-mediated endocytosis,» *J. Cell Biol.,* vol. 103, pp. 535-544, 1986.

[55] J. Strotmeier, «Botulinum neurotoxin serotype D attacks neurons via two carbohydrate-binding sites in a ganglioside-dependent manner,» *Biochem. J.,* vol. 431, pp. 207-216, 2010.

[56] A. P. Karalewitz, Z. Fu, M. R. Baldwin, J. J. Kim e J. T. Barbieri, «Botulinum neurotoxin serotype C associates with dual ganglioside receptors to facilitate cell entry,» *J. Biol. Chem.,* vol. 287, pp. 40806-40816, 2012.

[57] J. Strotmeier, «The biological activityof botulinum neurotoxin type C is dependent upon novel types of ganglioside binding sites,» *Mol. Microbiol.,* vol. 81, pp. 143-156, 2011.

[58] B. C. Yowler, R. D. Kensinger e C. L. Schengrund, «Botulinum neurotoxin A activityis dependent upon the presence of specific gangliosides in neuroblastoma cells expressing synaptotagmin I,» *J. Biol. Chem.,* vol. 277, pp. 32815-32819, 2002.

[59] T. Nishiki, «Identification of protein receptor for Clostridium botulinum type B neurotoxin in rat brain synaptosomes,» *J. Biol. Chem.,* vol. 269, pp. 10498-10503, 1994.

[60] M. Dong, «Synaptotagmins I and II mediate entry of botulinum neurotoxin B into cells,» *J. Cell. Biol.,* vol. 162, pp. 1293-1303, 2003.

[61] A. Rummel, «Identification of the protein receptor binding site of botulinum neurotoxins B and G proves the double-receptor concept,» *Proc. Natl Acad. Sci. USA,* vol. 104, pp. 359-364, 2007.

[62] L. Peng, «Botulinum neurotoxin D-C uses synaptotagmin I and II as receptors, and human synaptotagmin II is not an effective receptor for type B, D-C and G toxins,» *J. Cell Sci.,* vol. 125, pp. 3233-3242, 2012.

[63] M. Dong, «SV2 is the protein receptor for botulinum neurotoxin A,» *Science.,* vol. 312, pp. 592-596, 2006.

[64] M. Dong, «Glycosylated SV2A and SV2B mediate the entry of botulinum neurotoxin E into neurons,» *Mol. Biol. Cell,* vol. 19, pp. 5226-5237, 2008.

[65] S. Mahrhold, «Identification of the SV2 protein receptor-binding site of botulinum neurotoxin type E,» *Biochem. J.,* vol. 453, pp. 37-47, 2013.

[66] S. Mahrhold, A. Rummel, H. Bigalke, B. Davletov e T. Binz, «The synaptic vesicle protein 2C mediates the uptake of botulinum neurotoxin A into phrenic nerves,» *FEBS Lett.,* vol. 580, pp. 2011-2014, 2006.

[67] R. M. Benoit, «Structural basis for recognition of synaptic vesicle protein 2C by botulinum neurotoxin A,» *Nature,* vol. 505, pp. 108-111, 2014.

[68] G. Schiavo, «Structural biology: dangerous liaisons on neurons,» *Nature,* vol. 444, pp. 1019-1020, 2006.

[69] C. Colasante, «Botulinum neurotoxin type A is internalized and translocated from small synaptic vesicles at the neuromuscular junction,» *Mol. Neurobiol.,* vol. 48, pp. 120-127, 2013.

[70] S. Takamori, «Molecular anatomy of a trafficking organelle,» *Cell,* vol. 127, pp. 831-846, 2006.

[71] Y. Saheki e P. De Camilli, «Synaptic vesicle endocytosis,» *Cold Spring Harb. Perspect. Biol.,* vol. 4, pp. a005645-, 2012.

[72] K. Wohlfarth, H. Goschel, J. Frevert, R. Dengler e H. Bigalke, «Botulinum A toxins: units versus units,» *Naunyn. Schmiedebergs. Arch. Pharmacol.,* vol. 355, pp. 335-340, 1997.

[73] C. Rasetti-Escargueil, Y. Liu, P. Rigsby, R. G. Jones e D. Sesardic, «Phrenic nerve hemidiaphragm as a highly sensitive replacement assay for determination of functional botulinum toxin antibodies,» *Toxicon,* vol. 57, pp. 1008-1016, 2011.

[74] S. Sun, W. H. Tepp, E. A. Johnson e E. R. Chapman, «Botulinum neurotoxins B and E translocate at different rates and exhibit divergent responses to GT1b and low pH,» *Biochemistry,* vol. 51, pp. 5655-5662, 2012.

[75] L. K. Koriazova e M. Montal, «Translocation of botulinum neurotoxin light chain protease through the heavy chain channel,» *Nature Struct. Biol.,* vol. 10, pp. 13-18, 2003.

[76] M. Dalla Serra, «Conductive properties and gating of channels formed by syringopeptin 25A, a bioactive lipodepsipeptide from Pseudomonas syringae pv. syringae, in planar lipid membranes,» *Mol. Plant. Microbe Interact.,* vol. 12, pp. 401-409, 1999.

[77] A. Fischer, «Molecular architecture of botulinum neurotoxin E revealed by single particle electron microscopy,» *J. Biol. Chem.,* vol. 283, pp. 3997-4003, 2008.

[78] S. Bade, «Botulinum neurotoxin type D enables cytosolic delivery of enzymatically active cargo proteins to neurones via unfolded translocation intermediates,» *J. Neurochem.,* vol. 91, pp. 1461-1472, 2004.

[79] M. Montal, «Botulinum neurotoxin: a marvel of protein design,» *Annu. Rev. Biochem.,* vol. 79, pp. 591-617, 2010.

[80] A. Fischer, «Synchronized chaperone function of botulinum neurotoxin domains mediates light chain translocation into neurons,» *Curr. Top. Microbiol. Immunol.,* vol. 364, pp. 115-137, 2013.

[81] D. H. Hoch, «Channels formed by botulinum, tetanus, and diphtheria toxins in planar lipid bilayers: relevance to translocation of proteins across membranes,» *Proc. Natl Acad. Sci. USA,* vol. 82, pp. 1692-1696, 1985.

[82] J. J. Donovan e J. L. Middlebrook, «Ion-conducting channels produced by botulinum toxin in planar lipid membranes,» *Biochemistry,* vol. 25, pp. 2872-2876, 1986.

[83] M. Pirazzini, O. Rossetto, P. Bolognese, C. C. Shone e C. Montecucco, «Double anchorage to the membrane and intact inter-chain disulfide bond are required for the low pH induced entry of tetanus and botulinum neurotoxins into neurons,» *Cell. Microbiol.,* vol. 13, pp. 1731-1743, 2011.

[84] M. Pirazzini, «Neutralisation of specific surface carboxylates speeds up translocation of botulinum neurotoxin type B enzymatic domain,» *FEBS Lett.,* vol. 587, pp. 3831-3836, 2013.

[85] A. Fischer e M. Montal, «Crucial role of the disulfide bridge between botulinum neurotoxin light and heavy chains in protease translocation across membranes,» *J. Biol. Chem.,* vol. 282, pp. 29604-29611, 2007.

[86] G. Miesenbock, D. A. De Angelis e J. E. Rothman, «Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins,» *Nature,* vol. 394, pp. 192-195, 1998.

[87] S. Sankaranarayanan e T. A. Ryan, «Real-time measurements of vesicle-SNARE recycling in synapses of the central nervous system,» *Nature Cell Biol.,* vol. 2, pp. 197-204, 2000.

[88] M. Eisenberg, T. Gresalfi, T. Riccio e S. McLaughlin, «Adsorption of monovalent cations to bilayer membranes containing negative phospholipids,» *Biochemistry,* vol. 18, pp. 5213-5223, 1979.

[89] P. Nordera, M. D. Serra e G. Menestrina, «The adsorption of Pseudomonas aeruginosa exotoxin A to phospholipid monolayers is controlled by pH and surface potential,» *Biophys. J.,* vol. 73, pp. 1468-1478, 1997.

[90] J. W. Deutsch e R. B. Kelly, «Lipids of synaptic vesicles: relevance to the mechanism of membrane fusion,» *Biochemistry,* vol. 20, pp. 378-385, 1981.

[91] R. W. Ledeen, M. F. Diebler, G. Wu, Z. H. Lu e H. Varoqui, «Ganglioside composition of subcellular fractions, including pre- and postsynaptic membranes, from Torpedo electric organ,» *Neurochem. Res.,* vol. 18, pp. 1151-1155, 1993.

[92] O. Rossetto, M. Pirazzini e C. Montecucco, «Botulinum neurotoxins: genetic, structural and mechanistic insights.,» *Nature reviews. Microbiology,* vol. 12, n. 8, pp. 535-49, Aug 2014.

[93] V. E. Bychkova, R. H. Pain e O. B. Ptitsyn, «The 'molten globule' state is involved in the translocation of proteins across membranes,» *FEBS Lett.,* vol. 238, pp. 231-234, 1988.

[94] O. B. Ptitsyn, R. H. Pain, G. V. Semisotnov, E. Zerovnik e O. I. Razgulyaev, «Evidence for a molten globule state as a general intermediate in protein folding,» *FEBS Lett.,* vol. 262, pp. 20-24, 1990.

[95] F. G. van der Goot, J. M. Gonzalez-Manas, J. H. Lakey e F. Pattus, «A 'molten-globule' membrane-insertion intermediate of the pore-forming domain of colicin A,» *Nature,* vol. 354, pp. 408-410, 1991.

[96] R. Kukreja e B. Singh, «Biologically active novel conformational state of botulinum, the most poisonous poison,» *J. Biol. Chem.,* vol. 280, pp. 39346-39352, 2005.

[97] M. Pirazzini, «The thioredoxin reductase-thioredoxin system is involved in the entry of tetanus and botulinum neurotoxins in the cytosol of nerve terminals,» *FEBS Lett.,* vol. 587, pp. 150-155, 2013.

[98] T. C. Sudhof e J. Rizo, «Synaptic vesicle exocytosis,» *Cold Spring Harb. Perspect. Biol.,* vol. 3, pp. a005637-, 2011.

[99] T. Hayashi, «Synaptic vesicle membrane fusion complex: action of clostridial neurotoxins on assembly,» *EMBO J.,* vol. 13, pp. 5051-5061, 1994.

[100] R. B. Sutton, D. Fasshauer, R. Jahn e A. T. Brunger, «Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution,» *Nature,* vol. 395, pp. 347-353, 1998.

[101] A. Megighian, «Evidence for a radial SNARE super-complex mediating neurotransmitter release at the Drosophila neuromuscular junction,» *J. Cell Sci.,* vol. 126, pp. 3134-3140, 2013.

[102] F. Anniballi, E. Chironna, S. Astegiano, A. Fiore, B. Auricchio, G. Buonincontro, M. Corvonato, V. Segala, G. Mandarino, D. De Medici e L. Decastelli, «Foodborne botulism associated with home-preserved turnip tops in Italy.,» *Annali dell'Istituto superiore di sanita,* vol. 51, n. 1, pp. 60-1, 2015.

[103] E. Augustynowicz, A. Gzyl e J. Slusarczyk, «Detection of enterotoxigenic Clostridium perfringens with a duplex PCR.,» *Journal of medical microbiology,* vol. 51, n. 2, pp. 169-72, Feb 2002.

[104] S. Andrews, «FastQC: a quality control tool for high throughput sequence data.,» 2010.

[105] N. Joshi e J. Fass, «Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files,» 2011.

[106] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones e I. Birol, «ABySS: a parallel assembler for short read sequence data.,» *Genome research,* vol. 19, n. 6, pp. 1117-23, Jun 2009.

[107] A. C. E. Darling, B. Mau, F. R. Blattner e N. T. Perna, «Mauve: multiple alignment of conserved genomic sequence with rearrangements.,» *Genome research,* vol. 14, n. 7, pp. 1394-403, Jul 2004.

[108] S. Balzer, K. Malde e I. Jonassen, «Systematic exploration of error sources in pyrosequencing flowgram data.,» *Bioinformatics,* vol. 27, n. 13, p. i304–i309, 2011.

[109] *454 Sequencing System Software Manual Version 2.6.,* 454 Life Sciences Corp. , 2011.

[110] M. Sebaihia, M. W. Peck, N. P. Minton, N. R. Thomson, M. T. G. Holden, W. J. Mitchell, A. T. Carter, S. D. Bentley, D. R. Mason, L. Crossman, C. J. Paul, A. Ivens, M. H. J. Wells-Bennik, I. J. Davis, A. M. Cerdeno-Tarraga, C. Churcher, M. A. Quail, T. Chillingworth, T. Feltwell, A. Fraser, I. Goodhead, Z. Hance, K. Jagels, N. Larke, M. Maddison, S. Moule, K. Mungall, H. Norbertczak, E. Rabbinowitsch, M. Sanders, M. Simmonds, B. White, S. Whithead e J. Parkhill, «Genome sequence of a proteolytic (Group I) Clostridium botulinum strain Hall A,» *Genome research,* vol. 17, n. 7, pp. 1082-92, Jul 2007.

[111] K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell e E. W. Sayers, «GenBank.,» *Nucleic acids research,* 20 Nov 2015.

[112] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller e D. J. Lipman, «Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,» *Nucleic acids research,* vol. 25, n. 17, pp. 3389-402, 1 Sep 1997.

[113] F. Sanger e A. Coulson, «A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.,» *Journal of molecular biology,* vol. 94, n. 3, pp. 441-338, 25 May 1975.

[114] F. Sanger, S. Nicklen e A. Coulson, «DNA sequencing with chain-terminating inhibitors.,» *Proceedings of the National Academy of Sciences of the United States of America,* vol. 74, n. 12, pp. 5463-5467, Dec 1977.

[115] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, S. Ciufo e W. Li, «Prokaryotic Genome Annotation Pipeline.,» *The NCBI Handbook [Internet]. 2nd edition.,* 10 Dec 2013.

[116] D. Bentley e S. Balasubramanian, «Accurate whole human genome sequencing using reversible terminator chemistry.,» *Nature,* vol. 456, n. 7218, pp. 53-59, 2008.

[117] K. V. Voelkerding, S. Dames e D. J.D., «Next-generation sequencing: from basic research to diagnostics.,» *Clinical chemistry,* vol. 55, n. 4, pp. 641-658, Apr 2009.

[118] Z. Zheng, A. Advani, Ö. Melefors, S. Glavas, H. Nordström, W. Ye e A. F. Andersson, «Titration-free massively parallel pyrosequencing using trace amounts of starting material.,» *Nucleic Acids Research,* vol. 38, n. 13, p. e137, 2010.

[119] C. King e T. Scott-Horton, «Pyrosequencing: a simple method for accurate genotyping.,» *Journal of visualized experiments,* vol. 8, n. 11, p. 630, 2008.

[120] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben e J. M. Rothberg, «Genome Sequencing in Open Microfabricated High Density Picoliter Reactors.,» *Nature,* vol. 437, n. 7057, p. 376–380, 2005.

[121] X. Didelot e D. Falush, «Inference of bacterial microevolution using multilocus sequence data.,» *Genetics,* vol. 175, n. 3, pp. 1251-66, Mar 2007.

[122] N. Wang, S. Gottesman, M. C. Willingham, M. M. Gottesman e M. R. Maurizi, «A human mitochondrial ATP-dependent protease that is highly homologous to,» *Proceedings of the National Academy of Sciences of the United States of America,* vol. 90, n. 23, pp. 11247-51, 1 Dec 1993.

[123] X. Didelot, G. Meric, D. Falush e A. E. Darling, «Impact of homologous and non-homologous recombination in the genomic evolution of,» *BMC genomics,* vol. 13, p. 256, 2012.

[124] K. S. Lole, R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R. Ingersoll, H. W. Sheppard e S. C. Ray, «Full-length human immunodeficiency virus type 1 genomes from subtype C-infected,» *Journal of virology,* vol. 73, n. 1, pp. 152-60, Jan 1999.

[125] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel e A. Bairoch, «ExPASy: The proteomics server for in-depth protein knowledge and analysis.,» *Nucleic acids research,* vol. 31, n. 13, pp. 3784-8, 1 Jul 2003.

[126] «The PyMOL Molecular Graphics System, Version 1.7.4.,» Schrödinger, LLC..

[127] E. Bramucci, A. Paiardini, F. Bossa e S. Pascarella, «PyMod: sequence similarity searches, multiple sequence-structure alignments, and,» *BMC bioinformatics,* vol. 13 Suppl 4, p. S2, 2012.

[128] D. Seeliger, «Autodock/Vina plugin for PyMOL,» 2009.

[129] C. Chothia e A. Lesk, «The relation between the divergence of sequence and structure in proteins.,» *The EMBO journal,* vol. 5, n. 4, pp. 823-826, 1986.

[130] M. Martí-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo e A. Sali, «Comparative protein structure modeling of genes and genomes.,» *Annual Review of Biophysics and Biomolecular Structure,* n. 29, pp. 291-325, 2000.

[131] S. Kaczanowski e P. Zielenkiewicz, «Why similar protein sequences encode similar three-dimensional structures?,» *Theoretical Chemistry Accounts,* vol. 125, n. 3, pp. 643-650, 2009.

[132] A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin e S. F. Altschul, «Improving the accuracy of PSI-BLAST protein database searches with,» *Nucleic acids research,* vol. 29, n. 14, pp. 2994-3005, 15 Jul 2001.

[133] I. N. Shindyalov e P. E. Bourne, «Protein structure alignment by incremental combinatorial extension (CE) of the,» *Protein engineering,* vol. 11, n. 9, pp. 739-47, Sep 1998.

[134] J. D. Thompson, D. G. Higgins e T. J. Gibson, «CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment choice.,» *Nucleic acids research,* vol. 22, n. 22, pp. 4673-80, 11 Nov 1994.

[135] K. Joo, J. Lee, J.-H. Seo, K. Lee, B.-G. Kim e J. Lee, «All-atom chain-building by optimizing MODELLER energy function using,» *Proteins,* vol. 75, n. 4, pp. 1010-23, Jun 2009.

[136] A. Sali e T. Blundell, «Comparative protein modelling by satisfaction of spatial restraints.,» *Journal of Molecular Biology,* vol. 234, n. 3, pp. 779-815, 1993.

[137] A. Fiser, R. Do e A. Sali, «Modeling of loops in protein structures.,» *Protein Science ,* n. 9, pp. 1753-1773, 2000.

[138] B. Webb e A. Sali, «Comparative Protein Structure Modeling Using MODELLER.,» *Current Protocols in Bioinformatics,* 2014.

[139] V. Roberts, B. Iverson, S. Iverson, S. Benkovic, R. Lerner, E. Getzoff e J. Tainer, «Antibody remodeling: a general solution to the design of a metal-coordination site in an antibody binding pocket.,» *Proceedings of the National Academy of Sciences of the United States of America ,* vol. 87, n. 17, p. 6654–6658, 1990 .

[140] G. Ramachandran, C. Ramakrishnan e V. Sasisekharan, «Stereochemistry of polypeptide chain configurations.,» *Journal of Molecular Biology,* n. 7, pp. 95-99, 1963.

[141] R. Laskowski, M. W. MacArthur, D. S. Moss e J. M. Thornton, «PROCHECK - a program to check the stereochemical quality of protein structures.,» *Journal of Applied Crystallography,* n. 26, pp. 283-291, 1993.

[142] P. Benkert, M. Kunzli e T. Schwede, «QMEAN server for protein model quality estimation.,» *Nucleic acids research,* vol. 37, n. Web Server issue, pp. W510-4, Jul 2009.

[143] P. Benkert, S. C. E. Tosatto e D. Schomburg, «QMEAN: A comprehensive scoring function for model quality assessment.,» *Proteins,* vol. 71, n. 1, pp. 261-77, Apr 2008.

[144] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov e P. E. Bourne, «The Protein Data Bank.,» *Nucleic acids research,* vol. 28, n. 1, pp. 235-42, 1 Jan 2000.

[145] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell e A. J. Olson, «AutoDock4 and AutoDockTools4: Automated docking with selective receptor,» *Journal of computational chemistry,* vol. 30, n. 16, pp. 2785-91, Dec 2009.

[146] J. Gasteiger e M. Marsili, «Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges.,» *Tetrahedron,* vol. 36, n. 22, pp. 3219-3228, 1980.

[147] R. Agarwal, J. Schmidt, R. Stafford e S. Swaminathan, «Mode of VAMP substrate recognition and inhibition of Clostridium botulinum neurotoxin F.,» *Nature structural & molecular biology ,* vol. 16, n. 7, pp. 789-794, 2009.

[148] D. Santos-Martins, S. Forli, M. J. Ramos e A. J. Olson, «AutoDock4(Zn): an improved AutoDock force field for small-molecule docking to,» *Journal of chemical information and modeling,* vol. 54, n. 8, pp. 2371-9, 25 Aug 2014.

[149] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew e A. J. Olson, «Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function.,» *Journal of Computational Chemistry,* vol. 19, n. 14, p. 1639–1662, 1998.

[150] C. R. Hutchinson, «Polyketide and non-ribosomal peptide synthases: falling together by coming apart.,» *Proceedings of the National Academy of Sciences of the United States of America,* vol. 100, n. 6, pp. 3010-2, 18 Mar 2003.

[151] G. Franciosa, L. Fenicia, M. Pourshaban e P. Aureli, «Recovery of a strain of Clostridium botulinum producing both neurotoxin A and,» *Applied and environmental microbiology,* vol. 63, n. 3, pp. 1148-50, Mar 1997.

[152] G. Franciosa, A. Maugliani, C. Scalfaro e P. Aureli, «Evidence that plasmid-borne botulinum neurotoxin type B genes are widespread,» *PloS one,* vol. 4, n. 3, p. e4829, 2009.

[153] M. Brigulla e W. Wackernagel, «Molecular aspects of gene transfer and foreign DNA acquisition in prokaryotes,» *Applied microbiology and biotechnology,* vol. 86, n. 4, pp. 1027-41, Apr 2010.

[154] E. Stackebrandt, I. Kramer, J. Swiderski e H. Hippe, «Phylogenetic basis for a taxonomic dissection of the genus Clostridium.,» *FEMS immunology and medical microbiology,* vol. 24, n. 3, pp. 253-8, Jul 1999.

[155] S. R. Kalb, «Discovery of a novel enzymatic cleavage site for botulinum neurotoxin F5,» *FEBS Lett.,* vol. 586, pp. 109-115, 2012.

[156] S. R. Kalb, J. Baudys, J. C. Rees, T. J. Smith, L. A. Smith, C. H. Helma, K. Hill, S. Kull, S. Kirchner, M. B. Dorner, B. G. Dorner, J. L. Pirkle e J. R. Barr, «De novo subtype and strain identification of botulinum neurotoxin type B through,» *Analytical and bioanalytical chemistry,* vol. 403, n. 1, pp. 215-26, Apr 2012.

[157] K. K. Hill, T. J. Smith, C. H. Helma, L. O. Ticknor, B. T. Foley, R. T. Svensson, J. L. Brown, E. A. Johnson, L. A. Smith, R. T. Okinaka, P. J. Jackson e J. D. Marks, «Genetic diversity among Botulinum Neurotoxin-producing clostridial strains.,» *Journal of bacteriology,* vol. 189, n. 3, pp. 818-32, Feb 2007.

[158] X. Didelot e M. C. J. Maiden, «Impact of recombination on bacterial evolution.,» *Trends in microbiology,* vol. 18, n. 7, pp. 315-22, Jul 2010.

[159] J. G. Lawrence, «Horizontal and vertical gene transfer: the life history of pathogens.,» *Contributions to microbiology,* vol. 12, pp. 255-71, 2005.

[160] S. Abby e V. Daubin, «Comparative genomics and the evolution of prokaryotes.,» *Trends in microbiology,* vol. 15, n. 3, pp. 135-41, Mar 2007.

[161] N. Gonzalez-Escalona, R. Timme, B. H. Raphael, D. Zink e S. K. Sharma, «Whole-genome single-nucleotide-polymorphism analysis for discrimination of,» *Applied and environmental microbiology,* vol. 80, n. 7, pp. 2125-32, Apr 2014.

[162] A. J. Vogler, F. Chan, D. M. Wagner, P. Roumagnac, J. Lee, R. Nera, M. Eppinger, J. Ravel, L. Rahalison, B. W. Rasoamanana, S. M. Beckstrom-Sternberg, M. Achtman, S. Chanteau e P. Keim, «Phylogeography and molecular epidemiology of Yersinia pestis in Madagascar.,» *PLoS neglected tropical diseases,* vol. 5, n. 9, p. e1319, Sep 2011.

[163] M. Vos e X. Didelot, «A comparison of homologous recombination rates in bacteria and archaea.,» *The ISME journal,* vol. 3, n. 2, pp. 199-208, Feb 2009.

[164] M. J. Jacobson, G. Lin, T. S. Whittam e E. A. Johnson, «Phylogenetic analysis of Clostridium botulinum type A by multi-locus sequence,» *Microbiology (Reading, England),* vol. 154, n. Pt 8, pp. 2408-15, Aug 2008.

[165] J. S. Olsen, H. Scholz, S. Fillo, V. Ramisse, F. Lista, A. K. Tromborg, T. Aarskaug, I. Thrane e J. M. Blatny, «Analysis of the genetic distribution among members of Clostridium botulinum group,» *Journal of microbiological methods,* vol. 96, pp. 84-91, Jan 2014.

[166] T. C. Bruen, H. Philippe e D. Bryant, «A simple and robust statistical test for detecting the presence of recombination.,» *Genetics,* vol. 172, n. 4, pp. 2665-81, Apr 2006.

[167] M. D. Collins e A. K. East, «Phylogeny and taxonomy of the food-borne pathogen Clostridium botulinum and its,» *Journal of applied microbiology,* vol. 84, n. 1, pp. 5-17, Jan 1998.

[168] T. J. Smith, K. K. Hill, B. T. Foley, J. C. Detter, A. C. Munk, D. C. Bruce, N. A. Doggett, L. A. Smith, J. D. Marks, G. Xie e T. S. Brettin, «Analysis of the neurotoxin complex genes in Clostridium botulinum A1-A4 and B1,» *PloS one,* vol. 2, n. 12, p. e1271, 2007.

[169] K. M. Marshall, M. Bradshaw e E. A. Johnson, «Conjugative botulinum neurotoxin-encoding plasmids in Clostridium botulinum.,» *PloS one,* vol. 5, n. 6, p. e11087, 2010.

[170] A. T. Carter, C. J. Paul, D. R. Mason, S. M. Twine, M. J. Alston, S. M. Logan, J. W. Austin e M. W. Peck, «Independent evolution of neurotoxin and flagellar genetic loci in proteolytic,» *BMC genomics,* vol. 10, p. 115, 2009.

[171] J. G. Lawrence e H. Hendrickson, «Genome evolution in bacteria: order beneath chaos.,» *Current opinion in microbiology,* vol. 8, n. 5, pp. 572-8, Oct 2005.

[172] A. Selvapandiyan, N. Arora, R. Rajagopal, S. K. Jalali, T. Venkatesan, S. P. Singh e R. K. Bhatnagar, «Toxicity analysis of N- and C-terminus-deleted vegetative insecticidal protein,» *Applied and environmental microbiology,* vol. 67, n. 12, pp. 5855-8, Dec 2001.

[173] M. Nishie, J.-I. Nagao e K. Sonomoto, «Antibacterial peptides "bacteriocins": an overview of their diverse,» *Biocontrol science,* vol. 17, n. 1, pp. 1-16, Mar 2012.

[174] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson e D. G. Higgins, «Fast, scalable generation of high-quality protein multiple sequence alignments,» *Molecular systems biology,* vol. 7, p. 539, 2011.

[175] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate e A. Bateman, «The Pfam protein families database: towards a more sustainable future.,» *Nucleic acids research,* 15 Dec 2015.

[176] R. Engh e R. Huber, «Accurate bond and angle parameters for X-ray protein structure refinement.,» *Acta Crystallographica ,* vol. A47, n. Part 4, pp. 392-400, 1991.

[177] R. A. Laskowski, «Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature.,» *Bioinformatics,* vol. 23, n. 14, pp. 1824-1827, 2007.

[178] T. Castrignano, P. D. De Meo, D. Cozzetto, I. G. Talamo e A. Tramontano, «The PMDB Protein Model Database.,» *Nucleic acids research,* vol. 34, n. Database issue, pp. D306-9, 1 Jan 2006.

[179] C. Montecucco e G. Schiavo, «Tetanus and botulism neurotoxins: a new group of zinc proteases.,» *Trends in biochemical sciences,* vol. 18, n. 9, pp. 324-7, Sep 1993.

PUBLISHED ARTICLES

This section contains scans of the following papers published during the doctorate:

1. Giordani F, Fillo S, Anselmo A, Palozzi AM, Fortunato A, Gentile B, Azarnia Tehran D, Ciammaruconi A, Spagnolo F, Pittiglio V, Anniballi F, Auricchio B, De Medici D, Lista; Genomic characterization of Italian Clostridium botulinum group I strains,"Infection, Genetics and Evolution", 36, 62–71,2015,Elsevier.
   Not printed in the present thesis because not published in open access mode.
2. Fillo S, Giordani F, Anselmo A, Fortunato A, Palozzi AM, De Santis R, Ciammaruconi A, Spagnolo F, Anniballi F, Fiore A, Auricchio B, De Medici D, Lista F.; ,Draft Genome Sequence of Clostridium botulinum B2 450 Strain from Wound Botulism in a Drug User in Italy,Genome Announcements, 3, 2, 2015,American Society for Microbiology.

3. Giordani F, Fillo S, Anselmo A, Palozzi AM, Fortunato A, Gentile B, Pittiglio V, Spagnolo F, Anniballi F, Fiore A, Auricchio B, De Medici D, Lista F.; ,"Whole-Genome Sequence of Clostridium botulinum A2B3 87, a Highly Virulent Strain Involved in a Fatal Case of Foodborne Botulism in Italy",Genome Announcements,3,2,,2015,American Society for Microbiology.

genomeA
Journals.ASM.org

Whole-Genome Sequence of *Clostridium botulinum* A2B3 87, a Highly Virulent Strain Involved in a Fatal Case of Foodborne Botulism in Italy

Francesco Giordani,[a] Silvia Fillo,[a] Anna Anselmo,[a] Anna Maria Palozzi,[a] Antonella Fortunato,[a] Bernardina Gentile,[a] Valentina Pittiglio,[a] Ferdinando Spagnolo,[a] Fabrizio Anniballi,[b] Alfonsina Fiore,[b] Bruna Auricchio,[b] Dario De Medici,[b] Florigio Lista[a]

Histology and Molecular Biology Section Army Medical and Veterinary Research Center, Rome, Italy[a]; National Reference Center for Botulism, Department of Veterinary Public Health and Food Safety, Istituto Superiore di Sanità (ISS), Rome, Italy[b]

Here, we report the genome sequence of a rare bivalent strain of *Clostridium botulinum*, A2B3 87. The strain was isolated from a foodborne botulism case that occurred in Italy in 1995. The case was characterized by rapid evolution of the illness and failure of conventional treatments.

Received 20 February 2015 **Accepted** 23 February 2015 **Published** 26 March 2015

Citation Giordani F, Fillo S, Anselmo A, Palozzi AM, Fortunato A, Gentile B, Pittiglio V, Spagnolo F, Anniballi F, Fiore A, Auricchio B, De Medici D, Lista F. Whole-genome sequence of *Clostridium botulinum* A2B3 87, a highly virulent strain involved in a fatal case of foodborne botulism in Italy. Genome Announc 3(2):e00237-15. doi:10.1128/genomeA.00237-15.

Address correspondence to Florigio Lista, romano.lista@gmail.com.

Clostridium botulinum is an anaerobic spore-forming bacterium capable of synthesizing the botulinum neurotoxin (BoNT), a powerful and highly lethal poison (1). BoNTs are the only toxins which are tier 1 select agents (select agents and toxins lists are available at http://www.selectagents.gov). The intoxication by BoNT causes a neuromuscular paralysis produced by the block of peripheral cholinergic synapses. Botulism occurs from ingestion of preformed BoNT within contaminated food (foodborne) or by infections with bacterial spores and consequently toxin formation *in situ* (intestinal or wound) (1). The *C. botulinum* taxon has been divided into four groups (I to IV), as demonstrated by rRNA 16S gene comparison, amplified fragment length polymorphism (AFLP), and other techniques (2–4). On the basis of their serological activity, *Clostridium botulinum* strains are classified by 8 serotypes of BoNT (A to H), which are further divided into subtypes (A1 to A5, B1 to B7, E1 to E9, F1 to F7). Sixteen *C. botulinum* strains were sequenced and analyzed and several draft assemblies are available. These data show significant differences between the four groups (5–7).

A2B3 87 *C. botulinum* strain was isolated from a clinical case of foodborne botulism that occurred in a 76-year-old woman in Italy in 1995. The patient died 4 days after the hospital admission after being treated with the polyvalent antiserum and supported by respiratory aid. The bacterium was isolated in a sample of canned macrobiotic food based on "seitan," a traditional Eastern recipe (8). The strain was found to produce both A and B toxin serotypes (ratio 10/1) (8).

The A2B3 87 genome was sequenced with the Roche 454 GS FLX Titanium and Illumina MiSeq platforms. From the 454 sequencing, a ~25× coverage was obtained (103,330,725 total sequenced bases, 272,724 total reads), while MiSeq sequencing reached ~246× coverage (970,106,826 total sequenced bases, 3,350,829 total paired reads). Illumina reads were used to cover the gaps in the 454 sequencing assembly and to correct the homopolymers length inaccuracies produced by 454 sequencing (9). The final draft assembly, with a G+C content of 27.9%, consists of 13 contigs. The 11 representing the chromosomal sequence are 3,847,714 bp long, while the 2 that constitute the plasmids are 275,568 and 45,268 bp long. The chromosomal gaps are caused by unresolved repeated sequences: the nine copies of the rRNA genes operon (total length of ~43 kb) and the two copies of the beta-*N*-acetyl-glucosamidase gene.

The bigger plasmid contains the two (A and B) BoNT genes. The BoNT/A gene is an A2 subtype, with a similarity of 99.85% (2 amino acid different) with A2 Kyoto BoNT sequence (YP_002803127.1) (3). BoNT/B is a B3 subtype but shows a considerable number of amino acid mutations compared to the other B3s; the similarity with CDC 795 (EF028400.1) is 98.22% and there are 21 different amino acids. More studies are needed to really understand the role played by the amino acid substitutions in this BoNT sequence. Moreover, the smaller plasmid showed no homologies with any other plasmid sequenced to date.

Nucleotide sequence accession number. The genome sequence of *C. botulinum* A2B3 87 is available in DDBJ/EMBL/GenBank under the accession no. AUZB00000000.

ACKNOWLEDGMENT

This work was supported by the Italian Ministry of Defense, SEGREDIFESA/DNA-5 Department of Technological Innovation (EBLN project).

REFERENCES

1. Arnon SS, Schechter R, Inglesby TV, Henderson DA, Bartlett JG, Ascher MS, Eitzen E, Fine AD, Hauer J, Layton M, Lillibridge S, Osterholm MT, O'Toole T, Parker G, Perl TM, Russell PK, Swerdlow DL, Tonat K, Working Group on Civilian Biodefense. 2001. Botulinum toxin as a biological weapon: medical and public health management. JAMA 285:1059–1070. http://dx.doi.org/10.1001/jama.285.8.1059.
2. Collins MD, East AK. 1998. Phylogeny and taxonomy of the foodborne

pathogen *Clostridium botulinum* and its neurotoxins. J Appl Microbiol 84: 5–17. http://dx.doi.org/10.1046/j.1365-2672.1997.00313.x.
3. Hill KK, Smith TJ, Helma CH, Ticknor LO, Foley BT, Svensson RT, Brown JL, Johnson EA, Smith LA, Okinaka RT, Jackson PJ, Marks JD. 2007. Genetic diversity among botulinum neurotoxin-producing clostridial strains. J Bacteriol 189:818–832. http://dx.doi.org/10.1128/JB.01180-06.
4. Olsen JS, Scholz H, Fillo S, Ramisse V, Lista F, Trømborg AK, Aarskaug T, Thrane I, Blatny JM. 2014. Analysis of the genetic distribution among members of *Clostridium botulinum* group I using a novel multilocus sequence typing (MLST) assay. J Microbiol Methods 96:84–91. http://dx.doi.org/10.1016/j.mimet.2013.11.003.
5. Sebaihia M, Peck MW, Minton NP, Thomson NR, Holden MTG, Mitchel WJ, Carter AT, Bentley SD, Mason DR, Crossman L, Paul CJ, Ivens A, Wells-Bennik MHJ, Davis IJ, Cerdeño-Tárraga AM, Churcher C, Quai MA, Chillingworth T, Feltwell T, Fraser A, Goodhead I, Hance Z, Jagels K, Larke N, Maddison M, Moule S, Mungall K, Norbertczak H, Rabbinowitsch E, Sanders M, Simmonds M, White B, Whithead S, Parkhill J. 2007. Genome sequence of a proteolytic (group I) *Clostridium botulinum* strain hall A and comparative analysis of the clostridial genomes. Genome Res 17:1082–1092. http://dx.doi.org/10.1101/gr.6282807.
6. Peck MW, Stringer SC, Carter AT. 2011. *Clostridium botulinum* in the post-genomic era. Food Microbiol 28:183–191. http://dx.doi.org/10.1016/j.fm.2010.03.005.
7. Skarin H, Håfström T, Westerberg J, Segerman B. 2011. *Clostridium botulinum* group III: a group with dual identity shaped by plasmids, phages and mobile elements. BMC Genomics 12:185. http://dx.doi.org/10.1186/1471-2164-12-185.
8. Franciosa G, Fenicia L, Pourshaban M, Aureli P. 1997. Recovery of a strain of *Clostridium botulinum* producing both neurotoxin A and neurotoxin B from canned macrobiotic food. Appl Environ Microbiol 63: 1148–1150.
9. Balzer S, Malde K, Jonassen I. 2011. Systematic exploration of error sources in pyrosequencing flowgram data. Bioinformatics 27:i304–i309. http://dx.doi.org/10.1093/bioinformatics/btr251.

genomeA
Journals.ASM.org

Draft Genome Sequence of *Clostridium botulinum* B2 450 Strain from Wound Botulism in a Drug User in Italy

Silvia Fillo,[a] Francesco Giordani,[a] Anna Anselmo,[a] Antonella Fortunato,[a] Anna Maria Palozzi,[a] Riccardo De Santis,[a] Andrea Ciammaruconi,[a] Ferdinando Spagnolo,[a] Fabrizio Anniballi,[b] Alfonsina Fiore,[b] Bruna Auricchio,[b] Dario De Medici,[b] Florigio Lista[a]

Histology and Molecular Biology Section, Army Medical and Veterinary Research Center, Rome, Italy[a]; National Reference Center for Botulism, Department of Veterinary Public Health and Food Safety, Istituto Superiore di Sanità (ISS), Rome, Italy[b]

Here, we report the draft genome sequence of *Clostridium botulinum* B2 450, responsible for the first reported case of wound botulism in a drug user in Italy.

Received 20 February 2015 **Accepted** 23 February 2015 **Published** 2 April 2015

Citation Fillo S, Giordani F, Anselmo A, Fortunato A, Palozzi AM, De Santis R, Ciammaruconi A, Spagnolo F, Anniballi F, Fiore A, Auricchio B, De Medici D, Lista F. 2015. Draft genome sequence of *Clostridium botulinum* B2 450 strain from wound botulism in a drug user in Italy. Genome Announc 3(2):e00238-15. doi:10.1128/genomeA.00238-15.

Address correspondence to Florigio Lista, romano.lista@gmail.com.

Clostridium botulinum is a microorganism able to produce the botulinum neurotoxin (BoNT), a powerful poison that causes botulism, a serious neuroparalytic disease. There are 3 mainly clinical manifestations of botulism: foodborne (ingestion of BoNT-contaminated foods), intestinal (BoNT is produced by intestinal colonization of *C. botulinum*), and wound (*C. botulinum* spores germinate and synthesize BoNT in contaminated wounds) (1). Based on its physiological characteristics, *C. botulinum* is divided into four different groups (I to IV), so phylogenetically different that they can be considered separate species (2, 3). Moreover, other species, such as *Clostridium baratii* and *Clostridium butyricum*, are able to produce botulinum neurotoxins type F and E, respectively. The BoNT is classified by 8 serotypes (A to H), each of which is divided into subtypes. To date, 16 full genomes and several draft assemblies of *C. botulinum* are available, prevalently isolated from food borne and infant cases. A5(B3′)H04402 065 (accession no. FR773526) is the only full genome sequence originating from a wound botulism case (4, 5).

The B2 450 strain was isolated in 2009 from wound exudate of a heroin user patient, in Messina, Sicily, Italy (6). Sequencing was performed using both Roche 454 GS FLX Titanium and Illumina MiSeq platforms. Roche sequencing generated 286,392 single-end reads, with 112,174,234 total sequenced bases, ~28-fold coverage. The reads were *de novo* assembled with the GS Assembler software (Newbler package) into 160 contigs, and 2,430,881 paired-end reads were produced by Illumina sequencing with 731,610,908 total sequenced bases and ~185-fold coverage. Illumina reads were assembled with Abyss-pe producing 497 contigs. 454 and Miseq contigs were combined using Minimus2 software, and 18 contigs were obtained.

A well-known flaw of the 454 platform is the erroneous determination of homopolymer lengths (7); 1,218 homopolymeric stretches were corrected according to Illumina sequences. Nine gaps were closed using 454 and Miseq reads or Sanger sequencing.

The final draft assembly consists of 9 chromosomal contigs, for a total length of 4,070,655 bp, and one plasmidic contig, 250,014 bp long, containing the BoNT/B2 gene. The genome has a G+C content of 27.8%. The 8 remaining gaps are due to repeated sequences (rRNA operon and beta-*N*-acetyl-glucosamidase genes).

The 16S rRNA gene sequence of B2 450 belongs to group I showing 99.8% similarity with the A1 ATCC 3502 16S sequence (NC_009495) (3). Comparing some gene sequences (rpoB-mdh-aroE-hsp60-aceK-oppB-recA) (8), the B2 450 strain appears phylogenetically closer to *Clostridium sporogenes*, with 99.99% similarity to ATCC 15579 (ABKW0200000), than to all other *C. botulinum* strains (96.14% to A1 ATCC 3502 and 95.79% to A3 Loch Maree; NC_010520). *C. sporogenes* phylogenetically belongs to group I (9). The 450 strain may represent a *C. sporogenes* lineage that has recently acquired the BoNT/B gene (perhaps through the plasmid) (10).

Nucleotide sequence accession number. The genome sequence of *C. botulinum* B2 450 was deposited at DDBJ/EMBL/GenBank under the accession no. JXSU00000000. The version described in this paper is the first version.

ACKNOWLEDGMENT

This work was supported by Italian Ministry of Defense, SEGREDIFESA/DNA-5 Department of Technological Innovation (EBLN project).

REFERENCES

1. Sobel J. 2005. Botulism. Clin Infect Dis 41:1167–1173. http://dx.doi.org/10.1086/444507.
2. Collins MD, East AK. 1998. Phylogeny and taxonomy of the foodborne pathogen *Clostridium botulinum* and its neurotoxins. J Appl Microbiol 84:5–17. http://dx.doi.org/10.1046/j.1365-2672.1997.00313.x.
3. Hill KK, Smith TJ, Helma CH, Ticknor LO, Foley BT, Svensson RT, Brown JL, Johnson EA, Smith LA, Okinaka RT, Jackson PJ, Marks JD. 2007. Genetic diversity among botulinum neurotoxin-producing clostridial strains. J Bacteriol 189:818–832. http://dx.doi.org/10.1128/JB.01180-06.
4. Peck MW, Stringer SC, Carter AT. 2011. *Clostridium botulinum* in the post-genomic era. Food Microbiol 28:183–191. http://dx.doi.org/10.1016/j.fm.2010.03.005.
5. Carter AT, Pearson BM, Crossman LC, Drou N, Heavens D, Baker D,

Febrer M, Caccamo M, Grant KA, Peck MW. 2011. Complete genome sequence of the proteolytic *Clostridium botulinum* type A5 (B3′) strain H04402 065. J Bacteriol 193:2351–2352. http://dx.doi.org/10.1128/JB.00072-11.

6. Rodolico C, Barca E, Fenicia L, Anniballi F, Sinardi AU, Girlanda P. 2010. Wound botulism in drug users: a still underestimated diagnosis. Neurol Sci 31:825–827. http://dx.doi.org/10.1007/s10072-010-0350-1.
7. Balzer S, Malde K, Jonassen I. 2011. Systematic exploration of error sources in pyrosequencing flowgram data. Bioinformatics 27:i304–i309. http://dx.doi.org/10.1093/bioinformatics/btr251.
8. Jacobson MJ, Lin G, Whittam TS, Johnson EA. 2008. Phylogenetic analysis of *Clostridium botulinum* type A by multi-locus sequence typing. Microbiology 154:2408–2415. http://dx.doi.org/10.1099/mic.0.2008/016915-0.
9. Olsen JS, Scholz H, Fillo S, Ramisse V, Lista F, Trømborg AK, Aarskaug T, Thrane I, Blatny JM. 2014. Analysis of the genetic distribution among members of *Clostridium botulinum* group I using a novel multilocus sequence typing (MLST) assay. J Microbiol Methods 96:84–91. http://dx.doi.org/10.1016/j.mimet.2013.11.003.
10. Smith TJ, Hill KK, Foley BT, Detter JC, Munk AC, Bruce DC, Doggett NA, Smith LA, Marks JD, Xie G, Brettin TS. 2007. Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3, /Ba4 and /B1 clusters are located within plasmids. PLoS One 2:e1271. http://dx.doi.org/10.1371/journal.pone.0001271.

ACKNOWLEDGEMENTS

This work was supported by Italian Secretariat General of Defence & National Armaments Directorate - 5th Department – Technological Innovation and was part of the European Biodefence Laboratory Network (EBLN) coordination work (project n. EDA B-1325-ESM4-GP) on dangerous pathogens involving biodefence institutions from Norway, the Netherlands, Germany, France, Sweden, Norway, Belgium, Poland, Austria, and Italy.
I express all my gratitude to Col. Lista, Prof. Brancolini and Prof. Pascarella for the useful scientific support.
During the development of this work I knew many researchers holding a strong scientific background and humanity. Among them I would thank Dr. Faggioni, Dr. Fillo, and Dr. Giordani for their support.
Thank you, Dr. Anselmo, for being patient and altruist!

Academic Year 2015/2016, Udine, Italy

Ferdinando Spagnolo