UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede amministrativa: Università degli Studi di Padova

Dipartimento di Biologia

CORSO DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE
INDIRIZZO: BIOTECNOLOGIE
CICLO XXIX

TOWARDS "SYSTEMS BIOTECHNOLOGY": IDENTIFICATION, CHARACTERIZATION AND DESIGN/ENGINEERING OF PROTEIN INTERACTION MOTIFS/DOMAINS MEDIATING REGULATORY SIGNALS

Coordinatore: Ch.mo Prof. Paolo Bernardi
Supervisore: Ch.mo Prof. Francesco Filippini
Co-supervisore: Ch.mo Prof. Stefano Mammi

Dottorando: Irene Righetto

“Somewhere, something incredible is waiting to be known” – Carl Sagan

ACKNOWLEDGEMENTS

I want to express my gratitude to all the people participating in my PhD projects: without you this dissertation could not have been possible!

In particular:

My supervisor, Prof. Francesco Filippini, together with Dr. Valeria Rossi for their patience and ability to make a stimulating and friendly working environment: you brought out the best in me!

My labmates and collaborators, especially Dr. Giorgia Scapin, Dr. Yuriko Suemi Hernandez Gomez, Dr. Adelaide Milani, Dr. Giovanni Cattoli, Dr. Rosa Di Liddo, Prof. Enzo Menna, Prof. Vincenzo De Filippis

Prof. Francesco Zonta and Dr. Caterina Lupini as external advisors, for useful suggetions having improved the draft.

This dissertation is also based upon the material as it appears in:

Righetto I, Milani A, Cattoli G, Filippini F. Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features. BMC Bioinformatics. 2014 Dec 10;15:363. doi: 10.1186/s12859-014-0363-5.

Heidari A, Righetto I, Filippini F. Comparative Structural Analysis of Hemagglutinin for Unveiling Fingerprints in the Evolution and Spreading of Avian Influenza Viruses. (Submitted)

DEDICATION

I dedicate this thesis to Prof. Francesco Filippini and Dr. Valeria Rossi, thanking them for always believing in me and giving me the opportunity to spend other years of my life at their side.

SUMMARY

RIASSUNTO

Gli studi *in silico* aventi per oggetto la struttura di domini proteici e di motif sia strutturali che lineari, sono in grado di fornire un importante apporto in termini di comprensione di funzione e nelle biotecnologie.

Lo studio delle caratteristiche a carico della superficie proteica si rivelano essenziali nella comprensione delle Interazioni Proteina-Proteina (PPI); in particolare, la conservazione e variazione della superficie proteica e delle relative cavità in termini di idrofobicità, ingombro sterico e caratteristiche elettrostatiche, possono essere considerate come la forza in grado di guidare l'evoluzione e la specializzazione funzionale delle proteine stesse.

Alla luce di quanto sopra esposto, tecniche come la Modellistica Molecolare ed il confronto tra strutture giocano un ruolo importante nel chiarire il *modus operandi* delle proteine e questo progetto di Dottorato ha proprio sfruttato l'approccio integrato di alcune ben note tecniche di biologia computazionale basate sulla Modellistica Molecolare come, ad esempio, *Homology Modeling, Fold Recognition, Ab initio Modeling, PBE (Poisson-Boltzmann Electrostatics), Protein-peptide Docking e Hydropathy Analysis* con confronto di sequenze e strutture. Elemento indispensabile e prezioso, ovviamente, il feedback ottenuto dagli esperimenti al banco effettuati dai nostri collaboratori.

Questo approccio integrato è stato dunque applicato a differenti sistemi biologici:

- **Individuazione di determinanti di superficie in virus influenzali di tipo A H5N1**: è stata effettuata un'analisi dei determinanti di superficie a carico dell'emoagglutinina proveniente dal virus influenzale H5N1, coinvolta nell'interazione virus-ospite. Questo lavoro ha già condotto ad una pubblicazione. La variazione genomica è elevata nei virus influenzali di tipo A. L'evoluzione e la diffusione dei virus sono molto influenzate dalle caratteristiche immunogeniche e dalla capacità del virus, di interagire con le cellule dell'ospite tramite le due più importanti proteine presenti sul capside virale: l'emoagglutinina e la neuraminidasi. Le analisi oggi a disposizione sono basate sul confronto dell'attività sierologica e di sequenze primarie; alla luce di ciò, l'analisi strutturale di queste proteine capsidiche può essere in grado di svelare delle conoscenze a riguardo di certe regioni presenti sulla superficie proteica che possono essere cruciali per l'antigenicità e per il legame alle cellule dell'ospite. L'emoagglutinina, sezionata nei suoi domini e subdomini, è stata da noi studiata con metodi di Modellistica Molecolare e sottoposta a confronti strutturali fini, per individuare quelle variazioni che potessero risultare tipo/dominio specifiche. Abbiamo

evidenziato che la vicinanza strutturale e la similarità di sequenza primaria non sempre sono correlate; in più, caratteristiche tipo-specifiche di sottoregioni dell'emoagglutinina, monomeri e trimeri, possono essere rivelate grazie al confronto delle loro proprietà di superficie, (in termini di elettrostatica ed idrofobicità) appartenenti a sottoregioni dell'emoagglutinina, monomeri e trimeri. In questo lavoro ci siamo focalizzati sul virus H5N1 e abbiamo scoperto che il dominio di legame recettoriale dell'emoagglutinina (RBD) presenta delle variazioni tra clade circolanti e non più circolanti.

Le recenti scoperte riguardanti l'associazione tra la disposizione delle cariche al RBD ed il successo in termini evolutivi e di diffusione del virus H5N1 ci hanno spinto ad eseguire analisi integrate di filogenesi e biologia strutturale a carico dei virus H9N2. Infatti, l'influenza A è un agente zoonotico in grado di produrre un grosso impatto sia sulla salute pubblica che sull'industria del pollame, avendo la capacità di effettuare il salto d'ospite, come riportato proprio per H5N1 ed H9N2. Abbiamo effettuato un'analisi evoluzionistica su un grande dataset non ridondante di ceppi virali e questo ci ha consentito di individuare cinque gruppi di virus H9N2. In accordo con le precedenti analisi effettuate per H5N1, abbiamo ottenuto accordo tra i dati filogenetici con quelli ottenuti dalle analisi di confronto strutturale. In particolare, emerge che la variazione della disposizione delle cariche coincide con quella di siti noti dell'emoagglutinina coinvolti nell'evasione al sistema immunitario e nella specificità d'ospite. I risultati ottenuti da questo secondo lavoro pongono l'accento sull'importanza dell'integrazione tra analisi di tipo filogenetiche e di biologia strutturale nella scoperta di nuovi meccanismi evolutivi dei virus dell'influenza.

- **Variazione dell'architettura di domini in proteine di mammifero coinvolte nel traffico vescicolare**: la proteina umana VAMP7b è la più interessante tra quelle prodotte per splicing alternativo del gene SYBL1. La produzione di VAMP7b è causata dal salto dell'esone 6 che si traduce in uno slittamento della sequenza codificante. Abbiamo scoperto che questo evento è conservato in altre specie di mammiferi. VAMP7b condivide con l'isoforma principale il dominio inibitorio l*ongin* N-terminale e la prima metà dello SNARE motif. Nei mammiferi, VAMP7b è una proteina tronca in cui al C-terminale metà dello SNARE motif e la regione transmembrana sono sostituite da peptidi corti e variabili. È molto interessante notare come negli uomini e nelle scimmie antropomorfe lo slittamento della regione codificante determinato dal salto dell'esone 6 abbia prodotto un nuovo dominio di funzione

sconosciuta: proprio per questo VAMP7b umana non è tronca, ma addirittura 40 residui più lunga rispetto all'isoforma principale. Dal momento che l'esistenza di questa isoforma "lunga" ed il suo nuovo dominio sono stati confermati a livello proteico grazie all'ausilio di specifici anticorpi, abbiamo effettuato una dissezione *in silico* del nuovo dominio adoperando un'analisi di sequenza di tipo matrice posizione-specifica (PSI-BLAST), seguita da da Modellistica Strutturale di tipo *ab initio*. In più, dal momento che la regione N-terminale dello SNARE motif è conservata ed è nota nel mediare il legame intramolecolare al dominio *Longin*, abbiamo appurato la conservazione della conformazione chiusa sia in vivo (saggio del doppio ibrido in lievito) che in vitro (analisi NMR). Inoltre, la localizzazione subcellulare (SCL) di VAMP7b e Ykt6b è stata studiata adoperando chimere contenenti GFP e RFP. Non ultimo, le isoforme b dei geni *longin* sono stati analizzati tramite qPCR e si è scoperto essere regolate durante lo sviluppo.

- **Motif di legame con azione regolatoria sulla crescita e l'indirizzamento neuronale**: La regolazione fine delle interazioni proteina-proteina che avviene grazie alle variazioni nell'architettura dei domini o dal cambiamento di motif locali indotto dalla modulazione di caratteristiche di superficie, è in grado di regolare i percorsi di segnalazione sia a livello intra- che extracellulare. Le interazioni proteina-proteina extracellulari possono giocare un ruolo fondamentale nel riconoscimento eterologo (es. ospite-patogeno) come nella segnalazione omologa tra cellule appartenenti allo stesso organismo. Le proteine esposte in membrana plasmatica (PM) possono interagire le une con le altre e con la matrice extracellulare (ECM) per consentire informazioni posizionali e segnali di indirizzamento. Le molecole di adesione cellulare (CAMs) sono proteine della membrana plasmatica in grado di mediare segnali sia di natura attrattiva che repulsiva grazie ad interazioni omo- ed eterofiliche a carico dei loro domini extracellulari (EDs). Questi ultimi sono composti per la magior parte da domini ripetuti aventi fold di tipo Ig o Fibronectina di tipo III. Le attuali conoscenze suggeriscono che i 4 domini extracellulari N-terminali di tipo Ig siano importanti nelle interazioni omo- o eterofiliche ed in modo particolare il dominio Ig2 è provvisto di un importante motif di interazione. Nel nostro laboratorio abbiamo sviluppato dei peptidi biomimetici che riproducono i motif di interazione conosciuti o predetti appartenenti al dominio Ig2 di L1CAM umana e al singolo dominio Ig di LINGO1 umana, proteine, queste, che giocano un ruolo fondamentale nella crescita, nell'indirizzamento e nel differenziamento neuronale. Sulla base della conservazione strutturale della regione

del motif (anche tra proteine con architetture molto diverse dei loro EDs), abbiamo iniziato a studiarne la variazione di sequenza mediante analisi per omologia e per espressioni regolari, per infine tornare al livello strutturale mediante Modellistica Molecolare. I risultati preliminari indicano una forte conservazione dell'Arginina centrale presente nel motif d'interazione, mentre nelle altre posizioni del motif si osserva la conservazione di proprietà dei residui piuttosto che la presenza di specifici residui. Questa evidenza è in accordo con il dato di fatto che la mutazione dell'Arginina in L1CAM è responsabile di un serio disordine neurologico, mentre mutazioni a carico di altri residui del motif causano un fenotipo meno grave. Questo suggerisce che il motif è un epitopo posizionalmente conservato attorno all'Arginina centrale in grado di consentire una variabilità limitata ma significativa nella sequenza circostante. Per verificare quest'ipotesi è stata effettuata una superimposizione strutturale dei domini Ig contenenti il motif d'interazione: il risultato ha confermato che il peptide contenente il motif è di per sé conservato posizionalmente e che la conservazione maggiore sia a livello posizionale che struturale è a carico del residuo centrale di Arginina. Esperimenti con peptidi mutati nell'Arginina centrale hanno dimostrato un'attività in termini di segnalazione nella neuritogenesi.

Questi lavori hanno consentito di sviluppare un protocollo bioinformatico per la caratterizzazione di determinanti d'interazione e della loro modulazione funzionale, facilmente trasportabile su altre proteine.

ABSTRACT

In silico investigation on protein domains structure and linear/structural motifs can strongly boost functional analyses and technological design.

Protein surface features study is crucial to understanding Protein-Protein Interactions (PPI); in particular, surface and pockets conservation and variation, in terms of hydrophobicity, steric hindrance and electrostatics can act as driving forces in protein evolution and functional specialization.

Therefore, molecular modeling and structure comparison techniques play an important role in shedding light on "protein behavior" and this PhD work took advantage from integrating computational approaches based on some known molecular modeling methods, such as e.g. Homology Modeling, Fold Recognition, Ab initio Modeling, PBE (Poisson-Boltzmann Electrostatics), Protein-peptide Docking and Hydropathy Analysis with structure and sequence comparison and scanning tools and, of course, with feedback from wet lab analyses performed by co-workers.

Such an integrative approach was followed along investigations on a number of different biological systems:

- **Surface determinants in H5N1 type A Influenza viruses**: Here, an analysis of surface determinants from H5N1 haemagglutinin, involved in host-viral interaction, was completed and then published. Genomic variation is very high in influenza A viruses. However, viral evolution and spreading are strongly influenced by immunogenic features and capacity to bind host cells, depending in turn on the two major capsidic proteins (haemagglutinin and neuraminidase). Current analyses of viral evolution are based on serological and primary sequence comparison; however, comparative structural analysis of capsidic proteins can provide functional insights on surface regions possibly crucial to antigenicity and cell binding. We performed molecular modeling and extensive structural comparison of influenza virus haemagglutinin and of their domains and sub-regions to investigate type- and/or domain specific variation. We found that structural closeness and primary sequence similarity are not always tightly related; moreover, type-specific features could be inferred when comparing surface properties of haemagglutinin subregions, monomers and trimers, in terms of electrostatics and hydropathy. Focusing on H5N1, we found that the variation at the receptor binding domain (RBD) surface intriguingly

relates to branching of still circulating clades from those ones that are no longer circulating.

Recent evidence on the association between electrostatic fingerprints at the haemagglutinin receptor binding surface and the evolutionary success and spreading of H5N1 avian influenza clades prompted us to perform further integrated phylogenetic and structural bioinformatic analysis in H9N2 viruses. In fact, influenza A virus is a zoonotic agent with a significant impact both on public health and poultry industry and switch to human host has been reported for both H5N1 and H9N2 viruses. We performed the evolutionary analysis of a large and non-redundant viral strain dataset, leading to clustering of H9N2 viruses in five groups. Then and according to recent evidence on H5N1, congruence resulted among phylogenetic data and surface electrostatic fingerprints from structural comparison. In particular, surface feature fingerprints could be inferred that relate group specific variation in electrostatic charges and isocontour to well-known hemagglutinin sites involved in modulation of immune escape and host specificity. Results from this second work strengthen suggestion that when integrating up-to-date phylogenetic analyses with sequence-based and structural investigation of surface features may represent a front-end strategy for inferring trends and relevant mechanisms in influenza virus evolution.

- **Domain architecture variation in mammalian protein trafficking**: Human VAMP7b is the most interesting variant among those produced by alternative splicing of the encoding gene SYBL1. Production of VAMP7b variants is determined by skipping of exon 6 which in turn results in coding sequence frameshift. We found that this event is conserved in other mammalian species. VAMP7b shares with the main isoform the N-terminal, inhibitory longin domain and the first half of the SNARE motif. In mammals, VAMP7b is a truncated protein in which the C-terminal half of the SNARE motif and the transmembrane region are replaced by short and variable peptides. Intriguingly instead, only in human and apes sequence frameshift determined by exon 6 skipping results in the creation of a novel unique domain of unknow function, hence human VAMP7b is not truncated but even 40 residues longer than the main isoform. Since existence of such “long” isoform and of its unique domain at protein level were confirmed by specific antibodies, we embarked on in silico dissection of the novel domain by position specific matrix sequence analysis and by *ab initio* structural modeling. Moreover, since the N-terminal region of the

SNARE motif is conserved and it is known to mediate intramolecular binding to the Longin domain, we investigated both in vivo (by two-hybrid in yeast analysis) and in vitro (by NMR analysis) on conservation of the closed conformation. Furthermore, SCL of both VAMP7b and Ykt6b was investigated using GFP and RFP chimeras. Last but not least, b isoforms of the longin genes were analyzed by qPCR and found to be developmentally regulated.

- **Binding motif regulating neurite outgrowth and guidance**: Fine tuning of PPIs by variation in domain architecture or by changing local motifs by surface features modulation can regulate both extracellular and intracellular signaling pathways. Extracellular PPIs can play a central role in heterologous recognition (e.g. host-pathogen) as well as in homologous signaling among cells from the same organism. Proteins exposed at the plasma membrane (PM) can interact each other and with the extracellular matrix (ECM) to provide positional information and guidance cues. Cell adhesion molecules (CAMs) are PM proteins mediating either attractive or repulsive signals by homo- and heterophilic interactions of their extracellular domains (Eds). CAM EDs are most often composed by Ig-like or Fibronectin type III fold repeats. Current evidence suggests that the four N-terminal Ig 1-4 domains of CAM EDs play a major role in such homo- or heterophilic interactions and in particular an important interaction motif is contributed by repeat Ig2. In our lab, biomimetic peptides have been developed by reproducing the known or predicted interaction motifs from the Ig2 domain of human L1CAM and the single Ig domain of human LINGO1, i.e. two proteins that play a crucial role in neurite outgrowth and guidance and in neuronal differentiation. Based on the somehow surprising structural and sequence conservation of the motif region (even when proteins show very different ED architectures), we started investigating on variation and conservation of the putative motif region by means of homology search, regular expression and finally by structural modeling and comparison. Preliminary results highlighted strong conservation of the central Arg residue in the interaction motif, while in other positions of the motif residue properties rather than specific residues are conserved. Such evidence is in agreement with finding that mutation of such residue in L1CAM is responsible for a severe neurological disorder, while mutations at other residues of the motif, results in less severe phenotype. This suggests the motif is an epitope positionally conserved around the central Arg allowing limited, but significant structural variability in surrounding sequence. In order to check such a hypothesis,

a structural superposition of the Ig domains containing the interaction motif was performed, confirming that the peptide motif itself is positionally conserved but the highest positional and structural conservation concerns the central Arg residue. Experiments with peptides mutated in the central Arg showed biological activity of these peptides in terms of neuritogenesis signalling.

These works carry out a bioinformatic protocol for the characterization of interaction determinants and their functional modulation, easily transportable to other proteins.

AIMS OF THE THESIS

This PhD thesis is focused on the study of protein-protein interactions in different biological systems.

Influenza A (AI) viruses are zoonotic agents, having a high impact on humans and animals due to infectious and contagious diseases caused. Therefore, great interest turns on AI viruses in terms of surveillance and vaccination strategies. Features common to AI viruses are the high mutation rates, rapid replication and infection. As a consequence, viral population is characterized by related co-existing variants undergoing the environmental pressure. In order to escape host immunity response, novel viral strains arise due to mutations or recombination of their genome. These events lead to variations in surface proteins such as haemagglutinin. Therefore, vaccines become ineffective. A combination of different bioinformatic approaches can help in understanding viral evolution and genetic variability. Structural analyses show us that mutations are not the same: in fact they can be silent, compatible or producing huge effects on protein folding and/or surface features (e.g. charges disposition). Fortunately, an increased number of available protein 3D structures allowed to perform wide comparisons.

SNARE proteins take on great interest in the lab hosting me. In fact, Longins and LD were discovered and characterized here (Filippini *et al*., 2001). We also showed that different VAMP7 isoforms exist, due to the alternative splicing of SYBL1 gene. LD and SNARE motifs are linked to neuronal diseases, as revealed by the implication of VAMP7 in neuronal plasticity and potential mental illness (e.g. bipolar disorders). Again, a bioinformatic integrated approach permitted to shed light on VAMP7b isoform. This variant is characterized by having a LD and the N-ter part of the SNARE motif followed by a 116 residues novel region of unknown function that - when considering the VAMP7 involvement in brain development and cognitive functions - is intriguingly specific to humans and apes. Given that no 3D structure of this isoform is available, structural modelling (using all main approaches) was adopted as a tool for obtaining functional inference; some predictions have already been confirmed while further lab work is planned for validating latest suggestions.

Proteins from the L1 family are involved in neuritogenesis and axon guidance. L1-CAM proteins are able to mediate both homo- and heterophilic interactions with Ig repeats in their extracellular domain. In particular, the second repeat (ig2) plays a major role in such interactions. Given that expression and purification of whole proteins ectodomains is quite expensive and time consuming, we aimed at designing, producing and testing synthetic

peptides reproducing the active motif involved in homo/heterophilic interactions of the Ig-like domains, hopefully able to mime neuritogenic and guidance cues from the original domains. Once again, several predictions led to successful experimental validation and last wet lab experiments are ongoing to get this work complete.

This thesis is presented as follows: an introduction section explaining the importance of protein surfaces from a structural point of view and bioinformatics approaches used to study it, then specific workpackages and results are presented and discussed as chapters. Published or submitted manuscripts are included. For readers' convenience, a list of manuscripts and a short description of each chapter is presented hereafter:

CHAPTER 1

Manuscript: Righetto I, Milani A, Cattoli G, Filippini F. Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features. BMC Bioinformatics. 2014 Dec 10;15:363.

Summary: In this work, we performed extensive structural comparison of influenza virus haemagglutinins and of their domains and subregions to investigate type- and/or domain-specific variation. We found that structural closeness and primary sequence similarity are not always tightly related; moreover, type-specific features could be inferred when comparing surface properties of haemagglutinin subregions, monomers and trimers, in terms of electrostatics and hydropathy. Focusing on H5N1, we found that variation at the receptor binding domain surface intriguingly relates to branching of still circulating clades from those ones that are no longer circulating. This work suggests that integrating phylogenetic and serological analyses by extensive structural comparison can help in understanding the 'functional evolution' of viral surface determinants. In particular, variation in electrostatic and hydropathy patches can provide molecular evolution markers: intriguing surface charge redistribution characterizing the haemagglutinin receptor binding domains from circulating H5N1 clades 2 and 7 might have contributed to antigenic escape hence to their evolutionary success and spreading.

Manuscript: Heidari A[#], Righetto I[#], Filippini F. Comparative structural analysis of hemagglutinin for unveiling fingerprints in the evolution and spreading of avian influenza viruses (Submitted) #: Equal contribution

Summary: In this work, we reported that comparative structural analysis of hemagglutinin can provide relevant evolutionary fingerprints to integrate sequence-based analyses. Phylogenetic analyses, carried out with different methods, of H9N2 viral strains from wild birds and poultry reliably led to clustering of viruses into five main groups. Then, structural features comparison showed congruence among such a clustering and surface fingerprints. These latters relate group specific variation in electrostatic charges and isocontours to well-known hemagglutinin sites involved in the modulation of immune escape and host specificity. This work suggests that integrating structural and sequence comparison may boost investigation on trends and relevant mechanisms in viral evolution.

CHAPTER 2

Summary: After an introduction about SNARE proteins structure and role in subcellular trafficking, we focused on their importance in neurosciences. In particular, we investigated on Human VAMP7b, a most interesting variant among those produced by alternative splicing of the encoding gene SYBL1. Production of VAMP7b variants is determined by skipping of exon 6 which in turn results in coding sequence frameshift. We found that this event is conserved in other mammalian species as well as in the other two prototypical longin genes Ykt6 and Sec22. VAMP7b shares with the main isoform the N-terminal, inhibitory longin domain and the first half of the SNARE motif. In mammals, VAMP7b is a truncated protein in which the C-terminal half of the SNARE motif and the transmembrane region are replaced by short ad variable peptides. Intriguingly instead, only in humans and primates sequence framshift determined by exon 6 skipping results in the creation of a novel unique domain of unknown function, hence human VAMP7b is not truncated but even 40 residues longer than the main isoform. Since existence of such “long” isoform and of its unique domain at protein level were confirmed by specific antibodies, we embarked on in silico dissection of the novel domain by position specific matrix sequence analysis and by ab initio structural modeling. Moreover, since the N-terminal region of the SNARE motif is conserved and it is known to mediate intramolecular binding to the Longin domain, we investigated both in vivo (by two-hyb in yeast analysis) and in vitro (by NMR and chemical shift analysis) on conservation of

the closed conformation. Furthermore, SCL of both VAMP7b and Ykt6b was investigated using GFP and RFP chimeras. Last but not least, b isoforms of the three longin genes were analyzed by qPCR and found to be developmentally regulated.

CHAPTER 3

Summary: In this workpackage I provided a description of Cellular Adhesion Molecules (CAMs) in terms of structures and functions, focusing on their role in neuroscience in promoting neurite outgrowth and guidance (NOG). We investigated the possibly general conservation of a functional NOG motif derived from CAMs ectodomain, in order to characterize and test biomimetic peptides able to reproduce the neuritogenic effect of the whole proteins. Wet lab analyses on the set of studied CAM/ECM proteins and peptides were derived from have already confirmed several predictions from this *in silico* work. When novel peptides likely to be biomimetic were used with neuronal precursors, they were all confirmed to mediate comparable neuritogenic effects.

Introduction

1. *PROTEIN SURFACE PROPERTIES AS DRIVING FORCES IN PROTEIN FUNCTION*

The existence of the correlation between protein surface features and their "behavior" is well known. Understanding the structure-function relationship is essential for unravelling interaction mechanisms. On the other hand, protein surface comparison may be more useful than structural comparison in the case of proteins with different folds but sharing similar chemical/physical features at their surface.

Exploring protein surface is very important but conventional wet lab techniques such as NMR and X-ray crystallography are not sufficient to achieve this goal due to several limitations (i.e. transient complexes are difficult to obtain) (de Vries *et al*., 2006). In fact, a huge amount of interacting complexes are not retrieved by experimentally determined 3D structures (Mosca et al., 2013). Therefore, bioinformatic tools (i.e. clustering techniques) are created in order to fulfill this gap (Baldacci *et al*., 2006).

When protein structures were not retrieved in PDB, I used molecular modeling techniques (i.e. homology modeling, *ab initio* modeling) to obtain protein models.

1.1 Surface properties

Hydrophobicity

Defined as "the major force which stabilizes protein-protein associations" (Chotia and Janin, 1975), hydrophobicity is based on interactions between polar/non-polar atoms at the protein surface with the solvent. While polar atoms are able to make hydrogen bonds with surrounding water, non-polar atoms are not. So, water molecules arrange hydrogen bonds between them in order to reduce the contact with non-polar surface, giving birth to the hydrophobic effect (Fig 1). Subsequently, water molecules close to the surface are more ordered than in a bulk solvent, and both a local decrease in entropy and an unfavorable free energy of solvation occur (Gruber *et al*., 2007). The free energy (ΔG) associated with hydrophobic effect is responsible in defining the structure of globular proteins and in governing protein interactions. As reported by Pace *et al*, 2014, the contribution ($\Delta(\Delta G)$) of hydrophobic interactions to protein stability was assessed through experiments in which protein variants showing a loss of a hydrophobic buried group are compared with the wild type. The results from these experiments indicate that $\Delta(\Delta G)$ values are determined by a constant term depending on the difference in hydrophobicity between the wild type and

variant side chains and a variable term that depends on the difference in the Van der Waals interactions of the side chains. However, the treatment of hydrophobic effect is not addressed with a defined framework as well as for continuum electrostatics (Gruber et al., 2007). Hydrophobicity can be measured by scales commonly derived from the portioning of model compounds between an aqueous and an oil-like phase; in my PhD project, I used the Kyte-Doolittle one. However, other methods have been recently developed for hydropathy calculations: based on Molecular Dynamics simulations (Schauperl *et al*., 2016), or on hydrophobicity score assignment based on the hydrogen-bonding capacity of atoms or functional groups (Gruber *et al*., 2007).

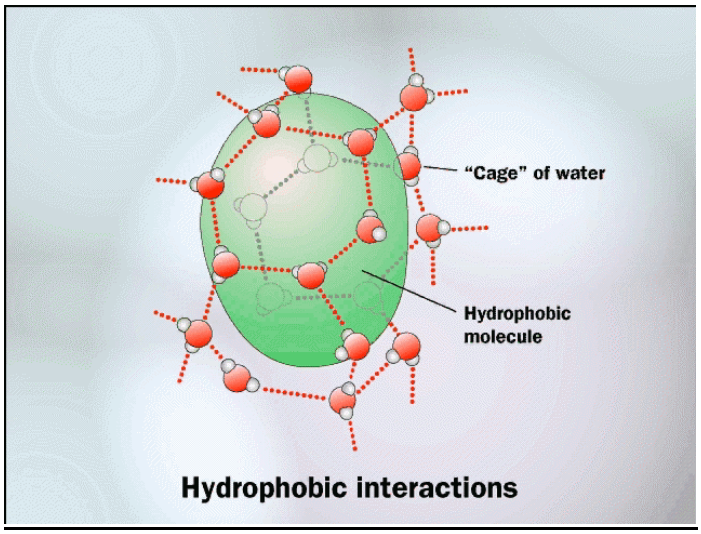


Fig.1. Graphical representation of the hydrophobic effect. High ordered water surrounds the non polar molecule. From http://myhome.sunyocc.edu/~weiskirl/water.htm

Electrostatics

As for hydrophobicity, electrostatic features play an important role in macromolecular interactions (Ritchie and Webb, 2015; Sheinerman *et al*., 2000). As shown by Honig and Nicholls in 1995, electrostatics is fundamental in determining specificity, thermodynamics and kinetics of protein binding. Therefore, modeling electrostatic potential in proteins is crucial to understanding their function, stability, and interactions. For example, electrostatics can explain why binding of a certain ligand to a protein can be affected even if a mutation is far from the binding interface. Electrostatic properties depend on the distribution of whole and partial charge across the protein structure. Charge interactions are long-ranged, whereas interactions between uncharged atoms are short-ranged and diminish with

distance as r^{-6}. On the other hand, Coulombic interactions diminish with distance as r^{-1}. Charge interactions are weaker in liquids than in vacuum. Liquid molecules tend to reorient their charges when two fixed charges (one positive and one negative) are separated by a distance r in this liquid. Positive charges in the medium orient toward the negative fixed charge, and vice versa. Therefore, the liquid can be polarized. This event shields and weakens the interactions between the two fixed charges and the dielectric constant describes the weakening of the coulombic interactions (Dill and Bromberg, 2003). Media having high dielectric constants are able to strongly mask charge interactions. Coulomb's law allows us to describe electrostatic potential V(r) as follows:

$$V(r) = \frac{1}{4\pi\varepsilon_0\varepsilon}\frac{q}{r}$$

where:

q = charge

r = distance

ε = dielectric constant relative to vacuum permettivity

ε_0= dielectric medium

However, Coulomb's law is not suitable for describing electrostatics in proteins: in fact this expression describes a system with a single dielectric medium whereas proteins show a hydrophobic core (with a dielectric constant similar to those in vacuum) enveloped by solvent. Therefore, electrostatic calculations are carried out using LPBE (Linear Poisson-Boltzmann Equation), implemented in APBS:

$$-\nabla\varepsilon\cdot(r)\nabla\phi(r) + \varepsilon_0\varepsilon(r)k^2(r)\phi(r) = \frac{4\pi e^2}{\varepsilon_0 k_B T}\sum_{i=1}^{F} z_i\delta(r-r_i)$$

Where:

ϕ = electrostatic potential

ε = dielectric coefficient

κ = ion accessibility

The parameter κ describes the quantity and type of mobile ions and it is dependent on the ionic strength (I):

$$k^2(r) = \frac{4e^2 I}{\varepsilon_0 k_B T}$$

Explicit model systems are represented by many atoms; some of them are connected via chemical bonds, others are able to interact via van der Waals or electrostatic non-bonded interactions. In this thesis, the system is treated as implicit: this way, dynamic effects of water are not directly internalized, leading to a better analysis of electrostatics (Gorham *et al*, 2011) (Fig.2, Fig.3).

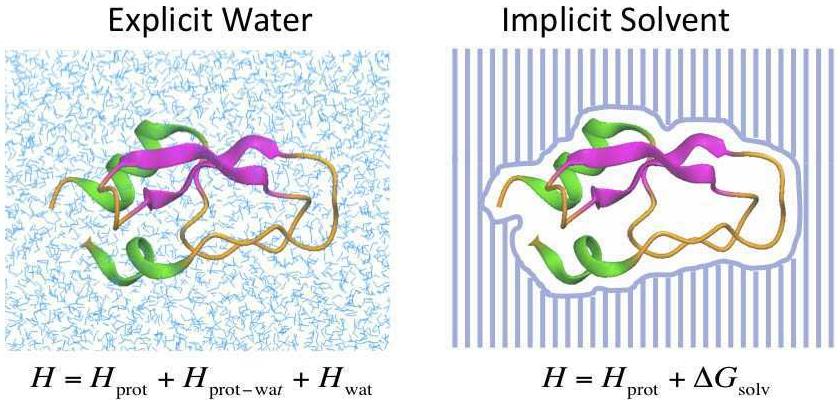


Fig.2. Implicit vs. Explicit solvent; water dynamic effects are not taken into account in implicit solvent. From http://www.k-state.edu/bmb/labs/jc/research.html

Protein dielectric constants express the effect of the protein environment, reflecting protein structure and sequence properties (Schutz and Warshel, 2001). Implicit solvent calculations (continuum electrostatics) provide a water phase atomic detail reduction and intrinsically an equilibrium solution (Li *et al*. 2014). In these workpackages, the medium dielectric constant was set at 80 (ε = 80, water), whereas a dielectric constant of 2 (ε = 2) is used for protein. This latter value should account for electronic polarization and small backbone fluctuations.

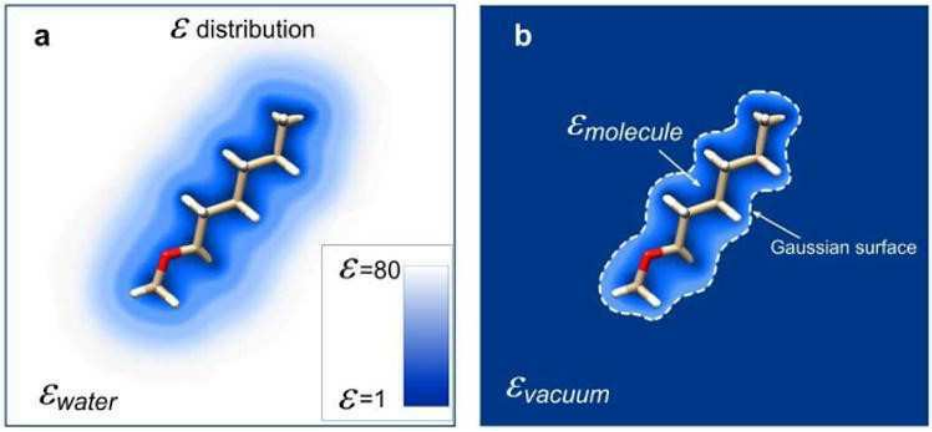


Fig.3. Protein dielectric constant. From: Li L, Li C, Alexov E. On the Modeling of Polar Component of Solvation Energy using Smooth Gaussian-Based Dielectric Function. *J Theor Comput Chem*. 2014 May;13(3).

Choosing the right value of the protein dielectric constant is not a trivial task. No optimal value for that can be retrieved in literature (i.e. protein stability is investigated using different dielectric constant values, from ε=1 (Mobley *et al*., 2008) to ε=40 (Vicatos *et al*., 2009)). Values of ε=1 or ε=2 were used in Molecular Dynamics (MD) studies to deliver the energies via Molecular Mechanics Poisson-Boltzmann (MMPB) or Generalized Born (MMGB) methods (Gouda *et al*., 2003; Kollman *et al*., 2000). Works by Schutz et Warshel 2001 and Warshel et al., 2006 considered the solute and the water phase conformational reorganization crucial in choosing the optimal value of the internal dielectric constant; a large dielectric constant can be used when a large conformational change is involved; otherwise, reactions not inducing big conformational changes can be modeled with low dielectric constant.

At first, partial charges and van der Waals radii were assigned using PDB2PQR (Dolinsky *et al*., 2004) and the PARSE force field (Sitkoff et al., 1994). A force field is made up of equations used to calculate the potential energy and forces from particle coordinates. In all force fields, potential functions are subdivided into "bonded interactions" (i.e. covalent bond-stretching, angle-bending, torsions potential, out-of-plane improper torsion) and "non-bonded interactions" (Lennard-Jones repulsion and dispersion and Coulomb electrostatics):

$$E_{bonded} = \sum_{bonds} K_b(b - b_0)^2 + K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi[1 + \cos(n\chi - \sigma)] \quad \text{Eq. 1}$$

$$E_{nonbonded} = \sum_{\substack{nonbonded \\ pairs\ ij}} \left(\varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{R_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{R_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \right) \quad \text{Eq. 2}$$

The total energy is:

$$E_{total} = E_{bonded} + E_{nonbonded} + E_{other} \quad \text{Eq. 3}$$

E_{other} is referred to any force field-specific terms.

In Eq.1, b represents the bond length, whereas K_b and b_0 describe the stiffness and the equilibrium length of the bond. The second term is referred to the bending of angles. Here, triplets of atoms are involved (i.e. A, B, C, where A is bonded to B and B is bonded to C). θ is the angle formed by the two bond vectors, K_θ and θ_0 describe the stiffness and equilibrium geometry of the angle. The final term involved four atoms (i.e. A, B, C, D, where A is bonded to B, B to C and C to D) defining a dihedral angle. This term describes the energetics associated with the rotation of this dihedral. χ is the dihedral value, K_χ is the energetic parameter determining barrier heights, n is the periodicity and σ is the phase. The function

cos is the expression of the periodicity of the rotation. Due to the addition of 1 in this final term, the energy is equal to or greater than zero.

Eq.2 is made up of two parts; the first one is referred to the Lennard-Jones (LJ) equation, modeling attractive dispersion and repulsive Pauli exclusion interactions. LJ part is known as the van der Waals term. Dispersion is modeled by the negative part of the LJ equation: when two atoms are brought together from infinite separation, this negative part dominates the interaction and the attraction is higher with decreasing distance. During this process, an energy minimum is reached and, at closer distance, the positive term starts to dominate and repulsion occurs. The pre-factor ε_{ij} parameter depends on the types of the two interacting atoms i and j. At increasing values, the interaction minimum becomes deeper and the repulsive wall steeper. $R_{min,ij}$ is a parameter defining the distance at which the LJ energy is at minimum. The second part of Eq.2 is Coulomb's law, modeling electrostatic interections between pairs of non-bonded atoms. q_i and q_j describe the charges on atoms i and j. These are partial atomic charges with noninteger values, selected to represent the overall molecule charge distribution. In the case of metal ions, the charge assigned is the formal one (Kukol, 2008)

Once the protein is prepared this way via PDB2PQR, PBE calculations are ready to be carried out. APBS superimposes the molecule onto a 3D grid, assigning values for charge, dielectric coefficient and ionic strength at every grid point. The protein is then wrapped by two surfaces: the ε one defines the dielectric coefficient boundary. It is defined by a water-sized rolling sphere on the protein van der Waals surface and the center of this sphere defines a new surface. The κ surface is referred to ion accessibility and the rolling sphere defining it has the size of an ion. At the end of calculations, a surface electrostatic map (.dx file) is computed and it can be visualized onto the molecule as isopotential contour (Fig.4):

Fig.4. Electrostatic representations: surface projection (left) and isopotential contours (right). Positive charges are highlighted in blue, negative ones in red. Images obtained via UCSF Chimera (Pettersen *et al.*, 2004).

Isopotential contours are plotted at levels of different $\pm nk_BT/e$; this terminology means that the electrostatic component of the potential energy of interaction between the protein field and the elementary charge of +1*e* located somewhere on the + nk_BT/e isopotential surface, would equal:

$$\frac{(n \cdot k_B T)}{(+1 \cdot 1.602E{-}19\ Coulombs)} \qquad \text{Eq.4}$$

$$= n \cdot (8.314\ Joules/Kelvin/6.022E + 23) \cdot 298K) \ / \ (+1 \cdot 1.602E - 19\ Coulombs)$$

$$= n\ Volts$$

k_BT/e can be considered as an energy density.

Surface conservation

Here, the underlining idea is that conserved residues in a protein are fundamental to its function / structure: slow-evolving sites on the protein surface are usually important for function (Lichtarge *et al*., 1996; Nimrod *et al*., 2005), whereas slow-evolving portions at the core are important for structural stability and folding (Kessel *et al*, 2010). Many studies confirmed that active- and ligand-binding sites residues are more conserved than general surface residues in different protein families (Grishin and Phillips, 1994; Ouzounis et al., 1998; Bartlett et al., 2002; Caffrey et al., 2004). By estimating the evolutionary rates of amino- and nucleic acids, ConSurf software is able to detect crucial sites within the query macromolecule.

Surface conservation analysis can be addressed using different approaches: sequence-based (Watson *et al*., 2005), structure-based (Laskowsky et al., 2005) or mixed (Watson *et al*., 2005). Software as Consurf (Celniker *et al*., 2013) are able to project conservation scores onto the molecule. Obviously care must be taken with sequences sharing 30% ide because of the uncertainty of the sites prediction.

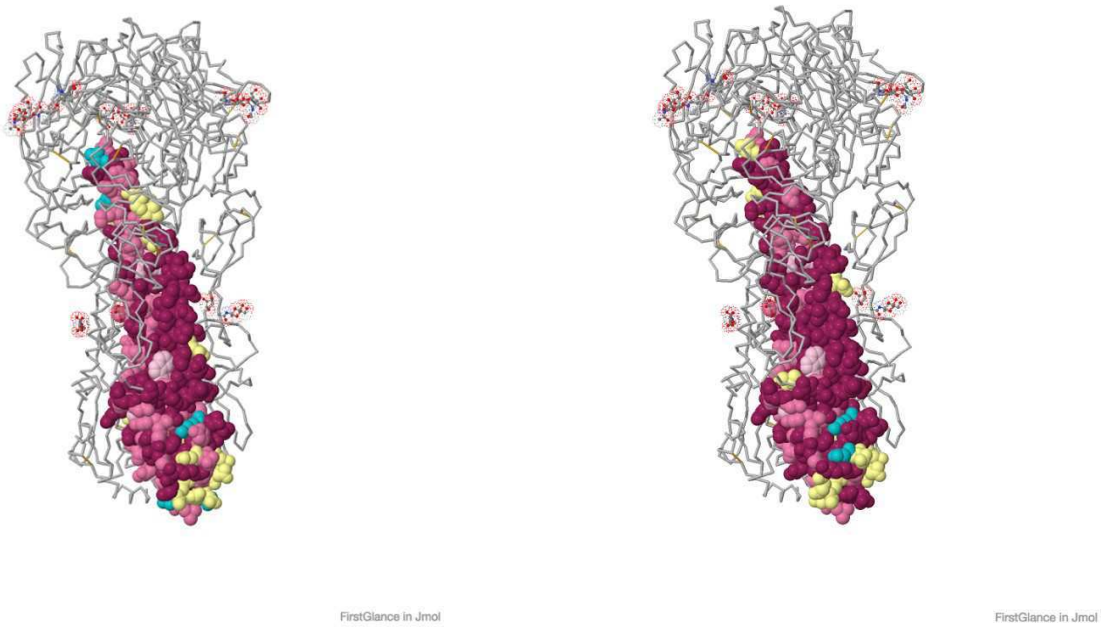


Fig.5. Consurf results: residues are colored according to the conservation scale from dark turquoise (variable) to dark red (conserved). Insufficient data are highlighted in yellow. At left, H5N1 avian stem, at right H5N1 mammal stem. Image obtained via CONSURF.

1.2 Surface description

A protein surface can be described in more than one way:

Van der Waals surface

In this surface representation, atoms making up proteins are represented via their van der Waals radius. It could be thought as the surface through which the molecule might be conceived as interacting with other molecules. The van der Waals surface is the basis for other surfaces developing.

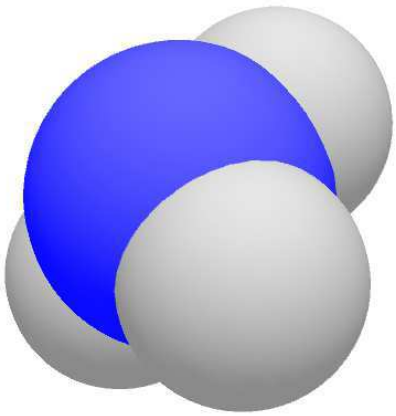


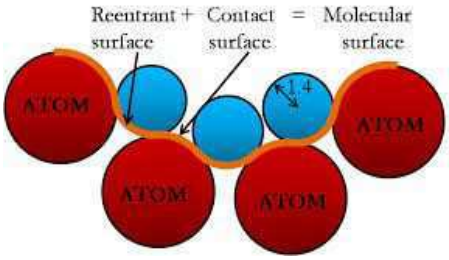
Fig.6. Van der Waals surface. Atoms in the molecule are represented by overlapping spheres.

Solvent Excluded Surface (SES)

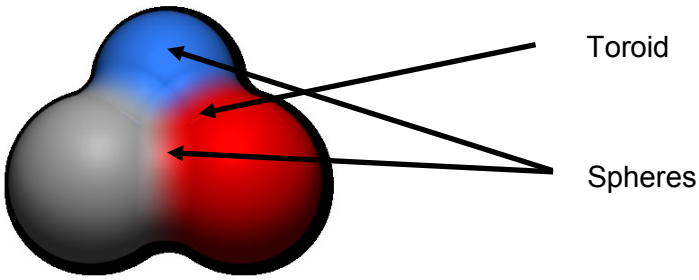
This kind of representation is also known as Connolly surface or molecular surface. This surface is generated by a spherical probe, standing for the solvent (water), of radius 1.4 Å, rotating over the Van der Waals surface. Here, the spaces inaccessible to the solvent are covered with a reentrant surface (Fig.7.a.). As van der Waals surface, also SES represents the interacting surface of a molecule. The ratio contact surface to re-entrant surface can be a measure of the molecular surface roughness. Connolly surfaces are complementary at the interface between two molecules (eg. ligand/binding pocket). A Connolly surface is made up by (Fig.7.b,c.):

- Contact surface elements, defined by convex spherical ones
- Reentrant surface elements defined by saddle-shaped toroidal and concave spherical ones.

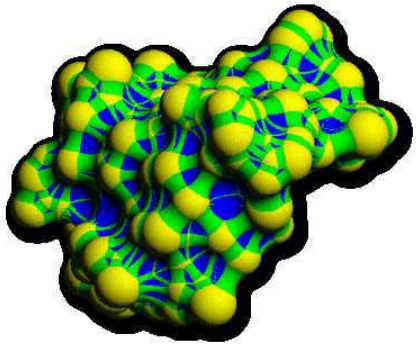
Elements geometric properties calculations are carried out via equations implying van der Waals radii and probe radius. Each element surface is defined by a set of circular arcs, for which the center, the radius and the end points need to be determined (Connolly, 1985).



a.



b.



c.

Fig.7. Connolly surface (SES) representation. Panel c depicts convex spherical patches colored yellow, saddle-shaped pieces of tori colored green and the concave reentrant surface colored blue. From: https://wiki.cmbi.ru.nl/index.php?title=File:Connolly.png&limit=50
http://compbio.biosci.uq.edu.au/mediawiki/upload/8/8a/MolecularSurfaces.pdf

Solvent Accessible Surface (SAS or ASA)

SAS is also known as Lee and Richards surface and this is the most commonly used representation in structural biology. It is very similar to the Connolly surface, but here the solvent accessible surface is generated from the center of the solvent probe. This feature is

relevant because the SAS of a given atom is proportional to the number of simultaneously contacting it solvent molecules. So, this characteristic becomes important when dividing the surface based on property. In fact, whereas in Connolly representation the reentrant regions of a molecular surface have a property associated with two or more atoms, SAS is associated with only a single protein atom (Fig.8.). This is the reason for the success of this surface representation.

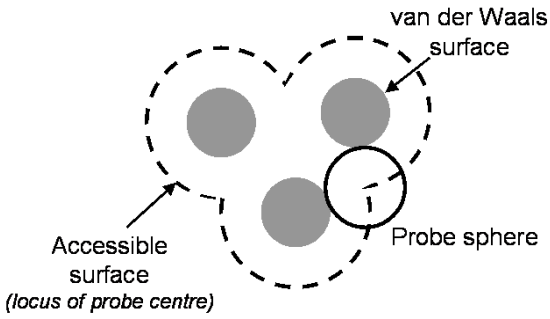


Fig.8. Solvent Accessible Surface (SAS), defined as the locus of the centre of a probe sphere rolling over the Van der Waals surface.

Various software are able to calculate surfaces (MSMS, MS, Molecular Surface, SURF, etc). In particular MSMS is implemented in UCSF CHIMERA, the visualization package adopted in this thesis.

Managing molecular surfaces is fundamental for different purposes as:

- Exploring variations with time, during simulations
- Comparing between different states as folding or binding
- Calculating contact/interaction surface area between proteins or domains (as shown in Fig.9.)

$$C_{A,B} = S_A + S_B - S_{A \cup B}$$

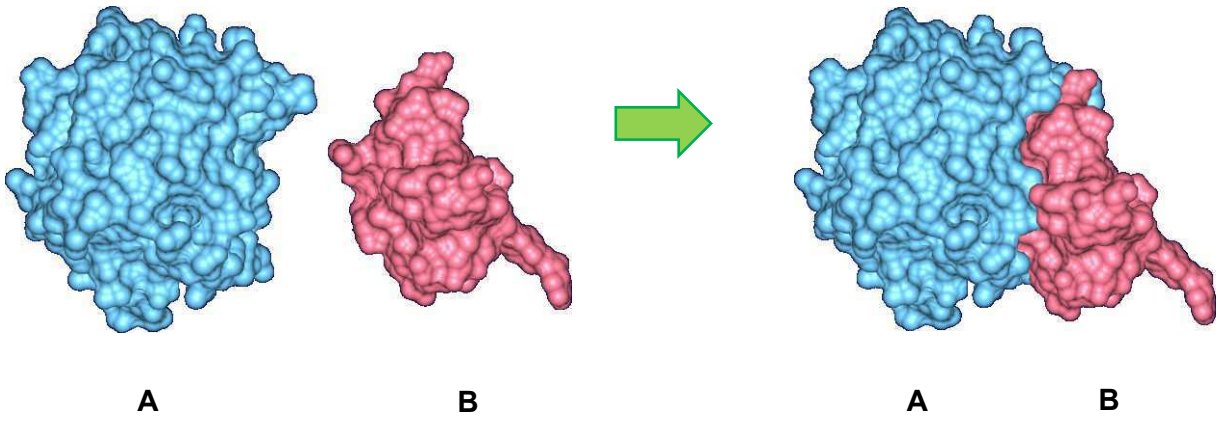


Fig.9. Alpha-chymotrypsinogen complex with Trypsin inhibitor. From: http://www.cs.umd.edu/gvil/rimages/cgi_undock.gif

- Studying hydration in implicit solvation models (GBSA)

$$\Delta G_{solv} = \sum_{i=1}^{n} a_i \sigma_i \qquad \text{Eq. 5}$$

where:

ΔG_{solv} = free energy of solvation of a solute (n atoms)

a_i = accessible surface area of atom *i*

σ_i = solvation parameter of atom *i* (contribution to the free energy of solvation of atom *i* per surface unit area)

- Studying protein-ligand and protein-protein interactions (e.g. docking of a ligand in a binding pocket, identification of possible antigenic determinants on viruses).

1.3 Protein-protein interactions (PPIs) and Protein-ligand interactions

The interactions of a protein with other molecules are perhaps the most important features linked to the protein surface. Accessible surface area (ASA), developed by Lee and Richards in 1971, was used by Jones and Thornton in 1995 and 1996 to explore protein-protein interfaces. Assuming the hydrophobic interaction (Chotia and Janin, 1975) – the gain in free energy which occurs when non-polar residues of proteins associate in an aqueous environment (Kauzmann, 1959) – and the complementarity in terms of shape and electrostatics – charge of groups of the surface – as driving force and specificity of PPIs (Fersht, 1984), PPI interfaces can be defined at two levels:

- *The residue level*: The interface residues were defined as those residues within side chain possessing an ASA that decreased by >1Å^2 on dimerization. This value was used to account for errors in computational inaccuracies in the calculation of the ASAs. Exterior and interior residues were based on their relative ASA (0% for residues with no atom contact with the solvent, 100% for fully accessible residues). Miller *et al*., 1987 defined a cutoff of 5% to show exterior (having a relative accessibility > 5%) and interior residues (having a relative accessibility $\leq$ 5%)
- *The atom level*: The interface was defined as those atoms that showed a decrease of 0.01 Å^2 in ASA on dimerization. The interior was defined as those atoms with an atomic ASA of zero and the exterior as those atoms with an atomic ASA > 0.

PPIs can also be categorized in different types (Nooren and Thornton, 2003):

- *Homo- and hetero-oligomeric complexes*: PPIs occur between identical or different chains
- *Non obligate and obligate complexes*: Components of obligate PPIs are not found as stable structures *per se* in vivo (eg. Arc repressor dimer, essential for DNA binding), on the other hand non obligate complexes are made up independently existing protomers.
- *Transient and permanent complexes*: A permanent interaction only exists in its complexed form, on the other hand a transient interaction associated and dissociated in vivo.

Moreover, PPIs can be evaluated from an evolutionary perspective (Jones and Thornton, 1996). Residues in the enzyme active site are pressed by evolution whereas protein families characterized by a non-catalytic activity obviously lack this evolutionary pressure. From a structural point of view, geometry (in terms of size, shape and complementarity) and chemical nature (in terms of amino acid composition, hydrophobicity, electrostatic interactions, hydrogen bonds) underlie protein-protein recognition sites (Chackrabarti and Janin, 2002). Protein-protein interfaces can be taught as patches having a "standard size" of 1200-2000 $Å^2$ bury surface area (Lo Conte *et al*, 1999), surrounded by a rim of residues having a protein surface-like composition. Trp and Tyr are favored in the core of recognition site, while Ser and Thr are not. Moreover, Ala-scan experiments revealed that charges in binding energies ($\Delta\Delta G$) > 2 kcal/mol characterize residues clustered at the center of the recognition sites (Bogan and Thorn, 1998).

Protein-ligand / peptide interactions are crucial in a wide range of biological process (eg. signalling pathways) and represent an interesting target for potential therapeutic applications. Protein-ligand interactions are defined by some features:

- *The presence of pockets in the protein receptor*: This is the preferred way, for the peptide, to form extensive contact. The largest pockets on the protein surface receptor are the favorite for the binding. (London *et al*, 2010)
- *Different peptide motifs adopted after the interaction with a protein*: The conformation of the bound peptide often reflects a low-energy form. Adopted motifs (Stanfield and Wilson, 1995) are the followings:
    - Extended chain (eg. Neurofascin and biomimetic peptides)
    - β-turn (eg. Fab-peptide structures)
    - α-helix (eg. peptides bound to Ca^{2+}-dependent Calmodulin)
- *The lack of conformational changes in protein receptor after the peptide docking and enthalpy maximization via hydrogen bonds* (London *et al*., 2010)
- *The presence of hot spots in peptide interfaces* (London *et al*., 2010): This feature plays the major role in stabilizing protein-peptide interactions. Hot spots are enriched in Leu, Phe, Tyr, Trp, Ile.
- *The presence of some peptide positions able or not to contribute to the binding affinity*: It is well-known the ability of many domain families to recognize peptide motifs (e.g. SH3 domains are able to bind sequences containing P-X-X-P, where X is any amino acid and P is proline).
- *Peptide-protein interface has a size of ~ 500 $Å^2$* (London *et al*., 2010).

Protein complexes can be modeled *in silico* via docking experiments. All docking software are characterized by two phases, named as sampling and scoring.

Sampling conformational space is a big challenge. In fact, taking into account the conformation flexibility and at the same time an exhaustive search of all possible protein-protein interfaces is unrealistic. Computational costs can be reduced if interaction sites are known, otherwise proteins are considered as rigid bodies by docking programs. However, methods as FFT correlation approach and Geometric Hashing are able to cover the entire accessible interacting surface (Soni and Madhusudhan, 2017). Conformational space can be sample also by Molecular Dynamics (MD) simulations. Even if I didn't use this approach as primary choice in this PhD thesis, MD takes place in other software I used (i.e. *ab initio* modeling) thus talking briefly about this computational technique is appropriate. Using Newton's equations of motion, MD enables us to follow the thermal motion of a protein (in this case). This powerful toolbox represents atoms as hard spheres whereas bonds as springs. Potential energy from particle coordinates are calculated via forcefields (see pag. 16 for explanations). Then, forces driving atoms motions can be obtained by potential energy applying the first derivative of it:

$$F_i = -\frac{dV}{dr_i} \qquad \text{Eq. 6}$$

where V is the potential energy.

The position of an atom after a Δt (time step) can be computed using the Verlet algorithm:

$$x(t + \Delta t) = x(t) + \frac{dx(t)}{dt}\Delta t + \frac{d^2x(t)}{dt^2}\frac{\Delta t^2}{2} + \cdots \qquad \text{Eq. 7}$$

where:

$x(t)$ = position

$\frac{dx(t)}{dt}\Delta t$ = velocities resulting from kinetic energy

$\frac{d^2x(t)}{dt^2}\frac{\Delta t^2}{2}$ = acceleration

Scoring scheme implemented in docking software should distinguish the native, or near native structure from non-native conformations. To address this goal, scoring techniques can combine different features such as solvation energy, electrostatics, van der Waals interaction, hydrogen bonds, clashes, etc (Soni and Madhusudhan, 2017).

CHAPTER 1

Surface determinants in H5N1/H9N2 type A Influenza viruses

State of the art

1. *INFLUENZA A VIRUSES*

Influenza viruses are RNA viruses and belong to the *Orthomyxoviridae* family; they are classified into three types: A, B, C and D on the basis of antigenic differences in their nucleoprotein (NP) and matrix proteins (MP) (Hamilton *et al*., 2012; Han and Marasco, 2011; Ferguson *et al*., 2016). Seasonal flu epidemics are caused especially by A and B influenza viruses, but type A is responsible for more severe clinical effects in human (Han and Marasco, 2011) and also in animal population. Influenza virus C is not responsible for epidemics but only for mild respiratory illness and type D affects primarily cattles (Ferguson *et al*., 2016).

Their rapid evolution due to genetic shift (i.e. one or more gene segments is exchanged between different virus subtypes) and genetic drift (i.e. mutations accumulation in viral gene) caused by the relatively error-prone replication of the viral RNA, makes them a challenge for vaccine design (Stray and Pittman, 2012).

Wild water birds are thought to be the natural reservoir for influenza A virus (Munster *et al*., 2007), able to infect other avian or mammal hosts (Sriwilaijaroen and Suzuki, 2012). However, other species have been described to act as an infection pool, such as bats (Tong *et al*., 2013).

Influenza A virus is characterized by 18 haemagglutinin (HA) subtypes and 11 neuraminidase (NA) subtypes. 16 HA and 9 NA subtypes were found in wild avian species whereas H17N10 and H18 N11 in bats (Tong *et al*., 2013). Usually, avian influenza viruses circulate among the natural host birds, but sometimes they are able to infect different animal hosts (Vandegrift *et al*., 2010; Sriwilaijaroen and Suzuki, 2012). In particular, H5N1, H7N2, H7N3, H7N7 and H9N2 are of particular interest because of their animal-human and human-human transmission. Every year, seasonal influenza epidemics are responsible for 250.000-500.000 deaths (Han and Marasco, 2011; Sriwilaijaroen and Suzuki, 2012). The famous “Spanish” influenza pandemic outbreak occurred in 1918-1919 was responsible for about 100 million deaths worldwide (Johnson *et al*., 2002). Moreover, influenza viruses heavily affected poultry industry worldwide (see the H7N1 epidemic in Italy in 1999-2001, the epidemic in the Netherlands in 2003 and the H7N3 in Canada in 2004). Therefore, studying the structure-based mechanisms contributing to the viral evasion of the host immune response and viruses high adaptability is crucial to develop vaccines and/or drugs.

1.1 Influenza virus structure description

Influenza A viruses are pleomorphic and enveloped; they are roughly spherical (80-120 nm of diameter) or can exhibit a filamentous form (>300 nm). The genome is made up of 7 (in influenza virus C) - 8 (in the other viruses) negative single-strain RNA segments, coding for nine structural (PB1, PB1-F2, PB2, PA, HA, NA, NP, M1, M2) and non-structural proteins (NS1, NS2) (Fig.10, A)

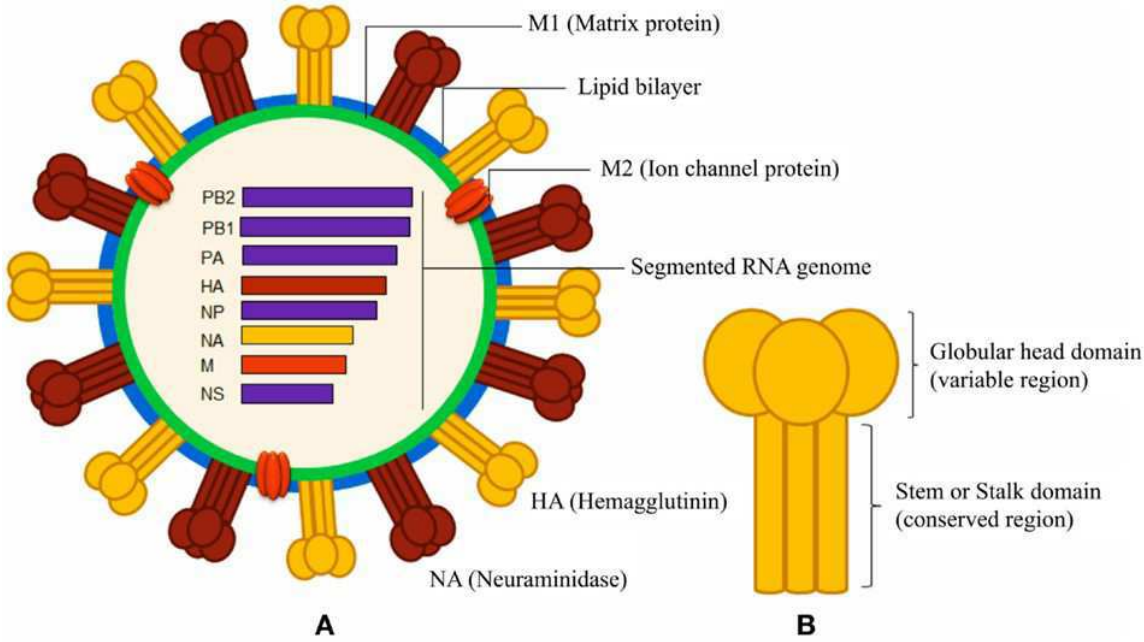


Fig.10. A: Overview of Influenza A virion. Genome segments and structural and non structural proteins are reported; B: Haemagglutinin. From: https://www.asm.org/index.php/mbiosphere/item/326-one-antibody-to-neutralize-them-all-a-human-igg1-is-effective-against-multiple-influenza-virus-subtypes

The genome is associated to polymerase complex proteins (PB1, PB2, PA) and enveloped in nucleoprotein (NP). The inner side of the viral envelope shows M1(matrix protein) and M2 (ion channel) proteins. Haemagglutinin (HA) and Neuraminidase (NA) form the surface proteins.

Structural proteins

- *M1 protein*: Also known as matrix protein (Fig.11.), it is the most abundant protein in viral particles (Skehel and Schild, 1971). M1 mediates the encapsidation of RNA-nucleoprotein (NP) cores into the membrane envelope through electrostatic interactions (Sha and Luo, 1997) and during infection these NP cores are transported by M1 into or out of the nucleus: M1 dissociates from NP cores in the endosomes (pH 4-5), allowing them to enter the nucleus. The exit of NP out of the nucleus is also mediated by M1 (Martin and Helenius, 1991). The interaction of M1 with NP is able to inhibit viral transcription and replication (Wakefield and Browniee, 1989; Winter and Fields, 1981). Moreover, M1 protein is responsible for the structural integrity of the viral particle via hydrophobic interactions (Fujiyoshi *et al*., 1994; Gregoriades and Frangione, 1981). Because of its interaction with HA, NA and M2 proteins, M1 protein is important in virus budding from the host cell; finally, only M1 is sufficient for vesicle formation, thanks to its viral self-assembly property.

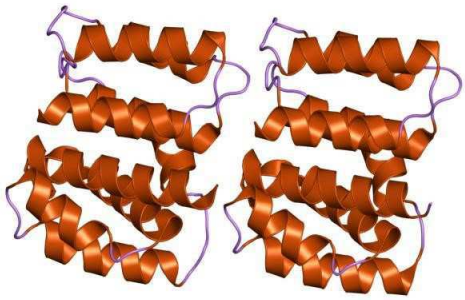


Fig.11. Influenza A matrix protein M1. It consists of two domains connected by a linker sequence. N-ter domain has a multi helical structure divided into two subdomains. Also the C-terminal contains alpha helices.

- *M2 protein*: This is an ion (proton) channel (Fig.12.), responsible for the passage through low-pH compartments during viral entry and maturation. M2 protein is located on the inner side of viral envelope (Schnell and Chou, 2008; Stouffer *et al*., 2008) Moreover, M2 can play a role in virus budding interacting with M1 in virus morphology determination.

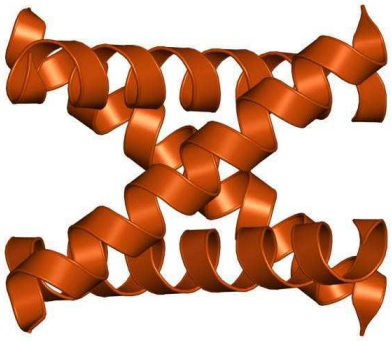


Fig.12. Influenza A protein M2. Here the ion channel is depicted in a closed conformation.

- *NP protein*: Is made up by a complex of genomic RNA segments associated with trimeric RNA polymerase (PB1, PB2, PA) and stoichiometric quantities of NP (Portela and Digard, 2002) (Fig.13.). After fusion, NPs are released into the cytoplasm and carried to the nucleus (Herz *et al*., 1981). Once viral mRNA transcription is ended, newly synthetized virions emerged from the apical surface of plasma membrane showing NPs previously exported from the nucleus and then incorporated into budding virions. NP proteins are coded by RNA segment 5 and the resulting polypeptide is 498 amino acids in length, rich in Arginine, Glycine and Serine residues. The net charge is positive at neutral pH. NP is able to interact with protein of viral and cellular origin and can self-associate to form large oligomeric proteins (Portela and Digard, 2002). The viral polymerase is characterized by having a high mutation rate, ensuring a rapid evolution of the virus and leading seasonal pandemics and epidemics. Thanks to Pflug's et al work published in 2004 on Nature, we can now speculate on the crystal structure of the bat influenza virus polymerase, able to replicate efficiently in human cells. So, this structure seems to be a good model for all Flu A polymerases.

  PB1 (polymerase basic 1) shows a similarity with that of HCV; it shows a central region (21-669 positions) carrying a RdRp fold. Residues 641-657 form a conserved anti-parallel β-loop involved in anti-genome replication by the polymerase. PB1 shows N- and C-terminal extension making inter subunit contacts with PA and PB2. Moreover, a flexible hinged β-ribbon (strands β6-β7) is characterized by PB1-NLS motifs. The internal cavity of PB1 contains the catalytic centre responsible for template-directed nucleotide addiction.

  PB2 (polymerase basic 2) contains N-ter and C-ter domains, each formed by several folded subdomains: PB2-N carries linked modules wrapping part of PB1 and

interacting with PB2-C; PB2-C forms an arc-shaped unit composed by two interacting sub-domains: the PB2-mid-domain and the cap-627 linker.

PA (polymerase acidic) comprises two domains: PA-Nter (endonuclease) and PA-C-ter, linked via the PA-linker wrapping around the external face of the PB1 fingers and palm domain. PA-N-ter is anchored to the rest of the polymerase through contacts with the same helical region of PB1-Cter that interacts with PB2-Nter, so that all three subunits are involved in positioning the endonuclease.

PB1-F2 is a polypeptide discovered in 2001 (Chen *et al*., 2001), acting as a regulator of the Influenza A viral polymerase activity (Mazur *et al*., 2008; Ueda *et al*., 2014). Its activity depends on viral strain and host cell type: can show pro-apoptotic function (Yamada *et al*.,2004; Zamarin *et al*.,2005; Mitzner *et al*., 2009) and it is able to modulate the immune response also by inhibition of type I interferon (Dudek *et al*., 2011). *In vitro* findings reveal an enhancing activity of the viral polymerase by PB1-F2 protein (Mazur *et al*., 2008; Ueda *et al*., 2014).

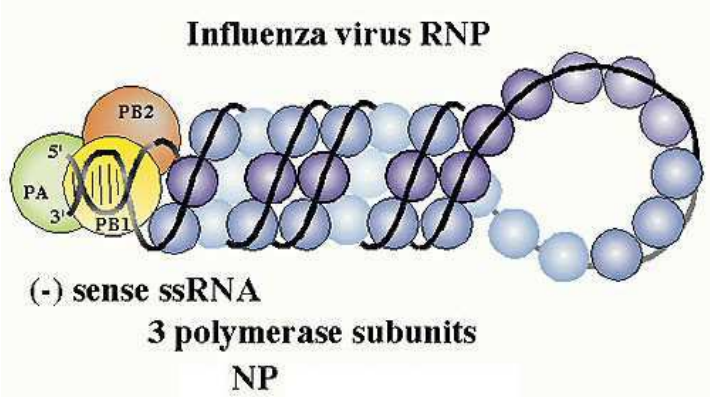


Fig.13. Influenza virus NP protein. The RNP (RNA + nucleoprotein, N) is in a helical form with the 3 polymerase polypeptides associated with each segment.

- *Neuraminidase (NA)*: Is a tetrameric surface glycoprotein (Fig.14.), bound to the viral membrane via its hydrophobic tail end. Neuraminidase contains antigenic and enzymatically active sites. This protein is characterized by a sialidase activity, that is it is able to cause hydrolysis of sialic acid residues present on the glycoprotein receptors on red cells. This way, progeny virions from infected hos cells can be released. Moreover, Neuraminidase is responsible for the degradation of the mucus layer of the respiratory tract, exposing the epithelial membrane for viral infection (Subhash 2012).

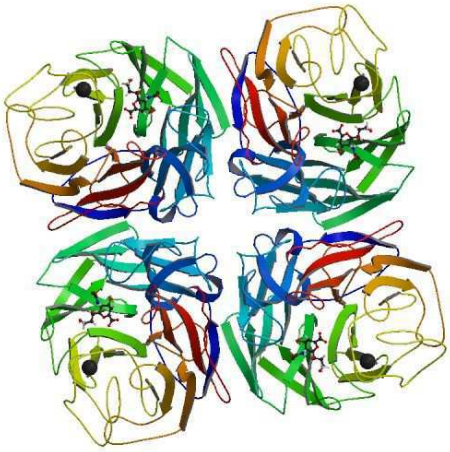


Fig.14. Influenza virus neuraminidase. The protein is coloured accordingly to rainbow color code: blue (N-ter) to red (C-ter).

- *Haemagglutinin (HA)*: is a homo-trimeric surface glycoprotein, responsible for viral attachment to host cells and subsequently viral internalization and endosomal membrane fusion functions (Skehel and Wiley, 2000). It has a cylindrical shape with dimensions of 135Å (length) x 35-70Å (radius) (Isin *et al*., 2002). Haemagglutinin is synthesized as a trimeric precursor called H0, in order to prevent premature fusion and Haemagglutinin activation throughout the secretory pathway (Hamilton et al., 2012). This precursor is then activated via membrane fusion by a post-translational cleavage thanks to intracellular trypsin-like proteases. These enzymes are able to recognize a basic cleavage site, located in a loop near a cavity in H0 (Klenk *et al*., 1975). The cleavage site differs among the Haemagglutinin subtypes: HPAI viruses (Highly Pathogenic Avian Influenza) carry a polybasic sequence whereas LPAI viruses (Low Pathogenic Avian Influenza) a monobasic one (Arg) (Hamilton *et al*., 2012; Sriwilaijaroen and Suzuki, 2012). After this process Haemagglutinin undergoes a structural rearrangement, in which the fusion peptide is allocated in the trimer. This way, ionizable residues involved in conformational changes in the endosome are buried. This event leads to two subunits (HA1 and HA2 of 36 and 27 KDa) linked by a disulfide bond. This cleavage is the start point of the Haemagglutinin fusion and the virus infectivity. Moreover, as seen before, it is a determinant of pathogenicity (Rott *et al*., 1987; Webster *et al*., 1987; Garten et al., 2008). From a structural point of view, Haemagglutinin consists of the two aforementioned subunits: HA1 and HA2. HA1 contains the receptor binding domain (RBD). This globular head domain is folded into a jelly-roll motif of eight stranded antiparallel β-sheets and into a shallow pocket (sialic acid binding pocket) surrounded by antigenic sites (130-loop, 190-helix, 220-loop).

These sites contain residues interacting with sialic acid located on the surface of epithelial cells (Weis *et al*., 1988; Martin *et al*., 1988). Moreover, HA1 also contains the vestigial esterase domain (Sriwilaijaroen and Suzuki, 2012). HA2, also called stem, shows a helical coiled-coil structure. The fusion peptide is located in this subunit. HA2 is able to anchor the membrane via a 10-residue cytosolic tail. This subunit is characterized by fusion activity.

Antigenic sites mapped onto H3 subtype (A, B, C, D, E) (Wiley *et al*., 1981) were used for mapping antigenic sites onto H1 and H2. H1 sites are named as Sa, Sb, Ca1, Ca2, Cb. H3 numbering were used to map antigenic sites onto H5: residues 140-145 (Site 1) correspond to antigenic site A of H3 and Ca2 of H1, residues 156-157 (Site 2) correspond to site B of H3, residues 129-133 (Site 3) correspond to Sa in H1 subtype (Peng *et al*., 2014).

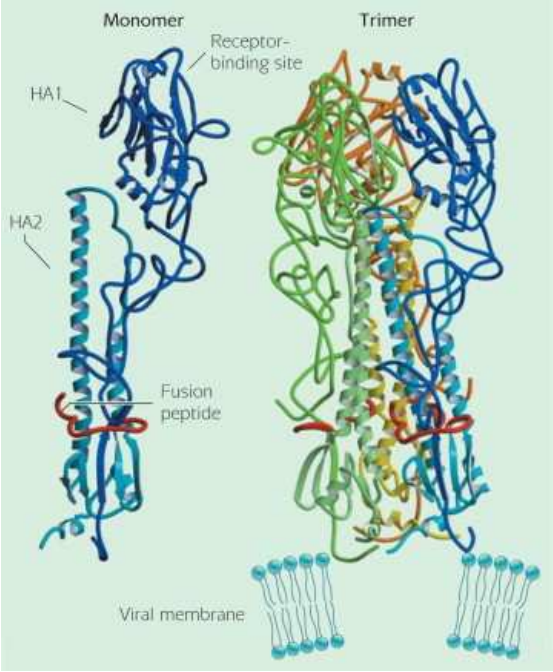


Fig.15. Influenza virus haemagglutinin. Monomer and trimer are depicted. From: https://www.rapidreferenceinfluenza.com/chapter/B978-0-7234-3433-7.50009-8/aim/introduction

Non-structural proteins

- *NS1*: This protein of 26 kDa (Fig.16.) is able to sequester the RNAs formed during the viral life cycle, preventing the recognition of these RNA elements by cellular RNA helicases, PKR (double-stranded RNA-activated kinase), TL3 and Dicer-mediated RNA-silencing pathways. This way, the antiviral immune response is overcome. The activity of NS1 leads to the abrogation of cell apoptosis and enables the virus to complete its life cycle and spread (DeFranco, Locksley, Robertson Immunity: The immune response in infectious and inflammatory disease).

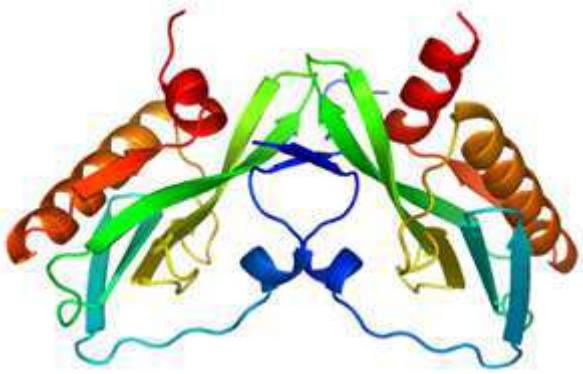


Fig.16. Influenza virus NS1 protein. This protein shows a dimeric form. The representation is coloured in rainbow manner from blue (N-ter) to red (C-ter). From: https://www1.aps.anl.gov/APS-Science-Highlight/2006/A-Bird-Flu-Protein-Link-to-Virulence

- *NS2*: This protein of 11 kDa is also named as NEP (Nuclear Export Protein); it is able to mediate the nuclear export of v-RNA by acting as an adaptor between RNP complexes and the nuclear machinery of the cell. NS2 contains a nuclear export signal interacting with cellular nucleoporins (O'Neill *et al*., 1998). NS1 and NS2 are encoded by the same genetic segment via alternative splicing (Sriwilaijaroen and Suzuki, 2012).

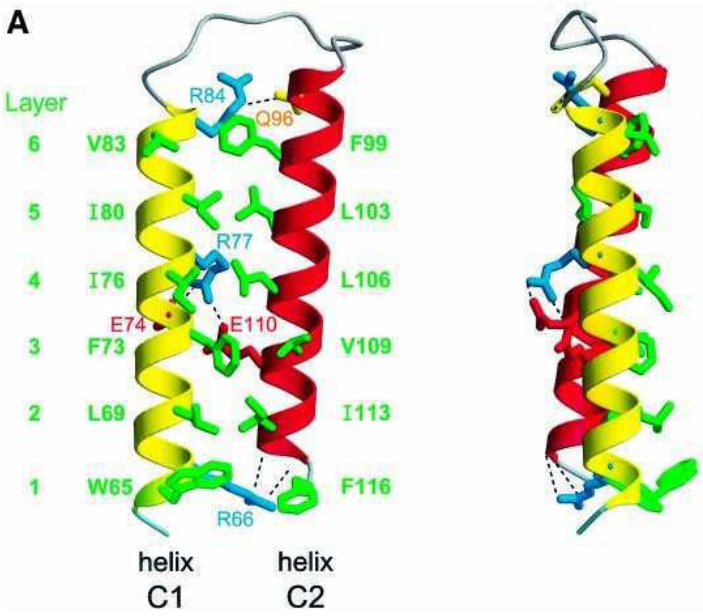


Fig.17. Influenza virus NS2 protein. Hydrophobic residues involved in interhelical contacts are highlited. From: Akarsu H, Burmeister WP, Petosa C, Petit I, Müller CW, Ruigrok RW, Baudin F. Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). *EMBO J*. 2003 Sep 15;22(18):4646-55.

1.2 Influenza virus life cycle

The first step in influenza virus infection (Fig.19.) starts with haemagglutinin binding to sialic receptors on the surface of the host cell. It is shown that human-adapted haemagglutinins preferentially bind to the α-2,6-sialic acid linkage, whereas the avian-adapted haemagglutinins bind the α-2,3-sialic acid linkage (Garcia-Sastre, 2010).

After the attachment, the virus is internalized into an endosome via chlatrin-dependent or independent endocytosis (Lakadamyali *et al*., 2004; Lakadamyali *et al*., 2006; Sieczkarski and Whittaker, 2002; Chen and Zhuang, 2008) or via macropinocytosis (De Vries *et al*., 2011; De Conto *et al*., 2011).

The low pH into the endosome is able to trigger dramatic conformational changes in the haemagglutinin (Fig.18. left panel): HA exposes the fusion peptide and the HA1 trimer becomes divided into its monomers and separated from the HA2 domain (the stem) except for the disulfide bridge between residues 14 and 137 (H3 numbering). On the other side HA2 shows a coiled-coil extension due to a loop-to-helix transition at positions 55-76 and a helix-to-loop transition at positions 106-112. This extended intermediate is responsible for the viral-endosomal fusion, starting with the exposure of the hydrophobic fusion peptide, located in a pocket near the viral membrane, to the endosomal membrane. Once the endosomal membrane approaches the viral one, the fusion peptide attaches the opposite

membrane thanks to strong hydrophobic interactions with the lipid acyl chains. After this process, several haemagglutinins seem to form a fusogenic unit: the stalks collapse by zipping up N- and C- terminal membrane anchors together leading the formation of a fusion pore (Sriwilaijaroen and Suzuki, 2012; Hamilton *et al*., 2012) (Fig.18. right panel).

After membrane fusion, the viral genome and associated proteins are released into the host cytosol, the viral RNPs are transported to the nucleus (via the nuclear pore complex) where replication takes place. This process is error prone, showing error frequencies similar to those of DNA transcription (1 error every 104nt synthesized). Positive sense mRNAs are produced during transcription (RNPs are used as template by v-RNA polymerase complex). These mRNAs are then exported and translated by the host cell machinery in the cytoplasm in order to produce vral proteins. The mRNA nuclear export is under NS1 control.

Viral proteins (PB1, PB2, PA, PB1-F2, NP, M1, NS1) are imported into the nucleus. Synthesis of integral membrane proteins (HA, NA and M2) takes place on the RER and the maturation in the Golgi complex (Sriwilaijaroen and Suzuki, 2012).

Mature HA and NA, M2, NS2-M1-vRNPs and other M1 and NS2 are assembled into a virion at the apical cell membrane. The forming bud contains coating proteins immersed in the host cell lipid bilayer.

Finally, the budding step occurs: here, viral particles are released from host cells. Eventually host proteins present in the plasma membrane are excluded.

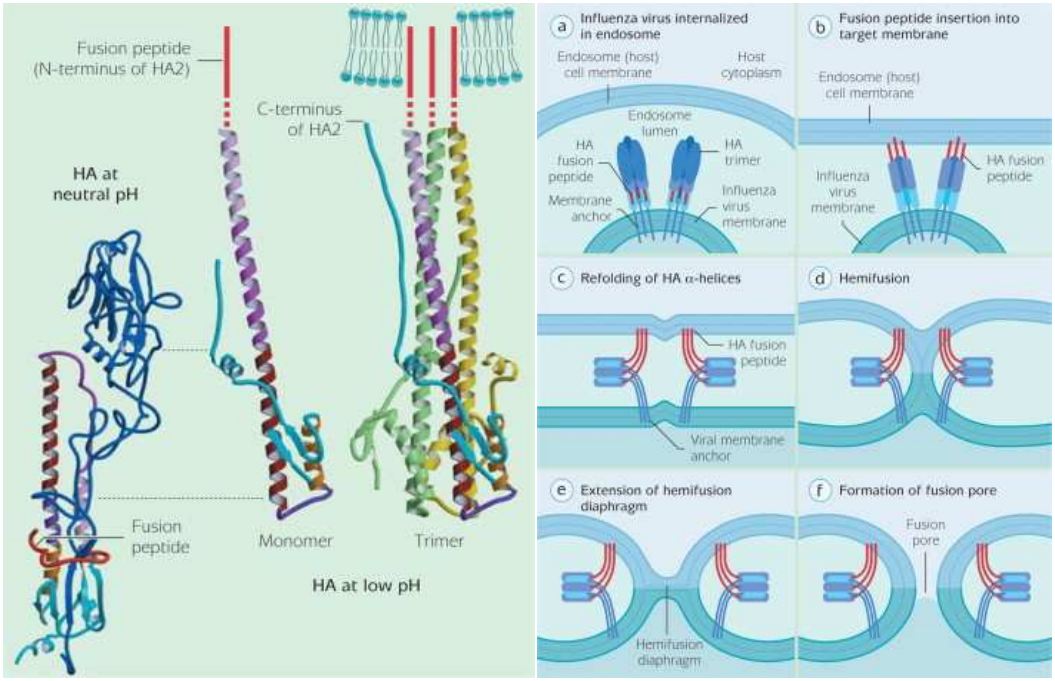


Fig.18. HA conformational changes during fusion (left panel) and the fusion process (right panel). From: https://www.rapidreferenceinfluenza.com/chapter/B978-0-7234-3433-7.50009-8/aim/introduction

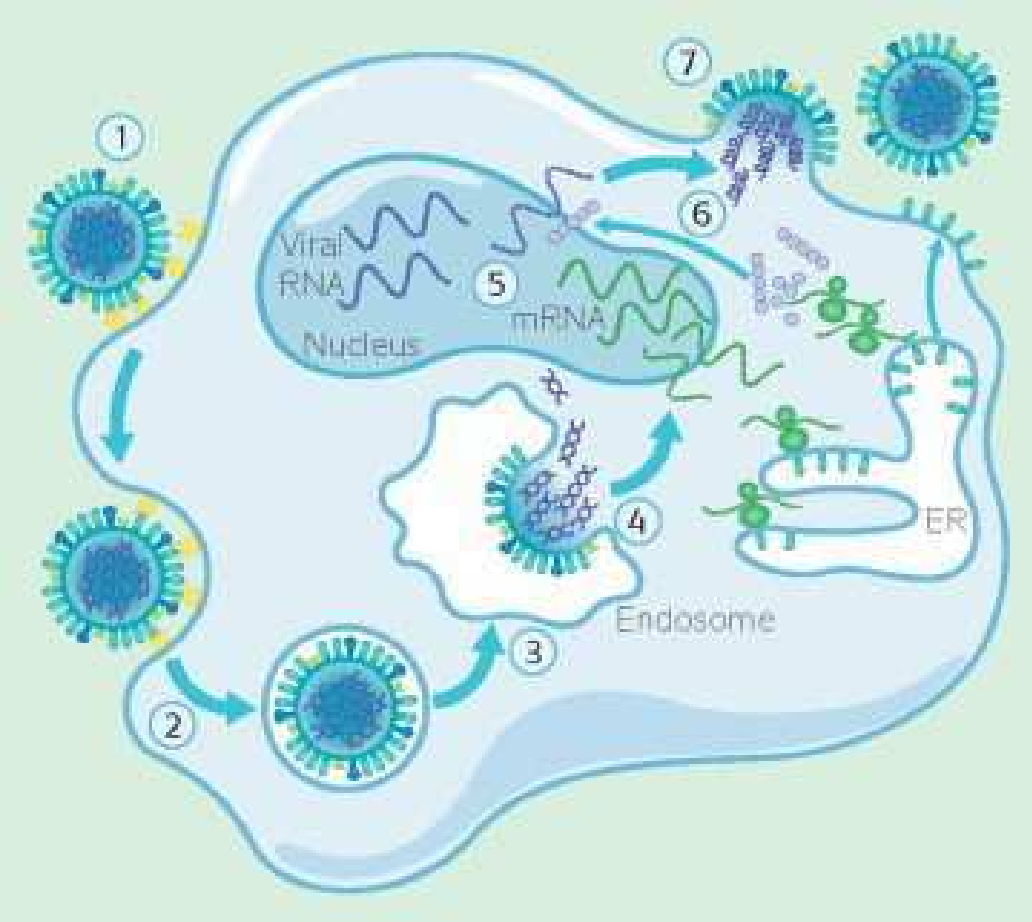


Fig.19. Influenza virus life cycle. 1) Virus attachment to a sialic acid receptor via spikes on the viral envelope; 2) Virus engulfment and formation of an endocytic vesicle; 3) Delivery of the virus in the endosomal cell compartment; 4) Fusion of the viral membrane with the endosomal membrane: here, the low pH induces conformational changes in HA able to drive the fusion process; 5) After fusion vRNPs are released into the cytoplasm and enter the host nucleus, where viral transcription and replication occur. Transcription produces positive sense mRNAs in order to build viral proteins by the host cell machinery in the cytoplasm. Viral proteins PB1, PB2, PA and NP making up vRNP, NS2 involved in the control of viral transcription and regulation of vRNP export, and M1 related to nuclear import and export of vRNP are imported into the nucleus. NS1 protein is able to inhibit host immune response and PB1-F2 enhances apoptosis; 6) HA, NA and M2 proteins are synthesized on the RER and their maturation occurs in the Golgi complex; 7) The mature glycosylated HA and NA, non-glycosylated M2, NS2-M1-vRNPs, and other M1 and NS2 are transported to the apical cell membrane, where they are assembled into a progeny virion that finally buds from the host cell surface. From: https://www.rapidreferenceinfluenza.com/chapter/B978-0-7234-3433-7.50009-8/aim/introduction

1.3 Influenza virus evolution

Influenza viruses undergo rapid evolution and adaptation, allowing viruses to escape immune surveillance, causing annual epidemics and periodic pandemics. These mechanisms arise due to the low fidelity RNA polymerase, rapid replication and infection of large population size. Antigenic drift and shift are the two events able to explain the viral evolution. Antigenic variation is a common feature of influenza A and B viruses and concerns the two major glycoproteins on the viral surface: haemagglutinin and neuraminidase.

Antigenic drift

Antigenic drift occurs about every 2-8 years and this is the gradual evolution of viral strains due to frequent mutations. We may think at antigenic drift to as the answer to the selection pressure to escape immunity. During antigenic drift HA and NA carry point mutations at their antibody binding sites. For example, as compared with the previously circulating H3 viruses represented by A/Panama/2007/99, the A/Fujian/411/2002 virus has 13 amino acid changes in different antigenic sites (Treanor, 2004). Usually these mutations are not responsible for conformational changes but some of them are able to inhibit host antibody binding. As a consequence, infecting viruses can spread more rapidly among the population. Antigenic drift can be different between strains: for example, H1 (Influenza A) and B are characterized by co-circulating drift variants with multiple co-existing lineages (the re-emergence of old strains can occur) whereas H3 (Influenza A) is more subjected to antigenic drift and new variants are prone to replace the old ones (Carrat and Flahault, 2007). Influenza epidemics are related to antigenic drift. Determination of new variants HA nucleotide sequence and evaluation of the old virus antiserum inhibition of the new virus are useful to evaluate the extent of antigenic drift. The event of antigenic drift is important in vaccine development. HA1 region of haemagglutinins is sequenced and tested with serum from infected ferrets. In case of identification of a new variant, this can be used as an influenza vaccine (Treanor, 2004).

Antigenic shift

Antigenic shift refers only to Influenza A and occurs approximately three times every 100 years. It results from the replacement of HA, (and sometimes of NA), subtypes with novel ones. The consequence is the creation of new viruses never seen before. This event can lead to pandemics or worldwide epidemics characterized with million deaths. Genetic reassortment (mixing of genetic material between different viral strains), occurring due to the co-circulation of different Influenza A subtypes (even from different species), can be considered as a reason for antigenic shift. The virus emerged from antigenic shift can undergo to antigenic drift as demonstrated by the fact that all current circulating influenza viruses are drift variants of previously pandemic influenza strains. Focusing on HPAI A/H5N1, the avian influenza strain undergoes antigenic drift making it possible the human-to-human transmissibility, resulting in a major worldwide human pandemic. (Carrat and Flahault, 2007).

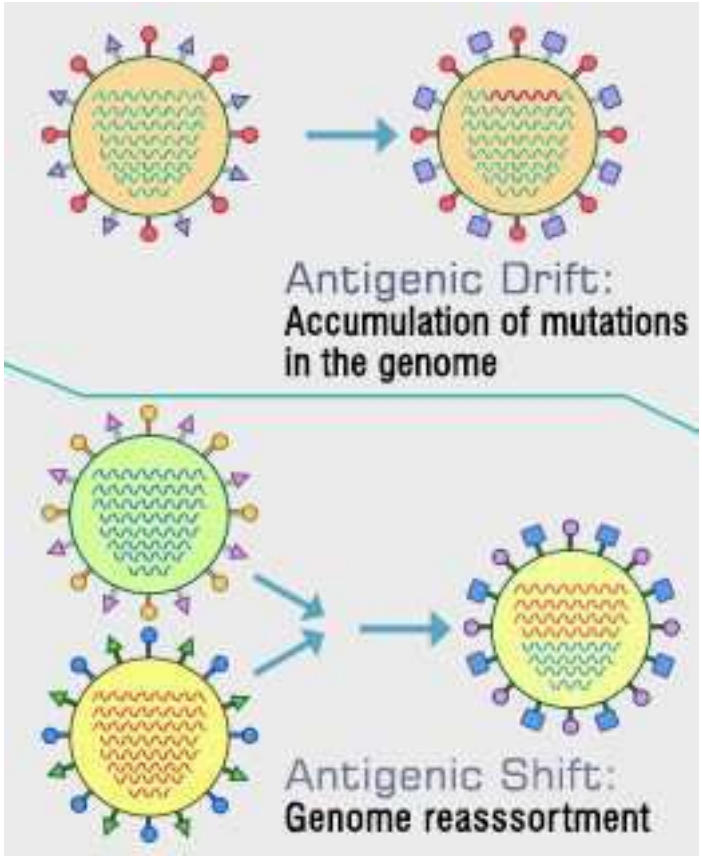


Fig.20. Antigenic drift and shift as mechanism for viral evolution. Drift produces point mutations in HA and NA at their Ab binding sites, shift produces new viruses by the replacement of HA and NA via genome reassortment. From: http://www.buzzle.com/articles/antigenic-drift-vs-antigenic-shift.html

1.4 Receptor binding specificity

Host cells exhibit two types of sialic acid (SA), both of them being recognized by Influenza viruses: N-acetylneuraminic acid (NeuAc) and N-glycolylneuraminic acid (NeuGc), attached to galactose via α-2,3 (SAα-2,3Gal) and α-2,6 (SAα-2,6Gal) linkages. SA types influence the ability of the virus to replicate in different host species: avian viruses show a propensity for SAα-2,3Gal linkage, whereas human viruses for SAα-2,6Gal one (Baigent and McCauley 2003). α-2,3 and α-2,6 are not the only two types of glycosidic linkages recognized by Influenza virus RBD, but also α-2,8 linkage present in many glycoproteins (such as N-CAMS) and some gangliosides (eg. GD3 Neu5Acα2-8Neu5Acα2-3Galβ1-ceramide) (Wu and Air, 2004; Childs et al., 2009). Sulfated glycans present in human mucins seem to inhibit the attachment of influenza viruses to target cells (Stevens *et al*., 2006). SAs correct orientation in the RBD is ensured by two specific positions in the RBD: 226 and 228. Gln 226, found in avian viruses, correlates with SAα-2,3Gal receptor specificity, whereas Leu226 with SAα-2,6Gal in human viral subtypes H2 and H3, but not H1. In H1 viruses (swine and human) positions 190 and 225 are required for the acquisition of SAα-2,6Gal specificity. H1 HA carrying E190 and G225 preferentially binds to α-2,3 receptors in birds, H1 HA showing D190 and G225 binds both SAα-2,3Gal and SAα-2,6Gal in pigs whereas H1 HA containing D190 and D225 binds to SAα-2,6Gal in humans (Sriwilaijaroen and Suzuki, 2012). Not only the presence of different SAs can affect the specificity of viral binding. The site and the temperature of replication can play a pivotal role. Avian viruses replicate in the intestinal tract, in contrast human-adapted viruses in the respiratory tract, even if SA is present in gut too. In their work, Kobase et al. in 2001 proved the inability of human viruses to replicate in duck intestine, despite their NAs were SAα-2,3Gal cleavage featured. By contrast, H5N1 virus, directly transmitted from poultry to humans without adaptation in other mammals, is able to replicate in the human intestine. Therefore, one can infer that a biological difference exists between human-adapted and avian-adapted viruses in replication in gut tissue. Temperature is fundamental in the activity of HA and NA. Avian NAs work better at higher temperature and lower pH than do mammalian ones. Moreover, human viruses replicate better a 37°C while avian strains at 40°C. A shift of the receptor binding specificity from avian to human seems to lead to human pandemic (Baigent and McCauley 2003).

1.5 Influenza virus pathogenicity

Pathogenicity is the ability of a virus to produce disease in a host, when compared with similar agents (Digard *et al*., 2005). Influenza virus pathogenicity depends on a combination ov viral and host determinants. Influenza viruses can be subdivided in two groups on the basis of their pathogenicity: HPAI (high pathogenicity avian influenza) viruses are responsible for severe diseases with high mortality (90-100% in chickens in 48 hours) whereas LPAI (low pathogenicity avian influenza) cause no disease (only mild illness). Both of HPAI and LPAI viruses are able to spread rapidly across the population (https://www.cdc.gov/flu/avianflu/). Different strains differ in pathogenicity and haemagglutinin seems to play a fundamental role in this picture. However, haemagglutinin is not the main actor but the combination of PB2 and NS1 in addition to HA is necessary in determining the degree of pathogenicity.

The pivotal role of haemagglutinin is linked directly with its cleavage site, cleaved by host cell proteases (Klenk *et al*., 1975; Klenk and Garten, 1994; Chen *et al*., 1998; Steinhauer, 1999). Cleavage site shows two forms:

- *Mono-basic cleavage sites*: these cleavage sites are cleaved by few cellular proteases and contain one basic amino acid in the critical position (eg. PEKQTR/GLF). Viruses carrying these cleavage sites can grow generally only in poultry intestinal and respiratory tracts.
- *Multi-basic cleavage sites*: these cleavage sites are cleaved by several common cellular proteases and contain several basic amino acids in the critical positions (eg. PQRESRRKK/GLF). Viruses carrying these cleavage sites can grow throughout the body of the host.

HPAI viruses carry a polybasic cleavage site cleaved by the ubiquitous subtilisin-like proteases furin and PC6 (Horimoto et al., 1994). Pathogenicity is also regulated by acid stability of the hemagglutinin. For example, even if HPAI H5N1 (A/chicken/Hong Kong/YU562) and LPAI H5N1 (A/goose/Hong Kong/437-10) are expressed and cleaved in similar amounts, and both RBDs are similar featured, these two HA are expressed at different pH: 5.7 for HPAI virus and 5.3 for LPAI virus. This behavior can be explained due to amino acid variations at positions 104 and 115 at N- and C- termini of the 110-helix in the vestigial esterase subdomain of RBD, interacting with the B-loop of the stem (HA2) (DuBois *et al*., 2011). Carbohydrate residues next to cleavage site are able to affect the enzymes capability of cleavage, due to steric hindrance of those molecules (Kawaoka *et al*., 1984). The HA of LPAI is less cleavable and less accessible to the activating enzymes. A low

pathogenic virus can adopt the strategy of recombination to become high pathogenic. This process, involving the insertion of genetic material from the host or other viral strains, is typical for H7 but not for H5. For example, A/chicken/Chile/4322/02 (H7N3) became HPAI via insertion of 30nt into the HA gene encoding the cleavage site (Digard *et al*., 2005).

The contribution of NS1 protein to pathogenesis lies in the inhibition of antiviral immune response, by blocking the interferon production (Garcia-Sastre, 2001). Reassortant virus carrying H5N1/97 NS gene was able to increase the production of inflammatory cytokines and chemokines in infected mouse lungs.

The role of PB2 in increasing virulence can be explained by the presence of the residue Lys at position 627. This residue seems to be a necessary condition for high virulence and systematic replication of H5N1 virus in mice. It was inferred that PB2 Lys_{627} is responsible for viral replication in mouse cells but not in avian ones, but this residue doesn't correlate with the viral tropism toward different organs in mouse (Shinya *et al*., 2004).

Results and discussion

Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features

Righetto I., Milani A., Cattoli G., Filippini F.

BMC Bioinformatics 2014 Dec; 15:363

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features

Irene Righetto[1], Adelaide Milani[2], Giovanni Cattoli[2] and Francesco Filippini[1*]

Abstract

Background: Genome variation is very high in influenza A viruses. However, viral evolution and spreading is strongly influenced by immunogenic features and capacity to bind host cells, depending in turn on the two major capsidic proteins. Therefore, such viruses are classified based on haemagglutinin and neuraminidase types, e.g. H5N1. Current analyses of viral evolution are based on serological and primary sequence comparison; however, comparative structural analysis of capsidic proteins can provide functional insights on surface regions possibly crucial to antigenicity and cell binding.

Results: We performed extensive structural comparison of influenza virus haemagglutinins and of their domains and subregions to investigate type- and/or domain-specific variation. We found that structural closeness and primary sequence similarity are not always tightly related; moreover, type-specific features could be inferred when comparing surface properties of haemagglutinin subregions, monomers and trimers, in terms of electrostatics and hydropathy. Focusing on H5N1, we found that variation at the receptor binding domain surface intriguingly relates to branching of still circulating clades from those ones that are no longer circulating.

Conclusions: Evidence from this work suggests that integrating phylogenetic and serological analyses by extensive structural comparison can help in understanding the 'functional evolution' of viral surface determinants. In particular, variation in electrostatic and hydropathy patches can provide molecular evolution markers: intriguing surface charge redistribution characterizing the haemagglutinin receptor binding domains from circulating H5N1 clades 2 and 7 might have contributed to antigenic escape hence to their evolutionary success and spreading.

Keywords: Haemagglutinin, Avian influenza virus, Viral evolution, H5N1, Antigenic drift, Receptor binding domain, Homology modeling, Isopotential contour, Hydropathy analysis

Background

Influenza caused by influenza A viruses occurs in both birds and mammals. In humans, influenza A viruses infect hundreds of millions individuals, causing a high number of deaths per year. Indeed, influenza A outbreaks occurred in 1918, 1957 and 1968 resulted in death for ~100 million people worldwide [1]. However, seasonal epidemic outbreaks cause estimated 250.000 to 500.000 yearly deaths worldwide [2] (data from the World Health Organization (WHO) [3] and from the Center for Disease Control and prevention [4]). The largest reservoir of all subtypes of influenza A is found in wild water avian species and some viruses can infect different hosts [5,6]. Classification of influenza type A virus subtypes is based on antigenic and genetic differences in the two surface spike proteins: haemagglutinin (HA) and neuraminidase. For instance, H5N1 viruses combine the haemagglutinin of the H5 subtype with neuraminidase of the N1 subtype. A wide interest for haemagglutinin depends on evidence that this protein (i) is crucial to the attachment and penetration into the host cell, (ii) represents the main viral surface antigen, and (iii) is a major player in the stimulation of the neutralizing antibody response [7]. Haemagglutinin is synthesized as a precursor and then processed by cellular proteases to yield mature polypeptide subregions. In order to provide unambiguos information, hereafter acronyms

* Correspondence: francesco.filippini@unipd.it
[1]Molecular Biology and Bioinformatics Unit (MOLBINFO), Department of Biology, University of Padua, via U. Bassi 58/B, 35131 Padova, Italy
Full list of author information is available at the end of the article

for haemagglutinin are the followings: 'HA' for haemagglutinin in general; HA0 for the precursor; HA1 and HA2 for the two subregions and 'H' followed by progressive numbering (H1 to H16) for each haemagglutinin subtype. Influenza virus haemagglutinin is a type I transmembrane glycoprotein that is exposed at the viral surface as a homotrimer. Trimerization is possible once proteolytic cleavage of the unfolded HA0 precursor occurs hence allows for folding of monomers, each consisting of two mature chains: HA1 and HA2 [7]. Structurally, each monomer consists of a globular 'head' (part of chain HA1) and of a 'stem' region (contributed by both chains HA1 and HA2). The head includes a receptor-binding domain (RBD) and a vestigial esterase domain (VED), whereas the stem is structured as a mainly α helical, coiled coil region. Functionally, the RBD mediates docking to the host cell by binding sialic acids as cell entry receptors, whereas the stem domain mediates both tethering and membrane fusion once conformational change is occurred, caused by pH decreasing along the endosomal route. For several years, classification of HA from influenza viruses was mainly based upon serological and/or phylogenetic analysis [8]. However, structural genomics projects are providing the scientific community with an increasing number of structural templates, while contemporary reverse genetics, immunogenomics investigations and improved sequencing technologies are producing a high number of mutant sequences. Changes in serological specificity depend on variation of epitopes recognized by the specific antibody rather than on the extent of sequence divergence, meaning i.e. that (i) two proteins with highly similar sequences may show quite different properties when considering recognition of specific epitopes and (ii) two proteins may share antigenic properties even when having highly divergent sequences, if epitopes involved in the specific recognition were conserved. Variation of some protein properties sometimes may depend only on 'local and limited changes', e.g. mutation of a few - or even only one – residue(s) within linear or conformational motifs. In fact, even when local variation in sequence is seemingly poorly evident, it may result in 'locally dramatic' changes in accessible surface area, electrostatic potential, hydropathy or hydrophilicity features that can deeply change motif functionality. It is common knowledge that variation in surface features of a protein can modulate 'recognition' interactions of the protein itself. Since variation often depends on mutation of a number of residues and changes in side chains can vary multiple biochemical features, it is difficult or even nonsense trying to establish *a priori* which specific property (among e.g. surface area and shape, electrostatics or hydrophobicity) should be more relevant than others in modulating recognition interactions. In fact, changes in each specific property can result in such modulation, and this can be independent on variation of other features, or modulation can result from the aggregate or synergistic effect of multiple feature changes. So far, several sequence-based studies on variation could provide valuable phylogenetic evidence; however, such studies are of minor help in inferring variation at protein regions including amino acids that are far each other in the primary sequence and quite close within the 3D protein structure (conformational epitopes). In practice, while sequence-based investigation can be good in highlighting very evident changes at individual positions of a protein chain, in general they fail in highlighting meaningful 'group variation', i.e. in identifying - especially when the overall variation is relevant and spread - relationship of specific multiple changes to variation in conformational epitopes hence in interactions they mediate.

Once solved structures are available, presence of one or more structural templates allows for shifting to 'conformational epitope based' studies on variation and, in particular, to investigating on surface region variation. Stressing relevance of local surface variation is particularly important when considering special constraints addressing viruses evolution: keeping basic properties in simplified but complex pathogenic systems while simultaneously varying - as much as possible - all variable epitopes, in order to escape the immune responses of their hosts. Therefore, viral genome evolution runs along two parallel tracks, both of which, like in railways, must be followed: (i) mutations in sites crucial to protein machinery mediating basic functions (e.g. in motifs relevant to host recognition or cell entrance) are not allowed because they strongly impair viral fitness, and at the same time, (ii) hyper-variability is needed to escape recognition by neutralizing antibodies ('antigenic drift', [7]). Given that surface viral proteins do not interact only with antibodies (as their original function is to contact the host), in addition to determining antigenic drift, variation can also influence pathogenicity (because e.g. of modified interaction with cell receptors in different tissues and organ districts) or host specificity. Influenza viruses do not escape such a two-tracks rule, hence while global structure conservation ensures basic functions, limited or even subtle changes in local structural features may modulate interactions of the viral proteins with the host molecules/cells and thus mechanisms underlying antigenic drift, pathogenicity shifts and host specificity change. Phylogenetically and serologically, haemagglutinins are divided into either two supergroups or four groups: Group 1 (H1, 2, 5, 6, 11, 13 and 16); Group 2 (H8, 9 and 12); Group 3 (H3, 4 and 14) and Group 4 (H7, 10 and 15). The two supergroups consist of Groups 1 + 2 and 3 + 4, respectively [9,10]. Thanks to the availability of thousands of viral genomes/gene sequences and of several specific antibodies/vaccines, a large number of sequence-

based/phylogenetic and serological analyses of avian flu viruses have been performed and published so far. This notwithstanding, mechanisms in viral evolution are still elusive, as genome/proteome-wide analyses on sequence variation or antigenic features are able to only partially unveil a number of relevant changes, because of the overall mutational noise. Therefore, structural 'zoom in' is needed to integrate such analyses by identifying 'meaningful' variation. This prompted us to take advantage from availability of structural templates to perform structural comparison among different HA subtypes, in order to identify subtype- and subregion-specific feature variation suggestive for possible involvement in antigenic recognition, or pathogenicity and host specificity. Last but not least, evidence from structural comparison can check relationship among serological, phylogenetic and structural closeness.

We started our analyses using six currently available solved HA structures; then, in order to investigate structural variation possibly underlying H5N1 clades evolution and spreading, we also created clade models by homology modeling. The six HA structures solved so far: H1 [11], H2 [12], H3 [13], H5 [14], H7 [9], H9 [15], all concern mature proteins, consisting of the two HA1 and HA2 parts of haemagglutinin. Solved structure of H16 [16] was not considered for this analysis because it corresponds to the HA0 precursor. Comparative analysis of structural features unveiled that some discrepancy may occur with respect to a generally observed agreement between sequence and structural closeness, because of subregion local variation. Structural analysis was performed by comparison of secondary structure topology and surface analysis, in terms of both electrostatic and hydropathy analysis.

Results and discussion

Comparison among solved HA structures

Prior to creating models, preliminary analysis of the six available HA structures was performed in order to evaluate intra- and inter-group structural variation by superposition of all structure pairs and computation of their Root Mean Square Deviation (RMSD). Indeed, the RMSD of two superposed structures indicates their 'structural divergence' from one another. As both sequence mutation and conformational variation inflate the RMSD, values up to 2 Ångstrom indicate structural similarity [17]. Structural superposition of each possible combination of two different HA molecules (hereafter referred to as 'pairs') and RMSD computing were performed using Chimera 1.8.1 software [18]. Pair-wise method was chosen to calculate RMSD because all superpositions only compared pairs in order to properly relate a structural closeness index for a pair to identity/similarity values (commonly reported as an index to state closeness) from the corresponding aligned sequences. Fold comparison method based on sequence fragmentation and order-independent resorting was not considered because order-dependent global alignment is an established standard for comparing highly similar sequences in structural biology and the alignment of sequence blocks for phylogenetic analyses is also order-dependent.

In addition to superposing structures of HA monomers, also corresponding structures of their Receptor Binding domains (RBDs) were superposed. Results are summarized in Table 1. Evidence that RMSD values for monomer pairs are lower than those ones for corresponding HA1 or RBD regions is not surprising, because RBDs are major determinants in antigenic variation [9]. Moreover, HA2 'stem' region of the monomer is structurally less variable than HA1 [19], hence its contribution results in decreasing the overall monomer RMSD value. RMSD values for HA1 pairs are higher than corresponding RBDs because of unstructured regions connecting RBDs to stems. Group 1 is - at least to date - the only HA group in which multiple structures (in particular, H1, H2 and H5) are solved. Structural comparison within this group highlights some intriguing evidence. When comparing monomers amino acid sequences, H5 results to be closer to H2 than to H1, independently on identity (roughly 73% vs. 63%) or similarity (approximately 86% vs. 81%) is considered. Such relationship is confirmed for both HA1 and RBD sequences, as shown by identity and similarity values in Table 1. However, when comparing structures, H5 is closer to H1 than H2, as in all comparisons, H5:H1 superposition RMSD values are lower than H5:H2 ones. Commonly, % identity is taken into account as an index for relationship among proteins [20]. However, from a structural point of view, 'type' of mutations occurred - rather than the overall sequence divergence - is very important: a few mutations (or even a single one) to some specific residues in 'critical' regions can result in dramatic structural changes. Structural fold and architecture can be highly conserved even among proteins and protein domains showing no sequence homology because of either long evolutionary divergence or even convergent evolution [21]. At the same time, within such families, fold can be disrupted (resulting in loss of function and disease) by single or few specific mutation(s), which indeed result in keeping 99% or higher sequence identity values [22,23]. In the structural comparison of H5 to haemagglutinins from different groups (represented by H9, H3 and H7) further interesting points emerge. In the monomer comparison, % identity approximately ranges from 41 to 49%. The same 8% difference in % identity is retrieved in % similarity (ranging from 64 to 72%). However, RMSD for corresponding monomer pairs keep quite similar values, i.e. they are not impaired by lower %

Table 1 Structural and sequence closeness among pairs of haemagglutinin proteins with solved structures

	RBD				
	H2	**H5**	**H9**	**H3**	**H7**
H1	r:1.343	r:0.918	r:1.249	r:2.292	r:2.784
	i:55.4 s:78.4	i:52.0 s:78.3	i:45.7 s:69.7	i:38.0 s:61.1	i:37.2 s:63.7
H2		r:1.130	r:1.636	r:2.083	r:1.772
		i:65.6 s:83.7	i:41.4 s:66.8	i:36.8 s:57.3	i:33.5 s:60.7
H5			r:1.498	r:2.241	r:3.085
			i:41.4 s:66.4	i:37.3 s:61.4	i:38.4 s:67.4
H9				r:1.983	r:2.069
				i:36.9 s:60.4	i:33.9 s:58.4
H3					r:1.429
					i:35.0 s:63.6
	HA1				
	H2	**H5**	**H9**	**H3**	**H7**
H1	r:1.476	r:1.065	r:1.563	r:2.548	r:2.941
	i:56.7 s:78.7	i:56.6 s:79.2	i:46.4 s:69.4	i:37.1 s:62.9	i:36.1 s:63.3
H2		r:1.527	r:2.087	r:3.253	r:3.025
		i:67.7 s:83.3	i:43.5 s:65.3	i:35.3 s:58.3	i:34.5 s:60.6
H5			r:1.680	r:3.043	r:2.755
			i:43.5 s:67.0	i:37.2 s:61.9	i:36.9 s:66.7
H9				r:2.320	r:3.672
				i:35.8 s:60.9	i:33.5 s:59.8
H3					r:1.631
					i:37.8 s:64.0
	Monomer				
	H2	**H5**	**H9**	**H3**	**H7**
H1	r:1.180	r:0.98	r:1.350	r:1.710	r:1.780
	i:64.2 s:82.9	i:62.8 s:81.5	i:50.4 s:71.3	i:40.0 s:61.6	i:42.4 s:67.1
H2		r:1.100	r:1.450	r:1.760	r:1.730
		i:73.0 s:85.7	i:49.0 s:69.6	i:37.6 s:59.6	i:40.6 s:66.5
H5			r:1.686	r:1.680	r:1.620
			i:48.7 s:72.0	i:40.2 s:63.9	i:42.3 s:69.9
H9				r:1.760	r:1.850
				i:37.9 s:61.7	i:40.8 s:66.1
H3					r:1.250
					i:44.0 s:66.2

Within each cell, the upper value is RMSD (r) for the superposed pair and lower values (in %) are identity (i) and similarity (s) for corresponding, aligned amino acid sequences.

identity or similarity values. This is not surprising, because - as shown by aforementioned example (and by many others in literature) - very ancient divergence or convergence can result in fold conservation among proteins without significant sequence similarity. Structural differences become clearly evident when comparison focuses on HA1 and RBD regions: H5 is quite closer to H9 than H3 and H7 (roughly doubled RMSD) and in this instance substantial agreement between structural and sequence divergence is found. Once again, a rationale for this is found when considering common properties of protein domains. Different subregions of the same protein are involved in different interactions and pathways. Therefore, molecular evolution can locally change subregion structures to modulate specific interactions and pathways, without affecting those ones mediated from other subregions of the same protein. In practice, only when structural variation analysis is performed at both overall and local level (i.e. focusing on individual domains and/or domain motifs), it is possible to boost subsequent experimental work. In fact, subregion analysis allows for shedding light on specific molecular properties that are likely to underlie different functions of the protein. In conclusion, agreement between sequence homology and structural closeness which is generally observed [20] has not to be strictly interpreted as 'a rule' to be followed. Values from Table 1 show that, in most instances, such an agreement is found. However, in several examples and depending on local variation, superimpositions between pairs with quite comparable % identity and similarity may show very different RMSD values and vice versa.

Comparative analysis of secondary structure elements

Available structures were superposed and then tiled using UCSF Chimera 1.8.1 to keep the same orientation and to avoid visual superposition. This way, variation of secondary structure elements among individual structures can be clearly distinguished and viewed. In order to exclude any artifact from modeling, only the six available solved structures were compared. In terms of secondary structure, three subregions can be distinguished within the HA2 stem [see Additional file 1, panel A]: an α subregion and two β subregions (being either proximal or distal to the VED). The former consists of α helices A-C-D and the B loop (that upon fusion becomes B helix [1]). No meaningful variation - in terms of secondary structure - is found in the α subregion of the stem, because structural changes only concern the B loop [see Additional file 1, panel B], which indeed is unfolded in the pre-fusion state. The B loop coordinates depend on crystallization conditions and in particular on pH [14]. The VED-proximal and distal β subregions are recognized by respectively antibodies CR6261 and CR8020 [24]. The VED-proximal β subregion shows a varying number (zero, two or four) of β strands [see Additional file 1, panel C] and such variation is not relevant to antibody recognition specificity. For instance, a four-strands structure is shared between H5 (recognized by CR6261) and H3 (not recognized); moreover, a two-strands structure is shared between H2 (recognized) and H7

(not recognized). Secondary structure variation is evident also in the distal β subregion [see Additional file 1, panel D], but once again it does not relate to antibody recognition: e.g., CR8020 recognizes subregion from H7 but not corresponding one from H5. Given that subregions recognized by each antibody are clearly different (CR6261 recognizes H1, H2, H5 and H9 independently on they are showing either zero, two or four β strands) such a preliminary analysis demonstrates that secondary structure variation as viewed by cartoon representation is not indicative for epitope variation. Secondary structure variation in the globular RBD-VED region is poorly evident, according to the aforementioned 'two-tracks' rule: mutations altering the overall backbone/fold of the RBD would impair binding to host cells hence conservation (track 1) is needed to keep such basic function. However, local variation (track 2) is needed to modulate surface features hence interactions. Therefore, we did not further investigate secondary structure variation and moved instead to surface analysis, considering both most relevant features: (i) electrostatic charge distribution and (ii) hydropathy/hydrophilicity patches.

Comparative analysis of electrostatic potentials

In order to perform analyses taking into account the influence of ionic strength (I), the spatial distribution of the electrostatic potential was calculated at both I = 0 mM (Coulombic interactions unscreened by counter-ions) and I = 150 mM (physiological), assuming +1/-1 charges for the counter-ions. Prior to electrostatic potential calculations, partial charges and van der Waals radii were assigned with PDB2PQR [25,26]; then, linear Poisson-Boltzmann (PB) equation calculations were carried out by using Adaptive PB Solver (APBS) [27] through Opal web service (see Methods). The spatial distribution of the electrostatic potential was determined for each HA subregion, monomers and trimers, comparing the six available HA structures to identify possible HA-specific signatures. In particular, we focused on the role of charge distribution as visualized by isopotential contours within the tertiary structure and on classifying conservation and divergence among the different HAs. In order to evaluate electrostatic distance (ED) also in a quantitative way, clustering of the spatial distributions of the electrostatic potentials was obtained by WebPIPSA (Protein Interaction Property Similarity Analysis; [28], having the use of Hodgkin and Carbo similarity index (SI) [29] (see Methods). The Carbo SI is sensitive to the shape of the potential being considered but not the magnitude, whereas the Hodgkin SI is sensitive to both shape and magnitude. Therefore, WebPIPSA results obtained using the Hodgkin SI are shown in Figures 1, 2, 3, 4 and 5, and evidence from analyses performed using the Carbo SI is cited to confirm parameter independent data.

Stem subregions

The electrostatic patches at ionic strength I = 0 mM clearly show for all six stems preferential side disposition (Figure 1, top left), as observed for SNAREs [30]. In particular, density of negative potential (red) at the 0° side is higher than at the 180° side; positive potential (blue) shows a reverse distribution, highest density being at the 180° side. At physiological ionic strength (Figure 1, top right), preferential distribution of the positive potential (180° side) is more evident, whereas higher density in negative potential (0° side) is less evident, because most Coulombic interactions are masked by counter-ions. When considering individual stem variation, net charge roughly doubles from the −8 *e* value of H1 and H9 to −15 *e* of H7. However, similar net charge does not necessarily correspond to similar distribution (along the stem) of the potential, that can preferentially locate at either the VED-distal stem subregion (left side in figure) or at the VED-proximal one (right side). This is the case for H1 and H9 stem, sharing net charge −8 *e*, and showing (more evident at I = 0 mM) preferential VED-distal and VED-proximal negative potential, respectively. Such preferential VED-distal location of the negative potential shown by H1 is conserved also in the other two stems from Group 1, in spite of their different net charge (−10 *e*). Positive potential is more homogeneously distributed along all stems. Heat maps and corresponding density plots (Figure 1, bottom) depict the overall similarity among HA stem electrostatic profiles. Comparison between the density plots at I = 0 mM and I = 150 mM highlights a general increase in distance, i.e. a peak shift from middle ED (green region) to high ED (cyan/blue region). When comparing Group 1 stems to those from other groups it can be noticed that - at both ionic concentrations - H3 is slightly closer to Group 1 than H7, while H9 is far apart. However, H9 distance is not homogeneous with respect to the three Group 1 stems, as it is closer to H2 than to H1 and H5. Indeed, H9 stem is also quite far from H7 because it shows the highest overall distance, with respect to other stem structures. When using WebPIPSA, the distance matrix of the electrostatic potential can also be displayed as a tree referred to as 'epogram' (electrostatic potential diagram). Epograms [see Additional file 2] further highlight at both ionic concentrations that: (i) H9 stem shows unique electrostatic features (i.e., the highest ED with respect to other stems) and (ii) H7 is closer to H3 than to other stems. This clustering is confirmed when using Carbo SI. The highest electrostatic distance shown by H9 might depend on its mammalian (swine) rather than avian origin. Therefore, structural models were obtained by homology modeling for avian H9 (A/Chicken/Jiangsu/H9/2010(H9N2), UniProtKb AC: G8IKB3) and horse H3 (A/Equine/Mongolia/56/2011(H3N8); UniProtKb AC: J9TJ60),

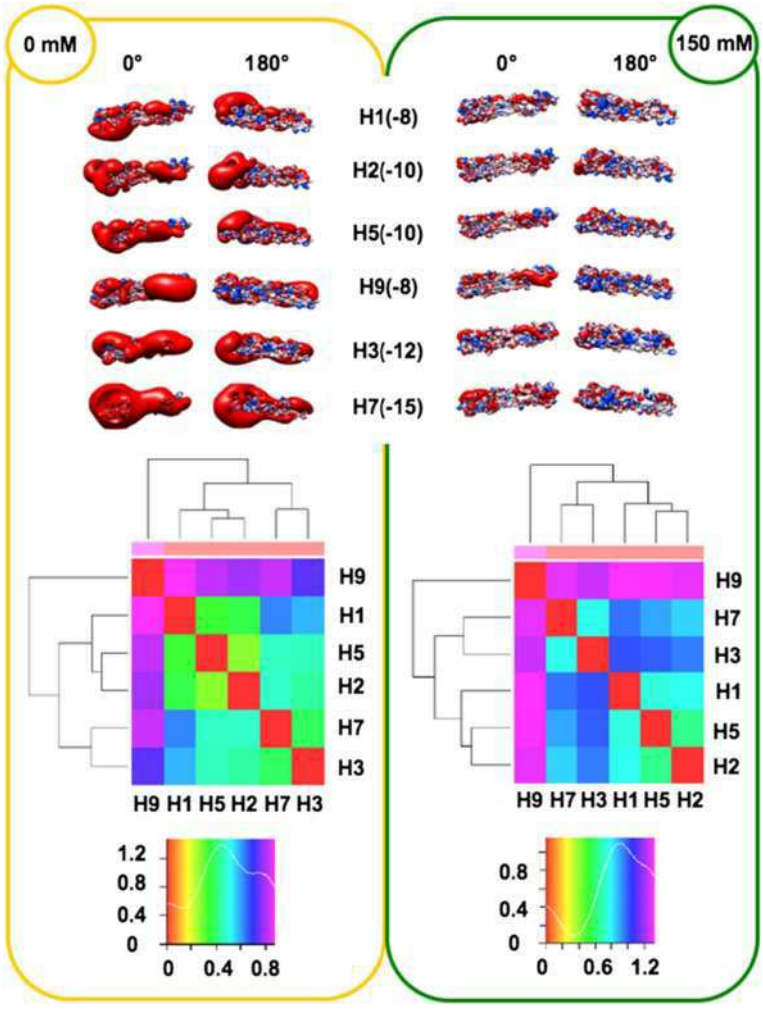


Figure 1 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA stems. Electrostatic features are shown at I = 0 mM (left panel, yellow) and I = 150 mM (right panel, green). Electrostatic potentials (blue for positive and red for negative) are presented in two orientations (0° and 180°). Isopotential contours are plotted at $\pm 3k_BT/e$. In heat maps, red, warm and cold colors correspond to identity, low and high Electrostatic Distances (ED), respectively. The number of hits (pairs of compared structures) is plotted and referred to as 'Density'.

using as structural templates 1JSD (H9) and 1MQL (H3), respectively and investigated using WebPIPSA. Comparison of epograms alternatively including either the avian H9 model or the swine template showed conservation of the highest distance observed for H9: at I = 0 mM, swine/avian epogram clustering was congruent; at I = 150 mM, avian H9 sorted with H3 and H7; this notwithstanding, highest distance of H9 from other HAs was anyway kept [see Additional file 3]. Concerning equine H3, it sorted like avian H3 at both I = 0 mM and I = 150 mM (congruent epograms see Additional file 3). In conclusion, electrostatic distance is not significantly influenced by taxonomy hence segregation depends on HA-specific features.

RBD subregions

As with the stem subregion, charge separation onto the RBD surface is more evident at I = 0 mM. Group 1 RBDs have an overall slightly negative (H1 and H2) or neutral (H5) net charge, which is positive (up to +3*e* in H3) in other groups. At large, the RBD net charge is less negative than stems (Figure 2, top). Side disposition in RBDs is not 'side preferential' as for stems, and no meaningful difference is observed when comparing the 0° and 180° views. However, preferential local distribution is clearly apparent also for RBDs, when a roughly orthogonal axis is considered: negative charges are densely distributed at the VED-proximal region (left side in figure), whereas charge of the VED-distal region (right side) is more positive. This is particularly evident for Group 1 RBDs at I = 0 mM. At physiological ionic strength, such preferential distribution is less evident, in particular for H3, where differently charged patches are interspersed. Peaks at the blue/purple regions in density plots (Figure 2, bottom) depict high electrostatic distances

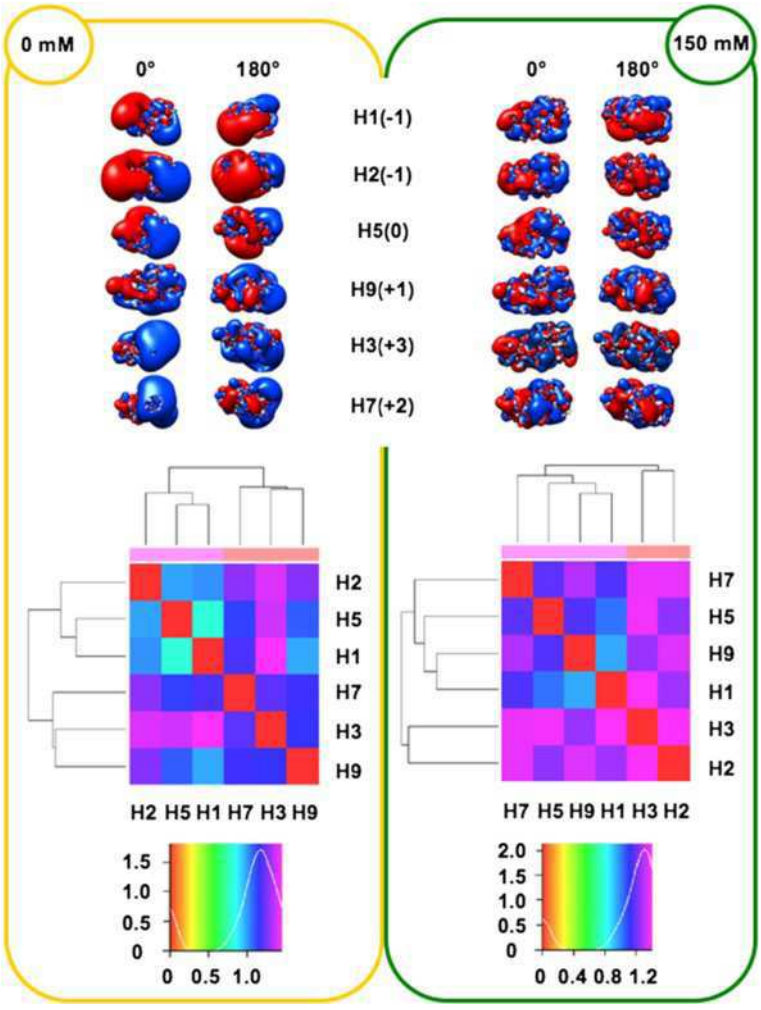


Figure 2 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA RBDs. See Figure 1 caption for color code and definitions. Isopotential contours are plotted at $\pm 1 k_B T/e$.

at both ionic strengths. Surprisingly - and independently on using either Hodgkin or Carbo SI - at I = 150 mM, the electrostatic potential of the H5 RBD is closer to H9 and H7 than to RBDs from H2, in spite H5 and H2 belong to the same Group. Splitting of Group 1 is confirmed by epogram [see Additional file 2] at I = 150 mM: H5 and H1 create a new cluster with H7 and H9.

HA1 subregions

Once the electrostatic analysis is repeated for the whole HA1 region, including the VED and F' subregions in addition to the RBD [14], the most evident difference is an overall shift towards net positive charge (see upper panels in Figures 2 and 3), according to the presence of basic patches in F' subregions [2,6]. Comparison of density plots (RBD vs. HA1) shows that peaks similarly locate at the high distance blue/purple regions (see lower panels in Figures 2 and 3) but, at I = 150 mM, Group 1 no longer splits, as H1, H2 and H5 form a cluster including H9. Resembling RBD distances, it also occurs with HA1 that members from Group 1 (H1 and H5) can be closer to an outgroup (H9) than to a member of the same group (H2) (see at I = 150 mM both heat map in Figure 3 and epogram in Additional file 2). This parameter independent evidence further highlights the relevance of counter-ions to shape the final electrostatic profile, as well as the possible disagreement between classic clustering (based on phylogenetic and serologic data) and electrostatics of the RBDs.

Monomers

The net charge is negative for all monomers, ranging -4*e* to -11*e* (Figure 4, top). Evidence that the net charge is quite negative for all stems (−8*e* to -15*e*) while being close to 0 for RBDs (−1*e* to +3*e*), stresses the total charge balancing by local basic patches in VED and F'

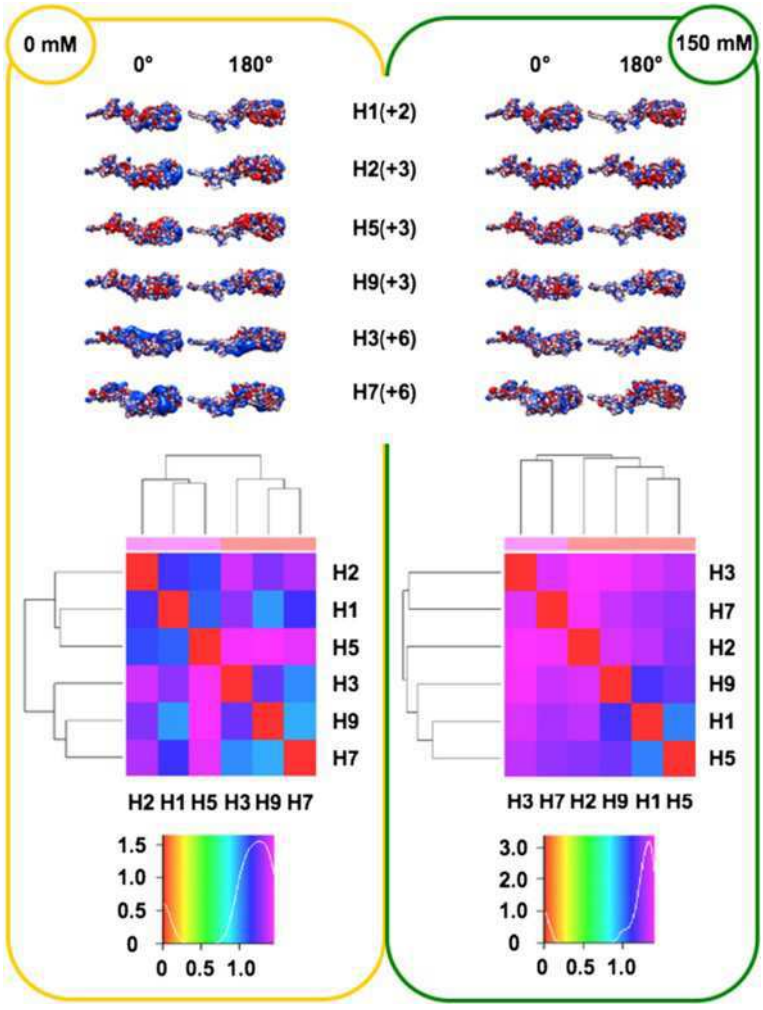


Figure 3 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA1 subregions. See Figure 1 caption for details. Isopotential contours are plotted at $\pm 2k_BT/e$.

subregions. Once again, peculiar electrostatic features are evident (and SI independent) for H9, characterized by the less negative net charge and forming its own branch at both I = 0 mM and I = 150 mM (heat maps in Figure 4, bottom, and epograms in Additional file 2). Disagreement with serological and phylogenetic data is less evident when performing electrostatic analysis with entire monomer structures, as shown by clustering of Group 1 members in Figure 4 and Additional file 2.

Trimers

Once the entire haemagglutinin functional unit is analyzed, disagreement with serological and phylogenetic clustering is highlighted again by Group 1 splitting; in particular (and independently on which SI is used) at I = 0 mM, H1 sorts separately from H2 and H5 (see Figure 5, trimer heat maps and Additional file 2, trimer epograms). Such splitting is also observed at I = 150 mM, as H5 and H1 sort with H9 and H7, whereas H2 sorts out with H3. Comparison of net charges from monomers and corresponding trimers unveils striking doubling vs. triplication mechanisms: trimer net charge values for H1 and H3 is roughly three-fold with respect to corresponding monomers, or even more (–37*e* vs. -11*e*) for H5. Instead, trimer values are only roughly twofold increased for H2, H7 and H9. Therefore, different orientations of monomers within corresponding trimers results in significant modulation of the trimer surface electrostatic charge and this in turn can be quite relevant to HA interactions. Different HA clustering at I = 0 mM and I = 150 mM may highlight the importance of ionic screening of coulombic interactions [31,32]. As a final remark, based on absence of net charge-based clustering in any executed electrostatic analyses, the spatial distribution of electrostatic potential is suggested to be more suitable than net charge alone for eventual use as a further 'signature' for protein/domain function.

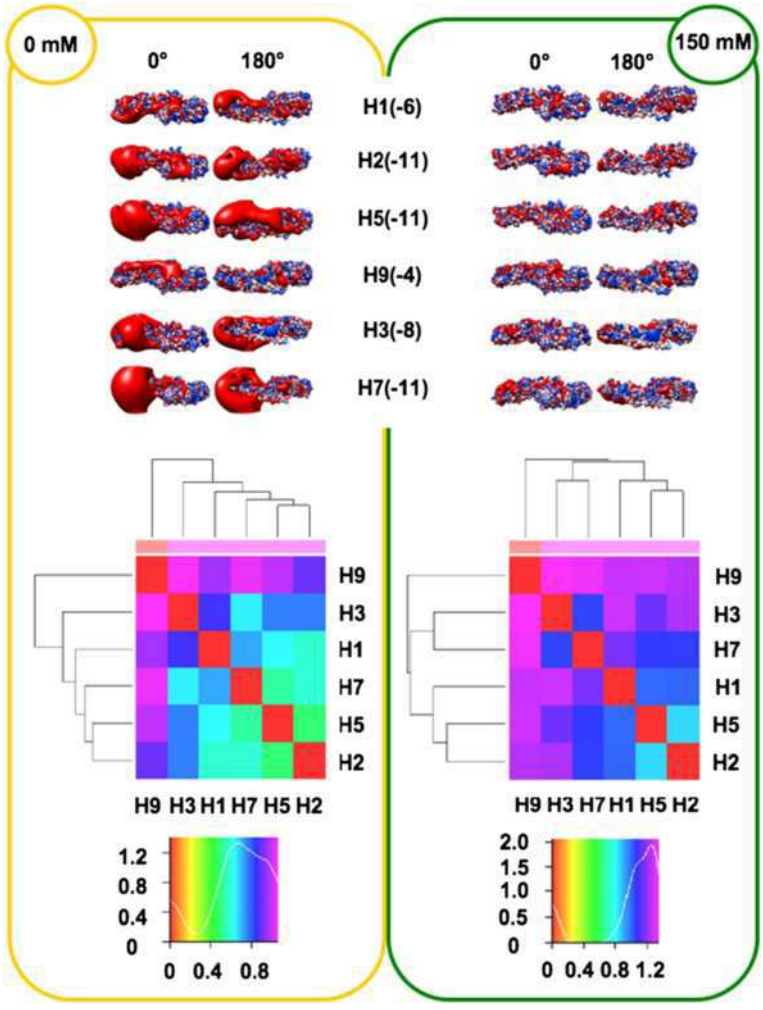


Figure 4 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA monomers. See Figure 1 caption for details. Isopotential contours are plotted at $\pm 2k_BT/e$.

Hydrophobicity analysis

Search for HA-specific motifs/signatures can be integrated by hydropathy analysis. Both electrostatics and hydrophobicity are key determinants in surface properties hence in regulating protein interactions. In particular, hydrophobic patches located at the protein surface create unstable areas. The identification of well-defined patches rather than a 'patchwork surface' of hydrophobic and hydrophilic areas can thus shed light on molecular evolution of haemagglutinin. Stem, RBD and HA1 profiles were obtained and compared using ProtScale [33] and Protein Hydrophobicity Plots [34]. Profiles from the stem subregions did not unveil any clearly meaningful difference and thus are not shown here.

RBD subregions

Figure 6 shows GRand AVerage hYdrophobicity (GRAVY) indexes, Kyte-Doolittle plots and 0° +180° surface hydropathy views for the RBDs from the six available HA structures. Similar to total electrostatic charges, GRAVY indexes are reported here for completeness of information; however, they are not suitable for use as evolutionary or functional fingerprint. In fact, variation of GRAVY values amongst the six RBDs does not correspond to high conservation and fine tuning of their surface patches as depicted in 0° and 180° views. However, comparison of Kyte-Doolittle plots could infer variation at specific positions. Plots in Figure 6 always start by residue 1 because the default numbering system from the software refers to analyzed sequence fragments (RBDs in this case); therefore, for Reader's convenience, hereafter we report both real numbers (referring to complete protein sequences) and software output numbers (between parentheses). Within Group 1, the highest intra-group hydrophilicity is shown by H1 positions Arg223 (160) of the 220-loop and by H2 at

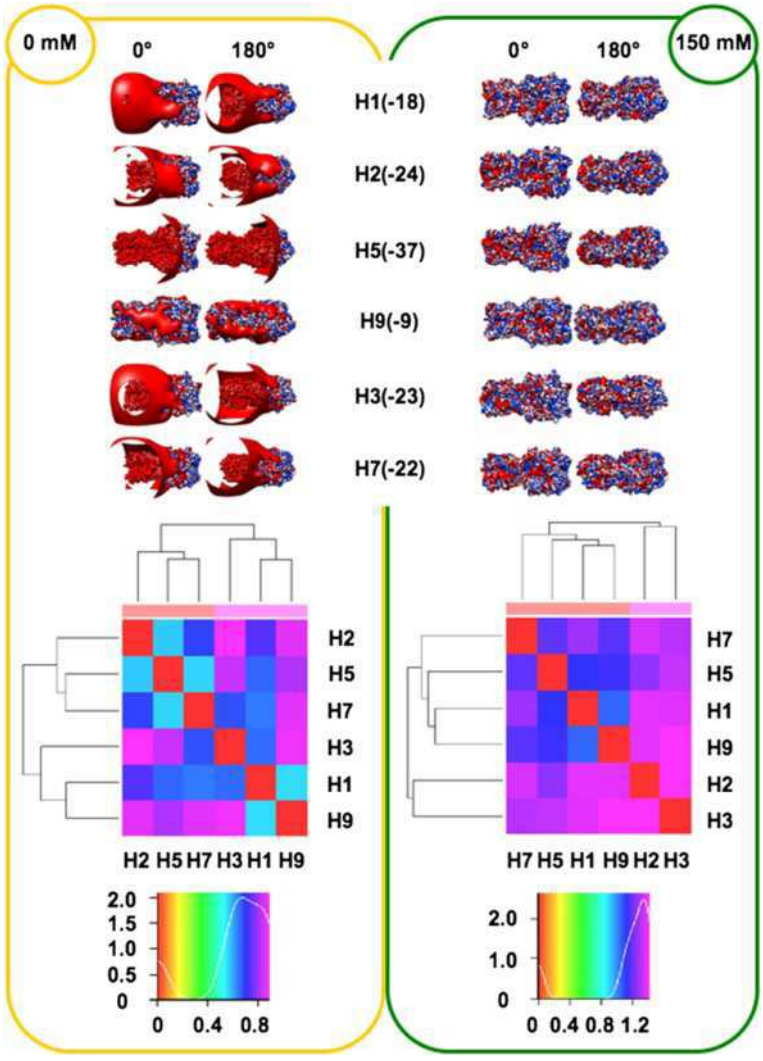


Figure 5 Isopotential contours (top), heat maps (middle) and density plots (bottom) of HA trimers. See Figure 1 caption for details. Isopotential contours are plotted at $\pm 3k_BT/e$.

positions Asn80, Ser136 and Glu202 (17, 73 and 139). At position 112 (49), H1 is significantly more hydrophobic (Ile) than H2 and H5 (Asn). Inter-group comparison highlights in H3 three hydrophilic peaks centered on residues Asp191, Thr208 and Gln227 (114, 135 and 154), as well as increased hydrophobicity of H7 in subregion 105–155 (50–100). Comparative analysis of surface patches unveiled possible HA-specific fingerprints. Within Group 1, variation concerns both the VED and RBD subregions. Such variation is even more evident when extending comparison to H9, H3 and H7. Hydrophobic patches (light and dark orange) are variable in terms of position and area. Comparison of 0° views highlights a large orange surface encompassing the VED-RBD border, specific to H9. Moreover, H5 and H7 show at the VED subregion a hydrophilic (violet) surface (green ovals) that in other HAs includes at least one small orange patch. Comparison of 0° views shows that H2 and H3 share three hydrophobic spots in an RBD subregion (blue circles) where other HAs can lack one, two or even all such spots. Further variation can be observed, and in general it seems to concern 'position-shifting' rather than significant difference in the total ratio of hydrophilic/hydrophobic surfaces. Therefore, combined variation in both electrostatic and hydropathy features is likely to fine tune local interaction properties of the different HA RBDs.

HA1 subregions

Apart from differences already observed in the RBD subregion, no further meaningful variation was found among HA1 hydropathy profiles. The only relevant evidence concerns the hydrophilicity peak at position 297 in H3 haemagglutinin (not shown).

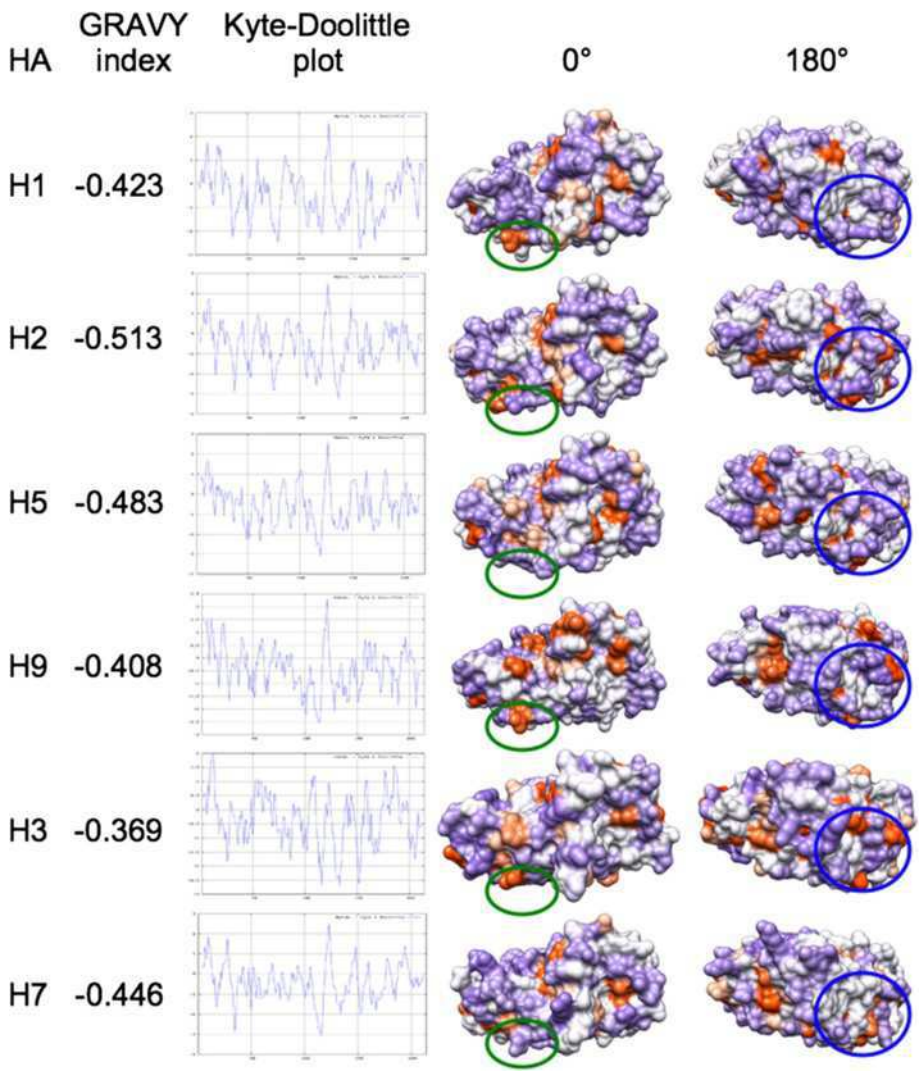


Figure 6 Hydrophobicity analysis of the RBD subregions from the six available HA structures. GRAVY Index, Kyte-Doolittle plots and surface hydrophobic (orange) and hydrophilic (violet) patches (as both 0° and 180° views) are depicted.

Structural modeling of H5N1 clades and electrostatic features comparison

Electrostatic features can vary among different types of haemagglutinins (see above). This prompted us to further investigate on differential electrostatic features as a possible fingerprint for monitoring viral evolution, i.e. as a tool to distinguish among circulating/spreading and extinguished H5N1 clades. Table 2 resumes relevant data concerning the ten clades used for this analysis; their geographical spread is shown in Figure 7. Spreading of no longer circulating clades (0, 3, 4, 5, 6, 8 and 9) is restricted to the eastern part of China and to Vietnam (see Figure 7, zoom in map); noticeably, all such clades share one or more outbreak areas with the most ancient clade (clade 0, black spots). Among circulating clades, clade 7 was also found in western China and clade 1 also spread towards India and Indochina countries (Thailand, Laos, Cambodia and Malaysia). The widest spreading concerns circulating clade 2 (red dots in the upper map of Figure 7), having reached Japan and Korea, Mongolia, Russia, several countries from Middle-East and Europe (including UK) as well as a number of African countries from the Northern hemisphere. So far, spreading of H5N1 viruses neither concerns Americas nor any country from the Southern hemisphere (Oceania and sub-equatorial Africa).

Based on a very high, average % identity (over 90%) of the clade target sequences with the available structural H5 template (PDB: 3S11), structural models for clades 0 to 9 were obtained by homology. Given that distribution of surface charge is strongly influenced by the orientation of side chains, models refinement was performed using a number of tools based on different algorithms: SCWRL [35,36], ModRefiner [37] and SCit [38]. Then, QMEAN server was used to check model quality;

Table 2 H5N1 clades

Clade	Year	Strain name	Genomic Ac	Protein Ac
0	1996-2002	A/Goose/Guangdong/1/1996	AF144305.1	AAD51927.1
1 (c)	2002-2003	A/Quail/Shantou/3054/2002	CY028946.1	ACA47648.1
2 (c)	2005	A/Bar-headed Gooze/Qinghai/75/2005	DQ095619.1	AAZ16276.1
3	2000-2001	A/Duck/Hong Kong/2986.1/2000	AY059481.1	AAL31387.1
4	2002-2003 2005-2006	A/Duck/Shantou/700/2002	CY028943.1	ACA47615.1
5	2000-2003 2004	A/Duck/Zhejiang/52/2000	AY585377.1	AAT12042.1
6	2002-2004	A/Duck/Hubei/wg/2002	DQ997094.1	ABI94747.1
7 (c)	2002-2004 2005-2006	A/Chicken/Shanxi/2/2006	DQ914814.3	ABK34764.2
8	2001-2004	A/Chicken/Hong Kong/61.9/2002	AY575876.1	AAT39076.1
9	2003-2005	A/Duck/Guangxi/50/2001	AY585375.1	AAT12040.1

Periods (years) of circulation, strain names (based on year and location of identification) and accession numbers (for both genomic and protein data) are reported for each clade. Circulating clades are marked by (c).

QMEAN is a scoring function that measures multiple geometrical aspects of protein structure, ranging 0 to 1 with higher values indicating more reliable models [39]. QMEAN scores for each refined or not refined model (mQMEAN) and the average QMEAN score for each ten clades model series (aQMEAN) was calculated. Models refined by SCWRL showed the highest aQMEAN (0.734), with highest mQMEAN for clades 0, 1, 2, 3 and 5. However, quality was similarly good when models were not refined (aQMEAN: 0.724; highest mQMEAN for clades 6 and 7) or refined by ModRefiner (aQMEAN: 0.720; highest mQMEAN for clades 4, 8 and 9), confirming once again reliability and robustness of the SWISS-MODEL homology modeling method [40]. SCit refined models showed the lowest average quality (aQMEAN: 0.702). Therefore, electrostatic analyses were performed thrice, using the ten clades models: (i) refined by SCWRL, (ii) refined by ModRefiner and (iii) not refined.

Preliminary comparison at trimer and monomer level showed meaningful variation only at the VED-RBD subregion. In fact, direct comparison of stems did not allow for inferring any clade-specific signature as all clades were found to share - at both I = 0 mM and I = 150 mM - the typical isocontour of the H5 stem (see Figure 1, top). Moreover, apart from electrostatic differences in the VED-RBD subregion, no further meaningful variation was observed among HA1 isocontours. This prompted us to 'zooming in' variation analysis at the RBD subregion level.

Figure 8 illustrates local charge variation in RBD isocontours among H5N1 clades. Even though variation is more evident at I = 0 mM, meaningful difference is kept hence highlighted at physiological ionic strength. It is noteworthy that, independently on models are refined or not and on algorithm used for refinement, the same relevant local changes in RBD isopotential contours are found (see Figure 8, panels A to C). Early clades evolution is characterized by a charge shift event at the 220-loop: in the most ancient clade (clade 0), the side chain of amino acid 228 shows either negative (Glu: 50/89 and Asp: 1/89 sequences) or positive (Lys: 38/89 sequences) charge. The positive charge is 'fixed' in the most recent, and still circulating clades 2 (Lys: 308/310, Glu or Asp: 0/310 sequences) and 7 (Lys: 25/26; Glu: 1/26 sequences) (see Figure 8 and Table 3). Further loss of a negative residue (Asp) concerns the VED isocontour at the 110-helix region. Table 3 shows that in clade 0, position 110 is negatively charged (Glu or Asp: 67/89 sequences) or polar, non-charged (Asn: 22/89 sequences). This negative charge is almost completely lost in clade 2 (Asp: 3/310, Glu: 0/310), while being retained (Asp: 26/26) in clade 7; however, this latter clade shows ongoing loss of the negative charge at position 104 (Asp: 15/26; Gly: 11/26), that is positively charged in 100% of clade 0 and clade 2 sequences (Figure 8 and Table 3). In clades 2 and 7, such 'denegativization' of the VED isocontour is somehow counterbalanced by negativization (or depositivization) at the properly receptorial part of the RBD. In clade 2, this depends on Asn140Asp mutation (in 307/310 sequences) while in clade 7 both depositivization (Arg178Val in 8/26 sequences) and negativization (Ala200Glu in 12/26 sequences) mutations are observed (Figure 8 and Table 3). Intriguingly, when considering aforementioned replacements altogether, evolution of H5N1 still circulating clades seems having been characterized by an isocontour rearrangement based on a VED-to-RBD flow of negative charges; this process is 'partial' hence seemingly in progress in clade 7 (mutation arose in the clade and it is present, at least so far, in less than 50%

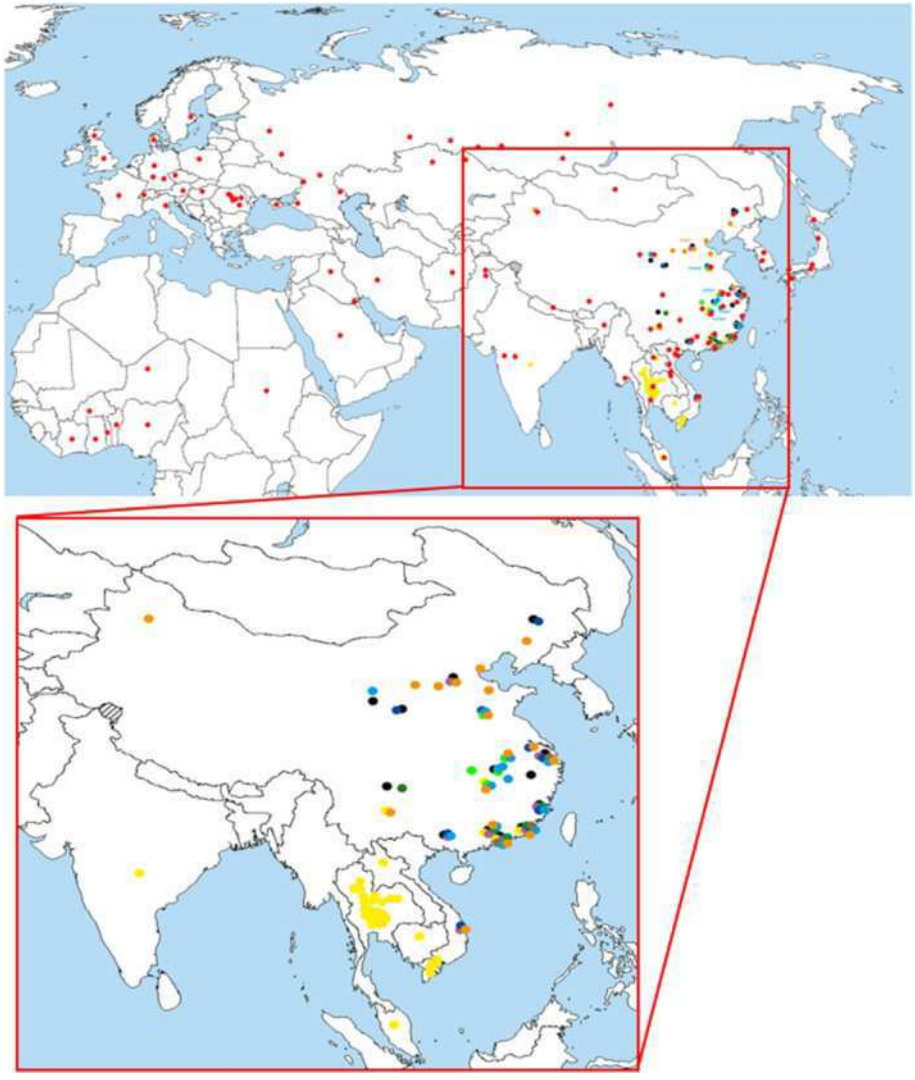


Figure 7 Geographical spread of H5N1 clades. Outbreak areas for each clade are color coded as follows: 0, black; 1, yellow; 2, red; 3, violet; 4, dark green; 5, dark blue; 6, light green; 7, orange; 8, brown; 9, cyan.

sequences) whereas it is complete and 'fixed' (99% sequences) in clade 2. Given that comparison of the six different HA structures identified HA-specific variation in both electrostatic and hydropathy features, and that specific electrostatic signatures of the RBD could also be associated to the ten H5N1 clades, clades analysis was integrated by comparison of the RBD surface hydropathy profiles (Figure 9). As for electrostatic analysis, the most ancient clade (clade 0) is the reference for tracking hydropathy profile variation along clades evolution. As previously explained, hereafter both real protein sequence numbering and (between parentheses) software output numbering is reported for Reader's convenience. Clade 3 shows no substantial difference with respect to clade 0, at least in terms of hydropathy plots. Instead, clade 4 shows increased hydrophilicity at position Asn211 (148). Clade 1 shows increased hydrophobicity around position Ser140 (77). Replacement at position 124 of a polar residue in clade 0 by Ile in all other clades results in increased hydrophobicity. Intriguingly, the hydropathy profile of clade 7 resembles the one of H3 haemagglutinin, including its aforementioned three hydrophilicity peaks. Please note that the apparent disagreement among positions of the three H3 peaks in Figure 6 and those from Clade 7 in Figure 9 is not confirmed in real numbering, as plot shift is determined by ten extra residues present in the really N-terminal region of H3. Apart from difference illustrated so far for the RBD, no further meaningful variation was observed when comparing other HA1 subregions or the stem profiles (not shown).

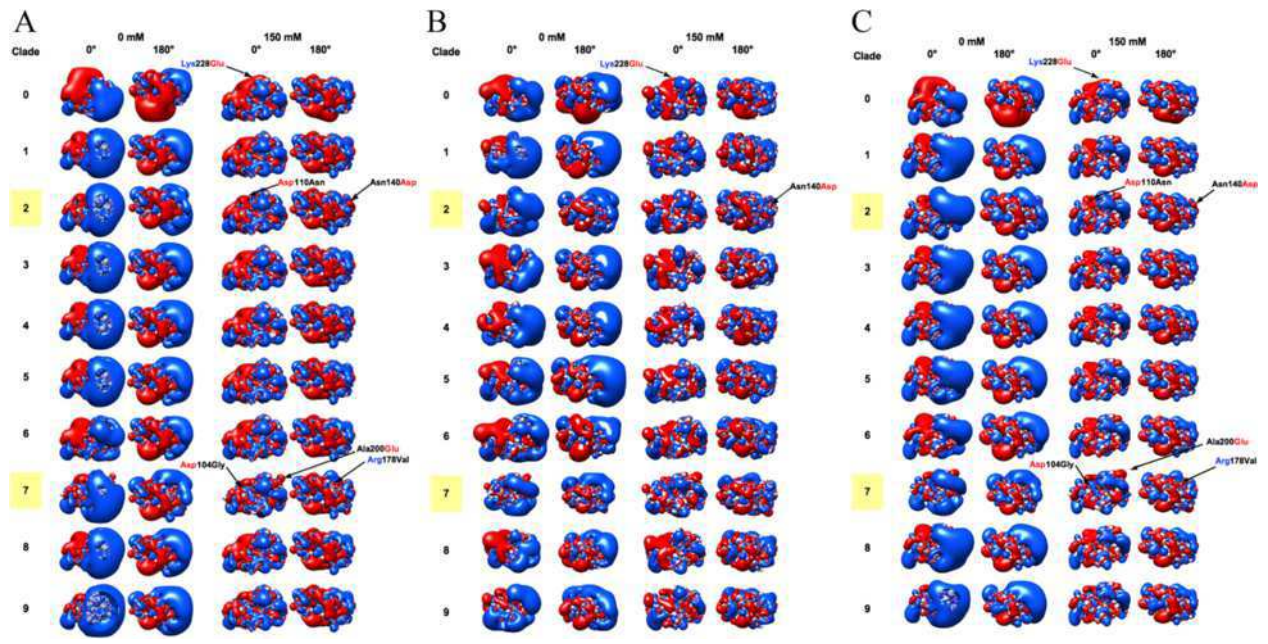


Figure 8 Isopotential contours of the RBD subregions from H5N1 clades 0 to 9. See Figure 1 caption for color code and definitions. Isopotential contours are plotted at $\pm 1 k_B T/e$. Specific mutations discussed in the text are highlighted. Positively charged residues are written in blue and negatively charged ones in red. Panels: models refined using SCWRL **(A)**, ModRefiner **(B)** or not refined **(C)**.

Conclusions

Evidence from this work shows that sequence homology is often, but not always, related to structural similarity and vice versa. In fact, in some instances, protein domains with less related sequences can show intriguing structural closeness. Therefore, in order to obtain a more complete view of the 'functional evolution', phylogenetic analyses based on sequence comparison and resulting in trees, might be integrated taking into account information from structural comparison. Dissimilarity in secondary structure elements does not always result in different antigenic properties. Sometimes, secondary structure is not prominent to the molecule antigenicity. Indeed, electrostatic features are crucial to interactions and in fact electrostatic profiles of the RBD subregion varies amongst different HAs. On the other hand, stems, HA1, monomers and trimers topology appears to be variable. As shown by H9 and H3 modeled structures, electrostatic profiles seem to depend on HA type rather than organism source. Hydrophobicity analysis reveals that local, 'spot' variation especially concerns the RBD subregion. No flow of hydrophobicity/hydrophilicity

Table 3 Mutations in H5N1 clades 0, 2 and 7

Clade	Sequences	Position					
		104	110	140	178	200	228
0	89	Asp = 89	Asp = 64	Asn = 86	Arg = 89	Ala = 89	Glu = 50
			Asn = 22	Asp = 3			Lys = 38
			Glu = 1				Asp = 1
2.2	310	Asp = 310	Asn = 302	Asp = 307	Arg = 284	Ala = 307	Lys = 308
			Lys = 4	Asn = 2	Ile = 26	Gly = 3	Asn = 1
			Asp = 3	Gly = 1			Gln = 1
			Ser = 1				
7	26	Asp = 15	Asp = 26	Asn = 24	Arg = 16	Ala = 14	Lys = 25
		Gly = 11		Asp = 2	Val = 8	Glu = 12	Glu = 1
					Gly = 2		

For each clade, the number of analyzed available sequence is shown. For each position (numbering refers to clade 0 sequence), the type of present residues and corresponding number of sequences showing that residue is shown.

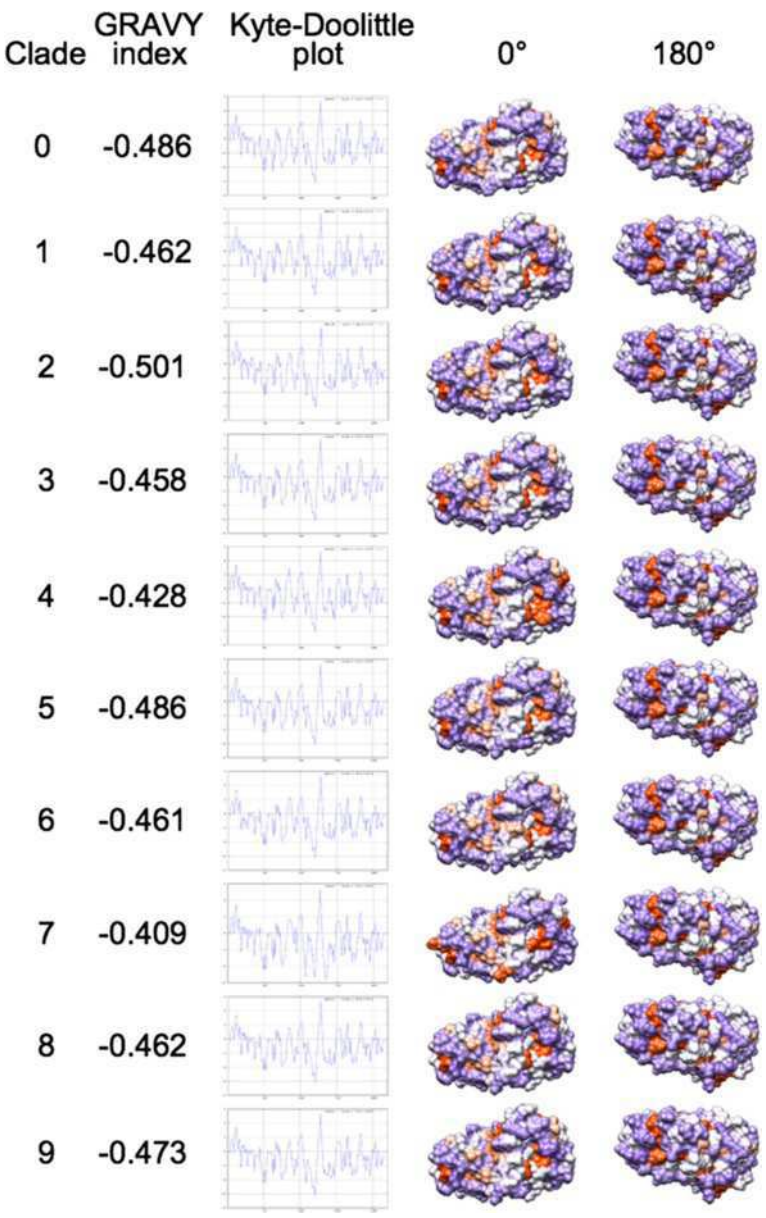


Figure 9 Hydrophobicity analysis of the RBD subregions from H5N1 clades 0 to 9. See Figure 6 caption for color code and definitions.

is observed as for charge flow in the electrostatic analysis. In H5N1 clades comparison, from an electrostatic point of view, meaningful variation concerns only the VED-RBD subregion. Intriguingly, a charge flow specifically concerns still circulating clades 2 and 7, where 'denegativization' of the VED isocontour is counterbalanced by negativization in the RBD. It is noteworthy (and a 'positive mark' for robustness of the observation) that the same specific differences are found when comparing refined or not refined clade models or models refined using different algorithmic strategies (as SCWRL is rotamer library-based [35,36] while ModRefiner is based on two-step atomic-level energy minimization [37]). Given that local charge concentration is typical for antigenic epitopes, it is tempting to speculate that charge redistribution in such clades might have contributed to antigenic escape hence to their evolutionary success and spreading. Indeed, such an hypothesis is in agreement with evidence that charge redistribution on the RBD characterizes the two clades (2 and 7) which were able to spread over the largest geographical distribution and that, in particular, such redistribution is fixed in sequences from clade 2, which is the world most spread clade. It is noteworthy that also variation in hydrophobic patches is especially observed in the RBD subregion.

Methods

Structural templates and target sequences

The following structures from the Protein Data Bank (PDB) were used as templates for modeling: H1, PDB 1RUZ, from viral strain A/South Carolina/1/1918(H1N1); H2, PDB 2WR5, from Asian pandemic influenza virus of 1957; H3, PDB 1MQL, from viral strain A/duck/Ukraine/1963 (H3N8); H5, PDB 3S11, from viral strain A/Goose/Guangdong/1/1996 (H5N1); H7, PDB 1TI8, from viral strain A/turkey/Italy/214845/2002(H7N3); H9, PDB 1JSD, from viral strain A/swine/Hong Kong/9/98(H9N2). UniProtKb accession codes (AC) of target sequences modeled by H.M. and corresponding viral strains (VS) are the followings: H4, AC F2NZ53, VS A/duck/Guangxi/912/2008(H4N2); H6, AC H8PBW2, VS A/duck/Fujian/6159/2007(H6N6); H8, AC D4NQL7, VS A/northern pintail/Alaska/44420-106/2008(H8); H10, AC P12581, VS A/Chicken/Germany/n/1949 (H10N7); H11, AC D5LPX8, VS A/turkey/Almaty/535/2004(H11N9); H12, AC E6XYK2, VS A/mallard/Interior Alaska/9BM1907R1/2009(H12); H13, AC P13101, VS A/Gull/Astrakhan/227/1984 (H13N6); H14, AC P26136, VS A/Mallard/Astrakhan/263/1982 (H14N5); H15, AC Q82565, VS A/duck/Australia/341/1983(H15N8); H16, AC Q5DL23, VS A/black-headed gull/Sweden/3/99(H16N3). Given that original UniProtKb sequences indeed correspond to H0 precursors, sequence fragments missing in mature chains were manually removed to avoid improper structural alignment.

Structural superpositions, Homology Modeling, model refinement and quality check

Structural superpositions were performed and viewed using UCSF Chimera [18] v. 1.8.1 (free download from [41]). Target protein sequences were modeled on best available structure templates using SWISS-MODEL [40]. Then, model structures were refined using SCWRL [35,36], ModRefiner [37] or SCit [38]. Model quality was checked via QMEAN server [39].

Electrostatic surface analysis

Isopotential contours were calculated using UCSF Chimera 1.8.1: the software utility allows for connecting - through Opal web server - to the Adaptive Poisson-Boltzmann Solver (APBS) server [42]. Isopotential contours were then plotted at $\pm 3k_BT/e$, $\pm 2k_BT/e$ and $\pm 1k_BT/e$ (RBDs). PDB2PQR was used to assign partial charges and van der Waals radii according to the PARSE force field [43]. Interior $\varepsilon_p = 2$ and $\varepsilon_s = 78.5$ were chosen for respectively the protein and the solvent [30,44,45], T = 298.15 K. Probe radius for dielectric surface and ion accessibility surface were set to be r = 1.4 Å and r = 2.0 Å, respectively. Electrostatic distance was calculated using the Hodgkin index and the Carbo index at the WebPIPSA server [46]. Rigid-body superposition was performed and electrostatic potential was computed using Chimera 1.8.1.

Hydropathy analysis

Hydropathy analysis was performed using the Kyte-Doolittle scale implemented in Protein Hydrophobicity Plots [34] and in ProtScale at the ExPASy server [47,48]. In order to highlight hydrophilic regions likely exposed on the surface, a seven amino acids window was chosen; regions with score >0 are hydrophobic [33]. Hydrophobic/hydrophilic patches were plotted onto structures through Chimera 1.8.1.

Additional files

Additional file 1: Two-pages figure relating HA stem secondary superstructures to immunogenic epitopes.

Additional file 2: Multi-page figure reporting epograms for each analyzed HA subregions (stem, RBD, HA1) and for HA monomers and trimers.

Additional file 3: Reports comparison amongst epograms for stem subregions obtained performing the WebPIPSA analyses with solved PDB structures or replacing either H9 or H3 templates by modeled structures.

Abbreviations

AC: Accession code; APBS: Adaptive PB Solver; ED: Electrostatic distance; Epogram: Electrostatic potential diagram; GRAVY: GRand AVerage hYdrophobicity; HA: Haemagglutinin; I: Ionic strength; N: Neuraminidase; PB: Poisson-Boltzmann; PDB: Protein data bank; PIPSA: Protein Interaction Property Similarity Analysis; RBD: Receptor-binding domain; RMSD: Root mean square deviation; SI: Similarity index; VED: Vestigial esterase domain; VS: Viral strain; WHO: World Health Organization.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FF and GC conceived the study. FF oversaw the study. IR performed most of bioinformatic analyses (modeling, electrostatics, hydropathy). IR and FF interpreted the data. AM performed part of the bioinformatic analyses on H5N1 clades and provided other authors with help in data interpretation. IR and FF wrote the paper with input from GC and AM. All authors read and approved the final manuscript.

Authors' information

IR is a PhD student and a bioinformatician; AM is a staff technician at the IZSVe, currently performing the PhD course, and a molecular virologist; GC is the Head of Research and Development Department, Division of Biomedical Science, OIE/FAO and National Reference Laboratory for Newcastle Disease and Avian Influenza, IZSVe; FF is Associate Professor of Molecular Biology and Bioinformatics and the PI of the MOLBINFO Unit at the Department of Biology, University of Padua.

Acknowledgements

We thank Stefan Richter for helpful information on WebPIPSA, Walter Rocchia and Sergio Decherchi for expert suggestions on electrostatic analyses, Stefano Vanin and Isabella Monne for useful discussions. This work was supported by basic funding ('ex 60%') from the Italian Ministry for University and Research (MIUR) to FF.

Author details

[1]Molecular Biology and Bioinformatics Unit (MOLBINFO), Department of Biology, University of Padua, via U. Bassi 58/B, 35131 Padova, Italy. [2]FAO-OIE and National Reference Laboratory for Newcastle Disease and Avian Influenza, Istituto Zooprofilattico delle Venezie (IZSVe), viale dell'Università 10, 35020 Legnaro, Italy.

Received: 14 July 2014 Accepted: 28 October 2014
Published online: 10 December 2014

References

1. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA: **Antibody recognition of a highly conserved influenza virus epitope.** *Science* 2009, **324:**246–251.
2. Han T, Marasco WA: **Structural basis of influenza virus neutralization.** *Ann N Y Acad Sci* 2011, **1217:**178–190.
3. **World Health Organization** [http://www.who.int/research/en/]
4. **Center for Disease Control and prevention** [http://www.cdc.gov/datastatistics/]
5. Hamilton BS, Whittaker GR, Daniel S: **Influenza virus-mediated membrane fusion: determinants of hemagglutinin fusogenic activity and experimental approaches for assessing virus fusion.** *Viruses* 2012, **4:**1144–1168.
6. Sriwilaijaroen N, Suzuki Y: **Molecular basis of the structure and function of H1 hemagglutinin of influenza virus.** *Proc Jpn Acad Ser B Phys Biol Sci* 2012, **88:**226–249.
7. Velkov T, Ong C, Baker MA, Kim H, Li J, Nation RL, Huang JX, Cooper MA, Rockman S: **The antigenic architecture of the hemagglutinin of influenza H5N1 viruses.** *Mol Immunol* 2013, **56:**705–719.
8. Stanekova Z, Vareckova E: **Conserved epitopes of influenza A virus inducing protective immunity and their prospects for universal vaccine development.** *Virol J* 2010, **7:**351.
9. Russell RJ, Gamblin SJ, Haire LF, Stevens DJ, Xiao B, Ha Y, Skehel JJ: **H1 and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes.** *Virology* 2004, **325:**287–296.
10. Gamblin SJ, Skehel JJ: **Influenza haemagglutinin and neuraminidase membrane glycoproteins.** *J Biol Chem* 2010, **285:**28403–28409.
11. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, Ha Y, Vasisht N, Steinhauer DA, Daniels RS, Elliot A, Wiley DC, Skehel JJ: **The structure and receptor binding properties of the 1918 influenza haemagglutinin.** *Science* 2004, **303:**1838–1842.
12. Xu R, Wilson IA: **Structural characterization of an early fusion intermediate of influenza virus haemagglutinin.** *J Virol* 2011, **85:**5172–5182.
13. Sauter NK, Hanson JE, Glick GD, Brown JH, Crowther RL, Park SJ, Skehel JJ, Wiley DC: **Binding of influenza virus haemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography.** *Biochemistry* 1992, **31:**9609–9621.
14. DuBois RM, Zaraket H, Reddivari M, Heath RJ, White SW, Russell CJ: **Acid stability of the haemagglutinin protein regulates H5N1 influenza virus pathogenicity.** *PLoS Pathog* 2011, **7**(12):e1002398.
15. Ha Y, Stevens DJ, Skehel JJ, Wiley DC: **H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes.** *EMBO J* 2002, **21:**865–875.
16. Lu X, Shi Y, Gao F, Xiao H, Wang M, Qi J, Gao GF: **Insights into avian influenza virus pathogenicity: the haemagglutinin precursor HA0 of subtype H16 has an alpha-helix structure in its cleavage site with inefficient HA1/HA2 cleavage.** *J Virol* 2012, **86:**12861–12870.
17. Carugo O, Pongor S: **A normalized root mean square distance for comparing protein three dimensional structures.** *Protein Sci* 2001, **10:**1470–1473.
18. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera–a visualization system for exploratory research and analysis.** *J Comput Chem* 2004, **25:**1605–1612.
19. Wang W, Anderson CM, De Feo CJ, Zhuang M, Yang H, Vassell R, Xie H, Ye Z, Scott D, Weiss CD: **Cross-neutralizing antibodies to pandemic 2009 H1N1 and recent seasonal H1N1 influenza A strains influenced by a mutation in haemagglutinin subunit 2.** *PLoS Pathog* 2011, **7**(6):e1002081.
20. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5:**823–826.
21. De Franceschi N, Wild K, Schlacht A, Dacks JB, Sinning I, Filippini F: **Longin and GAF domains: structural evolution and adaptation to the subcellular trafficking machinery.** *Traffic* 2014, **15:**104–121.
22. Jang SB, Kim YG, Cho YS, Suh PG, Kim KH, Oh BH: **Crystal structure of SEDL and its implications for a genetic disease spondyloepiphyseal displasia tarda.** *J Biol Chem* 2002, **277:**49863–49869.
23. Jeyabalan J, Nesbit MA, Galvanovskis J, Callaghan R, Rorsman P, Thakker RV: **SEDLIN forms omodimers: characterisation of SEDLIN mutations and their interactions with transcription factors MBP1, PITX1 and SF1.** *PLoS One* 2010, **5**(5):e10646.
24. Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M, Yu W, Ophorst C, Cox F, Korse HJ, Brandenburg B, Vogels R, Brakenhoff JP, Kompier R, Koldijk MH, Cornelissen LA, Poon LL, Peiris M, Koudstaal W, Wilson IA, Goudsmit J: **A highly conserved neutralizing epitope on group 2 influenza A viruses.** *Science* 2011, **333:**843–850.
25. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA: **PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations.** *Nucleic Acids Res* 2004, **32**(Web server issue): W665–W667.
26. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA: **PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations.** *Nucleic Acids Res* 2007, **35**(Web server issue):W522–W525.
27. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA: **Electrostatics of nanosystems: application to microtubules and the ribosome.** *Proc Natl Acad Sci U S A* 2001, **98:**10037–10041.
28. Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade R: **WebPIPSA: a web server for the comparison of protein interaction properties.** *Nucleic Acid Res* 2008, **36**(Web Server Issue):W276–W280.
29. Hodgkin EE, Richards WG: **Molecular similarity based on electrostatic potential and electric field.** *Int J Quant Chem* 1987, **32**(Suppl 14):105–110.
30. Guo T, Gong LC, Sui SF: **An electrostatically preferred lateral orientation of SNARE complex suggests novel mechanisms for driving membrane fusion.** *PLoS One* 2010, **5**(1):e8900.
31. Lee KK, Fitch CA, Garcìa-Moreno EB: **Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein.** *Protein Sci* 2002, **11:**1004–1016.
32. López de Victoria A, Kieslich CA, Rizos AK, Krambovitis E, Morikis D: **Clustering of HIV-1 Subtypes Based on gp120 V3 Loop electrostatic properties.** *BMC Biophys* 2012, **5:**3.
33. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157:**105–132.
34. **Protein Hydrophobicity Plots** [http://arbl.cvmbs.colostate.edu/molkit/hydropathy/]
35. Bower M, Cohen FE, Dunbrack RL Jr: **Sidechain prediction from a backbone-dependent rotamer library: A new tool for homology modeling.** *J Mol Biol* 1997, **267:**1268–1282.
36. Canutescu AA, Shelenkov AA, Dunbrack RL Jr: **A graph theory algorithm for protein side-chain prediction.** *Protein Sci* 2003, **12:**2001–2014.
37. Xu D, Zhang Y: **Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization.** *Biophys J* 2011, **101:**2525–2534.
38. Gautier R, Camproux AC, Tufféry P: **SCit: web tools for protein side chain conformation analysis.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W508–W511.
39. Benkert P, Künzli M, Schwede T: **QMEAN Server for Protein Model Quality Estimation.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W510–W514.
40. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T: **Protein structure homology modeling using SWISS-MODEL workspace.** *Nat Protoc* 2009, **4**(1):1–13.
41. **UCSF Chimera** [http://www.cgl.ucsf.edu/chimera/]
42. **APBS server** [http://www.poissonboltzmann.org]
43. Sitkoff D, Sharp K, Honig B: **Accurate calculation of hydration free energies using macroscopic solvent models.** *J Phys Chem* 1994, **98:**1978–1988.
44. Schutz CN, Warshel A: **What are the dielectric 'constants' of proteins and how to validate electrostatic models?** *Proteins* 2001, **44:**400–417.
45. Gorham RD Jr, Kieslich CA, Morikis D: **Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization.** *Ann Biomed Eng* 2011, **39:**1252–1263.
46. **WebPIPSA** [http://pipsa.eml.org/pipsa]
47. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: **Protein identification and analysis tools on the ExPASy server.** In *The Proteomics Protocols Handbook.* Edited by Walker JM.: Humana Press; 2005:571–607.
48. **ExPASy server** [http://www.expasy.org]

doi:10.1186/s12859-014-0363-5
Cite this article as: Righetto *et al.*: Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features. *BMC Bioinformatics* 2014 **15**:363.

Comparative Structural Analysis of Hemagglutinin for Unveiling Fingerprints in the Evolution and Spreading of Avian Influenza Viruses

Heidari A., Righetto I., Filippini F.

(Submitted)

OXFORD UNIVERSITY PRESS | Bioinformatics

Comparative Structural Analysis of Hemagglutinin for Unveiling Fingerprints in the Evolution and Spreading of Avian Influenza Viruses

Journal:	*Bioinformatics*
Manuscript ID	Draft
Category:	Discovery Note
Date Submitted by the Author:	n/a
Complete List of Authors:	Heidari, Alireza; Università degli Studi di Padova, Dipartimento di Biomedicina Comparata e Alimentazione, viale dell'Università 16, 35020 Legnaro (PD) Righetto, Irene; Università degli Studi di Padova, Dipartimento di Biologia Filippini, Francesco; Università degli Studi di Padova, Dipartimento di Biologia
Keywords:	Structural bioinformatics, Protein structure prediction, Protein evolution, Influenza Virus, Electrostatics, Molecular modeling

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Structural Bioinformatics

Comparative Structural Analysis of Hemagglutinin for Unveiling Fingerprints in the Evolution and Spreading of Avian Influenza Viruses

Alireza Heidari[1,#], Irene Righetto[2,#,*] and Francesco Filippini[2]

[1]Department of Comparative Biomedicine and Food Science, University of Padua, viale dell' Università 16, 35020 Legnaro (PD), Italy, [2]Department of Biology, University of Padua, via U. Bassi 58/B, 35131 Padova, Italy.

[#]Equal contributions; *To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Avian influenza virus is a zoonotic agent with a significant impact on public health and poultry industry hence monitoring its evolution and spreading is needed. Current surveillance and tracing are mainly based on serological or DNA sequencing-phylogenetics analysis. However, virus-host interaction, antigenic drift and viral clades spreading are strongly influenced by variation in spike proteins surface features. We report here that comparative structural analysis of hemagglutinin can provide relevant evolutionary fingerprints to integrate sequence-based analyses.

Results: Phylogenetic analyses, carried out with different methods, of H9 viral strains from wild birds and poultry reliably led to clustering of viruses into five main groups. Then, structural features comparison showed congruence among such a clustering and surface fingerprints. These latters relate group specific variation in electrostatic charges and isocontours to well-known hemagglutinin sites involved in the modulation of immune escape and host specificity. This work thus suggests that integrating structural and sequence comparison may boost investigation on trends and relevant mechanisms in viral evolution.

Contact: irene.righetto@bio.unipd.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Wild waterfowl are primary reservoirs of avian influenza (AI) viruses that can also sporadically infect domestic birds and mammalian/human hosts (Nelson and Vincent, 2015). Therefore, setting up a coordinated global surveillance network and studying viral evolution is crucial to monitor genetic changes and predict 'evolutionary trends', especially when considering viral clades for which avian to mammalian/human host switch was reported (Al-Tawfiq et al., 2014). Risk for human/animal health also depends on the emergence of novel reassortant viruses, especially where multiple strains and clades co-circulate (Su et al., 2015). H5N1 AI viruses are unique in their ecological success, broad host range and geographical spreading (Guan and Smith, 2013); however, recently reassorted subtypes (H7N9, H9N2, H10N8) may also jump the host-species barrier, increasing concern for pandemic risk (Trombetta et al., 2015). Studying virus variation possibly related to low pathogenic (LPAI) to high pathogenic (HPAI) shift and antigenic drift is quite relevant to human/animal health, poultry industry and vaccine efficacy. To date, H5 and H7 AI viruses were reported to evolve from a LPAI to HPAI form after their introduction into poultry from the wild bird reservoir (Alexander 2007), and H9N2 viruses could occasionally be transmitted from poultry to mammals/humans. Therefore, H5, H7 and H9 viruses are deeply studied as top pandemic agents (Lin et al., 2000; Butt et al., 2005). Current AI vaccines are based upon the elicitation of a neutralizing antibody (Ab) response against the major hemagglutinin (HA) epitopes. HA - the main viral surface antigen - plays a central role in AI virus evolution by mediating attachment and penetration into the host cell; mutations in HA immune-dominant regions may result in antigenic

drift allowing the virus to escape Ab neutralization (Velkov et al., 2013). In H5N1 viruses, charge redistribution at the surface of the receptor binding domain (RBD) sub-region of HA relates to branching of still circulating clades with respect to no longer circulating ones (Righetto et al., 2014). In order to assess whether electrostatic variation is a general fingerprint for AI virus evolution, we screened a large dataset of H9 viruses. Comparative analysis of groups and clades confirmed congruence between phylogenetic and electrostatic clustering, suggesting these latters as widely applicable fingerprints for avian flu viral evolution.

2 Methods

Only methods underlying the structural approach presented in this paper are reported hereafter in full, whereas phylogenetics methods are reported in Supplementary file 1 together with indications on collected sequences, databases, web sites, references etc.

2.1 Structural Modeling and Model Refinement

Structural models for the RBD regions of target HA proteins were obtained by homology modelling using SWISS-MODEL (Bordoli et al., 2009) and PDB 1JSH structure as H9N2 HA template; such models were then refined using SCWRL (Krivov et al., 2009) and their quality was checked via QMEAN server (Benkert et al., 2009). Protein structures were viewed using UCSF Chimera (Pettersen et al., 2004) v. 1.11.2 (http://www.cgl.ucsf.edu/chimera/).

2.2 Analysis of Electrostatic Potentials

Comparative analysis of electrostatic potentials was performed through Opal web server connected to the Adaptive Poisson-Boltzmann Solver (APBS) server (http://www.poissonboltzmann.org/apbs), calculating the spatial distribution of the electrostatic potential at physiological ionic strength (I) = 150 mM, assuming +1/-1 charges for the counter-ions. Partial charges and van der Waals radii were assigned via PDB2PQR according to the PARSE force field (Sitkoff et al., 1994), choosing interior ε_p=2 and ε_s=78.5 for protein and solvent, respectively. Contouring was done at $\pm 1 k_B T/e$ and viewed using UCSF Chimera. Electrostatic distance (ED):

$$\text{Electrostatic Distance } D_{a,b} = \sqrt{2 - 2SI_{a,b}}$$

was calculated with Hodgkin and Carbo indexes at the WebPIPSA server (http://pipsa.eml.org/pipsa). Rigid-body superposition was performed was computed using UCSF Chimera.

3 Results

3.1 Phylogenetic Clustering of AI H9 HA

As the major spike protein and surface antigen, hemagglutinin is most commonly used for inferring AI virus phylogenetic trees (Velkov et al., 2013). Genetic diversity among AI H9 viruses was assessed to identify main groups and clades; to this aim, three methods were used: Neighbor Joining (NJ), Maximum-likelihood (ML) and Bayesian method. Such analysis reliably separated AI H9 strains into five monophyletic groups labelled as A, B, C, D, and E. See Supplementary file 1 for details on methods, approach, trees, groups and clades separation criteria, as well as a list of 'representative' viruses (from the different H9 groups and clades) characterizing this work. Group A (mostly containing strains isolated by the wild birds reservoir) shows the highest genetic heterogeneity whereas the other two large groups (B+C) show a lower intra-group nucleotide distance and are isolated from poultry. Two small groups (D+E) just contain a few strains isolated in Malaysia and the USA, respectively.

3.2 Clustering by Electrostatic Distance for AI H9 viruses: Heatmaps

A preliminary report on H5N1 related surface electrostatics to clades evolution and spreading (Righetto et al., 2014); we report here that clustering H9 viruses by electrostatic distance also shows substantial agreement to phylogenetic grouping, suggesting this as a general hallmark for AI virus evolution. In particular, semi-quantitative ED evaluation and clustering of the spatial distribution of RBD electrostatic potentials (see ED color coding in heatmaps and density plots in Supplementary file 1) shows that the electrostatic distance among H9 representative viruses from 'wild birds' groups A+D+E and those ones from 'poultry birds' groups B+C is high, whereas the distance among A, D and E is lower as well as that between B and C.

3.3 Variation in Electrostatic and Hydrophobicity Features Among AI H9 Groups and Clades

In depth analysis of charged residues distribution, their variation in number and position, and isocontours in hemagglutinin, further confirmed that variation especially concerns its RBD sub-region and suggested electrostatic fingerprints do relate to different H9 groups. For Readers' convenience, Table 1 compares most recently published HA mature chain numberings for H1, H3, H5, H7 and H9 (Burke and Smith, 2014), and highlights group and sub-group specific variation within RBD. In particular, group-associated 'charge redistribution' is observed at RBD positions 135, 146 and 162 (H9 numbering): the net charge for these positions is zero in all groups (being the sum of two opposite charges and a neutral residue). However, the charge distribution pattern shared by 'wild bird' groups A+D+E is neutral-positive-negative, whereas it is negative-neutral-positive in 'poultry' groups B+C. RBD compensatory mutations seem to keep group-specific fingerprints and net charge, while progressively 'sliding' charges over RBD sites in the viral population. For space constrains, the in-depth analysis of variation at each position is presented in Supplementary file 1.

Table 1. Variation in the distribution of negatively (red), positively (blue) charged and hydrophobic (yellow) residues among H9N2 RBDs.

HA subtype numbering		HA (mature chain) position number										
H9Nx		131	135	146	161	162	165	180	186	198	216	217
H7Nx		127	134	145	162	163	166	181	187	199	217	218
H5Nx		133	141	152	167	168	171	186	192	204	222	223
H3Nx		137	145	156	171	172	175	190	196	208	226	227
H1Nx		134	142	153	168	169	172	187	193	205	223	224
Clade	Strains	Fully / most conserved amino acid for each clade										
A.1	4	R	N	H	N	E	N	E	K	D	Q	Q
A.2	3	K	N	H	T	E	N	E	K	D	Q	Q
A.3	12	K	N	H	N	E	N	E	K	D	Q	Q
A.4	4	K	N	H	N	E	N	E	K	D	Q	Q
A.5.1	6	A	N	H	N	E	N	E	K	D	Q	Q
A.5.2	2	A	N	Q	N	N	S	E	K	D	Q	Q
A.5.3	24	K	N	H	N	E	N	E	K	D	Q	Q
A.5.4	7	R	N	H	N	E	N	E	E	D	Q	Q
A.5.5	9	K	G	H	D	W	N	E	K	D	Q	Q
D	3	K	N	H	N	E	S	E	K	D	Q	Q
E	5	K	N	H	N	E	S	E	K	D	Q	Q
B.1.1	17	K	D	Q	N	R	S	A	I	D	L	Q
B.1.2	18	K	D	Q	N	R	S	A	I	D	L	Q
B.3	7	R	G	Q	N	R	S	E	I	D	L	Q
B.4	10	K	D	Q	N	R	S	A	T	D	L	Q
B.2.1	9	K	D	Q	N	R	D	A	T	N	L	I
B.2.2	17	K	D	Q	N	R	D	A	T	N	L	I
B.2.3	4	K	N	Q	N	R	D	A	T	N	L	I
B.2.4	6	K	D	Q	N	Q	D	A	T	N	L	I
B.2.5	9	K	D	Q	N	R	D	A	T	N	L	I
B.2.6	14	K	E	Q	N	R	D	T	T	N	L	I
B.2.7	20	K	D	Q	N	R	D	A	T	N	L	I
C.1	8	N	N	Q	N	R	S	V	T	D	L	T
C.2	54	K	D	Q	N	R	N	T	T	D	Q	Q
C.2.1	12	K	D	Q	N	R	N	A	T	D	L	Q
C.2.2	67	K	D	Q	N	Q	N	T	T	D	L	Q
C.2.3	3	K	D	Q	N	R	N	V	T	D	L	Q

Altogether, counterbalancing mutations observed at 131-135 (C.1), 135-180 (B.3), 146-162 (A.5.2) and 161-162 (A.5.5) seem to support the compensatory mechanism keeping the overall net charge while sliding charges over the RBD itself, i.e. for keeping group specific landmarks along with contemporary creation of novel fingerprints. In all H9 groups, the net charge at 180-186 is zero (except for clade B.3 where A180E is indeed compensatory for D135G). In A+D+E viruses, neutral charge results from the sum of opposites (+1 -1 = 0), while depending in B+C viruses on replacement of both charges by neutral residues (0 + 0 = 0). This way, the net charge is kept based on decreased percentage of charged residues. Intriguing variation also concerns 216 and 217: at 216, A+D+E clades share a highly conserved (>99% strains) polar residue (Gln), while most (86%) of B+C viruses show polar to hydrophobic Q216L transition. Instead, Gln is still the major residue in clades B.1.2 and C.2. At 217, sub-group variation is observed: only B.2.x viruses share a hydrophobic residue (Ile), while Gln is common to all other A+B+C+D+E clades. Such sub-group specific variation is not limited to hydrophobic patches, as 'charge sliding' also occurs between 165 and 198: H9 clades are negatively charged at 198, except for B.2.x ones, showing a polar residue (mostly, Asn). Such 'denegativization' is however compensated in B.2.x by an equally peculiar acquisition of a negative charge at 165, where Asp is 100% conserved.

3.4 Residues Involved in Changes at the H9N2 RBD are Surface Exposed

The RBD from the solved structure of the H9 HA was viewed to highlight the nine positions involved in group or sub-group specific variation. In figure 1A, the RBD surface is grey and antigenic sub-regions mediating SA binding (130-loop, 190-helix and 220-loop) are highlighted in yellow. The isopotential contours of the viral strains A.1_AtkCA66 and C.2.2_AquSh01, well representative for electrostatic fingerprints from 'wild bird' A+D+E viruses and 'poultry' B+C ones, are depicted in figure 1B. Residues 135, 146 and 162 involved in group specific 'charge redistribution' are surface exposed (orange); in particular, 146 is close to 190-helix and 135 is part of 130-loop. Positions 180-186 (mediating the 'charge loss' observed in the A+D+E to B+C transition) are surface exposed as well (purple) and part of 190-helix. The four positions involved in group and sub-group variation are also surface exposed (green), 216 and 217 are part of 220-loop, while 165 and 198 protrude at the other RBD 'side'. Finally, 131 and 161 involved in compensatory variation (Table 1) were also confirmed as surface exposed (not shown). Therefore, a deeper view of surface variation among H9 groups, sub-groups and clades could be obtained via electrostatic analysis of RBD models refined for side chains orientation. Both strain A.1_AtkCA66 and C.2.2_AquSh01 match in all positions patterns (Table 1) typical for A+D+E or B+C, respectively. At 162, they clearly show opposite charges; at 135, the expected contours are found again, as A.1_AtkCA66 shows no charge while in C.2.2_AquSh01 a negative protrusion is found in the corresponding area. Comparative analysis at 180-186 shows that the loss of both charged residues in the A+D+E to B+C transition does not result in 'neutralization' of the corresponding surface area. Instead of just an expected negative (Glu)-to-. neutral (Thr) shift, a seeming 'positivization' (increased blue area) is observed at 180 in C.2.2_AquSh01, depending on enlargement of neighboring electrostatic contours. For completeness of information, the isopotential contours of the RBDs from all representative strains used for creating heatmaps are presented in Supplementary file 1.

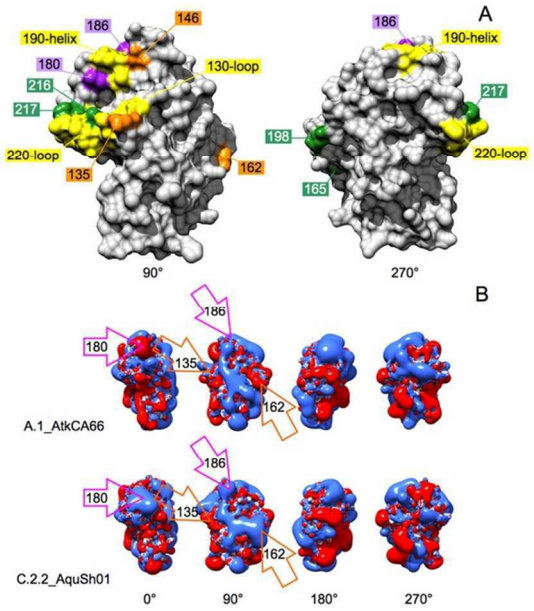


Fig. 1. Relevant surface epitopes and electrostatic variation at the RBD sub-region of H9N2 hemagglutinin

4 Concluding Remarks

'Classic' studies on AI virus evolution are based on antigenic and phylogenetic analyses (Stanekova and Vareckova, 2010). Recently, fingerprints for H5N1 evolution and spreading were obtained via deep analysis of the HA RBD surface: when comparing RBD structures, H5 was found to be quite closer to H9 than H3 and H7 and, in electrostatics, even closer to H9 (from a different phylogenetic group) than to H2 (same group) (Righetto et al., 2014). This prompted us to investigate whether similar mechanisms might underlie H5 and H9 evolution and spreading. Indeed, this work indicates for the H9 HA gene a remarkable similarity to the H5 in circulation and evolution and notable difference from the typical evolution of H3. AI H9 strains show extended branches as these viruses continue to co-circulate in different regions and host species and this allows the clades to further evolve and differentiate. Such H5-H9 structural agreement suggests RBD 'charge redistribution' as a general landmark for AI virus evolution and spreading. Intriguingly, most of changes observed in H9 RBD occur at sub-regions crucial to SA binding and host specificity and to immune escape/antigenic drift (Kobayashi et al., 2012): 130-loop (129-132), 190-helix (180-188) and 220-loop (211-218). Intriguingly, positions 162 and 217 are involved in immune escape (Peacock et al., 2016) and in group specific charge redistribution: 'denegativization' at 162 in two clades from group A, is compensated by either 'depositivization' at 146 (in A.5.2) or 'negativization' at 161 (in A.5.5). Sub-group specific, polar to hydrophobic transition occurs instead at 217, likely involved (as 220-loop) in increased binding to α2-6 SA and thus in improved affinity to the human host. Residue 224 in H1N1 (corresponding to 217 in H9) mediates hydrogen bond interactions with α2,6-SA (Chutinimitkul et al., 2010); the involvement of H9 217 in host range modulation is also based on studies on H5N1 (Gambaryan et al., 2006). H3 227 (corr. to H9 217) is located between H3 226 and 228, playing a key role (as 220-loop) in host range restriction (Vines et al.,

1998). In this work, group-specific variation in H9 216 (corr. to H3 226) is also observed. In H9 180-186 (190-helix) the conserved dual opposite charges pair in groups A+D+E shifts to a non-charged pair in B+C. Contemporary loss of the two opposite charges is somehow 'compensatory' for the original RBD net charge. It is noteworthy that in AI viruses, mutations increasing or decreasing the charge of the SA binding RBD region can modulate binding avidity and affinity, and thus frequently observed charge counterbalancing is likely to compensate for gain and loss effects hence keeping the HA-NA charge balance (Kobayashi et al., 2012). However, in A+D+E to B+C group transition, compensation only keeps the net charge, while the overall loss of two charged residues from the RBD in B+C viruses occurs. This in turn is likely to favor immune escape, because of both the location of the two residues and the well-known role of charged amino acids in modulating protein antigenicity and immunogenicity. In fact, it is well known that both positively and negatively charged residues improve the antigenic recognition (up to several folds, depending on their number in the antigenic site) by creating further salt bridges with the recognizing Ab complementary surface (Farber et al., 2007). Charge variation also occurs at 146 (involved with 135 and 162 in charge redistribution), which is exposed at the RBD surface close to 190-helix (Table 1). Therefore, changes like the observed 'depositivization' at 146 might influence SA binding affinity and specificity. Considering that chickens possess both α-2′3′ and α-2′6′ SA receptors, it is tempting to speculate that such changes could be linked to host adaptation and species specificity (Perez et al., 2003). Sequence comparison was able to infer group and sub-group specific fingerprints presented in Table 1 as sequence patterns; however, when complementing sequence analysis by the structural approach, a real-estate picture of the system is available. Comparison of the electrostatic isocontours unveiled combined mutations resulting in modulation of relevant surface features by altering local equilibrium in surrounding areas (salt bridges or repulsions, hydrophobicity changes, decreased or increased charge density etc.). Even though further work is needed to fully clarify the complex network of equilibria that can be altered by specific mutations, evidence from this study supports the integration of up-to-date phylogenetic analyses with sequence-based and structural investigation of surface features as a front-end strategy for inferring trends and relevant mechanisms in influenza virus evolution.

Acknowledgements

We thank Adelaide Milani, Alice Fusaro and Isabella Monne for unpublished phylogenetic trees and sequence analysis, Giovanni Cattoli and Alessandra Piccirillo for useful discussions. This work was performed in the framework of the Doctoral Schools of 'Veterinary Science' (AH) and of 'Biosciences and Biotechnologies' (IR) at the affiliation Departments of the University of Padua. In addition, we gratefully acknowledge the contributing authors and the originating and submitting laboratories for the sequences from the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database.

Funding

This work was supported by 'DOR' local fund from the University of Padua to FF; AH fellowship was supported by the NoFlu project, Fondazione Cariplo Vaccine Program (grant number 2009-3594).

Conflict of Interest: none declared.

References

Alexander, D.J. (2007) An overview of the epidemiology of avian influenza. *Vaccine* 25(30):5637-5644.

Al-Tawfiq, J.A., *et al*. (2014) Surveillance for emerging respiratory viruses. *Lancet Infect Dis* 14(10):992-1000.

Benkert, P., *et al*. (2009) QMEAN server for protein model quality estimation. *Nucleic Acids Res* 37(Web Server issue):W510-514.

Bordoli, L., *et al.* (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nat Protoc* 4(1):1-13.

Burke, D.F. and Smith, D.J. (2014) A recommended numbering scheme for influenza A HA subtypes. *PLoS One* 9(11):e112302.

Butt, K.M., *et al.* (2005) Human infection with an avian H9N2 influenza A virus in Hong Kong in 2003. *J Clin Microbiol* 43(11):5760-5767.

Carugo, O. and Pongor, S. (2001) A normalized root mean square distance for comparing protein three dimensional structures. *Protein Sci* 10:1470-1473.

Chutinimitkul, S., *et al.* (2010) Virulence-associated substitution D222G in the hemagglutinin of 2009 pandemic influenza A(H1N1) virus affects receptor binding. *J Virol* 84(22):11802-11813.

Farber, D.L. *et al*. (2007) Immune response: Antigens, Lymphocytes and Accessory Cells. in: Medical Immunology, Sixth Edition; Chapter 4:35-54.

Gambaryan, A., *et al.* (2006) Evolution of the receptor binding phenotype of influenza A (H5) viruses. *Virology* 344(2):432-438.

Gasteiger, E., *et al.* (2005) Protein Identification and Analysis Tools on the ExPASy Server. in: The Proteomics Protocols Handbook Edited by Walker JM. Humana Press:571-607.

Guan, Y. and Smith, G.J. (2013) The emergence and diversification of panzootic H5N1 influenza viruses. *Virus Res* 178(1):35-43.

Hu, M., *et al.* (2015) Coexistence of Avian Influenza Virus H10 and H9 Subtypes among Chickens in Live Poultry Markets during an Outbreak of Infection with a Novel H10N8 Virus in Humans in Nanchang, China. *Jpn J Infect Dis* 68(5):364-369.

Kobayashi, Y. and Suzuki, Y. (2012) Compensatory evolution of net-charge in influenza A virus hemagglutinin. *PLoS One* 7(7):e40422.

Krivov G.G. *et al* (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77(4):778-95.

Lin, Y.P., *et al.* (2000) Avian-to-human transmission of H9N2 subtype influenza A viruses: relationship between H9N2 and H5N1 human isolates. *Proc Natl Acad Sci U S A* 97(17):9654-8.

Nelson, M.I. and Vincent, A.L. Reverse zoonosis of influenza to swine: new perspectives on the human-animal interface. *Trends Microbiol* 23(3):142-53.

Peacock, T., *et al.* (2016) Antigenic mapping of an H9N2 avian influenza virus reveals two discrete antigenic sites and a novel mechanism of immune escape. *Sci Rep* 6:18745.

Perez, D.R., *et al.* (2003) Role of quail in the interspecies transmission of H9 influenza A viruses: molecular changes on HA that correspond to adaptation from ducks to chickens. *J Virol* 77(5):3148-3156.

Pettersen, E.F., *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605-1612.

Righetto, I., *et al.* (2014) Comparative structural analysis of haemagglutinin proteins from type A influenza viruses: conserved and variable features. *BMC Bioinformatics* 15:363.

Sitkoff, D., *et al.* (1994) Accurate calculation of hydration free energies using macroscopic solvent models. *J Phys Chem* 98:1978-1988.

Stanekova, Z. and Vareckova, E. (2010) Conserved epitopes of influenza A virus inducing protective immunity and their prospects for universal vaccine development. *Virol J* 7:351.

Su, S., *et al.* (2015) Epidemiology, Evolution, and Recent Outbreaks of Avian Influenza Virus in China. *J Virol* 89(17):8671-8676.

Trombetta, C., *et al.* Emerging Influenza Strains in the Last Two Decades: A Threat of a New Pandemic? *Vaccines (Basel)* 2015;3(1):172-185.

Velkov, T., *et al.* (2013) The antigenic architecture of the hemagglutinin of influenza H5N1 viruses. *Mol Immunol* 56(4):705-719.

Vines, A., *et al.* (1998) The role of influenza A virus hemagglutinin residues 226 and 228 in receptor specificity and host range restriction. *J Virol* 72(9):7626-7631.

Heidari, A. *et al*., Supplementary file 1

Phylogenetic methods

HA gene nucleotide sequences of H9N2 subtype were retrieved from the NCBI and GISAID (Global Initiative on Sharing Avian Influenza Data) EpiFlu database (http://www.gisaid.org). Nucleotide sequences of at least 1500 bp length were selected. Multiple sequence alignment of HA sequences was performed with MAFFT version 7 (http://mafft.cbrc.jp/alignment/server). Redudant isolates with 100% sequence similarity (i.e., redundant sequences) were identified and removed, giving a final HA dataset and alignment of 1669 sequences that was subjected to phylogenetic trees reconstruction. The neighbor-joining (NJ), maximum-likelihood (ML), and Bayesian methods were used to construct three different phylogenetic trees for comparison. Analysis of the best-fit substitution model was performed using MEGA5 (Tamura et al., 2011), and the goodness-of-fit of each model was measured by Bayesian Information Criterion and corrected Akaine Information Criterion (AICc). The General Time Reversible (GTR) model with a discrete gamma distribution (+Γ) allowing for invariant sites (+I) was selected based on AICc and used in all data analyses. MEGA5 was also used to perform phylogenetic analysis and the evolutionary history was inferred by both NJ and ML methods (Tamura and Kumar, 2002), with standard errors being calculated based on 1000 bootstrap replicates. Furthermore, PhyML (version 2.4.4) (Guindon et al., 2003) was used to create ML trees was used to create ML trees. The GTR + Γ + I model of nucleotide substitution was used for the analysis, with an estimated gamma shape parameter. Robustness of the groups was assessed using the bootstrap approach with 100 replicatest. Bayesian phylogenetic tree was inferred using MrBayes software (Ronquist and Huelsenbeck, 2003) and applied to generate the dendrograms as well as to assess statistical supports for the branches from the trees generated by the original dataset. For ease of display, and also to ensure that the clade topology would be maintained when fewer isolates are used, a small representative dataset of 360 H9N2 HA sequences was created and analyzed by the same aforementioned phylogenetic models (seed tree in this work). Phylogenetic trees were visualized using FigTree version 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

The largest HA gene dataset alignment (n = 1669; length ≥ 1500 bp) used for the phylogenetic reconstruction, was also used to infer evolutionary distances (within and between groups) by pair-wise analysis. The number of base substitutions per site was calculated by two different methods. The simplest one (uncorrected pairwise distance) was performed by averaging all sequence pairs between groups, while the second method followed the Maximum Composite Likelihood model. Variation rate among sites was modelled with a Γ distribution value = 9.4 (calculated by preliminary estimation from our dataset) and the differences in the composition bias among sequences were considered in the evolutionary comparisons. The C-value ratio used in the H9N2 clades partitioning - i.e. the ratio of the average pairwise distance between a particular taxon and its closest neighboring group divided by the average pairwise distance within that selected clade - was used to confirm the clades partitioning.

Phylogenetic clustering of AI H9N2 HA

We used the genetic correlation to follow objective criteria able to properly sort strains based on the phylogenetic topology and on specific evolutionary distances that reflect the diversity of the AI H9N2 subtype. Independently on algorithm used, i.e. NJ, ML or Bayesian, H9N2 viruses sorted into five different monophyletic groups, referred to as A, B, C, D and E; their clades - identified by numbers - are separated based on inter-clade average distance ≥5% and intra-clade average distance <5%; separation for each identified clade is confirmed by C-value ≥ 1. Groups and clades were assigned when at least three isolates with different epidemiological history formed a distinct taxonomic group with bootstrap value at the defining node ≥60%. Clades separation based on distance value cut off was confirmed using two different calculation algorithms as described in the methods section above. Phylogenetic groups A to E, as well as all clades but C.2.3, were recently confirmed using a further and larger HA gene dataset alignment (n = 2813; length ≥ 1500 bp; personal communication of unpublished results from Adelaide Milani, Alice Fusaro and Isabella Monne). Much further work would be needed to provide a comprehensive 'classification' for H9N2, which however is not the aim of this work.

Suplementary references

Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696-704.

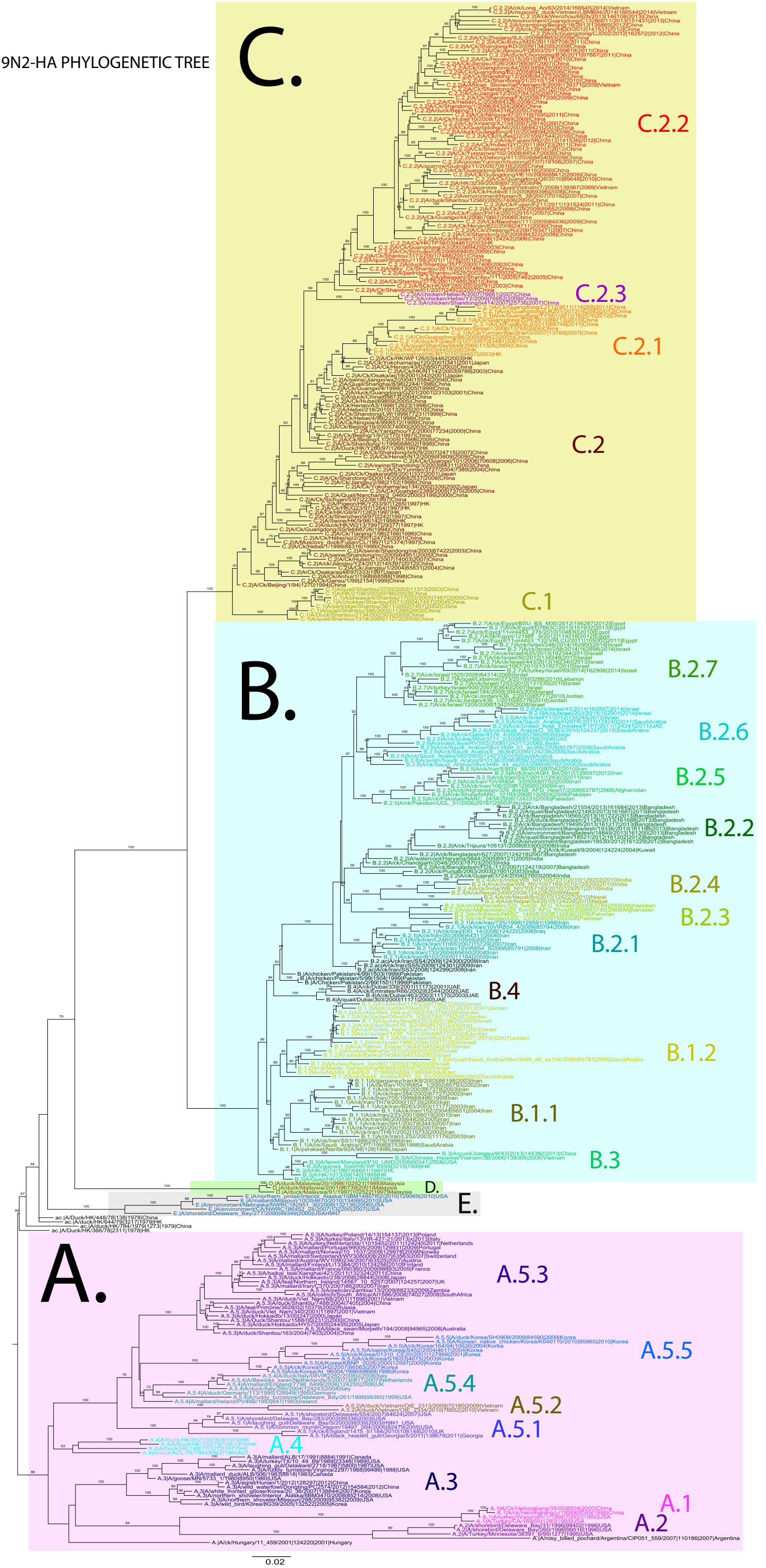
Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-4.

Tamura, K. and Kumar, S. (2002) Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol* 19(10):1727-36.

Tamura, K. *et al*. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731-9.

Clustering AI H9 viruses by Genetic Distances: Phylogenetic (seed) tree

H9N2-HA PHYLOGENETIC TREE

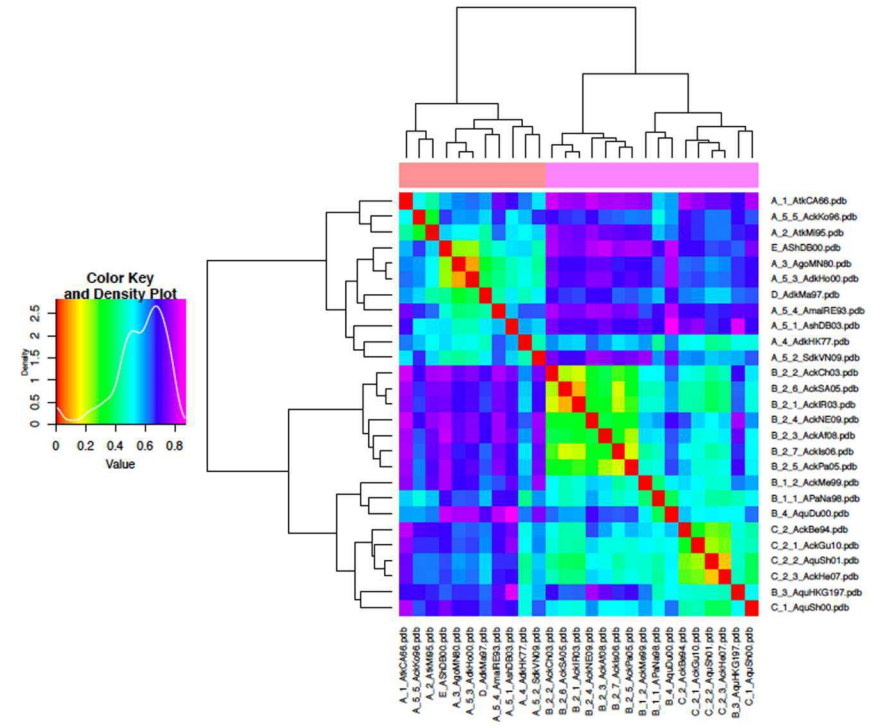


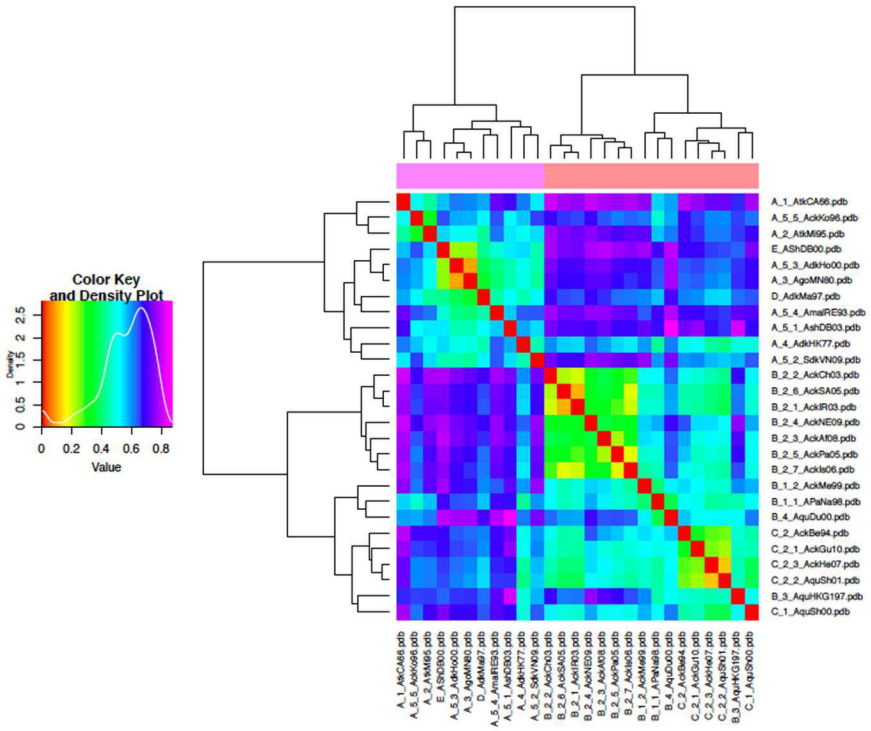
ABOVE: Maximum-likelihood short alignment tree with 360 H9 isolates constructed by PhyML. The different classes and clades are color coded. Estimates of the statistical significance of phylogenies were calculated by performing 100 bootstrap replicates. Numbers in the tree nodes represent the bootstrap support (≥60). **BELOW: 'Representative' viruses (from each group and clade) used for electrostatic analyses.**

Group	Clade	Countries	Hosts	Full name	Short name	NCBI AC
A	A.1	USA, China	Avian	A/turkey/CA/189/66	A.1_AtkCA66	AAD49000
	A.2	USA	Avian	A/turkey/Minnesota/38391-6/95	A.2_AtkMi95	AAD48997
	A.3	China, Korea, USA, Canada	Avian	A/goose/MN/5733-1/1980	A.3_AgoMN80	ABB88390
	A.4	Hong Kong, New Zealand	Avian	A/Duck/HK/168/77|1272|1977	A.4_AdkHK77	AF156382.1
	A.5.1	USA, Georgia, UK	Avian	A/shorebird/Delaware Bay/283/2003	A.5.1_AshDB03	AET77176
	A.5.2	Vietnam	Avian	A/duck/Vietnam/OIE_2313/2009|70180|2009	A.5.2_AdkVN09	AB639356.1
	A.5.3	Japan, Vietnam, Iran, Australia, China, UK, Russia, Netherlands, Italy, France, Finland, Austria, Switzerland, Norway, Portugal, South Africa, Zambia	Avian	A/duck/Hokkaido/13/00	A.5.3_AdkHo00	AAQ97383
	A.5.4	Ireland, UK, Italy, Netherlands, Germany, USA	Avian	A/mallard/Ireland/PV46B/1993	A.5.4_AmaIRE93	AB303077
	A.5.5	Korea	Avian, Swine	A/chicken/Korea/AI-96004/1996	A.5.5_AckKo96	ACZ48629
D	D	Malaysia	Avian	A/duck/Malaysia/91/1997	D_AdkMa97	AEY75592.1
E	E	USA	Avian, environ.	A/shorebird/Delaware_Bay/277/2000	E_AshDB00	AET77024
B	B.1.1	Saudi Arabia, Japan, Iran	Avian	A/parakeet/Narita/92A/98	B.1.1_APaNa98	AB049160
	B.1.2	Lebanon, Israel, Jordan, UAE, Iraq, Saudi Arabia	Avian	A/chicken/Middle East/ED-1/1999	B.1.2_AckME99	GU053201
	B.2.1	Iran, Iraq	Avian	A/chicken/Iran/L248/2003	B.2.1_AckIR03	EF063514
	B.2.2	India, Bangladesh, Kuwait	Avian	A/chicken/Chandigarh/2048/2003	B.2.2_AckCh03	ADL64047
	B.2.3	Afghanistan, Pakistan, Iran	Avian	A/chicken/Afghanistan/329-6vir09-AFG-Khost9/2008	B.2.3_AckAf08	EPI_ISL_63785
	B.2.4	Nepal, India	Avian	A/chicken/Nepal/2490/2009	B.2.4_AckNE09	AFO83282
	B.2.5	Pakistan, Afghanistan, Iran	Avian	A/chicken/Pakistan/UDL-01/2005	B.2.5_AckPA05	ACP50642
	B.2.6	Saudi Arabia, UAE, Qatar, Israel, Libyan	Avian	A/chicken/Saudi Arabia/582/2005	B.2.6_AckSA05	AFO83289
	B.2.7	Israel, Egypt, Lebanon, Jordan	Avian	A/chicken/Israel/1525/2006	B.2.7_AckIS06	ACJ68774
	B.3	Hong Kong, USA, Vietnam	Avian, Human	A/Quail/Hong Kong/G1/97	B.3_AquHKG197	AF156378/AAF00706
	B.4	UAE	Avian	A/quail/Dubai/303/2000	B.4_AquDu00	EF063512/ABM21877
C	C.1	Hong Kong, China	Avian/Human	A/quail/Shantou/1318/2000	C.1_AquSh00	EF154910/ABM46230
	C.2	China, Hong Kong, Japan	Avian, Swine, Human, Environ.	A/chicken/Beijing/1/1994	C.2_AckBe94	KF188294/AGO17871
	C.2.1	China	Avian	A/Ck/Guangdong/ZDR/2010|	C.2.1_AckGu10	JF715016.1
	C.2.2	China	Avian, Environ.	A/quail/Shantou/1158/2001	C.2.2_AquSh01	EF154916.1
	C.2.3	China	Avian	A/chicken/A/Hebei/2007/76661/2007/China	C.2.3_AckHe07	GQ202056

Clustering AI H9 viruses by Electrostatic Distances: Heatmaps

This section contains supplementary details and figures about 'Clustering by Electrostatic Features' evidence presented in section 3.2 of the main text. Semiquantitative ED evaluation and clustering of the spatial distributions of the electrostatic potentials were obtained by WebPIPSA (Protein Interaction Property Similarity Analysis) (Richter *et al*., 2008). In heatmaps below, high ED (dark blue, violet or magenta colors, see density plots) clearly separates 'wild birds' classes A+D+E from 'poultry' B+C ones, whereas lower ED among either A, D, E or between B and C is highlighted by prevalence of the light blue color. No meaningful difference is observed when using either Hodgkin (top) or Carbo (bottom) index.





The only exception to the overall substantial agreement of electrostatic clustering to phylogenetic grouping is represented by clades B3 and B4 (both isolated from the same host bird -- quail), which are closer to C2 (light blue) than to B2 strains.

Supplementary Reference (Electrostatics)

Richter, S., *et al.* (2008) WebPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acid Res* 36(Web Server Issue):W276-W280

Variation in electrostatic and hydrophobicity features among AI H9 classes and clades

This section contains more details about charge variation presented in Table 1 and section 3.3 of the main text.

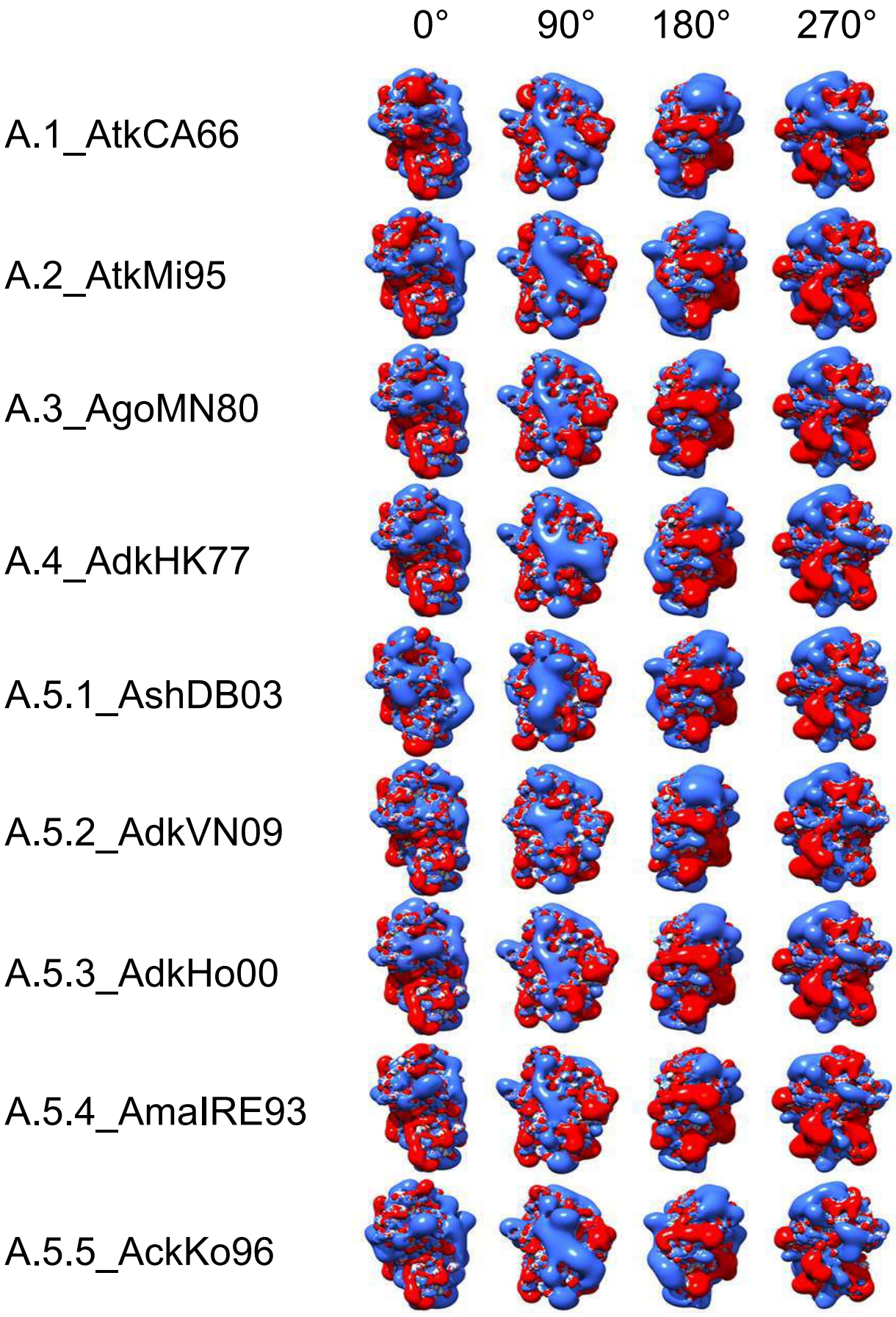
At position 135, almost all (94%) Class A viruses share a non charged residue with prevalence (72%) of Asn, which is 100% conserved in classes D+E; mutation to charged residue (N135D) only concerns 6% viruses from clades A.5.3, A.5.4 and A.5.5. Instead, negativization at position 135 is most often observed in both classes B (85%) and C (92%), with prevalence of Asp/Glu over other amino acids in almost all B+C clades. A compensatory mechanism is observed for exceptions, i.e. for clades not sharing a negative charge at 135. For example, clade C.1 lacks the negative charge of classes B+C and shares instead (100% sampled viruses) N135 with classes A+D+E; however, this is compensated as C.1 is also the only B+C clade missing a positive charge at 131. Similarly, B.3 (showing prevalence of Gl35) is also the only B+C clade with a negative charge (Glu) replacing a non charged residue at 180. Residue 146 is His in 95% class A viruses and in all D+E strains, and Gln in almost all B (>99%) and C (98%) clades. The only class A exception is clade A.5.2, showing Q146 (like B+C) instead of H146 (common to A+D+E). However, once again a counterbalancing event is observed: depositivization at position 146 of A.5.2 is compensated by peculiar denegativization at 162 (E162N).

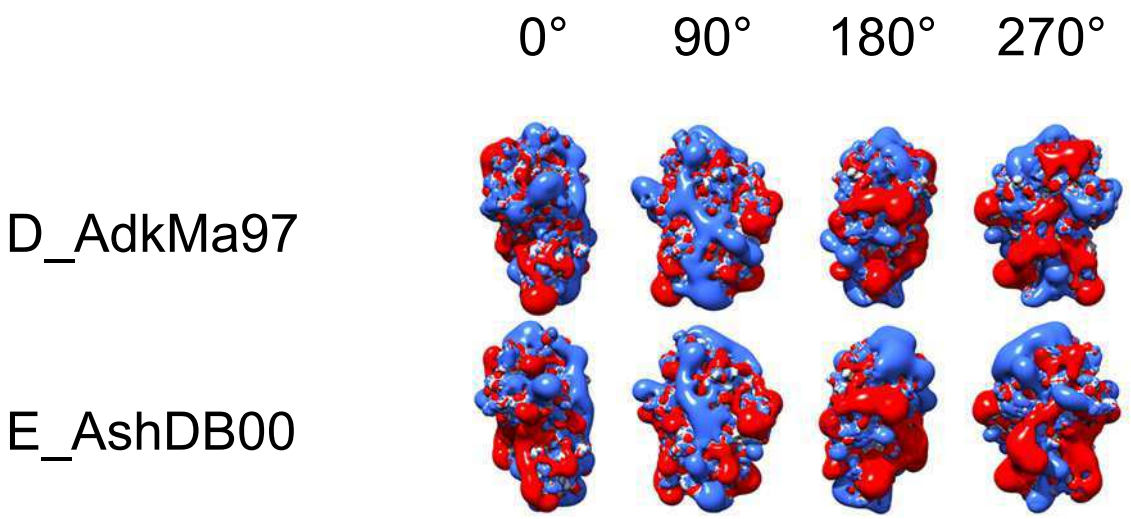
Loss of negative charge at E162 (otherwise shared by A+D+E groups) is also shown by clade A.5.5 (E162W in 100% viruses); however intriguingly, the lost negative charge is rescued at the contiguous amino acid position by the equally conserved (100%) and peculiar mutation N161D. A negative charge at position 162 (or 161) is thus a A+D+E groups landmark. In B+C groups, major residues at 162 are Arg and Gln, with prevalence of the former over the latter in all clades but B.2.4, where reverse prevalence is observed. Therefore, ongoing positivization of position 162 seems to be a landmark as well for viruses circulating in poultry.

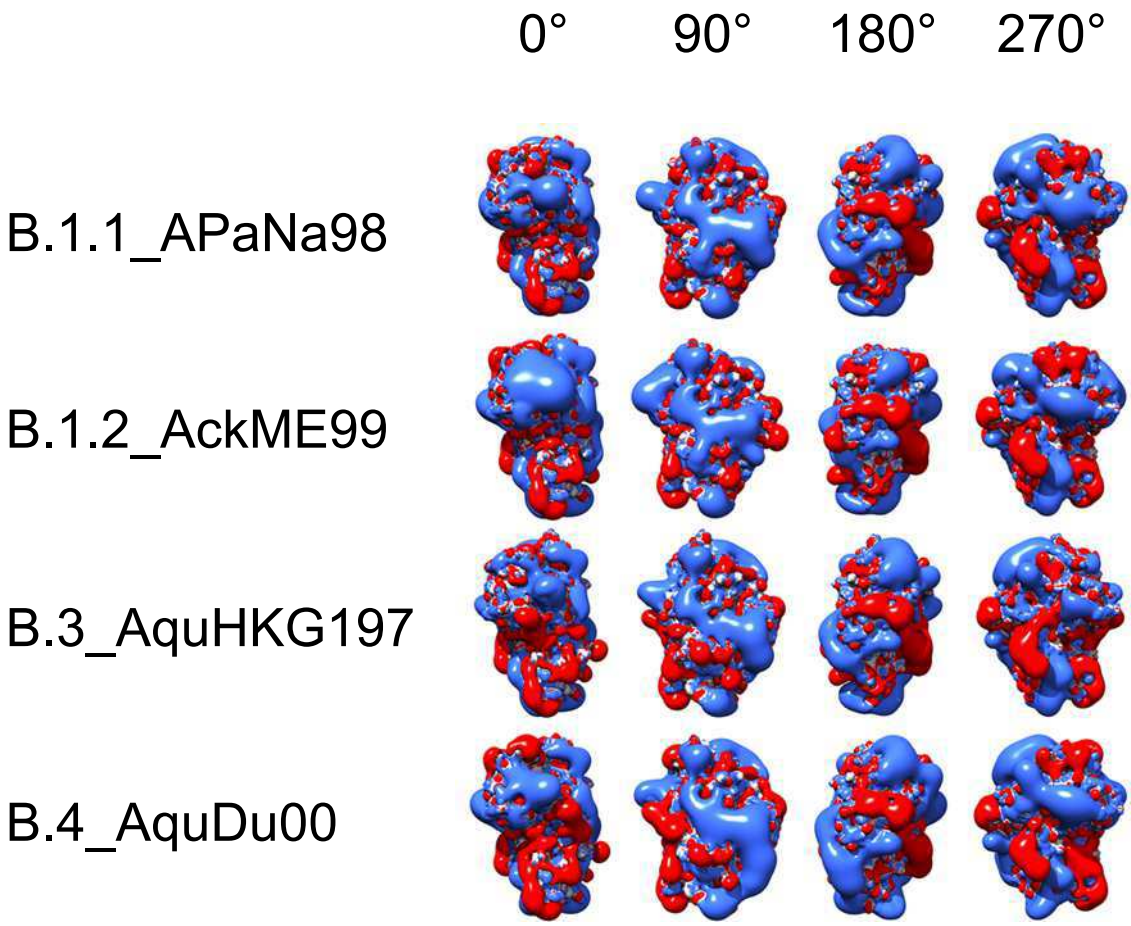
Deep inspection at position 217 shows meaningful difference with respect to 216. In addition to A+D+E groups (>99% strains), the 'original' Gln is highly conserved also in group C (82% strains) and in C.1 the major residue is anyway polar (Thr, in 97% strains). Instead, Gln is 100% conserved in clades B.1.1, B.1.2 and B.3, whereas polar to hydrophobic transition is ongoing in clade B.4 (Gln however being still the major residue) and fully fixed (100%) in the whole B.2.x sub-group (sharing Ile as major residue).

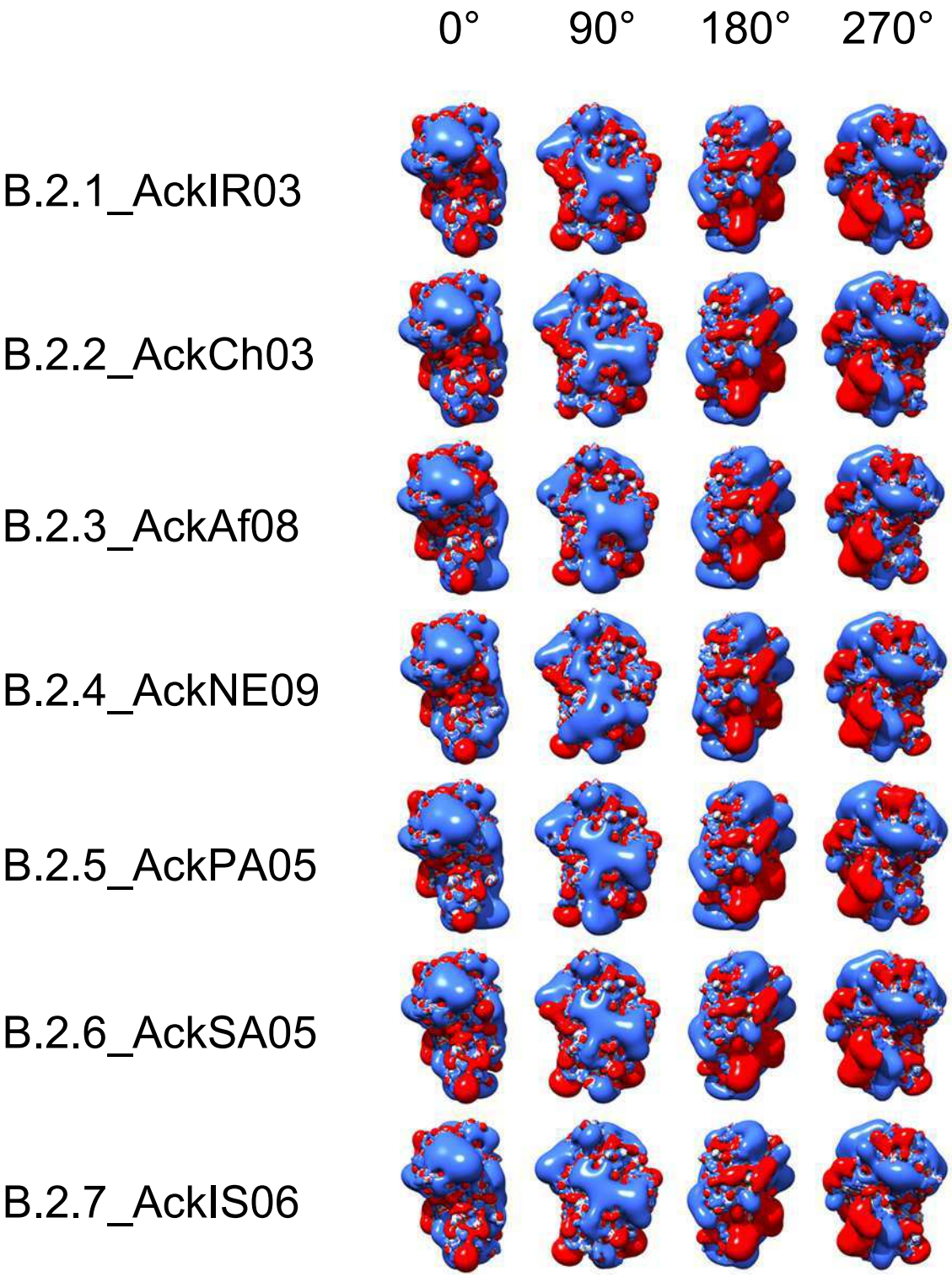
Isopotential contours of the RBDs from all representative H9N2 virus studied in this work

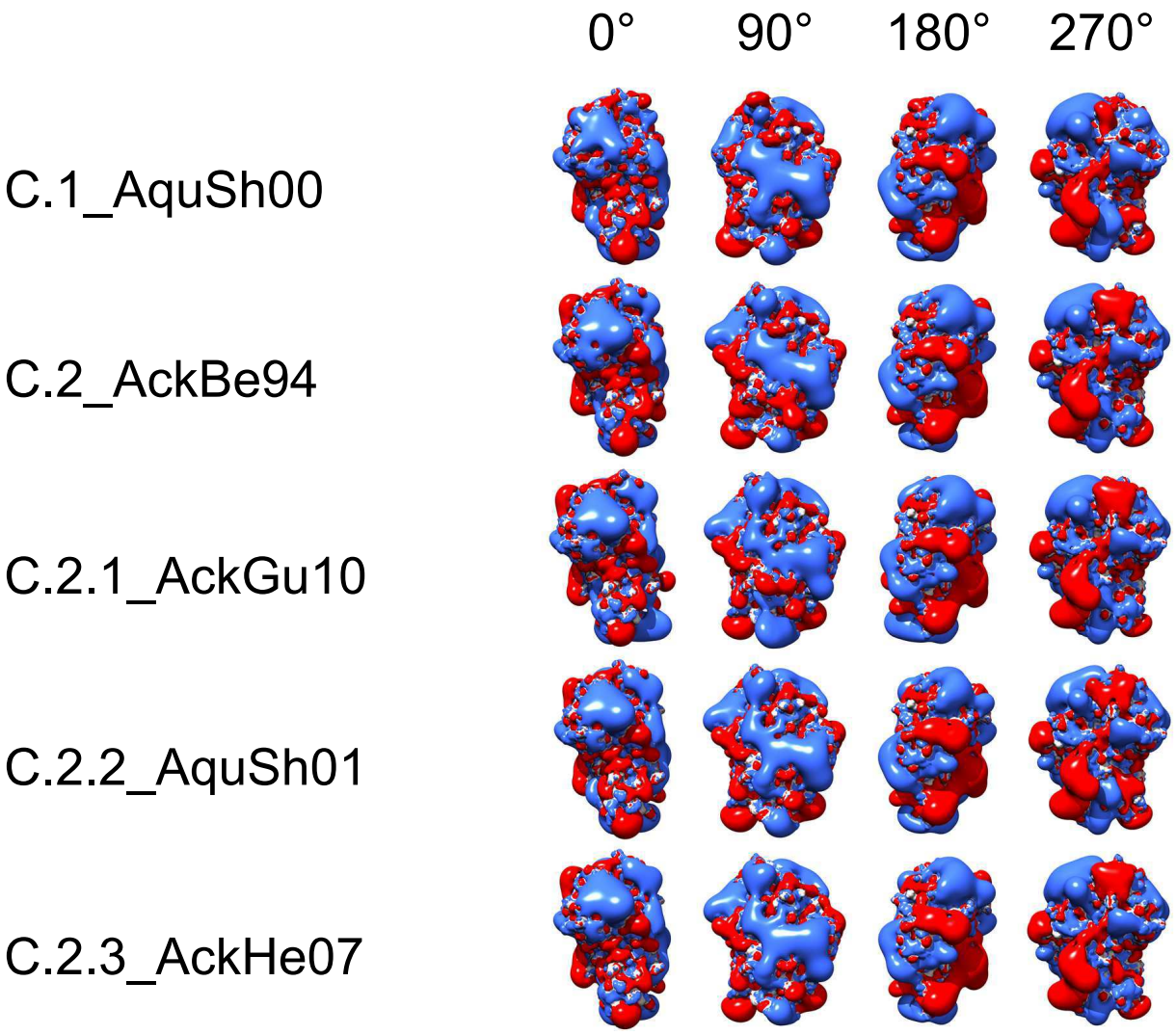
Four 90° stepwise rotation views are presented for each representative RBD electrostatic isocontour. Names of the H9N2 virus strains are the same as in table and heatmaps above.


0°
90°
180°
270°
A.1_AtkCA66
A.2_AtkMi95
A.3_AgoMN80
A.4_AdkHK77
A.5.1_AshDB03
A.5.2_AdkVN09
A.5.3_AdkHo00
A.5.4_AmaIRE93
A.5.5_AckKo96


0°
90°
180°
270°
D_AdkMa97
E_AshDB00


0°
90°
180°
270°
B.1.1_APaNa98
B.1.2_AckME99
B.3_AquHKG197
B.4_AquDu00


0°
90°
180°
270°
B.2.1_AckIR03
B.2.2_AckCh03
B.2.3_AckAf08
B.2.4_AckNE09
B.2.5_AckPA05
B.2.6_AckSA05
B.2.7_AckIS06


0°
90°
180°
270°
C.1_AquSh00
C.2_AckBe94
C.2.1_AckGu10
C.2.2_AquSh01
C.2.3_AckHe07

CHAPTER 2

Domain architecture variation in mammalian protein trafficking

State of the art

1. *SNARE PROTEINS*

In eukaryotic cells, several biological processes such as intracellular transport, neurotransmitter release, cell fertilization and viral infection require membrane fusion. Communication amongst cellular compartments is strongly linked to membrane fusion and SNARE (Soluble N- ethylmaleimide-sensitive factor attachment protein receptor) proteins play a pivotal role in this event, thought to be the principal machinery of the membrane fusion in the cell. Membrane fusion is energetically not favoured because of the electrostatic and hydration repulsive forces between the approaching bilayers. Another problem to manage is the lateral tension. SNARE proteins help to overcome energetic barrier, via the formation of a *trans*-SNARE complex; these proteins are involved even in endo- and exocytosis. However, the action mechanism of SNAREs during exocytosis stages is still unclear (Han et al., 2017). SNARE proteins are retrieved in both mammals and yeast cells for a total amount of more than 60 members. SNAREs can be divided into the large family proteins of VAMPs (Vesicle Associated Membrane Proteins), syntaxins and SNAP-25. Moreover, on the basis on their localization, SNAREs are also referred to as v-SNAREs (v= vesicles) and t-SNAREs (located at the target membrane). A feature common to SNAREs is the presence of a conserved stretch of ~70 residues, named SNARE motif, structured as a coiled coil domain (CCD), showing a hydrophobic heptad register, interrupted by a conserved polar residue at the ionic zero layer (Fig.23.). This zero layer is important in setting the right register for SNARE motifs assembling (Fasshauer *et al*., 1998; Kloepper *et al*., 2007). Membrane curvature can affect the secondary structure of the SNARE motif region (Liang *et al*., 2014). The so called SNARE complex is built up by several SNARE motifs initially assembled at the N-ter domain toward the C-ter domain (Sutton *et al*., 1998). The energy release produced by the formation of the SNARE complex, is then used to put the membranes in contact (Lu *et al*., 2008; Hernandez *et al*.,2012). Figure 22.A depicts the *trans*-SNARE complex, made up of a helix bundle consisting of four parallel SNARE motif helices (3 Q-SNAREs and 1 R-SNARE): one from Synaptobrevin (VAMP) and Syntaxin and two from SNAP-25 (Fasshauer *et al*., 1998). When the type of polar residue is considered, SNAREs can also be grouped in Q- or R-SNAREs (Fasshauer *et al*., 1998). Q-SNAREs can be further subdivided in Syntaxin, SNAP-25 N-ter CCD and SNAP-25 C-ter CCD, whereas R-SNAREs in short VAMPs (brevins) and long VAMPs (longins).

Longins share a conserved N-terminal domain named the Longin Domain (LD) (Fig.21.). This LD is characterized by having a conserved fold (De Franceschi *et al*., 2014) and it is able to regulate membrane fusion and subcellular localization (SCL) (Martinez-Arca *et al*., 2003).

The lab hosting me was able to investigate the functional differences amongst SNAREs through a bioinformatic approach: pattern and profile-based screenings with similarity-based analyses were helpful for discovering variation of residues at specific positions of the CCD, thought to be crucial to the assembly/disassembly of the fusion complex.

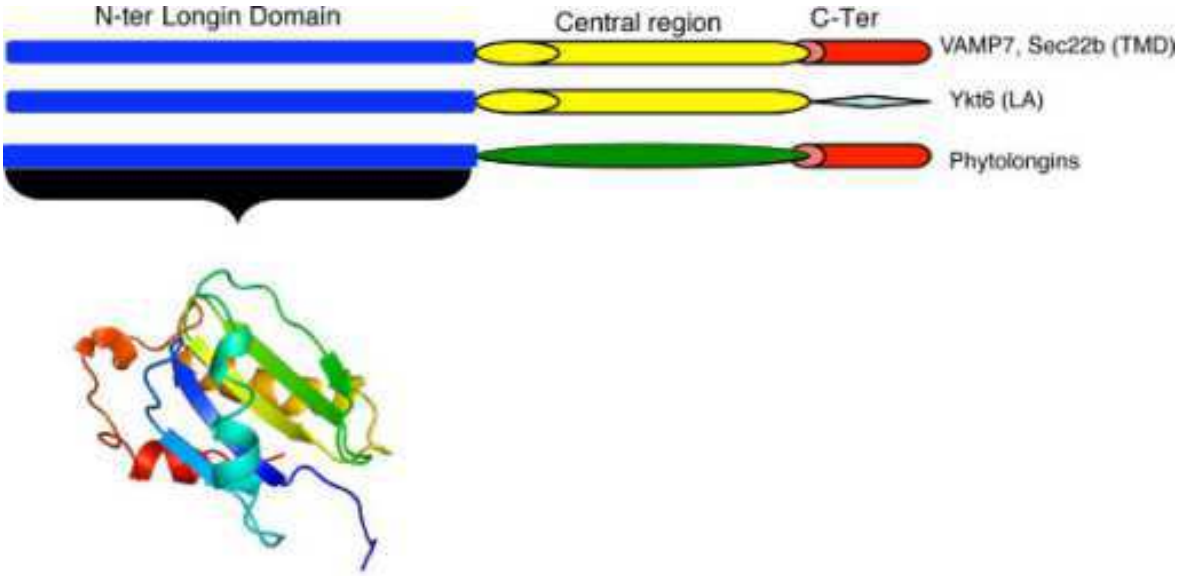


Fig.21. Domain architecture of longin proteins. SNARE motif correspond to the yellow central region, replaced by PhyL region (green) in Phytolongins. Except for Ykt6, other longin proteins possess a CTD region organised as a transmembrane domain (TMD) highlited in red. From: Vedovato M, Rossi V, Dacks JB, Filippini F. Comparative analysis of plant genomes allows the definition of the “Phytolongins”: a novel non-SNARE longin domain protein family. *BMC Genomics*. 2009

These analyses revealed an additional classification of R-SNARE subfamilies based on the conservation of residues at positions of the CCD other than the zero and the hydrophobic layer:

- *RD-SNAREs*: These SNARE proteins present an aspartic acid residue (D) at the position C-ter to the zero layer (R).
- *RG-SNAREs*: These SNARE proteins present a glycine residue (G) at the position C-ter to the zero layer (R).

Moreover, two pattern signatures with standard PROSITE syntax were developed in order to represent specific tags for RD- or RG-SNAREs (Fig.23.). Only a number of short VAMPs (brevins) belongs to RD-SNAREs, while RG-SNAREs group contains the whole longins subfamily, non-neuronal brevins, mammalian VAMP4, VAMP5 and VAMP8. Different VAMP subfamilies were analyzed for CCD sequence variation and these analyses identified

conserved couples of residues at layer -3 to -2, positively charged + hydrophobic in RD-brevins, hydrophobic + charged in VAMP7, polar and/or charged in RG-brevins and in Ykt-like longins. Moreover, nearby layer +6, three positions are different between RD- and RG-SNAREs. A Glu residue is conserved in all RD-SNAREs and it is replaced by a positively charged residue in most RG-brevins and longins or by Tyr in Ykt-like longins. RD-SNARE CCDs are common in Metazoa whereas RG-SNARE signature in yeast brevins, VAMP4, VAMP8 and longins. Moreover, v-SNAREs involved in the neurotransmitter releasing are all RD-SNAREs and this could be an evidence that these SNAREs evolved to specialize into the most rapid fusion reactions taking place in animals (Rossi *et al*., 2004).

Syntaxins carry the independently folded Habc domain at their N-ter, bound to the SNARE motif via a short linker; When the Habc domain folds back to the SNARE motif and this closed conformation is stabilized by the interaction with Munc 18-1, participation of syntaxin to the SNARE complex is not allowed and thus membrane fusion is inhibited (Furgason *et al*., 2009).

The *trans*-SNARE complex can switch to a *cis*-SNARE complex where proteins are fully folded in the same membrane (Stein *et al*., 2009) (Fig.24.). This configuration seems to make possible the formation and the expansion of the fusion pore (Han *et al*., 2017) as also revealed by coarse grained (CG) simulations (Risselada *et al*., 2011). The folding order used to build the SNARE complex is linked to different stages of synaptic vescicle fusion (Lu *et al*., 2008). The spontaneous folding at the N-ter region allows to juxtapose vesicles membranes. This first step is followed by a fast zipping toward the C-ter domain and finally by the fusion pore formation and expansion. In Ca^{2+} dependent neurotransmitter releasing the zippering is controlled by regulatory protein Synaptotagmin (Han *et al*., 2017). Studies from Martens *et al*., 2009; Hui *et al*., 2009; MacMahon *et al*., 2010 suggest a putative role of regulatory proteins in destabilizing membrane by increasing its local curvature.

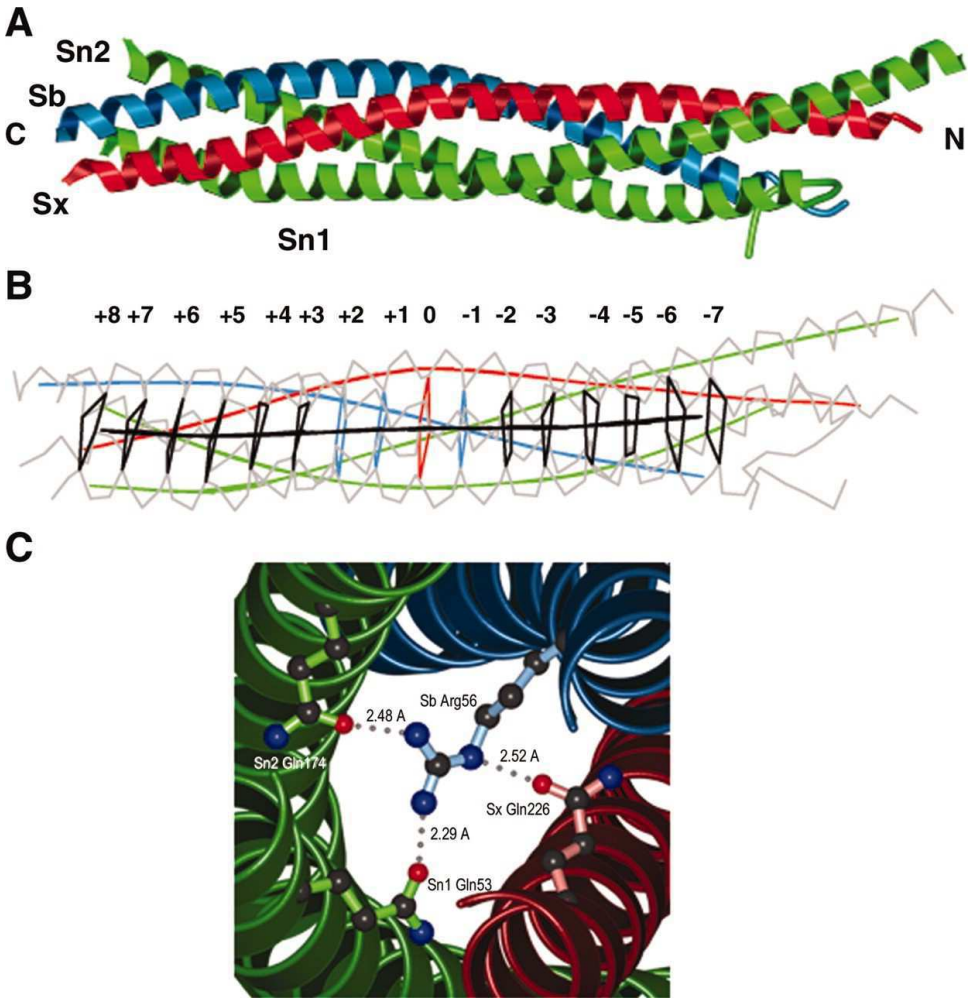


Fig.22. A. Structure of the core SNARE complex: Sb (VAMP), Sn1 and Sn2 (SNAP-25), Sx (Syntaxin); B. Layers organisation in the core SNARE complex; C. Configuration of the zero layer, completely conserved between different cell types and species. From: Joseph G. Duman, John G. Forte. What is the role of SNARE proteins in membrane fusion? *American Journal of Physiology - Cell Physiology* Aug 2003, 285 (2) C237-C249

```
RD-SNARE signature [LIVM]-x-[VTND]-[NTHA]-x(3)-[LIVT]-x(2)-[RK]-[DE]-[QSTVKA]-x-
[LIS]-x(2)-[LIVM]-x(3)-[ASTIN]-x(5)-[GQ]-x(3)-[FYMNS]-[EQ]-x(3)-[AGRS]

TP = 66, of which: 13 (+ 3 isoforms) from SwissProt(a) and 50 from
TrEMBL+TrEMBLnew(b) FP = 0; FN = 0; Precision = 100.00 %; Recall = 100.00 %

RG-SNARE signature [LIVMA]-x-[DENQSTKRAG]-[NTHAID]-x(3)-[LIVTMA]-x(2)-[RKV]-G-
[DEQTAV]-x-[LISV]-x(2)-[LIVM]-x(3)-[ASTIN]-x(2)-[LM]-x(2)- {CFHIKPRVW}-x(3)-
[FYMNS]-[NSKRHYAQ]-x(2)-[GASTN]- [QNKRTVFYS]

TP = 90, of which: 22 (+ 1 isoform) from SwissProt(a) and 67 from
TrEMBL+TrEMBLnew(b) FP = 0; FN = 2; Precision = 100.00 %; Recall = 97.82 %
```

Fig.23. RD- and RG-SNARE signatures in standard PROSITE syntax. True (TP) or false (FP) positive hits and false negatives (FN) concerns the scannings of databases released on 15-Dec-2003: (a) SwissProt 42.7 (141681 entries) and (b) TrEMBL+TrEMBLnew 25.7 (1078339 entries). Indexes of both “Precision” and “Recall” were calculated following PROSITE definition. From: Rossi, V., Picco, R., Vacca, M., D’Esposito, M., D’Urso, M., Galli, T. and Filippini, F. (2004), VAMP subfamilies identified by specific R-SNARE motifs. Biology of the Cell, 96: 251–256.

In both Synaptobrevin and Syntaxin, the SNARE motif is followed by a short linker region, the TMD and a topological domain, while the two SNARE motifs of SNAP-25 are attached to the plasma membrane by multiple palmitotyl tails (Han *et al*., 2017). The linker region plays a pivotal role in the fusion process as it is able to transduct the stress from the assembly of the cytosolic complex toward the membrane interface, trigging the membrane fusion; moreover, its positively charged residues can drive the transition from hemifusion to full fusion (Hernandez *et al*., 2012). The length and the folding of this linker region can affect the fusion efficiency: insertion of extra amino acids generally decreases the fusion efficiency (Van Komen *et al*., 2005; Deàck *et al*., 2006; Kesavan *et al*., 2007; Zhou *et al*., 2016). The degree of linker flexibility is important in the fusion process (Han *et al*., 2016).

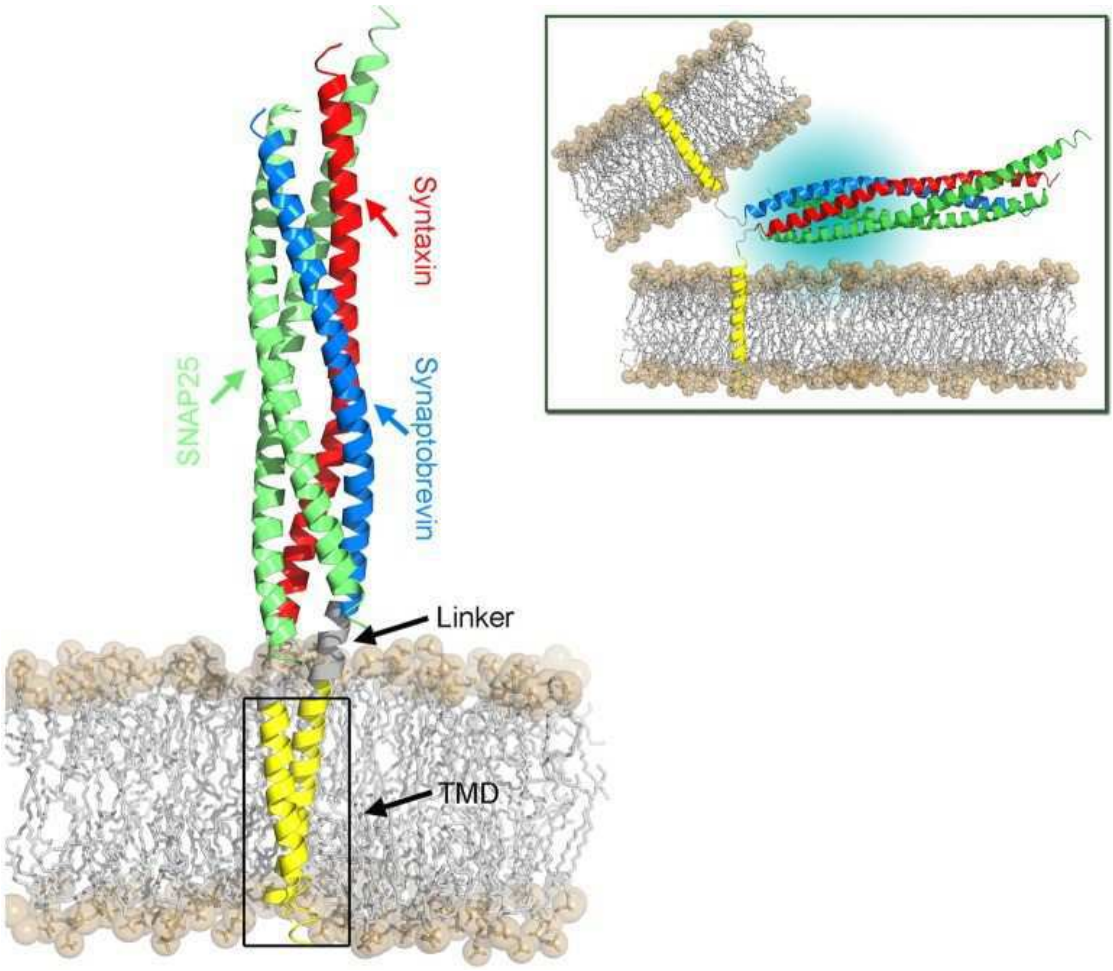


Fig.24. The *cis*-SNARE complex at the post-fusion stage. Syntaxin and Synaptobrevin are characterized by a SNARE motif, a short linker and a transmembrane domain (TMD). The SNARE complex in the pre-fusion stage is depicted in the top right box: the TMD of Syntaxin and Synaptobrevin are immersed in host membranes. From: Han J, Pluhackova K, Böckmann RA. The Multifaceted Role of SNARE Proteins in Membrane Fusion. *Frontiers in Physiology*. 2017;8:5.

The neuronal SNARE complex is SDS-resistant and has been solved by X-ray crystallography (Sutton *et al*., 1998; Han *et al*., 2017). Membrane fusion process in endo- or exocytosis can be explained in several steps as depicted in Fig. 25.

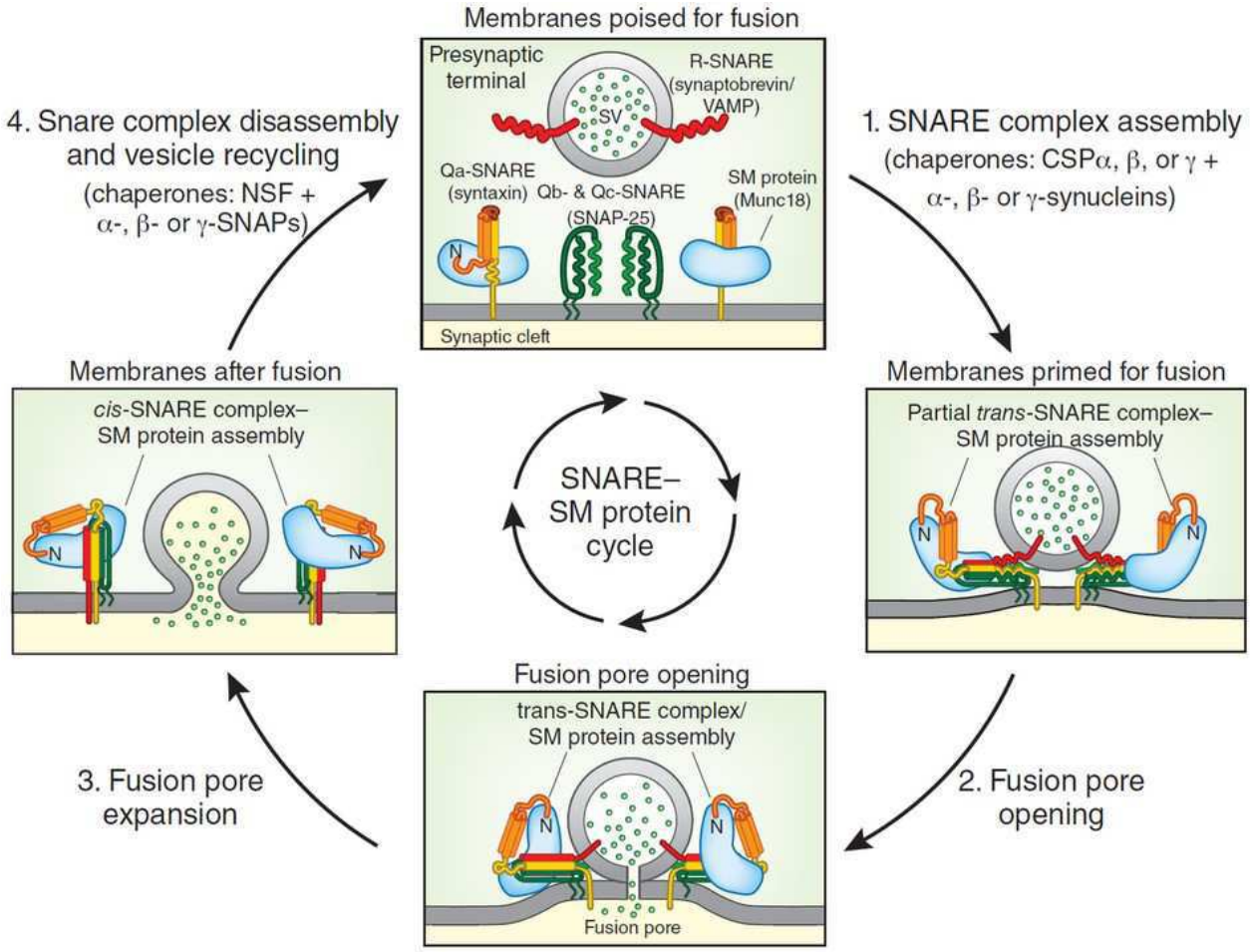


Fig.25. Pathway of a membrane fusion event. 1) Priming of vesicles for fusion. During this step Syntaxin changes its conformation from closed to open. Cysteine string proteins (CSPs) and Synucleins act as chaperones in order to facilitate the *trans*-SNARE complex assembly. Approaching of the two membranes is made possible by tethering proteins bound by a membrane-anchored complex of an activate GTPase (Rab) with its cognate effector; 2) The full *trans*-SNARE complex is assembled with accessory proteins such as Sec/Munc (SM) proteins, Ca^{2+}- and/or lipid-binding proteins and the fusion pore opens; 3) The fusion pore expands, converting *trans*-SNARE into *cis*-SNARE complex; 4) NSF and SNAPs mediate disassembly of the SNARE complex, leading to vescicle recycling. From: Sudhof, T. A molecular machine for neurotransmitter release: synaptotagmin and beyond. *Nat med*. 2013;1227-1231; Wickner W, Schekman R. Membrane fusion. *Nat Struct Mol Biol*. 2008 Jul;15(7):658-64.

At first, membranes are tethered thanks to membrane-anchored Rab GTPase proteins, Rab effectors and tether proteins charachterized by having a coiled-coil structure. Also SNARE proteins recruitment is Rab-mediated. The formation of *trans*-SNARE complex takes advantage of three SNARE motifs from the acceptor membrane and one from the donor vesicle. The membranes hemifusion is then driven by the free energy from the SNARE complex assembly. After the formation of the fusion pore and its expansion, SNARE

complex becomes in a *cis* configuration and the mixing of the inner lipid layer and the luminal content will happen. At the end of the fusion process, NSF protein, bound to α-SNAP, will disassemble the cis-SNARE complex.

Longin Domain (LD)

As previously shown, the SNARE motif tunes the fusion process together with other elements, including the Longin domain Discovered in 2001 by a bioinformatic approach (Filippini *et al*.,2001), the LD is not only a SNARE regulatory domain, as it is involved in further steps along the life cycle of vescicles. In fact, domains with the LD fold appear in the σ and μ subunits of the AP2 complex involved in vescicle formation (Collins *et al*., 2002), in the SEDL subunit of the TRAPP complex responsible for vescicle docking (Kim *et al*., 2006) and in the SNARE complex related to the membrane fusion event. Indeed, the LD is found in further subcellular trafficking routes other than the vescicle life cycle: for example, in the case of SRX subunit of the SRP receptor, LD is able to act as potential small-GTPase effector (Schlenker et al., 2006). LD proteins belong to seven homologous superfamilies (De Franceschi *et al*., 2014):

- *Sensu strictu* Longins: Ykt6p, Sec22b and VAMP7.
- Adaptins: LD is retrieved in the σ an μ subunits of AP complexes (1-4) involved in vescicle budding at the Golgi, the endosomal compartment and the plasma membrane.
- Sedlins: LD takes part in TRAPP I (at the *cis*-Golgi surface) and TRAPP II (at the *trans*-Golgi surface) complexes involved in tethering ER-derived vescicles to the Gogli membrane and in Golgi trafficking.
- SANDs: This superfamily contains in *H. sapiens* HPS-1/4 (components of BLOC-3), Mon1A/B and Ccz1/C7orf28A proteins (acting as a GEF dimeric complex in yeast).
- Targetings: They are homologues of the N-ter region of the α subunit of the SRP receptor and they are able to mediate targeting of the ribosome to the ER by association with SRb.
- DENNs: They act as regulators of Rab GTPases function. Examples of DENNs are Avl9, FAM45A, FAM45B, LCHN.

AVLs

SNAREs carrying LD are referred to as *sensu stricto* Longins and are divided in three subfamilies on the basis of homology: Ykt6p, Sec22b and VAMP7 (Rossi *et al*., 2004).

Competitive binding to the SNARE motif, in yeast, can prevent Ykt6p participation to fusion bundle, thus regulating the membrane fusion. Moreover, the LD of VAMP7 is crucial in neurite outgrowth as demonstrated by the overexpression of a Δ-longin fragment increasing neurite outgrowth whereas the LD alone expression carries outgrowth inhibition (Martinez-Arca *et al*., 2000; Martinez-Arca *et al*., 2001).

From a structural point of view, the LD consists of approximately 120 amino acids. As revealed by published N-ter domain structures of yeast SNARE protein Ykt6 and murine SNARE Sec22b (Tochio *et al*., 2001; Gonzalez et al., 2001), the LD fold (Fig. 21) consists of five antiparallel β-strands (β1-β5) located in between an α-helix (α1) on one side and two helices (α2-α3) on the opposite side.

In addition to aforementioned regulation of fusion, the LD is able to mediate a number of different functions:

- Targeting R-SNAREs to the right subcellular compartment: for example, VAMP7 is targeted to late endosome by binding to the δ subunit of the AP3 complex (Martinez-Arca et al., 2003).
- Selecting cargo into sorting vescicles: this function is made possible thanks to the interaction with vescicular coat proteins. For example, the export of Sec22b from the ER to the Gogli is ensured by the interaction between the α2-α3 interface of the LD and the COPII subcomplex Sec23/24. This subcomplex is able to recognize a conformational epitope on the α2-α3 surface, formed by the N-ter of SNARE motif folded on the α1-β3 interface of the LD. This folded back conformation prevents unspecific SNARE binding to other partners (Mancias and Goldberg, 2007).

1.1 VAMP7 and its isoforms

Also known as Tetanus Insensitive VAMP (TI-VAMP) because of its resistance to tetanus and botulinum neurotoxin (Galli *et al*., 1998), VAMP7 is a member of the Longins R-SNARE family. Human VAMP7 is encoded by SYBL1 gene, located in the Xq/Yq pseudoautosomal region (PAR); this gene is transcriptionally repressed on the Yq PAR region (D'Esposito et al., 1996). Both human and mouse tissues ubiquitously exhibit VAMP7 proteins (D'Esposito *et al*., 1996; Matarazzo *et al*., 1999), also involved in different cell pathways. VAMP7 can adopt a closed conformation via a LD-SNARE motif binding. Its LD is bound by the chlatrin adaptor Hrb and ArfGAP and this way SNARE motif competes for the same groove, indicating that Hrb-mediated endocytosis of VAMP7 occours only when it takes place into a

cis-SNARE complex (Pryor et al., 2008). However, VAMP7 functions retrieved in brain, such as control of neurite outgrowth, neuronal plasticity and morphogenesis have great biological relevance. Infact, VAMP7 is able to interact with VARP, a GEF of the small GTPase Rab21, acting as a positive factor for neurite growth (Burgo *et al*., 2009; Tamura *et al*., 2009). Like other genes involved in subcellular trafficking, SYBL1 undergoes to alternative splicing (AS), resulting in the production of isoforms having different SCL and properties. SYBL1 gene is made up of 8 exons: exon 1 is non coding (5'-UTR), exons 2 to 4 encode LD, whereas 5 to 8 encode both SNARE motif and TMD (Fig. 26).

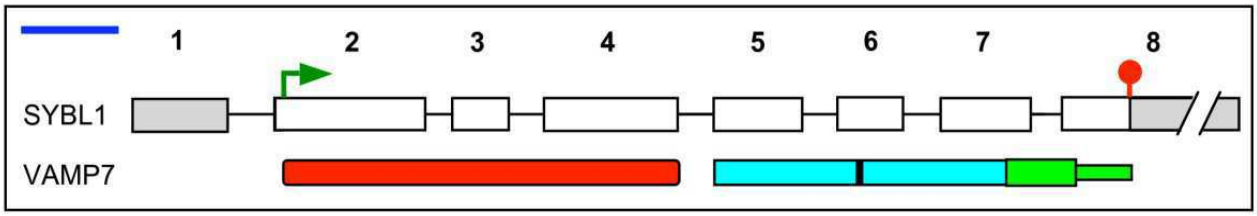


Fig.26. Genomic structure of the SYBL1 gene and protein domain architecture of VAMP7. SYBL1 exons are numbered: coding regions are white and non-coding ones are grey. Green arrow and red circle indicate the start and stop codon, respectively. VAMP7 LD is highlited in red, SNARE motif in cyan (the black vertical bar indicates the conserved arginine of the polar layer) and the TM region in light green. From: Vacca M, Albania L, Della Ragione F, Carpi A, Rossi V, Strazzullo M, De Franceschi N, Rossetto O, Filippini F, D'Esposito M. Alternative splicing of the human gene SYBL1 modulates protein domain architecture of Longin VAMP7/TI-VAMP, showing both non-SNARE and synaptobrevin-like isoforms. *BMC Mol Biol*. 2011 May 24;12:26.

AS of SYBL1 produces two kinds of isoforms, on the basis of different exon skipping: “non-longin” isoforms and “non-SNARE” isoforms (Fig.27.):

<u>“Non-longin” isoforms</u>

These isoforms are produced by exon skipping at the 5’ half of SYBL1 and are also referred to as synaptobrevin-like. “Non-longin” isoforms are characterized by the absence of the LD and retain all the canonical exons from 4 to 8. In their work published in 2011, Vacca *et al*. named these variants as “c”, “d”, and “h”.

- *VAMP7c*: Originates from a splicing event skipping out approximately 40 residues of the N-ter region but the mRNA presents the correct reading frame; the variant protein lacks the LD but SNARE motif, TMD and intravescicular tail are shared with the main isoform.
- *VAMP7d/h*: VAMP7d and h are different mRNAs which are able to encode, via translation reinitiation from an alternative TIS in exon 5, the same polypeptide with SNARE motif, TMD and intravescicular tail.

“Non-SNARE” isoforms

Skipping events of SYBL1 exons 5 and/or 6 are responsible for the production of these isoform, named as “b”, “i” and “j”. These proteins contain the LD but not the SNARE motif.

- *VAMP7b*: Represents the study object of this workpackage. It is 40 residues longer than the main isoform (VAMP7a) and retains the LD and N-ter part of the SNARE motif. A novel, C-ter region of 116 residues of unknown function arises from a frameshift downstream of exon 5.
- *VAMP7i*: Consists in the LD alone.
- *VAMP7j*: Originates from the skipping of both exons 5 and 6 preserving the original frame: this isoform comprises LD followed by a short hinge region preceding the original TMD and intravesicular tail as reported in VAMP7a.

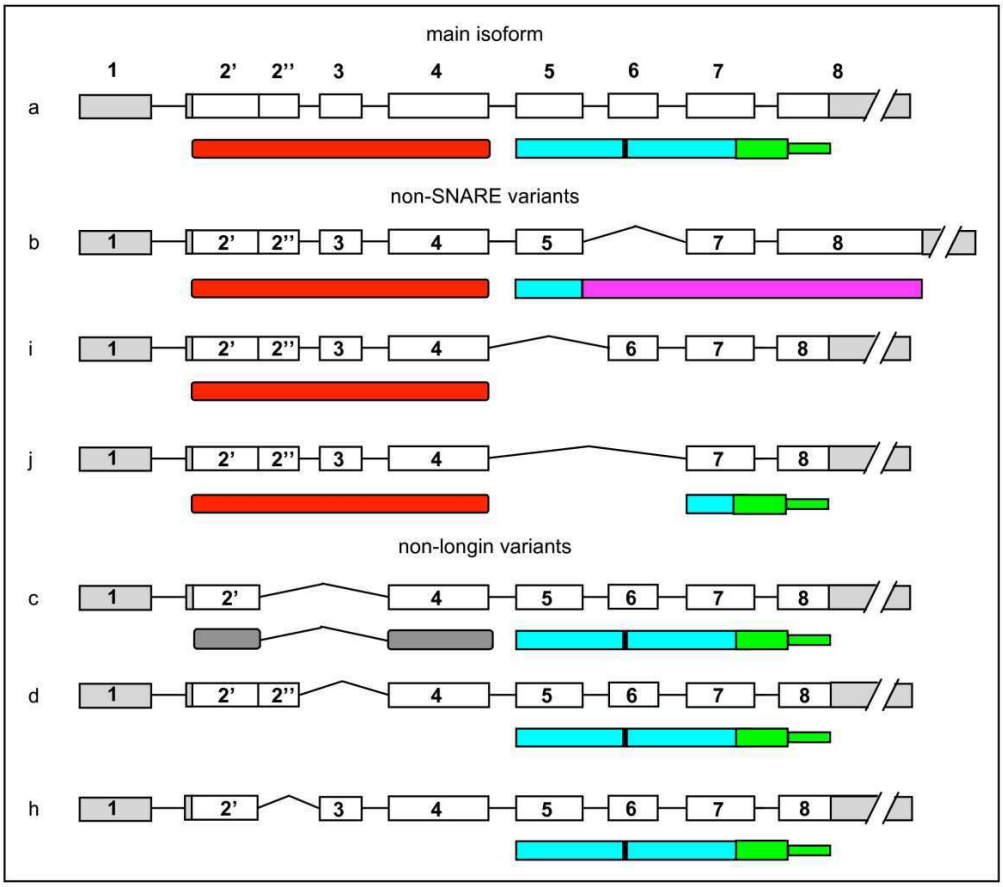


Fig.27. Alternative splicing of human SYBL1 gene and corresponding products. Coding exons are numbered and white, non coding ones are gray. LD is highlited in red, SNARE motif in cyan (the conserved arg in the polar layer is represented by a vertical black bar), TMD and intravescicular tail in green. The magenta region indicates the unknown function area of VAMP7b. From: Vacca M, Albania L., Della Ragione F, Carpi A, Rossi V, Strazzullo M, DeFranceschi N, Rossetto O, Filippini F, D'Esposito M. Alternative splicing of the human gene SYBL1 modulates protein domain architecture of Longin VAMP7/TI-VAMP, showing both non-SNARE and synaptobrevin-like isoforms. *BMC Mol Biol*. 2011 May 24;12:26.

As previously mentioned, VAMP7b isoform is produced by exon 6 skipping and a following coding sequence frameshift. In general, sequence frameshifts result in the appearance of a stop codon and thus in premature truncation of the polypetide chain. Instead, intriguingly in VAMP7b such a stop codon is not found and conversely translation goas ahead along part of the original 3'-UTR, resulting in a protein that is even longer (260 aa vs. 220 aa) than the main isoform. Given that this isoform is the subject of investigation of this part of the thesis work, further details on its sequence and architecture are found in the results section, while it has to be stressed here that the VAMP7b protein existence (and existence of the novel, 116 aa region specific to VAMP7b) was experimentally confirmed by using specific antibodies (Vacca *et al*., 2011).

Results and discussion

Preliminary structural characterization of VAMP7b isoform

1. VAMP7b

When insertions or deletions - depending on either mutations or alternative splicing - result in shifting the protein coding frame, most often this determines premature termination of translation, because several stop codons are commonly present in the two possible shifted frames.

The splice variant VAMP7b is characterized instead by “elongated translation”, as its novel 116 residues region is even longer than the original regions (C-ter half of the SNARE motif, transmembrane domain and intravesicular tail) it replaced (Vacca *et al*., 2011).

In order to distinguish between a possible random event and the result of functional selection, we investigated on the evolutionary conservation of such sequence feature in VAMP7 genes from other species. In public databases, most of cDNA sequences for VAMP7 isoforms are computationally predicted mRNAs, of which some might be wrong predictions. Therefore, either presence or absence of early or intermediate stop codons in the frame coding for the 116 aa homologous sequence was assessed by translating VAMP7 genomic sequences. In apes, no such stop codons were found, confirming the elongated translation observed with the corresponding human 'long' VAMP7b. Instead, premature truncation was found in new world monkeys as well as in other mammals and vertebrates, having 'short' VAMP7b variants (Fig. 52).

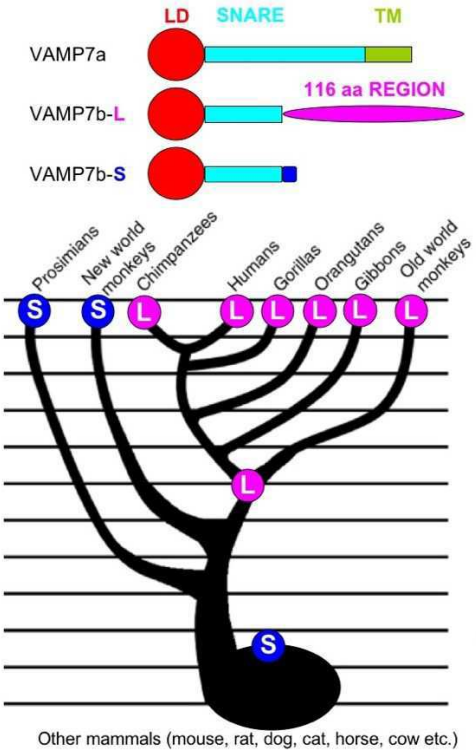


Fig.52. Domain architecture and distribution in mammals of long (L) or short (S) VAMP7b variants.

Wet lab preliminary evidence from coauthors suggests that VAMP7b mRNA might exist e.g. in mouse; however, in this case premature truncation is found and the novel domain is not produced.

When considering that (i) VAMP7 is involved in neurite outgrowth and neuronal function (Martinez-Arca *et al*., 2000 and 2001), and (ii) its alternative splicing is modulated along neuronal and brain development (Vacca *et al*., 2011), it is tempting to speculate that the special architecture of VAMP7b in humans and primates might account at least in part for their very complex brain evolution and cognitive functions.

The real existence (i.e., at protein level) of the novel, b-specific 116 aa region has already been confirmed by using antibodies specific to this domain (Vacca *et al*., 2011) and this further prompted us to perform in silico investigations on the novel VAMP7b architecture and on possible structure and function of its specific C-ter region.

Because no template with ≥ 30% of sequence identity could be identified, preliminary modelling work started by fold recognition, using PHYRE 2 webserver (Kelley *et al*., 2015). Threading softwares evaluate target sequence against a library of unique fold representatives according to residue by residue similarity in terms of spatial site, hydropathy and helix- or sheet-forming propensity.

PHYRE 2 modeled the VAMP7b part shared with VAMP7a (1-144) in close conformation, whereas only a small, central fragment within the novel C-ter region (residues 188-205 of VAMP7b, corresponding to residues 44-61 of the novel region) showed similarity to a Ferredoxin helical region.

It has to be noticed that the 144 N-ter residues shared by VAMP7a and VAMP7b includes both the Longin domain (LD) and the N-ter part of the SNARE motif until the first two residues (Asn, Ile) of the so-called 'NIE' motif that in Longin Sec22b was found to mediate intramolecular binding to the α1-β3 region of the LD (Mancias and Goldberg, 2007). Closed conformation depending on such LD-SNARE intramolecular binding in turn allows Sec22b to bind the COPII subcomplex Sec23/24 complex and specifies its endoplasmic reticulum (ER) exit as an unassembled SNARE (Mancias and Goldberg, 2007). The closed conformation and intramolecular LD-SNARE binding in VAMP7 was characterized via a collaboration of our team and the lab of Axel Brunger (Stanford, USA), finding that the full cytoplasmic region (1-180) and fragment 1-160 of VAMP7 can mediate intramolecular bnding and stable closed conformation; if fragment 1-150 is used instead, intramolecular binding still occurs but the closed conformation is less stable (Vivona *et al*., 2010).

Alignment of VAMP7a and VAMP7b each other and to the NIE regions of the characterized longins Ykt6 and Sec22b shows that crucial residues in the NIE motif region are intriguingly conserved, in spite after aa 144 they have different sequences (Fig. 53).

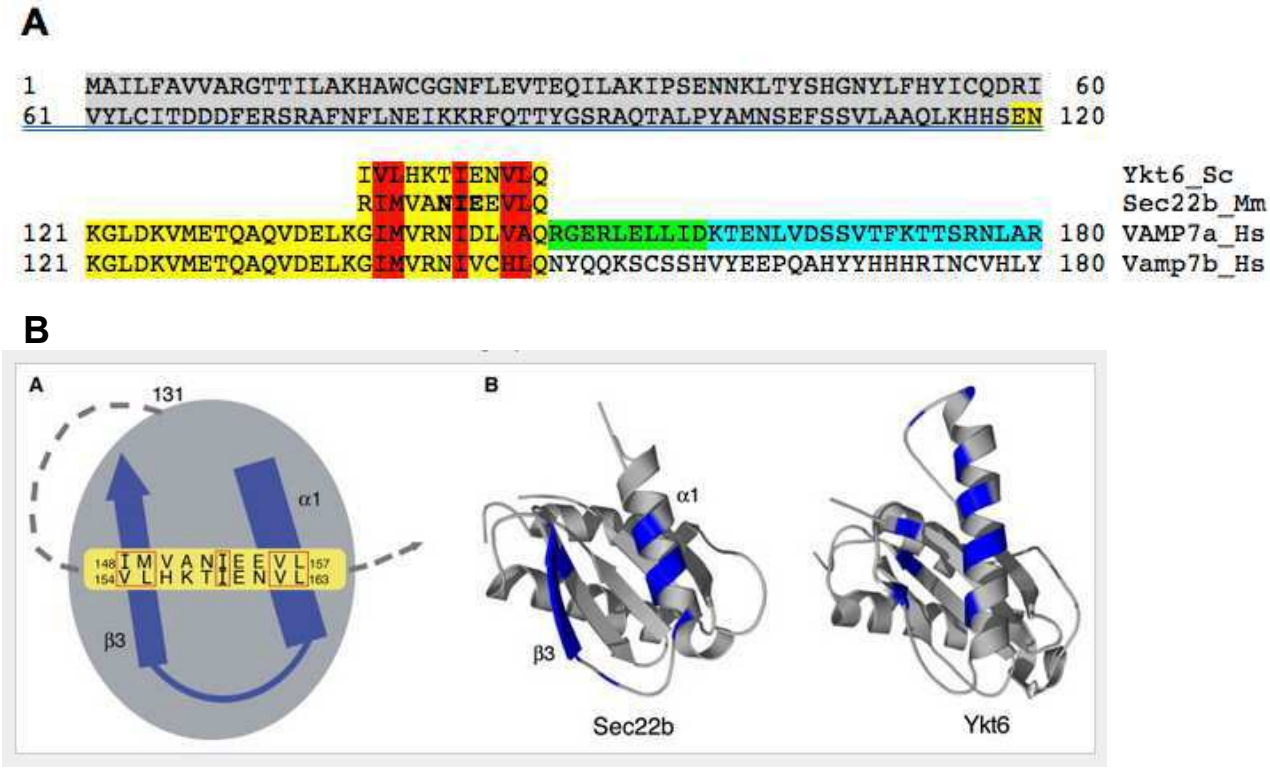


Fig.53. Conservation of the LD-binding SNARE residues among Longins Ykt6, Sec22b and VAMP7. In panel A, human VAMP7 sequence regions are highlighted as follows: grey, LD; yellow, SNARE motif until residue 150; green, SNARE residues 151-160; cyan, residues 161-180. The SNARE residues in the NIE motif region that are crucial to intramolecular binding to the LD are red boxed in panel B and highlighted in red in panel A. Panel B corresponds to Fig. 4 from Mancias and Goldberg, 2007.

This suggested VAMP7b 1-150 might be functionally equivalent to VAMP7a 1-150, i.e. equally (or similarly) able to mediate a somehow less stable, but anyway closed conformation by sharing the NIE-LD intramolecular binding. The 1-150 region of VAMP7b was more correctly modeled using a homology model approach with SWISS-MODEL (Biasini *et al*., 2014). As the best template SWISS-MODEL (automatic mode) choose the VAMP7 closed conformation in complex with VARP (Schäfer *et a*l., 2012) (Fig. 54), further suggesting that the conservation of the N-ter part of the SNARE motif and in particular of the NIE region might allow VAMP7b to share modulation of the open-closed conformation with the main isoform, eventually being stabilized by interacting partners.

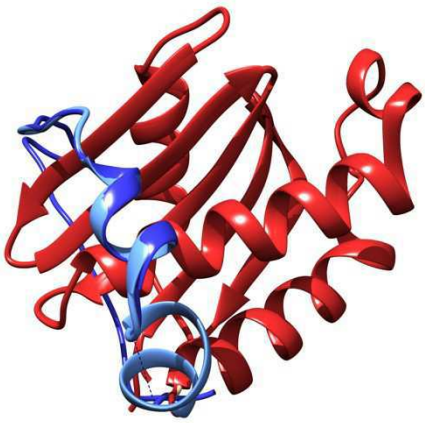


Fig. 54. Superposition between homology model VAMP7b 1-150 region and its template (4B93). Color code: LDs of VAMP7b and 4B93 are in firebrick, SNARE motif of VAMP7b is in cornflower blue and SNARE motif of 4B93 is in medium blue.

Such hypothesis has been double and reliably confirmed by wet-lab coworkers, both in vitro (NMR) and in vivo (via two-hybrid interaction assay in yeast).

Then, we focused our attention on the novel 116 aa region/domain of unknown function and, in the absence of neither available models for H.M. approach nor similar folds for threading, it was modeled *ab initio* using I-TASSER/QUARK server (Zhang *et al*., 2012). *Ab initio* modeling flowchart is used when structural homologues are missing. In general, this method provides a designed energy function-driven conformational search. As a consequence, a number of possible conformations, known as structure decoys, are generated and final models are selected from them. Three features are responsible for a successful *ab initio* prediction (Ridgen, 2009):

- An accurate energy function able to correlate the protein native structure to the most thermodynamically stable state between decoys;
- An effecient search method identifying the low energy states via conformational search;
- Selection of native-like models from a pool of decoy structures.

However, prior to starting *ab initio* modelling, this unique 116 residues region was used as a sequence probe for a number of analyses with standard (blastp, tblastn) and special (PSI- and PHI-blast) local alignment search tools, with the aim to 'dissect' such region in eventual subregions hence providing some hints to the next modelling steps. PSI-BLAST is an extension of BLAST that uses position-specific scoring matrices (PSSMs) to assign a score

to matches between the query and database sequence; a multiple alignment of high scoring sequence will be used iteratively to generate a new PSSM used in the next round of searching. PSI-BLAST is more sensitive than BLAST, so it might be used to find distantly related sequences not retrieved in a BLAST search. Pattern Hit Initiated BLAST (PHI-BLAST) instaed allows to select sequences sharing both similarity and a pattern of residues likely (or known) to represent a motif.

Iterations of position-specific matrix and pattern hit initiated sequence analysis using the whole 116 aa sequence or subsequences allowed to identify three subregions as presented in Fig. 55: 145-184 (magenta), 185-231 (green) and 232-260 (cyan). In particular, each of these regions showed similarity to a domain or to protein regions sharing a binding function or a structure (Fig. 55).

VCHLQNYQQKSCSSHVYEEPQAHYYHHHRINCVHLYHCFTSLWWIYMAKLCEEIGKKKLPLTKDMR
EQGVKSNPCDSSLSHTDRWYLPVSSTLFSLFKILFHASRFIFVLSTSLFL

Fig. 55. Subregions of the novel domain of VAMP7b isoform, as retrieved by PSI-BLAST: metal ion binding (magenta), JMJ-C-like (green), helical structure (cyan).

The *ab initio* modeling using QUARK server (Xu and Zhang, 2012) resulted in model shown in figures 56 and 57:

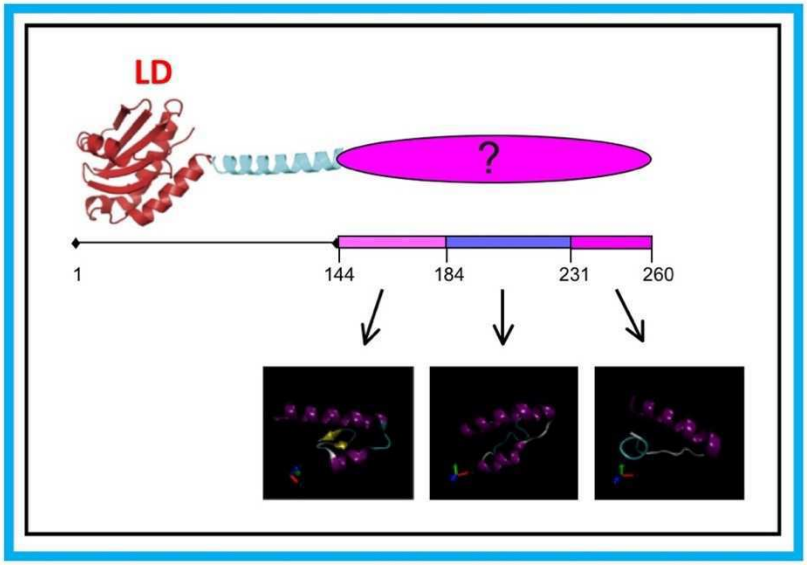


Fig.56. Domain architecture of VAMP7b-L and in silico dissection of its 116 aa region of unknown (?) structure and function. The three subregions identified by PSI-BLAST and corresponding *ab initio* predicted structures.

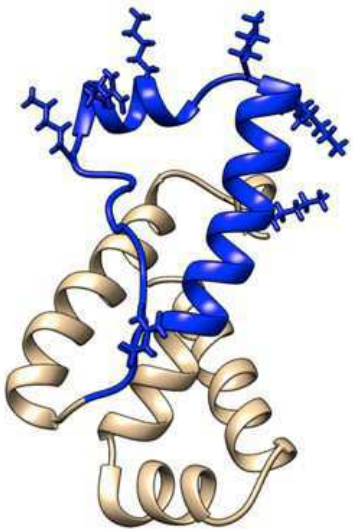


Fig. 57. Cartoon representation of the novel C-ter region of VAMP7b. A putative RNA binding regioni is highlited in dark blue.

Hits retrieved by the N-terminal fragment 145-184 showed as a common feature binding to divalent cations; in addition, a slightly degenerate Zinc finger C2H2 type domain signature (Prosite AC PS00028) was found by PROscan analysis to correspond to region 156-178. Hits extracted by the central fragment 185-231 share nucleic acid binding regions and in particular, homology to JmjC was found. This domain is highly conserved among proteins with cupin fold (Clissold and Ponting, 2001), which bind divalent cations, in agreement with function suggested for the contiguous fragment 145-184. JmjC domain is involved in chromatin remodelling and indeed fragment 185-231 shows homology to the JmjC region of histone demethylase, playing a pivotal role in epigenetic regulation (Hancock *et al*., 2017). Intriguingly, the JmjC domain of KDM4D lysine demethylase has been found to mediate RNA binding (which is demethylase-independent) via an α helical motif exposing positively charged residues (Zoabi *et al*, 2014). JmjC domains are also found in pre-mRNA splicing factors such as JMJD6, a nuclear protein involved in histone modification, transcription and RNA processing for adipogenic gene expression (Hu *et al*., 2015). It is noteworthy that both these types of proteins are known to bind divalent cations. Features common to Zn-binding regions were found again in the most C-terminal region, 232-260, by PROscan analysis, as partially overlapping sequences 239-248 and 246-256 regions are highly similar to the Zn-binding signatures PS00216 and PS00142, respectively.

The mRNA transport and local protein synthesis play a vital role in the control of polarity, synaptic plasticity and growth cone motility. RNA-binding proteins, which form the transported ribonucleoparticle (RNP), control mRNA stability and local translation. Recently, the existence of processing bodies (P-bodies), in which mRNA decapping and degradation

take place, was revealed in neurons. It was suggested that P-bodies serve as a transient storage compartment for mRNAs, which can be released and, upon stimulation, resume translation. It is noteworthy that Zinc binding and RNA binding relate as Zinc is a translation regulator in neurons as being involved in disruption of polysomes, aggregation of P-bodies in neurons and impairment of the RNP-polysome interaction (Blumenthal and Ginzburg, 2008). Containing mRNA granules mammalian FMRP is associated with the neuronal specific kinesin KIF3C; FMRP seems to act as a molecular link between microtubule-based transport and mRNA cargo. Loss of function mutations of *FMR1* gene, encoding FMRP, are responsible for the Fragile X syndrome (FraX). This pathology is characterized by severe neurological disorder, abnormal neuronal morphology and defects in the number and function of synapses (Estes *et al*., 2008). Increasing evidences suggest that axon and growth cone mRNAs play an important role in axon extension and pathfinding via local translation. Moreover, microRNAs (miRNAs) in axons are able to control local protein synthesis during axon development. The regulation of mRNA localization (and local protein synthesis) is necessary for axon guidance, regeneration and synaptic plasticity (Sasaki *et al*., 2014). Moreover, subcellular transcriptome analysis of neural projections and soma revealed that alternative last exons (ALEs) often confer isoform-specific localization: in poarticular, gene-distal ALE isoforms are four times more often localized to nurite than gene-proximal isoforms. These localized isoforms are induced during neuronal differentiation and enriched for motifs associated with muscle-blind-like (Mbnl) family RNA-binding proteins (Taliaferro *et al*., 2016).

The RNA-binding propension of this new 116aa region of VAMP7b was evaluated also by other bioinformatic analyses: bind-N and electrostatic analyses are reported in Fig. 58 and Fig. 59, respectively.

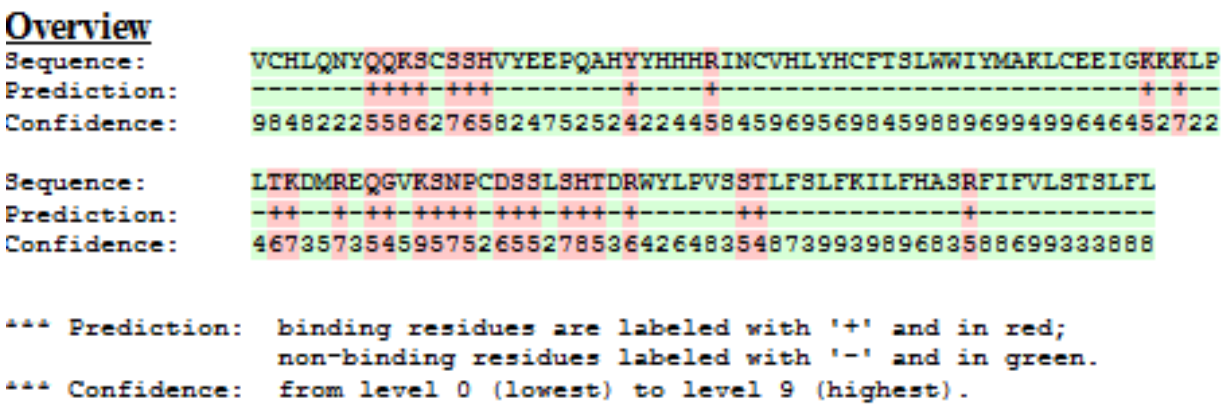


Fig. 58. BindN prediction of the 116aa new region of VAMP7b.

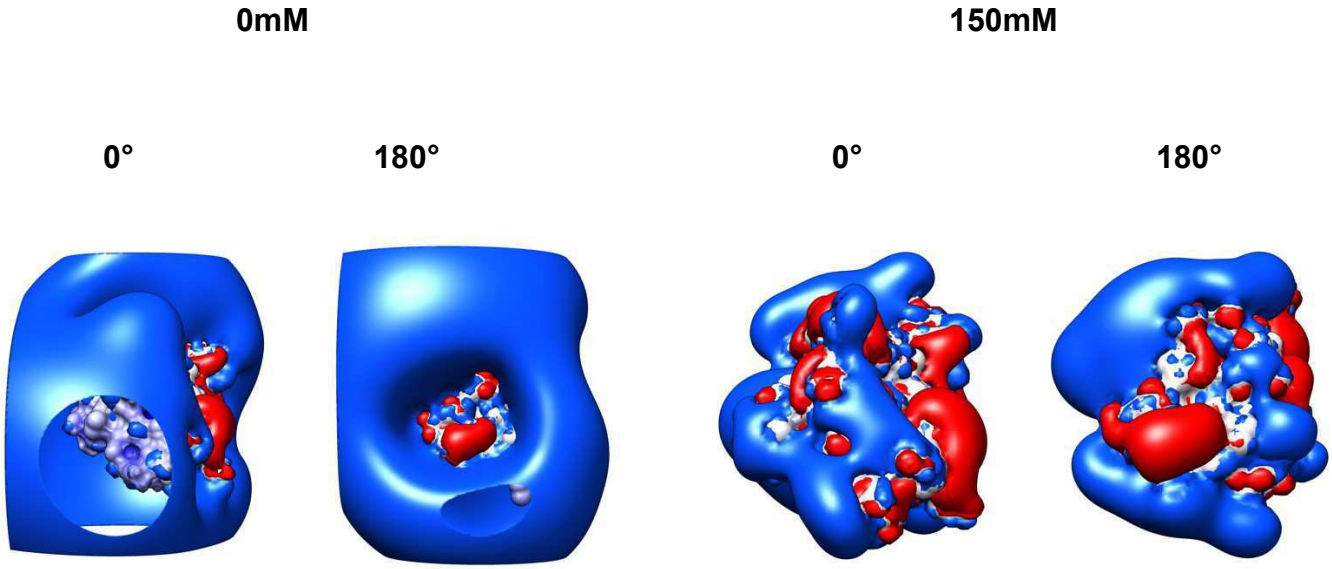


Fig. 59. Isopotential contour of the 116aa new region of VAMP7b

Finally, docking simulations using NPDock (http://genesilico.pl/NPDock) (Tuszynska *et al*., 2015) between RNA and VAMP7b were carried out (Fig. 60). RNA structure (4msr) was retrieved at http://rna.bgsu.edu/rna3dhub/nrlist/release/1.89. Wet experiments to prove such RNA interaction are ongoing.

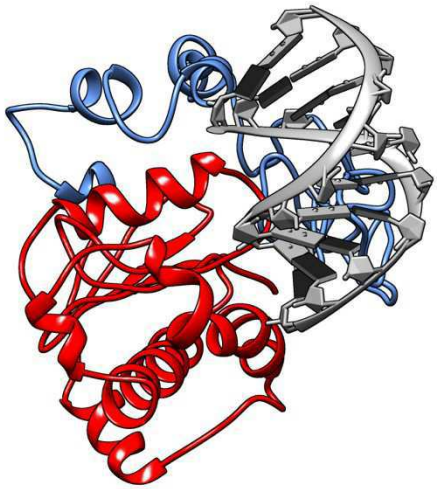


Fig. 60. VAMP7b-RNA (4msr) complex via NPDock. The novel C-ter region of VAMP7b is highlited in cornflower blue.

Interacting residues were evaluated using UCSF Chimera and RING at http://protein.bio.unipd.it/ring/results/58f9bdca23dd205d5b69c370; crossing data from these two softwares and bind-N revealed that positions 168, 212, 215, 217 seem to be important in VAMP7b-RNA binding.

CHAPTER 3

Binding motifs regulating neurite outgrowth and guidance

State of the art

1. *CELL ADHESION MOLECULES (CAMs)*

Cell adhesion molecules (CAMs) are proteins located at the cell surface and involved in binding to other cells or the extracellular matrix (ECM). For example, some CAMs such as Neurexin and Neuroligins are able to facilitate cell-cell interactions (Craig and Kang, 2007), whereas Integrins are proned to interact with ECM.

These proteins are made up of three domains: intracellular, transmembrane and extracellular. Roughly speaking, CAMs belong to two classes (Fig. 28):

- Ca^{2+} dependent: Cadherins, Selectins, Integrins
- Ca^{2+} independent: Ig superfamily

Moreover, CAMs are able to mediate two kinds of interactions: homophilic or heterophilic. Homophilic binding is defined *trans* when one protein attached to the cell surface interacts with an identical protein protruding from an opposite cell surface. NCAM1 and N-cadherin are involved in such kind of interactions. In addition, many CAMs can bind heterophilically with ECM partners and CAM proteins at the plasma membrane (Comoglio *et al*., 2003), as well as with intracellular proteins of the cytoskeleton and with enzymes (Mège *et al*., 2006; Takai *et al*., 2008; Buttner and Horstkorte, 2010) (Fig.29.). Thanks to their cytoplasmic domain, CAMs are able to interact with signaling molecules in the cell, thus participating in signal transduction and adhesive regulation. However, also ectodomains of many CAMs enable these proteins to modulate signal transduction (Cavallaro and Dejana, 2011).

Even if CAMs are involved in many biological processes and come in different types, this workpackage focused on neural cell adhesion molecules of the Ig superfamily. Here, CAMs are involved in the migration of neural crest cells (McKeown *et al*., 2013), the growth, guidance and regulation of axons (Kamiguchi, 2007; Zhang *et al*., 2008). As a consequence, mutations in CAM genes are linked to neurological disorders (Sytnyk *et al*., 2017) such as the CRASH syndrome (Zhang, 2010). CAMs are required both during development and in the adult nervous system. Different types of CAM interactions lead to contact-mediated attraction or repulsion (Maness and Schachner, 2007).

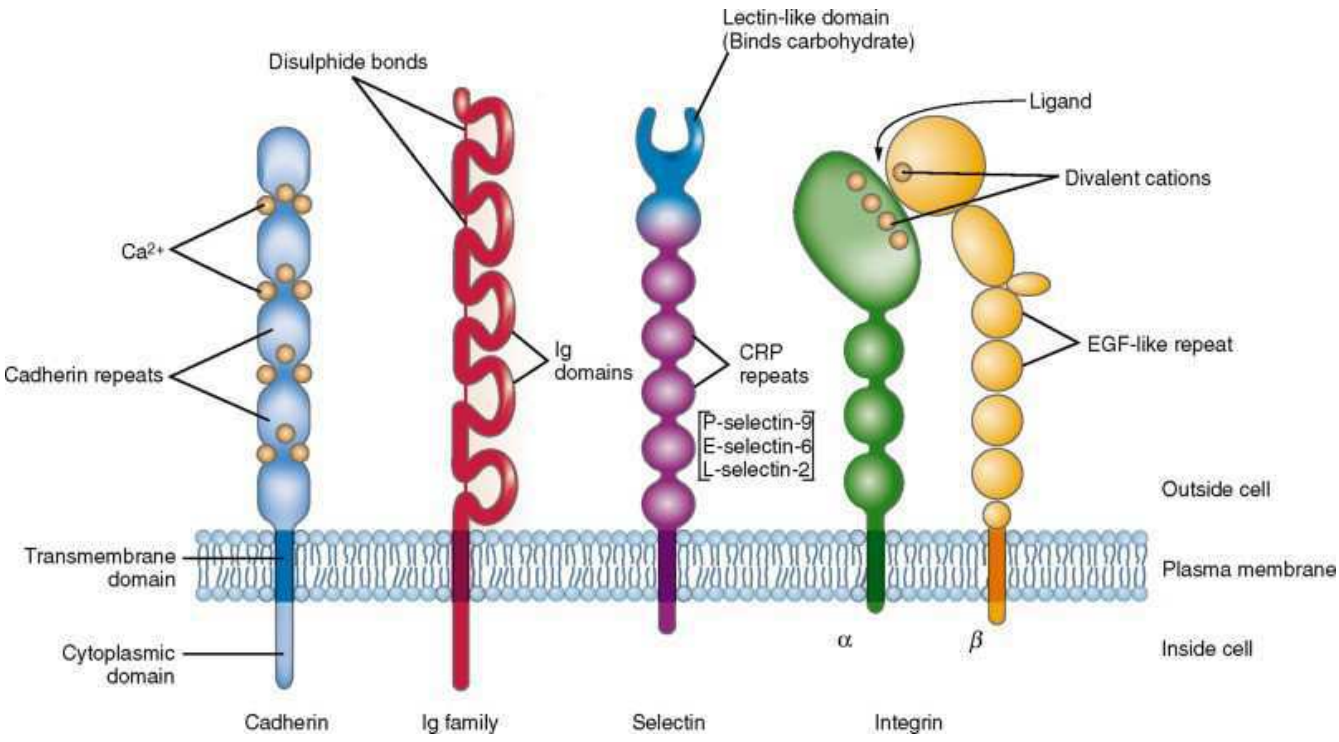


Fig.28. Diagram of the four major CAMs molecular structures: calcium-dependent CAMs exhibit sites for divalent cations (yellow dots). From: http://www.sciencedirect.com/topics/page/Single-pass_transmembrane_proteins.

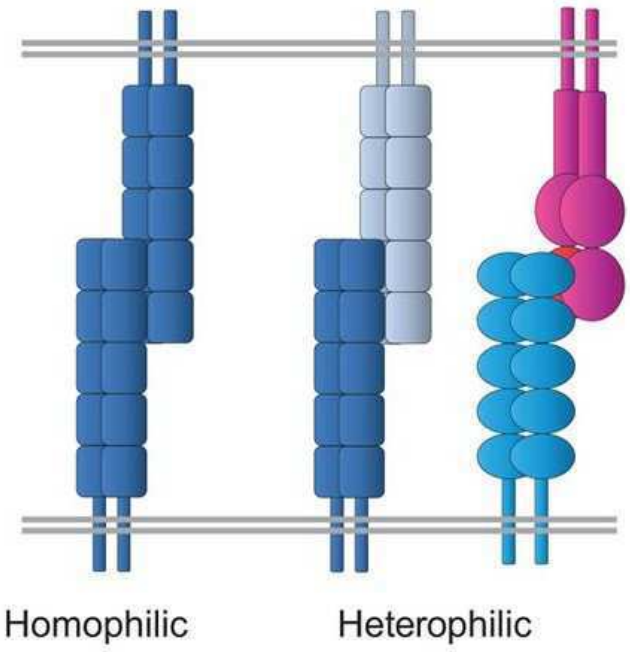


Fig.29. Diagram of CAM mediated interactions. In homophilic (or heterophilic) adhesion, a molecule at the cell surface binds to another identical (or different) molecule from the opposite cell surface. Cell-cell adhesion is often homophilic and cell-ECM adhesion is always eteropohilic. Adapted from: Fagotto F. The cellular basis of tissue separation. *Development*. 2014 Sep;141(17):3303-18.

1.1 L1 subfamily

The L1 subfamily encompasses L1, CHL1, Neurofascin and NrCAM (Fig. 30). These molecules are crucial to axon outgrowth and fasciculation, neuronal migration and survival, synaptic plasticity and regeneration after trauma (Maness and Schachner, 2007).

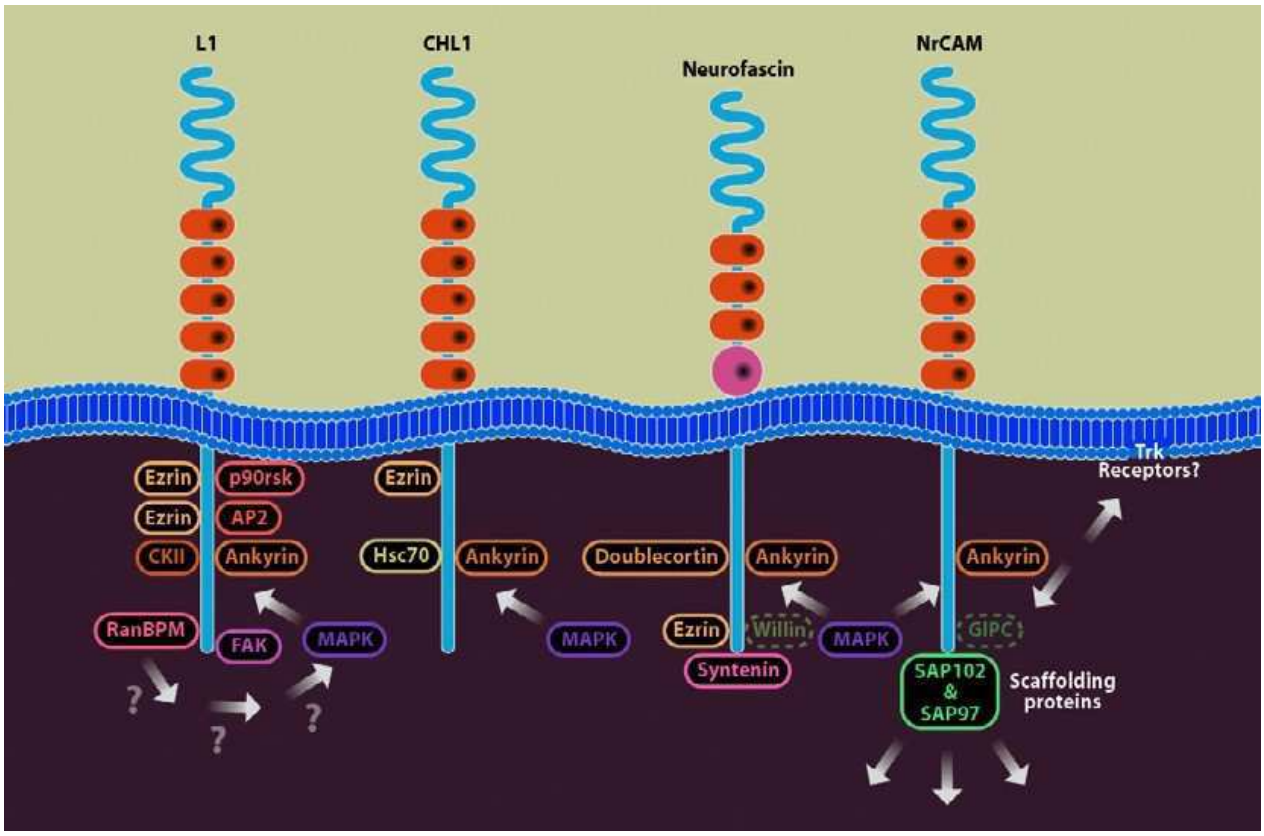


Fig.30. L1 subfamily members and their cytoplasmic interactors. From: Herron LR, Hill M, Davey F, Gunn-Moore FJ. The intracellular interactions of the L1 family of cell adhesion molecules. *Biochem J*. 2009 May 1;419(3):519-31.

L1 subfamily members are linked to neurological disorders such as: CRASH syndrome, foetal alcohol syndrome, increased vulnerability to autism and addiction, multiple sclerosis, schizophrenia and 3p syndrome. L1 non-nervous proteins are also involved in diseases, including cancers of the lung, pancreas, kidney, colon, melanoma, uterine and ovarian carcinomas. From a structural point of view L1, CHL1, Neurofascin and NrCAM share an overall similar structure: a large ectodomain made up of several Ig and Fibronectin type III (FnIII) repeats, and a single transmembrane hydrophobic region followed by a ~120 amino acid cytoplasmic tail able to interact with different partners, potentially at the same time. This tail contains an Ankyrin binding motif (SFIGQY): this binding is tuned through the potentially at the same time. This tail contains an Ankyrin binding motif (SFIGQY): this binding is tuned through the phosphorylation of the motif Tyr residue. Infact, this event abolishes Ankyrin

binding to both Neurofascin and L1; Tyr phosphorylation is crucial to the embryonic brain development and growth. Ser and Tyr residues in the Ankyrin binding motif are often mutated in patients with CRASH syndrome: Ser is mutated to Leu and Tyr to His, thus reducing the ability of L1 to recruit Ankyrin. Ankyrins are a family of cytoplasmic proteins coupling L1 subfamily members and ion channels to the spectrin cytoskeleton. This plays an important role in the formation of initial segments and nodes of Ranvier and in the growth cone initiation (Herron *et al*, 2009). The Ig fold is a highly conserved structure involved in driving homophilic and heterophilic protein-protein interactions. It is an all β-strand structure composed of 7-10 strands arranged into a 2 sheet "β-sandwich" (Haspel and Grumet, 2003). The Ig1-Ig4 domains of the L1 subfamily members can form horseshoe structures where the domains Ig1 and Ig2 interact with Ig4 and Ig3, respectively, with homophilic adhesion mediated by the Ig2 domain (Sytnyk *et al*., 2017).

L1

L1 is a protein of ~200 kDa encoded by *L1CAM* gene located in the long arm of the X chromosome (Xq28 position). L1 is endowed with six Ig domains and five FnIII repeats (Fig. 31). This protein can interact with the cytoskeleton via either Ankyrin or Ezrin-Radixin-Moesin (ERM) proteins. In the developing central nervous system (CNS) of mammals, L1 is primarily expressed at the surface of growth cones and axons of both developing and differentiated neurons and on Schwann cells of the peripheral nervous system. It ensures neural development by binding different set of molecules on neighboring neurons, glial cells and the ECM. On differentiated neurons, L1 localizes at contact regions between neighboring axons and on the growth cones (Kenwrick *et al*., 2000). Neuronal L1 is produced by alternative splicing and contains two specific sequences: (i) a RSLE motif in the cytoplasmic domain that mediates the AP-2-clathrin adaptor recruitment for endocytosis and (ii) an insertion in the Ig2 domain that increases homophilic binding. Neuronal L1 is internalized by clathrin-mediated endocytosis within the central domain of the growth cone and it is recicled to the front, promoting motility through new adhesive contacts at the leading edge and the detachment of old adhesions. Endocytosis of L1 is regulated by $pp60^{c\text{-}src}$, which can phosphorylate the YSRLE motif, inhibiting L1 binding to the AP2-clathrin complex in a possible feedback loop (Maness and Schachner, 2007; Herron *et al*., 2009). Neuritogenic activities and homophilic binding are features of L1 Ig2. Cell transfection studies revealed that neurite outgrowth is possible in transfectant expressing intact L1 and

not allowed in L1Δ2 transfectants. Moreover, competition experiments with wt Ig2 fusion protein led to L1-dependent cell aggregation inhibition, whereas an Ig2 fusion protein containing the hydrocephalus R184Q did not. Oligopeptide L1-A (His^{178} - Gly^{191}: HIKQDERVTMGQNG), derived from Ig2, inhibited homophilic binding, thus abolishing L1-dependent neurite outgrowth. The sequence of L1-A is suggested to contain an homophilic binding site, crucial in promoting neurite outgrowth. Arg184 seems to play a pivotal role in homophilic interaction: the substitution of this residue with Gln (HSAS mutation) leads to a drastic reduction in its ability to compete for the homophilic binding site (Zhao *et al.*, 1998).

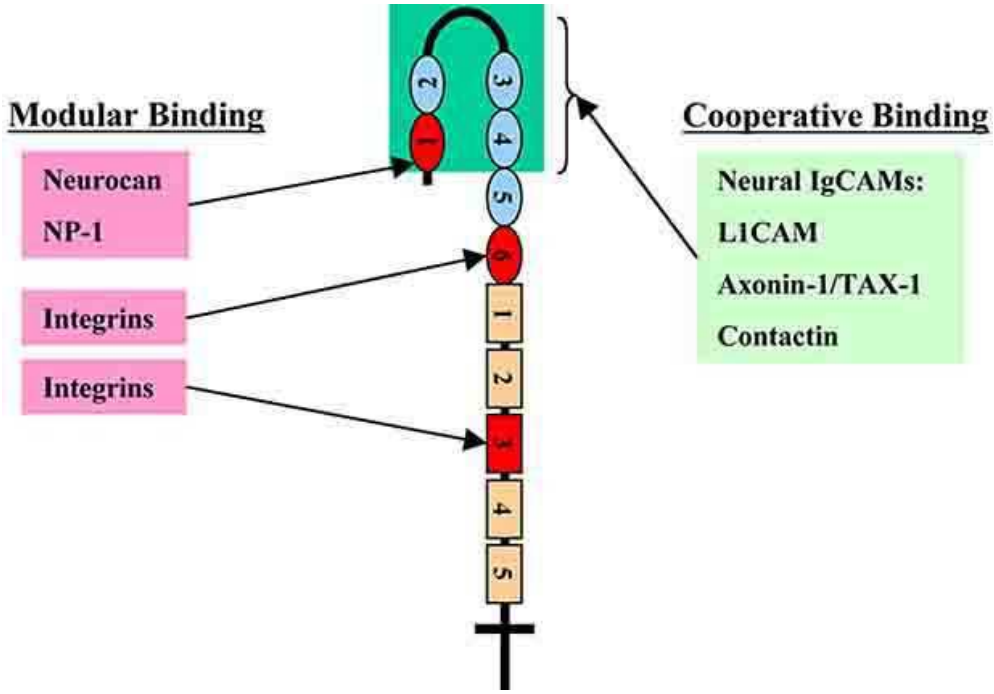


Fig.31. Interaction map for the L1 Ig domains (ovals) and FnIII domains (rectangles). Ig domains horshoe is highlited in green. From: Haspel J, Grumet M. The L1CAM extracellular region: a multi-domain protein with modular and cooperative binding modes. Front Biosci. 2003 Sep 1;8:s1210-25.

L1 is able to recruit other CAMs and signaling receptor at the neuronal membrane and, at the same time, to organize cytoskeletal and signaling proteins. L1 ectodomain exhibits three notable peptide sequences not included in Ig-like repeats (Fig. 32):

- *Leader sequence at the N terminus*: this sequence, containing the motif YEGHH encoded by exon 2, is a feature of the neuronal L1 isoform;
- *7-aa insertion between Ig2 and Ig3*: this ATNSMID sequence has hydrophilicity and flexiblility features likely to favour Ig1-Ig2 to pivot independently of the rest of L1;
- *Sequence between Fn5 and the TMD*: this is a proteolytic cleavage site.

The L1 ectodomain undergoes two covalent modifications: (i) extensive Asn-linked glycosylation accounting for 25% of the mass of L1 itself and (ii) proteolytic cleavage within

Fn3 and the aforementioned sequence in between Fn5 and TMD (Haspel and Grumet, 2003). Cleavage at the Fn3 site is operated by metalloprotease PC5A proprotein convertase (Maness and Schachner,2007) and produces fragments of 140 kDa and 85 kDa. 140 kDa fragment, containing Ig1-Fn3, may also be shed from the cell surface and can be recovered from human cerebrospinal fluid (CSF) (Haspel and Grumet, 2003). Cleavage at the site distal to Fn5 is mediated by the metalloprotease ADAM10 or ADAM 17 (Haspel and Grumet, 2003; Maness and Schachner, 2007), resulting in ~200 kDa and 32 KDa fragments. The first one contains the entire extracellular region and can also be recovered from the CSF.

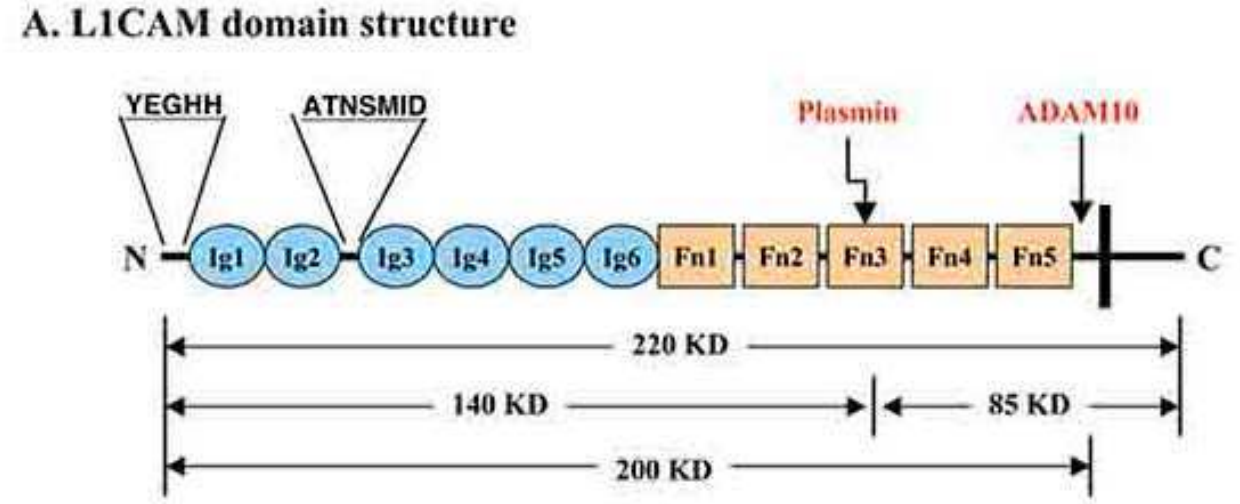


Fig.32. L1 domain architecture. The TMD (vertical bar) is followed by the cytoplasmic region. Proteolytic sites are denoted by arrows. Adapted from: Haspel J, Grumet M. The L1CAM extracellular region: a multi-domain protein with modular and cooperative binding modes. Front Biosci. 2003 Sep 1;8:s1210-25.

ADAM-mediated cleavage is regulated by $pp60^{c\text{-}src}$, ERK, and PKC and calcium. Growth factor activation enhances this event. L1 cleavage stimulates migration, adhesion and neurite outgrowth. Released extracellular fragments can improve signal transduction or adhesion reduction, thus improving motility (Maness and Schachner, 2007).

The L1 ectodomain is able to bind different partners: neural IgCAMs (such as L1 itself, NrCAM and axonin-1), non-Ig family CAMs (e.g. Integrins), ECM constituents (laminin, phosphacan, neurocan), signaling receptors (NP-1). Binding to these molecules can occur in either *cis* or *trans*. Analysis of the *L1CAM* gene from an evolutionary perspective can shed light on L1 ectodomain flexibility to accomodate many different binding partners. It was theorized that an ancestral duplication event produced multiple L1 family genes and the addition of new domains to the extracellular region added new functionality. Therefore, each Ig-like domain is able to create a multifunctional extracellular region due to their unique binding activities. As reported in Fig. 39, L1 is able to interact with different binding molecules in a modular or cooperative way. The “modular model” is based upon evidences

showing that certain binding activities of L1CAM segregated to individual domains: for example Ig1 is responsible for Neurocan and NP-1 binding or Ig2 proposed in 1998 by Zhao *et al.* for L1 homophilic binding in *trans*. Instead, Fn3 was proposed by Silletti *et al*., 2000 to mediate homophilic binding in *cis*. In the “cooperative model” Ig1-Ig4 mediate binding together, while Ig5-Ig6 and Fn2 enhance the interaction. Moreover this model seems to underlie the interactions between L1 and other neural IgCAMs as well as it happens for Nr-CAM and Axonin-1. Therefore, the Ig1-Ig4 segment may represent a conserved functional unit able to drive protein-protein interactions among members of the neural IgCAM family. L1 Ig1-Ig4 adopts a horseshoe-shaped structure (Fig. 39, Fig. 33) as revealed by electron microscopy (EM) studies. L1 seems to naturally interconvert between folded (horseshoe) and extendended conformations as reported by EM data; however, a crystallographic structure of L1 is neither available yet, nor is it clear which is the active conformation (Fig. 41). When Ig1-Ig4 adopt the extended conformation L1 can switch to modular mode binding (such as binding to Neurocan or Integrins). Since L1 has multiple binding opportunities at any given time, binding selection depends on the L1 ectodomain affinity for the different partners. One mechanism employs the short N-ter sequence YEGHH, important for both homo- and heterophilic binding, whereas clustering of extracellular regions at the cell surface can reorganize binding preferences, and it is mediated by interactions with Ankyrin and associated cytoskeletal proteins that can modulate cell adhesion. Finally, a third mechanism takes advantage from cooperation among L1 ectodomains in selecting one kind of binding partner over another (Haspel and Grumet, 2003).

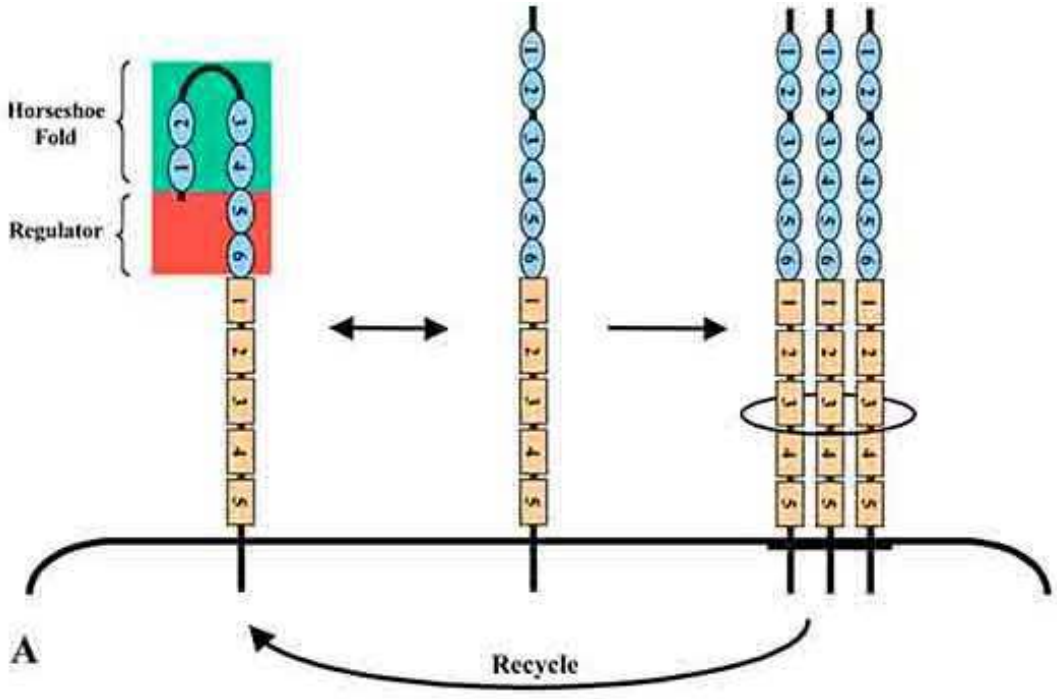


Fig.33. Model for L1 molecular interactions. At the cell surface, Ig1-Ig4 adopt a horseshoe-shaped conformation (highlited in green) that can reversibly convert into an extended conformation. Adapted from: Haspel J, Grumet M. The L1CAM extracellular region: a multi-domain protein with modular and cooperative binding modes. *Front Biosci*. 2003 Sep 1;8:s1210-25.

Similar to other transmembrane cell surface receptors, L1 acts like a signal transducer. L1 is able to bind Integrins inducing cell adhesion and directional motility (Fig. 34). Even if the mechanism underlying these events is still unclear, L1 transiently activates pp60$^{c\text{-}src}$, PI3 kinase, the Vav2 guanine nucleotide exchange factor, the Rac1 GTPase and PAK1 in a pathway culminating in MEK and ERK activation. The Ig1 motif of L1 (FASNKL) or CHL1 (FASNRL) can also bind the Semaphorin 3A receptor Neuropilin-1 to promote growth collapse. A conserved motif in the cytoplasmic domain of L1 (FIGQY) or CHL1 (FIGAY) recruits Ankyrin, which couples to F-actin through direct spectrin association. The phosphorilation of the Y of the motif induces the microtubule-associated protein doublecortin (DSX) recruitment, potentially able to couple L1 to microtubules. Similarly, β1 Integrins through RGD (L1) or DGEA (CHL1) motifs in their respective Ig6 domains, transduce adhesion signals via pp60$^{c\text{-}src}$, PI3 kinase, Rac1 and PAK1. In the cerebellum, L1 partecipates in radial migration of granule neurons, possibily in conjunction with NrCAM. *Cis* or *trans* binding of L1 and NP-1 to Sema3A results in axon repulsion or attraction, respectively. L1-deficient mice show guidance errors of corticospinal, retino-collicular, thalamocortical and callosal axons due to the lack of responsiveness to the repellent effects of Sema3A. An increasing amount of cGMP accompanies the switch from repulsion to

attraction and receptor endocytosis seems to favour axon repulsion: infact, Sema3A promotes L1 and NP-1 co-internalization. The conversion of repulsion to attraction blocks endocytosis and growth cone collapse.

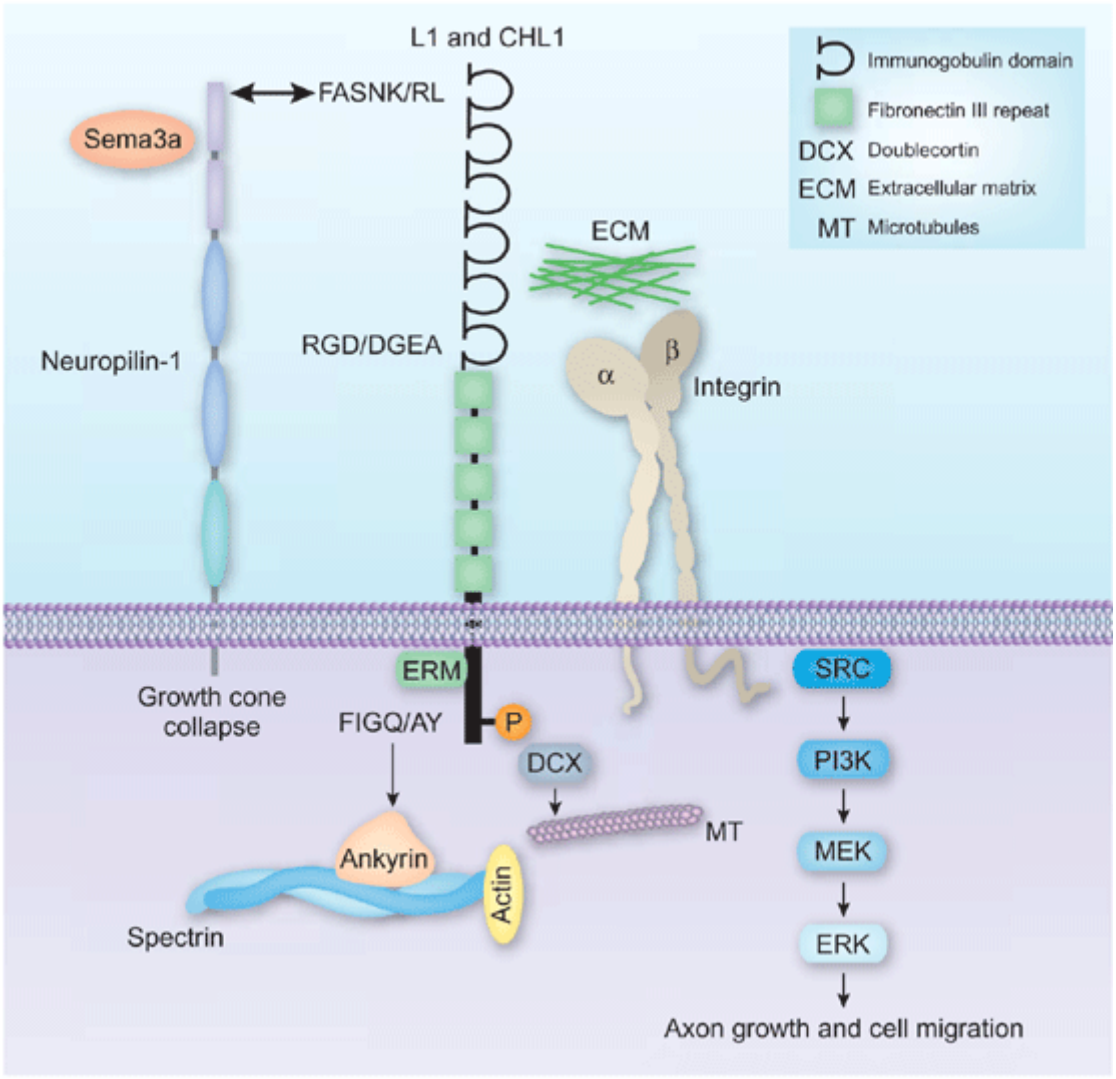


Fig.34. Signaling pathways downstream of L1/CHL1. Proteins involved are shown. From: Maness PF, Schachner M. Neural recognition molecules of the immunoglobulinsuperfamily: signaling transducers of axon guidance and neuronal migration. NatNeurosci. 2007 Jan;10(1):19-26. Review. Erratum in: Nat Neurosci. 2007Feb;10(2):263.

L1-deficient mice exhibit dendritic misorientation of cortical pyramidal neurons, smaller hippocampus, abnormal position of dopaminergic neurons, abnormal cerebellar development and deficits in spatial learning and sensorimotor gating (Maness and Schachner, 2007). Schwann cell-axon interactions are also disrupted, leading to abnormal myelination (Herron *et al*., 2009). In humans, mutations of *L1CAM* gene are responsible for severe neurological disorders such as lower limb spasticity, mental retardation, hydrocephalus and flexion deformity of the thumbs. *L1CAM* gene is subjected to different

mutations: most of nonsense and frameshift mutations truncating L1 prior to the TMD can eliminate the surface expression of the protein. Mutation of the cytoplasmic domain are less likely to cause hydrocephalus than those abolishing the extracellular domain: the homophylic binding may be preserved in the absence of the cytoplasmic domain. Mutations at the cytoplasmic domain can eliminate at least part of the conserved Ankyrin-binding domain. Mutations deleting RSLE motif will also abolish trafficking of L1 in differentiated neurons, only allowing transport of protein to the cell soma. However, the most important mutations able to affect *L1CAM* gene function are the missense ones. The majority of these mutations interest residues involved in structural integrity of individual domains in the extracellular region whereas a smaller proportion affects residues affecting surface properties of L1. Interestingly, 50% of the human mutations takes place in the contact regions required for Ig1-Ig4 horseshoe. Missense mutations may lead to three consequences:

- *L1 folding or intracellular trafficking alteration*: mutation of the signal peptide W9S may affect cell surface expression;
- *Cys residue at the protein surface*: Y194C and Y1070C could be affecting L1 function or mobility because of intermolecular disulphides;
- *Effects on ligand binding*: for example missense changes in the L1 extracellular region have variable effects on binding L1 to itself or to the related CAMs TAG-1/Axonin-1 and F3/F11.

CHL1

CHL1 (Close Homolog of L1) is a protein of ~200 kDa encoded by the CHL1 gene located on chromosome 3; it consists of six Ig-like domains and four FnIII domains. CHL1 is involved with L1 in neuron survival and neurite outgrowth. In the brain, CHL1 is expressed in pyramidal cells of the hippocampus and thalamus. In contrast to L1, both the soluble 165 kDa and the transmembrane 180 kDa isoforms of CHL1 can promote neurite outgrowth. CHL1 is able to mediate axon repulsion through the cytoplasmic motif RGGKYSV recruiting Ezrin to the plasma membrane; Ezrin recruitment plays a role in growth cone collapse, neurite outgrowth and branching. CHL1 also recruits Hsc70 to synaptic vescicles for ADP-dependent clathrin uncoating. The loss of CHL1 leads to an accumulation of chlatrin coated synaptic vescicles and the decrease in the production of new vesicles. The abolishment of synaptic vescicle recycling may explain mental retardation and schizophrenia linked to

CHL1 loss and mutation (Herron *et al*., 2009). Morover, CHL1 intracellular domain is implicated in SNARE complex refolding (Sytnyk *et al*., 2017). CHL1 signaling pathway is reported in Fig. 34.

Neurofascin

Neurofascin is encoded by *NFASC* gene located on chromosome 1. Neurofascin is involved in neurite outgrowth and synapse formation and comes in different splicing variants. The two major ones are known as NF155 and NF186. NF155 is the “glial” isoform due to its presence in both oligodendrocytes and Schwann cells whereas NF186 is the “axonal” version, present at the node of Ranvier in neurons (Herron et al., 2009). These two variants exhibit different expression patterns and functions since neurite outgrowth is promoted by NF155 but inhibited by NF186. The ~200 kDa Neurofascin protein is characterized by six Ig-like domains and five FnIII domains; its crystal structure (PDB 3P3Y) reveals that the N-ter Ig-like domains (Ig1-Ig4) form a horseshoe shape (Fig. 35A,B).

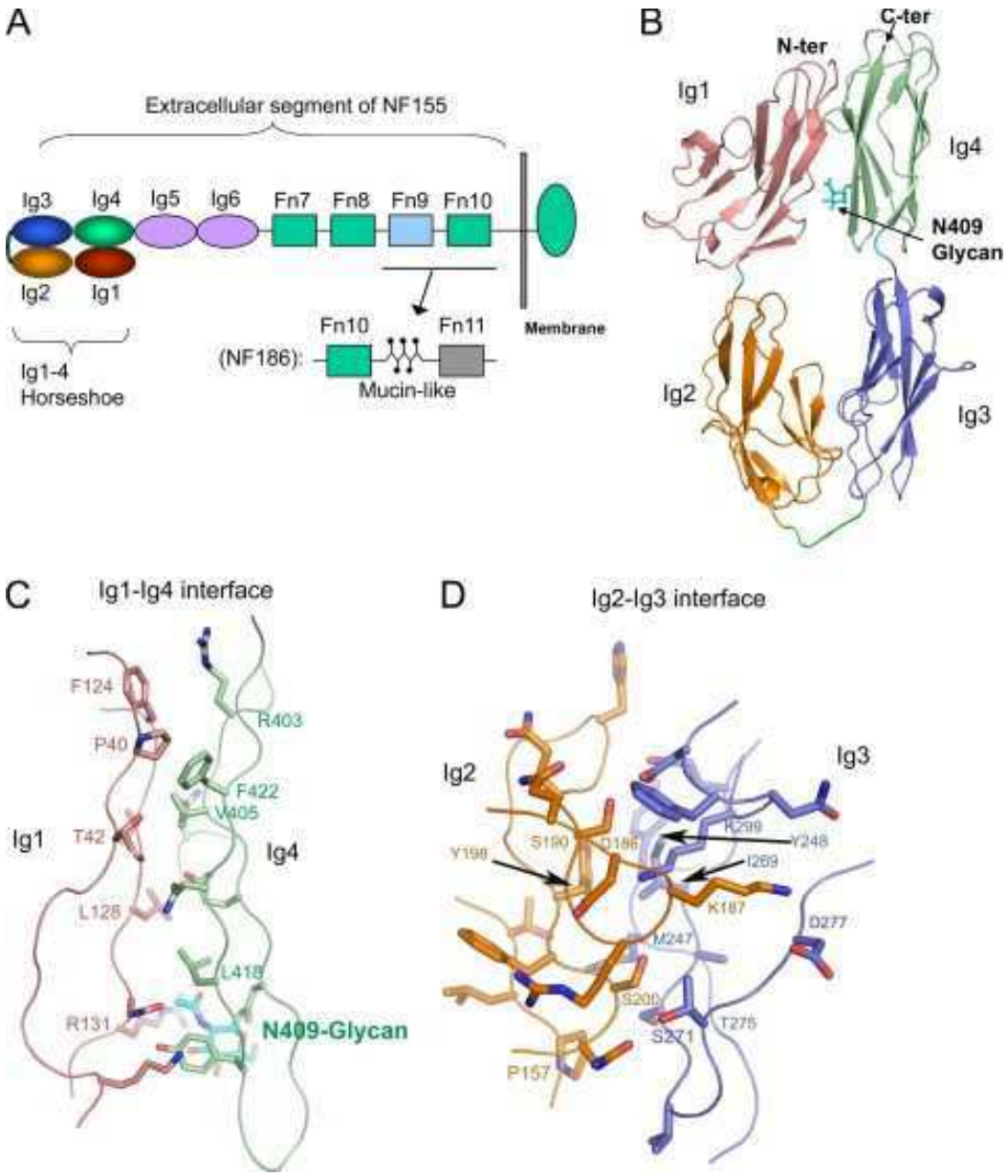


Fig.35. Structure of the horseshoe-shaped neurofascin headpiece. **A**. Diagram of the Neurofascin composition. **B**. Ribbon diagram of an NF_{Ig1-4} monomer, with each Ig domain colored as shown in A. the N-linked glycan is depicted as sticks. **C**. Ig1-Ig4 interface, with the main chain depicted as Cα trace and the side chains as sticks. **D**. Ig2-Ig3 interface. From: Liu H, Focia PJ, He X. Homophilic adhesion mechanism of neurofascin, a member of the L1 family of neural cell adhesion molecules. *J Biol Chem*. 2011 Jan 7;286(1):797-805.

The Neurofascin structure allowed to infer that protein horseshoe monomers interacts with each other forming a dimer in a *trans*-synaptic adhesion way. Ig2, as a part of the horseshoe-shaped headpiece, plays a pivotal role in such interaction. This domain features an intermolecular β-sheet formed by the joining of two individual GFC β-sheets and a large but loosely packed hydrophobic cluster. The orthogonal side-to side adhesion mode of the horseshoe resemble that of L1 as demonstered by cryo-electron tomography studies: paired

horseshoes rather than extended N-ter domain appeared. So, this adhesion mechanism can be easily transferred to other L1 family members. The sequence similarity among L1 family members is quite high: Neurofascin is 41, 56, and 42% identical to L1, CHL1 and NrCAM, respectively. Notably, residues involved in both super β-sheet interactions and hydrophobic cluster are highly conserved among these four members (Liu *et al*.,2011). The Neurofascin cytoplasmic tail is able to interact with many binding partners. Syntenin-1 binds only to Neurofascin via its PDZ domain but not to other L1 subfamily members. The role of Syntenin-1 is as yet inclear, but it seems to have a structural role with the NG2 proteoglycan receptor in migratory oligodendrocyte precursor cells. Neurofascin can also bind Ezrin via its last 28 amino acids, forming a FERM-binding motif. Another FERM-protein, called Willin, is able to bind to Neurofascin and it is thought to be involved in normal cell growth and development. Finally, the Tyr phosphorylation of the FIGQY motif plays a role in tuning Ankyrin and Doublecortin binding. Colocalization of phosphorylated Neurofascin and Doublecortin early occurs in the developing brain. Since Doublecortin can also bind to proteins associated with the cytoskeleton and clathrin adaptor, it could be involved in the trafficking of Neurofascin. Neurological mutations occurring within Doublecortin may abolish the binding to the phosphorilated form of Neurofascin (Herron *et al*., 2009).

NrCAM

The *NRCAM* gene is located on chromosome 7q31.0 and it undergoes AS: the most common isoform in the brain is charachterized by four, rather than five, FnIII repeats. *NrCAM* is expressed in both neurons and glial cells in the developing and adult nervous system. In the mouse CNS, it is expressed in specialized glial formations in the ventral midline throughout the nervous system, especially on crossing fibers present in these areas. *NrCAM* expression is also found in decussating pathways (anterior commissure, corpus callosum, posterior commissure) and in non decussating pathways (lateral olfactory tract, habenulo-interpeduncular tract). Moreover, *NrCAM* is expressed in Schwann cells and in the cortex at a time point when NrCAM-positive thalamic axons are growing into this region. Together with PTPRZ, NrCAM is expressed in adult white matter progenitor cells from human brain.
The ~200 kDa NrCAM protein exhibits six Ig-like domains but several amino acid insertions/deletions can occur in the regions upstream of the Ig domains, between Ig domains and between the Ig domains and the Fn type III repeats (Fig.36.).

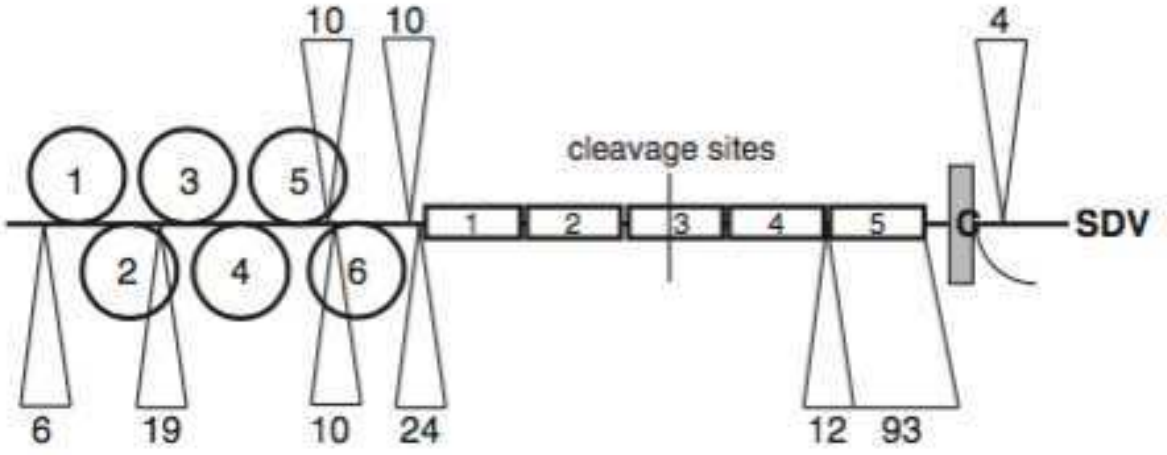


Fig.36. Schematic view of the NrCAM structure; Ig-like domains and Fn type III repeats are represented by numbered ovals and squares, respectively. Splicing insertions/deletions are shown by triangles and number of amino acids is indicated. Possible cleavage sites are in the middle of the third FnIII repeat. C in the TMD (grey box) region is Cys, a possible fatty acid modification site. At the C-ter, a PDZ domain binding site is shown as SDV. From: Sakurai T. The role of NrCAM in neural development and disorders-beyond a simple glue in the brain. Mol Cell Neurosci. 2012 Mar;49(3):351-63

In vitro evidence seems to support NrCAM activity in neurite outgrowth. This function becomes effective if NrCAM can interact with Contactin-1 (CNTN1), Contactin-2 (CNTN2) and L1 as revealed by studies on chick retinal explants. Moreover, NrCAM works as a receptor for outgrowth when NrCAM interacting molecules are used as substrate such as CNTN1, CNTN2, PTPRZ, Neurofascin.

NrCAM-mediated neurite outgrowth downstream signaling pathways are still unclear; however, *in vivo* evidence suggests that NrCAM-mediated interactions may be more relevant to axon guidance than to axon growth. As shown in Fig. 45, NrCAM plays important roles in brain wiring including cerebellar granule cell development, axon entry at the dorsal spinal cord and axon guidance at the ventral spinal cord, optic chiasm formation, and the formation of the thalamocortical projection. Both the L1 family of CAMs and Contactin family of CAMs are expressed on cerebellar granule cells in a temporally regulated manner, supporting the idea that specific CAMs may play distinct roles in the process of cerebellar granule cells development. NrCAM is also involved in C cell axons guidance at the CNS midline. Axon guidance can occur thanks to the interaction between positive and negative cues.

1.2 Contactins (CNTNs) subfamily

The contactin (CNTN) family consists of six structurally related members, sharing average 40-60% sequence identity: CNTN1 (F3/F9), CNTN2 (TAG-1), CNTN3 (BIG-1), CNTN4 (BIG-2), CNTN5 (NB-2), CNTN6 (NB-3). All Contactins are attached at the cell membrane via a GPI anchor and they can be available in both membrane-bound and soluble form. The extracellular part contains six Ig-like repeats followed by four FnIII domains (Fig. 37):

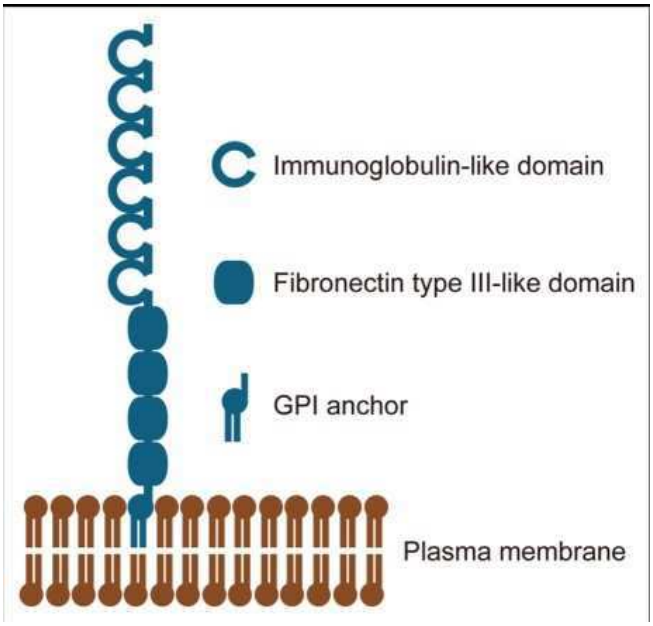


Fig.37. Contactins domain architecture. From: Shimoda Y, Watanabe K. Contactins: Emerging key roles in the development and function of the nervous system. *Cell Adhesion & Migration*. 2009;3(1):64-70.

CNTN1 and CNTN2 proper expression is essential for normal cerebellar morphogenesis. In the postnatal cerebellum, CNTN1 is expressed on migrating granule cells and CNTN2 on premigratory granule cells in the inner part of the external granule cell layer. Moreover, CNTN1 enhances voltage-gated sodium channel at nodes of Ranvier.

Little is known about CNTN3. This protein is abundantly expressed in the adult brain. Expression of CNTN3 is restricted to cerebellum Purkinje cells, hippocampal dentate gyrus granule cells and neurons in the superficial layers of the cerebral cortex.

In mice, CNTN4 expression increases after birth and reaches a maximum in adulthood. CNTN4 is expressed in different substes of neurons in different brain regions, including olfactory system.

Expression of CNTN5 becomes apparent after birth and reaches a maximum around a postnatal day (P) 14.

Expression of CNTN6 in the cerebrum is evident after birth, reaches a maximum around P7 and declines thereafter to low levels. (Shimoda and Watanabe, 2009).

1.3 DCC Netrin Receptor

Netrin-1 is a guidance cue able to interact simultaneously with two DCC molecules thanks to a DCC specific site and a unique generic receptor binding site, where sulfate ions staple together positively charged patches on both DCC and Netrin-1. Netrin-1 can act bifunctionally, triggering either attraction or or repulsion effects on migrating axons, depending on the receptor types exhibited on the growth cone. Netrins (Fig.38.A) consist of a laminin VI domain, a V domain containing three EGF repeats and a C-ter netrin-like domain. Netrin-1 binding to DCC induces chemoattraction whereas binding to UNC5 is responsible for repulsion. DCC (Fig.38.A) consists of four N-ter Ig-like domains, forming a horseshoe conformation. These domains are followed by six FnIII domains, a single transmembrane segment and a large cytosolic portion containing three highly conserved motifs called P1, P2 and P3. The absence of Netrin-1 leads to apoptosis. The ability of Neterin-1 to link two DCC receptor together enables the dimerization via P3 motif of the cytosolic domains of DCC. This recruits an intracellular signaling complex that leads to the release of calcium, kinase activation and cytoskeleton rearrangement.
The structure of the human Netrin-1/DCC complex (Fig.38.B) reveals two binding sitrs for DCC on the V domain of Netrin-1. The two DCC_{FN56} fragments are not directly linked and solution studies indicate different kinetics for these two sites. These data suggest a modular binding mode for the two Netrin-1 DCC_{FN56} binding sites. Binding of DCC and Netrin-1 is facilitated by sulfate ions and a chloride ion embedded onto the Netrin-1 surface, able to neutralize positively charged patches on both Netrin-1 and DCC. The chloride ion is coordinated by four residues from the Netrin-1 molecule and sulfate ions can interact with one of five Arg residues located closely together of Netrin-1. The embedded ion binding sites may provide the necessary flexibility to incorporate linear heparan sulfates, able to mediate receptor binding in Netrin-1. Fig.39.A depicts a model where one Netrin-1 molecule binds two DCC molecules along the V domain to form a signaling unit, and the DCC molecule that occupies the DCC specific binding site 1 on Netrin-1 engages another Netrin-1 molecule via Fn4 domain to stich different Netrin/DCC signaling units together. The DCC specific binding site 1 acts as an anchor. Whereas the generic receptor binding site is

occupied by DCC, UNC5 or another receptor (Li *et al*., 2014). Fig.39.B displays a model of Netrin-1 attraction or repulsion on the basis of heparan sulfate bound to Netrin-1.

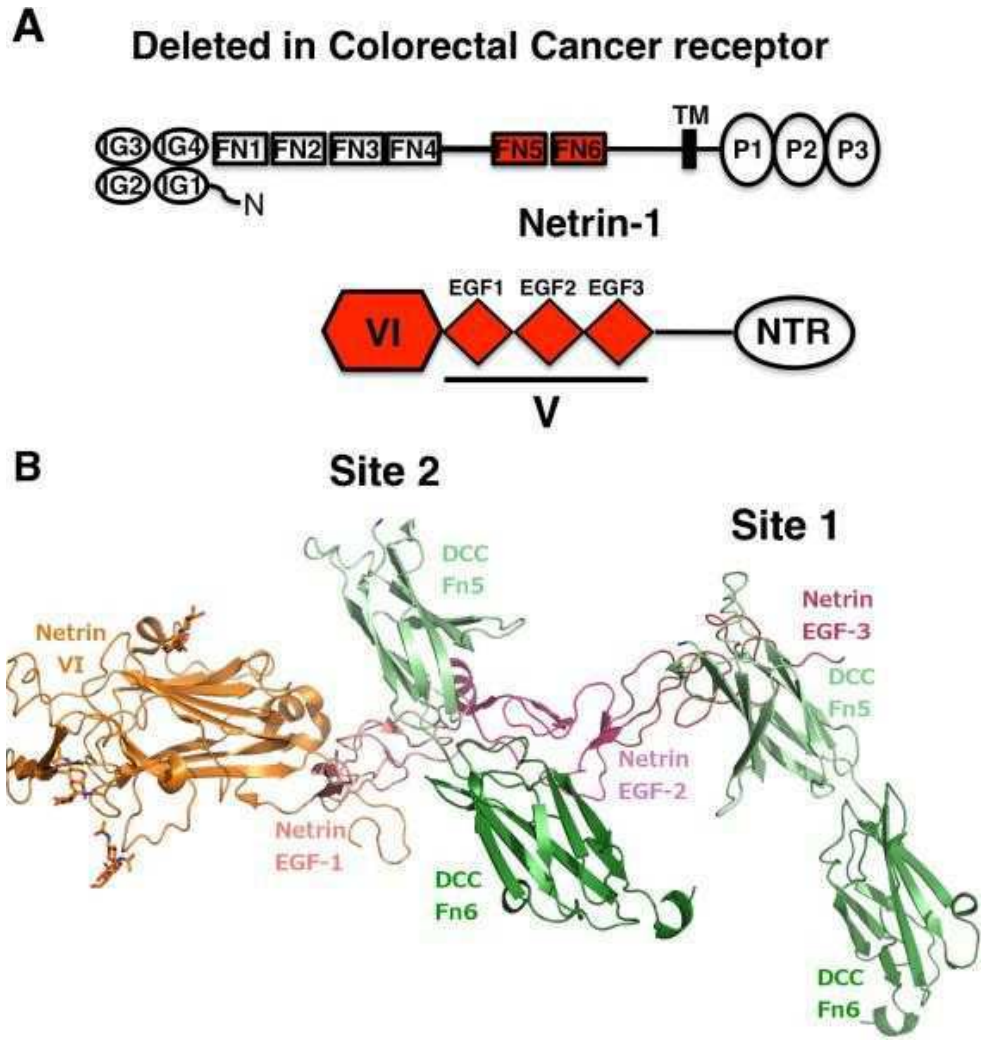


Fig.38. Domain architecture and crystal structure of the Netrin-1/DCC complex. **A**. Domain diagram of DCC and Netrin-1, with the domains present in the crystal structure colored red. **B**. Ribbon diagram of the $Netrin_{VI-V}/DCC_{FN56}$ complex is depicted showing two DCC fragments (in green) bound to the V domain (in salmon red) of one Netrin-1 molecule. Glycosylation sites on the VI domain of Netrin-1 (in orange) are shown as sticks. From: Finci LI, Krüger N, Sun X, Zhang J, Chegkazi M, Wu Y, Schenk G, Mertens HD, Svergun DI, Zhang Y, Wang JH, Meijers R. The crystal structure of netrin-1 in complex with DCC reveals the bifunctionality of netrin-1 as a guidance cue. *Neuron*. 2014 Aug 20;83(4):839-49.

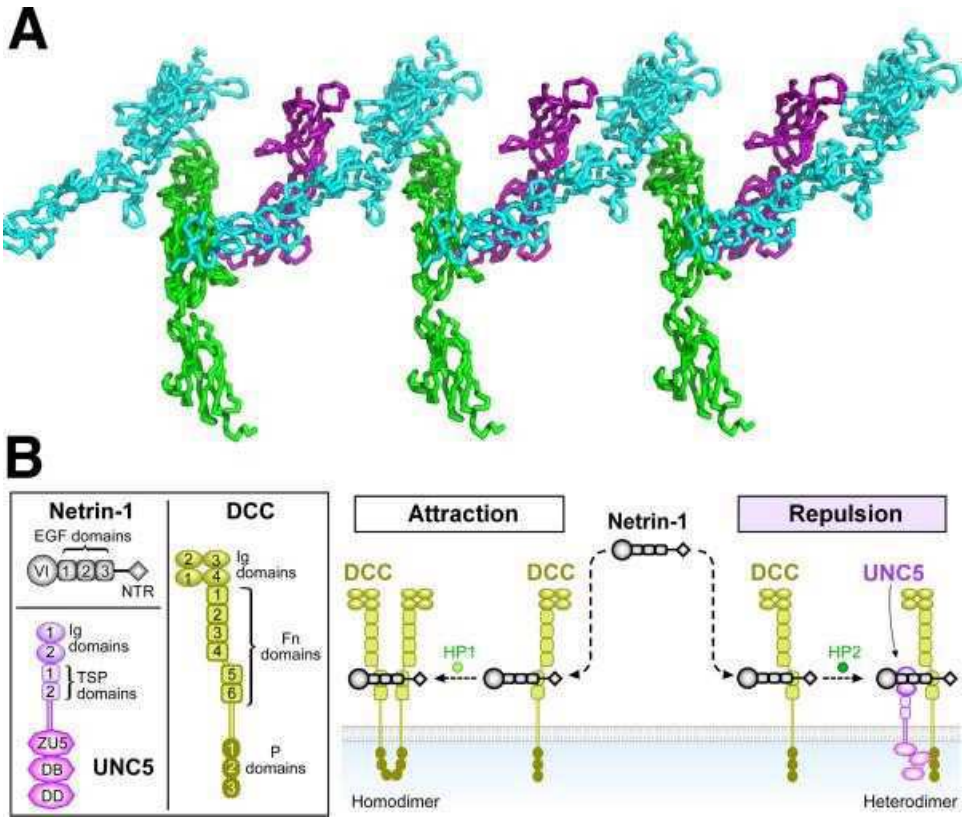


Fig.39. Model of heparan-sulfate-dependent formation of the DCC/DCC and DCC/UNC5 complex with Netrin-1. **A**. Composite model of an extended Netrin-1/DCC cluster based on the superimposition of the crystal structure presented here with the crystal structure of $Netrin_{VIV}$ in complex with the FN4 and FN5 domains of DCC. The $Netrin_{VIV}$ molecules are colored in cyan, the DCC molecules (FN4-FN5-FN6) occupying binding site 1 are colored green, and the DCC molecules (shown only FN5-FN6) occupying site 2 are colored purple. **B**. DCC binds specifically to the EGF-3 domain of Netrin-1 (binding site 1). Netrin-1 associates with a heparan sulfate molecule that is selective for DCC (HP1) or UNC5 (HP2). When heparan sulfate HP1 is bound to Netrin-1, a second DCC molecule is recruited to binding site 2, and the cytosolic P3 domain of the two DCC molecules associate to form a signaling complex, leading to attraction of the growth cone. When heparan sulfate HP2 binds to Netrin-1, UNC5A is recruited to the EGF-1/EGF-2 domains, and the cytosolic P1 domain of DCC and the region between the ZU5 and DB domain of UNC5A form a signalin complex, leading to repulsion of the growth cone. From: Finci LI, Krüger N, Sun X, Zhang J, Chegkazi M, Wu Y, Schenk G, Mertens HD, Svergun DI, Zhang Y, Wang JH, Meijers R. The crystal structure of netrin-1 in complex with DCC reveals the bifunctionality of netrin-1 as a guidance cue. *Neuron*. 2014 Aug 20;83(4):839-49.

1.4 Roundabout (ROBO) receptors subfamily

As for Netrin-1 and its receptor DCC, also Slit proteins and their ROBO receptors (Fig.40.) are involved in commissural axon developing at the ventral midline. Here, a finely regulated balance of specific cues is able to influence growth cones whether to cross. These guidance cues can be both attractive and repulsive (Brose *et al*., 1999). Slit-ROBO signaling complex is also central to the development of blood vessels and some organs, such as the heart. Three Slit proteins (Slit1-3) and four ROBO proteins (ROBO1-4) have been identified in mammals. Netrin and Slits are secreted by the midline cells, whereas DCC and ROBO1-3 are expressed on the surface of growing axons. ROBO1-3 receptors exhibit an ectodomain of five Ig-like domains (Ig1-Ig5) and three FnIII domains, whereas ROBO4 has only two Ig domains. Interaction between Slits1-3 and ROBO1-3 is mediated by the second Slits LRR domain and ROBOs first two Ig domains (Fig. 41) (Morlot *et al*., 2007). *In vivo* and in cell colture, Slit2 is cleaved into 140 kDa N-ter (Slit2-N) and 55-60 kDa C-ter (Slit2-C) fragments (Nguyen *et al*., 2001). Netrin-1 and Slit2 interact *in vitro* with an affinity that is comparable to the affinities of Slit2 for ROBO proteins and of Netrin-1 for its high-affinity receptors. This binding could be explained by the Slit1-3 and Netrin-1 coexpression both in the floorplate and in different locations throughout the nervous system (Brose *et al*.,1999).

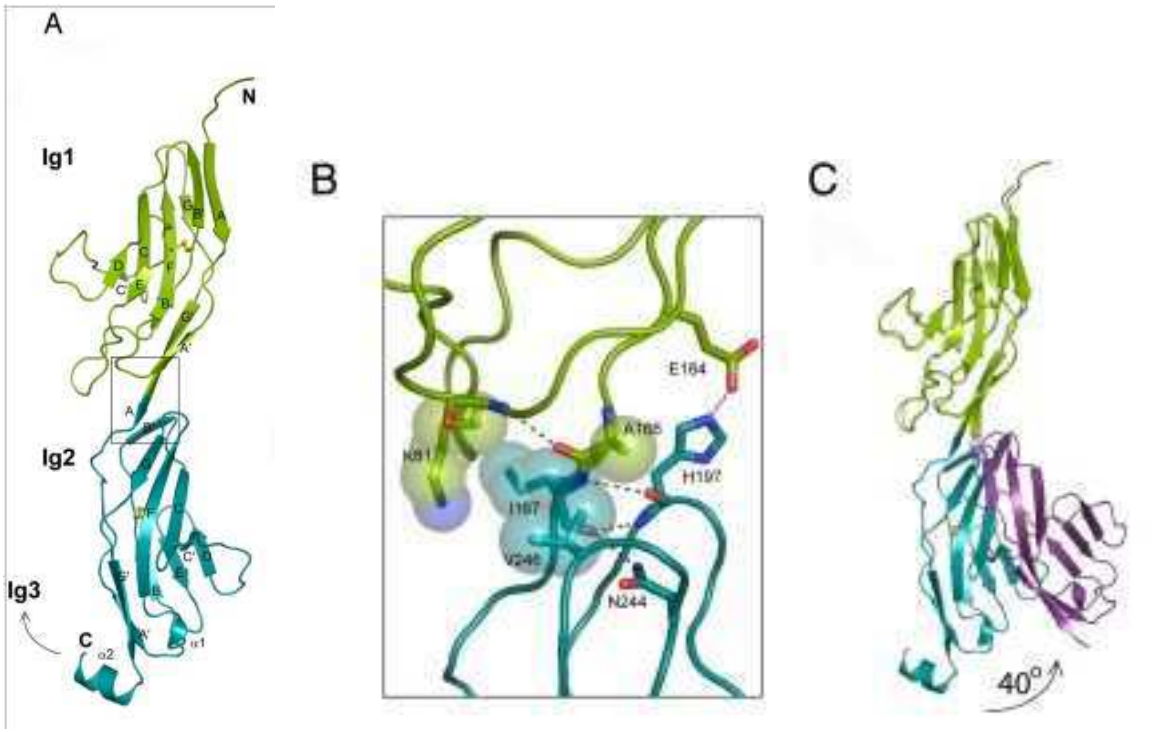


Fig.40. Structure of human ROBO1 Ig1-2. **A**. Ribbon diagram. The disulfide bridges are in yellow and the box indicates the region highlited in B. **B**. Residues involved in interdomain contacts of the Ig1-Ig2 interface. C. Ribbon diagram of the two Ig1-2 crystal form showing the hinge movement of Ig2. Adapted from: Morlot C, Thielens NM, Ravelli RB, Hemrika W, Romijn RA, Gros P, Cusack S, McCarthy AA. Structural insights into the Slit-Robo complex. *Proc Natl Acad Sci USA*. 2007 Sep 18;104(38):14923-8.

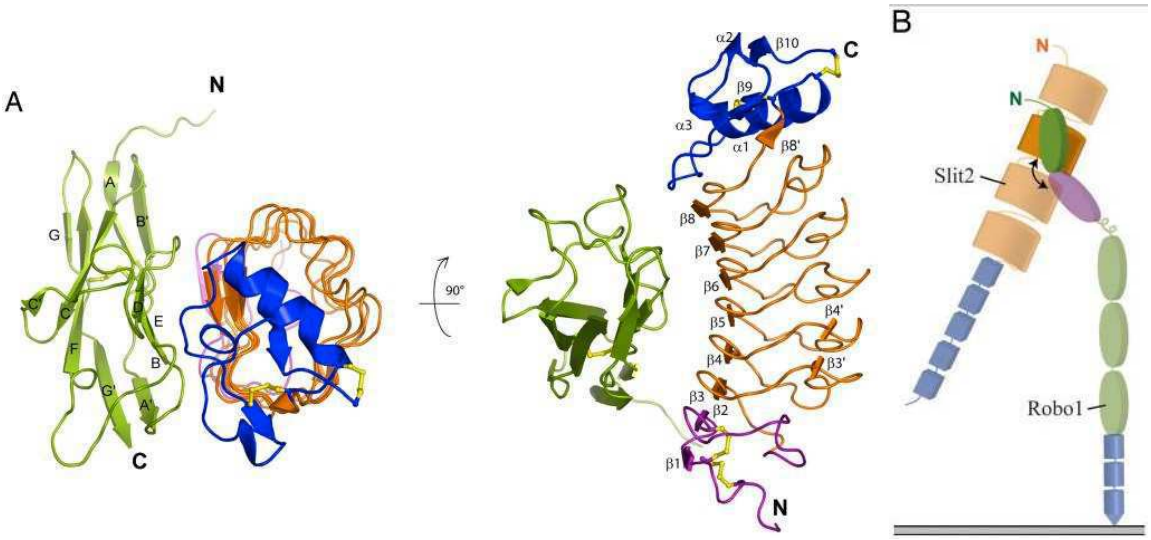


Fig.41. Structure of Slit2 D2 bound to ROBO Ig1. Ig1 is in green; Slit2 D2 N- and C- terminal caps are in purple and blue, respectively. LRR 1-6 are in orange and the disulfide bridges are in yellow. **A**. Ribbon diagram of the complex in two orthogonal orientations. **B**. Schematic of the Slit2-ROBO1 domain organization with the flexible linkage marked by a curved arrow. The ROBO1 Ig1 domain is shown in green, and the Slit2 D2 is in orange. All other domains are opaque. The ROBO1 Ig2 is in magenta, the other Ig domains are green and the FnIII domains are in blue. The Slit2 LRR domains are colored in orange and the EGF domains are in blue. Adapted from: Morlot C, Thielens NM, Ravelli RB, Hemrika W, Romijn RA, Gros P, Cusack S, McCarthy AA. Structural insights into the Slit-Robo complex. Proc Natl Acad Sci US A. 2007 Sep 18;104(38):14923-8.

1.5 LINGO subfamily

The LINGO subfamily comprises four paralogs: LINGO1, LINGO2, LINGO3 and LINGO4. Crystal structure is available only for LINGO1 (2ID5) and few literature is disposable on LINGO1 and LINGO2 whereas LINGO3 and LINGO 4 have not been characterized yet. These proteins share twelve extracellular LRRs, an Ig like domain and a short intracellular tail.

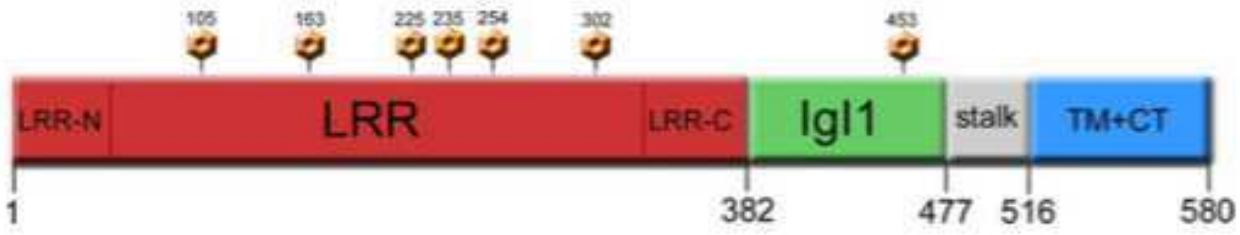


Fig.42. Schematic of the domain location of LINGO1 showing the N-terminal LRR module in red (1-382; LRR-N and LRR-C label the N- and the C-cap, respectively), the Ig1 domain in green (383-477). The stalk region in gray (478-516) and the transmembrane domain (TM) plus cytoplasmic (CT) C-terminal tail in blue (517-580). The circles respresent the occupied N-linked glycosylation sites identified in the three-dimensional structure. From: Mosyak L, Wood A, Dwyer B, Buddha M, Johnson M, Aulabaugh A, Zhong X, Presman E, Benard S, Kelleher K, Wilhelm J, Stahl ML, Kriz R, Gao Y, Cao Z, Ling HP, Pangalos MN, Walsh FS, Somers WS. The structure of the Lingo-1 ectodomain, a module implicated in central nervous system repair inhibition. *J Biol Chem*. 2006 Nov 24;281(47):36378-90.

LINGO1 (Fig. 42) is encoded by the *LINGO1* gene mapped to chromosome 15q24 (Zhou *et al*., 2012). This protein is a component of of the Nogo receptor complex (NgR), composed of the Nogo-66 receptor, p75 (or TROY) and LINGO1 (Fig. 43). This complex is responsible for a downstream RhoA-dependent signaling pathway able to inhibit neurite outgrowth. This explains why injuried neuron axons are unable to regrow in mature organisms after a Central Nervous System (CNS) damage. Infact, truncated LINGO-1 lacking the intracellular domain restores neurite outgrowth *in vitro* (Mosyak *et al*., 2006). Morover, via Fyn-RhoA signaling pathway LINGO1 acts as a negative regulator of oligodendrocyte maturation and myelination. Loss- and gain of function experiments allowed to infer that LINGO1 inhibitory signaling could be one of the factors controlling CNS myelination. Infact, inhibition of LINGO1 activity leads to outgrowth of oligodendrocyte processes and highly developed myelinated axons (Mosiak *et al*., 2006). LINGO1 is also involved in EGFR signaling pathway, acting as a negatively regulator via accelerating EGFR internalization and degradation. After binding with its ligands, EGFR forms dimers, and is phosphorilated and activated. The activated EGFR sends a signal to PI3K which in turn phosphorylates Akt, being responsible of cell survival and proliferation. The EGFR signaling pathway promotes proliferation, survival, migration of neuronal stem cells and also provides protection to post-mitotic cells

in vivo and *in vitro* against injuries. The lack of the EGFR can induce neurodegeneration in transgenic animals. The expression level of LINGO1 is higher in the substantia nigra of Parkinson's desease (PD) patients than in age-matched controls and in animal PD models after neurotoxic damage. Increase in LINGO1 expression might negatively regulate Purkinje neurons viability due to downregulation of EGFR-PI3K-Akt signaling. Both in vivo and in vitro data suggest that LINGO1 inhibition enhances and protects neurite growth of midbrain neurons (Zhou *et al*., 2012).

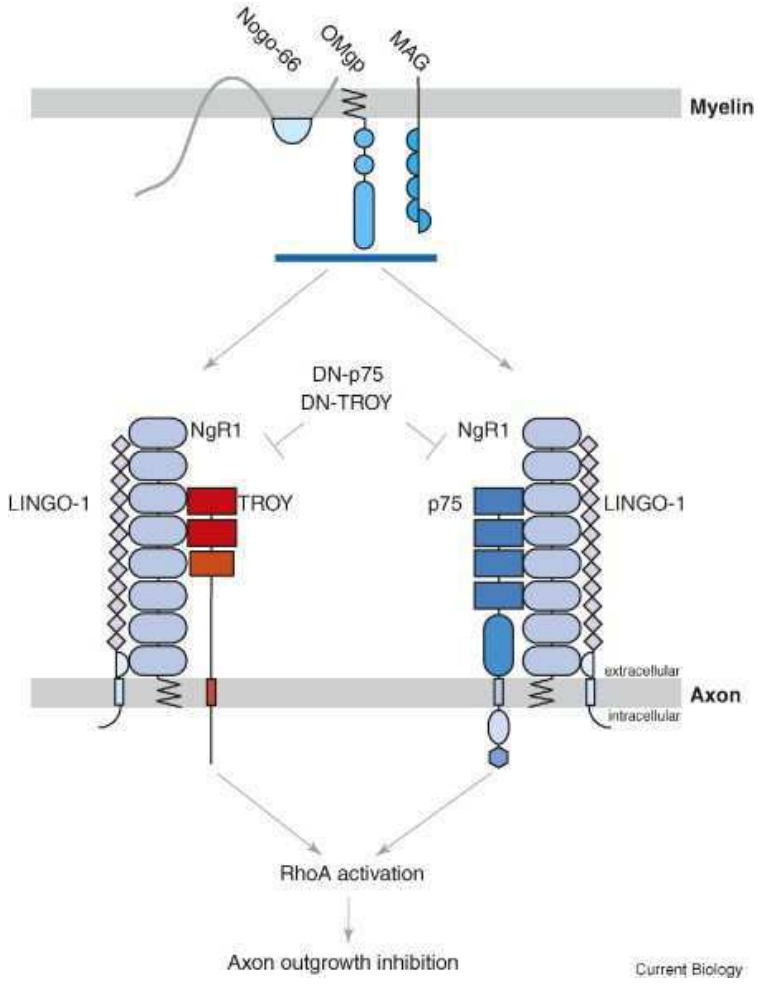


Fig.43. The myelin associated inhibitor factors Nogo-66, MAG and OMgp block regeneration of azons by binding to a shared receptor NgR1. In addition to an interaction between NgR1, LINGO1 and p75, a newly identified receptor complex consisting of NGR1, LINGO1 and TROY can transduce signaling upon binding of myelin associated inhibitor factor to NgR1, leading to RhoA activation and axon outgrowth inhibition. Blocking the formation of these receptor complexes by addition of dominant negative (DN) forms of either p75 or TROY (DN-p75, DN-TROY) antagonizes the axon outgrowth inhibitory effect of myelin associated inhibitor factor and myelin leading to greatly improved neurite outgrowth of DRG and cerebellar granule neurons *in vivo*. From: Mandemakers WJ, Barres BA. Axon regeneration: it's getting crowded at the gates of TROY. *Curr Biol*. 2005 Apr 26;15(8):R302-5.

From a structural point of view, the N-ter LRR module (residues 1-382) is an elongated arc with 15 parallel β-strands on the concave face and mostly irregular extended structures on the convex face. Each LRR begins with a β-strand and loops back by virtue of this repeated motif: XL^{2} XXL^{5} XL^{7} XXN^{10} XL^{12} XXL^{15} $XXXXF^{20}$ XXL^{23} X, where *X* can be any amino acid; **L** are hydrophobic residues (Leu, Ile, Val, Met, Phe, Thr); **N** includes mostly Asn, Cys, Asp, Leu, Trp; **F** represents Phe, Leu. The consensus residues at the indicated positions make up the interior of the LRR domain. The hydrophobic core of this LRR structure has a high occurrence of aromatic rings (Phe^{342}, Phe^{350}, Phe^{362}, Phe^{368}, Phe^{371}, Tyr^{379}, Phe^{380}). A short loop (residues 349-354) is integrated into the concave LRR structure with the apex Cα (Arg^{352}) bulging away from the β-sheet to the concave space (Fig.44.). Around this β-bulge structure there are an exposed Trp (Trp346) and a protruding Arg (Arg352). These two residues are likely part of binding epitopes, as Trp and Arg are frequently encountered in protein-protein recognition sites. The LRR module is followed by a single Ig-like domain (residues 383-477) exhibiting high structural homology to the Ig3 NrCAM domain. Superposition of these modules (Fig. 44) gives an r.m.s.d of 1.58Å and ~30% residue identity. Fig. 44 also highlits a cleft (22Å deep and 35Å long) originated from the arrangement of LRR and Ig-like domains, showing a 90° angle between them. This cleft extends on a glycan free surface and may be suitable for binding. The sides of the cleft are made up by repeats 10-12 plus helix α1 shaping one side and the concavity of the β-sheet A'CC'FG shaping the other one. The LRR face is polar and charged, whereas the CC'FG face is predominantly hydrophobic. Hence, the two modules appear not to interact with each other directly.

LINGO1 is a glycan decorated molecule. However, the glycan disposition is quite unexpected. It is known that the presence of extensive glycosylation inhibits molecule binding and binding surfaces are predicted to be glycan-free. In LINGO1, out of four LRR faces, only the convex surface is free of carbohydrates, whereas the concave and two major side surfaces are glycosylated. Glycans are retrieved also in the Ig-like domain, mapping to the A'CC'FG face.

As both crystals and solution studies revealed, LINGO1 cames in a ring-like tetrameric form (Fig. 45), not observed in LRR proteins before. It is thought that the protein rotation axis lies normal to the cell surface, the curved LRR domains lie horizontally, back-to-back, whereas the Ig-like C-ter ends extend vertically as if to continue toward the membrane. The total buried area within the tetramer is ~9.200$Å^2$ and extensive contacts in the interface support the finding that LINGO1 can exist as a tetramer, able to remain stable over a wide pH range

and at a very low ionic strength. Moreover, chemical cross-linking, gel filtration cromatography, dynamic light scattering and analytical ultracentrifugation experiments indicated the tetrameric form of LINGO1. These studies suggest that tetramer formation is not a consequence of crystal packing, but reflects tetramerization of LINGO1 in solution.
Analyses such as evolutionary conserved sequences, electrostatic surface potentials, carbohydrateexposure and comparison between common characteristics emerged from other LRR structures allowed to search for LINGO1 possible binding sites. LINGO1 exhibits a high degree of evolutional conservation, with 92.7-99.8% extracellular sequence identity among human and homologous monkey, mouse, rat, and chicken. Conserved patterns, with much of the concave face, the self-recognition motifs and glycosylation motifs identify surfaces that may be important for ligand binding, oligomerization or the structural integrity of folding topology. Electrostatic analysis on LINGO1 revealed that upon tetramerization, ~4000Å^2 of hydrophobic surface area become buried, which by itself can be a driving force for assembly: infact, the exposure of hydrophobic patches on a protein surface is in general energetically unfavorable. So the electrostatic component seems to be significantly involved in the interactions between LINGO1 and its non-self ligands due to the presence of few solvent-exposed hydrophobic residues. Moreover, electrostatic analysis showed that LINGO1 specificity may largely depend on its oligomeric structure rather than on individual binding sites. Infact, a positive charged remarkably large area (V-shaped positive potential) could constitute an essential binding site for acidic p75 (Fig. 55B, Mosyak *et al*., 2006). The interior of the ring exhibits a net positive potential, due to ripetitive clusters of Arg and His residues. A conspicuous conserved His ladder and a continuous area of negative charge (Fig. 45A) due to the presence of Glu and Asp characterized the LRR concave faces. In LINGO1 the corresponding ABDE side of the Ig-like module is not involved in the tetramer formation (Fig. 45B), so it is available for other interactions and suggesting that this module, like that of LRR, may play an integrated role in oligomer formation and the recognition of a co-receptor. The ABDE side presents mostly charged, highly conserved residues. The spacing between the Ig-like domains in the tetramer (~65Å between adiacent monomers) (Fig. 45B) seems appropriate, allowing each to be competent to bind ligand (Mosyak *et al*., 2006).

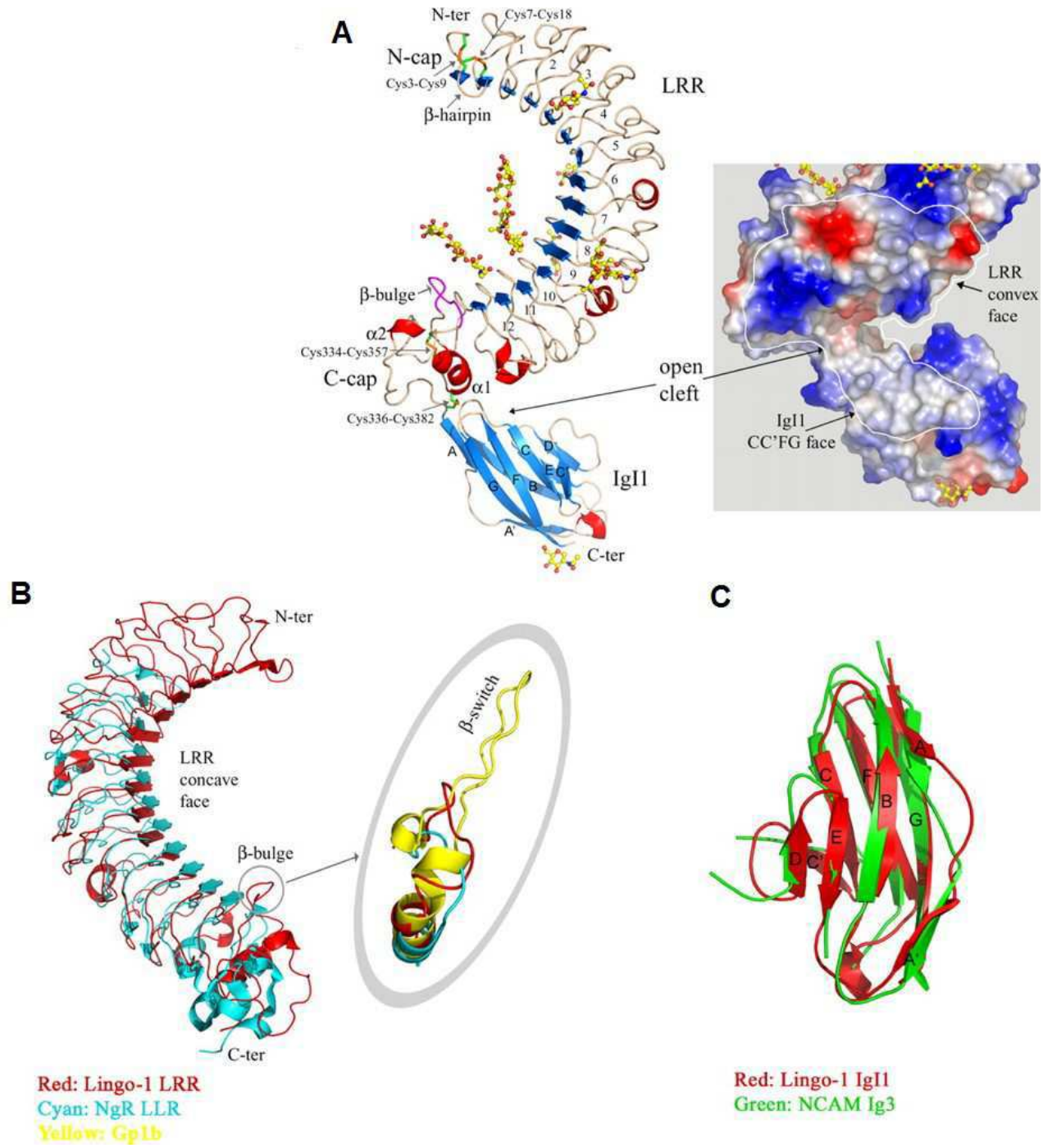


Fig.44. The protomer structure of LINGO1. **A**. Ribbon diagram showing the overall architecture of the LINGO1 monomer, colored according to secondary structure: beige, coil; blue, β strand; red, α-helix. Disulfide bonds are shown in green, and the N-linked carbohydrates are yellow. LRRs are numbered. Selected loops, α helices and β strands of the Ig-like domain are labeled. To the right is a close-up view of the cleft surface, marked with a withe line and colored by electrostatic potential (red for negative; blue for positive) to emphasize different chemical properties of the opposite surfaces. **B**. Superposition of the LRR structure of LINGO1 (red) and NgR (cyan). The small circle marks the location of the β-bulge in the LINGO1 structure; comparison (right) to the segment of glycoprotein Gp1bα is marked in yellow. **C**. Superposition of Ig-like domain of LINGO1 (red) with the Ig3 module of NrCAM (green). The view is from the face of the β sheet ABDE. Adapted from: Mosyak L, Wood A, Dwyer B, Buddha M, Johnson M, Aulabaugh A, Zhong X, Presman E, Benard S, Kelleher K, Wilhelm J, Stahl ML, Kriz R, Gao Y, Cao Z, Ling HP, Pangalos MN, Walsh FS, Somers WS. The structure of the Lingo-1 ectodomain, a module implicated in central nervous system repair inhibition. *J Biol Chem*. 2006 Nov 24;281(47):36378-90.

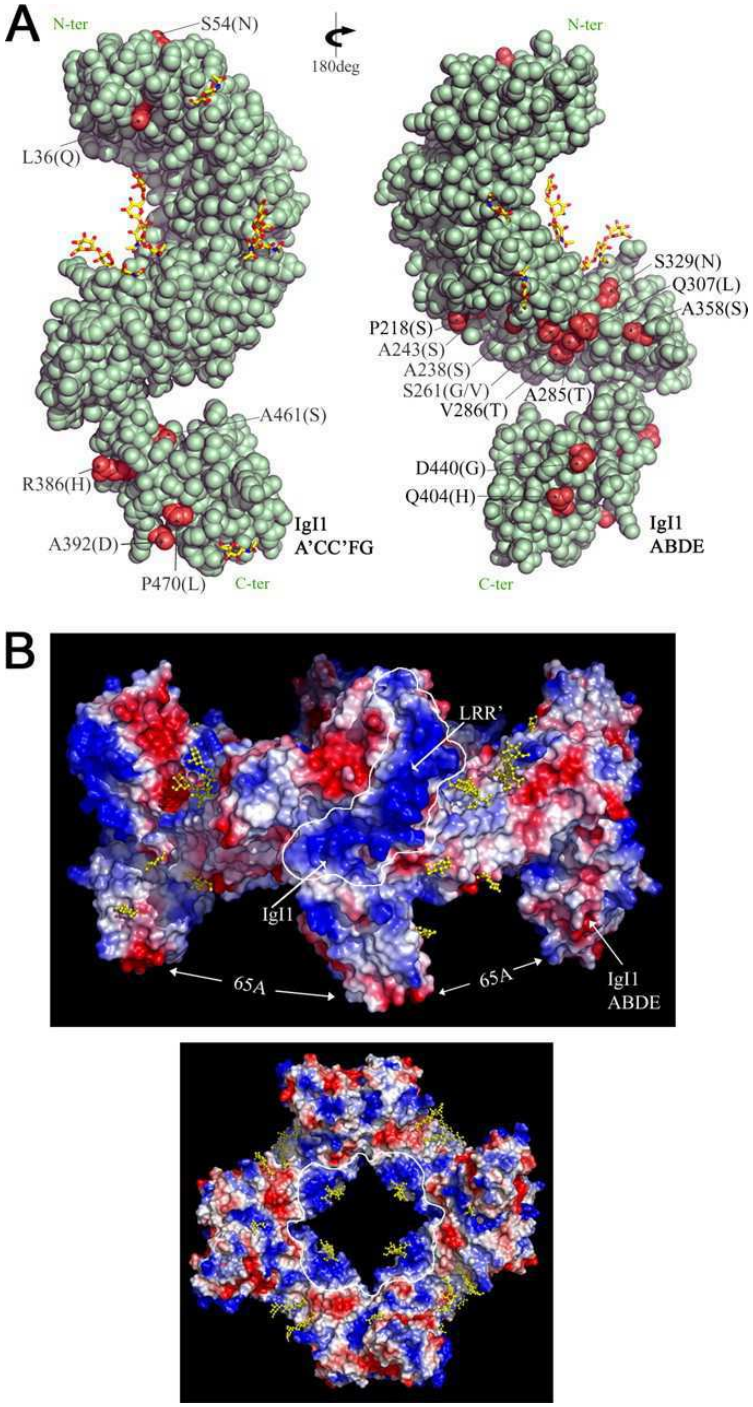


Fig.45. Sequence conservation and electrostatic surface potential. A. space-filling model showing the location of evolutionary mutations on the LINGO1 structure. The two side views rotated 180° highlit residues that vary in red (labeled) and conserved (green). The highest amino acid sequence variations (human and chicken sequences; 92.7% identity) map to the convex area of LRR and some of the regions of the Ig-like domain. Substitutions that do not change the chemical or aromatic character of the amino acid (e.g. Arg to Lys, Ser to Thr, or Tyr to Phe) are not included. B. Electrostatic potential of the tetramer. View of the front and top surfaces of the LINGO1 tetramer, with coloring ranging from dark blue (most positive) to red (negative). White lines delineate composite surfaces: the V-shaped blue potential (combined from LRR and the Ig-like domain) and the local positive potential inside the ring. From: Mosyak L, Wood A, Dwyer B, Buddha M, Johnson M, Aulabaugh A, Zhong X, Presman E, Benard S, Kelleher K, Wilhelm J, Stahl ML, Kriz R, Gao Y, Cao Z, Ling HP, Pangalos MN, Walsh FS, Somers WS. The structure of the Lingo-1 ectodomain, a module implicated in central nervous system repair inhibition. *J Biol Chem*. 2006 Nov 24;281(47):36378-90.

Results and discussion

Bioinformatic predictions on features and interactions on the conserved motif involved in neurite outgrowth and axon guidance

1. Conserved motif involved in neurite outgrowth and axon guidance

Amongst (several) players involved in neuronal differentiation and nervous system development, the ectodomains from several Cell Adhesion Molecule (CAM) and Extracellular Matrix (ECM) proteins provide neurite outgrowth and guidance (NOG) signals via homo- or hetero-philic interactions.

The modular architecture of such ectodomains often includes a number of N-ter Ig-like domains (6 repeats in neuronal L1 family CAMs: L1CAM, Neurofascin, NrCAM, CHL1).

The solved structure of a Neurofascin homodimer shows that Ig1:Ig4 and Ig2:Ig3 intramolecular binding results in two 'horseshoe' structures binding each other via the Ig2 surfaces. Such structure and binding mechanism is likely conserved also in other CAMs, including L1CAM.

The general mechanism of Ig2-Ig2 interaction is further supported by evidence that mutations in L1CAM Ig2 residues may result in severe neurological diseases such as the CRASH syndrome (see the introduction section for more details on literature).

We recently reported that peptide L1-A (identified by Zhao *et al*., 1998) can be used as a biomimetic tool for neural regenerative medicine as it mimics the L1CAM ability to improve neuritogenesis and neuronal differentiation (Scapin *et al*., 2015 and 2016). We also found that a peptide from the single Ig domain of the ECM protein LINGO1, representing a sequence and structural equivalent for L1-A in Ig2, is biomimetic as well (Scapin *et al*., 2015 and 2016).

Given that L1CAM and Lingo1 ectodomains show a really different architecture and function and the two proteins only share the involvement in NOG, this prompted us to investigate on the possible conservation in further ectodomains of a NOG motif coding for homo/heterophilic binding events that are crucial to nervous system development.

Furthermore, we were aimed at investigating on molecular mechanisms underlying the biomimetic activity of such motif, in terms of capacity to mimic the ectodomain (or to act via a different path) as well as to exploit its potential in designing agonist synthetic peptides as regenerative medicine and eventually therapeutic tools.

The two biomimetic peptides developed so far belong to L1CAM Ig2 domain and to Lingo1 single Ig domain; however, the extracellular domain of L1CAM and other CAMs are endowed with multiple Ig-like repeats (4 to 6).

Therefore, we were wondering to investigate on sequence and structural relationship among such domains.

When using Ig2 (from L1CAM or other CAMs involved in the NOG regulation) as blastp sequence queries against the UniprotKb animal proteome, other Ig2 sequences were found to be highest score hits, i.e., each Ig2 was found to be more similar to Ig2s from other CAMs than to other Igs (Ig1, Ig3-6) from the same extracellular domain.

Then, available PDB structures for CAM Igs were used for measuring the r.m.s.d. of superposed Ig twins in possible combinations (Fig. 61).

Once again and confirming the highest sequence similarity, Ig2 domains were found to be structurally closer each other than to other Igs from the same extracellular domain.

		Neurofascin Igs				**CNTN 1 Igs**	**CNTN2 Igs**				**ROBO1 Igs**		**LINGO 1 Ig**
		1	**2**	**3**	**4**	**2**	**1**	**2**	**3**	**4**	**1**	**2**	**1**
Neurofascin Igs	**1**		8.19	6.69	4.81	10.13	7.04	8.51	3.35	8.17	4.64	6.98	8.14
	2			5.33	4.03	2.17	6.32	2.65	5.04	3.38	4.43	3.15	3.01
	3				4.33	5.25	7.73	5.12	2.05	3.09	3.92	4.36	3.64
	4					4.76	4.38	6.23	1.73	1.97	4.46	3.28	7.02
CNTN1 Ig	**2**						7.52	1.54	5.33	3.55	4.65	3.08	9.12
CNTN2 Igs	**1**							4.82	5.83	4.67	3.02	2.32	15.81
	2								5.44	5.30	4.92	3.39	5.80
	3									3.17	4.99	2.07	2.37
	4										4.53	4.21	6.48
ROBO1 Igs	**1**											3.02	5.55
	2												2.79
LINGO1 Ig	**1**												

Fig.61. Table showing structural similarity among CAM Igs

When considering the biomimetic peptide position and the special sequence and structural conservation of the Ig2 (and of Ig2 with the Lingo1 single Ig), the compared Ig2 or Ig structures were superposed altogether, highlighting the position of their respective peptides corresponding to biomimetic ones from L1CAM and Lingo1 (Fig. 62).

Intriguingly, it was found that structural variation is very low at the peptide region, and that superposition is even best fitting in correspondence of the central, conserved Arg, suggesting this residue may play a special role that can explain the severe CRASH syndrome caused by its mutation.

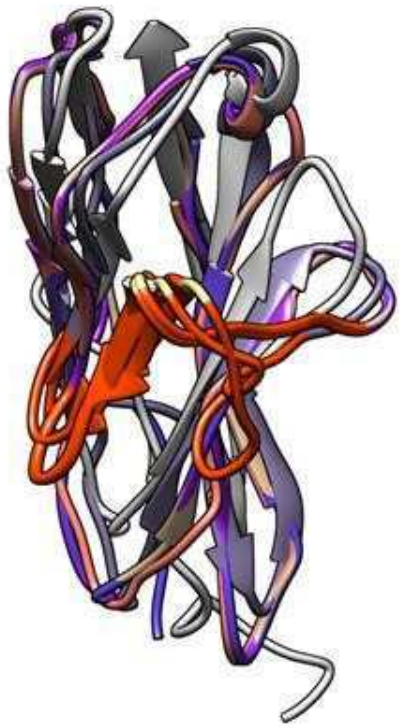


Fig.62. Cartoon representation of CAM Ig domains. Peptide region is highlited in orange red, conserved Arg in khaki.

In order to investigate on the possibly general conservation of such a functional NOG motif, and considering that the story started from L1-A and Lingo1-A, the other three members of the L1 family (CHL1, NrCAM and Neurofascin) were considered, together with other neuronal Lingo proteins, contactins, Roundabout (ROBO) receptors and the Deleted in Colon Cancer (DCC) Netrin receptor, which are known to be part of a complex regulatory network of actractive and repulsive signals mediated by homo- and heterophilic interactions. Fig. 63 shows the extracellular domain architecture of these proteins. Names are shown in red when a pdb structure is available and the twelve proteins chosen for the developing and testing biomimetic peptides are highlited in yellow.

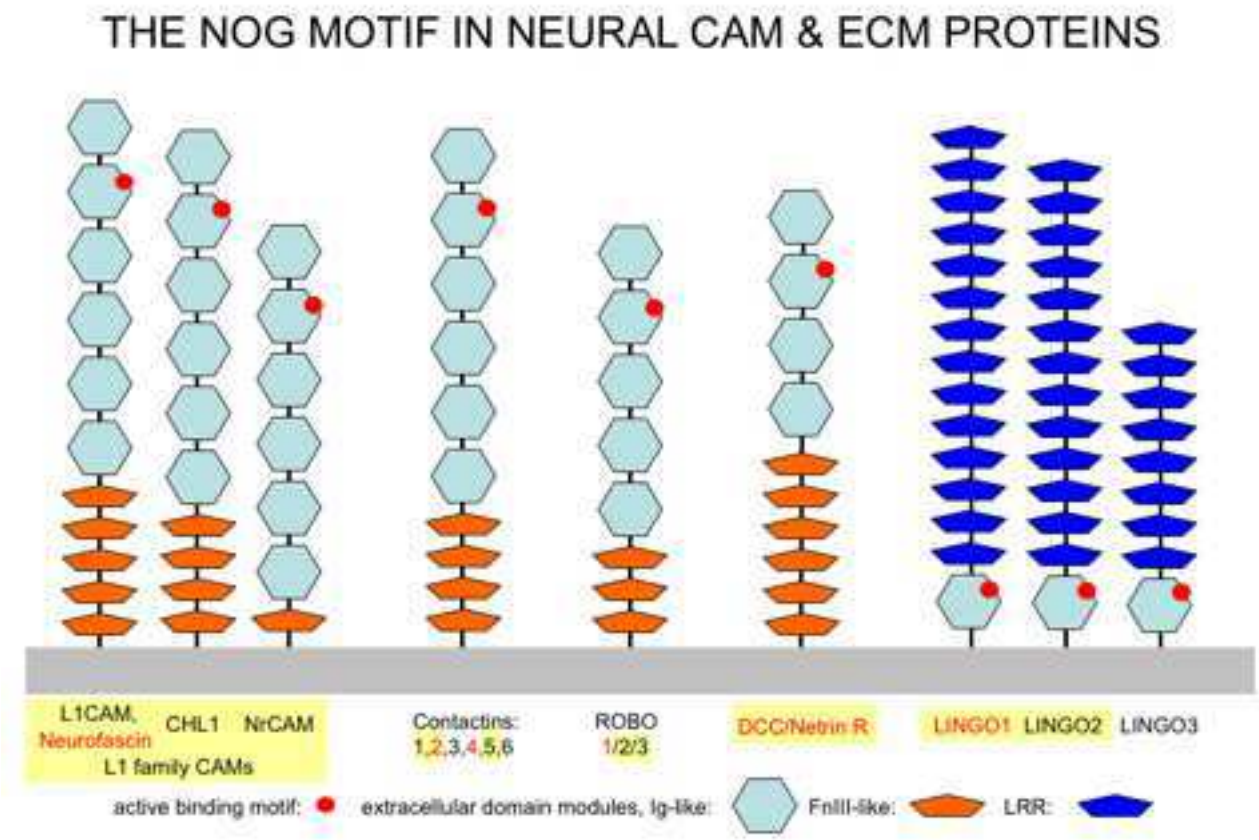


Fig.63. Domain architecture and ECM proteins sharing Ig domains.

In addition to the ten novel peptides, we also designed mutant versions of the L1-A and Lingo1-A peptides, to investigate on the functional relevance of their 100% conserved Arg residue.

In particular, in L1-A_R184A and Lingo1-A_R473A peptides, the conserved Arg is replaced by Ala, while L1-A_R184Q corresponds to natural L1CAM mutation found in the CRASH syndrome.

Moreover, in order to compare the activity of an already characterized biomimetic peptide (L1-A) and of its scramble version to the ectodomain by which it is derived, we also used a full recombinant L1CAM ectodomain (Thermofisher Sci).

Wet lab analyses on the set of studied CAM/ECM proteins and peptides were derived from have already confirmed several predictions from this *in silico* work.

First, surface plasmon resonance (SPR) analysis, performed using the Biacore SPR system and a commercially available recombinant L1CAM extracellular domain (from Thermofisher Sci), confirmed in our experimental conditions the homophilic L1CAM-L1CAM binding (with strong agreement with previously published data from Gouveia *et al*., 2008). This in turn

allowed us to validate the agonist-antagonist hypothesis inferred from docking simulations for the biomimetic and scramble/mutant peptides respectively. Indeed, both L1-A and L1-A_scr could inhibit the homophilic binding, but only the former could also stimulate neuronal differentiation, while neither neurite outgrowth nor axon elongation were improved by the scramble version. Similarly, the L1-A_R184A mutant peptide (Zhao mutant), which in docking simulations is predicted to bind a site other than that occupied by L1-A and of other biomimetic peptides, was as uneffective on cells as the scramble peptide. Instead, L1-A_R184Q peptide (CRASH mutation) showed an intermediate neuritogenic potential (with respect to L1-A and L1-A_scr), thus providing a rationale for the presence in human population of the CRASH mutation (resulting in a severe disease but still compatible with life) and the absence of any R184A mutation, which probably - having lost 100% of activity - is not viable.

When the ten novel peptides likely to be biomimetic were used with neuronal precursors, they were all confirmed to mediate comparable neuritogenic effects.

Evidence that L1-A is really 'biomimetic for' (i.e. active by mimicking) the L1CAM ectodomain it is derived from, was obtained in comparative dose-response experiments, which showed both molecules reach the same plateau effect, even if this is reached by the complete L1CAM ectodomain or by L1-A at nanomolar or micromolar concentration, respectively. In spite of the observed (and not surprising, considering literature on biomimetics) 1000X difference in NOG potential, the two molecules are likely to stimulate the cells via the same pathway, as their combination is unable to overpass the plateau effect: once binding sites are saturated, no additional effect is mediated and thus L1-A can be considered 'agonist'.

The 'antagonist' effect of the scramble peptide was confirmed as well by using L1-A_scr to impair the NOG stimulation mediated by either L1CAM ectodomain or by L1-A peptide: when the bioactive molecule (either the protein or biomimetic peptide) is added to the cell culture together with increasing concentrations of the scramble 'antagonist', this latter can progressively impair the biomimetic NOG effect, as suggested by binding to the horseshoe conformation in docking simulations.

Automated molecular docking is helpful in understanding and predicting molecular recognition and binding affinity between the "target" (protein, DNA, RNA) and the "ligand" (a much smaller molecule docked to the target). Docking approaches goal is to find the most favorable binding modes of a ligand to the target of interest. Binding modes are defined by ligand position, orientation and conformation respect to the target. Each of these state variables describes one degree of freedom in a multidimensional search space. Ligands are

often treated flexibly and this is time consuming. On the other hand, rigid body docking is faster than the flexible one because of the smaller dimension of the search space, but if the ligand conformation is not correct, the complementary fit will be less probable.
Among L1 subfamily, peptide-receptor interactions were investigated on Neurofascin horseshoe structure, the only available in PDB. Peptide sequences used in docking simulations are the following:

Neurofascin:	PITQDKRVSQGHNG
L1-A:	HIKQDERVTMGQNG
L1-A_scr:	IVDQGNREMGTKHQ
L1-A_R184A:	HIKQDEAVTMGQNG
LINGO1-A:	SAKSNGRLTVFPDG
LINGO1-A_scr:	TVFSRSKPLGNDGA

Docking results were evaluated for peptide-receptor interactions via UCSF Chimera (Pettersen *et al*., 2004) and RING2 (Piovesan *et al*., 2016) (Fig. 64-67). The analysis of residues involved in ligand-receptor binding suggests that the conserved Arg at peptide position 7 (in red) and hydrophobic residue at position 2 (in blue) are crucial to properly locate the peptide itself on Ig2 receptor. However, the importance of the other residues order is highlighted by evidence that L1-A_scr shares both positions 2 and 7 (and Lingo1-A_scr only position 2) with biomimetic peptides and this notwithstanding both scramble peptides have no NOG potential.
Arg at peptide position 7 interacts with the following residues in Neurofascin horseshoe:

Neurofascin: His182
L1-A: His218 and Phe219
L1-A_scr: Asp54 and Arg56
LINGO1-A: Asp54, Arg56, Phe219, Thr220, His221
LINGO1-A_scr: Arg56, Phe219, His221

Such interactions can be of different types: Van der Waals, ionic, hydrogen bonds.
Intriguingly, Ala184 in Zhao mutant peptide does not mediate interactions with the Neurofascin horseshoe.

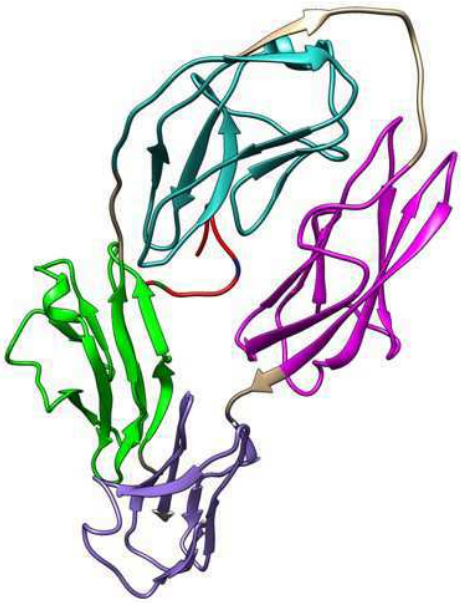
Peptide binding affininity (∆G) and K_d prediction were calculated using PRODIGY (Xue *et al*, 2016):

	∆G (kcal mol^{-1})	K_d (M)
Neurofascin-Neurofascin	-9.6	8.4 e^{-08}
L1-A-Neurofascin	-9.1	2.2 e^{-07}
L1-A_R184A-Neurofascin	-8.7	4.3 e^{-07}
L1-A_scr-Neurofascin	-11.2	5.9 e^{-09}
LINGO1-A-Neurofascin	-9.1	2.1 e^{-07}
LINGO1-A_scr-Neurofascin	-8.2	9.2 e^{-07}

∆G refers to Gibbs free energy. Even if these values are not largely negative, the binding reaction is likely to spontaneously occur. K_d values indicate that complexes can persist for seconds (Sanders, 2010).

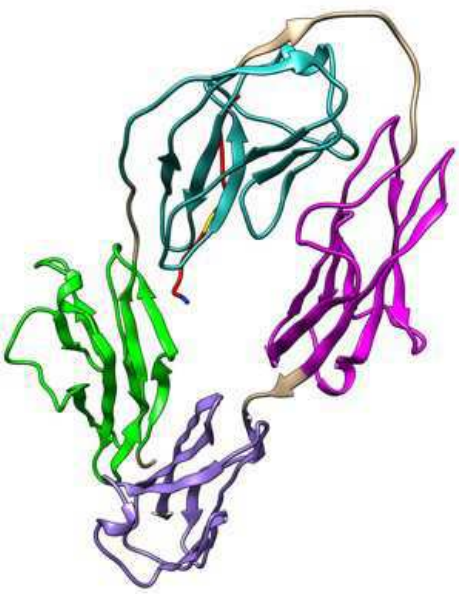
LINGO1-A

LINGO1-A_scr

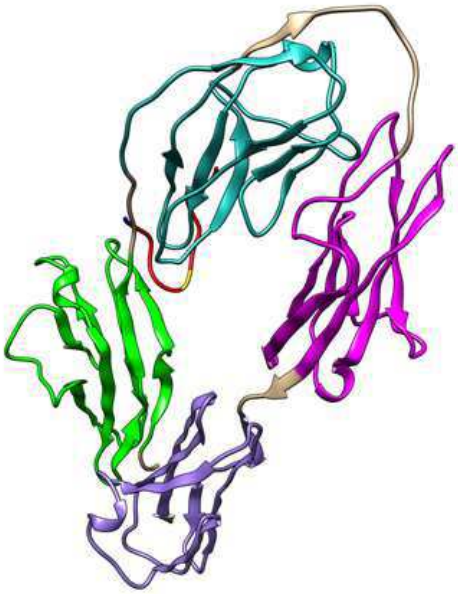


Neurofascin

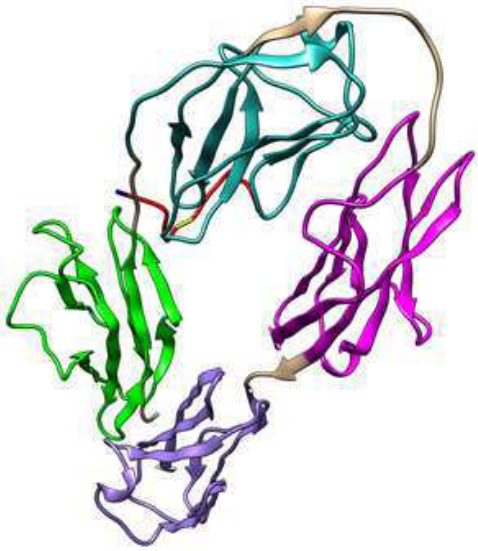
Fig.64 Neurofascin, LINGO1-A and LINGO1-A_scr peptides (red) docking to Neurofascin horseshoe



L1-A

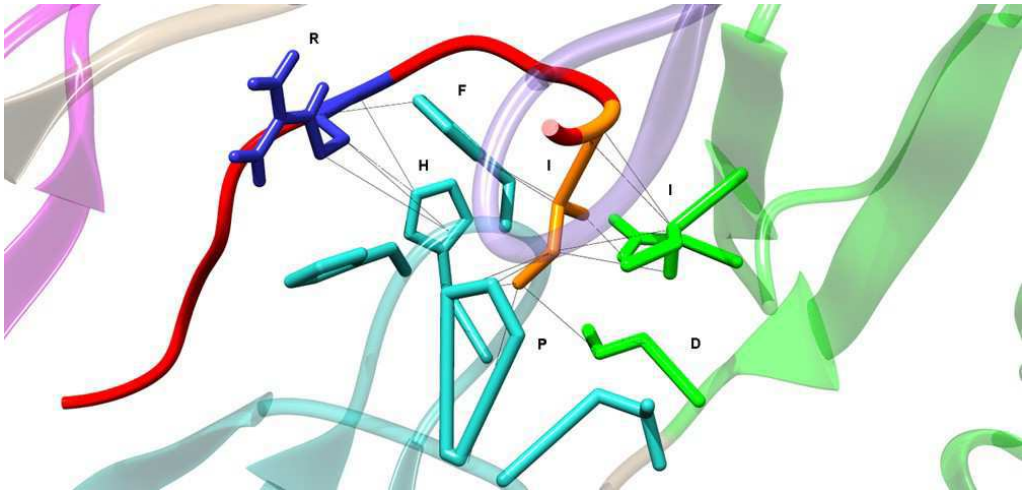


L1-A_R184A

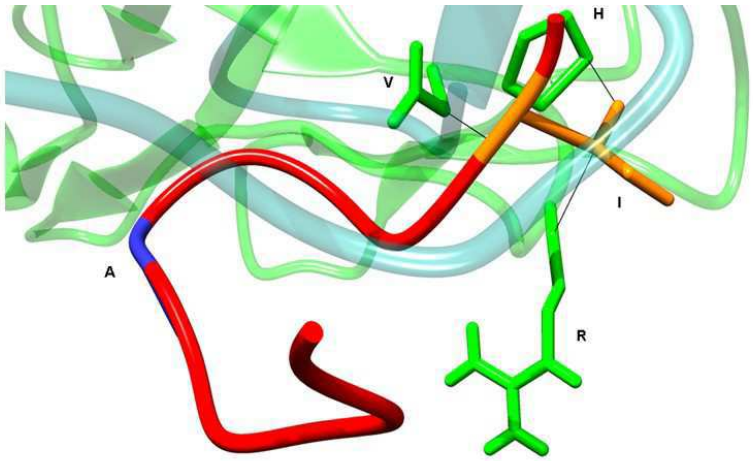


L1-A_scr

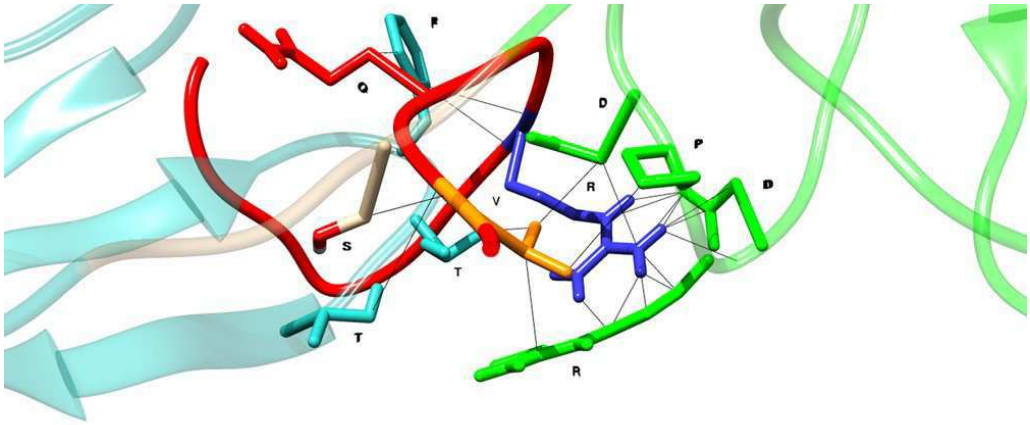
Fig.65. L1-A, L1-A_R184A and L1-A_scr peptides (red) docking to Neurofascin horseshoe.



L1-A

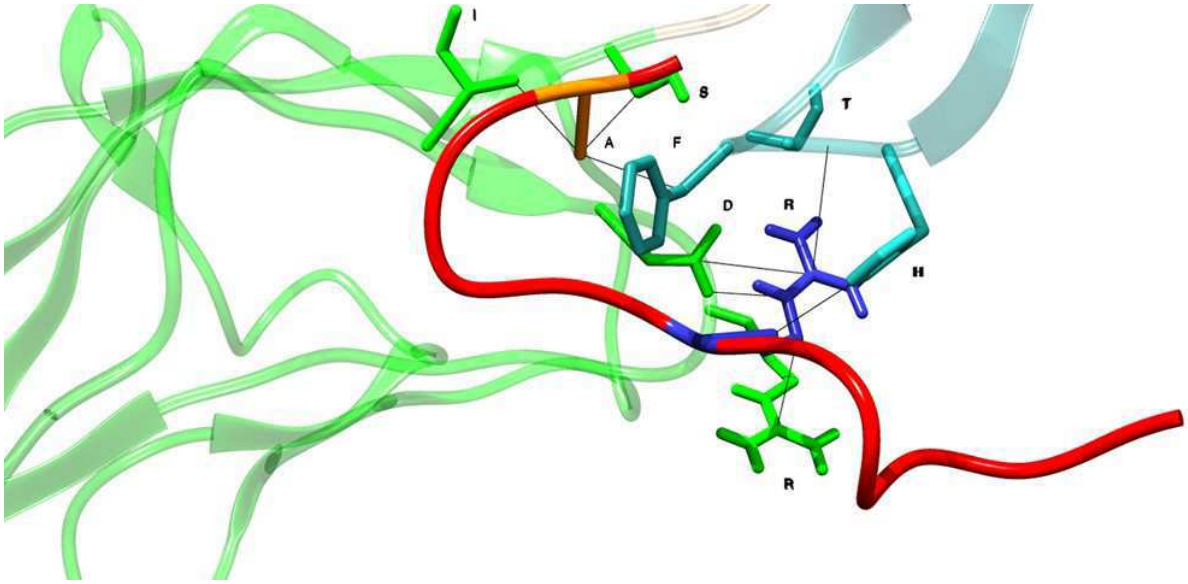


L1-A_R184A

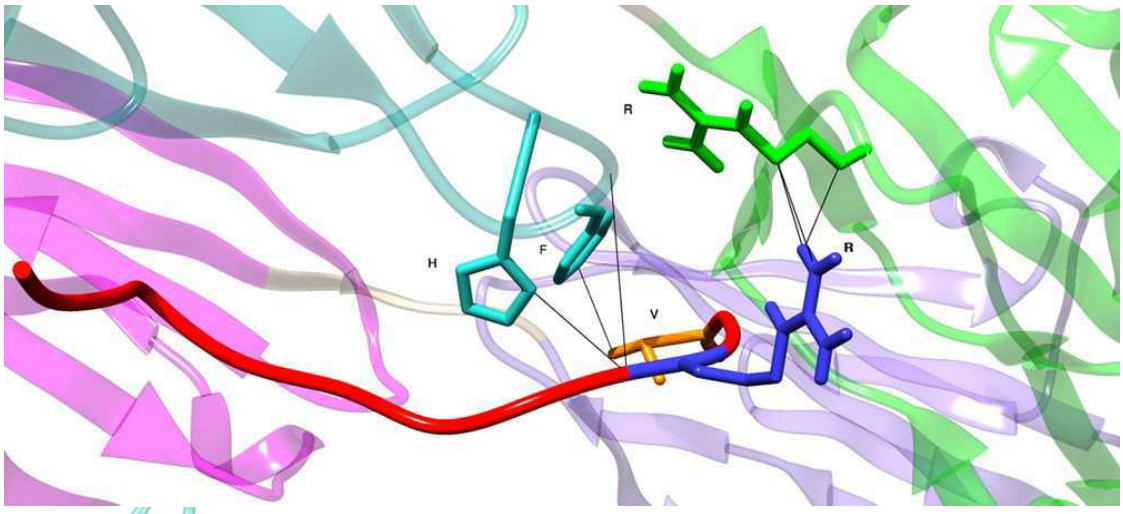


L1-A_scr

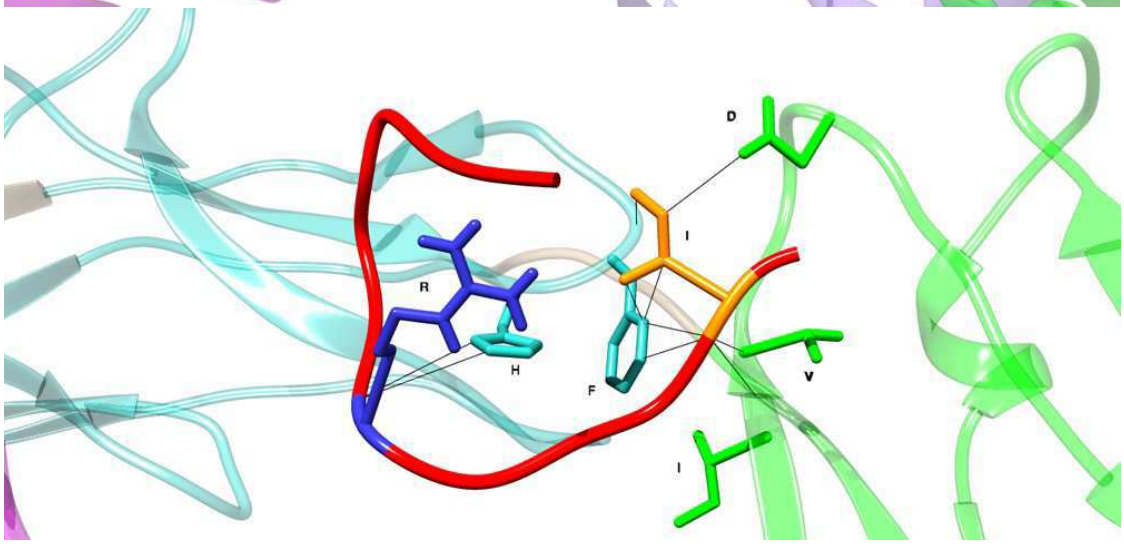
Fig.66. Zoom views of peptides interactions



LINGO1-A



LINGO1-A_scr



Neurofascin

Fig.67. Zoom views of peptides interactions.

Materials and Methods

Homology Modeling

3D structure prediction was performed by Homology modeling (HM) strategy (Fig. 46) using SWISS-MODEL (Biasini *et al*., 2014).

In general, a HM workflow can be summarized as follows:

1. Protein template identification on the basis of structural similarity to the target: this first step is crucial to ensuring a high quality model. Indeed, a reasonable template-target sequence identity can prevent sequence alignment errors and provide at the same time the best fold template for the target. In general, 3D prediction by HM is preferred when sequence identity on the global target-template alignment is ≥ 30%: this cutoff seems to offer good odds for reliable structure prediction (Forrest *et al*., 2006).
2. Target sequence aligment with template sequence: HM softwares generally produce a tentative sequence alignment relative to the target. When target-template global alignment exhibits a conservation >50%, it is generally assumed that a qualitatively reliable aligment is produced by the software, with only modest local misalignments.
3. Target spatial alignment: During this step a preliminary target structural model is constructed. However, the procedure is incorporated in a black-box located in HM softwares. Only inputs and outputs are known, but not black-box internal workings. Some HM workflows construct a backbone model first, and then incorporate side chains into the resulting framework. Other methods assemble the protein core region first, then the exposed loops. Most backbone modeling methods begin by superimposing all templates onto a common framework and computing a consensus framework defined by mean portions of corresponding C_αs.
4. Loop and gap modeling: This step occurs in case of poor sequence conservation between the template and the target or when the aligment presents gapped regions. 5 residues gaps are treated with reasonable accuracy via polypeptide structure libraries (Fernandez-Fuentes *et al*., 2006; Kolodny *et al*., 2002; Levitt, 1992) whereas gaps of grather length are difficult to reliably model: aforementioned peptide libraries or protein folding strategies (e.g. Molecular Dynamics and Monte Carlo conformational searches) can be used to achieve the goal. However, for gaps >10 residues, neither of these strategies is able to lead a model with close correspondence to its real optimal structure. Swiss Model utlizes ProMod-II (Guex and Peitsch, 1997) for loop modeling or MODELLER (Sali and Blundell, 1993) if ProMod-II rsults are not satisfactory.

5. Side chain modeling: As for step 3, also side chain modeling algorithms are implemented as black-box features in HM software. Side chain positions for highly conserved residues may be inferred from the template, as well as conserved disulfide bonds and salt bridges, incorporated into the target during step 3.
6. Refinement: Here, HM softwares tend to ameliorate clashes and molecular strains via conformational searches. A model obtained via HM may include some deviations comparable to an enviromental perturbation. Thus, reverting to normal structure (native conformation) is needed. MD simulations can be used to this task but they are computational expensive. An alternative is to perform simulated annealing calculations. The protein is gradually warmed up to 1000K and then slowly cooled back to ambient temperature. Simulated annealing technique is able to correct most errors in the original structure. MODELLER software can provide these kinds of structure refinement. SWISS-MODEL applies molecular mechanics-based energy minimization to regularize the geometry of the models (Guex *et al*., 2009), using the OpenMM molecular mechanics library (https://swissmodel.expasy.org/docs/help).
7. Validation: This is the last step evaluating the final relaxed model for physical tenability. Many tools are able to detect 3D model aspects differing conformationally from standard bond distance, angle, torsion, clashes. SWISS-MODEL assesses model quality via the composite scoring function QMEAN, taking into account the comparison between geometrical features of the model (pairwise atomic distances, torsion angles, solvent accessibility) and the statistical distributions obtained from experimental structures. Finally, these geometrical features are scored. Each residues is scored betweeen 0 and 1 where higher numbers indicate higher reliability of the residues. In addition, global QMEAN scores are calculated as indicators for the overall model quality, provided as a Z-score, which relates the obtained values to scores calculated from a set of high-resolution X-ray structures (Biasini *et al*., 2014)

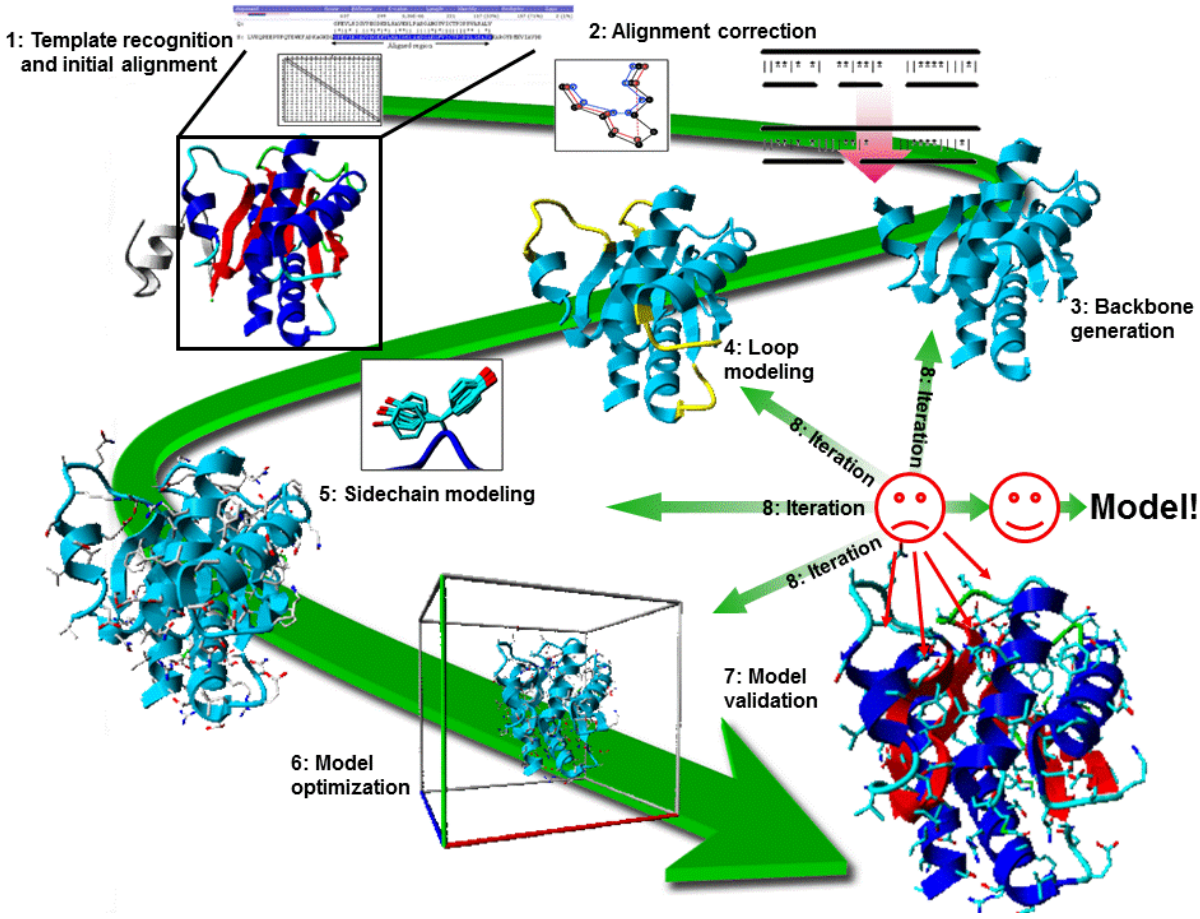


Fig.46. Homology Modeling workflow. From: http://swift.cmbi.ru.nl/teach/B4/IMAGE/Swirl.gif

Threading

Threading was performed using PHYRE 2 web portal (Kelley *et al*., 2015). Normal mode option was chosen. Threading strategy (Fig. 47) is summarized here:

1. Gathering homologous sequences: A query evolutionary profile is constructed by scanning protein sequence databases. The method HHblits is able to perform sequence profile matching. In addition, PSIPRED is used to obtain the secondary structure of the query.
2. Fold library scanning: The profile calculated in the previous stage is then converted to a hidden Markov Model (HMM). This HMM profile is then used as a probe for scanning a precompiled HMMs fold library, composed of a representative set of experimentally determined protein structures whose profiles have been calculated using the same approach in step 1. The alignment algorithm used is HHsearch and the final result of this scanning is a list of query-template alignments ranked by their posterior probabilities. Crude backbone models without side-chains are then generated using these alignments.
3. Loop modeling: A library of fragments of known protein structures (2-15 aa) is used to handle insertions and deletions in the models. After a sequence-profile search, these fragments are fitted to the crude model in order to minimize changes in the dihedral angles of the fragment. Finally fitted fragments are ranked via a combination of empirical energy terms and the top scoring model selected.
4. Side chain placement: This step is performed using the R3 protocol. A fast graph-based technique and a rotamer library allow side-chain placing and steric clashes avoiding.

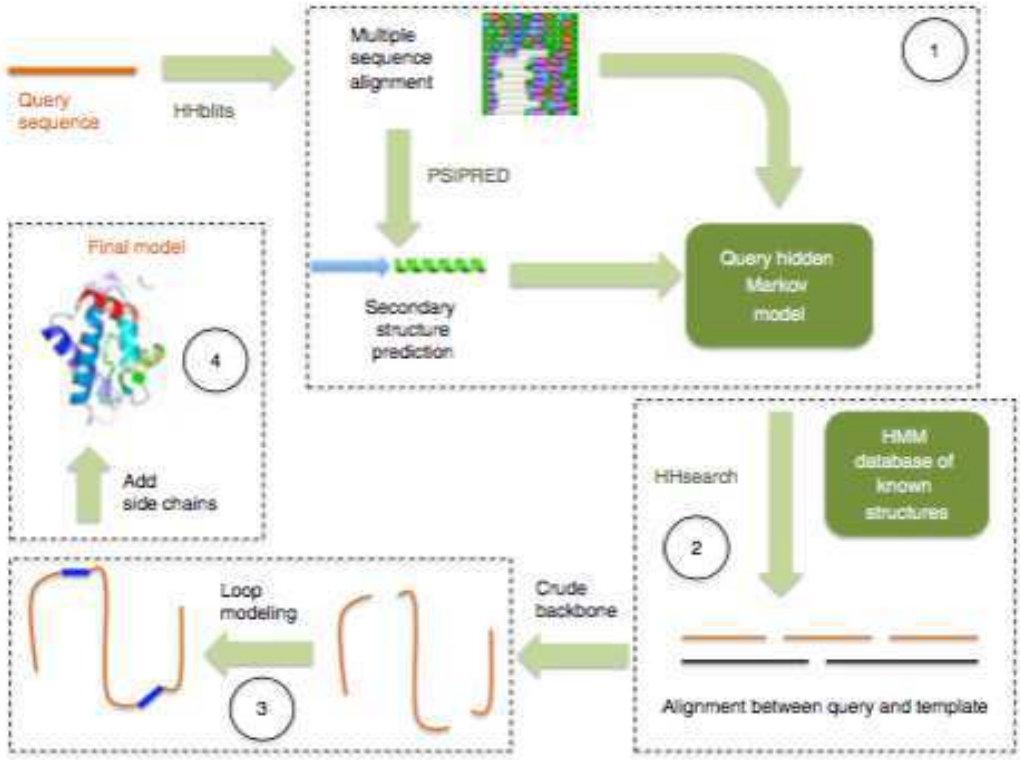


Fig 47. Threading workflow. From: Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun;10(6):845-58.

Ab initio protein structure prediction

Ab initio modelling was performed using I-TASSER/QUARK webservice available at http://zhang.bioinformatics.ku.edu/I-TASSER (Zhang, 2008). I-TASSER (Fig. 48) is an evolution of the TASSER pipeline (Zhang and Skolnick, 2004). Here, 3D models are obtained by knowledge-based approach, i.e. empirical energy terms derived from the statistics of the PDB solved structures. These terms can be generic and sequence-independent such as hydrogen bonding and local backbone stiffness of a polypeptide chain (Zhang *et al*., 2003) or sequence-dependent such as pair wise residue contact potential (Skolnick et al., 1997), distance dependent atomic contact potential (Samudrala and Moult., 1998; Lu and Skolnick, 2001; Zhou and Zhou, 2002; Shen and Sali., 2006), and secondary structure propensities (Zhang and Skolnick, 2005). TASSER pipeline first step is focussed on searching for target possible folds via LOMETS target sequence threading through a set of representative protein structures. Templates are ranked in terms of Z-score (the difference between the raw and average scores in the unit of standard deviation) and the top 10 templates are chosen. Contiguous fragments (>5 residues) are then excised from the threated aligned regions and used to build up full-length models, while unaligned regions are obtained by *ab initio* modelling (Zhang *et al*., 2003). Monte Carlo (MC) simulations are used in the reassembly process. The energy terms of TASSER express predicted secondary structure propensities, backbone hydrogen bonds, short- and long- range correlations and hydrophobic energy based on statistics from the PDB library. I-TASSER extracts spatial restraints from TASSER first round models and then tries to remove steric clashes and refine their topology. This goal is achieved by TM-align (Zhang and Skolnick, 2005) template structures searching from the PDB library, then exploited in the second round simulations. In the second round SPICKER (Zhang and Skolnick, 2004) decoys clustering identifies the low free energy states. Final models are generated by clustering of thousand of decoy models from MC simulations obtained in the second step and the lowest energy structures are selected. I-TASSER output consists of 5 or less models, quantitatively ranked for their confidence by C-score and TM-score (RMSD). C-score values are [-5,2] where higher value means a model with a high confidence. TM-score is a scale for measuring the structural similarity between two structures (the predicted model and the native strucures). A TM-score >0.5 indicates a correct topology model, whereas a TM-score <0.17 means a random similarity. C-score and TM-score are highly correlated. In the output section, onbly the quality prediction (TM-score and RMSD) is reported for the first model, because the

correlation between C-score and TM-score is weak for lower rank models. (Ridgen,2009; http://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html).

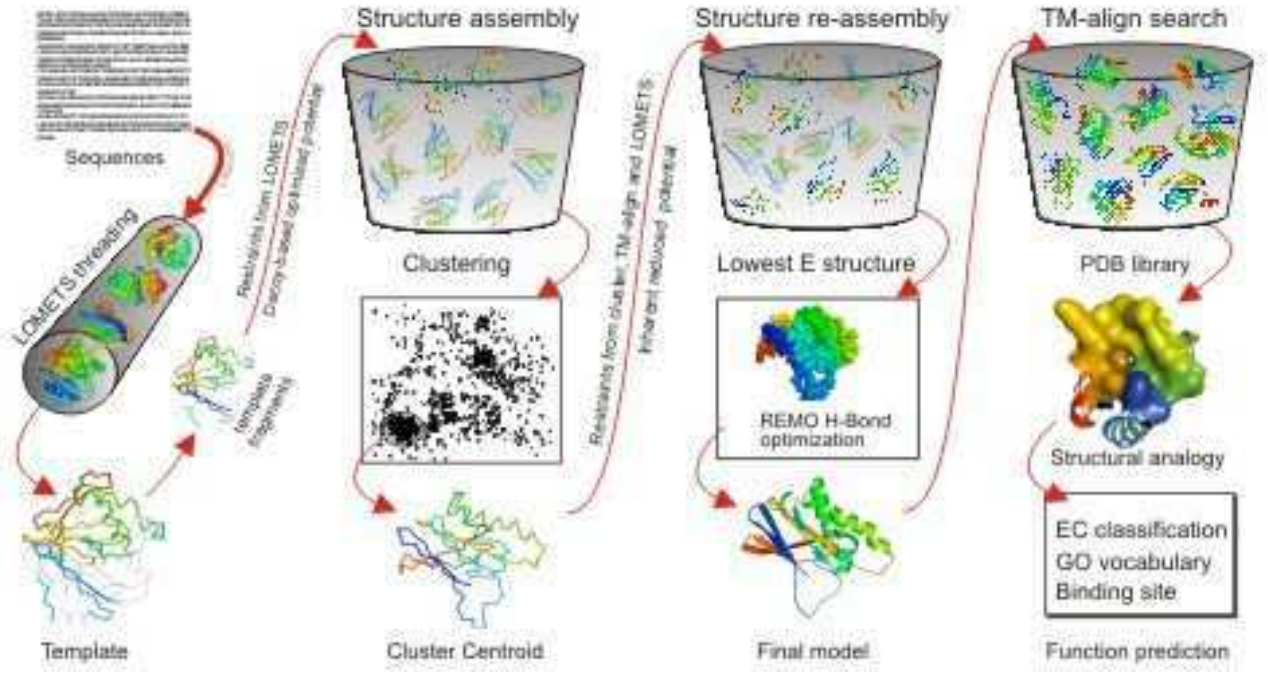


Fig.48. I-TASSER flowchart for protein structure modelling and function prediction. From: http://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html .

Ligand-protein docking

Docking simulations were performed using two softwares: GalaxyPepDock (Lee *et al*., 2015) and NPDock (Tuszynska *et al*., 2015).

GalaxyPepDock (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=PEPDOCK) is a similarity-based protein-peptide docking approach that performs additional flexible-structure energy-based optimization. GalaxyPepDock flowchart is represented in Fig. 49. Target proteins were prepared for docking using DockPrep in UCSF Chimera. PepBind database (Das *et al*., 2013) is used at first for templates selection with the following score for each complex structure in the database:

$$S_{complex} = Z_{TM} + Z_{Inter}$$

where:

Z_{TM} : measure of the protein structure similarity by the Z-score of the TM-score of a database protein structure when alignes to the target protein structure by TM-align (Zhang and Skolnick, 2005);

Z_{Inter} : measure of the interaction similarity of a database complex and the target complex when aligned to the former by the Z-score of the interaction similarity score S_{Inter}.

Up to 10 complexes with $S_{complex}$ >90% of the maximum value are selected as templates and then used in the model-building procedure. This step is provided by GalaxyTBM (*Ko et al*., 2012). This tool models first the more reliable core region from multiple templates whereas variable local regions are re-modeled by an *ab initio* method. Of the model structures generated by GalaxyTBM, 10 structures are selected by choosing the structures with the best energy values for each template and are further refined using GalaxyRefine tool. In this latter step backbone and side chains are adjusted via MD relaxations after side chain repacking. When tested on the CAPRI target 67, GalaxyPepDock generates models that are more accurate than the best server models submitted during the CAPRI blind prediction experiment.

NPDock is available at the following website: http://genesilico.pl/NPDock. As for the other docking tools, NPDock comprises scoring of poses, clustering of the best-scored models and refinement of the most promising solutions. Computational workflow (Fig. 50.) implements GRAMM program, DARS-RNP and QUASI-RNP/DNP statistical potentials for

scoring protein-RNA/DNA complexes with coarse-grained representation, and tools for clustering, selection and refinement of models. In the first step, GRAMM performs a rigid body global search generating decoys. Then, the aformentioned statistical potentials score and rank the obrained decoys. The best-scored decoys are then clustered and representatives of the three largest clusters are selected. Finally, protein-nucleic acid interactions are optimized via Monte Carlo Simulated Annealing procedure. As for GalaxyPepDock, template was prepared via UCSFChimera DockPrep.

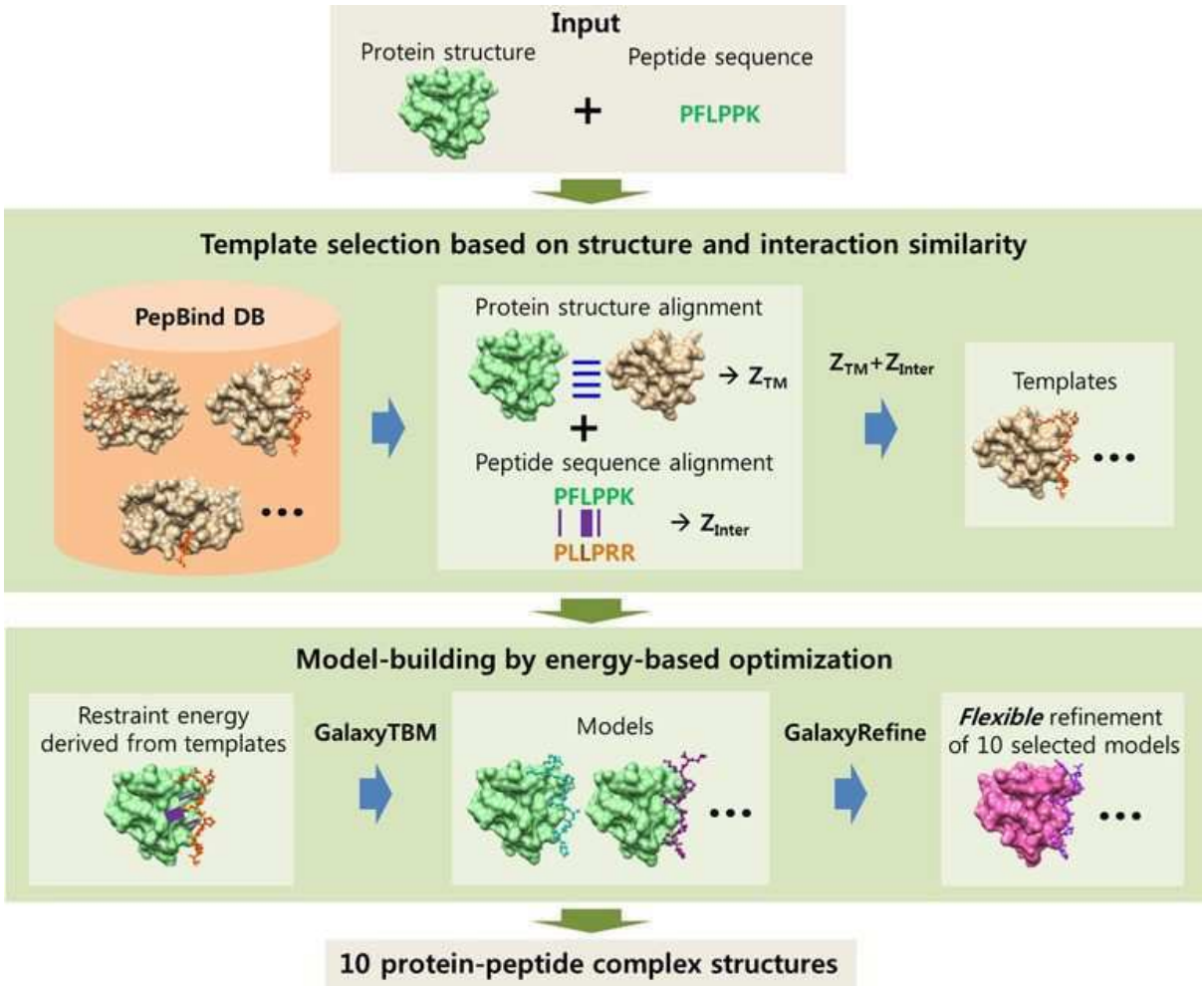


Fig.49. GalaxyPepDock flowchart. Given a protein structure and a peptide sequence, template complex structures are first selected from the PepBind database based on protein structure similarity and protein-peptide interaction similarity. Models are then built with GalaxyTBM, and the 10 models that are selected based on energy are returned after further optimization by GalaxyRefine tool. From: Lee H, Heo L, Lee MS, Seok C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res*. 2015 Jul 1;43(W1):W431-5.

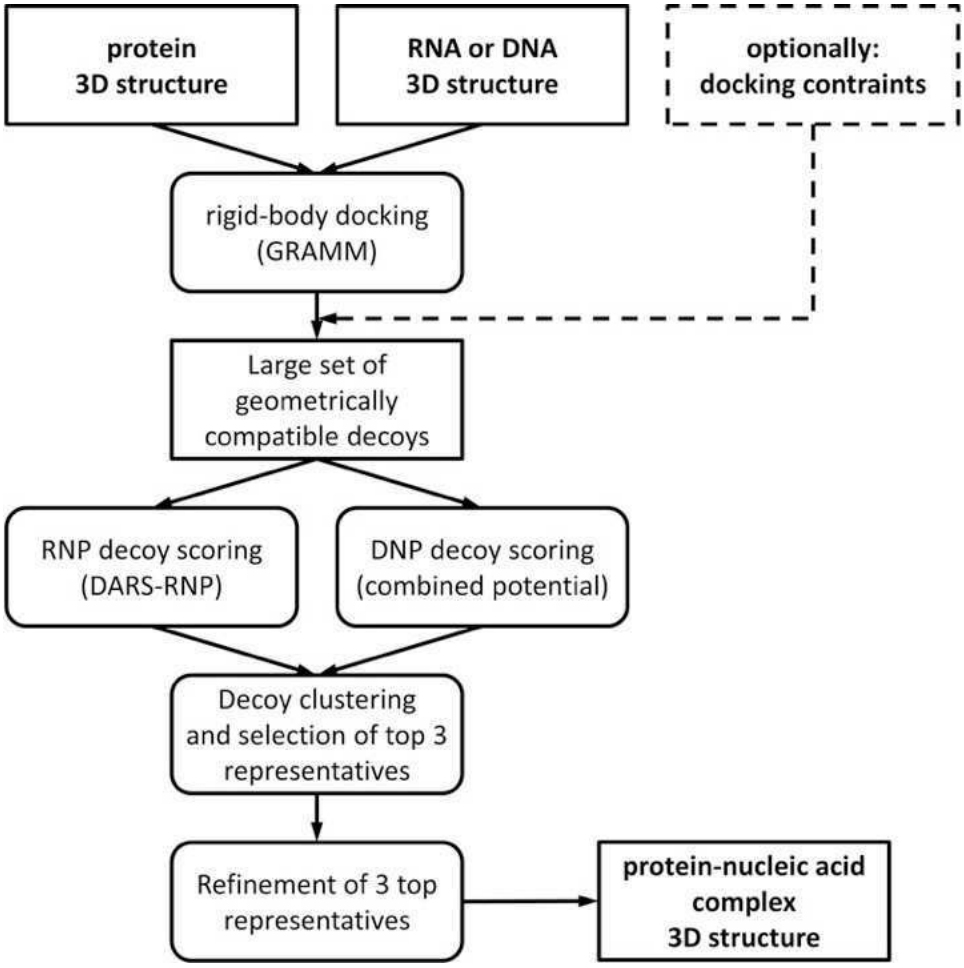


Fig.50. NPDock flowchart. From: Tuszynska I, Magnus M, Jonak K, Dawson W, Bujnicki JM. NPDock: a web server for protein-nucleic acid docking. *Nucleic Acids Res*. 2015 Jul 1;43(W1): W425-30.

Electrostatic calculations

Electrostatic analyses were carried out using two tools: APBS (Dolinsky *et al*., 2007) and WebPIPSA (Richter *et al*., 2008).

Adaptive Poisson-Boltzmann Solver (APBS) software package is able to solve PBE using electrostatic "focusing". This is a popular finite difference technique for generating accurate solutions to the PBE in subsets of the problem domain, such as a binding or titrable sites within a protein. This approach implies that charges and dielectric constants are discretized over a grid. Protein molecular surface (MS) is mapped onto a user-defined density 3D grid, then used to obtain the finite difference solutions of the PBE. The product of the electrical potential and charge at each voxel (grid point), where a real charge has been mapped, provides the electrostatic free energies. APBS performs calculations by using initially a coarser grid and then a finer one for the refinement. APBS carries out calculations with a grid spacing of 0.5Å. APBS first defines the solvent accessible regions of the protein and calculates the electrostatic potential for each of the grid point. Finally, the electrostatic potential is mapped onto the MS (.dx file). Electrostatic calculations can be applied on molecules of a wide size range. I used APBS through UCSF Chimera via Opal server. (Baker *et al*., 2001; Ul-Haq and Madura, 2015).

WebPIPSA (http://pipsa.eml.org) allows computing and comparing electrostatic potential among a large number of proteins. After structures upload, WebPIPSA workflow (Fig.51.) can be summarized as follow:

1. Structures superimposition: "sup2pdb" option was selected, in which the sequence of one structure, called template, undergoes a pairwise sequence alignment with the remaining coordinate files. Alignments were then used to perform structures superimpositions;
2. Polar hydrogens addiction: WHATIF (Vriend, 1990) adds polar hydrogen atoms to the structures. Protonation is executed at pH7 for all residue except for His, trated as singly or doubly protonated;
3. Electrostatic potentials calculations: I choose to execute this step using APBS. UHBD is the alternative choice. Electrostatic potentials are automatically calculated. Chosen parameters were the following: ionic strength of 150mM and a temperature of 300K. Solvent is treated as implicit;
4. Electrostatic potentials comparison: A probe of radius 2Å defines the protein surface. PIPSA compares potentials in the complete protein surface skins. The skin estends out from the protein surface with a thickness of 3Å. Electrostatic protein comparison

is possible due to the implementation of the Hodgkin or Carbo similarity indexes. The similarity indeces range from -1 (anti-correlated potentials) through 0 (uncorrelated) to +1 (identical potentials). These values are converted into distances given by $\sqrt{2-2SI}$ where SI stands for similarity index. Distance values are comprised between 0 (identical) and 2 (anti-correlated potentials);

5. Clustering analysis and epograms generation: This is the last step. WebPIPSA output consists in a heat map (representing the distance matrix) and an epogram, allowing the fast identification of inter-protein relations.

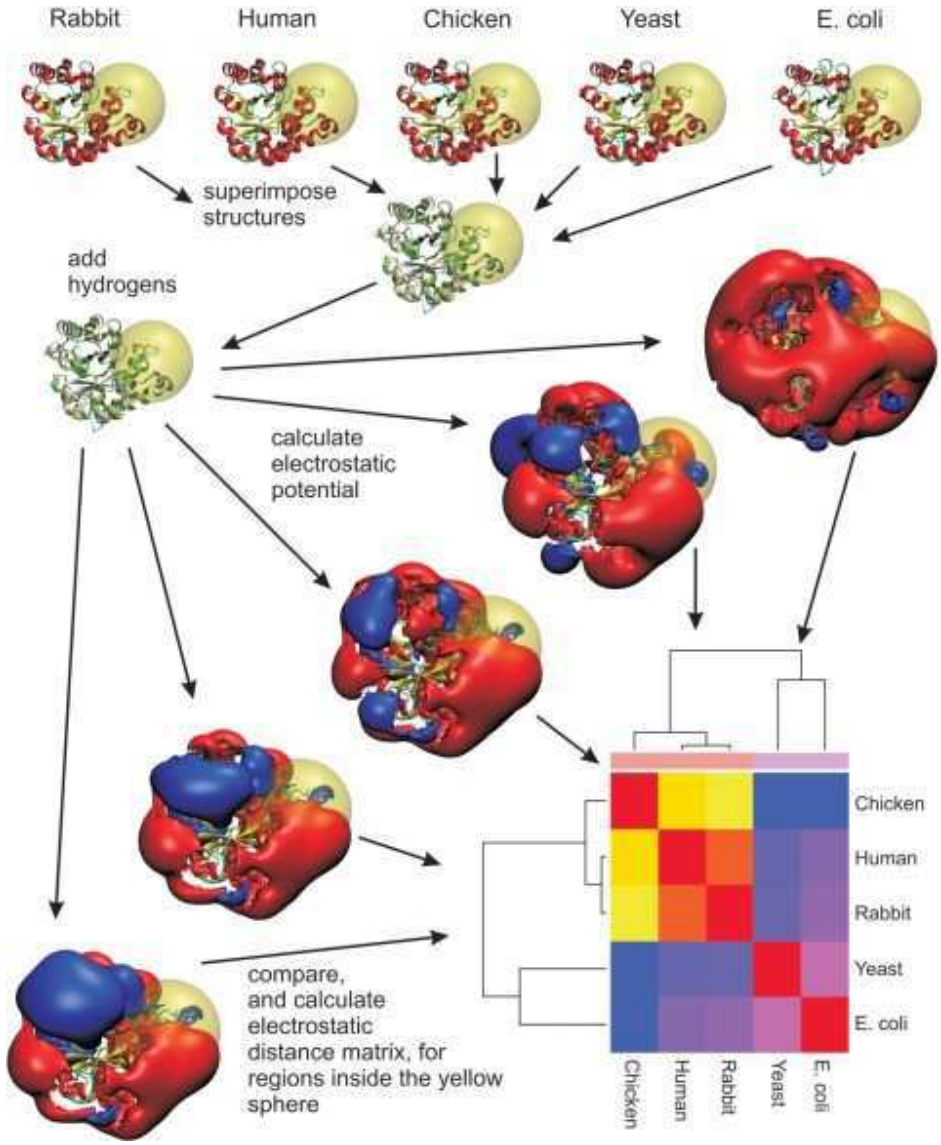


Fig.51. The WebPIPSA workflow. PIPSA is used to compare the electrostatic potentials and to calculate a distance matrix. These distances are used to cluster the proteins according to the relations between their electrostatic potentials and the clustering is displayed in a tree-like diagram (epogram). From: Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade RC. webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res*. 2008 Jul1;36(Web Server issue):W276-80.

Concluding remarks

Molecular modeling can strongly boost wet lab analyses for shedding light on biomolecular interactions, helping design of novel therapeutics, such as vaccines, drugs or regenerative medicine devices. In order to better understand protein “behavior” a bioinformatic integrate approach was used.

A bioinformatic protocol, published in Righetto *et al*., 2014, allowed at first to investigate on avian hemagglutinin evolution. This surface protein plays a pivotal role in the viral spreading among host population and in the rising of a heterogenous, genetically-related viral pool. The study of protein surface with a bioinformatic integrated approach resulted in interesting evidencies, helpful in understanding the acquisition of virulence determinants or host specificity and, more widely, a better comprehension of influenza virus evolutionary dynamics. Structural analyses were carried out on H5N1 and H9N2 subtypes haemagglutinin, in order to discover surface differences responsible for functional evolution. Respect to primary nucleotide- or sequence-based analyses, structural studies also allow to take into account amount and kind of mutations, therefore weighting them on the protein 3D surface. The study on H9N2 confirmed the wide applicability of the novel approach and in a manuscript that is quite close to submission.

Such a bioinformatic approach was also used to shed light on an intriguing isoform of VAMP7 gene, encoding variant VAMP7b. This variant, shared only by humans and apes, exhibits a novel 116 residues region/domain of unknown function. Structural modeling of VAMP7b was able to predict conservation of the closed conformation and this was confirmed by 'wet lab coauthors' both in vitro and in vivo by means of NMR and two-hybrid approach. Moreover, *ab initio* modeling of the new region/domain provided us with a possible rationale for a specific role of this isoform in the context of neuronal function and for its involvement, in the evenience of unbalanced splicing, in neurological disorders.

Last but not least, in the context of a work aimed at designing and characterizing novel biomimetics for neural regenerative medicine, structural modeling and superpositions with CAM/ECM human proteins and peptide-protein docking simulations could provide our team with useful suggestions and models for driving wet lab workpackages.

Both in silico tasks for VAMP7b and the CAM/ECM biomimetics works are going to be included in two corresponding manuscripts of which I am co-author and that are planned to be submitted by summer/fall in current year.

References

Baigent SJ, McCauley JW. Influenza type A in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission. Bioessays. 2003 Jul;25(7):657-71.

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A. 2001 Aug 28;98(18):10037-41. Epub 2001 Aug 21.

Baldacci L, Golfarelli M, Lumini A, Rizzi S. Clustering techniques for protein surfaces. Pattern Recognition 2006 February; 39(2370).

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol. 2002 Nov 15;324(1):105-21.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W252-8.

Blumenthal J, Ginzburg I. Zinc as a translation regulator in neurons: implications for P-body aggregation. J Cell Sci. 2008 Oct 1;121(Pt 19):3253-60.

Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998 Jul 3;280(1):1-9.

Brose K, Bland KS, Wang KH, Arnott D, Henzel W, Goodman CS, Tessier-Lavigne M, Kidd T. Slit proteins bind Robo receptors and have an evolutionarily conserved role in repulsive axon guidance. Cell. 1999 Mar 19;96(6):795-806.

Burgo A, Sotirakis E, Simmler MC, Verraes A, Chamot C, Simpson JC, Lanzetti L, Proux-Gillardeaux V, Galli T. Role of Varp, a Rab21 exchange factor and TI-VAMP/VAMP7 partner, in neurite growth. EMBO Rep. 2009 Oct;10(10):1117-24.

Büttner B, Horstkorte R. Intracelluar ligands of NCAM. Adv Exp Med Biol. 2010;663:55-66.

Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 2004 Jan;13(1):190-202.

Carrat F, Flahault A. Influenza vaccine: the challenge of antigenic drift. Vaccine. 2007 Sep 28;25(39-40):6852-62. Epub 2007 Aug 3.

Cavallaro U, Dejana E. Adhesion molecule signalling: not always a sticky business. Nat Rev Mol Cell Biol. 2011 Mar;12(3):189-97.

Celniker G., Nimrod G., Ashkenazy H., Glaser F., Martz E., Mayrose I., Pupko T., and Ben-Tal N. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function Isr. J. Chem. 2013 March 10

Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins. 2002 May 15;47(3):334-43.

Chandra Parija S. Microbiology&Immunology, 2012, Elsevier.

Chen J, Lee KH, Steinhauer DA, Stevens DJ, Skehel JJ, Wiley DC. Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and the origin of the labile conformation. Cell. 1998 Oct 30;95(3):409-17.

Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O'Neill R, Schickli J, Palese P, Henklein P, Bennink JR, Yewdell JW. A novel influenza A virus mitochondrial protein that induces cell death. Nat Med. 2001 Dec;7(12):1306-12

Chen YC. Beware of docking! Trends Pharmacol Sci. 2015 Feb;36(2):78-95

Childs RA, Palma AS, Wharton S, Matrosovich T, Liu Y, Chai W, Campanero-Rhodes MA, Zhang Y, Eickmann M, Kiso M, Hay A, Matrosovich M, Feizi T. Receptor-binding specificity of pandemic influenza A (H1N1) 2009 virus determined by carbohydrate microarray. Nat Biotechnol. 2009 Sep;27(9):797-9.

Chothia C, Janin J. Principles of protein-protein recognition. Nature. 1975 Aug 28;256(5520):705-8.

Clissold PM, Ponting CP. JmjC: cupin metalloenzyme-like domains in jumonji, hairless and phospholipase A2beta. Trends Biochem Sci. 2001 Jan;26(1):7-9.

Comoglio PM, Boccaccio C, Trusolino L. Interactions between growth factor receptors and adhesion molecules: breaking the rules. Curr Opin Cell Biol. 2003 Oct;15(5):565-71.

Connolly LM. Computation of molecular volume. J. Am. Chem. Soc. 1985 107(1118).

Craig AM, Kang Y. Neurexin-neuroligin signaling in synapse development. Curr Opin Neurobiol. 2007 Feb;17(1):43-52

Das AA, Sharma OP, Kumar MS, Krishna R, Mathur PP. PepBind: a comprehensivedatabase and computational tool for analysis of protein-peptide interactions. Genomics Proteomics Bioinformatics. 2013 Aug;11(4):241-6.

de Vries SJ, van Dijk AD, Bonvin AM. WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins. 2006 May 15;63(3):479-89.

D'Esposito M, Ciccodicola A, Gianfrancesco F, Esposito T, Flagiello L, Mazzarella R, Schlessinger D, D'Urso M. A synaptobrevin-like gene in the Xq28 pseudoautosomal region undergoes X inactivation. Nat Genet. 1996 Jun;13(2):227-9.

Digard P, Nash AA, Randall RE. Molecular Pathogenesis of Virus Infections, SGM symposium 64, 2005, Cambridge University Press.

Dill KA, Bromberg S. Molecular Driving Forces. Statistical Thermodynamics in Chemistry and Biology. Garland Sci. 2003.

Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W665-7.

Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W522-5.

DuBois RM, Zaraket H, Reddivari M, Heath RJ, White SW, Russell CJ. Acid stability of the hemagglutinin protein regulates H5N1 influenza virus pathogenicity. PLoS Pathog. 2011 Dec;7(12):e1002398.

Dudek SE, Wixler L, Nordhoff C, Nordmann A, Anhlan D, Wixler V, Ludwig S. The influenza virus PB1-F2 protein has interferon antagonistic activity. Biol Chem. 2011 Dec;392(12):1135-44.

Estes PS, O'Shea M, Clasen S, Zarnescu DC. Fragile X protein controls the efficacy of mRNA transport in Drosophila neurons. Mol Cell Neurosci. 2008 Oct;39(2):170-9.

Ferguson L, Olivier AK, Genova S, Epperson WB, Smith DR, Schneider L, Barton K, McCuan K, Webby RJ, Wan XF. Pathogenesis of Influenza D Virus in Cattle. J Virol. 2016 May 27;90(12):5636-42.

Fernandez-Fuentes N, Oliva B, Fiser A. A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res. 2006 Apr 14;34(7):2085-97. Print 2006.

Fersht AR. Basis of biological specificity. Trends in Biochemical Sciences. 1984 April; 9(4): 145-7.

Filippini F, Rossi V, Galli T, Budillon A, D'Urso M, D'Esposito M. Longins: a new evolutionary conserved VAMP family sharing a novel SNARE domain. Trends Biochem Sci. 2001 Jul;26(7):407-9.

Finci LI, Krüger N, Sun X, Zhang J, Chegkazi M, Wu Y, Schenk G, Mertens HD, Svergun DI, Zhang Y, Wang JH, Meijers R. The crystal structure of netrin-1 in complex with DCC reveals the bifunctionality of netrin-1 as a guidance cue. Neuron. 2014 Aug 20;83(4):839-49.

Forrest LR, Tang CL, Honig B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J. 2006 Jul 15;91(2):508-17.

Fujiyoshi Y, Kume NP, Sakata K, Sato SB. Fine structure of influenza A virus observed by electron cryo-microscopy. EMBO J. 1994 Jan 15;13(2):318-26.

Galli T, Zahraoui A, Vaidyanathan VV, Raposo G, Tian JM, Karin M, Niemann H, Louvard D. A novel tetanus neurotoxin-insensitive vesicle-associated membrane protein in SNARE complexes of the apical plasma membrane of epithelial cells. Mol Biol Cell. 1998 Jun;9(6):1437-48.

García-Sastre A. Inhibition of interferon-mediated antiviral responses by influenza A viruses and other negative-strand RNA viruses. Virology. 2001 Jan 20;279(2):375-84.

Garten W., Klenk H.-D. Cleavage Activation of the Influenza Virus Hemagglutinin and Its Role in Pathogenesis. In: Klenk H.-D., Matrosovich M.N., Stech J., editors. Avian Influenza. Karger; Basel, Switzerland: 2008.

Gorham RD Jr, Kieslich CA, Morikis D. Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. Ann Biomed Eng. 2011 Apr;39(4):1252-63.

Gouda H, Kuntz ID, Case DA, Kollman PA. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. Biopolymers. 2003 Jan;68(1):16-34.

Gregoriades A, Frangione B. Insertion of influenza M protein into the viral lipid bilayer and localization of site of insertion. J Virol. 1981 Oct;40(1):323-8.

Grishin NV, Phillips MA. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. Protein Sci. 1994 Dec;3(12):2455-8.

Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, Jahn D. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W526-31.

Gruber J, Zawaira A, Saunders R, Barrett CP, Noble ME. Computational analyses of the surface properties of protein-protein interfaces. Acta Crystallogr D Biol Crystallogr. 2007 Jan;63(Pt 1):50-7.

Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. 1997 Dec;18(15):2714-23.

Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis. 2009 Jun;30 Suppl 1:S162-73.

Hamilton BS, Whittaker GR, Daniel S. Influenza virus-mediated membrane fusion: determinants of hemagglutinin fusogenic activity and experimental approaches for assessing virus fusion. Viruses. 2012 Jul;4(7):1144-68.

Han T, Marasco WA. Structural basis of influenza virus neutralization. Ann N Y Acad Sci. 2011 Jan;1217:178-90.

Hancock RL, Masson N, Dunne K, Flashman E, Kawamura A. The Activity of JmjC Histone Lysine Demethylase KDM4A is Highly Sensitive to Oxygen Concentrations. ACS Chem Biol. 2017 Apr 21;12(4):1011-1019.

Haspel J, Grumet M. The L1CAM extracellular region: a multi-domain protein with modular and cooperative binding modes. Front Biosci. 2003 Sep 1;8:s1210-25.

Herron LR, Hill M, Davey F, Gunn-Moore FJ. The intracellular interactions of the L1 family of cell adhesion molecules. Biochem J. 2009 May 1;419(3):519-31.

Herz C, Stavnezer E, Krug R, Gurney T Jr. Influenza virus, an RNA virus, synthesizes its messenger RNA in the nucleus of infected cells. Cell. 1981 Nov;26(3 Pt 1):391-400.

Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science. 1995 May 26;268(5214):1144-9.

Horimoto T, Nakayama K, Smeekens SP, Kawaoka Y. Proprotein-processing endoproteases PC6 and furin both activate hemagglutinin of virulent avian influenza viruses. J Virol. 1994 Sep;68(9):6074-8.

Hu YJ, Belaghzal H, Hsiao WY, Qi J, Bradner JE, Guertin DA, Sif S, Imbalzano AN. Transcriptional and post-transcriptional control of adipocyte differentiation by Jumonji domain-containing protein 6. Nucleic Acids Res. 2015 Sep 18;43(16):7790-804.

Isin B, Doruker P, Bahar I. Functional motions of influenza virus hemagglutinin: a structure-based analytical approach. Biophys J. 2002 Feb;82(2):569-81.

Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol. 1995;63(1):31-65.

Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci U S A. 1996 Jan 9;93(1):13-20.

Kamiguchi H. The role of cell adhesion molecules in axon growth and guidance. Adv Exp Med Biol. 2007;621:95-103.

Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem. 1959;14:1-63.

Kawaoka Y, Naeve CW, Webster RG. Is virulence of H5N2 influenza viruses in chickens associated with loss of carbohydrate from the hemagglutinin? Virology. 1984 Dec;139(2):303-16.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun;10(6):845-58.

Kenwrick S, Watkins A, De Angelis E. Neural cell recognition molecule L1: relating biological complexity to human disease mutations. Hum Mol Genet. 2000 Apr 12;9(6):879-86.

Kessel A, Ben-Tal N. Introduction to proteins: structure, function, and motion. Boca Raton: CRC Press; 2010

Klenk HD, Rott R, Orlich M, Blödorn J. Activation of influenza A viruses by trypsin treatment. Virology. 1975 Dec;68(2):426-39.

Klenk HD, Garten W. Host cell proteases controlling virus pathogenicity. Trends Microbiol. 1994 Feb;2(2):39-43.

Ko J, Park H, Seok C. GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. BMC Bioinformatics. 2012 Aug 10;13:198.

Kobasa D, Wells K, Kawaoka Y. Amino acids responsible for the absolute sialidase activity of the influenza A virus neuraminidase: relationship to growth in the duck intestine. J Virol. 2001 Dec;75(23):11773-80.

Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE 3rd. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res. 2000 Dec;33(12):889-97.

Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. J Mol Biol. 2002 Oct 18;323(2):297-307. Erratum in: J Mol Biol. 2003 Feb 7;326(1):337.

Kukol A. Molecular modeling of proteins. Humana Press 2008.

Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol. 1971 Feb 14;55(3):379-400.

Lee H, Heo L, Lee MS, Seok C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. Nucleic Acids Res. 2015 Jul 1;43(W1):W431-5.

Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol. 1992 Jul 20;226(2):507-33.

Li L, Li C, Alexov E. On the Modeling of Polar Component of Solvation Energy using Smooth Gaussian-Based Dielectric Function. J Theor Comput Chem. 2014 May;13(3).

Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996 Mar 29;257(2):342-58.

Liu H, Focia PJ, He X. Homophilic adhesion mechanism of neurofascin, a member of the L1 family of neural cell adhesion molecules. J Biol Chem. 2011 Jan 7;286(1):797-805.

Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol. 1999 Feb 5;285(5):2177-98.

London N, Movshovitz-Attias D, Schueler-Furman O. The structural basis of peptide-protein binding strategies. Structure. 2010 Feb 10;18(2):188-99.

Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins. 2001 Aug 15;44(3):223-32.

Mancias JD, Goldberg J. The transport signal on Sec22 for packaging into COPII-coated vesicles is a conformational epitope. Mol Cell. 2007 May 11;26(3):403-14.

Mandemakers WJ, Barres BA. Axon regeneration: it's getting crowded at the gates of TROY. *Curr Biol*. 2005 Apr 26;15(8):R302-5.

Maness PF, Schachner M. Neural recognition molecules of the immunoglobulin superfamily: signaling transducers of axon guidance and neuronal migration. Nat Neurosci. 2007 Jan;10(1):19-26. Review. Erratum in: Nat Neurosci. 2007 Feb;10(2):263.

Martin K, Helenius A. Nuclear transport of influenza virus ribonucleoproteins: the viral matrix protein (M1) promotes export and inhibits import. Cell. 1991 Oct 4;67(1):117-30

Martín J, Wharton SA, Lin YP, Takemoto DK, Skehel JJ, Wiley DC, Steinhauer DA. Studies of the binding properties of influenza hemagglutinin receptor-site mutants. Virology. 1998 Feb 1;241(1):101-11.

Matarazzo MR, Cuccurese M, Strazzullo M, Vacca M, Curci A, Miano MG, Cocchia M, Mercadante G, Torino A, D'Urso M, Ciccodicola A, D'Esposito M. Human and mouse SYBL1 gene structure and expression. Gene. 1999 Nov 15;240(1):233-8.

Mazur I, Anhlan D, Mitzner D, Wixler L, Schubert U, Ludwig S. The proapoptotic influenza A virus protein PB1-F2 regulates viral polymerase activity by interaction with the PB1 protein. Cell Microbiol. 2008 May;10(5):1140-52.

Mège RM, Gavard J, Lambert M. Regulation of cell-cell junctions by the cytoskeleton. Curr Opin Cell Biol. 2006 Oct;18(5):541-8.

McKeown SJ, Wallace AS, Anderson RB. Expression and function of cell adhesion molecules during neural crest migration. Dev Biol. 2013 Jan 15;373(2):244-57.

Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol. 1987 Aug 5;196(3):641-56.

Mitzner D, Dudek SE, Studtrucker N, Anhlan D, Mazur I, Wissing J, Jänsch L, Wixler L, Bruns K, Sharma A, Wray V, Henklein P, Ludwig S, Schubert U. Phosphorylation of the influenza A virus protein PB1-F2 by PKC is crucial for apoptosis promoting functions in monocytes. Cell Microbiol. 2009 Oct;11(10):1502-16.

Mobley DL, Dill KA, Chodera JD. Treating entropy and conformational changes in implicit solvent simulations of small molecules. J Phys Chem B. 2008 Jan 24;112(3):938-46.

Morlot C, Thielens NM, Ravelli RB, Hemrika W, Romijn RA, Gros P, Cusack S, McCarthy AA. Structural insights into the Slit-Robo complex. Proc Natl Acad Sci US A. 2007 Sep 18;104(38):14923-8.

Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013 Jan;10(1):47-53. doi: 10.1038/nmeth.2289.

Mosyak L, Wood A, Dwyer B, Buddha M, Johnson M, Aulabaugh A, Zhong X, Presman E, Benard S, Kelleher K, Wilhelm J, Stahl ML, Kriz R, Gao Y, Cao Z, Ling HP, Pangalos MN, Walsh FS, Somers WS. The structure of the Lingo-1 ectodomain, a module implicated in central nervous system repair inhibition. J Biol Chem. 2006 Nov 24;281(47):36378-90.

Munster VJ, Baas C, Lexmond P, Waldenström J, Wallensten A, Fransson T, Rimmelzwaan GF, Beyer WE, Schutten M, Olsen B, Osterhaus AD, Fouchier RA. Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds. PLoS Pathog. 2007 May 11;3(5):e61.

Nguyen Ba-Charvet KT, Brose K, Ma L, Wang KH, Marillat V, Sotelo C, Tessier-Lavigne M, Chédotal A. Diversity and specificity of actions of Slit2 proteolytic fragments in axon guidance. J Neurosci. 2001 Jun 15;21(12):4281-9.

Nick Pace C, Scholtz JM, Grimsley GR. Forces stabilizing proteins. FEBS Lett. 2014 Jun 27;588(14):2177-84.

Nimrod G, Glaser F, Steinberg D, Ben-Tal N, Pupko T. In silico identification of functional regions in proteins. Bioinformatics. 2005 Jun;21 Suppl 1:i328-37.

Nooren IM, Thornton JM. Diversity of protein-protein interactions. EMBO J. 2003 Jul 15;22(14):3486-92.
Peng Y, Zou Y, Li H, Li K, Jiang T. Inferring the antigenic epitopes for highly pathogenic avian influenza H5N1 viruses. Vaccine. 2014 Feb 3;32(6):671-6.

O'Neill RE, Talon J, Palese P. The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. EMBO J. 1998 Jan 2;17(1):288-96.

Ouzounis C, Pérez-Irratxeta C, Sander C, Valencia A. Are binding residues conserved? Pac Symp Biocomput. 1998:401-12.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605-12.

Pflug A, Guilligay D, Reich S, Cusack S. Structure of influenza A polymerase bound to the viral RNA promoter. Nature. 2014 Dec 18;516(7531):355-60.

Piovesan D, Minervini G, Tosatto SC. The RING 2.0 web server for high quality residue interaction networks. Nucleic Acids Res. 2016 Jul 8;44(W1):W367-74.

Portela A, Digard P. The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. J Gen Virol. 2002 Apr;83(Pt 4):723-34.

Pryor PR, Jackson L, Gray SR, Edeling MA, Thompson A, Sanderson CM, Evans PR, Owen DJ, Luzio JP. Molecular basis for the sorting of the SNARE VAMP7 into endocytic clathrin-coated vesicles by the ArfGAP Hrb. Cell. 2008 Sep 5;134(5):817-27.

Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade RC. webPIPSA: a web server for the comparison of protein interaction properties. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W276-80.

Rigden DJ. From protein structure to function with bioinformatics. Springer 2009

Ritchie AW, Webb LJ. Understanding and Manipulating Electrostatic Fields at the Protein-Protein Interface Using Vibrational Spectroscopy and Continuum Electrostatics Calculations. J Phys Chem B. 2015 Nov 5;119(44):13945-57.

Rott R, Klenk HD. Significance of viral glycoproteins for infectivity and pathogenicity. Zentralbl Bakteriol Mikrobiol Hyg A. 1987 Aug;266(1-2):145-54.

Sakurai T. The role of NrCAM in neural development and disorders--beyond a simple glue in the brain. Mol Cell Neurosci. 2012 Mar;49(3):351-63.

Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993 Dec 5;234(3):779-815.

Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol. 1998 Feb 6;275(5):895-916.

Sanders CR. Biomolecular Ligand-receptor Binding Studies: theory, practice, and analysis. Vanderbilt University, 2010. pdfs.semanticscholar.org.

Sasaki T, Aoi H, Oga T, Fujita I, Ichinohe N. Postnatal development of dendritic structure of layer III pyramidal neurons in the medial prefrontal cortex of marmoset. Brain Struct Funct. 2015 Nov;220(6):3245-58.

Scapin G, Salice P, Tescari S, Menna E, De Filippis V, Filippini F. Enhanced neuronal cell differentiation combining biomimetic peptides and a carbon nanotube-polymer scaffold. Nanomedicine. 2015 Apr;11(3):621-32.

Scapin G, Bertalot T, Vicentini N, Gatti T, Tescari S, De Filippis V, Marega C, Menna E, Gasparella M, Parnigotto PP, Di Liddo R, Filippini F. Neuronal commitment of human circulating multipotent cells by carbon nanotube-polymer scaffolds and biomimetic peptides. Nanomedicine (Lond). 2016 Aug;11(15):1929-46.

Schäfer IB, Hesketh GG, Bright NA, Gray SR, Pryor PR, Evans PR, Luzio JP, Owen DJ. The binding of Varp to VAMP7 traps VAMP7 in a closed, fusogenically inactive conformation. Nat Struct Mol Biol. 2012 Dec;19(12):1300-9.

Schauperl M, Podewitz M, Waldner BJ, Liedl KR. Enthalpic and Entropic Contributions to Hydrophobicity. J Chem Theory Comput. 2016 Sep 13;12(9):4600-10.

Schnell JR, Chou JJ. Structure and mechanism of the M2 proton channel of influenza A virus. Nature. 2008 Jan 31;451(7178):591-5.

Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins. 2001 Sep 1;44(4):400-17.

Sha B, Luo M. Structure of a bifunctional membrane-RNA binding protein, influenza virus matrix protein M1. Nat Struct Biol. 1997 Mar;4(3):239-44.

Sha B, Luo M. Crystallization and preliminary X-ray crystallographic studies of type A influenza virus matrix protein M1. Acta Crystallogr D Biol Crystallogr.1997 Jul 1;53(Pt 4):458-60.

Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res.1987 Feb 11;15(3):1281-95.

Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol. 2000 Apr;10(2):153-9.

Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006 Nov;15(11):2507-24.

Shimoda Y, Watanabe K. Contactins: emerging key roles in the development and function of the nervous system. Cell Adh Migr. 2009 Jan-Mar;3(1):64-70.

Shinya K, Hamm S, Hatta M, Ito H, Ito T, Kawaoka Y. PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice. Virology. 2004 Mar 15;320(2):258-66.

Sitkoff D, Sharp KA, Honig B. Accurate calculation of free energy using macroscopic solvent models. The Journal of Physical Chemistry 1994 98 (7), 1978-1988.

Skehel JJ, Schild GC. The polypeptide composition of influenza A viruses. Virology. 1971 May;44(2):396-408.

Skehel JJ, Wiley DC. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem. 2000;69:531-69.

Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci. 1997 Mar;6(3):676-88.

Soni N, Madhusudhan MS. Computational modeling of protein assemblies. Curr Opin Struct Biol. 2017 Jun;44:179-189. doi: 10.1016/j.sbi.2017.04.006. Epub 2017 May 12.

Sriwilaijaroen N, Suzuki Y. Molecular basis of the structure and function of H1 hemagglutinin of influenza virus. Proc Jpn Acad Ser B Phys Biol Sci. 2012;88(6):226-49.

Stanfield RL, Wilson IA. Protein-peptide interactions. Curr Opin Struct Biol. 1995 Feb;5(1):103-13.

Steinhauer DA. Role of hemagglutinin cleavage for the pathogenicity of influenza virus. Virology. 1999 May 25;258(1):1-20.

Stevens J, Blixt O, Glaser L, Taubenberger JK, Palese P, Paulson JC, Wilson IA. Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. J Mol Biol. 2006 Feb 3;355(5):1143-55.

Stouffer AL, Acharya R, Salom D, Levine AS, Di Costanzo L, Soto CS, Tereshko V, Nanda V, Stayrook S, DeGrado WF. Structural basis for the function and inhibition of an influenza virus proton channel. Nature. 2008 Jan 31;451(7178):596-9.

Stray SJ, Pittman LB. Subtype- and antigenic site-specific differences in biophysical influences on evolution of influenza virus hemagglutinin. Virol J. 2012 May 8;9:91.

Subhash Chandra Parija, Microbiology&Immunology, 2012, Elsevier.

Sytnyk V, Leshchyns'ka I, Schachner M. Neural Cell Adhesion Molecules of the Immunoglobulin Superfamily Regulate Synapse Formation, Maintenance, and Function. Trends Neurosci. 2017 May;40(5):295-308.

Takai Y, Ikeda W, Ogita H, Rikitake Y. The immunoglobulin-like cell adhesion molecule nectin and its associated protein afadin. Annu Rev Cell Dev Biol. 2008;24:309-42.

Taliaferro JM, Vidaki M, Oliveira R, Olson S, Zhan L, Saxena T, Wang ET, Graveley BR, Gertler FB, Swanson MS, Burge CB. Distal Alternative Last Exons Localize mRNAs to Neural Projections. Mol Cell. 2016 Mar 17;61(6):821-33.

Tamura K, Ohbayashi N, Maruta Y, Kanno E, Itoh T, Fukuda M. Varp is a novel Rab32/38-binding protein that regulates Tyrp1 trafficking in melanocytes. Mol Biol Cell. 2009 Jun;20(12):2900-8.

Tong S, Li Y, Rivailler P, Conrardy C, Castillo DA, Chen LM, Recuenco S, Ellison JA, Davis CT, York IA, Turmelle AS, Moran D, Rogers S, Shi M, Tao Y, Weil MR, Tang K, Rowe LA, Sammons S, Xu X, Frace M, Lindblade KA, Cox NJ, Anderson LJ, Rupprecht CE, Donis RO. A distinct lineage of influenza A virus from bats. Proc Natl Acad Sci U S A. 2012 Mar 13;109(11):4269-74.

Treanor J. Influenza vaccine--outmaneuvering antigenic shift and drift. N Engl J Med. 2004 Jan 15;350(3):218-20.

Tuszynska I, Magnus M, Jonak K, Dawson W, Bujnicki JM. NPDock: a web server for protein-nucleic acid docking. Nucleic Acids Res. 2015 Jul 1;43(W1):W425-30.

Ueda Y, Tanaka M, Kyan Y, Yoshida M, Sasahara K, Shinya K. PB1-F2 amino acids regulate influenza A viral polymerase activity. Journal of Basic & Applied Sciences 2014 10: 1-6.

Ul-Haq DZ and Madura JD. Frontiers in Computational Chemistry: Volume 2: Computer applications for drug design and biomolecular systems. Elsevier 2015.

Vacca M, Albania L, Della Ragione F, Carpi A, Rossi V, Strazzullo M, De Franceschi N, Rossetto O, Filippini F, D'Esposito M. Alternative splicing of the human gene SYBL1 modulates protein domain architecture of Longin VAMP7/TI-VAMP, showing both non-SNARE and synaptobrevin-like isoforms. BMC Mol Biol. 2011 May 24;12:26.

Vandegrift KJ, Sokolow SH, Daszak P, Kilpatrick AM. Ecology of avian influenza viruses in a changing world. Ann N Y Acad Sci. 2010 May;1195:113-28.

Vicatos S, Roca M, Warshel A. Effective approach for calculations of absolute stability of proteins using focused dielectric constants. Proteins. 2009 Nov 15;77(3):670-84.

Vivona S, Liu CW, Strop P, Rossi V, Filippini F, Brunger AT. The longin SNARE VAMP7/TI-VAMP adopts a closed conformation. J Biol Chem. 2010 Jun 4;285(23):17965-73.

Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph. 1990 Mar;8(1):52-6, 29.

Wakefield L, Brownlee GG. RNA-binding properties of influenza A virus matrix protein M1. Nucleic Acids Res. 1989 Nov 11;17(21):8569-80.

Warshel A, Sharma PK, Kato M, Parson WW. Modeling electrostatic effects in proteins. Biochim Biophys Acta. 2006 Nov;1764(11):1647-76. Epub 2006 Aug 25.

Webster RG, Rott R. Influenza virus A pathogenicity: the pivotal role of hemagglutinin. Cell. 1987 Aug 28;50(5):665-6.

Weis W, Brown JH, Cusack S, Paulson JC, Skehel JJ, Wiley DC. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. Nature. 1988 Jun 2;333(6172):426-31

Wiley DC, Wilson IA, Skehel JJ. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature. 1981 Jan 29;289(5796):373-8.

Winter G, Fields S. The structure of the gene encoding the nucleoprotein of human influenza virus A/PR/8/34. Virology. 1981 Oct 30;114(2):423-8.

Wu W, Air GM. Binding of influenza viruses to sialic acids: reassortant viruses with A/NWS/33 hemagglutinin bind to alpha2,8-linked sialic acid. Virology. 2004 Aug 1;325(2):340-50. Erratum in: Virology. 2004 Nov 10;329(1):213-4.

Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins. 2012 Jul;80(7):1715-35.

Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. Bioinformatics. 2016 Dec 1;32(23):3676-3678.

Yamada H, Chounan R, Higashi Y, Kurihara N, Kido H. Mitochondrial targeting sequence of the influenza A virus PB1-F2 protein and its function in mitochondria. FEBS Lett. 2004 Dec 17;578(3):331-6.

Zamarin D, García-Sastre A, Xiao X, Wang R, Palese P. Influenza virus PB1-F2 protein induces cell death through mitochondrial ANT3 and VDAC1. PLoS Pathog. 2005 Sep;1(1):e4.

Zhang L. CRASH syndrome: does it teach us about neurotrophic functions of cell adhesion molecules? Neuroscientist. 2010 Aug;16(4):470-4.

Zhang Y, Yeh J, Richardson PM, Bo X. Cell adhesion molecules of the immunoglobulin superfamily in axonal regeneration and neural repair. Restor Neurol Neurosci. 2008;26(2-3):81-96.

Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J. 2003 Aug;85(2):1145-64.

Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem. 2004 Apr 30;25(6):865-71.

Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A. 2004 May 18;101(20):7594-9.

Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A. 2005 Jan 25;102(4):1029-34.

Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005 Apr 22;33(7):2302-9.

Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008 Jan 23;9:40.

Zhao X, Yip PM, Siu CH. Identification of a homophilic binding site in immunoglobulin-like domain 2 of the cell adhesion molecule L1. J Neurochem. 1998 Sep;71(3):960-71.

Zhou ZD, Sathiyamoorthy S, Tan EK. LINGO-1 and Neurodegeneration: Pathophysiologic Clues for Essential Tremor. Tremor Other Hyperkinet Mov (N Y). 2012;2. pii: tre-02-51-249-1.

Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002 Nov;11(11):2714-26. Erratum in: Protein Sci. 2003 Sep;12(9):2121.

Zoabi M, Nadar-Ponniah PT, Khoury-Haddad H, Usaj M, Budowski-Tal I, Haran T, Henn A, Mandel-Gutfreund Y, Ayoub N. RNA-dependent chromatin localization of KDM4D lysine demethylase promotes H3K9me3 demethylation. Nucleic Acids Res. 2014 Dec 1;42(21):13026-38.